INSIGHTS IN LIFE-COURSE EPIDEMIOLOGY AND SOCIAL INEQUALITIES: 2021

EDITED BY: Cyrille Delpierre and Hilde Langseth

PUBLISHED IN: Frontiers in Public Health

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-905-6
DOI 10.3389/978-2-88976-905-6

INSIGHTS IN LIFE-COURSE EPIDEMIOLOGY AND SOCIAL INEQUALITIES: 2021

Topic Editors:
Cyrille Delpierre, INSERM U1027 Epidémiologie et analyses en santé publique: Risques, Maladies Chroniques et Handicap, France
Hilde Langseth, Cancer Registry of Norway, Norway

Citation: Delpierre, C., Langseth, H., eds. (2022). Insights in Life-course Epidemiology and Social Inequalities: 2021. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-905-6

Table of Contents

frontiers | Frontiers in Public Health

TYPE Editorial
PUBLISHED 29 July 2022
DOI 10.3389/fpubh.2022.980547

Editorial: Insights in life-course epidemiology and social inequalities: 2021

Cyrille Delpierre[1] and Hilde Langseth[2,3]*

[1]CERPOP, UMR 1295, Inserm, Paul Sabatier University, UPS, Toulouse, France, [2]Department of Research, Cancer Registry of Norway, Oslo, Norway, [3]Department of Epidemiology and Biostatistics, Imperial College London, London, United Kingdom

KEYWORDS

social determinants, quality data, methodological approaches, life-course epidemiology, social inequalities

Editorial on the Research Topic
Insights in life-course epidemiology and social inequalities: 2021

OPEN ACCESS

EDITED AND REVIEWED BY
MinJae Lee,
University of Texas Southwestern Medical Center, United States

*CORRESPONDENCE
Hilde Langseth
hilde.langseth@kreftregisteret.no

SPECIALTY SECTION
This article was submitted to Life-Course Epidemiology and Social Inequalities in Health, a section of the journal Frontiers in Public Health

RECEIVED 28 June 2022
ACCEPTED 18 July 2022
PUBLISHED 29 July 2022

CITATION
Delpierre C and Langseth H (2022) Editorial: Insights in life-course epidemiology and social inequalities: 2021. *Front. Public Health* 10:980547. doi: 10.3389/fpubh.2022.980547

The aim of this Research Topic was to focus on important research challenges in the field of life course epidemiology and social inequalities in health. This volume gathers thirteen papers dealing with aspects that represent current and future challenges in this field of research. More specifically (1) determinants beyond social position and in new contexts, (2) access to high quality data and biospecimens and how to share harmonized datasets, (3) methodological approaches to analyse complex datasets from different sources.

Which social and living conditions that may act as a mechanism to explain the construction of social inequalities over the life course is an important challenge. The paper by Dong et al. highlights the importance of focus on childhood conditions, in particular child malnutrition, which is common in developing countries, as well as developed countries. This is particularly in challenging in context of the global warming. Based on more than 13,000 elderly Chinese people aged 65–99 years, the authors show that childhood starvation is associated with socioeconomic determinants (age, gender, residency, education level, income level) and find a persistent negative cumulative effect of childhood starvation on the quantity and quality of life. Haugland et al. reminds us of the importance of Adverse Childhood Experiences (ACEs) as exposures that should be more targeted by public health strategies. This study including 28,047 adults, shows that the prevalence of ACEs (family conflict, lack of adult support, struggle with bad memories, and difficult childhood) varies with socio-demographic factors (age, gender, marital status, and history of divorced parents) and that exposure to ACEs is associated with low socio-economic status in adulthood (low educational attainment, perceived financial difficulties, receipt of social benefits).

The influence of ethnicity on health is a well-known but the drivers behind this remain less understood. The systematic review by Rubin et al. examines the potential role of genetics by analyzing the research that has been published in the US about the genetic factors of Alzheimer's Disease and Related Dementias (ADRDs) among racial/ethnic minorities, which are disproportionality affected. This review of 66 articles highlights that well-established ADRD genetic risk factors for Caucasian populations have not been studied to the same degree in minority U.S. populations which are underrepresented. The study of Athavale et al. analyse the social conditions associated with ethnicity and show that infection and mortality due to COVID-19 infection among Non-Hispanic Black (NHB) and Hispanics was considerably higher than Non-Hispanic White (NHW) mainly because of social unfavorable conditions more likely to concern racial minorities poverty.

Social inequalities in health are not just an issue in high-income countries. The study of Akokuwebe et al. highlights the need to analyze social inequalities in low-income countries as well. The increased burden of non-communicable disease and the double burden of malnutrition (undernutrition and over-nutrition) in low-income countries due to the epidemiological transition have become a public health concern, which has not received the attention that this problem deserves. This study established the synchrony of a double burden of underweight and overweight/obesity among 3,263 women of reproductive age in South Africa, although the prevalence of underweight was declining, while overweight/obesity increased significantly over the study period. It also highlights the influence of social determinants (age, marital status, education, employment status, wealth, ethnicity, and residence), with more advantaged people more likely to be overweight/obese and people from rural areas or of non-African/Black ethnicity more likely to be underweight. Farias et al. describes the temporal trend of stomach cancer mortality in Brazil which is a middle-income country characterized by great internal socioeconomic heterogeneity. The study shows a decline in stomach cancer over time with periods of variation similar to the behavior observed in both high and low-income countries. The findings point to the need of understanding the behavior of stomach cancer mortality in different geographical regions, since they present different socioeconomic characteristics.

Investigating the complex interplay between life-course exposures and disease requires access to high quality data and biospecimens. In the last decades there has been large initiatives in establishing cohort studies across the world, however the efforts to make the collected data available to the scientific community has been sparse. In a data report by Rodriguesz-Laso et al. they provided a map of initiatives that harmonize patient cohorts across the world. Most initiatives are partnered with universities, hospitals and research institutions. The paper focus on the strengths of integration of cohort studies to take the advantage of already collected information to increase the sample size in studies of uncommon exposures, rare diseases, less strong associations, or very restricted populations like in personalized medicine. O'Leary et al. write about development of a multi-study repository to support research on veteran health. The study aimed to describe the selection of studies included in the repository, the design of metadata-driven architecture for secure storing and tracking of data and biospecimens and development of a process to review the scientific and ethical merit of data and specimen request. This is a good example of the importance of using data and biospecimens from several cohorts to get sufficient statistical power to study rare exposures. This multi-study repository provides a structure that can be used to support the sharing of data and specimens across multiple content areas for different types of research studies.

Research in life course epidemiology also raises important methodological issues. In particular, the question of causality is a major challenge in observational studies. Three studies use Mendelian randomization (MR) to analyze causal effects of their exposure on health outcomes. The study of Probst-Hensch et al. examines the effect of BMI on lung function (LF). A negative causal BMI LF effect is observed with a stronger effect for childhood BMI highlighting the importance of a life course perspective in studies using MR method. The two other studies use MR to analyze the causal effect of education on health. The study of Yoshikawa et al. which is one of the first investigating the association between education and COVID-19 severity, shows that education is associated with a lower risk of COVID-19 severity. The study of Wang et al. investigates the causal effect of education on 14 urological and reproductive health outcomes. Education is associated with a higher or lower risk according to health outcomes. However, in both studies, the mechanisms underlying the associations found are unknown and unexplored. Furthermore, the SNPs used as instrumental variables for education are different between the two studies, raising questions that deserve further investigation about how best to use (if relevant) MR for analyzing social traits. Other approaches may be relevant especially when we are interested in mechanisms and multiple mediators may exist, a situation that is not uncommon in life course epidemiology. Tai et al. proposes a method using G-computation algorithm to conduct causal mediation analysis in the presence of multiple ordered mediators. Their approach is powerful and versatile for settings with multiple mediators. An application of the method is proposed to investigate the mediating role of early and late hepatitis B virus (HBV) viral load in the effect of hepatitis C virus (HCV) infection on hepatocellular carcinoma (HCC). Another methodological issue is no longer causality but prediction, in particular how best to predict a disease using the large amount of data available in the most relevant way. Lufkin et al. proposes a Bayesian regression model to characterize the risk of Rheumatoid Arthritis (RA) from common comorbidities, demographic, socioeconomic, and behavioral factors that are known to associate with RA. The model demonstrates a high predictive accuracy in comparison with other models reported in the literature and model is able to identify important second- and third-order interactions between the risk factors, which may have important clinical relevance and stimulate further research to understand the mechanisms underlying such interactions.

In summary we hope that the papers put together in this Research Topic will be helpful and raise awareness on important scientific challenges and opportunities in future life course epidemiology research.

Author contributions

CD and HL contributed as monitoring editors for this Research Topic. Both authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Public Health

ORIGINAL RESEARCH
published: 17 February 2021
doi: 10.3389/fpubh.2021.612806

Development of a Multi-Study Repository to Support Research on Veteran Health: The VA Cooperative Studies Program Epidemiology Center-Durham (CSPEC-Durham) Data and Specimen Repository

Meghan C. O'Leary[1]*, *R. Lawrence Whitley*[2], *Ashlyn Press*[1], *Dawn Provenzale*[1,3], *Christina D. Williams*[1,3], *Blair Chesnut*[1,4], *Rodney Jones*[1,4], *Thomas S. Redding IV*[1] *and Kellie J. Sims*[1]*

[1] Cooperative Studies Program Epidemiology Center-Durham, Durham Veterans Affairs Health Care System, Durham, NC, United States, [2] RTI International, Research Triangle Park, NC, United States, [3] Duke University Medical Center, Durham, NC, United States, [4] Duke Molecular Physiology Institute, School of Medicine, Duke University, Durham, NC, United States

OPEN ACCESS

Edited by:
Randi Elin Gislefoss,
Cancer Registry of Norway, Norway

Reviewed by:
Tony Kuo,
UCLA Fielding School of Public Health, United States
Cláudia De Freitas,
University of Porto, Portugal
Karina Standahl Olsen,
Arctic University of Norway, Norway

***Correspondence:**
Meghan C. O'Leary
meghan.o'leary@va.gov
Kellie J. Sims
kellie.sims@va.gov

Specialty section:
This article was submitted to Life-Course Epidemiology and Social Inequalities in Health, a section of the journal Frontiers in Public Health

Received: 30 September 2020
Accepted: 13 January 2021
Published: 17 February 2021

Citation:
O'Leary MC, Whitley RL, Press A, Provenzale D, Williams CD, Chesnut B, Jones R, Redding TS IV and Sims KJ (2021) Development of a Multi-Study Repository to Support Research on Veteran Health: The VA Cooperative Studies Program Epidemiology Center-Durham (CSPEC-Durham) Data and Specimen Repository. Front. Public Health 9:612806. doi: 10.3389/fpubh.2021.612806

Federal agencies, including the Department of Veterans Affairs (VA), have prioritized improved access to scientific data and results collected through federally funded research. Our VA Cooperative Studies Program Epidemiology Center in Durham, North Carolina (CSPEC-Durham) assembled a repository of data and specimens collected through multiple studies on Veteran health issues to facilitate future research in these areas. We developed a single protocol, request process that includes scientific and ethical review of all applications, and a database architecture using metadata (common variable descriptors) to securely store and share data across diverse studies. In addition, we created a mechanism to allow data and specimens collected through older studies in which re-use was not addressed in the study protocol or consent forms to be shared if the future research is within the scope of the original consent. Our CSPEC-Durham Data and Specimen Repository currently includes research data, genomic data, and study specimens (e.g., DNA, blood) for three content areas: colorectal cancer, amyotrophic lateral sclerosis, and Gulf War research. The linking of the study specimens and research data can support additional genetic analyses and related research to improve Veterans' health.

Keywords: repository, data sharing, veteran, biospecimen, genomics

INTRODUCTION

Biobanking encompasses all procedures needed to collect, process, store, and share specimens collected from human subjects as well as the policies that govern these activities (1). Stored biospecimens along with linked clinical and research data can and have been used to advance translational and population health research and support personalized medicine within clinical care (1, 2). The development and maintenance of data and specimen repositories commonly involves substantial resources, including dedicated staff, laboratory space and equipment, creation of standard operating procedures or related protocols, information technology systems, and funding (3). In addition, many ethical considerations are involved in obtaining participants' informed consent to use and share their data and specimens through biobanks and other repositories; examples include determining the appropriate type of consent to use and ensuring participants understand all procedures and potential risks and benefits (1–4). Despite these challenges, prior research has shown that participants are generally willing to have their information shared for future research purposes (5, 6). Potential benefits of data and specimen sharing include increasing efficiency of limited research resources, minimizing the burden of research participants and potential risks of research participation, and contributing to more generalizable knowledge intended to improve patient health and care (7).

Among federal agencies, increased transparency of and access to federally funded research results and scientific data have been prioritized over the past decade. In February 2013, the White House's Office of Science and Technology Policy issued a memorandum requiring federal agencies to make the results of their research collected with federal funds publicly available to support future research and innovations (8). This directive also required agencies to make their scientific data available to the public to the extent possible (8). In response, the U.S. Department of Veterans Affairs (VA) issued guidance on increasing public access to research data while continuing to protect the privacy of its Veteran patients. Beginning in December 2015, VA researchers were required to submit written data management plans with their protocols outlining how the data would be made available and describing the mechanisms for ensuring privacy, confidentiality, and long-term preservation and storage of the data (9).

As of July 2020, the VA's Office of Research and Development (ORD) included data and specimen sharing within two of its three strategic priorities for VA research (10). These priorities include ensuring that research findings are translated into clinical applications that improve the care of Veterans, and facilitating larger-scale research that can benefit Veterans and the general public (10). Noted activities to achieve these goals include the curation of linked and standardized data sources and the collection of biospecimens for genomic analysis (10).

The VA Cooperative Studies Program Epidemiology Center located in Durham, North Carolina (CSPEC-Durham) is one of many research programs under VA ORD oversight (11). We aimed to develop a repository to enable data and specimen sharing that was consistent with VA and ORD guidance and priorities and that would support additional epidemiologic and genomic research specific to the health needs of Veterans.

In this paper, we describe our center's process for developing a repository of research data and biological specimens collected from Veterans with and without chronic disease for sharing with investigators with approved research protocols. The CSPEC-Durham Data and Specimen Repository, subsequently referred to as the CSPEC-Durham Repository, houses data and specimens collected from multiple research studies on diverse Veteran health issues for the purpose of facilitating future research intended to improve the health of Veterans. The purpose of this report is to describe: (1) the selection of studies included in this repository, (2) the design of metadata-driven architecture for securely storing and tracking data and specimens, and (3) the development of a process to review the scientific and ethical merit of data and specimen requests.

MATERIALS AND METHODS

Identifying Feeder Studies and Potential Sharing Restrictions

We first identified all studies conducted by members of the CSPEC-Durham research team for possible inclusion in the repository. These studies were evaluated as potential feeder studies, defined as individual research studies with a protocol approved by an Institutional Review Board (IRB) for which the collected data and, if applicable, specimens would be stored and available for sharing through the CSPEC-Durham Repository. We considered active studies with data collection and analysis still in progress, as well as legacy studies for which data collection and analysis had already been completed. Each research study focused on Veteran health issues and enrolled all or predominantly Veteran participants. We included studies that addressed different types of chronic disease areas affecting Veterans, as well as studies that enrolled Veterans with or without a particular illness to support research on risk factors, early detection, and progression of these illnesses.

We then developed and implemented a formal process for determining whether each study's data and specimens could be shared for future research, and if there were any restrictions on data and specimen sharing (**Figure 1**). Following guidance from our local IRB, all IRB-approved versions of the study protocol, informed consent form (ICF), Health Insurance Portability and Accountability Act (HIPAA) authorization, ICF waiver, and/or HIPAA waiver were obtained for each feeder study. Two CSPEC-Durham study coordinators reviewed these study documents and documented their findings related to sharing permissions. They recorded whether the study participants had previously consented to the use of their data and specimens for future research. If the participants agreed to future sharing, the reviewers documented any restrictions; e.g., only sharing the data and specimens with researchers within the VA, or for particular research questions (e.g., future research on the causes or treatment of the disease only). The reviewers also noted whether study participants had consented to be re-contacted for future research studies.

Since we included older legacy studies in our repository, some studies did not explicitly address use of the data and specimens for future research in the study consent forms or other documents. For example, a study evaluating the prevalence of colorectal cancer in an average-risk cohort recruited all Veteran participants from 1994 to 1997; the study's ICF was developed prior to the enactment of HIPAA in 1996 (12), and therefore did not include language about the future use of participants' data.

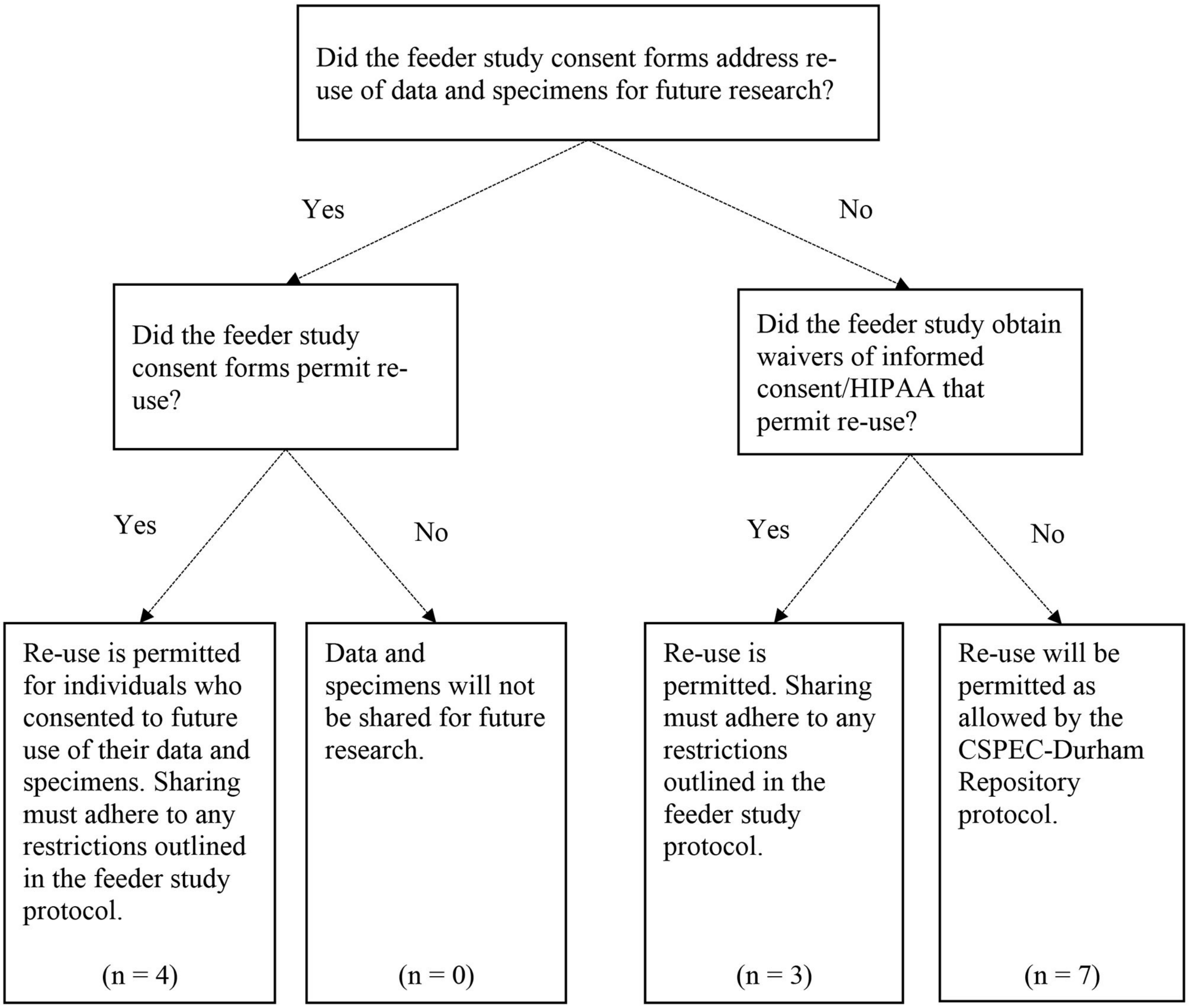

Did the feeder study consent forms address re-use of data and specimens for future research?
Yes
No
Did the feeder study consent forms permit re-use?
Did the feeder study obtain waivers of informed consent/HIPAA that permit re-use?
Yes
No
Yes
No
Re-use is permitted for individuals who consented to future use of their data and specimens. Sharing must adhere to any restrictions outlined in the feeder study protocol.
(n = 4)
Data and specimens will not be shared for future research.
(n = 0)
Re-use is permitted. Sharing must adhere to any restrictions outlined in the feeder study protocol.
(n = 3)
Re-use will be permitted as allowed by the CSPEC-Durham Repository protocol.
(n = 7)

FIGURE 1 | Data and specimen re-use decision points for the CSPEC-Durham Repository. The CSPEC-Durham Repository includes $N = 15$ total feeder studies; however, one feeder study uses administrative data only and does not involve re-use of Veteran data and/or specimens.

Studies in which data and specimen sharing was not explicitly addressed were noted in the review document. The Durham VA IRB approved the inclusion of these legacy studies in our repository and the future sharing of data and specimens if the future research to be conducted was within the scope of the original consent.

We created a study protocol, as well as standard operating procedures, to outline the administration of the repository, types of data and specimens to be shared, data access, methods of data and specimen storage and transfer, and mechanisms for protecting the participants' identities and information. In this protocol, we identified all feeder studies, and categorized each feeder study based on the extent to which it permitted re-use of study data and specimens and/or future re-contact of study participants. In total, we evaluated 15 feeder studies for potential inclusion in our repository, and all 15 studies met our criteria for inclusion, although these studies varied in their restrictions for how data and specimens can be re-used. Of these 15 studies, six were active studies and nine were legacy studies. The CSPEC-Durham Repository protocol was approved by the Durham VA Health Care System IRB in August 2016.

Database Development and Request Tracking

We designed our repository database to be structured around metadata (i.e., common set of variable descriptors applicable for any study). The metadata-driven architecture is used to manage the data and specimens across all feeder studies and to support the sharing of these data and specimens for future use by approved investigators. While the specific variables differ by study, the metadata across all feeder studies include items such as the dictionary ID, variable label, value description, and data types.

TABLE 1 | Common variable descriptors used across feeder studies and examples by feeder study.

Descriptors	
Variable Name:	Name of the variable
Dictionary ID:	Variable ID number
Variable Label:	A short description of the variable; the variable label only appears when no survey question is available
Survey Question:	Number and text of survey question from which variable is derived; the survey question only appears when it is available
Value Descriptions:	Description of possible values for categorical data
Value Min:	Minimum value possible
Value Max:	Maximum value possible
Data Types:	Data types, listed for Generic, SAS, R, SQL, C#, and XML
Is Nullable:	If true (i.e., is Nullable = 1), a null value is possible
Form Section:	Section on form or survey from which variable is derived
Table Name:	Name of database table from which variable is derived

Variable descriptor*	**Colorectal cancer**	**Amyotrophic lateral sclerosis**	**Gulf War research**
Variable name	Colonoscopy	Speech	Act mod days
Dictionary ID	110238	120294	100694
Variable label		Speech	
Survey question	15. Have you ever had a colonoscopy (tube with a light inserted into colon after you are given medicine to make you sleepy)?		18. On how many days did you engage in moderate physical activity (like a brisk walk) in the last 7 days?
Variable description		Please indicate the category that most describes your current state of health: Speech	
Value descriptions	1 = Yes 2 = No	0 = Loss of usual speech 1 = Speech combined with non-vocal communication 2 = Intelligible with repeating 3 = Detectable speech disturbance 4 = Normal speech processes	
Value min			0
Value max			7
Data types	Generic [integer], SAS [4.], R [int], SQL [tinyint], C# [Byte?], XML [xsd:integer]	Generic [integer], SAS [6], R [int], SQL [smallint], C# [Int16?], XML [xsd:integer]	Generic [integer], SAS [4.], R [int], SQL [tinyint], C# [Byte?], XML [va:tinyint]
Is nullable	True	True	True
Form section	Form 01 Clinic Survey Form—Medical History	Veterans ALS Registry Questionnaire—B. ALS Functional Rating Scale	Baseline survey—lifestyle and activities
Table name	AllForm01	ALS Questionnaire	Survey Parent

**The variable descriptors used are survey item/question dependent and, therefore, some fields are blank for particular variables of each study.*

As shown in **Table 1**, using metadata allows us to share the same types of data across feeder studies of diverse topics and designs.

The common set of variable descriptors are used to generate application code for data entry and validation, creation of data dictionaries, and data extracts used to fulfill specific data sharing requests. The use of metadata was intended to eliminate repetitive and error-prone manual steps, to ensure data provenance, and to create a common structure despite differences in the types of feeder studies. Since requestors typically only need access to a subset of the data collected for a particular feeder study for their own analyses, the use of metadata allows us to create individualized data dictionaries for each requestor and to track all transfers of data to each requestor. We used a similar process to facilitate specimen sharing; common descriptors across feeder study specimens were used to develop specimen-specific applications, inventories, and shipping manifests.

The metadata tables are stored in a relational database and data maintenance history logs, data extract snapshots, and histories of all source code are retained. The repository data are stored on a Microsoft SQL Server behind VA firewalls and access is controlled through active directory security groups for study-specific IRB-approved personnel. Microsoft SQL Server Management Studio is used to work with the data (e.g., data updates, data pulls for sharing). While metadata for the feeder studies are comingled, each feeder study's data are stored in a separate database (with access controlled by the IRB staff list),

TABLE 2 | Sample terms of agreement for data use agreements (DUAs) and material transfer agreements (MTAs).

Topic	Terms of agreement
DUA terms of agreement	
General	The Requestor represents that CSP Data will be used solely for the purpose of the Study as specified.
Data ownership	The Requestor is designated as Custodian of the CSP Data provided under this Agreement and does not own the data.
Data management	The Requestor affirms that the requested CSP Data is the minimum necessary to achieve Study goals involving CSP Data.
Unauthorized disclosure	The Requestor shall immediately report any use or disclosure of CSP Data not provided for in this Agreement or any non-compliance with this Agreement to the CSP Center Contact.
Institution approvals	CSP will be provided with written evidence of the IRB determination before release of CSP Data.
Products	The Requestor shall present any product resulting from the CSP Data in aggregated form.
MTA terms of agreement	
Research materials	The Research Materials will only be used for research purposes by the Recipient of the Biological Materials in his/her laboratory, for the research project described under suitable containment conditions.
Commercialization	The Human Biological Materials shall not be used for any commercial purposes, including selling, commercial screening, or transfer of the Human Biological Materials to a third party for commercial purposes.
Data management	The Recipient agrees to retain control over this Material and further agrees not to transfer the Material to other people not under his or her direct supervision without advance written approval of the Provider.
Intellectual property	The Recipient acquires no intellectual property as a result of the transfer of the Materials identified under this Agreement.

CSP, Cooperative Studies Program.

and there is no comingling of the feeder study data. We adhere to all VA directives about how to securely store and work with Veteran data.

We also created a Research Electronic Data Capture (REDCap) database (13) to track the study documentation for all researchers who submit a formal application to use data and specimens from the CSPEC-Durham Repository. The study documentation includes the requestor's contact information, application materials, and IRB approval letters; evaluations, scores, and recommendations for each request; dates of all study agreements executed; and details and dates of all data and specimen transfers. In addition, the REDCap system is used to track communications with the requestor from the initial inquiry through study completion. The REDCap database is behind VA firewalls and can only be accessed by IRB-approved study staff.

Data and Specimen Sharing

We developed a comprehensive process for reviewing requests from VA and non-VA researchers to use the data and specimens stored in the CSPEC-Durham Repository. The review process begins when an investigator submits a full application, comprised of a data and specimen request form, documentation of IRB approval, documentation of funding support, and biosketches for all co-investigators and biostatisticians. In the request form, the investigator identifies the feeder study of interest, which variables and/or biosamples are requested, and whether Veterans have been consulted in the study design, among other details.

Following receipt of a full application, we convene the CSPEC-Durham Repository's Scientific and Ethical Oversight Committee (SEOC) to review the request. For each request, the SEOC is comprised of a minimum of two content reviewers (i.e., subject matter experts) who evaluate the proposed study's scientific and ethical merit; at least one statistical reviewer who focuses primarily on the study design, statistical analysis plan, and considerations of the implications of the sample size for the proposed study; and at least one Veteran representative who considers the relevance of the research question to Veterans and the extent to which Veterans have been consulted or engaged in the study design (which is a dedicated section of the application). The Veteran representative is invited from a larger team of Veterans who take turns reviewing each request based on their availability and interest. Each reviewer is asked to independently review all materials, evaluate the request on a series of criteria using a web-based evaluation tool, and provide an overall score of the request that reflects the quality of the application and the prioritization of the specific request. The level of prioritization is particularly critical for specimen requests because there are finite amounts of most specimen types. Once the independent evaluations are completed, the SEOC reviewers and repository administrators hold a review meeting to discuss the reviewers' comments and determine if the request should be approved, approved conditionally with revisions, recommended for resubmission, or declined.

If a request is approved, the CSPEC-Durham Repository team works directly with the requestor and the requestor's institution to execute a data use agreement (DUA) and, if specimens will be used, a material transfer agreement (MTA). Data and specimens will only be shared with approved investigators once these agreements are fully executed to ensure the security of the data and specimens during transfer, storage, and analysis. The agreements specify all terms of the data and/or specimen sharing, including who will have access, methods of transfer and storage, ownership, reporting of results, and destruction or return of the data and/or specimens following study completion. Examples of these terms are presented in **Table 2**. The requested data and specimens are then securely transferred to the investigator for the approved research study.

TABLE 3 | Data and specimens collected for three primary content areas of the CSPEC-Durham Repository.

Study title	Subjects (*N*)	Eligibility criteria	Data collected	Timing of data collection	Samples collected*	Timing of specimen collection
Prospective Evaluation of Risk Factors for Large Colonic Adenomas in Asymptomatic Subjects (CSP #380)	3,121	Veterans ages 50–75 who underwent screening colonoscopies from 1994 to 1997	Results of GI exams, medical history, family history, lifestyle factors, GWAS results	Baseline: 1994–1997 5-year GI exams 10-year GI exams	Blood, tissue	Baseline: 1994–1997 Longitudinal: 1994-Present
National Registry of Veterans with Amyotrophic Lateral Sclerosis (CSP #500A)	1,225	Veterans with a verified diagnosis of ALS in 2003–2007, regardless of VA user status	ALS functional rating score, family history, lifestyle factors, use of ventilatory or feeding support, GWAS results	Baseline: 2003–2007 Every 6 months for up to 5 years	Blood	Baseline: 2003–2007
Gulf War Era Cohort and Biorepository (CSP #585)	1,274	Veterans who served between July 1990 and August 1991, regardless of deployment or VA user status	Prior exposures if deployed to the Gulf region, family history, physical and mental health, lifestyle factors, GWAS results	Baseline: 2014–2016	Blood	Baseline: 2014–2016

GI, gastrointestinal; GWAS, genome-wide association study; ALS, amyotrophic lateral sclerosis.
**DNA samples have been extracted from the blood samples for each of these feeder studies.*

Return of Derived Data

Since the objective of the CSPEC-Durham Repository is to support additional research on Veteran health, we require all approved investigators to return the data derived from their analyses to the repository. This includes analytic data derived from the study data and assay data derived from use of the study specimens. We further developed this process in August 2020 by standardizing the requirements related to the return of derived data. These requirements include returning data in a mutually agreed upon timely manner after publication of results, providing a codebook or related documentation that describes any new or collapsed variables in the analytic dataset, and, if specimens were shared, providing an assay protocol that describes how the specimens were stored and analyzed. The returned data can then be made available to other researchers for validation and subsequent analyses.

RESULTS

The CSPEC-Durham Repository includes Veteran data and specimens from 15 feeder studies with a focus on three primary disease areas: colorectal cancer, amyotrophic lateral sclerosis (ALS), and Gulf War research (**Table 3**). Seven of these 15 feeder studies relate to these three primary content areas, and each of these seven feeder studies were funded by the VA Cooperative Studies Program (CSP). While we do not currently anticipate requests for the other feeder studies, we included them as a means for long-term storage and security of the previously collected data.

For the three primary disease areas, the repository contains data and biospecimens such as Veterans' demographic, military service, healthcare utilization, and clinical data, as well as tissue and blood samples. The data and specimens were collected longitudinally at multiple time points for the first two of these three disease areas, allowing for research on disease progression and how risk factors differentially affect clinical and survival outcomes, and cross-sectionally for the third disease area. In each of these cases, the research data and specimens can be linked with the participants' VA medical records to assess longer-term clinical and survival outcomes. The ability to link the feeder study specimens with rich clinical and research data provides opportunities for genetic and molecular association analyses to inform Veteran care.

Colorectal Cancer

Asymptomatic Veterans aged 50–75 years were enrolled in the study, "Prospective Evaluation of Risk Factors for Large Colonic Adenomas in Asymptomatic Subjects," (CSP #380) between 1994 and 1997 at 13 geographically diverse VA medical centers (14). Each of the 3,121 study participants underwent a baseline screening colonoscopy as part of the study and were followed for 10 years or until death. The cohort's clinical outcomes, including prevalence of advanced colorectal neoplasia and colorectal cancer, at the time of the study (14), after 5 years (15), and after 10 years (16) were previously reported.

The study data stored in the repository includes the results of the baseline colonoscopies as well as other gastrointestinal (GI) exams completed during the longitudinal follow-up period. Sixty-one percent ($N = 1,915$) of this cohort had at least one surveillance colonoscopy within 10 years of their baseline exam (16). Survey data, including medical history, family history, and lifestyle factors, such as tobacco use, alcohol use, physical activity, and diet, are also stored.

The specimen repository includes colorectal tissues biopsied during colonoscopies and other GI exams completed as part of the study and as part of routine clinical care. These formalin-fixed paraffin-embedded (FFPE) and Bouin's-fixed tissue samples are stored in VA pathology labs until they are ready to be discarded or used for future research according to VA policy. The study team works with the local sites to retrieve these tissue

TABLE 4 | Glossary of terms.

Term	Definition
Active study	Feeder study in which collection and/or analysis of the data/specimens is occurring currently by the study team
Data use agreement	A legal document describing the terms of agreement for the transfer and use of data between the institution providing the data and the institution/investigator requesting to use the data for research purposes
Feeder study	An individual research study with an IRB-approved protocol for which the collected data and/or specimens are stored and available for sharing through our CSPEC-Durham Data and Specimen Repository
Gulf War Illness	Chronic, multi-symptom health condition affecting Veterans who served in the 1990–1991 Gulf War that is not explained by other medical diagnoses or standard laboratory tests. Common symptoms include fatigue, cognitive impairment, chronic pain, sleep problems, gastrointestinal issues, and skin problems. Multiple diagnostic definitions are used to identify cases of Gulf War Illness (21, 22)
Legacy study	Feeder study in which the collection and analyses of data/specimens have been completed by the study team
Material transfer agreement	A legal document describing the terms of agreement for the transfer and use of human biological specimens between the institution providing the specimens and the institution/investigator requesting to use the specimens for research purposes
Metadata	Common set of variable descriptors (e.g., IDs, variable labels, value descriptions, etc.) for data and specimens collected across feeder studies that is used to structure the repository database

samples and have them transferred to the Southern Arizona VA Healthcare System (SAVAHCS) in Tucson, Arizona for long-term storage. Tissue samples from some local sites may be stored temporarily at the Durham VA Health Care System for coding purposes. To date, more than 1,800 of these tissues have been added to the specimen repository governed by the CSPEC-Durham Repository and physically located at SAVAHCS. Additional tissues will be retrieved and added to the repository over time as the tissues become available for research purposes. DNA will be extracted from these tissue samples and made available in the repository as well.

The repository also includes frozen blood and tissue samples collected from 815 study participants during their baseline colonoscopy exams, and DNA extracted from these samples. Serum and lymphocytes were collected from those participants with a large polyp (i.e., at least 1 cm); serum and lymphocytes were also collected from age- and sex-matched participants with no polyps detected. Normal-appearing tissue samples and polyp tissues were biopsied from these participants and stored for future use. Each of these cross-sectionally collected specimens have been frozen since baseline and are currently stored at the Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC) in Boston, Massachusetts. A genome-wide association study (GWAS) of these DNA samples has been conducted (17), and the results will be made available through the repository.

Amyotrophic Lateral Sclerosis

The "National Registry of Veterans with Amyotrophic Lateral Sclerosis" (CSP #500A) enrolled 2,068 Veterans with an ALS diagnosis between April 2003 and September 2007 (18). Each ALS diagnosis was confirmed by a neurologist, providing information on the type of ALS diagnosis, site of onset, and date of diagnosis. Participants, who were recruited from all 50 states, self-reported their symptoms and the severity of their symptoms through phone interviews at baseline and every 6 months for up to 5 years. The ALS Functional Rating Score was used to monitor their health and functional status over time. Additional survey data included in the repository include family history, smoking status, medications, comorbidities, surgical history, and use of ventilatory or feeding support.

Each participant in the ALS Registry was asked to provide a DNA sample to be included in the study's DNA Bank for future research. More than half of the participants ($N = 1{,}168$) provided a DNA sample, most commonly by a blood sample (85% vs. 15% with a saliva sample) (18). These cross-sectional blood and DNA samples are all governed by our repository, physically stored at MAVERIC, and can be used for future research on ALS causes and treatment. The repository also contains the results of a GWAS of ALS diagnosis and survival using these samples (19).

Gulf War Research

Veterans who served during the 1990–1991 Gulf War era were enrolled in the "Gulf War Era Cohort and Biorepository" (GWECB, also referred to as CSP #585) between 2014 and 2016 (20). The goal of the GWECB was to collect data to be used for future research on diverse health concerns specific to this cohort of Veterans, including Gulf War Illness (**Table 4**). A total of 1,344 Veterans were enrolled in the GWECB, including 1,275 for whom we have survey data, health records, and a blood sample (i.e., fully enrolled) and 69 for whom we have surveys and health records. The GWECB sample reflected the geographic distribution of Veterans across the four U.S. Census regions and included Gulf War era Veterans regardless of their deployment, health status, or use of VA healthcare.

The cross-sectional survey data in the repository includes prior exposures during military service if deployed to the Gulf region, family history, physical and mental health, including the severity, frequency, and functional impact of specific conditions, and lifestyle factors, such as physical activity, tobacco use, and alcohol use. The participants also consented to be re-contacted by the GWECB team for possible participation in future studies.

The repository also includes plasma and buffy coat samples, as well as extracted DNA, for each study participant. These samples were collected at baseline. Our repository governs these specimens, which are physically stored at MAVERIC. A GWAS using these DNA samples has been conducted. The analysis of the GWAS data is in progress; the GWECB team plans to publish the results and make them available for sharing through the repository in the future. In addition, an algorithm is being developed by the GWECB team that will help to identify cases of Gulf War Illness in this cohort using the self-reported survey data; manuscripts describing the results and the methodology are forthcoming.

Data and Specimen Requests

As of December 2020, we have received 10 formal applications to use data and/or specimens from the CSPEC-Durham Repository. Seven of the 10 applications were approved for data and/or specimen sharing following the scientific and ethical review process. Of these 7 approved requests, 3 requests have been fulfilled, 2 requests have a fully executed DUA/MTA but the data/specimens have not yet been transferred, and 2 requests have a DUA/MTA in progress and the data/specimens will be transferred thereafter. The amount of time required to review each request, execute the DUA/MTA, and transfer the data/specimens has varied by request due to the complexity of each request, review and negotiations of all legal agreements, and other factors.

DISCUSSION

Through the development of the CSPEC-Durham Repository, we created a mechanism for facilitating future research on diverse health topics affecting Veterans. This multi-study repository provides a single structure that can be used to support the sharing of data and specimens across multiple content areas, for different types of research studies (e.g., active vs. legacy studies, cross-sectional vs. longitudinal data collection, etc.), and across types of data (e.g., survey data, medical record data, genomic data) and specimens (e.g., blood, tissue, DNA) collected. This resource can support additional research, including genetic and molecular association analyses, aimed to better understand, diagnose, and treat chronic diseases affecting Veterans and the general population, which closely aligns with current ORD, VA, and national research priorities.

The CSPEC-Durham Repository adds to the growing number of data repositories and biorepositories within the VA, reflecting the high prioritization of research collaboration to improve care delivery. The VA's Million Veteran Program (MVP) has enrolled more than 825,000 Veterans since 2011 in order to facilitate research assessing genetic influences on health and disease to develop precision medicine (23, 24). The Veterans Precision Oncology Data Commons similarly aims to support research in precision oncology through the sharing of clinical and genomic data available for cancer patients in the VA (25). Other examples of repositories in the VA focused on specific topic areas include the Mental Illness Research Education and Clinical Center (MIRECC) (26) and the VA Biorepository Brain Bank for ALS research (27). The CSPEC-Durham Repository is unique in that it is a center-wide repository, not specific to a single health topic, and allows for data and specimen sharing across legacy studies for which data and specimen sharing would otherwise not be possible.

One of the strengths of our repository is the metadata-driven database architecture, which has automated steps across the data life cycle, including data entry and data extraction for approved investigators. We also successfully applied this metadata-driven approach to the specimens stored in our repository. This approach has increased efficiency from a repository management perspective and allowed for improved safeguarding of the study data and specimens. Another asset has been the inclusion of legacy studies in the repository. Given that Veteran participants of these studies provided their time and efforts to research on particular health issues, it is important to be able to use the information and specimens they shared to advance research and innovations in these areas (within the scope of their original consent). Including these studies reflects the trend over time toward increased transparency and access to research data.

There are also some limitations. The number of participants across the repository feeder studies is relatively small when compared to biobanks such as MVP (23, 24). For this reason, researchers requesting the data are asked to provide their statistical plan and reflect on the implications of the sample size for their particular study. The merits of their statistical plan and plans to address any data limitations are reviewed and evaluated by one or more statistical reviewers on the SEOC as part of the review process. In addition, now that we have developed the repository structure and database architecture, we have a well-established mechanism to adopt additional feeder studies, including those that may be actively recruiting participants. This may help to increase the number of study participants for whom we have data and specimens available for each content area. A second limitation is that the data and specimens for the GWECB were collected at a single point in time, and the specimens collected from ALS Registry participants were also cross-sectional. However, the participants from the GWECB consented to be re-contacted for additional studies related to Gulf War research, which may allow for collection of data at subsequent time points. In addition, while the specimens were collected at a single point in time from the ALS Registry participants, their surveys were completed at multiple time points.

Our repository team has taken steps to integrate our resources with other repository initiatives within the VA to increase efficiencies for our staff and researchers alike and to improve visibility of our resources. As one key example, our team works closely with the Integrated Veteran Epidemiologic Study Data Resource (INVESTD-R) team, which has created a publicly available web-based tool to describe the resources available for continued research within the VA CSP (28). The feeder studies included in the CSPEC-Durham Repository are highlighted on this resource, allowing us to potentially reach more diverse researchers and consolidate our resources within the context of the larger CSP research program. We continue to work with the INVESTD-R team to streamline review processes and other documentation for researchers requesting data and specimens across the VA CSP. In addition, following existing models within the VA, all study specimens in our repository are physically stored at approved VA biorepositories. While our team governs all aspects of data management for these specimens, including the request process, crosswalk between the specimens and the corresponding study data, and the sharing of specimens with approved researchers, the laboratory personnel at the VA biorepositories ensure secure storage and maintenance of the physical specimens. These collaborations help to leverage our respective areas of expertise and available resources to best support continued Veteran health research. Within the larger VA ORD, there are ongoing discussions across the program about how to further integrate existing repository resources while still adhering to all VA data sharing requirements and adhering to the permissions documented in the original consent forms.

There are continued opportunities to advance Veteran health research and delivery of care through collaboration with other VA repositories. As one example, we hope to create a streamlined review process for requests to use ALS specimens with the VA Biorepository Brain Bank, which stores central nervous system (CNS) tissues for Veterans with ALS. Creating a joint process will allow interested investigators to simultaneously request DNA samples from our repository and tissue samples from the Brain Bank for the same individuals. Furthermore, there is opportunity to link our GWECB with additional Gulf War research resources in the VA. These collaborative activities can create further efficiencies in the storage and sharing of Veteran data and specimens, with the overarching goal of sharing VA data nationally and using this information to improve the health and care of Veterans.

DATA AVAILABILITY STATEMENT

The original contributions generated in the study are included in the article. Details regarding the data and specimens included in the CSPEC-Durham Data and Specimen Repository are available on the INVESTD-R website (https://www.vacsp.research.va.gov/CSPEC/Studies/INVESTD-R/Main.asp). Inquiries about requesting to use the data and specimens can be directed to the corresponding authors.

ETHICS STATEMENT

The feeder studies included in the CSPEC-Durham Data and Specimen Repository involving human participants were reviewed and approved by the Durham VA Health Care System Institutional Review Board and/or the VA Central Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

DISCLOSURE

The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs, the VA Cooperative Studies Program, or the U.S. government.

AUTHOR CONTRIBUTIONS

MO'L: conceptualization, methodology, project administration, writing, and revising the manuscript. RLW and CW: conceptualization, methodology, and revising the manuscript. AP: conceptualization, methodology, project administration, and revising the manuscript. DP and KS: conceptualization, methodology, revising the manuscript, and supervision. BC, RJ, and TR: methodology and revising the manuscript. All authors: contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Cooperative Studies Program in the Office of Research and Development, U.S. Department of Veterans Affairs.

ACKNOWLEDGMENTS

We thank the following individuals for their contributions to the development of the CSPEC-Durham Data and Specimen Repository: Elizabeth R. Hauser and Rebecca B. McNeil for the design and development of the repository database; Mary E. Grewe for the development of the repository protocol and application process; A. Jasmine Bullard for quality improvement activities; Kristina Felder for project management; and Ashton N. Madison and Reana S. Thomas for the collection and retrieval of specimens. We are grateful to the members of the Scientific and Ethical Oversight Committee for their review of data and specimen requests. We also thank the Massachusetts Veterans Epidemiology Research and Information Center (MAVERIC) and the Southern Arizona VA Healthcare System (SAVAHCS) for their oversight and management of all specimen storage and sharing. We would also like to thank the Veterans who participated in the research studies included in our repository for their time and willingness for their data and specimens to be used to support future research on Veteran health.

REFERENCES

1. Vaught J. Biobanking comes of age: the transition to biospecimen science. *Ann Rev Pharmacol Toxicol.* (2016) 56:211–28. doi: 10.1146/annurev-pharmtox-010715-103246
2. Coppola L, Cianflone A, Grimaldi AM, Incoronato M, Bevilacqua P, Messina F, et al. Biobanking in health care: evolution and future directions. *J Transl Med*. (2019) 17:172. doi: 10.1186/s12967-019-1922-3
3. Harati MD, Williams RR, Movassaghi M, Hojat A, Lucey GM, Yong WH. An introduction to starting a biobank. *Methods Mol Biol (Clifton, NJ)*. (2019) 1897:7–16. doi: 10.1007/978-1-4939-8935-5_2
4. Eisenhauer ER, Tait AR, Rieh SY, Arslanian-Engoren CM. Participants' understanding of informed consent for biobanking: a systematic review. *Clin Nurs Res*. (2019) 28:30–51. doi: 10.1177/1054773817722690
5. Mello MM, Lieou V, Goodman SN. Clinical trial participants' views of the risks and benefits of data sharing. *N Engl J Med*. (2018) 378:2202–11. doi: 10.1056/NEJMsa1713258
6. Kaufman D, Bollinger J, Dvoskin R, Scott J. Preferences for opt-in and opt-out enrollment and consent models in biobank research: a national survey of Veterans Administration patients. *Genet Med*. (2012) 14:787–94. doi: 10.1038/gim.2012.45
7. Institute of Medicine. *Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk*. Washington, DC: The National Academies Press (2015).

8. Holdren JP. *Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research.* Washington, DC: Executive Office of the President, Office of Science and Technology Policy (2013). Available online at: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf (accessed January 22, 2021).
9. U.S. Department of Veterans Affairs. *Policy and Implementation Plan for Public Access to Scientific Publications and Digital Data from Research Funded by the Department of Veterans Affairs*. Washington, DC: U.S. Department of Veterans Affairs (2015). Available online at: https://www.va.gov/ORO/Docs/Guidance/VA_RSCH_DATA_ACCESS_PLAN_07_23_2015.pdf (accessed January 22, 2021).
10. U.S. Department of Veterans Affairs Office of Research and Development. *Strategic Priorities for VA Research.* (2018). Available online at: https://www.research.va.gov/about/strategic_priorities.cfm (accessed January 22, 2021).
11. U.S. Department of Veterans Affairs Office of Research and Development. *Cooperative Studies Program Epidemiology Center (CSPEC)-Durham*. Available online at: https://www.research.va.gov/programs/csp/cspec/default.cfm (accessed January 22, 2021).
12. U.S. Department of Health and Human Services. *Summary of the HIPAA Privacy Rule*. Available online at: https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html (accessed January 22, 2021).
13. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data Capture (REDCap)–a metadata-driven methodology and workflow process for providing translational research informatics support. *J Biomed Inform*. (2009) 42:377–81. doi: 10.1016/j.jbi.2008.08.010
14. Lieberman DA, Weiss DG, Bond JH, Ahnen DJ, Garewal H, Harford WV, et al. Use of colonoscopy to screen asymptomatic adults for colorectal cancer. *N Engl J Med*. (2000) 343:162–8. doi: 10.1056/NEJM200007203430301
15. Lieberman DA, Weiss DG, Harford WV, Ahnen DJ, Provenzale D, Sontag SJ, et al. Five-year colon surveillance after screening colonoscopy. *Gastroenterology*. (2007) 133:1077–85. doi: 10.1053/j.gastro.2007.07.006
16. Lieberman D, Sullivan BA, Hauser ER, Qin X, Musselwhite LW, O'Leary MC, et al. Baseline colonoscopy findings associated with 10-year outcomes in a screening cohort undergoing colonoscopy surveillance. *Gastroenterology*. (2020) 158:862–74.e8. doi: 10.1053/j.gastro.2019.07.052
17. Sullivan BA QX, Redding TS, Gellad ZF, Stone A, Weiss D, Madison AN, et al. Genetic colorectal cancer and adenoma risk variants are associated with increasing cumulative adenoma counts. *CEBP*. (2020) 29:2269–76. doi: 10.1158/1055-9965.EPI-20-0465
18. Allen KD, Kasarskis EJ, Bedlack RS, Rozear MP, Morgenlander JC, Sabet A, et al. The National Registry of Veterans with Amyotrophic Lateral Sclerosis. *Neuroepidemiology*. (2008) 30:180–90. doi: 10.1159/000126910
19. Kwee LC, Liu Y, Haynes C, Gibson JR, Stone A, Schichman SA, et al. A high-density genome-wide association screen of sporadic ALS in US Veterans. *PLoS ONE*. (2012) 7:e32768. doi: 10.1371/journal.pone.0032768
20. Khalil L, McNeil RB, Sims KJ, Felder KA, Hauser ER, Goldstein KM, et al. The Gulf War Era Cohort and Biorepository: a longitudinal research resource of Veterans of the 1990-1991 Gulf War era. *Am J Epidemiol*. (2018) 187:2279–91. doi: 10.1093/aje/kwy147
21. Fukuda K, Nisenbaum R, Stewart G, Thompson WW, Robin L, Washko RM, et al. Chronic multisymptom illness affecting Air Force Veterans of the Gulf War. *JAMA*. (1998) 280:981–8. doi: 10.1001/jama.280.11.981
22. Steele L. Prevalence and patterns of Gulf War Illness in Kansas Veterans: association of symptoms with characteristics of person, place, and time of military service. *Am J Epidemiol*. (2000) 152:992–1002. doi: 10.1093/aje/152.10.992
23. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, et al. Million Veteran Program: a mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol*. (2016) 70:214–23. doi: 10.1016/j.jclinepi.2015.09.016
24. U.S. Department of Veterans Affairs Office of Research and Development. *Million Veteran Program (MVP)*. Available online at: https://www.research.va.gov/mvp/ (accessed January 22, 2021).
25. Do N, Grossman R, Feldman T, Fillmore N, Elbers D, Tuck D, et al. The Veterans Precision Oncology Data Common: transforming VA data into a national resource for research in precision oncology. *Semin Oncol*. (2019) 46:314–20. doi: 10.1053/j.seminoncol.2019.09.002
26. U.S. Department of Veterans Affairs. *MIRECC/CoE*. Available online at: https://www.mirecc.va.gov (accessed January 22, 2021).
27. Brady CB, Trevor KT, Stein TD, Deykin EY, Perkins SD, Averill JG, et al. The Department of Veterans Affairs Biorepository Brain Bank: a national resource for amyotrophic lateral sclerosis research. *Amyotroph Lateral Scler Frontotemporal Degener*. (2013) 14:591–7. doi: 10.3109/21678421.2013.822516
28. U.S. Department of Veterans Affairs Cooperative Studies Program. *Integrated Veteran Epidemiologic Study Data Resource (INVESTD-R).* (2018). Available online at: https://www.vacsp.research.va.gov/CSPEC/Studies/INVESTD-R/Main.asp (accessed January 22, 2021).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

ORIGINAL RESEARCH
published: 11 May 2021
doi: 10.3389/fpubh.2021.584955

Causal Effects of Body Mass Index on Airflow Obstruction and Forced Mid-Expiratory Flow: A Mendelian Randomization Study Taking Interactions and Age-Specific Instruments Into Consideration Toward a Life Course Perspective

Nicole Probst-Hensch[1,2*], *Ayoung Jeong*[1,2], *Daiana Stolz*[3], *Marco Pons*[4], *Paola M. Soccal*[5], *Robert Bettschart*[6], *Deborah Jarvis*[7,8], *John W. Holloway*[9], *Florian Kronenberg*[10], *Medea Imboden*[1,2], *Christian Schindler*[1,2] and *Gianfranco F. Lovison*[1,2,11]

[1] Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Basel, Switzerland, [2] Department of Public Health, University of Basel, Basel, Switzerland, [3] Clinic of Pulmonary Medicine and Respiratory Cell Research, University Hospital Basel, Basel, Switzerland, [4] Division of Pulmonary Medicine, Regional Hospital of Lugano, Lugano, Switzerland, [5] Division of Pulmonary Medicine, Geneva University Hospitals, Geneva, Switzerland, [6] Lungenpraxis Aarau, Hirslanden Klinik, Aarau, Switzerland, [7] Medical Research Council-Public Health England, Centre for Environment and Health, Imperial College London, London, United Kingdom, [8] Population Health and Occupational Disease, National Heart and Lung Institute, Imperial College London, London, United Kingdom, [9] Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, United Kingdom, [10] Division of Genetic Epidemiology, Department of Medical Genetics, Molecular and Clinical Pharmacology, Medical University of Innsbruck, Innsbruck, Austria, [11] Department of Economics, Business and Statistics, University of Palermo, Palermo, Italy

OPEN ACCESS

Edited by:
Konstantin G. Arbeev,
Duke University, United States

Reviewed by:
Ronja Foraita,
Independent Researcher,
Bremen, Germany
Fulvio Ricceri,
University of Turin, Italy

***Correspondence:**
Nicole Probst-Hensch
Nicole.probst@swisstph.ch
orcid/org.0000-0001-8561-5976

Specialty section:
This article was submitted to Life-Course Epidemiology and Social Inequalities in Health, a section of the journal Frontiers in Public Health

Received: *07 September 2020*
Accepted: *01 April 2021*
Published: *11 May 2021*

Citation:
Probst-Hensch N, Jeong A, Stolz D, Pons M, Soccal PM, Bettschart R, Jarvis D, Holloway JW, Kronenberg F, Imboden M, Schindler C and Lovison GF (2021) Causal Effects of Body Mass Index on Airflow Obstruction and Forced Mid-Expiratory Flow: A Mendelian Randomization Study Taking Interactions and Age-Specific Instruments Into Consideration Toward a Life Course Perspective. Front. Public Health 9:584955. doi: 10.3389/fpubh.2021.584955

Obesity has complex links to respiratory health. Mendelian randomization (MR) enables assessment of causality of body mass index (BMI) effects on airflow obstruction and mid-expiratory flow. In the adult SAPALDIA cohort, recruiting 9,651 population-representative samples aged 18–60 years at baseline (female 51%), BMI and the ratio of forced expiratory volume in 1 second (FEV_1) to forced vital capacity (FVC) as well as forced mid-expiratory flow (FEF25–75%) were measured three times over 20 follow-up years. The causal effects of BMI in childhood and adulthood on FEV1/FVC and FEF25–75% were assessed in predictive (BMI averaged over 1st and 2nd, lung function (LF) averaged over 2nd and 3rd follow-up; $N = 2,850$) and long-term cross-sectional models (BMI and LF averaged over all follow-ups; $N = 2,728$) by Mendelian Randomization analyses with the use of weighted BMI allele score as an instrument variable and two-stage least squares (2SLS) method. Three different BMI allele scores were applied to specifically capture the part of BMI in adulthood that likely reflects tracking of genetically determined BMI in childhood. The main causal effects were derived from models containing BMI (instrumented by BMI genetic score), age, sex, height, and packyears smoked as covariates. BMI interactions were instrumented by the product of the instrument (BMI genetic score) and the relevant concomitant variable. Causal effects of BMI on FEV1/FVC and FEF25–75% were observed in both the predictive and long-term cross-sectional models. The causal BMI- LF effects were negative and attenuated with increasing age, and stronger if instrumented by gene scores associated with childhood BMI. This non-standard MR approach interrogating causal effects of multiplicative interaction suggests that the genetically rooted part of BMI patterns in childhood may be of particular relevance for the level of small airway function and airflow obstruction later in life. The methodological relevance of the results is first to point to the importance of a life course perspective in studies on the etiological role of BMI in respiratory health, and second to point out novel methodological aspects to be considered in future MR studies on the causal effects of obesity related phenotypes.

Keywords: Mendelian randomization, body mass index, genetic score, lung function, COPD, longitudinal cohort

INTRODUCTION

Obesity, mostly measured as body mass index (BMI) is an established asthma risk factor. Its etiological role with regard to other respiratory phenotypes including chronic obstructive pulmonary disease (COPD) remains unclear (1–4). Observational evidence on the association of obesity with spirometry-derived lung function (LF) is inconclusive (5–12). In adulthood, increasing BMI has been often, but not exclusively, associated with lower forced expiratory volume in 1 second (FEV1) and forced vital capacity (FVC). Bariatric surgery improved FVC and FEV1 in asthmatics over 5 years (13). FEV1/FVC was sometimes preserved or even increased in the presence of excess body weight, but overall the association with airflow obstruction (AO) remains unclear (12). Inconsistencies between studies reflect differences in the study populations (age, health state, ethnicity, lifestyles, environments, socio-economic profile), differences in obesity parameters studied, and statistical models (confounders and effect modifiers considered).

Mechanisms by which obesity in adults can impair LF include increased abdominal pressure due to fat mass, a related decrease in the recoil properties of the chest wall, distal airway closure and lung volume reduction. In addition, excess fat mass may exacerbate systemic and airway inflammation (1, 14–16). In fact, adipose tissue associated immunological and pro-inflammatory factors may already impact on respiratory health during childhood. Weight change patterns in early life were recently associated with dysanapsis in which FVC is higher relative to FEV_1 as a result of a possible imbalance between alveolar and airway growth (17). Although no study was able to investigate the association of early life weight change patterns with respiratory health in older adults, small airways are known to be frequently involved at a very early stage of COPD and possibly asthma (18).

Small airways are more difficult to study in the absence of a gold standard for measuring their dysfunction. Forced mid-expiratory flow (FEF25–75%, abbreviated as FEF2575 hereafter) is thought to better capture small airways dysfunction than FEV1/FVC (19). It may therefore be more sensitive to reflect chronic effects of obesity on small airways. Few observational BMI–LF studies in adults have considered FEF2575 (20–22). But impulse oscillometry (IOS) studies, more reliable in assessing distal airway function, found increased airway resistance and decreased airway reactance with elevated BMI (15).

Further insight into the causality of the BMI-LF association can be gained by Mendelian randomization (MR) studies (23). Increasingly larger genome-wide association studies (GWAS), primarily in adults, have identified more and more loci associated with BMI at effect sizes and allele frequencies becoming smaller and smaller (24–26), enabling derivation of an instrumental variable. The different GWAS, conducted in adults or in children, allow deriving instrumental variables more specifically targeting either BMI in adulthood or BMI in childhood and thereby reflecting age-related differences in pathways to BMI, an aspect largely ignored in previous studies on BMI and lung function. While the largest BMI GWAS in adults to date (24) (named "Yengo score" in this paper) was not tested for association with childhood BMI, the single nucleotide variants (SNPs) identified in the earlier adult BMI GWAS (named "Speliotes score" in this paper) were explicitly confirmed for association with childhood BMI (26). Yet, the correlation between this latter genetic score with one derived from a recent GWAS meta-analysis on BMI of more than 40 000 children (named "Felix score") was reported at only 0.73 (25), pointing to differences in genetic pathways determining childhood vs. adulthood BMI.

Only one, large MR meta-analysis has investigated the causal effect of BMI on adult LF and it applied the adult BMI-derived genetic score ("Speliotes score") (25, 26), but not the childhood BMI-derived genetic score ("Felix score") (25). This study relied on FEV1, FVC and BMI measured at a single time point in almost 500,000 participants, and supported a causal effect of BMI (2). The causal effect of BMI on other LF parameters relevant to asthma and COPD such as FEV1/FVC, the physiological parameter used to define AO, and FEF2575 (27–30), has not been investigated using an MR approach.

The SAPALDIA cohort with 20 years of BMI and LF follow-up offered the opportunity to study the chronicity of BMI-LF association over an extended period of time in the context of an MR study. We evaluated causal effects (a) of BMI averaged over time points 1 and 2 on lung function averaged over time points 2 and 3 (predictive model) and (b) of BMI and lung function averaged over 3 time points (long-term cross-sectional model). Since BMI fluctuates over time, we instrumented long-term average BMI as a more meaningful exposure measure than

BMI from a single time point. Similarly, since lung function fluctuates over time and is measured with error (compliance of participants, field worker effects, spirometry device effects), we focused on long-term average LF as a more meaningful outcome phenotype than level or change in lung function. We applied MR to BMI, even though BMI is only an imprecise measure of adiposity for the following reasons: First, better instrument is available for BMI thanks to large GWAS, compared to other adiposity metrics. Second, SAPALDIA has not longitudinally measured other adiposity metrics as complete as BMI. Third, BMI has been the most common obesity metric associated with lung function in previous studies. The study *a priori* focused on FEV1/FVC and FEF2575 as outcomes and *a priori* instrumented BMI in three different ways (Yengo score, Speliotes score, Felix score) in an attempt to specifically capture the part of BMI in adulthood that reflects the tracking of genetically determined BMI in childhood.

METHODS

Study Population

SAPALDIA has been described previously (31). Random population samples aged 18–60 years were invited in eight Swiss study areas for the baseline survey in 1991 (SAP1). Of the 9,651 baseline participants, 8,047 (83.4%) participated in follow-up SAP2 (2001/3) and 6,139 (63.6%) in follow-up SAP3 (2010/11). This paper was restricted to participants in all three surveys with complete spirometry, BMI, genotype and covariate data for the respective causal model (**Supplementary Figures 1A,B**).

Ethical approval was obtained for each survey and study area from the central ethics committee of the Swiss Academy of Medical Sciences and the Cantonal Ethics Committees. Participants provided informed consent. All methods were performed in accordance with the relevant guidelines and regulations.

Lung Function

Spirometry was conducted with heated-wire spirometers (SensorMedics, Yorba Linda, California) (SAP1 & SAP2), and by portable, ultrasonic EasyOne spirometer (ndd medizintechnick AG, Zürich, Switzerland) (SAP3), according to American Thoracic Society recommendations (32) (see **Supplementary Material**). The LF parameters considered for this study are the ratio FEV1/FVC, forced mid-expiratory flow FEF2575, and FEF2575/FVC (results in **Supplementary Material**), derived from pre-bronchodilation spirometry. FEV1 and FVC decline as airway narrows. A reduced FEV1/FVC defines AO, resulting if the decline in FEV1 is out of proportion to the decline in FVC, while reduced FVC indicates restriction. FEF2575 is an early indicator of AO and sensitive to small airway dysfunction. Reduced FEF2575/FVC is an indicator of dysanapsis where lung volume increases as a result of air trapping in the presence of AO. SAP3 measurements were re-calibrated to assure comparability with SAP1 and SAP2 measurements (33).

BMI and Covariates

Height was measured. Weight was asked for at baseline, but measured at follow-up. BMI was calculated in kg/m^2. Exact age was calculated based on birth and examination dates. Sex was self-reported. Smoking was self-reported and measured as pack-years smoked up to baseline and during the two follow-up periods. Non-asthmatics where defined as those who never reported a doctor diagnosis of asthma.

Genotyping

DNA was extracted from EDTA blood. 570k SNPs were genotyped for 1,612 SAPALDIA samples by Human610-Quad BeadChip (Illumina, San Diego, CA, USA) (34) and ~1 million SNPs were genotyped for additional 3,015 SAPALDIA samples by Infinium Human OmniExpressExome-8 (Illumina, San Diego, CA, USA) (35). Samples with call rate <0.97 or population outliers were excluded. Markers with call rate <0.95, minor allele frequency <0.05, or out of Hardy-Weinberg equilibrium ($p < 10^{-6}$) were excluded. The genotype datasets were then phased using ShapeIT (v2.r790) (36) and imputed using MiniMac2 (version 2014) (37) to 1,000 Genome phase 1 reference panel comprising of 1,092 samples. The imputed datasets were merged to yield 38 million markers.

BMI Allele Score

The genetic instruments for BMI were single-nucleotide polymorphisms (SNPs) independently [linkage disequilibrium (LD) R^2 measure < 0.2] associated in Caucasians with the BMI at a genome-wide level ($P < 5 \times 10^{-8}$). Three scores were derived, i.e., "Speliotes Score" (adult BMI GWAS also associated with childhood BMI, used in the only previous BMI-LF MR study); "Felix Score" (childhood BMI GWAS); and "Yengo Score" (largest adult BMI GWAS, unknown association with childhood BMI). They were computed as weighted sum of 32, 12, and 862 BMI-increasing alleles reported by Speliotes et al. (26), Felix et al. (25), and Yengo et al. (24), respectively, using the reported coefficients for each SNP as weights, following the same approach as earlier MR studies of BMI (2, 38). We excluded SNPs with poor imputation quality ($r^2 < 0.3$) or with known association with smoking phenotypes in PhenoScanner. rs13387838 for which visual inspection of MR Egger regression results clearly indicated pleiotropy was further excluded from Felix Score. **Supplementary Table 1** describes the 32, 12, and 862 SNPs used to construct Speliotes et al. (26), Felix et al. (25), and Yengo et al. (24) scores, respectively.

As the weights' sum is bounded by the number of SNPs considered, the effect size of each score can be interpreted as average effect per one BMI-increasing allele. The three scores were only moderately correlated (0.28–0.55). The correlation was smallest between the Yengo and the Felix scores (0.28) (**Supplementary Table 2**).

Statistical Analysis

Statistical analyses were performed in R, version 3.4.3 for Windows (http://www.r-project.org/) (see the "Statistical analysis" section in the **Supplementary Material**).

Analysis Scheme

We investigated the causal association in a *predictive model* (exposure: BMI averaged over SAP1 and SAP2, referred to as SAP1-SAP2; outcome: LF averaged over SAP2 and SAP3, referred to as SAP2-SAP3) and in a *long-term cross-sectional model* (BMI and LF both averaged over SAP1, SAP2, and SAP3; referred to as SAP1-SAP2-SAP3).

Descriptive Analysis

Characteristics of study participants were summarized for the combinations of surveys involved in the modeling phase: SAP1-SAP2, SAP2-SAP3, and SAP1-SAP2-SAP3. Partial correlation coefficients were computed: (i) between the same LF variables over the 3 occasions in time to assess temporal auto-correlation; and (ii) between the different LF variables (and their derived averages) at each occasion in time to assess their degree of (linear) relationship. Partial correlations were computed using residuals of each LF variable from models that regress them on Age, Age^2, Height, Height^2, Sex, and all their first-order interactions. Pairwise complete cases analysis was performed, to accommodate the differential presence of missing values in the variables involved. BMI distribution at each survey was visualized as histograms.

Checking MR Assumptions

In preparing for MR analysis, a set of assumptions as highlighted in VanderWeele et al. (39) were checked:

(1) The genetic score is associated with the exposure. In the context of our study, this requires testing the presence of an association of BMI genetic score with BMI;
(2) The genetic score is not associated with confounders of the exposure–outcome relationship. In the context of our study, this required various actions: (2.1) testing that the BMI genetic score is not associated with the observed confounders Packyears and Height; (2.2) SNPs identified in PhenoScanner (40) as associated with smoking phenotypes, were excluded from computing genetic scores; (2.3) MR-Egger regression (41) was conducted to check for pleiotropy; (2.4) We interrogated whether age or sex modify the influence of BMI genetic score on phenotypic BMI by regressing the BMI averages on a linear predictor including Age (averaged over SAP1-SAP2 and SAP1-SAP2-SAP3, respectively, and centered at 18 years), Sex, BMI genetic score and all their interactions;
(3) The genetic score is not associated with the outcome, conditional on the exposure and confounders of the exposure–outcome relationship. In the context of our study, this requires testing the absence of a BMI genetic score association with LF, conditional on BMI and (observed) confounders of the BMI-LF relationship.

In all these checks, the models used were chosen through a selection procedure carried out within the class of (extended) Generalized Linear Models, with the aim of making the choice more flexible and finding the model most appropriate in terms of both distribution of response and possible non-linearity of the relationship of the response with the predictors (see **Supplementary Material** for details).

Mendelian Randomization Analysis

As MR assumptions appeared to be satisfied in our data, instrumental variable (IV) analyses were carried out to test and estimate the causal effects of BMI on LF in the context of Linear Gaussian models (42). Estimation was carried out using the two-Stage Least Squares (2SLS) method. In the first-stage of 2SLS, the exposure is regressed on the genetic score to give fitted values of the exposure ("Exposure models"). In the second-stage, the outcome is regressed on the fitted values for the exposure from the first stage regression, along with other covariates ("Causal model"). The causal estimate is this second-stage regression coefficient for the change in outcome caused by a unit change in the exposure. Details can be found in Burgess and Thompson (43) (ch. 4.2). All MR analyses were carried out using the ivreg command of the R library AER.

The first- and second-stage analyses were based on identical data. The response variables were LF parameters averaged over either SAP2-SAP3 (predictive model) or SAP1-SAP2-SAP3 (long-term cross-sectional model). The causal variable (instrumented by the respective BMI genetic score) was the logarithm of the BMI averages over either SAP1-SAP2 (predictive model) or SAP1-SAP2-SAP3 (long-term cross-sectional model). The choice of log-transforming the BMI averages was made through an AIC-based selection procedure. This transformation appeared to be the best linear predictor for all LF outcomes and the best choice in checking MR Assumption 1 (see the Results for details).

Explanatory variables for each LF variable were chosen through a model selection procedure. The initial (maximal), and *a priori* sparse, model contained the following covariates: (instrumented) BMI, Age (centered at 18 years, the minimal admission age at SAP1), Sex, Height, and Packyears smoked, along with all their pairwise interactions. Physical activity was *a priori* not included in the model due to its potential role as mediator of the BMI-LF association. We decided not to include study center and educational level after we observed adding them to the final causal and observational models did not materially alter the effect estimates.

It is to be stressed that the inclusion of interactions implies that all the interaction parameters between BMI and all other variables must also be considered as causal, and must be themselves instrumented; this represents an innovative aspect of this paper, since models used in MR studies are usually assumed to be additive, and no attempt is made to check the appropriateness of this assumption. In a non-standard MR approach and following a suggestion by Bun and Harrison (44), the interrogation of causal interactions was instrumented by the product of the instrument (BMI genetic score) and the relevant concomitant variable. Given that age can be neither genetically determined nor confounded, BMI:Age interaction is a special case and our approach cannot be generalized into other interaction MR analyses.

Starting from the maximal model, a model selection procedure, based on AIC comparisons, provided the final model which retained (instrumented) BMI and its interaction with Age, as well as Age, Sex, Height, Packyears smoked, Age × Sex and Age × Height interactions. Standard errors for the causal parameter IV estimates were obtained by second order delta method. Wald confidence intervals were derived based on asymptotic Normality. In all models, the error distribution was assumed to be Normal, so that in all exposure models the response on the original scale ($BMI_{s1,s2}$ and $BMI_{s1,s2,s3}$) was assumed to be logNormal (see **Supplementary Material** for details).

The same final models were selected for the Mendelian Randomization analyses on the two lung function variables of main interest in this paper (FEV1/FVC and FEF2575). Their IV representation is as follows.

Predictive Model

$$\begin{aligned}
\text{Causal model}: E[LF_{s2,s3}] &= \beta_0 + \beta_{c1}\log(BMI_{s1,s2}) \\
&+\beta_1 Age\_c_{s1,s2} + \beta_2 Sex_{s2} \\
&+\beta_3 Height_{s1,s2} + \beta_4 PackYrs_{s2} \\
&+\beta_{c2}\log(BMI_{s1,s2})\times Age\_c_{s1s2} \\
&+\beta_5 Age\_c_{s1,s2}\times Sex_{s2} + \beta_6 Age\_c_{s1,s2} \\
&\times Height_{s1,s2} \qquad (1) \\
\text{Exposure models}: & \quad E[\log(BMI_{s1,s2})] = \alpha_0 + \alpha_1 BMI_{gs} \qquad (2) \\
& \quad E[\log(BMI_{s1,s2}) : Age\_c_{s1,s2}] \\
& \quad = \gamma_0 + \gamma_1 BMI_{gs}\times Age\_c_{s1,s2} \qquad (3)
\end{aligned}$$

Long-Term Cross-Sectional Model

$$\begin{aligned}
\text{Causal model}: E[LF_{s1,s2,s3}] &= \beta_0 + \beta_{c1}\log(BMI_{s1,s2,s3}) \\
&+\beta_1 Age\_c_{s1,s2,s3} + \beta_2 Sex_{s2} \\
&+\beta_3 Height_{s1,s2,s3} + \beta_4 PackYrs_{s3} \\
&+\beta_2\log(BMI_{s1,s2,s3})\times Age\_c_{s1,s2,s3} \\
&+\beta_5 Age\_c_{s1,s2,s3}\times Sex_{s2} \\
&+\beta_6 Age\_c_{s1,s2,s3}\times Height_{s1,s2,s3} \qquad (4) \\
\text{Exposure models}: & \quad E[\log(BMI_{s1,s2,s3})] = \alpha_0 \qquad (5) \\
& \quad +\alpha_1 BMI_{gs} \\
& \quad E[\log(BMI_{s1,s2,s3}) : Age\_c_{s1,s2,s3}] \\
& \quad = \gamma_0 + \gamma_1 BMI_{gs}\times Age\_c_{s1,s2,s3} \qquad (6)
\end{aligned}$$

where: LF (Lung Function) is either FEV1/FVC or FEF2575;

β_{c1} and β_{c2} are the causal effect parameters;
all variables with multiple subscripts are averages over the relevant SAPALDIA surveys (e.g., $BMI_{s1,s2}$ is the average of BMI_{s1} and BMI_{s2});
Age_c is Age averaged over either SAP1-SAP2 or SAP1-SAP2-SAP3 and centered at 18 years;
$PackYrs_{si}$ = Pack-years smoked up to SAP_i (i = 2 or 3)
BMI_{gs} is the BMI genetic score (either Speliotes, Felix, or Yengo score).

Observational Association Analysis

The BMI-LF associations were analyzed using linear regression analyses adjusted for Sex, Age, Height, and Packyears smoked. For comparability with the MR results, the same final models [(1) and (4)] were re-fitted, using observed BMI (and observed Age × BMI interaction) instead of instrumenting them, and estimated by Ordinary Least Squares.

TABLE 1 | Characteristics of study participants included in the sample: **(A)** used to fit the predictive model; **(B)** used to fit the long-term cross-sectional model.

(A) Sample of the predictive model

	SAP1, SAP2	SAP2, SAP3
	***N* = 2,850**	
Sex at s2, % female	49.35	
Mean (s1, s2) Age, years (mean; SD)	44.71 (10.81)	
Mean (s1, s2) Height, cm (mean; SD)	170.11 (8.85)	
Mean (s1, s2) Weight, kg[a] (mean; SD)	71.64 (13.25)	
Mean (s1, s2) BMI, kg/m^2 (mean; SD)	24.44 (3.54)	
Packyears of cigarettes at s2 (mean; SD)	10.47 (17.13)	
Mean (s2, s3) FEF2575, ml[b] (mean; SD)		2.58 (1.08) (*N* = 2,936)
Mean (s2, s3) FEV1/FVC^2 (mean; SD)		0.74 (0.07) (*N* = 2,939)
Asthma up to s2 (% doctor diagnosed asthma)	10.30	

(B) Sample of the long-term model

	SAP1, SAP2, SAP3
	***N* = 2,728**
Sex at s2, % female	50.53
Mean (s1, s2, s3) Age, years (mean; SD)	49.43 (10.78)
Mean (s1, s2, s3) Height, cm (mean; SD)	169.63 (8.87)
Mean (s1, s2, s3) Weight, kg[a] (mean; SD)	72.02 (13.21)
Mean (s1, s2, s3) BMI, kg/m^2 (mean; SD)	24.95 (3.68)
Packyears of cigarettes at s3 (mean; SD)	11.48 (18.74)
Mean (s1, s2, s3) FEF2575, ml[b] (mean; SD)	2.89 (1.07)
Mean (s1, s2, s3) FEV1/FVC^2 (mean; SD)	0.76 (0.06)
Asthma (% doctor diagnosed asthma ever)	13.35

[a] *Weight was self-reported at baseline, and measured at follow-up.*
[b] *Lung function at SAP3 was corrected for change in spirometry device (33).*

Sensitivity Analysis

The reliability of self-reported, instead of measured, weight at SAP1 was assessed by comparing the estimated regression coefficient and the estimated determination coefficient R^2 of the BMI_{s1} vs. BMI_{s2} and BMI_{s3} relationships with BMI genetic scores. In order to check the possible effects due to the non-Normality of the LF variables we re-fitted the final models (1–6) employed in MR analysis using log-transformed (FEF2575) and logit-transformed (FEV1/FVC) parameters as outcomes. MR analysis was repeated using the ratio FEF2575/FVC as outcome (20) and for non-asthmatics, again re-fitting the final models (1–6).

In a preliminary analysis, we also investigated the association between *changes* in BMI and *changes* in LF, to check if this was a better way of exploiting the longitudinal nature of our data, compared to the use of medium- and long-term averages. Notice that we could perform this analysis only in observational association terms, since no genetic variants for BMI change are available.

The attrition bias due to potentially disproportionate lost to follow up over 20 years was interrogated by replicating the observational association analysis using Inverse Probability Weighted analysis, where the weights were either (1) the probability of participation in SAP2 and SAP3 given the variables used in the models (BMI, LF (either FEF2575 or FEV1/FVC), Age, Sex, Height, Packyears) measured at SAP1; or (2) the probability of participation in SAP3 given the variables used in the models (BMI, LF (either FEF2575 or FEV1/FVC), Age, Sex, Height, Packyears) averaged over SAP1-SAP2.

As a *post-hoc* analysis, we conducted stratified analysis by fitting the same final models, except for the Age×BMI interaction, in the strata defined based on tertiles of age at SAP1, using Speliotes score as instrument.

RESULTS

Descriptive Analysis

Characteristics of the study samples used for fitting the predictive and the long-term cross-section model are presented in **Table 1**. Variability of the LF variables, both within and between SAPALDIA surveys, and stratified by obesity is graphically depicted in **Figures 1A,B**. LF was lower among obese persons, but the difference became weaker (FEF2575) or disappeared (FEV1/FVC), as participants aged. Inverse associations not dependent on age were observed for FEV1 and FVC (**Supplementary Figure 2**). Partial correlations between the LF parameters, and the derived means, are presented in **Supplementary Table 3**. Histograms of BMI at each survey are presented in **Supplementary Figure 3**.

Checking the MR Assumptions

MR assumptions appeared to be satisfied (for details see **Supplementary Material** and **Supplementary Tables 4A,B**). The three BMI genetic scores derived from different life-course specific BMI variants were predictors of adult BMI, with the Yengo score being the strongest instrument (F-statistics 182 and 233 for long-term cross-sectional and predictive models, respectively). They were not associated with Packyears or Height. None of the SNPs included in the BMI genetic scores overlap with 154 smoking-related SNPs (45–47). One of the BMI SNPs (rs10767664) was in high LD with several smoking initiation associated SNPs in BDNF (brain derived neurotrophic factor) ($R^2 = 0.681 \sim 0.911$), but was not associated with smoking initiation in SAPALDIA, irrespective of adjustment for Age, Sex, and BMI. None, two, and sixteen SNPs were excluded from Speliotes, Felix, and Yengo Score, respectively, due to known association with smoking phenotypes in PhenoScanner. MR Egger regression did not indicate potential pleiotropy for main BMI effects. Slight indication of pleiotropy for the Age × BMI interaction was observed in FEV1/FVC and FEF2575 prediction models for Speliotes Score and in FEF2575 cross-sectional model for Felix Scores (see **Supplementary Table 5**, **Supplementary Figures 4–6**). We did not observe Age, Sex, or their combination to modify the association of BMI genetic score with phenotypic BMI (data not shown).

Mendelian Randomization Analysis

Causal effects of BMI on FEV1/FVC and FEF2575 were observed in the predictive and long-term cross-sectional models (**Tables 2–4** for Speliotes, Felix and Yengo Scores, respectively). For the Speliotes Score the causal effect of BMI on these two LF parameters was negative, but attenuated with increasing Age. **Figure 2** illustrates the age-BMI interaction with a 175 cm tall male never smoker as a reference individual. If he is 18 years old at SAP1, and hence his average age is 28 over the 20 years period between SAP1 and SAP3, and during this period his BMI changes from 25 to 30, employing the estimates in **Table 2** we can predict he will experience on average a decrease in his FEV1/FVC ratio approximately equal to 0.10. On the other hand, if he is 48 years old at SAP1, and hence his average age is 58 over the 20 years period between SAP1 and SAP3, the same change of BMI from 25 to 30 will cause an increase in his FEV1/FVC ratio ≈0.016. Causal effects were in the same direction, but with confidence intervals covering no effect, for the Yengo and the Felix Scores. Effect estimates of the Felix Score were about as large as for the Speliotes Score, whereas effect estimates for the Yengo score were considerably smaller. Irrespective of the genetic score, no BMI interactions with covariates other than Age were present. No causal effect of BMI on FEV1 or FVC was observed (results not presented).

Observational Association Analysis

The Ordinary Least Squares estimates of the observational associations of BMI with FEF2575 and FEV1/FVC, in terms of both main effects and interactions with Age, are presented in **Table 5**. Associations of BMI with FEV1/FVC and FEF2575 were in opposite directions: negative main effects and positive interactions with Age for FEV1/FVC, positive main effects and negative interactions with Age for FEF2575.

The comparison of MR causal effects and observational associations is visually helped by the forest plots in **Figures 3A–D**. While directions of MR causal effects and observational associations were consistent for FEV1/FVC, they were opposite for FEF2575. Confidence intervals were considerably wider for causal effects compared to observational associations.

To further investigate possible sources of the considerable discrepancy between causal and observational BMI effects, we

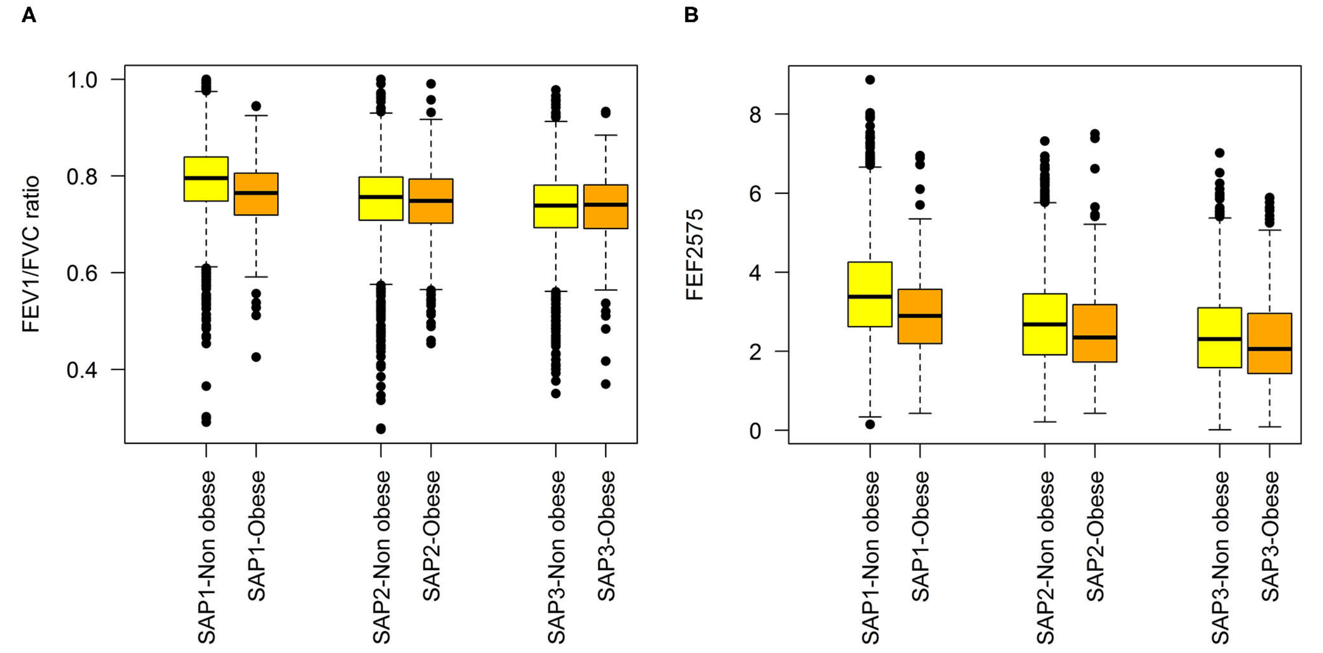


FIGURE 1 | Distributions of lung function variables at each SAPALDIA survey, by obesity state (BMI < 30 kg/m^2 vs. ≥30 kg/m^2: **(A)** FEV1/FVC, **(B)** FEF2575.

tried to check whether this could be due to the "composite" nature of BMI as a measure of obesity [for a related discussion of the difficulty of conducting causal inference with composite exposures, see (48)]. To this goal (see more detailed explanation in the **Supplementary Material**), we refitted the second stage IV models that contained both BMI$_{\text{instrumented}}$ and BMI$_{\text{residual}}$ = BMI-BMI$_{\text{instrumented}}$, derived from the IV first stage. BMI$_{\text{residual}}$, which reflects the non-genetically determined BMI variability, explained over 90% of observed BMI variability. We confirmed the positive association of BMI$_{\text{residual}}$ on FEF2575 and the lack of its association with FEV1/FVC, consistent with our observational analysis (see results in **Supplementary Table 6**).

Sensitivity Analysis

The comparison of regression estimates for BMI$_{s1}$, BMI$_{s2}$, and BMI$_{s3}$ on the three BMI genetic scores confirmed the reliability of BMI$_{s1}$ derived from self-reported weight (**Supplementary Table 7**). Irrespective of genetic score, the regression results for log-transformed (for FEF2575) and logit-transformed (for FEV1/FVC) outcomes were not materially different from those obtained using non-transformed parameters (**Supplementary Table 8**). No material changes in causal BMI effects were observed in models using FEF2575/FVC as outcome (**Supplementary Table 9**), in models restricted to non-asthmatics (**Supplementary Table 10**), or in models adjusting for study area and education (**Supplementary Table 11**). No association between change in BMI and change in lung function was observed. The Inverse Probability Weighted analyses did not show material changes in the associations of BMI with FEV1/FVC and FEF2575, although slightly attenuated associations were observed (**Supplementary Table 12**). Stratified analysis showed negative causal effects in younger age tertiles [(18.2, 35.2) and (35.2, 46.6)] but not in the oldest tertile [(46.6, 61.7)], confirming the MR results found for the Age × BMI interaction (**Supplementary Table 13**).

DISCUSSION

The results of this long-term study are consistent with a causal effect of BMI on AO and possibly small airway dysfunction. Higher levels of BMI cause lower levels of FEV1/FVC and FEF2575 up to middle-age, but the effect lessens with aging. The observed Age × BMI interaction, together with the stronger effects observed when instrumenting BMI with SNPs associated with childhood BMI, reflect the complexity of the BMI phenotype in adults. Adult BMI is the result of tracking of BMI over the life course and of genetic influences as well as non-genetic influences on weight change in both childhood and adulthood. Our results suggest that the genetically rooted part of BMI patterns in childhood may be of particular relevance for the level of small airway function and AO later in life, but that this effect diminishes with aging, when exogenous influences on BMI become more relevant.

The observational association between BMI and AO or COPD has not been well-studied. Results from the two largest, post-bronchodilation spirometry based studies are contradictory. In the world-wide BOLD study obesity was less common in persons with AO (12). The opposite was

TABLE 2 | Causal effects[a] of BMI on FEV1/FVC and FEF2575 in predictive and in long-term cross-sectional models.

	N	**β_{c1}, β_{c2}**	**SE**	***p*-value**
FEV1/FVC				
Predictive model				
BMI main effect $\log(BMI_{s1s2}) \rightarrow FEV1/FVC_{s2,s3}$	2,853	−0.561	0.256	0.029
BMI*Age interaction effect $\log(BMI_{s1s2}){:}Age_{s1s2} \rightarrow FEV1/FVC_{s2,s3}$		0.019	0.010	0.065
Long-term cross-sectional model				
BMI main effect $\log(BMI_{s1s2s3}) \rightarrow FEV1/FVC_{s1,s2,s3}$	2,731	−0.752	0.314	0.017
BMI*Age interaction effect $\log(BMI_{s1s2s3}){:}Age_{s1,s2,s3} \rightarrow FEV1/FVC_{s1,s2,s3}$		0.021	0.010	0.040
FEF2575				
Predictive model				
BMI main effect $\log(BMI_{s1s2}) \rightarrow FEF2575_{s2,s3}$	2,850	−7.152	3.457	0.038
BMI*Age interaction effect $\log(BMI_{s1s2}){:}Age_{s1s2} \rightarrow FEF2575_{s2,s3}$		0.222	0.141	0.116
Long-term cross-sectional model				
BMI main effect $\log(BMI_{s1,s2,s3}) \rightarrow FEF2575_{s1,s2,s3}$	2,728	−9.251	4.433	0.037
BMI*Age interaction effect $\log(BMI_{s1,s2,s3}){:}Age_{s1,s2,s3} \rightarrow FEF2575_{s1,s2,s3}$		0.242	0.146	0.096

BMI genetic score: (Speliotes; 32 SNPs).

[a] *β_{c1}, causal BMI main effect per one BMI-increasing allele; β_{c2}, causal BMI*Age interaction effect per one BMI-increasing allele.*

*The negative sign of the causal main effect means that, keeping all other predictors fixed, at age 18 (which has been chosen as the origin in our analysis) BMI has a causal negative effect on LF. The positive sign of the Age × BMI causal interactive effect implies that, as age increases, the detrimental effect of BMI on LF decreases. As a consequence, the total effect of BMI becomes null at middle ages and protective at older ages; for a graphical representation of such BMI total effects by selected ages see **Figure 2**.*

observed in PLATINO study, conducted in Latin American cities (49). The two studies differ in terms of environment, lifestyle and adiposity patterns, but their modifying effect on the BMI-AO association was not reported. SAPALDIA and comparable cohorts previously pointed to important interactions between BMI, physical activity and air pollution with regard to FEV1/FVC and FEF2575 (50–53). In contrast to the BOLD and PLATINO studies, this MR study was based on pre-bronchodilation spirometry. But the observed causal effects of BMI are possibly valid for post-bronchodilation LF, because results did not change after excluding asthmatics (54, 55).

Composite Nature of BMI Explains the Discrepancy Between Causal and Observational Effects

The current novel results are consistent with confounding in observational obesity- airflow obstruction links. MR and observational regression coefficients were consistent in direction for FEV1/FVC, but not for FEF2575. These parameter-specific differences between observational and causal effects could reflect differences in unmeasured positive confounders. FEV1/FVC and FEF2575, with potentially different etiology, may have different confounders with regard to the association with BMI. The sparsity of model, which included a minimal set of covariates, may be responsible in part for the large difference between observed and causal BMI effects. We cannot exclude entirely that the observed causal interaction with Age may be the result of confounding.

But residual confounding unlikely explains most of the observed difference between causal and observational associations. Another possible explanation of this discrepancy in our data is the composite nature of BMI, which is well-known to be an imprecise measure of different adiposity phenotypes (56), each with distinct genetic and non-genetic components, the contribution of which may vary over the life course. This is a form of measurement error with regard to the true adiposity measure and susceptible time window of interest. In MR studies it is usually assumed an exposure has the same impact on health outcomes, regardless of whether it is due to genetics, or to other sources. This may only hold true for well-defined biological traits, but not for composite exposures like BMI. By refitting the second stage IV model including terms for both, $BMI_{instrumented}$ (genetically determined BMI) and $BMI_{residual}$ (non-genetically determined BMI), we assumed a measurement error model that considers misclassification of the true adiposity measure of interest (see **Supplementary Material** for a more formalized

TABLE 3 | Causal effects[a] of BMI on FEV1/FVC and FEF2575 in predictive and in long-term cross-sectional models.

	N	β_{c1}, β_{c2}	SE	*p*-value
FEV1/FVC				
Predictive model				
BMI main effect log(BMI$_{s1s2}$)→ FEV1/FVC$_{s2,s3}$	2,853	–0.468	0.455	0.300
BMI*Age interaction effect log(BMI$_{s1s2}$):Age$_{s1s2}$→ FEV1/FVC$_{s2,s3}$		0.011	0.015	0.490
Long-term cross-sectional model				
BMI main effect log(BMI$_{s1s2s3}$)→ FEV1/FVC$_{s1,s2,s3}$	2,731	–0.479	0.565	0.400
BMI*Age interaction effect log(BMI$_{s1s2s3}$):Age$_{s1,s2,s3}$→ FEV1/FVC$_{s1,s2,s3}$		0.006	0.016	0.730
FEF2575				
Predictive model				
BMI main effect log(BMI$_{s1s2}$)→ FEF2575$_{s2,s3}$	2,850	–9.932	6.622	0.134
BMI*Age interaction effect log(BMI$_{s1s2}$):Age$_{s1s2}$→ FEF2575$_{s2,s3}$		0.269	0.225	0.231
Long-term cross-sectional model				
BMI main effect log(BMI$_{s,1s2,s3}$)→ FEF2575$_{s1,s2,s3}$	2,728	–9.111	8.117	0.262
BMI*Age interaction effect log(BMI$_{s1,s2,s3}$):Age$_{s1,s2,s3}$→ FEF2575$_{s1,s2,s3}$		0.163	0.234	0.486

Childhood BMI genetic score: (Felix; 12 SNPs).

[a] *β_{c1}, causal BMI main effect per one BMI-increasing allele; β_{c2}, causal BMI*Age interaction effect per one BMI-increasing allele.*

The negative sign of the causal main effect means that, keeping all other predictors fixed, at age 18 (which has been chosen as the origin in our analysis) BMI has a causal negative effect on LF. The positive sign of the Age×BMI causal interactive effect implies that, as age increases, the detrimental effect of BMI on LF decreases. As a consequence, the total effect of BMI becomes null at middle ages and protective at older ages.

illustration). The fact that BMI$_{residual}$ showed association with FEF2575 but not with FEV1/FVC, consistent with our observational analysis, supports our measurement error model and points to different effects of non-genetically determined BMI on FEV1/FVC and FEF2575. It is conceivable that genetically determined BMI has negative causal effect, while non-genetically determined BMI has positive effect, and the measurement error due to the metric "BMI" as a mixture of the two components can result in such a discrepancy. A recent metabolomics study reported that genetic score of BMI predicted actual BMI but not the metabolic signature of obesity, indicating that the genetic score captures anthropometric phenotype rather than obesity as a disease trait (57).

Given: (a) the observed negative effect of the genetic, but not of the non-genetic, component of BMI on lung function, (b) that the causal BMI effects were strongest for long-term cross-sectional models, (c) that genetic scores derived from SNPs associated with BMI in childhood led to stronger causal BMI effects, and (d) the observed BMI gene score-age interaction with inverse associations in the younger age groups, our results are consistent with the hypothesis that:

1. BMI in childhood impacts on lung function growth and affects the level of lung function in the first half of life (17), thereby leading to lower levels of attained lung function later in life and increasing the risk of chronic respiratory diseases.
2. BMI in adulthood is increasingly (with age) likely to reflect lifestyle rather than genetic background, which may lead to a different phenotype not well-captured by genetic instruments. This phenotype may have no, or even a positive, effect on lung function, following current discussions about what should be considered a healthy BMI cutoff for older persons.

Age-Dependent Causal Effects of BMI: Life Course Perspective of Lung Function

As some SNPs were reported to have specific effects on BMI in childhood or divergent BMI effects across the life course (25), the current results may point to specific BMI-related pathways affecting lung function early in life. Besides age-specific genetic effects on BMI (25), age-related differences in the distribution of fat and muscle mass and also in their association with the course of lung function have been reported (56, 58). These age-related differences may reflect changes in gene-environment interactions and the relative contribution of heritability and lifestyle to BMI over the life course (59, 60). The relative contribution of genetically determined BMI to lung function may decrease with aging and the accumulation of molecular damage due to BMI, determined by lifestyle and environmental risks may become more relevant.

Besides the above argued potential effect of BMI in early childhood on lung function growth and its trajectories into

TABLE 4 | Causal effects[a] of BMI on FEV1/FVC and FEF2575 in predictive and in long-term cross-sectional models.

	N	β_{c1}, β_{c2}	SE	*p*-value
FEV1/FVC				
Predictive model				
BMI main effect $\log(BMI_{s1s2}) \rightarrow FEV1/FVC_{s2,s3}$	2,853	–0.140	0.117	0.233
BMI*Age interaction effect $\log(BMI_{s1s2}){:}Age_{s1s2} \rightarrow FEV1/FVC_{s2,s3}$		0.005	0.004	0.255
Long-term cross-sectional model				
BMI main effect $\log(BMI_{s1s2s3}) \rightarrow FEV1/FVC_{s1,s2,s3}$	2,731	–0.226	0.132	0.088
BMI*Age interaction effect $\log(BMI_{s1s2s3}){:}Age_{s1,s2,s3} \rightarrow FEV1/FVC_{s1,s2,s3}$		0.005	0.003	0.173
FEF2575				
Predictive model				
BMI main effect $\log(BMI_{s1s2}) \rightarrow FEF2575_{s2,s3}$	2,850	–2.715	1.594	0.089
BMI*Age interaction effect $\log(BMI_{s1s2}){:}Age_{s1s2} \rightarrow FEF2575_{s2,s3}$		0.087	0.056	0.118
Long-term cross-sectional model				
BMI main effect $\log(BMI_{s1,s2,s3}) \rightarrow FEF2575_{s1,s2,s3}$	2,728	–3.073	1.885	0.103
BMI*Age interaction effect $\log(BMI_{s1,s2,s3}){:}Age_{s1,s2,s3} \rightarrow FEF2575_{s1,s2,s3}$		0.076	0.056	0.175

BMI genetic score: (Yengo; 862 SNPs).

[a] *β_{c1}, causal BMI main effect per one BMI-increasing allele; β_{c2}, causal BMI*Age interaction effect per one BMI-increasing allele.*

The negative sign of the causal main effect means that, keeping all other predictors fixed, at age 18 (which has been chosen as the origin in our analysis) BMI has a causal negative effect on LF. The positive sign of the Age × BMI causal interactive effect implies that, as age increases, the detrimental effect of BMI on LF decreases. As a consequence, the total effect of BMI becomes null at middle ages and protective at older ages.

adulthood, additional, not mutually exclusive explanations for the observed evidence of causal Age × BMI interactive effects on LF outcomes apply. *First*, it may be a chance finding. *Second*, the results may reflect age-related differences in prevalence and severity of AO. According to the obesity paradox in COPD, excess weight has an adverse effect on the disease course in the early stages. But at more advanced stages for the same degree of AO, obese COPD patients fare better on average than non-obese patients with regard to mortality and hospital admission (4). BMI was positively associated with FEF2575/FVC in heavy smokers with AO (20). *Third*, age-related changes in inflammation, immunologic responses and mechanical lung properties could alter the susceptibility of the airways to obesity (61). Challenges in interpreting low FEV1/FVC in the elderly have been discussed (62). *Fourth*, the observed age-interaction could in part be explained by survivor bias, if survivors with high BMI are those most resistant to the adverse LF effects of obesity. *Finally*, this study does not allow differentiating between causal biological BMI effects on LF and causal BMI effects on phenotypes that are comorbid with LF. The increasing number, but with decreasing effect size, of BMI associated SNPs arising from ever larger GWAS is likely to increase the number of comorbidity signals (24). Although the genetic scores we used in this study did not show association with height, we cannot rule out that the causal BMI effects are in part due to height.

BMI Effects on FEF2575, a Potential Early Indicator of Small Airway Dysfunction

A causal effect on FEF2575 is of interest, as small airways are frequently involved at an early stage in COPD and asthma (18) and they have been shown to be adversely affected by weight and growth patterns in early childhood (17). Adverse peripheral airway effects of excess weight were demonstrated by impulse oscillometry (15, 63). The insensitivity of spirometry to peripheral airway abnormalities may in part explain the contradictory findings on the BMI- LF association (63). The value of FEF2575 for early detection of small airway dysfunction has been questioned (64, 65), and attributed to the parameter's wide variability in healthy subjects (66). But several aspects of this study justify the consideration of FEF2575 as an independent phenotype. The partial correlations with FEV1, FVC, and FEV1/FVC were between 0.182 (FEF2575: FVC at SAP1) and 0.868 ($FEF2575_{s1,s2,s3}$: FEV1/FVC $_{s1,s2,s3}$) across phenotypes and time points. The intra-individual variability of FEF2575 was smaller than that of FEV1/FVC. FEF2575 was previously correlated with functional imaging assessment of small airway function (67). In obliterative bronchiolitis, the paradigm of small airway disease, FEF2575 is considered a sensitive diagnostic marker (68). FEF2575 was correlated with smooth muscle α-actin in the small airways, a marker of airway remodeling (69), and predicted mortality from COPD after 20 years of follow-up (70).

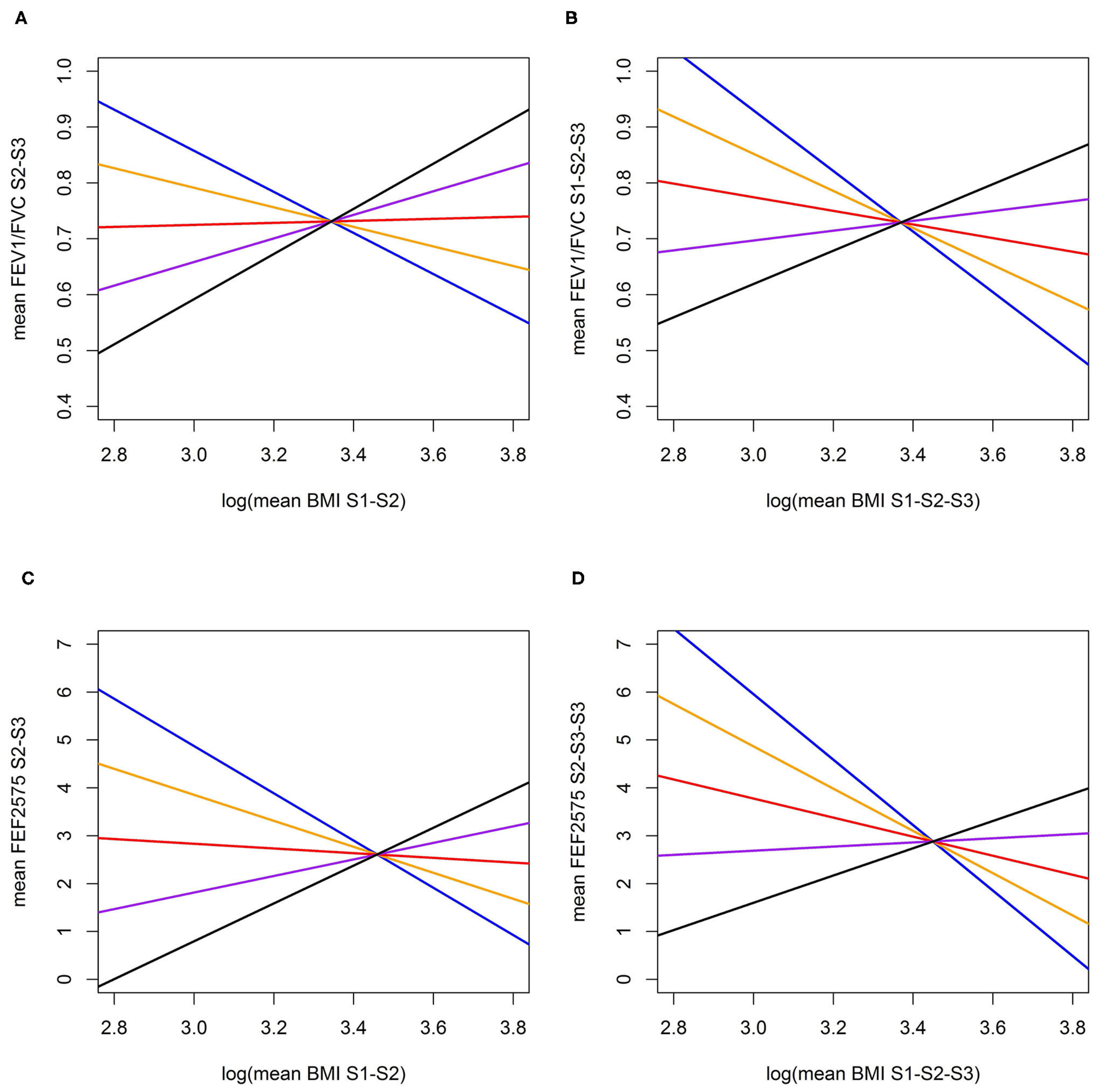


FIGURE 2 | Predictive and long-term cross-sectional total causal (26) effect of BMI on FEV1/FVC **(A,B)** and FEF2575 **(C,D)** for a reference individual (Male, Height = 175 cm., Never Smoker) at specific ages (Blue: at age 28, Orange: at age 38, Red: at age 48; Purple: at age 58; Black: at age 68). **(A)** Total predictive effect of BMI (log mean over SAP1-SAP2) on FEV1/FVC ratio (mean over SAP2-SAP3); **(B)** Total long-term cross-sectional effect of BMI (log mean over SAP1-SAP2-SAP3) on FEV1/FVC ratio (mean over SAP1-SAP2-SAP3); **(C)** Total predictive effect of BMI (log mean over SAP1-SAP2) on FEF2575 (mean over SAP1-SAP2); **(D)** Total long-term cross-sectional effect of BMI (log mean over SAP1-SAP2-SAP3) on FEF2575 (mean over SAP1-SAP2-SAP3).

Strengths and Limitations

As in any study, results have to be evaluated in the light of strengths and limitations. The assumptions of MR appeared to be satisfied, strengthening the choice of carrying out an MR study. The MR assumptions could still be violated by unobserved confounders, though. Statistical power of this study was limited, but the choice of considering medium- and long-term averages, for both exposure and outcome, alleviated this problem and allowed studying the stability and age dependency of causal effects. This study did not confirm previously reported causal effects of BMI on FEV1 and FVC (2) that were strictly cross-sectional and based on data from a single time point.

TABLE 5 | Observational associations of BMI with FEV1/FVC and FEF2575 in predictive and in long-term cross-sectional models.

	N	β_1, β_2^a	SE	*p*-value
FEV1/FVC				
Predictive model				
BMI main effect log(BMI_{s1s2})→ $FEV1/FVC_{s2,s3}$	2,853	−0.006	0.025	0.803
BMI*Age interaction effect log(BMI_{s1s2}):Age_{s1s2}→ $FEV1/FVC_{s2,s3}$		0.001	0.001	0.378
Long-term cross-sectional model				
BMI main effect log(BMI_{s1s2s3})→ $FEV1/FVC_{s1,s2,s3}$	2,731	−0.047	0.026	0.076
BMI*Age interaction effect log(BMI_{s1s2s3}):$Age_{s1,s2,s3}$→ $FEV1/FVC_{s1,s2,s3}$		0.001	0.001	0.179
FEF2575				
Predictive model				
BMI main effect log(BMI_{s1s2})→ $FEF2575_{s2,s3}$		0.746	0.330	0.024
BMI*Age interaction effect log(BMI_{s1s2}):Age_{s1s2}→ $FEF2575_{s2,s3}$	2,850	−0.019	0.011	0.085
Long-term cross-sectional model				
BMI main effect log($BMI_{s1,s2,s3}$)→ $FEF2575_{s1,s2,s3}$		0.525	0.375	0.162
BMI*Age interaction effect log($BMI_{s1,s2,s3}$):$Age_{s1,s2,s3}$→ $FEF2575_{s1,s2,s3}$	2,728	−0.017	0.011	0.114

$^a\beta_{c1}$, *associational BMI main effect;* β_2, *associational BMI*Age interaction effect.*

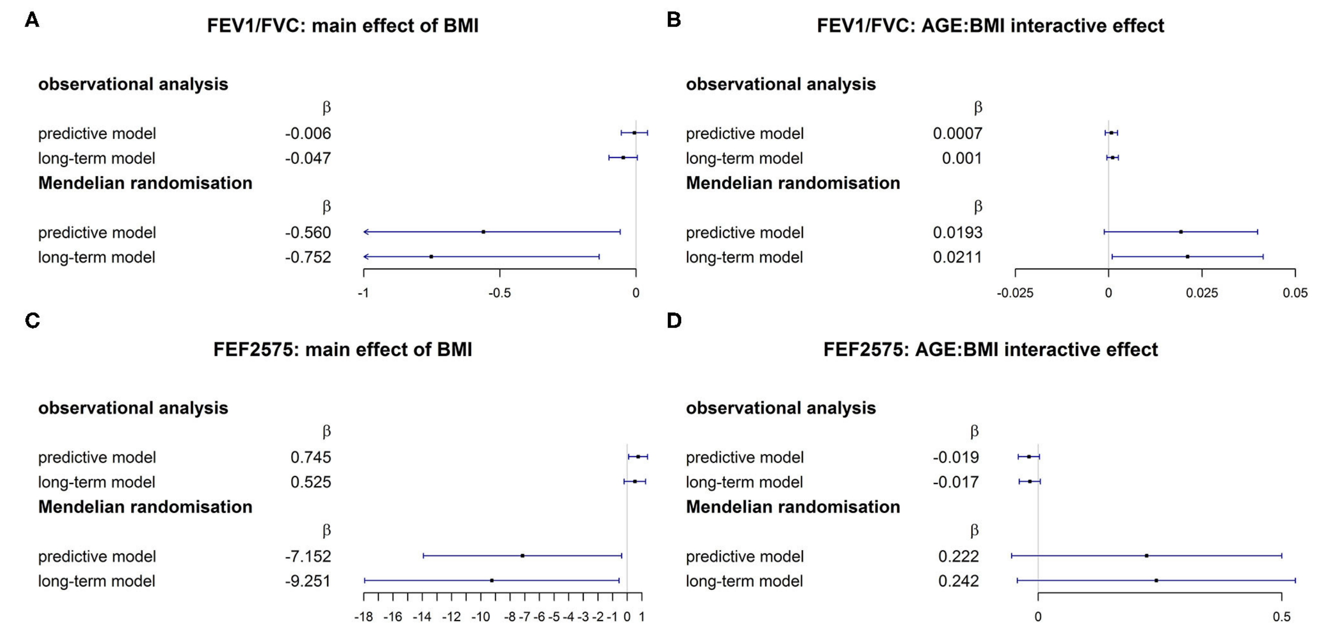


FIGURE 3 | Comparison of associational and MR causal (26) effects for FEV1/FVC [**(A)**: main effect of BMI; **(B)** AGE × BMI interaction] and FEF2575 [**(C)** main effect of BMI; **(D)** AGE × BMI interaction].

The sample size did not allow studying the causality of BMI effects in respiratory health subgroups, e.g., COPD patients. But the detailed participant characterization in SAPALDIA, an internationally renowned respiratory cohort (33, 71–73) allowed excluding persons with a self-report of doctor-diagnosed asthma at any point during 20 years. Additional limitations include the restriction of LF to pre-bronchodilation, whereas post-bronchodilation FEV1/FVC forms the basis for diagnosing COPD (74), and of obesity assessment to BMI in the absence of visceral adiposity indicators (68). No measurements of BMI in childhood of SAPALDIA participants were available, which would have allowed to instrument childhood BMI. We were limited in assessing longitudinal effects of BMI or its change on LF decline in adults. Genetic variants to instrument BMI change do not exist. Many more than three time points would be needed to truly assess causal BMI effects on LF change over time. But biological pathways underlying level of LF and LF decline may differ. BMI and lung function averaged over a certain time period as in this study may be better measures for assessing chronic long-term associations between the two, given the intra-individual volatility of these parameters over time. This is supported by the fact that we found stronger associations by using medium- and long-term averages, compared to single time point associations, and a higher predictive ability when compared with that of BMI change with lung function change. We acknowledge that by taking averages of BMI and averages of lung function we are faced with the problem that persons with higher BMI at baseline and lower BMI at follow-up may have the same long-term BMI average as persons with lower BMI at baseline and higher BMI at follow-up. The same caveat may apply for two people with the same average of lung function. Because this adds to the problem of reverse causation (and would most likely bias the associations toward the null), we were also taking a predictive approach of investigating the associations of BMI averaged over SAP1 and SAP2 with lung function averaged over SAP2 and SAP3. As another limitation, we acknowledge that our study did not investigate non-linearity of the causal effects. Finally, we cannot exclude that the complete case analysis led to some bias due to other sources of missingness, although attrition seems to be by far the most important mechanism generating missingness in our data. Although the Inverse Probability Weighted analysis considered bias due to the most important attrition factor, and for that matter a major mortality determinant, namely smoking, not all factors influencing non-participation could be considered. However, the attrition bias would likely bias the associations toward the null, given that the dropouts would more likely have experienced increase in BMI and decline in LF.

CONCLUSION

The results of this study suggest that AO and possibly small airways disease may, in part, be the result of excess weight in young and middle-aged adults, or even in children. The results need to be confirmed in the context of a larger MR study involving tests reflecting small airway dysfunction and more specific parameters for adiposity at different stages in life. In addition, the study points to important methodological needs in future studies on the causal effects of obesity and lung health, namely to consider adiposity- and lung phenotype-specific associations from a life course perspective and to derive and apply genetic instruments reflecting more specific obesity phenotypes.

DATA AVAILABILITY STATEMENT

Data included in this manuscript is available from the corresponding author upon a justified request.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Swiss Academy of Medical Sciences and the Cantonal Ethics Committees. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

NP-H designed the SAPALDIA cohort. NP-H, AJ, CS, and GL developed the analysis plan for this paper. AJ and GL conducted the statistical analysis. NP-H and GL drafted the manuscript. NP-H, MI, DS, MP, PS, and RB conducted the SAPALDIA study. All authors read and corrected the manuscript and approved of the final version.

FUNDING

This work was supported by the European Union's Horizon 2020 Research and Innovation programme under grant agreement no. 633212 (ALEC Study). SAPALDIA has been supported since its onset by the Swiss National Science Foundation (SNF Grant Nos. 33CS30-177506/1, 33CS30-148470/2, 33CS30-148470/1, 33CSCO-134276/1, 33CSCO-108796, 324730_135673, 3247BO-104283, 3247BO-104288, 3247BO-104284, 3247-065896, 3100-059302, 3200-052720, 3200-042532, 4026-028099, PMPDP3_129021/1, and PMPDP3_141671/1); the Swiss Federal Office for the Environment; the Federal Office of Public Health; the Federal Office of Roads and Transport; the canton's government of Aargau, Basel-Stadt, Basel-Land, Geneva, Luzern, Ticino, Valais, and Zürich; the Swiss Lung League; the canton's Lung League of Basel Stadt/Basel Landschaft, Geneva, Ticino, Valais, Graubünden and Zurich; Stiftung ehemals Bündner Heilstätten; SUVA; Freiwillige Akademische Gesellschaft; UBS Wealth Foundation; European Commission [grant agreement . . . (GABRIEL)]; Wellcome Trust (grant agreement . . .).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.584955/full#supplementary-material

REFERENCES

1. Franssen FME. Overweight and obesity are risk factors for COPD misdiagnosis and overtreatment. *Chest.* (2014) 146:1426–8. doi: 10.1378/chest.14-1308
2. Skaaby T, Taylor AE, Thuesen BH, Jacobsen RK, Friedrich N, Mollehave LT, et al. Estimating the causal effect of body mass index on hay fever, asthma and lung function using Mendelian randomization. *Allergy.* (2018) 73:153–64. doi: 10.1111/all.13242
3. Uppalapati A, Gogineni S, Espiritu JR. Association between body mass index (BMI) and fraction of exhaled nitric oxide (FeNO) levels in the national health and nutrition examination survey (NHANES) 2007-2010. *Obes Res Clin Pract.* (2016) 10:652–8. doi: 10.1016/j.orcp.2015.11.006
4. Zewari S, Vos P, van den Elshout F, Dekhuijzen R, Heijdra Y. Obesity in COPD: revealed and unrevealed issues. *COPD.* (2017) 14:663–73. doi: 10.1080/15412555.2017.1383978
5. Jones RL, Nzekwu MM. The effects of body mass index on lung volumes. *Chest.* (2006) 130:827–33. doi: 10.1378/chest.130.3.827
6. Sin DD, Sutherland ER. Obesity and the lung: 4. Obesity and asthma. *Thorax.* (2008) 63:1018–23. doi: 10.1136/thx.2007.086819
7. Marcon A, Corsico A, Cazzoletti L, Bugiani M, Accordini S, Almar E, et al. Body mass index, weight gain, and other determinants of lung function decline in adult asthma. *J Allergy Clin Immunol.* (2009) 123:1069–74:74.e1–4. doi: 10.1016/j.jaci.2009.01.040
8. Al Ghobain M. The effect of obesity on spirometry tests among healthy non-smoking adults. *BMC Pulm Med.* (2012) 12:10. doi: 10.1186/1471-2466-12-10
9. Banerjee J, Roy A, Singhamahapatra A, Dey PK, Ghosal A, Das A. Association of body mass index (BMI) with lung function parameters in non-asthmatics identified by spirometric protocols. *J Clin Diagn Res.* (2014) 8:12–4. doi: 10.7860/JCDR/2014/7306.3993
10. Fenger RV, Gonzalez-Quintela A, Vidal C, Husemoen LL, Skaaby T, Thuesen BH, et al. The longitudinal relationship of changes of adiposity to changes in pulmonary function and risk of asthma in a general adult population. *BMC Pulm Med.* (2014) 14:208. doi: 10.1186/1471-2466-14-208
11. Hanson C, Rutten EP, Wouters EF, Rennard S. Influence of diet and obesity on COPD development and outcomes. *Int J Chron Obstruct Pulmon Dis.* (2014) 9:723–33. doi: 10.2147/COPD.S50111
12. Vanfleteren LE, Lamprecht B, Studnicka M, Kaiser B, Gnatiuc L, Burney P, et al. Body mass index and chronic airflow limitation in a worldwide population-based study. *Chron Respir Dis.* (2016) 13:90–101. doi: 10.1177/1479972315626012
13. Maniscalco M, Zamparelli AS, Vitale DF, Faraone S, Molino A, Zedda A, et al. Long-term effect of weight loss induced by bariatric surgery on asthma control and health related quality of life in asthmatic patients with severe obesity: a pilot study. *Respir Med.* (2017) 130:69–74. doi: 10.1016/j.rmed.2017.06.010
14. Mancuso P. Obesity and lung inflammation. *J Appl Physiol.* (2010) 108:722–8. doi: 10.1152/japplphysiol.00781.2009
15. van de Kant KD, Paredi P, Meah S, Kalsi HS, Barnes PJ, Usmani OS. The effect of body weight on distal airway function and airway inflammation. *Obes Res Clin Pract.* (2016) 10:564–73. doi: 10.1016/j.orcp.2015.10.005
16. Salome CM, King GG, Berend N. Physiology of obesity and effects on lung function. *J Appl Physiol.* (2010) 108:206–11. doi: 10.1152/japplphysiol.00694.2009
17. Casas M, den Dekker HT, Kruithof CJ, Reiss IK, Vrijheid M, Sunyer J, et al. The effect of early growth patterns and lung function on the development of childhood asthma: a population based study. *Thorax.* (2018) 73:1137–45. doi: 10.1136/thoraxjnl-2017-211216
18. Dilektasli AG, Porszasz J, Casaburi R, Stringer WW, Bhatt SP, Pak Y, et al. A novel spirometric measure identifies mild COPD unidentified by standard criteria. *Chest.* (2016) 150:1080–90. doi: 10.1016/j.chest.2016.06.047
19. McNulty W, Usmani OS. Techniques of assessing small airways dysfunction. *Eur Clin Respir J.* (2014) 1:25898. doi: 10.3402/ecrj.v1.25898
20. Abston E, Comellas A, Reed RM, Kim V, Wise RA, Brower R, et al. Higher BMI is associated with higher expiratory airflow normalised for lung volume (FEF25-75/FVC) in COPD. *BMJ Open Respir Res.* (2017) 4:e000231. doi: 10.1136/bmjresp-2017-000231
21. Yao TC, Tsai HJ, Chang SW, Chung RH, Hsu JY, Tsai MH, et al. Obesity disproportionately impacts lung volumes, airflow and exhaled nitric oxide in children. *PLoS ONE.* (2017) 12:e0174691. doi: 10.1371/journal.pone.0174691
22. Cibella F, Bruno A, Cuttitta G, Bucchieri S, Melis MR, De Cantis S, et al. An elevated body mass index increases lung volume but reduces airflow in Italian schoolchildren. *PLoS ONE.* (2015) 10:e0127154. doi: 10.1371/journal.pone.0127154
23. Haycock PC, Burgess S, Wade KH, Bowden J, Relton C, Davey Smith G. Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am J Clin Nutr.* (2016) 103:965–78. doi: 10.3945/ajcn.115.118216
24. Yengo L, Sidorenko J, Kemper KE, Zheng Z, Wood AR, Weedon MN, et al. Meta-analysis of genome-wide association studies for height and body mass index in approximately 700000 individuals of European ancestry. *Hum Mol Genet.* (2018) 27:3641–9. doi: 10.1093/hmg/ddy271
25. Felix JF, Bradfield JP, Monnereau C, van der Valk RJ, Stergiakouli E, Chesi A, et al. Genome-wide association analysis identifies three new susceptibility loci for childhood body mass index. *Hum Mol Genet.* (2016) 25:389–403. doi: 10.1093/hmg/ddv472
26. Speliotes EK, Willer CJ, Berndt SI, Monda KL, Thorleifsson G, Jackson AU, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet.* (2010) 42:937–48. doi: 10.1038/ng.686
27. Eschenbacher WL. Defining airflow obstruction. *Chronic Obstr Pulm Dis.* (2016) 3:515–8. doi: 10.15326/jcopdf.3.2.2015.0166
28. Rao DR, Gaffin JM, Baxi SN, Sheehan WJ, Hoffman EB, Phipatanakul W. The utility of forced expiratory flow between 25% and 75% of vital capacity in predicting childhood asthma morbidity and severity. *J Asthma.* (2012) 49:586–92. doi: 10.3109/02770903.2012.690481
29. Tagiyeva N, Sadhra S, Mohammed N, Fielding S, Devereux G, Teo E, et al. Occupational airborne exposure in relation to chronic obstructive pulmonary disease (COPD) and lung function in individuals without childhood wheezing illness: a 50-year cohort study. *Environ Res.* (2017) 153:126–34. doi: 10.1016/j.envres.2016.11.018
30. Williamson PA, Clearie K, Menzies D, Vaidyanathan S, Lipworth BJ. Assessment of small-airways disease using alveolar nitric oxide and impulse oscillometry in asthma and COPD. *Lung.* (2011) 189:121–9. doi: 10.1007/s00408-010-9275-y
31. Ackermann-Liebrich U, Kuna-Dibbert B, Probst-Hensch NM, Schindler C, Felber Dietrich D, Stutz EZ, et al. Follow-up of the Swiss cohort study on air pollution and lung diseases in adults (SAPALDIA 2) 1991-2003: methods and characterization of participants. *Soz Praventivmed.* (2005) 50:245–63. doi: 10.1007/s00038-005-4075-5
32. Miller MR, Hankinson J, Brusasco V, Burgos F, Casaburi R, Coates A, et al. Standardisation of spirometry. *Eur Respir J.* (2005) 26:319–38. doi: 10.1183/09031936.05.00034805
33. Bridevaux PO, Dupuis-Lozeron E, Schindler C, Keidel D, Gerbase MW, Probst-Hensch NM, et al. Spirometer replacement and serial lung function measurements in population studies: results from the SAPALDIA study. *Am J Epidemiol.* (2015) 181:752–61. doi: 10.1093/aje/kwu352
34. Imboden M, Bouzigon E, Curjuric I, Ramasamy A, Kumar A, Hancock DB, et al. Genome-wide association study of lung function decline in adults with and without asthma. *J Allergy Clin Immunol.* (2012) 129:1218–28. doi: 10.1016/j.jaci.2012.01.074
35. Jackson VE, Latourelle JC, Wain LV, Smith AV, Grove ML, Bartz TM, et al. Meta-analysis of exome array data identifies six novel genetic loci for lung function. *Wellcome Open Res.* (2018) 3:4. doi: 10.12688/wellcomeopenres.12583.3
36. Delaneau O, Marchini J, Genomes Project C, Genomes Project C. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun.* (2014) 5:3934. doi: 10.1038/ncomms4934
37. Fuchsberger C, Abecasis GR, Hinds DA. minimac2: faster genotype imputation. *Bioinformatics.* (2015) 31:782–4. doi: 10.1093/bioinformatics/btu704
38. Millard LA, Davies NM, Timpson NJ, Tilling K, Flach PA, Davey Smith G. MR-PheWAS: hypothesis prioritization among potential causal effects of body mass index on many outcomes, using Mendelian randomization. *Sci Rep.* (2015) 5:16645. doi: 10.1038/srep16645
39. VanderWeele TJ, Tchetgen Tchetgen EJ, Cornelis M, Kraft P. Methodological challenges in Mendelian randomization. *Epidemiology.* (2014) 25:427–35. doi: 10.1097/EDE.0000000000000081

40. Staley JR, Blackshaw J, Kamat MA, Ellis S, Surendran P, Sun BB, et al. PhenoScanner: a database of human genotype-phenotype associations. *Bioinformatics*. (2016) 32:3207–9. doi: 10.1093/bioinformatics/btw373
41. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol.* (2017) 46:1734–9. doi: 10.1093/ije/dyx034
42. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. *Stat Methods Med Res.* (2007) 16:309–30. doi: 10.1177/0962280206077743
43. Burgess S, Thompson SG. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. New York, NY: CRC Press. (2015). doi: 10.1201/b18084
44. Bun M, Harrison T. *OLS and IV Estimation of Regression Models Including Endogenous Interaction Terms.* Philadelphia, PA: School of Economics Working Paper Series; LeBow College of Business; Drexel University (2014) p. 2014-3.
45. Thorgeirsson TE, Gudbjartsson DF, Surakka I, Vink JM, Amin N, Geller F, et al. Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat Genet.* (2010) 42:448–53. doi: 10.1038/ng.573
46. Tobacco, Genetics C. Genome-wide meta-analyses identify multiple loci associated with smoking behavior. *Nature genetics.* (2010) 42:441–7. doi: 10.1038/ng.571
47. Liu JZ, Tozzi F, Waterworth DM, Pillai SG, Muglia P, Middleton L, et al. Meta-analysis and imputation refines the association of 15q25 with smoking quantity. *Nat Genet.* (2010) 42:436–40. doi: 10.1038/ng.572
48. VanderWeele TJ. Commentary: on causes, causal inference, and potential outcomes. *Int J Epidemiol.* (2016) 45:1809–16. doi: 10.1093/ije/dyw230
49. Montes de Oca M, Talamo C, Perez-Padilla R, Jardim JR, Muino A, Lopez MV, et al. Chronic obstructive pulmonary disease and body mass index in five Latin America cities: the PLATINO study. *Respir Med.* (2008) 102:642–50. doi: 10.1016/j.rmed.2007.12.025
50. Curjuric I, Imboden M, Adam M, Bettschart RW, Gerbase MW, Kunzli N, et al. Serum bilirubin is associated with lung function in a Swiss general population sample. *Eur Respir J.* (2014) 43:1278–88. doi: 10.1183/09031936.00055813
51. Downs SH, Schindler C, Liu LJ, Keidel D, Bayer-Oglesby L, Brutsche MH, et al. Reduced exposure to PM10 and attenuated age-related decline in lung function. *N Engl J Med.* (2007) 357:2338–47. doi: 10.1056/NEJMoa073625
52. Schikowski T, Schaffner E, Meier F, Phuleria HC, Vierkotter A, Schindler C, et al. Improved air quality and attenuated lung function decline: modification by obesity in the SAPALDIA cohort. *Environ Health Perspect.* (2013) 121:1034–9. doi: 10.1289/ehp.1206145
53. Laeremans M, Dons E, Avila-Palencia I, Carrasco-Turigas G, Orjuela-Mendoza JP, Anaya-Boig E, et al. Black carbon reduces the beneficial effect of physical activity on lung function. *Med Sic Sports Exerc.* (2018) 50:1875–81. doi: 10.1249/MSS.0000000000001632
54. Rogliani P, Ora J, Puxeddu E, Cazzola M. Airflow obstruction: is it asthma or is it COPD? *Int J Chron Obstruct Pulmon Dis.* (2016) 11:3007–13. doi: 10.2147/COPD.S54927
55. Postma DS, Reddel HK, Ten Hacken NH, van den Berge M. Asthma and chronic obstructive pulmonary disease: similarities and differences. *Clin Chest Med.* (2014) 35:143–56. doi: 10.1016/j.ccm.2013.09.010
56. Monnereau C, Santos S, van der Lugt A, Jaddoe VWV, Felix JF. Associations of adult genetic risk scores for adiposity with childhood abdominal, liver and pericardial fat assessed by magnetic resonance imaging. *Int J Obes.* (2018) 42:897–904. doi: 10.1038/ijo.2017.302
57. Cirulli ET, Guo L, Leon Swisher C, Shah N, Huang L, Napier LA, et al. Profound perturbation of the metabolome in obesity is associated with health risk. *Cell Metab.* (2019) 29:488–500.e2. doi: 10.1016/j.cmet.2018.09.022
58. Tsao YC, Lee YY, Chen JY, Yeh WC, Chuang CH, Yu W, et al. Gender- and age-specific associations between body fat composition and c-reactive protein with lung function: a cross-sectional study. *Sci Rep.* (2019) 9:384. doi: 10.1038/s41598-018-36860-9
59. Korkeila M, Kaprio J, Rissanen A, Koskenvuo M. Effects of gender and age on the heritability of body mass index. *Int J Obes.* (1991) 15:647–54.
60. He L, Sillanpaa MJ, Silventoinen K, Kaprio J, Pitkaniemi J. Estimating Modifying Effect of Age on Genetic and Environmental Variance Components in Twin Models. *Genetics.* (2016) 202:1313–28. doi: 10.1534/genetics.115.183905
61. Knudson RJ, Clark DF, Kennedy TC, Knudson DE. Effect of aging alone on mechanical properties of the normal adult human lung. *J Appl Physiol Respir Environ Exerc Physiol.* (1977) 43:1054–62. doi: 10.1152/jappl.1977.43.6.1054
62. Pedone C, Giua R, Scichilone N, Forastiere F, Bellia V, Antonelli Incalzi R. Difference in mortality risk in elderly people with bronchial obstruction diagnosed using a fixed cutoff or the lower limit of normal of the FEV1/FVC ratio. *Respiration.* (2017) 94:424–30. doi: 10.1159/000479285
63. Robinson PD, King GG, Sears MR, Hong CY, Hancox RJ. Determinants of peripheral airway function in adults with and without asthma. *Respirology.* (2017) 22:1110–7. doi: 10.1111/resp.13045
64. Quanjer PH, Weiner DJ, Pretto JJ, Brazzale DJ, Boros PW. Measurement of FEF25-75% and FEF75% does not contribute to clinical decision making. *Eur Respir J.* (2014) 43:1051–8. doi: 10.1183/09031936.00128113
65. Pellegrino R, Viegi G, Brusasco V, Crapo RO, Burgos F, Casaburi R, et al. Interpretative strategies for lung function tests. *Eur Respir J.* (2005) 26:948–68. doi: 10.1183/09031936.05.00035205
66. Kanchongkittiphon W, Gaffin JM, Kopel L, Petty CR, Bollinger ME, Miller RL, et al. Association of FEF25%-75% and bronchodilator reversibility with asthma control and asthma morbidity in inner-city children with asthma. *Ann Allergy Asthma Immunol.* (2016) 117:97–9. doi: 10.1016/j.anai.2016.04.029
67. Jain N, Covar RA, Gleason MC, Newell JD Jr, Gelfand EW, et al. Quantitative computed tomography detects peripheral airway disease in asthmatic children. *Pediatr Pulmonol.* (2005) 40:211–8. doi: 10.1002/ppul.20215
68. Sritippayawan S, Keens TG, Horn MV, Starnes VA, Woo MS. What are the best pulmonary function test parameters for early detection of post-lung transplant bronchiolitis obliterans syndrome in children? *Pediatr Transplant.* (2003) 7:200–3. doi: 10.1034/j.1399-3046.2003.00069.x
69. Bergeron C, Hauber HP, Gotfried M, Newman K, Dhanda R, Servi RJ, et al. Evidence of remodeling in peripheral airways of patients with mild to moderate asthma: effect of hydrofluoroalkane-flunisolide. *J Allergy Clin Immunol.* (2005) 116:983–9. doi: 10.1016/j.jaci.2005.07.029
70. Thomason MJ, Strachan DP. Which spirometric indices best predict subsequent death from chronic obstructive pulmonary disease? *Thorax.* (2000) 55:785–8. doi: 10.1136/thorax.55.9.785
71. Kunzli N, Kuna-Dibbert B, Keidel D, Keller R, Brandli O, Schindler C, et al. Longitudinal validity of spirometers–a challenge in longitudinal studies. *Swiss Med Wkly.* (2005) 135:503–8.
72. Kunzli N, Ackermann-Liebrich U, Keller R, Perruchoud AP, Schindler C. Variability of FVC and FEV1 due to technician, team, device and subject in an eight centre study: three quality control studies in SAPALDIA. Swiss study on air pollution and lung disease in adults. *Eur Respir J.* (1995) 8:371–6. doi: 10.1183/09031936.95.08030371
73. Gerbase MW, Dupuis-Lozeron E, Schindler C, Keidel D, Bridevaux PO, Kriemler S, et al. Agreement between spirometers: a challenge in the follow-up of patients and populations? *Respiration.* (2013) 85:505–14. doi: 10.1159/000346649
74. Marcoa R, Rodrigues DM, Dias M, Ladeira I, Vaz AP, Lima R, et al. Classification of chronic obstructive pulmonary disease (COPD) according to the new global initiative for chronic obstructive lung disease (GOLD) 2017: Comparison with GOLD 2011. *COPD.* (2018) 15:21–6. doi: 10.1080/15412555.2017.1394285

Conflict of Interest: DS obtained funding for grants from the following companies: Astra-Zeneca, PanGas, Wenmann, Curetis, Boston Scientific, Cucassia, and also from Swiss Lung Leagues. DS was paid for lectures including service on speaker's bureaus by the following companies: Astra-Zeneca, Novartis, GSK, Roche, Zambon, Pfizer, Schwabe Pharma and Vifor.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

ORIGINAL RESEARCH
published: 03 June 2021
doi: 10.3389/fpubh.2021.673451

Educational Attainment Decreases the Risk of COVID-19 Severity in the European Population: A Two-Sample Mendelian Randomization Study

Masahiro Yoshikawa[1]* *and Kensuke Asaba*[2]

[1] Division of Laboratory Medicine, Department of Pathology and Microbiology, Nihon University School of Medicine, Tokyo, Japan, [2] Department of Computational Diagnostic Radiology and Preventive Medicine, The University of Tokyo Hospital, Tokyo, Japan

OPEN ACCESS

Edited by:
Bryan Bassig,
National Cancer Institute, National Institutes of Health (NIH), United States

Reviewed by:
Alessandra Macciotta,
University of Turin, Italy
Diana A. van der Plaat,
Imperial College London, United Kingdom
Charles Breeze,
National Institutes of Health (NIH), United States

***Correspondence:**
Masahiro Yoshikawa
myosh-tky@umin.ac.jp

Specialty section:
This article was submitted to Life-Course Epidemiology and Social Inequalities in Health, a section of the journal Frontiers in Public Health

Received: *01 March 2021*
Accepted: *12 April 2021*
Published: *03 June 2021*

Citation:
Yoshikawa M and Asaba K (2021) Educational Attainment Decreases the Risk of COVID-19 Severity in the European Population: A Two-Sample Mendelian Randomization Study. Front. Public Health 9:673451. doi: 10.3389/fpubh.2021.673451

Observational studies have reported that the severity of COVID-19 depends not only on physical conditions but also on socioeconomic status, including educational level. Because educational attainment (EA), which measures the number of years of schooling, is moderately heritable, we investigated the causal association of EA on the risk of COVID-19 severity using the Mendelian randomization (MR) approach. A two-sample MR analysis was performed using publicly available summary-level data sets of genome-wide association studies (GWASs). A total of 235 single-nucleotide polymorphisms (SNPs) were extracted as instrumental variables for the exposure of EA from the Social Science Genetic Association Consortium GWAS summary data of 766,345 participants of European ancestry. The effect of each SNP on the outcome of COVID-19 severity risk was obtained from the GWAS summary data of 1,059,456 participants of European ancestry gathered from the COVID-19 Host Genetics Initiative. Using inverse variance weighted method, our MR study shows that EA was significantly associated with a lower risk of COVID-19 severity (odds ratio per one standard deviation increase in years of schooling, 0.540; 95% confidence interval, 0.376–0.777, $P = 0.0009$). A series of sensitivity analyses showed little evidence of bias. In conclusion, we show for the first time using a two-sample MR approach the associations between higher EA and the lower risk of COVID-19 severity in the European population. However, the genetic or epidemiological mechanisms underlying the association between EA and the risk of COVID-19 severity remain unknown, and further studies are warranted to validate the MR findings and investigate underlying mechanisms.

Keywords: Mendelian randomization, COVID-19, SARS-CoV-2, educational attainment, years of schooling

INTRODUCTION

The coronavirus disease 2019 (COVID-19), caused by a novel coronavirus SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2), was originally reported as an outbreak of atypical pneumonia cases in Wuhan in the Hubei Province of China in December 2019. As of March 2021, the COVID-19 death toll has topped 2.8 million worldwide according to the World Health Organization (1). Serious COVID-19 patients have pneumonia with hypoxia and may be critical with acute respiratory distress syndrome, pulmonary fibrosis, and other organ failures (2).

Observational studies report that the severity of COVID-19 depends not only on physical conditions such as age, cardiovascular disease, and obesity (3–8) but also on socioeconomic status (SES) indicators such as lower incomes and lower educational level among various populations (9–12). In the European population, lower education level was associated with a higher risk of severe COVID-19 cases that were confirmed either at emergency departments or as inpatients and, therefore, likely reflect severe illness as well as a higher risk of asymptomatic COVID-19 cases in a prospective cohort study using UK Biobank data (9). However, traditional observational studies lacking randomization designs are generally prone to bias by various factors, including confounders and reverse causations (13).

Mendelian randomization (MR) is an epidemiological method that mimics the design of randomized controlled studies using single-nucleotide polymorphisms (SNPs) as instrumental variables (IVs) and examines the causal effects of a risk factor on an outcome of interest. Because genetic variants, such as SNPs, are randomly assigned at conception according to Mendel's law, MR studies are not influenced by confounders or reverse causations and can overcome limitations of observational studies (13). Educational attainment (EA) is highly affected by environmental and social factors but is also moderately heritable as shown by genome-wide association studies (GWASs) (14, 15). Therefore, we were motivated to investigate in this study whether EA had a causal effect on the risk of COVID-19 severity using the MR approach.

METHODS AND MATERIALS

Study Design and Data Sources

We conducted a two-sample MR study using publicly available summary statistics from two GWASs to investigate whether EA was associated with risk of COVID-19 severity. In MR analysis, SNPs from the exposure data set are used as IVs. IVs must satisfy the following three assumptions: The IVs are associated with the exposure (IV assumption 1), the IVs affect the outcome only *via* the exposure (IV assumption 2), and the IVs are not associated with measured or unmeasured confounders (IV assumption 3) (16). For the exposure data set of EA, which measured the number of years of schooling that individuals had completed, the SNPs were obtained from the Social Science Genetic Association Consortium's GWAS summary data of 766,345 participants of European ancestry (13), which was a meta-analysis of 70 discovery cohorts (excluding 23andMe) as shown in **Supplementary Table 1**. This data set was publicly available from the MRC IEU Open GWAS database (17) and MR-Base (18) given as GWAS-ID of "ieu-a-1239." For the outcome data set of the risk of COVID-19 severity, the SNPs were obtained from summary-level GWAS data of COVID-19-hg GWAS meta-analyses (round 5) including 14 studies, but excluding the UK Biobank, with a total of 1,059,456 participants (4,792 very severe respiratory confirmed COVID-19 cases and 1,054,664 controls) of European ancestry by the COVID-19 Host Genetics Initiative (19) (**Supplementary Table 1**), which was released on January 18, 2021, and was also publicly available (20). Very severe respiratory confirmed COVID-19 cases were defined as hospitalization for laboratory confirmed SARS-CoV-2 infection with death or respiratory support (20).

Selection of Instrumental Variables

The SNPs were selected from the exposure GWAS summary data as IVs by clumping together all SNPs that were associated with EA at a genome-wide significance threshold ($P < 5.0 \times 10^{-8}$) and were not in linkage disequilibrium ($r^2 < 0.01$ and distance > 10,000 kb) with the other SNPs. Palindromic SNPs with minor allele frequency > 0.42 were excluded from the analyses (16, 21). As a sensitivity analysis, we also excluded all palindromic SNPs regardless of minor allele frequencies (22). We studied only SNPs that were present in both the exposure and outcome GWAS data sets and did not include proxy SNPs in the analysis (22, 23). To evaluate the strength of the IVs, we calculated the F-statistic of each SNP using the following formula: $F\text{-statistic} = R^2 \times (N-2)/(1-R^2)$, where R^2 is the variance of the phenotype explained by each genetic variant in exposure, and N is the sample size. R^2 was calculated using the following formula: $R^2 = 2 \times (\text{Beta})^2 \times \text{EAF} \times (1-\text{EAF})/[2 \times (\text{Beta})^2 \times \text{EAF} \times (1-\text{EAF}) + 2 \times (\text{SE})^2 \times N \times \text{EAF} \times (1-\text{EAF})]$, where Beta is the per allele effect size of the association between each SNP and phenotype, EAF is the effect allele frequency, and SE is the standard error of Beta (24). IVs with an F-statistic <10 were regarded as weak instruments (25).

Two-Sample Mendelian Randomization

The Wald ratio, which estimates causal effect for each IV, was calculated as the ratio of Beta for the corresponding SNP in the outcome data set divided by Beta for the same SNP in the exposure data set (26). Our main approach was to conduct a meta-analysis of each Wald ratio by inverse variance weighted (IVW) method using multiplicative random-effects model to estimate overall causal effect of the exposure on the outcome. The causal effects were calculated as the odds ratio (OR) for the risk of COVID-19 severity per one standard deviation (SD) increase in years of schooling (one SD is equivalent to 4.2 years) (15, 27). In addition, we conducted sensitivity analyses by MR-Egger regression, weighted median method, MR-PRESSO (Mendelian Randomization Pleiotropy RESidual Sum and Outlier) global test, and leave-one-out sensitivity analysis. The MR-Egger regression method is used to assess horizontal pleiotropy of IVs. When IV assumption 2 is violated, horizontal pleiotropy occurs, and MR-Egger regression intercept significantly differs from zero (28, 29). The weighted median method provides a valid causal estimate when more than half of the instrumental SNPs satisfy the IV assumptions (24). The MR-PRESSO global test investigates whether there are outlier SNPs whose variant-specific causal estimates differ substantially from those of other SNPs (30, 31). Leave-one-out sensitivity analysis was conducted to assess the reliability of the IVW method by removing each SNP from the analysis and reestimating the causal effect (31). Moreover, among SNPs associated with EA, we searched for SNPs associated with $P < 5.0 \times 10^{-8}$ with pleiotropic effects on body mass index (BMI), smoking, and other SES using the web tool PhenoScanner (version 2) (32, 33). The heterogeneity was also measured between the causal estimates across all SNPs in the IVW method calculating Cochran's Q statistic and I^2 statistic (34). Low heterogeneity provides more reliability for a causal effect (35). We conducted all the two-sample MR analyses using "TwoSampleMR" package (version 0.5.5) in R (version 4.0.3) (36). A P-value below 0.05 was considered statistically significant in all statistical analyses.

TABLE 1 | MR results of the causal effect of EA on the risk of COVID-19 severity.

IVW method	Weighted median method	MR-Egger regression method		Heterogeneity (IVW)	MR-PRESSO global test
OR (95% CI)	OR (95% CI)	OR (95% CI)	Intercept	Cochran's *Q*	
P-value	*P*-value	*P*-value	*P*-value	*P*-value	*P*-value
0.540 (0.376–0.777)	0.484 (0.283–0.826)	0.353 (0.084–1.483)	0.006	261.9	
$P = 0.0009$	$P = 0.008$	$P = 0.156$	$P = 0.548$	$P = 0.102$	$P = 0.115$

MR, Mendelian randomization; COVID-19, coronavirus disease 2019; EA, educational attainment; IVW, inverse variance weighted; OR, odds ratio; CI, confidence interval.

RESULTS

In total, 235 instrumental SNPs were identified for both EA and the risk of COVID-19 severity GWAS data sets. The characteristics of all the SNPs included in our analysis are shown in **Supplementary Table 2**. The F-statistic of every instrument was >29, thus suggesting that weak instrument bias was unlikely.

The IVW method showed that EA was significantly associated with a lower risk of COVID-19 severity [OR per 1-SD increase in years of schooling, 0.540; 95% confidence interval (CI), 0.376–0.777; $P = 0.0009$] in the European population (**Table 1**, **Figure 1**, and **Supplementary Figure 1**). Cochran's Q statistic and I^2 statistic for the IVW method were 261.9 ($P = 0.102$) and 0.110, indicating low heterogeneity and more reliability for the causal effect. Other MR methods also showed overall consistent protective effects for EA on the risk of COVID-19 severity although the MR-Egger regression estimate did not have statistical significance (**Table 1** and **Figure 1**). However, when

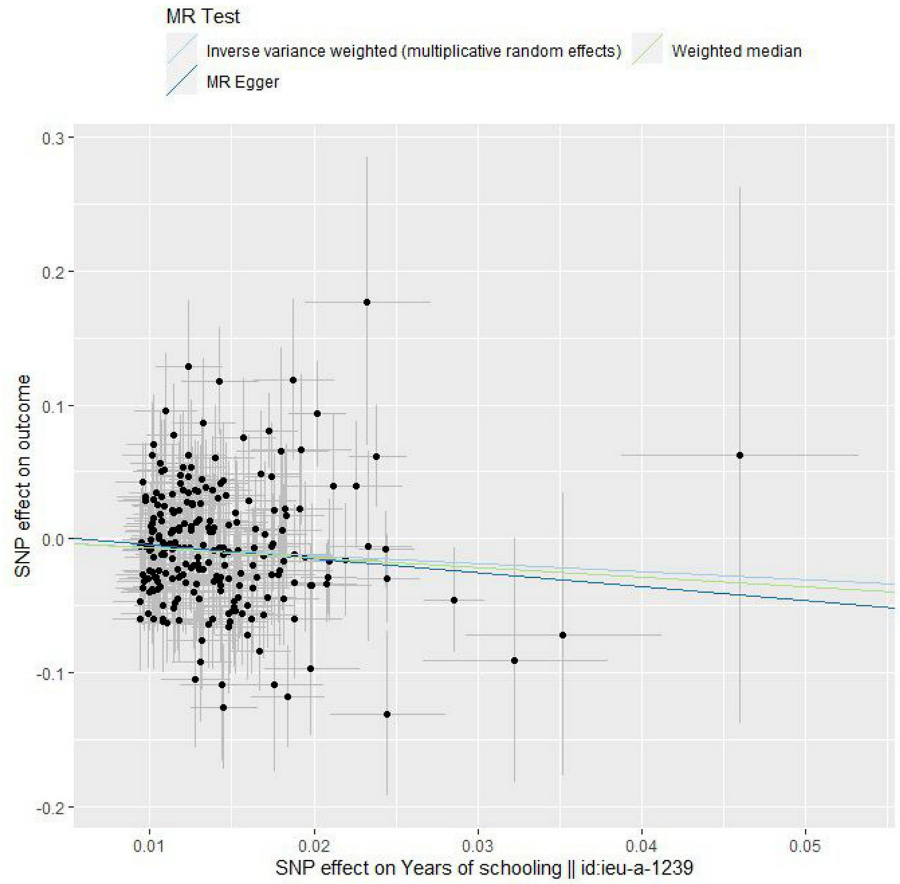


FIGURE 1 | Scatter plots. Each black point representing an SNP is plotted in relation to the effect size of the SNP on years of schooling (x-axis) and on the risk of COVID-19 severity (y-axis) with corresponding standard error bars. The slope of each line corresponds to the causal estimate using inverse variance weighted (light blue), weighted median (green), and MR-Egger regression (blue) method.

I^2 statistics are much <1, no measurement error assumption is violated and MR-Egger regression tends to underestimate the causal effect (34). In fact, Cochran's Q statistic and I^2 statistic for our MR-Egger regression were 261.5 ($P = 0.097$) and 0.113 (much <1), respectively. The MR-Egger intercept was 0.006 ($P = 0.548$), indicating little evidence of horizontal pleiotropy. The weighted median method indicated that more than half of the instrumental SNPs in our analysis satisfied the IV assumptions. The funnel plot showed general symmetry, suggesting little evidence of heterogeneity or horizontal pleiotropy (**Supplementary Figure 2**). MR-PRESSO global test ($P = 0.115$) and leave-one-out sensitivity analysis suggested the lack of an outlier SNP whose variant-specific causal estimate differed substantially from those of other SNPs (**Table 1** and **Supplementary Figure 3**). We excluded all palindromic SNPs regardless of minor allele frequencies from the IVW method, and then we obtained a comparable result to that of the original IVW method (OR, 0.540; 95% CI, 0.366–0.796, $P =$ 0.0019; number of instrumental SNPs; 209). Our search using PhenoScanner identified 17 SNPs that were associated with BMI (rs10073890, rs11123818, rs13090388, rs1334297, rs1566085, rs1618725, rs1689510, rs1964927, rs2725370, rs2820314, rs4787457, rs56391344, rs62444881, rs66568921, rs67890737, rs9372625, rs9384679), three SNPs that were associated with smoking traits (rs10240905 with pack years adult smoking as proportion of life span exposed to smoking, rs2179152 with pack years of smoking, and rs66568921 with ever smoked), and six SNPs that were associated with other SES traits (rs1008078 with Townsend deprivation index at recruitment, rs13090388, rs34316, and rs9372625 with job involving heavy manual or physical work, rs1391438 and rs2971970 with job involving mainly walking or standing), respectively. We excluded the 17 SNPs that were associated with BMI from the IVW method, and then we obtained a comparable result to that of the original IVW method (OR, 0.550; 95% CI, 0.373–0.810, $P = 0.0025$; number of instrumental SNPs; 218) (see section Discussion). When we excluded the three SNPs that were associated with smoking traits from the IVW method, we obtained a comparable result to that of the original IVW method (OR, 0.557; 95% CI, 0.386–0.803, $P = 0.0018$; number of instrumental SNPs; 232). Similarly, we excluded the six SNPs that were associated with other SES traits from the IVW method, and then we obtained a comparable result to that of the original IVW method (OR, 0.536; 95% CI, 0.372–0.773, $P = 0.0009$; number of instrumental SNPs; 229). Moreover, when we excluded all 23 SNPs that were associated with BMI, smoking traits, and other SES traits (three SNPs overlapped), we obtained a comparable result to that of the original IVW method (OR, 0.565; 95% CI, 0.382–0.835, $P = 0.0042$; number of instrumental SNPs; 212).

We noticed some possible overlap between the exposure GWAS participants and the outcome GWAS participants as shown in **Supplementary Table 1**. This might have led to bias in the causal estimate of EA on the risk of COVID-19 severity, but the bias was unlikely to be substantial because the possible overlap was small as discussed below.

DISCUSSION

To our knowledge, this is the first MR study to investigate the association between EA and the risk of COVID-19 severity. Observational studies report that a lower level of education influences the severity of COVID-19 among various populations (9–12). In the European population, those who had no qualification (equivalent to seven years of education) (37) had a higher risk of severe COVID-19 (i.e., a positive test for SARS-CoV-2 in a hospital setting either at emergency departments or as inpatients) than those who had college or university degree (equivalent to 20 years of education) (37) in fully adjusted model [risk ratio (RR), 1.58; 95% CI, 1.25–1.99; $p < 0.001$] in a prospective cohort study using UK Biobank data (9). Our two-sample MR approach supported, with little evidence of bias, the causal effect of higher EA on the risk of COVID-19 severity (OR, 0.540; 95% CI, 0.376–0.777; $P = 0.0009$) in the European population, which was consistent with the cohort study. In other populations, a risk-adjusted model of a large cohort, including 62,298 COVID-19 deaths, showed that lower education levels were strongly associated with the level of COVID-19 fatalities per 100,000 persons (rate ratio, 1.08; 95% CI, 1.05–1.11; $P < 0.0001$) in severely distressed counties in the United States (10). Another study in the United States showed that education level with a bachelor's degree was associated with a lower rate of mortality due to COVID-19 (estimate, −0.246; 95% CI, −0.388 to −0.103; $P =$ 0.0008) across various ethnicities in the seven most affected states (11). In São Paulo, Brazil, among patients under 60 years of age and living in areas with the lowest percentage (below 8.61%) of the population with a university degree, COVID-19 mortality was four times higher than that among those living in areas with the highest percentage (over 34.80%) of population with a university degree (rate ratio, 4.02; 95% CI, 3.42–4.72) (12). However, our MR analysis was based on populations of European ancestry, and the findings are unlikely to be generalized to other populations and ethnicities.

In our MR analysis, underlying genetic or epidemiological mechanisms of how EA lowered the risk of COVID-19 severity remain unknown. Therefore, although a range of sensitivity analyses indicated the robustness of our MR findings, we must pay careful attention to the possibility of unmeasured horizontal pleiotropy of genetic IVs for EA. Observational studies showed that higher EA was associated with decreased prevalence of smoking, physical inactivity, obesity, hypertension, and hypercholesterolemia (38). We infer that other risk factors, including BMI and lifetime smoking, were related to the causal effect of EA on the risk of COVID-19 severity in our analysis for the following reasons. First, MR studies have shown that EA has causal effects on decrease of BMI (39, 40). Second, MR studies have shown that BMI has a causal effect on the risk of COVID-19 severity (29, 41, 42). Consistent with the MR results showing the effect of BMI on the risk of COVID-19 severity, the risk-adjusted model showed that, in addition to the two socioeconomic factors of low level of education and a proportionally larger Black population, obesity was the only physical risk factor in the U.S. cohort (10). Other observational studies also have reported that BMI is a risk factor for hospital admission, disease severity, and in-hospital mortality due to COVID-19 (5–8). Consistently, our search using PhenoScanner identified 17 SNPs that were associated with both EA and BMI with $P < 5.0 \times 10^{-8}$. Similarly, MR studies have shown that EA has causal effects on increased lifetime smoking (39) and that lifetime smoking has a causal effect on the risk of COVID-19 severity (41) although the associations between smoking and the risk of COVID-19 remain controversial in observational studies (41). Therefore, the causal effect of EA on the risk of COVID-19 severity may be at least partly mediated through increases of BMI and lifetime smoking. Even if that is the case, EA would remain an intervention target for COVID-19 severity (43). In fact, when we excluded the 17 SNPs that were associated with BMI and the three SNPs that were associated with smoking traits, we obtained comparable results to the result of the original IVW method as described above. This supports the idea that EA remains an intervention target for COVID-19 severity because EA lowered the risk of COVID-19 severity to some extent independently of the effects of BMI and smoking.

Epidemiologically, the protective effect of EA on the risk of COVID-19 severity may be related to the social benefit of education. Observational studies showed that lower EA as well as other SES was associated with disparities in medical care (44). For example, counties in the United States with a higher percentage of people below the poverty level had a significantly lower percentage of the population with higher education as well as a lower percentage of people insured (11), and counties in the United States with higher income and education, a lower rate of disability, and a higher rate of the insured population were at a lower risk of COVID-19 mortality (11). However, we must pay attention to interpret the causal association between lower EA and the risk of COVID-19 severity because it remains unclear epidemiologically whether less educated people are more likely to develop severe COVID-19 symptoms. In other words, there is possibility that less educated people are more likely to be socioeconomically disadvantaged and to have an increased risk of SARS-CoV-2 transmission due to poor housing, overcrowding, and low-paid essential jobs that make social distancing more challenging (45). As a result of higher COVID-19 incidence, they may have a higher risk of COVID-19 severity. Ascertainment bias could also arise due to differential healthcare seeking, differential testing, and differential prognosis (9). We could not conduct an MR analysis investigating a causal effect of EA on the risk of COVID-19 incidence as described below. However, the prospective cohort study using UK Biobank data showed that both lower education and area-level socioeconomic deprivation by the Townsend index were associated with having a positive test including asymptomatic COVID-19 [RR 1.46 for no qualifications vs. degree (95% CI 1.19–1.79), and RR 1.39 for most deprived quartile vs. least (95% CI 1.12–1.71)] as well as a higher risk of testing positive in hospital (i.e., severe COVID-19 cases) [RR 1.58 for no qualifications vs. degree (95% CI 1.25–1.99), and RR 1.54 for most deprived quartile vs. least (95% CI 1.21–1.97)] in the fully adjusted model (9). The authors discussed that there remained the possibility that some socioeconomic groups had a poorer prognosis and were, therefore, more likely to be admitted to hospital and, therefore, to be tested (9).

The present study includes the following strengths. First, the samples used were gathered across populations with the same European ancestries, reducing substantial bias in our study. Among different genetic ancestries, effect sizes and allele frequencies can differ and lead to substantial bias (24). Second, we used the publicly available GWAS data sets with the largest sample sizes hitherto for both the exposure and outcome data sets. *F*-statistics were also large enough for weak instrument bias to be unlikely. Third, a range of sensitivity analyses relaxed the IV assumptions and supported the robustness of our MR findings.

However, we must pay attention to several major limitations. First, in the Geisinger Health System study, the participants in the exposure GWAS may have overlapped with the participants in the outcome GWAS as shown in **Supplementary Table 1**. This might have led to bias in the causal estimate of EA on the risk of COVID-19 severity (46). It was difficult for us to exclude the Geisinger Health System study because we used summary-level data for the exposure and outcome data sets. However, the participants in the Geisinger Health System study represented only 1.9% (14,562 out of 766,344) of those in the exposure GWAS data set. Moreover, the participants in the Geisinger Health System_EUR study represented only 1.2% (53 out of 4,392) of the severe COVID-19 cases, and most of them (10.7%, 112,862 out of 1,054,664) were controls in the outcome GWAS data set. If the data sets are of different sizes, the percentage overlap should be taken with respect to the larger data set (46). Therefore, vast majority of the participant overlap in the outcome GWAS data set occurred, if at all, among the controls. In that situation, the bias is unlikely to be substantial, and unbiased causal estimates are expectedly obtained in two-sample MR studies (41, 46). On the other hand, we could not conduct an MR analysis investigating a causal effect of EA on the risk of COVID-19 incidence because the summary-level GWAS data of COVID-19 incidence (i.e., 32,494 SARS-CoV-2 infection cases and 1,316,207 controls in the European population) by the COVID-19 Host Genetics Initiative (19, 20) had possible participant overlap [at most, 16.2% (5,270 out of 32,494) of the SARS-CoV-2 infection cases in the deCODE_EUR, the Geisinger Health System_EUR, and the Netherlands Twin Register_EUR studies] with the EA GWAS data set that could cause substantial bias (46) (**Supplementary Table 1**). Second, our MR findings might be affected by unmeasured horizontal pleiotropy as described above. As is the often the case with many MR studies, strictly satisfying all the IV assumptions can be challenging (47). Third, our MR analysis was based on populations of European ancestry, and the findings are unlikely to be generalized to other populations and ethnicities. Fourth, we could not conclude that the risk of COVID-19 severity could decrease simply by increasing years of schooling because the underlying genetic or epidemiological mechanisms remain unknown.

In conclusion, we have shown for the first time using a two-sample MR approach the associations between higher EA and the lower risk of COVID-19 severity in the European population that observational studies have reported. However, genetic or epidemiological mechanisms underlying the association between EA and the risk of COVID-19 severity remain unknown, and further studies are warranted to validate our MR findings and investigate underlying mechanisms.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://gwas.mrcieu.ac.uk/datasets/ieu-a-1239/, https://www.covid19hg.org/results/r5/.

ETHICS STATEMENT

Ethical approval was not provided for this study on human participants because we used only publicly available GWAS summary datasets. Written informed consent was not provided because we used only publicly available GWAS summary datasets.

AUTHOR CONTRIBUTIONS

MY designed this study, analyzed data, and wrote the draft of the manuscript. MY and KA discussed and reviewed the manuscript critically. Both authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We would like to thank the Social Science Genetic Association Consortium, the MRC IEU Open GWAS database, MR-Base, and the COVID-19 Host Genetics Initiative for making GWAS datasets publicly available. Our manuscript has been proofread by Editage English Editing Service.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.673451/full#supplementary-material

Supplementary Figure 1 | Forrest plots. Each black point represents the causal estimate of each SNP on the risk of COVID-19 severity per increase in years of schooling, and red points show the combined causal estimates using IVW and MR-Egger regression methods with horizontal lines denoting 95% confidence intervals.

Supplementary Figure 2 | Funnel plots. Each black point representing an SNP is plotted in relation to the estimate of years of schooling on the risk of COVID-19 severity (x-axis) and the inverse of standard error (y-axis). Vertical lines show the combined causal estimates using IVW (light blue) and MR-Egger regression (blue) methods.

Supplementary Figure 3 | Leave-one-out sensitivity analysis. Each black point represents the combined causal estimates on the risk of COVID-19 severity per increase in years of schooling using IVW methods with horizontal lines denoting 95% confidence intervals after removing the corresponding SNP from the analysis.

Supplementary Table 1 | Contributing studies of the exposure GWAS data and the outcome GWAS data.

Supplementary Table 2 | The characteristics of all the SNPs included in our analysis.

REFERENCES

1. *WHO Coronavirus Disease (COVID-19) Dashboard*. Available online at: https://covid19.who.int/ (accessed February 26, 2021).
2. Yuki K, Fujiogi M, Koutsogiannaki S. COVID-19 pathophysiology: a review. *Clin Immunol*. (2020) 215:108427. doi: 10.1016/j.clim.2020.108427
3. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using Open SAFELY. *Nature*. (2020) 584:430–6. doi: 10.1038/s41586-020-2521-4
4. Koh J, Shah SU, Chua PEY, Gui H, Pang J. Epidemiological and clinical characteristics of cases during the early phase of COVID-19 pandemic: a systematic review and meta-analysis. *Front Med (Lausanne)*. (2020) 7:295. doi: 10.3389/fmed.2020.00295
5. Palaiodimos L, Kokkinidis DG, Li W, Karamanis D, Ognibene J, Arora S, et al. Severe obesity, increasing age and male sex are independently associated with worse in-hospital outcomes, and higher in-hospital mortality, in a cohort of patients with COVID-19 in the Bronx, New York. *Metabolism*. (2020) 108:154262. doi: 10.1016/j.metabol.2020.154262
6. Popkin BM, Du S, Green WD, Beck MA, Algaith T, Herbst CH, et al. Individuals with obesity and COVID-19: a global perspective on the epidemiology and biological relationships. *Obes Rev*. (2020) 21:e13128. doi: 10.1111/obr.13128
7. Kalligeros M, Shehadeh F, Mylona EK, Benitez G, Beckwith CG, Chan PA, et al. Association of obesity with disease severity among patients with coronavirus disease 2019. *Obesity (Silver Spring)*. (2020) 28:1200–4. doi: 10.1002/oby.22859
8. Simonnet A, Chetboun M, Poissy J, Raverdy V, Noulette J, Duhamel A, et al. High prevalence of obesity in severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) requiring invasive mechanical ventilation. *Obesity (Silver Spring)*. (2020) 28:1195–9. doi: 10.1002/oby.22831
9. Niedzwiedz CL, O'Donnell CA, Jani BD, Demou E, Ho FK, Celis-Morales C, et al. Ethnic and socioeconomic differences in SARS-CoV-2 infection: prospective cohort study using UK Biobank. *BMC Med*. (2020) 18:160. doi: 10.1186/s12916-020-01640-8
10. Hawkins RB, Charles EJ, Mehaffey JH. Socio-economic status and COVID-19-related cases and fatalities. *Public Health*. (2020) 189:129–34. doi: 10.1016/j.puhe.2020.09.016
11. Abedi V, Olulana O, Avula V, Chaudhary D, Khan A, Shahjouei S, et al. Racial, Economic, and Health Inequality and COVID-19 Infection in the United States. *J Racial Ethn Health Disparities*. (2020) 8:732–42. doi: 10.1101/2020.04.26.20079756
12. Ribeiro KB, Ribeiro AF, de Sousa Mascena Veras MA, de Castro MC. Social inequalities and COVID-19 mortality in the city of São Paulo, Brazil. *Int J Epidemiol*. (2021) dyab022. doi: 10.1093/ije/dyab022. [Epub ahead of print].
13. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet*. (2014) 23:R89–98. doi: 10.1093/hmg/ddu328
14. Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, Rietveld CA, et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature*. (2016) 533:539–42. doi: 10.1038/nature17671
15. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nat Genet*. (2018) 50:1112–21. doi: 10.1038/s41588-018-0147-3
16. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. *Elife*. (2018) 7:e34408. doi: 10.7554/eLife.34408
17. *MRC IEU Open GWAS Project*. Available online at: https://gwas.mrcieu.ac.uk/ (accessed February 12, 2021).
18. *MR-Base*. Available online at: http://app.mrbase.org/ (accessed February 12, 2021).
19. COVID-19 Host Genetics Initiative. The COVID-19 host genetics initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. *Eur J Hum Genet*. (2020) 28:715–8. doi: 10.1038/s41431-020-0636-6
20. *COVID19-hg GWAS Meta-Analyses Round 5*. Available online at: https://www.covid19hg.org/results/r5/ (accessed February 3, 2021).

21. Liu HM, Hu Q, Zhang Q, Su GY, Xiao HM, Li BY, et al. Causal effects of genetically predicted cardiovascular risk factors on chronic kidney disease: a two-sample Mendelian randomization study. *Front Genet*. (2019) 10:415. doi: 10.3389/fgene.2019.00415
22. Gill D, Karhunen V, Malik R, Dichgans M, Sofat N. Cardiometabolic traits mediating the effect of education on osteoarthritis risk: a Mendelian randomization study. *Osteoarthritis Cartilage*. (2021) 29:365–71. doi: 10.1016/j.joca.2020.12.015
23. Xiuyun W, Qian WX, Weidong L, Lizhen L. Education and stroke: evidence from epidemiology and Mendelian ra Minjun ndomization study. *Sci Rep*. (2020) 10:21208. doi: 10.1038/s41598-020-78248-8
24. Gill D, Efstathiadou A, Cawood K, Tzoulaki I, Dehghan A. Education protects against coronary heart disease and stroke independently of cognitive function: evidence from Mendelian randomization. *Int J Epidemiol*. (2019) 48:1468–77. doi: 10.1093/ije/dyz200
25. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res*. (2017) 26:2333–55. doi: 10.1177/0962280215597579
26. Adams CD, Boutwell BB. Can increasing years of schooling reduce type 2 diabetes (T2D)? Evidence from a Mendelian randomization of T2D and 10 of its risk factors. *Sci Rep*. (2020) 10:12908. doi: 10.1038/s41598-020-69114-8
27. Dardani C, Howe LJ, Mukhopadhyay N, Stergiakouli E, Wren Y, Humphries K, et al. Cleft lip/palate and educational attainment: cause, consequence or correlation? A Mendelian randomization study. *Int J Epidemiol*. (2020) 49:1282–93. doi: 10.1093/ije/dyaa047
28. Walker VM, Davies NM, Hemani G, Zheng J, Haycock PC, Gaunt TR, et al. Using the MR-Base platform to investigate risk factors and drug targets for thousands of phenotypes. *Wellcome Open Res*. (2019) 4:113. doi: 10.12688/wellcomeopenres.15334.1
29. Aung N, Khanji MY, Munroe PB, Petersen SE. Causal inference for genetic obesity, cardiometabolic profile and COVID-19 susceptibility: a Mendelian randomization study. *Front Genet*. (2020) 11:586308. doi: 10.3389/fgene.2020.586308
30. Verbanck M, Chen CY, Neale B, Do R. Detecion of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet*. (2018) 50:693–8. doi: 10.1038/s41588-018-0099-7
31. Burgess S, Davey Smith G, Davies NM, Dudbridge F, Gill D, Glymour MM, et al. Guidelines for performing Mendelian randomization investigations. *Wellcome Open Res*. (2020) 4:186. doi: 10.12688/wellcomeopenres.15555.2
32. *PhenoScanner V2, A Database of Human Genotype-Phenotype Associations*. Available online at: http://www.phenoscanner.medschl.cam.ac.uk/ (accessed March 25, 2021).
33. Wood A, Guggenheim JA. Refractive error has minimal influence on the risk of age-related macular degeneration: a Mendelian randomization study. *Am J Ophthalmol*. (2019) 206:87–93. doi: 10.1016/j.ajo.2019.03.018
34. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int J Epidemiol*. (2016) 45:1961–74. doi: 10.1093/ije/dyw220
35. Greco MFD, Minelli C, Sheehan NA, Thompson JR. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Stat Med*. (2015) 34:2926–40. doi: 10.1002/sim.6522
36. R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
37. Hagenaars SP, Gale CR, Deary IJ, Harris SE. Cognitive ability and physical health: a Mendelian randomization study. *Sci Rep*. (2017) 7:2651. doi: 10.1038/s41598-017-02837-3
38. Hoeymans N, Smit HA, Verkleij H, Kromhout D. Cardiovascular risk factors in relation to educational level in 36 000 men and women in The Netherlands. *Eur Heart J*. (1996) 17:518–25. doi: 10.1093/oxfordjournals.eurheartj.a014903
39. Carter AR, Gill D, Davies NM, Taylor AE, Tillmann T, Vaucher J, et al. Understanding the consequences of education inequality on cardiovascular disease: mendelian randomisation study. *BMJ*. (2019) 365:l1855. doi: 10.1136/bmj.l1855
40. Cao M, Cui B. Association of educational attainment with adiposity, type 2 diabetes, and coronary artery diseases: a Mendelian randomization study. *Front Public Health*. (2020) 8:112. doi: 10.3389/fpubh.2020.00112
41. Li S, Hua X. Modifiable lifestyle factors and severe COVID-19 risk: a Mendelian randomisation study. *BMC Med Genomics*. (2021) 14:38. doi: 10.1186/s12920-021-00887-1
42. Ponsford MJ, Gkatzionis A, Walker VM, Grant AJ, Wootton RE, Moore LSP, et al. Cardiometabolic traits, sepsis, and severe COVID-19: a Mendelian randomization investigation. *Circulation*. (2020) 142:1791–3. doi: 10.1161/CIRCULATIONAHA.120.050753
43. Anderson EL, Howe LD, Wade KH, Ben-Shlomo Y, Hill WD, Deary IJ, et al. Education, intelligence and Alzheimer's disease: evidence from a multivariable two-sample Mendelian randomization study. *Int J Epidemiol*. (2020) 49:1163–72. doi: 10.1093/ije/dyz280
44. Woolf SH, Braveman P. Where health disparities begin: the role of social and economic determinants–and why current policies may make matters worse. *Health Aff (Millwood)*. (2011) 30:1852–9. doi: 10.1377/hlthaff.2011.0685
45. Khunti K, Singh AK, Pareek M, Hanif W. Is ethnicity linked to incidence or outcomes of covid-19? *BMJ*. (2020) 369:m1548. doi: 10.1136/bmj.m1548
46. Burgess S, Davies NM, Thompson SG. Bias due to participant overlap in two-sample Mendelian randomization. *Genet Epidemiol*. (2016) 40:597–608. doi: 10.1002/gepi.21998
47. Burgess S, Butterworth AS, Thompson JR. Beyond Mendelian randomization: how to interpret evidence of shared genetic predictors. *Clin Epidemiol*. (2016) 69:208–16. doi: 10.1016/j.jclinepi.2015.08.001

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

DATA REPORT
published: 29 June 2021
doi: 10.3389/fpubh.2021.677012

Are the Temporal Trends of Stomach Cancer Mortality in Brazil Similar to the Low, Middle, and High-Income Countries?

Samantha Hasegawa Farias, Wilson Leite Maia Neto, Katia Pereira Tomaz, Francisco Winter dos Santos Figueiredo and Fernando Adami*

Epidemiology and Data Analysis Laboratory, Centro Universitário Saúde ABC, Santo André, Brazil

Keywords: epidemiology, stomach neoplasms, socioeconomic status, mortality, social inequalities

OPEN ACCESS

Edited by:
MinJae Lee,
University of Texas Southwestern
Medical Center, United States

Reviewed by:
Reza Malekzadeh,
Tehran University of Medical
Sciences, Iran
Masoud Amiri,
Erasmus Medical Center, Netherlands

****Correspondence:***
Samantha Hasegawa Farias
samanthahf@gmail.com

Specialty section:
This article was submitted to
Life-Course Epidemiology and Social
Inequalities in Health,
a section of the journal
Frontiers in Public Health

Received: *07 March 2021*
Accepted: *14 May 2021*
Published: *29 June 2021*

Citation:
Farias SH, Maia Neto WL, Tomaz KP, Figueiredo FWdS and Adami F (2021) Are the Temporal Trends of Stomach Cancer Mortality in Brazil Similar to the Low, Middle, and High-Income Countries? Front. Public Health 9:677012. doi: 10.3389/fpubh.2021.677012

INTRODUCTION

Stomach cancer is the fourth most common malignant tumor in the world, and although numbers have fallen in recent years, mortality from this cause is still high (1–3). In Brazil, some studies have shown a reduction in mortality from stomach cancer since the 1980s (4, 5), which can be attributed to improved eating habits, food preservation, and treatment of Helicobacter pylori infection (6, 7).

In addition, there were significant advances related to socioeconomic development and the reduction of inequalities and socioeconomic inequities, which improved the population's access to health care and reduced the morbidity and mortality of diseases such as breast cancer (8–10).

Brazil is a middle-income country characterized by great internal heterogeneity (11, 12). It is notorious that poverty in Brazil has a location (13) and, in terms of disparities, the country has a very striking feature that is the regional inequalities, where the north and northeast regions have the worst indicators. The central region has intermediate rates, and the south and southeast are the best conditions, regardless of the socioeconomic indicator being evaluated (14). These social inequalities in the country still today directly reflect on health inequality, explaining the unfavorable scenarios for the north and northeast, and a very evident polarization in relation to the south and southeast (15).

The country presents regions with different socioeconomic characteristics, which impacts health services, lifestyle, and socio-cultural aspects. In other words, there are developed regions with high technology for cancer-oriented health services and underdeveloped regions that cannot properly treat and diagnose its citizens (16).

Thus, considering that Brazil is a country with territorial extension of continental characteristics and high socioeconomic plurality, and that the mortality due to stomach cancer is related to the socioeconomic status of the site, what level of development does the behavior of stomach cancer mortality in Brazil follow?

Thus, the aim of this study was to describe the temporal trend of stomach cancer mortality in Brazil from 1990 to 2016, analyzing its behavior in relation to low, middle, and high income countries.

METHODS

Study Design

Secondary data analysis performed based on data from 1990 to 2016 obtained from the Global Burden of Disease (GBD).

Data Source

The Global Burden of Disease database is coordinated by the Institute of Health Metrics and Evaluation (IHME) of the University of Washington and maintained through a partnership with researchers from 124 countries, with the objective of estimating the global burden of more than 300 diseases and injuries (17).

This database provides information from various sources based on official documents such as censuses, administrative databases, scientific publications, hospital and police records, among others. Through this information, there is a joint effort by scientific commissions from various countries to systematically quantify the magnitude of health loss due to diseases, injuries and risk factors by age, sex, and geographic location.

To facilitate the production of estimates and comparability of data, GBD researchers created a measure to classify the socio-demographic development of a locality, the Socio-demographic Index (SDI) (18), based on the average income per person, schooling, and total fertility rate to classify countries as low, medium low, medium, medium high, and high income.

Study Variables

The studied variables were deaths, age-standardized mortality, and proportional mortality from all cancer causes and proportional to all deaths. Data for Brazil and low, middle, and high income countries were adjusted for age and were expressed as rates (per 100,000 inhabitants). In the present study, only the low, middle, and high income classifications were evaluated, in order to better capture the differences between the analysis groups.

Statistical Analysis

Descriptive statistics were performed using the statistical program Stata® (StataCorp, L,C) version 11.0 and presented through absolute and relative frequency.

The time trend analysis was performed through the program Joinpoint Regression version 4.6.0 (Statistical Research and Applications Branch, National Cancer Institute, Rockville, EUA) (19). The joinpoint regression is a technique that explores the relationship between two variables by means of segmented linear regression. It determines the magnitude of change in the trend in percentage terms and verifies whether or not this change is statistically significant (20).

The final model chosen was the one with the highest number of points and maintained the statistical significance ($p < 0.05$). From the estimated slope for each straight line (regression coefficient), the Annual Percentage Change (APC) and Average Annual Percentage Change (AAPC) were calculated and its statistical significance was estimated by the Least Squares Method by a generalized linear model and for each straight line segment, with an estimated slope, and their 95% confidence intervals.

Ethical Aspects

According to Resolution No. 510 of April 7, 2016 of the National Health Council of Brazil, since these are public data and of free access. There is no need for ethical appreciation.

TABLE 1 | Mean mortality rates and age-adjusted mortality rates due to stomach cancer, proportional mortality for all deaths and proportional mortality for all cancers, 1990–2016.

Place	Mortality rate[a]	Age-standardized Mortality rate[a]	Proportional mortality (%)	
			All deaths	All cancers
Brazil	13,10	15,54	1,60	9,74
Socioeconomic status				
Low income	11,06	11,12	0,38	8,01
Middle income	17,16	24,70	2,31	12,37
High income	10,78	12,04	1,94	7,44

[a] *Per 100,000 inhabitants.*

RESULTS

There were 14,139,731 deaths from stomach cancer in the high, middle, and low income countries between 1990 and 2016, of which 612,818 were in low-income countries, 9,137,851 in middle-income countries and 4,389,062 in high-income countries. In Brazil, there were 449,682 deaths in the same period.

With regard to socioeconomic status, stomach cancer mainly affects middle-income countries. In these countries, ∼25 people die from stomach cancer per 100,000 inhabitants, representing 2.3% of all deaths from known causes and 12.3% of deaths from some form of cancer. In Brazil, the burden of stomach cancer appears to be lower than that observed in middle-income countries (15.5 deaths per 100,000 inhabitants, mortality proportional to all deaths of 1.6%, and all cancers of 9.7%) (**Table 1**).

It was observed that, regardless of the socioeconomic status, there is a decrease in the mortality rates due to stomach cancer in the studied sites. Throughout the study time, the rates decreased more in high income countries, while the middle income countries had greater variability (**Figure 1**).

In the first period of change corresponding to the years between 1990 and 2003, Brazil presented the annual percentage change (APC) of −1.8 (95% CI −1.9; −1.7), behavior of low and middle-income countries, which presented the same changes in their respective first periods of change. The second period of change observed in Brazil corresponded to the years of 2003–2015 and had APC of −2.8 (95% CI −3.0; −2.7), behavior close to high-income countries, which presented APC of −2.5 (95% CI −2.6; −2.4) (**Table 2**).

When analyzing the average of the annual percentage change, we observed that the low-income countries had the lowest fall with the AAPC of −1.4(95% CI −1.5; −1.3), followed by middle-income −2.1(95% CI −2.1; −2.0) and high income countries −2.7(95% CI −2.8; −2.6). Brazil presented AAPC of −2.3(95% CI −2.4; −2.2).

DISCUSSION

Between 1990 and 2016, there was a downward trend in age-adjusted mortality from stomach cancer in all socioeconomic

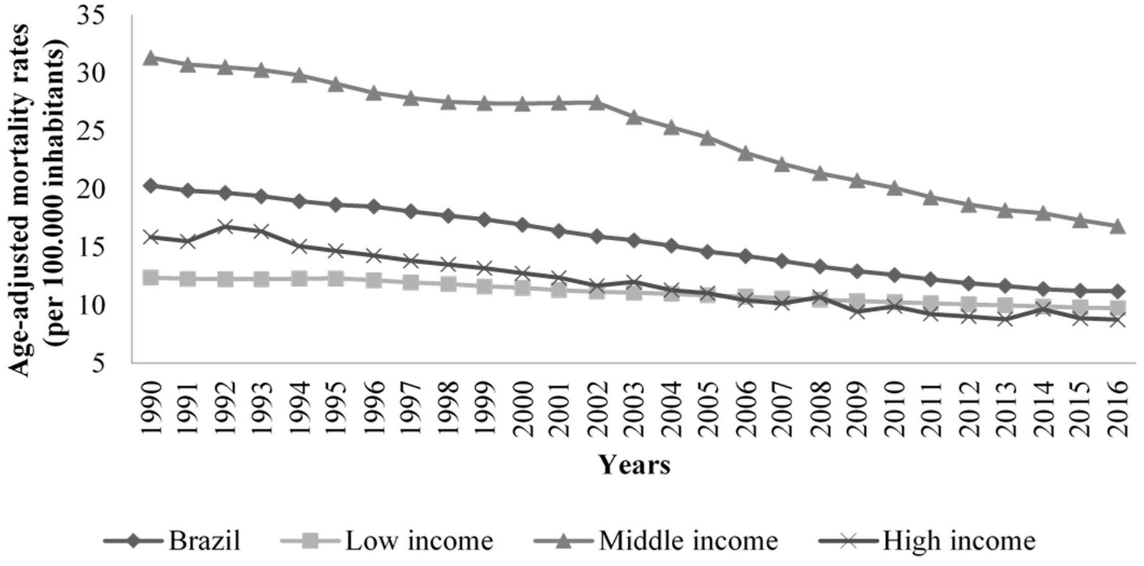


FIGURE 1 | Trends of age adjusted mortality rates related to stomach cancer (per 100,000 inhabitants) in Brazil and low, middle, high income countries, 1990 to 2016.

TABLE 2 | Estimates of temporal trend of specific mortality rates for stomach cancer according to cut-off points obtained through the joinpoint. 1990–2016.

	Period	AAPC (95%CI)	APC(95%CI)	*p*-value
Brazil	1990–2003	−2.3 (−2.4; −2.2)	−1.8 (−1.9; −1.7)	<0.001
	2003–2016		−2.8 (−3.0; −2.7)	<0.001
Low income	1990–2004	−1.4 (−1.5; −1.3)	−1.8 (−1.8; −1.7)	<0.001
	2004–2013		−1.2 (−1.3; −1.0)	<0.001
	2013–2016		−0.4 (−1.1; -0.4)[a]	0.3
Middle income	1990–1997	−2.1 (−2.1; −2.0)	−1.8 (−2.0; −1.7)	<0.001
	1997–2004		−0.9 (−1.1; −0.8)	<0.001
	2004–2007		−5.4 (−6.3; −4.6)	<0.001
	2007–2010		−1.8 (−2.7; −1.0)	<0.001
	2010–2013		−3.3 (−4.2; −2.5)	<0.001
	2013–2016		−0.7 (−1.2; −0.3)	<0.001
High income	1990–1995	−2.7 (−2.8; −2.6)	−2.6 (−3.0; −2.2)	<0.001
	1995–2006		−3.0 (−3.1; −2.9)	<0.001
	2006–2016		−2.5 (−2.6; −2.4)	<0.001

CI, confidence interval; APC, Annual Percent Change; AAPC, Average Annual Percent Change. [a] The APC is not statistically significant (p > 0.05).

statuses studied (low, middle, and high income) and in Brazil, which showed a similar trend to that observed in middle-income countries.

The decrease in mortality in all socioeconomic statuses studied can be explained by the improvement in the population living conditions. Even in poorer countries, there has been improvement in social and economic aspects in recent decades (21).

Despite the improvements, epidemiological studies have found relationships between low socioeconomic status in childhood and the development of stomach cancer in adult life. One of the possibilities would be an early infection by H. pylori bacteria (22, 23). In view of this, it is to be understood that changes in mortality rates in low- and middle-income countries still tend to bear the consequences of this socioeconomic condition over a given time, even if they have already been overcome.

Over time, Brazil presented similar variations to all high-income countries, and in some periods of the series studied, variations were found in both low-income and middle- and high-income countries.

However, the mortality rates presented in Brazil are similar to the rates of middle-income countries and higher than those of some high-income countries (5, 24). This is because despite the high incidence in countries such as Japan, China, and South Korea, the diagnosis of stomach cancer occurs early, which reduces mortality (25).

On the other hand, some factors may explain the higher mortality in Brazil. Cancers of infectious origin, such as the stomach, are common in Latin countries due to economic development, and the Brazilian health system has no guidelines for screening. One of the main aspects that is directly involved with cases of stomach cancer deaths in Brazil is the inequality related to economic, geographic, and socio-cultural issues (5, 26, 27).

Despite the drop in stomach cancer mortality in Brazil, the cases are still high and projections show an increase in the less developed regions of the country (4, 5). This fact underscores the importance of studies that take into account the geographical distribution, especially in countries such as Brazil, characterized by large socioeconomic discrepancies between regions.

It is important to emphasize that the territorial extension of Brazil also has an impact on the difficulty of professional qualification, access to health services and treatment funds, important factors for early detection, clinical management, and patient survival (26, 28).

Another important issue to consider in the current scenario of stomach cancer in the country and that is directly related to territorial extension was the lack of standardization in the diagnosis, staging, and treatment in the study period (26), key factors in achieving good treatment results (29). Only in 2018 did Brazil approve diagnostic and therapeutic guidelines for stomach adenocarcinoma, which is the most common type of gastric cancer, accounting for about 90% of diagnosed cases (30).

Brazil presents a process of demographic and epidemiological transition that occurs differently depending on its Federative Units due to its socioeconomic disparities (31).

The North and Northeast regions present characteristics of low and middle income countries, such as high mortality rates due to infectious diseases (32), worse sanitation conditions (33) and a larger proportion of population residing in rural areas (34).

In contrast, the Midwest, South, and Southeast regions have characteristics of high income countries, such as the increase of chronic diseases such as obesity (35), the increase in life expectancy and, therefore, a more aged population (36).

This scenario shows that Brazil encompasses several factors that may influence the burden of stomach cancer. It is important to identify what local socioeconomic characteristics are related to the disease, which is a crucial starting point for the change of scenery in the country.

The limitations of this study are related to the use of secondary data, in which the researcher does not have control of data quality. However, despite being a constraint, we believe that because it is a database produced by important institutions and the database is used in scientific articles published in high impact journals, the findings support the reliability and validity of this data.

CONCLUSION

Over time, Brazil shows a constant decline, with periods of variation similar to the behavior observed in both high and low income countries. Additionally, the findings of this study point to the need to understand the behavior of stomach cancer mortality in the regions and federal states of Brazil, since they present different socioeconomic characteristics.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: http://ghdx.healthdata.org/gbd-results-tool.

AUTHOR CONTRIBUTIONS

SF conceived the study, analyzed the data, constructed the results from the data, and wrote the manuscript. WM analyzed the data, constructed the results from the data, and wrote the manuscript. KT collected the data and constructed the results from the data. FF conceived the study, analyzed the data, and reviewed results. FA conceived the study and reviewed results. All authors read and approved the final manuscript.

FUNDING

This work was supported by National Council for Scientific and Technological Development (CNPq)-FA was CNPq research productivity scholarship (Process Number 307183/2018-1).

ACKNOWLEDGMENTS

Federal University of South and Southeast Pará (UNIFESSPA)-for the institutional support to SF.

REFERENCES

1. Kamangar F, Dores GM, Anderson WF. Patterns of cancer incidence, mortality, and prevalence across five continents: defining priorities to reduce cancer disparities in different geographic regions of the world. *J Clin Oncol.* (2006) 24:2137–50. doi: 10.1200/JCO.2005.05.2308
2. Fielding JW, Ellis DJ, Jones BG, Paterson J, Powell DJ, Waterhouse JA, et al. Natural history of" early" gastric cancer: results of a 10-year regional survey. *Br Med J.* (1980) 281:965–67. doi: 10.1136/bmj.281.6246.965
3. Ang TL, Fock KM. Clinical epidemiology of gastric cancer. *Singapore Med J.* (2014) 55:621. doi: 10.11622/smedj.2014174
4. Chatenoud L, Bertuccio P, Bosetti C, Levi F, Curado MP, Malvezzi M, et al. Trends in cancer mortality in Brazil, 1980–2004. *Europ J Cancer Prevent.* (2010) 19:79–86. doi: 10.1097/CEJ.0b013e32833233be
5. de Souza Giusti ACB, de Oliveira Salvador PTC, Dos Santos J, Meira KC, Camacho AR, Guimarães RM, et al. Trends and predictions for gastric cancer mortality in Brazil. *World J Gastroenterol.* (2016) 22:6527. doi: 10.3748/wjg.v22.i28.6527
6. Koifman S, Koifman RJ. Stomach cancer incidence in Brazil: an ecologic study with selected risk factors. *Cad de saude Publica.* (1997) 13:S85–92. doi: 10.1590/S0102-311X1997000500009
7. Muñoz N. Aspects of gastric cancer epidemiology with special reference to Latin America and Brazil. *Cad de saude Publica.* (1997) 13:S109. doi: 10.1590/S0102-311X1997000500013
8. dos Santos Figueiredo FW, Adami F. Income inequality and mortality owing to breast cancer: evidence from Brazil. *Clin Breast Cancer.* (2018) 18:e651–8. doi: 10.1016/j.clbc.2017.11.005
9. Atun R, De Andrade LOM, Almeida G, Cotlear D, Dmytraczenko T, Frenz P, et al. Health-system reform and universal health coverage in Latin America. *Lancet.* (2015) 385:1230–47. doi: 10.1016/S0140-6736(14)61646-9
10. Marmot M. Brazil: rapid progress and the challenge of inequality. *Int J Equity Health.* (2016) 15:1–2. doi: 10.1186/s12939-016-0465-y
11. Silva ICMD, Restrepo-Mendez MC, Costa JC, Ewerling F, Hellwig F, Ferreira L, et al. Measurement of social inequalities in health: concepts and methodological approaches in the Brazilian context. *Epidemiol Serv Saúde.* (2018) 27:e000100017. doi: 10.5123/S1679-49742018000100017
12. Santos AMAD, Jacinto PDA, Tejada CAO. Causalidade entre renda e saúde: uma análise através da abordagem de dados em painel com os estados do Brasil. *Estudos Econ.* (2012) 42:229–61. doi: 10.1590/S0101-41612012000200001
13. Beghin N. Notes on inequality and poverty in Brazil: current situation and challenges. In: Green D, editor. *From Poverty to Power: How Active Citizens and Effective States Can Change the World.* Oxford: Oxfam International (2008). p. 1–6.
14. Silva SAD. Regional inequalities in Brazil: divergent readings on their origin and public policy design. *EchoGéo.* (2017) 41:315–40. doi: 10.4000/echogeo.15060

15. Bucciferro JR, de Souza PHF. The evolution of regional income inequality in Brazil, 1872–2015. In: Deng K, editor. *Time and Space.* Cham: Palgrave Macmillan (2020). p. 131–56. doi: 10.1007/978-3-030-47553-6_6
16. dos Santos Figueiredo FW, do Carmo Almeida TC, Cardial DT, da Silva Maciel É, Fonseca FLA, Adami F. The role of health policy in the burden of breast cancer in Brazil. *BMC Women's Health*. (2017) 17:121. doi: 10.1186/s12905-017-0477-9
17. Institute for Health Metrics and Evaluation (IHME). *A Protocol for the Global Burden of Diseases, Injuries, and Risk Factors Study* (*GBD*). Seattle, WA: IHME, University of Washington (2020).
18. Global Burden of Disease Collaborative Network. *Global Burden of Disease Study 2015 (GBD 2015) Socio-Demographic Index (SDI) 1980–2015.* Seattle, WA: Institute for Health Metrics and Evaluation (IHME) (2016).
19. Joinpoint Regression Program, Version 4.8.0.1. *Statistical Methodology and Applications Branch, Surveillance Research Program.* Bethesda, MD: National Cancer Institute (2020).
20. Kim HJ, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. *Statist Med*. (2000) 19:335–51. doi: 10.1002/(sici)1097-0258(20000215)19:3<335::aid-sim336>3.0.co;2-z
21. Sinding SW. Population, poverty and economic development. *Philo Trans R Soc B Biol Sci*. (2009) 364:3023–30. doi: 10.1098/rstb.2009.0145
22. Zali H, Rezaei-Tavirani M, Azodi M. Gastric cancer: prevention, risk factors and treatment. *Gastroenterol Hepatol Bed Bench*. (2011) 4:175–85. doi: 10.22037/ghfbb.v4i4.193
23. Mihmanli M, Ilhan E, Idiz UO, Alemdar A, Demir U. Recent developments and innovations in gastric cancer. *World J Gastroenterol*. (2016) 22:4307. doi: 10.3748/wjg.v22.i17.4307
24. Sierra MS, Cueva P, Bravo LE, Forman D. Stomach cancer burden in Central and South America. *Cancer Epidemiol*. (2016) 44:S62–73. doi: 10.1016/j.canep.2016.03.008
25. Carcas LP. Gastric cancer review. *J Carcinog*. (2014) 13:14. doi: 10.4103/1477-3163.146506
26. Zilberstein B, Malheiros C, Lourenço LG, Kassab P, Jacob CE, Weston AC, et al. Consenso brasileiro sobre câncer gástrico: diretrizes para o câncer gástrico no Brasil. ABCD. *Arq Bras Cir Dig*. (2013) 26:2–6. doi: 10.1590/S0102-67202013000100002
27. De Martel C, Ferlay J, Franceschi S, Vignat J, Bray F, Forman D, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol*. (2012) 13:607–15. doi: 10.1016/S1470-2045(12)70137-7
28. Nagini S. Carcinoma of the stomach: a review of epidemiology, pathogenesis, molecular genetics and chemoprevention. *World J Gastrointestinal Oncol*. (2012) 4:156. doi: 10.4251/wjgo.v4.i7.156
29. Muraro CLPM. Câncer gástrico precoce: contribuição ao diagnóstico e resultado do tratamento cirúrgico. *Rev Col Bras Cir*. (2003) 30:352–8. doi: 10.1590/S0100-69912003000500005
30. Brasil. *Diretrizes Diagnósticas e Terapôuticas do Adenocarcinoma de Estômago.* Brasília: Ministério da Saúde (2018).
31. Marinho F, de Azeredo Passos VM, Malta DC, França EB, Abreu DM, Araújo VE, et al. Burden of disease in Brazil, 1990–2016: a systematic subnational analysis for the Global Burden of Disease Study 2016. *Lancet*. (2018) 392:760–75. doi: 10.1016/S0140-6736(18)31221-2
32. Kerr-Pontes LR, Barreto ML, Evangelista CM, Rodrigues LC, Heukelbach J, Feldmeier H. Socioeconomic, environmental, and behavioural risk factors for leprosy in North-east Brazil: results of a case–control study. *Int J Epidemiol*. (2006) 35:994–1000. doi: 10.1093/ije/dyl072
33. Mantovani SA, Delfino BM, Martins AC, Oliart-Guzmán H, Pereira TM, Branco FL, et al. Socioeconomic inequities and hepatitis A virus infection in Western Brazilian Amazonian children: spatial distribution and associated factors. *BMC Infect Dis*. (2015) 15:428. doi: 10.1186/s12879-015-1164-9
34. Maia AG, Buainain AM. O novo mapa da população rural brasileira. *Confins*. (2015) 25:1–17. doi: 10.4000/confins.10548
35. Santos LMP, De Oliveira IV, Peters LR, Conde WL. Trends in morbid obesity and in bariatric surgeries covered by the Brazilian public health system. *Obes Surg*. (2010) 20:943–8. doi: 10.1007/s11695-008-9570-3
36. Paes NA. Elderly mortality in Brazil: trends, differentials, causes and links with socioeconomic indicators. In: *24th IUSSP General Population Conference.* Salvador, BA (2001).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

ORIGINAL RESEARCH
published: 05 July 2021
doi: 10.3389/fpubh.2021.690645

Multi-State Analysis of the Impact of Childhood Starvation on the Healthy Life Expectancy of the Elderly in China

Huiling Dong [1], *Chunjing Du* [2], *Bingyi Wu* [2]* *and Qunhong Wu* [3]*

[1] *Department of Public Health, Weifang Medical University, Weifang, China,* [2] *Department of Management, Weifang Medical University, Weifang, China,* [3] *Department of Health Management, Harbin Medical University, Harbin, China*

Background: Child malnutrition is not only common in developing countries but also an important issue faced by developed countries. This study aimed to explore the influence and degree of childhood starvation on the health of the elderly, which provides a reference for formulating health-related policies under the concept of full lifecycle health.

Methods: Based on the Chinese Longitudinal Healthy Longevity Survey (CLHLS) in 2008, 2011, and 2014, this study took a total of 13,185 elderly people aged 65–99 years as the target population. By IMaCH software, with gender and income level as the control variables, the average life expectancy and healthy life expectancy of the elderly were measured. The x^2test was used to explore the differences in the socioeconomic status of elderly people with or without starvation in childhood. Statistical differences between average life expectancy and healthy life expectancy were analyzed by rank tests.

Results: (1) The results showed that there was a statistically significant difference in age, gender, residency, education level, and income level between the groups with or without starvation ($P < 0.05$). (2) Transition probabilities in health–disability, health–death, and disability–death all showed an upward trend with age ($P < 0.05$), where the elderly who experienced starvation in childhood were higher than those without such an experience ($P < 0.05$). However, the probability of disability–health recovery showed a downward trend with age ($P < 0.05$), in which the elderly who experienced starvation in childhood were lower than those without starvation ($P < 0.05$). (3) For the elderly who experienced starvation in childhood, the health indicators of the average life expectancy, healthy life expectancy, and healthy life expectancy proportion accounted for the remaining life were lower than those of the elderly without childhood starvation ($P < 0.05$).

Conclusions: The average life expectancy and healthy life expectancy of the elderly with childhood starvation are lower than those without childhood starvation. It shows that the negative impact of childhood starvation on health through the life course till old age has a persistent negative cumulative effect on the quantity and quality of life. Therefore, it is important to pay attention to the nutritional status of children in poor families from the perspective of social policymaking.

Keywords: the elderly, healthy life expectancy, risk transition probability, childhood starvation, life course

OPEN ACCESS

Edited by:
MinJae Lee,
University of Texas Southwestern
Medical Center, United States

Reviewed by:
Christiana Sandu,
Alexandru Ioan Cuza
University, Romania
Débora Martins Dos Santos,
Rio de Janeiro State University, Brazil

****Correspondence:***
Bingyi Wu
wuby369@163.com
Qunhong Wu
wuqunhong@163.com

Specialty section:
This article was submitted to
Life-Course Epidemiology and Social
Inequalities in Health,
a section of the journal
Frontiers in Public Health

Received: *03 April 2021*
Accepted: *08 June 2021*
Published: *05 July 2021*

Citation:
Dong H, Du C, Wu B and Wu Q (2021)
Multi-State Analysis of the Impact of
Childhood Starvation on the Healthy
Life Expectancy of the Elderly in
China. Front. Public Health 9:690645.
doi: 10.3389/fpubh.2021.690645

BACKGROUND

Since the 1900's, China has entered a violently turbulent modern society. During this period, the lives of people were generally difficult, and hunger caused a large number of deaths. After the establishment of the People's Republic of China in 1949, the country experienced a "three-year difficult period" from 1959 to 1961, which caused a nationwide "famine." The number of deaths caused by starvation had risen, and the lack of nutrition greatly affected the health of the population. In recent years, some studies analyzed the impact of the "Great Famine" in childhood on the health and economic status of their adulthood and found that infants and children who had grown up in that period had shorter life longevity and poorer health in old age (1). The factors show that the health status of the elderly is not only affected by elements of the old stage, but also earlier life experiences, especially childhood.

Life course theory has gradually become an important paradigm for the study of elderly health, emphasizing the long-term impact of life events in critical periods on the health outcomes of the elderly. Existing research shows that childhood is a critical period of growth and development (2). At this time, negative nutrition, and health shocks experienced may change the original development track of the individual and thus affect the health trajectory of their entire life course. The elderly who experienced starvation during childhood as a health disadvantaged group have current disadvantages that may depend on the previous unfavorable socioeconomic status. From a policy perspective, it is necessary to understand the long-term effects of malnutrition in earlier life, because child malnutrition is not only common in developing countries, but also an important issue faced by developed countries. Data show that there are approximately 1 billion people in the world who are malnourished, including 140 million preschool children under the age of 5, which will lead to permanent damage to their physical and cognitive development and even death due to nutritional diseases (3).

In recent years, economics has begun to pay attention to the long-term effects of life experiences in the fetus or early childhood on health, education, and labor market conditions, especially the health and nutritional status before the age of 5 (4). These studies generally used negative external shocks as an identification event, such as war, famine, rainfall, flu, etc., (5–7). Schellenberg J A, Victora C G, and other studies of low-income or middle-income countries such as Brazil, India, and South Africa have found that malnourished children had shorter height in adulthood, fewer years of education, and diminished labor productivity (8). Epidemiology and health economics in China are increasingly concerned about the impact of childhood health and developmental status on their health in the adult period. Scholars such as Chen and Zhou (9) analyzed the long-term effects of China's "Great Famine" on the health of those who experienced famine. Studies found that babies and children born or raised during this period had a lower height, poorer health, and economic status in adulthood.

Although, there are many literatures on widely acknowledged links between childhood hunger and a range of adverse health outcomes late in life, the reliability of research results and research conclusions is different because of different measurement indicators. A comprehensive evaluation of the long-term impacts of hunger on an individual's health capital is empirically difficult to conduct. Earlier investigation of this issue was hindered by several challenges including data restrictions. To determine the long-term consequences, we need both information about whether a person experienced hunger several decades ago and information about health status. Data tracking individual experiences for such a long period are not often available even in developed countries (10, 11). The most frequent method in the previous literature is to use exposure to shocks defined at a more aggregated level (12), taking famine as an indicator of having childhood starvation (13–15). But the problem is that exposure to famine and exposure to hunger are not equivalent. Famine and hunger belong to different levels of variables. Therefore, identification strategies that only exploit macro-level variations may obtain inconsistent estimates of the long-term effects of hunger.

We deal with this problem by exploiting retrospective data on the individual-level occurrence of hunger episodes during childhood, collected by the Chinese Longitudinal Healthy Longevity Survey (CLHLS). This kind of measurement method is more effective, which can conduct a micro-analysis of how the dilemma in the early life is transformed into the negative results in later life. There is a growing literature taking advantage of this self-reported measure to examine the long-term consequences of childhood hunger associated with World War II and several famines that happened in European countries (16–18).

In summary, some researchers have done a lot of exploration in this field, but there are still many limitations. First, the above-mentioned studies have paid more attention to the lasting impact of severe nutrition and health shocks on the health of economically active people (people who have not yet reached old age). Second, in most of the previous research studies, self-assessed health status of the elderly was used as the dependent variable, by which body function of the elderly cannot be reflected effectively due to stronger subjectivity. As for the measurement of healthy life expectancy, most international literature adopts the multi-state life table method based on cohort data, which can reflect the true health level of the study population, and the research conclusions are more reliable (19, 20). Third, due to the lack of high-quality cohort data in China, most previous studies were based on cross-sectional data to measure the relationship between child hunger and health and cannot make statistical inferences. Therefore, this study takes life course theory as the analysis framework based on strict cohort data, whether to have the childhood starvation as independent variables, to measure healthy life expectancy of the elderly in China, trying to give an answer to the question: to what extent does the accumulated disadvantage formed by childhood starvation affect the health of the elderly?

Abbreviations: LE, average life expectancy; HLE, healthy life expectancy; HLE/LE, healthy life expectancy accounted for the remaining life.

METHODS

Data

The data were derived from the CLHLS. This research project has conducted surveys eight times in 23 provinces in China. The project used the 2008 baseline data and the 2011 and 2014 follow-up data with a total of 13,185 people aged 65–99 years as samples. Since the survey did not select samples with equal probability, it was necessary to weigh the samples according to the actual composition of age, gender, and residence in the 23 provinces to represent the general population of the elderly in the country. Therefore, before calculating the relevant indicators, the study weighed the data according to the weight coefficients provided by the project team, after which the sample size was 16,200 people (based on the sixth census data of China) in order to better reflect the overall Chinese elderly population.

Variables

The explained variable in this article is the health status of the elderly. The CLHLS database uses the ADL to determine the health status of the elderly. The scale includes six measurement items for eating, dressing, indoor activities, going to the toilet, bathing, and controlling toilet. For the above six measurement items, if the research subjects select "can be completed," they are judged as "healthy"; if any one of the items is selected as "unable to complete," the sample is judged as "disabled" (21). The question about income in the original questionnaire was, "Compared with the local people, what is your life?" There are five options. The options are set to "very rich," "relatively rich," "average," "more difficult," and "very difficult." This study combined "very wealthy" and "relatively wealthy" into "rich" as a high-income group, and combined "average," "relatively difficult," and "very difficult" into "difficult" as a low-income group.

The core explanatory variable of this article is "Did you experience starvation in childhood," to which the answer is "yes or no." Childhood starvation in this study generally refers to physical hunger. Specifically, in childhood, due to food shortages, insufficient intake of energy and essential nutrients leads to changes in body structure and function. In the questionnaire of CLHLS, the exact age at which starvation occurred is not asked. Therefore, this article draws on the definition of children's physiological age in the Medicine, Education, and Labor Legislation field and defines childhood age from 0 to 18 years (22, 23).

Preparation and Calculation of Multi-State Life Table

IMaCh software is used for the estimation of healthy life expectancy in this study, which is an abbreviation of Interpolated Markov Chain and one of the first batches of software to provide multi-state life table estimation. Its main advantage is the direct use of the original survey data, and the use of multi-period (≥2 times) different longitudinal data at intervals, in which processing different health statuses are considered such as improvement, reduction, no change, and death. In this study, State 1 and State 2 are healthy and disabled, respectively, and State 3 is dead.

In the multi-state life table, the initial state of the cohort (2008) is "healthy" and "unhealthy"; the end state (2014) is "healthy," "unhealthy," and "death." Each state at the beginning of the period can be transited to any state at the end of the period. In this study, "whether childhood is starving" is defined as a binary variable, and "gender" and "income level" are included as control variables to calculate the multi-state healthy life expectancy of the elderly with or without starvation in childhood. The calculation formula of the main variables is as follows [Lievre et al. (24)]:

- Let X(x) denote the state of an individual aged x. After time h, this individual is in state X(x+h). Assume that X(x) is a non-homogeneous discrete parameter Markov chain on these three states with transition probabilities:

$$ {}_hp_x^{jk} = \Pr(X(x+h) = k/X(x) = j) \tag{1} $$

- If this individual is observed only once more at timet_3, and noted to be in state l, then a further contribution to the likelihood is $({}_{d2}p_{x2}^{kl})$. In this case, the component of the total likelihood due to individual i is:

$$ L^{(i)} = ({}_{d1}p_{x1}^{jk}) \times ({}_{d2}p_{x2}^{kl}). \tag{2} $$

One observes that the formation of the likelihood is no trivial matter since there is no simple analytical expression for the higher-order transition probabilities.

- If θ denotes the vector of parameters and $\hat{\theta}$ its maximum-likelihood estimator, then standard theory tells that for a large sample of size N, the MLE$\hat{\theta}$ is approximatively normally distributed with mean θ and covariance matrix V ($\hat{\theta}$):

$$ lim_{N\to\infty} E(\hat{\theta}) \ = \ \theta \tag{3} $$

$$ V(\hat{\theta}) = \frac{1}{N} I^{-1}(\theta) \tag{4} $$

Where, I(θ) is the information matrix computed at the true value θ. This implies the asymptotic normality of the estimates of the transition probabilities and health expectancies.

- The initial state was i. The proportion of outcome status was 1 (health) and 2 (disability). The prevalence ${}_tW^{i1}(x)$ among survivors at age x and in state 1 from a cohort of individuals in state i at age x - t (t years earlier) reads:

$$ {}_tw^{i1}\ (x) \ = \ \frac{{}_tp_{x-t}^{i1}}{{}_tp_{x-t}^{i1} + {}_tp_{x-t}^{i2}} \tag{5} $$

and the prevalence ${}_tW^{i2}(x)$:

$$ {}_tw^{i2}\ (x) \ = \ \frac{{}_tp_{x-t}^{i2}}{{}_tp_{x-t}^{i1} + {}_tp_{x-t}^{i2}} \tag{6} $$

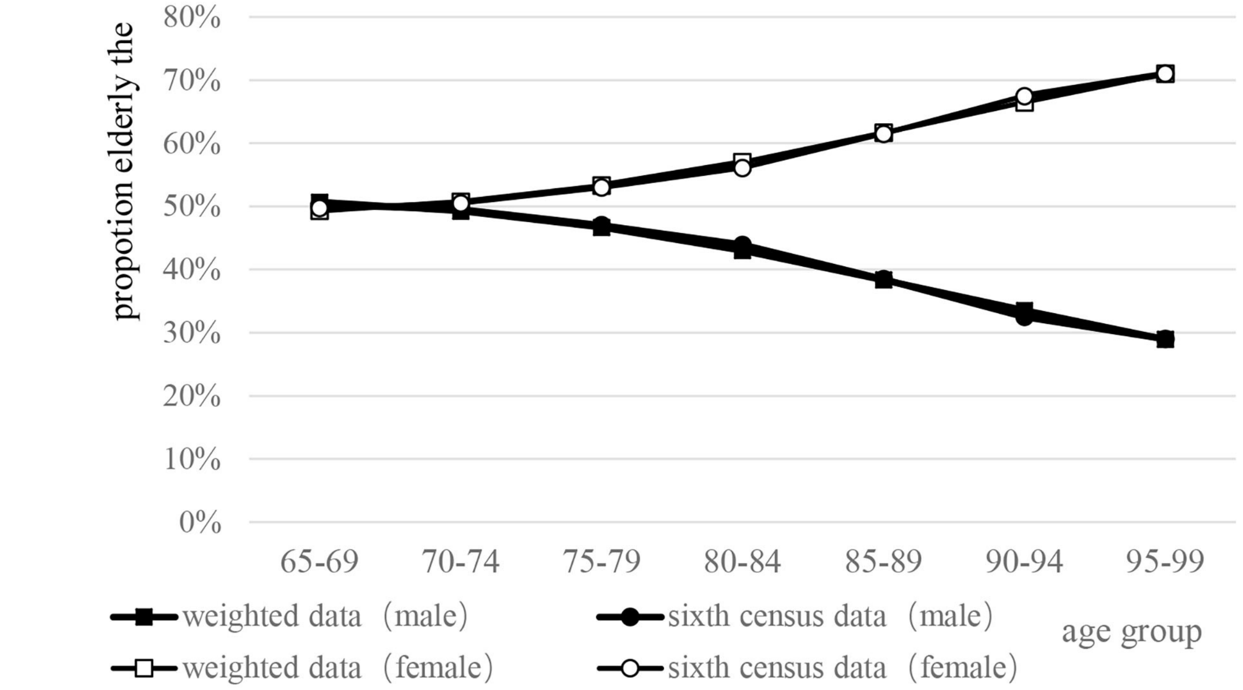


FIGURE 1 | Comparison between the sixth census data of China and weighted survey data by sex–age structure.

- Calculate the incidence rate of individual ending status j (the stable disability rate in the state j):

$$_{y}p_{x}^{.j}(\theta) = {}_{y}p_{x}^{1j}(\theta) + w^{2}(x,\theta)\left({}_{y}p_{x}^{2j}(\theta) - {}_{y}p_{x}^{1j}(\theta)\right) \quad (7)$$

- Over the interval (x, x + y), given the initial state i at age x, with y as the upper limit in the sums:

$$_{y}e_{x}^{ij} = \sum_{u=1}^{y} {}_{u}p_{x}^{ij} \quad (8)$$

- Total life expectancies respective of the initial state are:

$$e_{x}^{i.} = e_{x}^{i1} + e_{x}^{i2} \quad (9)$$

Statistical Analysis

First, the survey data were weighted using the weight coefficients provided by the CLHLS project team, and the weighted data were compared to the sixth census data of China from the age and sex, making the study sample more representative. Second, SPSS17.0 software was used to describe the frequency of different health statuses of the elderly. It analyzed the distribution of health status by age, gender, residency, education level, income level, and whether hungry or not. Third, x^2 test was used to explore the differences in the socioeconomic status of elderly people with or without starvation in childhood, taking $\alpha = 0.05$ as the inspection standard. Finally, by IMaCh software, the multi-state life table method was used to measure the average life expectancy and healthy life expectancy of the elderly people who have childhood starvation or not. Fourth, statistical differences between average life expectancy and healthy life expectancy were analyzed by rank tests.

RESULTS

Data Quality Assessment

Taking the sex–age structure of the elderly over 65 years in the sixth census in China as a reference, the data after weight adjustment of the cohort in the CLHLS database from 2008 to 2014 were compared. The results showed that the weighted adjusted data fitted well with the sixth census data of China (**Figure 1**).

Descriptive Analysis

With the weighted adjustment of the raw data, the baseline number in 2008 was 16,200. The remaining number in 2011 was 14,405, and the number of survivors in 2014 was 12,876. In 2008, the proportion of elderly people under 80 years accounted for 83.51%. In 2011 and 2014, the proportion under 80 years old increased to 86.01 and 88.05%, respectively. The proportion of elderly females was higher than 50%, which was slightly higher than the elderly males. More than 60%

TABLE 1 | Basic situation and health transition of the elderly.

	2008 Year		2011 Year		2014 Year	
Explanatory variable	***N***	**Proportion (%)**	***N***	**Proportion (%)**	***N***	**Proportion (%)**
Age group						
65–69	5,590	34.51	5,351	37.16	5,097	39.59
70–74	4,696	28.99	4,266	29.62	3,870	30.06
75–79	3,242	20.01	2,770	19.23	2,369	18.40
80–84	1,755	10.83	1,401	9.73	1,109	8.61
85–89	696	4.30	495	3.44	352	2.73
90–94	184	1.14	106	0.74	66	0.51
95–99	37	0.23	16	0.11	13	0.10
Gender						
Male	7,754	47.86	6,780	47.08	5,976	46.41
Female	8,447	52.14	7,622	52.93	6899	53.58
Residency						
Urban	10,033	61.9	8,764	65.1	7,758	60.3
Rural	5,135	38.1	4,695	34.9	4,251	39.7
Education level						
Primary schools and below	13,234	81.69	11,657	80.95	10,349	80.37
Above primary school	2,943	18.17	2,721	18.89	2,507	19.47
Economic level						
Low income	14,115	87.1	12,507	86.85	11,136	86.49
High income	2,086	18.3	1,876	13.03	1,721	13.37
Starvation						
Yes	10,709	66.1	9,489	65.9	8,442	65.6
No	5,492	33.9	4,912	34.1	4,434	34.4
Health condition						
Health	15,405	95.09	10,331	63.77	8,024	49.53
Disability	796	4.91	1,393	8.60	4,627	28.56
Death	0	0	1,800	11.11	3325	20.52
Missing visits	0	0	2,677	16.52	225	1.39

of the elderly were farmers or unemployed; over 80% were primary and lower in education and lower income, which presented a declining trend over time. The proportion of elderly people with hunger in 2008 was 66.1%, rising to 65.9% in 2011 and 65.6% in 2014, respectively. From the perspective of health status, the proportion of healthy elderly decreased year by year during the follow-up period. Meanwhile, the proportion of disabled and dead elderly showed an upward trend (**Table 1**).

Single-Factor Analysis

The differences in the socioeconomic status of the two groups of elderly were explored. The results showed that there was a statistically significant difference in age, gender, residency, education level, and income level between the groups with or without starvation ($P < 0.05$). Specifically, the main features of the elderly with starvation experience were as follows: mainly over 80 years old, female (52.9%), rural (88.4%), lower education level with primary schools and below (91.9%), and mainly low income (87.8%). The detailed results are shown in the **Supplementary Table 2**.

Risk Transition Probability

Health–Disability and Disability–Health Transition Probability

In general, the health–disability transition probability showed a linear upward trend with age [male: 95% CI (0.1244, 0.2715), female: 95% CI (0.1245, 0.2893), $P < 0.05$], whether male or female, whereas, the difference between the two groups gradually increased with age. On the contrary, the disability–health transition probability was linearly decreasing with age for all the elderly [male: 95% CI (0.094, 0.1716), female: 95% CI (0.0921, 0.1746), $P < 0.05$].

For the elderly males, the disability–health transition probability of the elderly who experienced starvation in childhood was lower than that of those without childhood starvation ($t = 0.440$, $P < 0.05$), especially for those over 80 years old. The difference between the two groups gradually widened with age. However, as for health–disability transition probability, the elderly who experienced starvation in childhood were higher than those without starvation ($t = 0.526$, $P > 0.05$). Elderly females ($t = 3.279$, $P < 0.05$) are similar to men. Meanwhile, for the probability of health–disability transition,

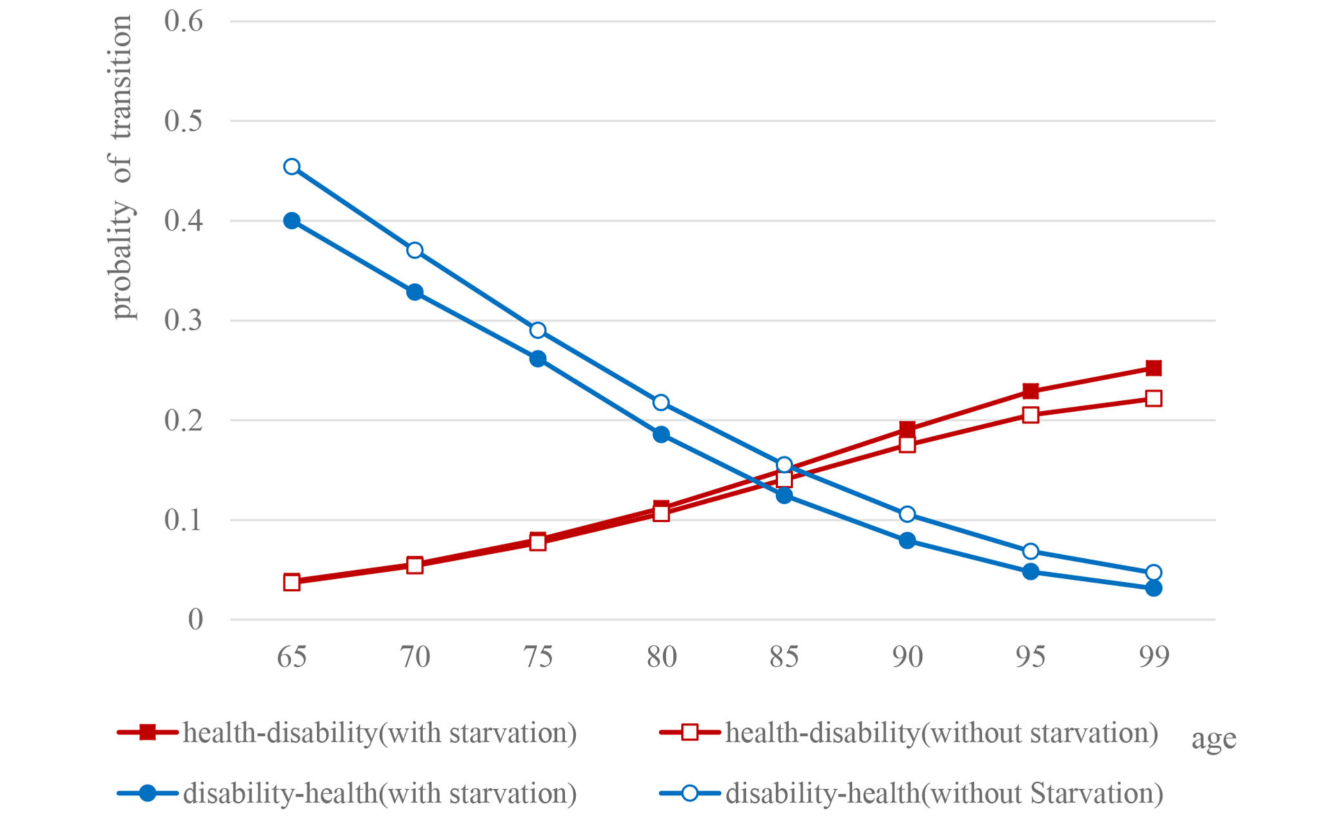


FIGURE 2 | Comparison of the probability curves of disability and health transition among elderly males.

the situation of the two groups was just the opposite ($t = 0.999$, $P > 0.05$) (**Figures 2**, **3**).

Health–Death and Disability–Death Transition Probability

Overall, not only health–death [male: 95% CI (0.0988, 0.2480), female: 95% CI (0.0973, 0.2534), $P < 0.05$] but also disability–death transition probability [male: 95% CI (0.2516, 0.4789), female: 95% CI (0.2526, 0.4717), $P < 0.05$] of the elderly with or without starvation in childhood have both shown a linear upward trend, whereas, the difference between the two groups gradually expanded with age.

Specifically, for elderly male people who experienced starvation in childhood, the probability of disability–death transition ($t = 8.140$, $P < 0.05$) and health–death transition probability ($t = 2.079$, $P > 0.05$) were both higher than for the elderly without experience of starvation. Similarly, as for the probability of both disability–death ($t = 8.135$, $P < 0.05$) and health–death ($t = 1.873$, $P > 0.05$), the elderly female who experienced starvation in childhood were higher than those without experience of starvation (**Figures 4**, **5**).

Analysis of Healthy Life Expectancy and Its Differences

Overall, regardless of male or female, the elderly who experienced starvation in childhood were lower than the elderly without starvation experience on such indicators as the average life expectancy, healthy life expectancy, and healthy life expectancy accounted for the remaining life, in which the difference between the two groups gradually decreasing with age on the average life expectancy [male: 95% CI (4.7241, 9.6559), female: 95% CI (5.8672, 11.6542), $P < 0.05$] and the healthy life expectancy [male: 95% CI (3.1915, 7.8079), female: 95% CI (3.7217,9.0226), $P < 0.05$], respectively.

For the elderly males, the HLE of the elderly between 65 and 69 years was 12.26 ± 0.26 years, while the LE was 14.36 ± 0.27 years, which meant that elderly males between 65 and 69 years were in a healthy state accounting for 85.30% of the time. In the same age group, the HLE of the elderly without hunger in childhood was 12.70 ± 0.21 years. Meanwhile, the LE was 14.78 ± 0.23 years, indicating that 65- to 69-year-old males without hunger had 85.92% in a healthy state for the rest of their lives. The paired t-test found that the HLE of the elderly without starvation in all age groups was higher than that of the elderly with starvation, with statistically significant difference ($P < 0.05$), while the HLE/LE of the elderly without starvation was also higher than that of elderly people with starvation, with statistically significant difference ($P < 0.05$).

For the elderly females, the HLE of those between 65 and 69 years with hunger experience was 14.06 ± 0.30 years, while the LE was 17.06 ± 0.33 years, which meant that the males between 65 and 59 years with hunger had 82.30% healthy state for the rest of their life. In the same age group, the HLE of the elderly without hunger in childhood was 14.48 ± 0.22 years. Meanwhile,

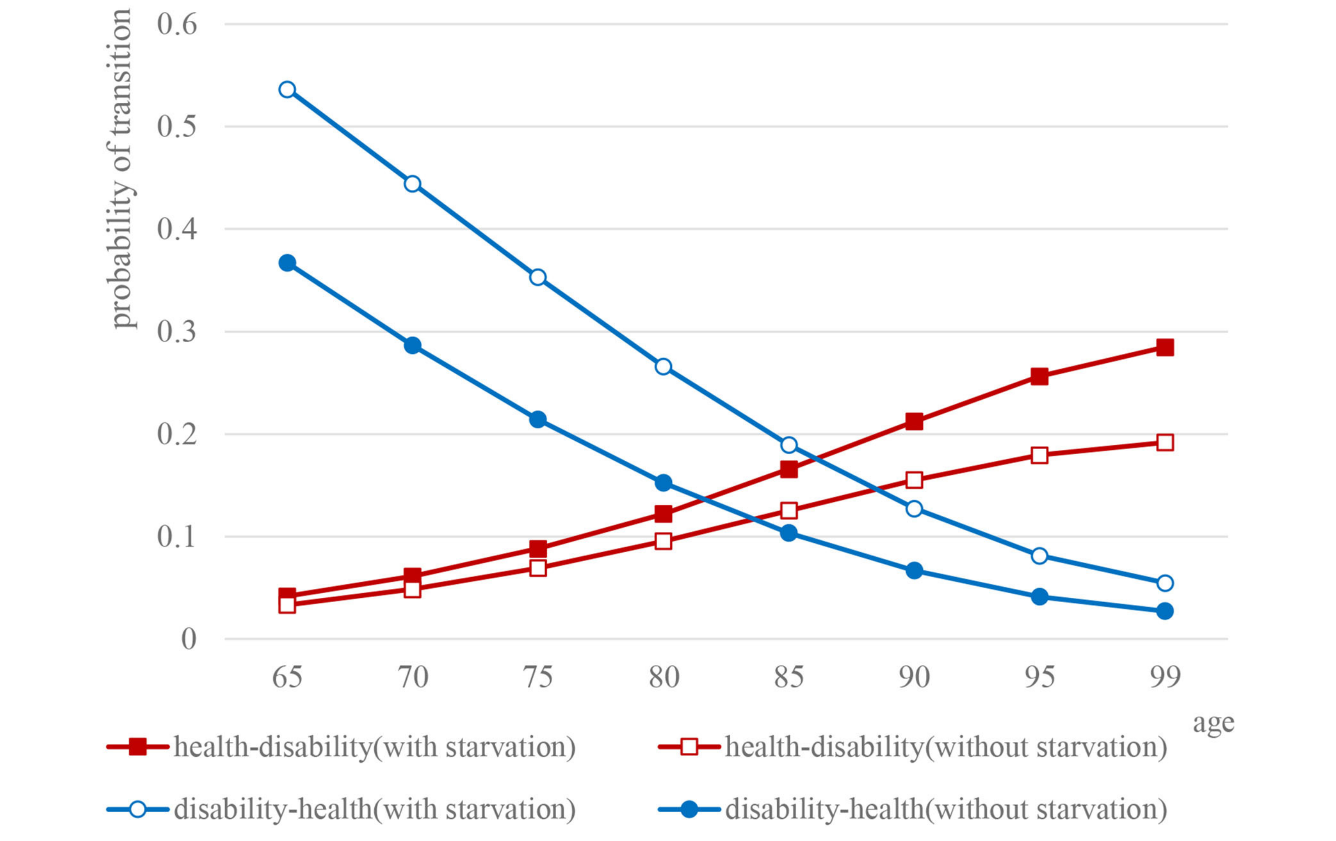


FIGURE 3 | Comparison of the probability curves of disability and health transition among elderly females.

the LE was 17.42 ± 0.25 years, indicating that the elderly males between 65 and 59 years without hunger had 83.1% in a healthy state of living (**Table 2**). The paired *t*-test found that the HLE of the elderly without starvation in all age groups was higher than that of the elderly with starvation, at which the difference was statistically significant ($P < 0.05$). The proportion (HLE/LE) of the elderly without starvation was also higher than that of elderly people who experienced starvation, at which the difference was statistically significant ($P < 0.05$).

DISCUSSION

The life course provides an important theoretical perspective for a comprehensive analysis of the health status of the elderly. The results of this study showed that the experience of starvation in childhood had a negative cumulative effect on the health in old age, which was related to the social and historical environment of the research group. The target group was born in 1908–1942, which was a special historical period of social transformation, political turmoil, and material deprivation. During the time, many elderly people had experienced starvation before the age of 12. Some scholars have studied the long-term negative effects of "great famine of China" on the health of famine-experienced people (25–27). However, as a rare historical event, the "Great Famine" has serious, extreme, and transient characteristics, which conclusions drawn have certain limitations in terms of external validity. In contrast, the adverse effects of childhood starvation on health in this study are more typical and more universal.

The experience of starving in childhood affects the socioeconomic status of adulthood, which in turn affects the health outcomes of the elderly. The results of this study showed that the elderly who have experienced starvation in childhood were in rural (88.4%), mostly primary school and below in education (91.9%), and lower income level (87.8%). The literature that examined the long-term effects of fetal or childhood health as independent variables found that chronically poor health or malnutrition in childhood had a significant negative impact on the years of education during adulthood (28). Qing He and Yuan Yan analyzed the data of CHNS to show that the overall health status during childhood had a significant positive effect on adult income (29). Specifically, people with low socioeconomic status usually have cumulative disadvantages in terms of work environment, access to medical services, and health risks, which can affect their availability of health resources and health protection capacity (30).

The multi-state transition probability is the basis for measuring healthy life expectancy. When calculating the healthy life expectancy, the transitions between different multiple health states and the death risk could be taken to consideration, in which

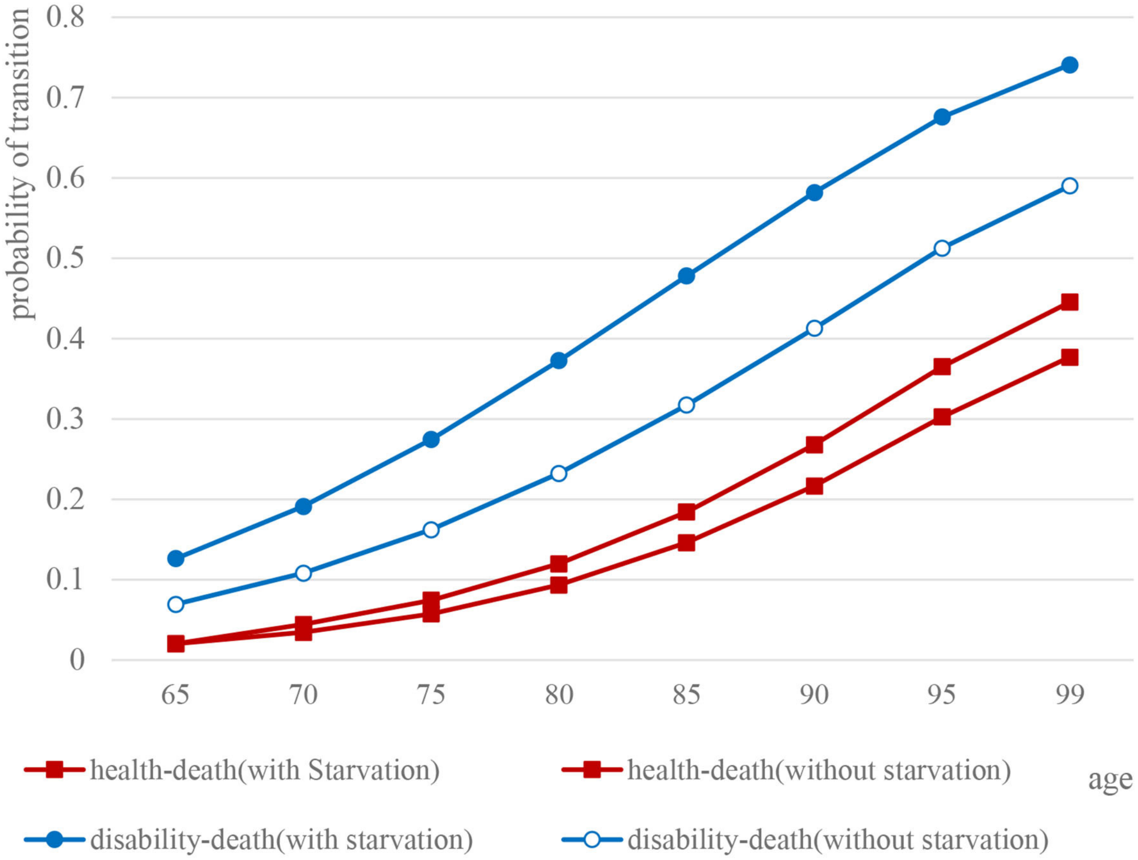


FIGURE 4 | Comparison of health and death transition probability curves of elderly males.

the result is closer to the health level of the crowd. This study found that regardless of the elderly with or without starvation in childhood, the transition probabilities showed an upward trend such as health–disability, health–death, and disability–death with age, while the probability of disability–health showed a downward trend. The elderly people with starvation in all age groups were lower than those who did not experience starvation in childhood. This result reflects that the impact of childhood nutrition on the health of people depends on the degree of hunger in childhood. Disability is a reversible state that can return to health, or it can lead to death. It means that we should pay more attention to the problem of malnutrition in childhood, earlier detection, earlier intervention, and earlier treatment, to not cause lasting adverse health effects.

The elderly who experienced starvation in childhood are lower than those without hunger in the three indicators, such as average life expectancy and healthy life expectancy, which result is closely related to the transition probability. Older people who experienced starvation in childhood had a higher probability of health–disability, but the probability of disability–health recovery was relatively low. Therefore, the probability of disability–health recovery is the key indicator to explain the difference between the two groups above. The lower health recovery rate may reflect lower utilization of medical services, on which the social status of education and economic status affect the conditions and quality of medical service utilization (31, 32). Therefore, good education and economic conditions can not only increase their utilization of health resources but also increase awareness of preventive healthcare, which can effectively reduce the possibility of disability and increase the rate of disability–health recovery (33). The elderly without childhood hunger has obvious advantages in this respect.

The policy enlightenment brought by the research is that the improvement of the material living standard in recent years, and the support of the social security system, cannot completely offset negative effects on the health status of the elderly with childhood starvation experience. Therefore, the government should strengthen nutrition and health interventions for poor children and effectively improve the nutrition and health status of children in poverty-stricken areas and families through the implementation of nutrition improvement programs for preschool children. At the same time, health investment on children is an important prerequisite for the elderly people bonus. At present, the delayed retirement age has taken shape in China, but the smooth implementation of this policy depends largely on the health of the elderly. The research in this article shows that the health problems of the elderly population should be considered from the perspective of the life course, and policymakers should

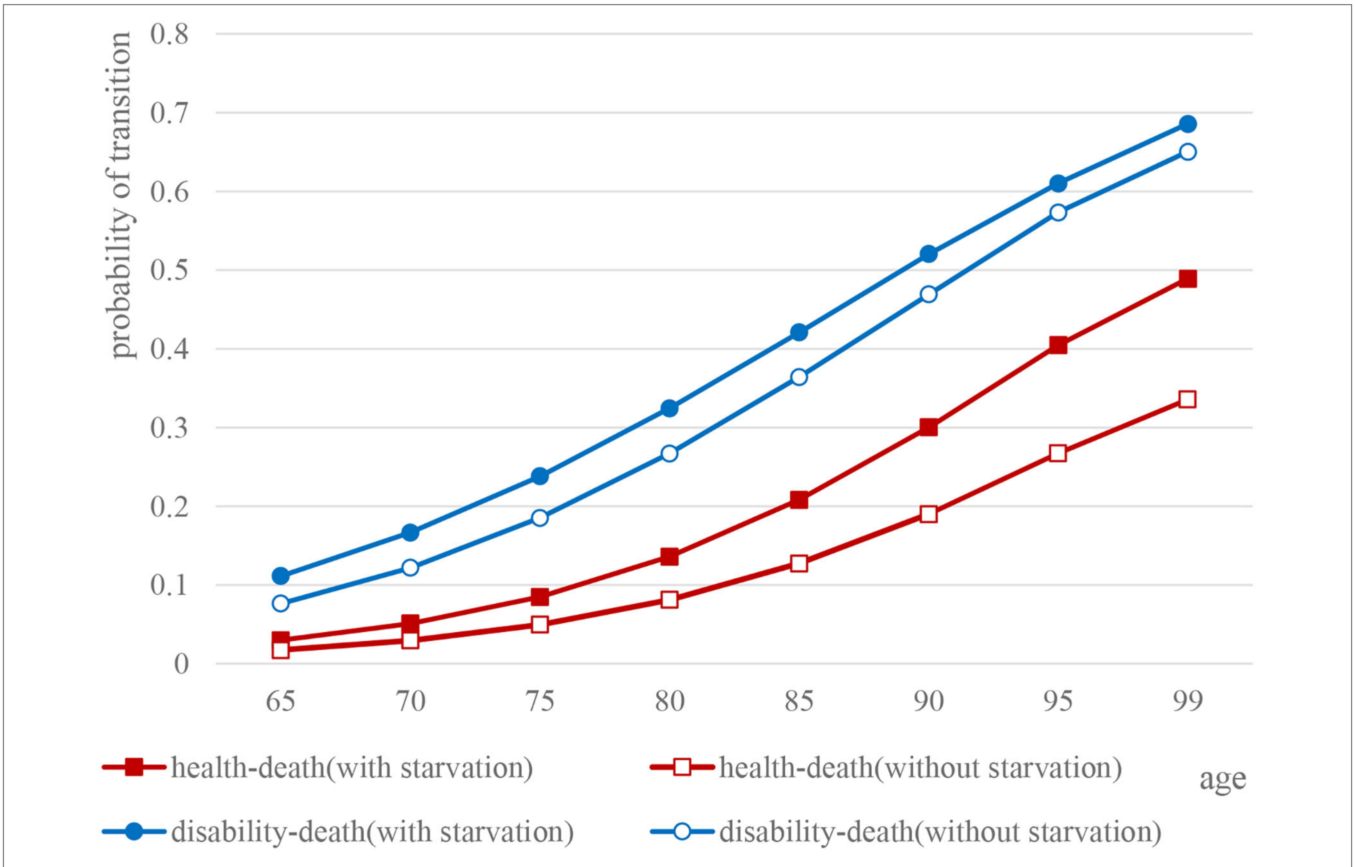


FIGURE 5 | Comparison of health and death transition probability curves of elderly females.

TABLE 2 | Comparison of healthy life expectancy among the elderly population (x ± s).

Age group	**With starvation**			**Without starvation**		
	LE	**HLE**	**HLE/LE (%)**	**LE**	**HLE**	**HLE/LE (%)**
Man						
65–69	14.36 ± 0.27	12.26 ± 0.26	85.3	14.78 ± 0.23	12.70 ± 0.21	85.92
70–74	11.02 ± 0.24	9.03 ± 0.23	82.55	11.34 ± 0.19	9.39 ± 0.18	82.72
75–79	8.23 ± 0.21	6.36 ± 0.19	77.19	8.46 ± 0.17	6.65 ± 0.16	78.54
80–84	6.02 ± 0.18	4.30 ± 0.17	71.22	6.18 ± 0.24	4.53 ± 0.16	73.18
85–89	4.38 ± 0.16	2.80 ± 0.15	63.68	4.48 ± 0.14	2.99 ± 0.13	66.48
90–94	3.22 ± 0.13	1.77 ± 0.15	54.51	3.28 ± 0.12	1.92 ± 0.13	58.39
95–99	2.44 ± 0.11	1.08 ± 0.15	44.03	2.47 ± 0.09	1.22 ± 0.12	49.08
Woman						
65–69	17.06 ± 0.33	14.06 ± 0.30	82.3	17.42 ± 0.25	14.48 ± 0.22	83.1
70–74	13.38 ± 0.30	10.51 ± 0.27	78.43	13.65 ± 0.22	10.87 ± 0.20	79.56
75–79	10.19 ± 0.38	7.50 ± 0.27	73.42	10.37 ± 0.20	7.79 ± 0.18	74.99
80–84	7.58 ± 0.24	5.09 ± 0.21	67.02	7.67 ± 0.19	5.32 ± 0.16	69.21
85–89	5.55 ± 0.22	3.29 ± 0.18	59.16	5.59 ± 0.17	3.48 ± 0.15	62.14
90–94	4.07 ± 0.19	2.04 ± 0.16	49.96	4.07 ± 0.15	2.20 ± 0.14	53.87
95–99	3.04 ± 0.16	1.22 ± 0.15	39.97	3.01 ± 0.13	1.36 ± 0.13	44.77

①LE: average life expectancy; ②HLE: healthy life expectancy; ③HLE/LE: healthy life expectancy accounted for the remaining life.

have a forward-looking awareness to strengthen the nutritional improvement and health promotion of vulnerable groups in the early life.

First, due to the limitations of IMaCH model, if all the individual and family socioeconomic status and environmental variables were included as control variables, it was very difficult to calculate. Therefore, this study only took gender and age as the basic variable and the current income level as the control variable in the model. The second is that if the nutritional status in childhood is too bad and leads to death, then these people will not appear in the sample of 2008–2014. Therefore, the estimates obtained from the sample used in this article may have survivor bias, which will underestimate the impact of childhood starvation experience on the health of the elderly. Third, to explore the impact of childhood hunger on the health of the elderly, there will be a problem of recall bias. But the way to correct recall bias is to expand the sample size. The cohort data used in this study have a larger sample size of 13,185, which can greatly reduce the bias of recall on the research results.

CONCLUSIONS

The negative impact of childhood starvation on health through life course till old age has a persistent negative cumulative effect on the elderly health. The average life expectancy and healthy life expectancy of the elderly with childhood starvation both are lower than those of the elderly without childhood starvation. This study meant that for the social groups with poor early nutritional status, the upward mobility of adult social class and the improvement of material living conditions can not completely offset the negative effects of the early hunger experience. Therefore, in order to achieve healthy aging, government decision-makers should have a sense of foresight and take systematic intervention of all factors affecting health from the early stage of life. The conclusions of this study are important to the comprehensive understanding of the elderly health impact mechanism and the evaluation of current child nutrition projects, such as child nutrition improvement projects in poor areas, nutrition improvement projects in preschool children, etc.,

DATA AVAILABILITY STATEMENT

The data used in this study are openly available in the Peking University open research data at: https://opendata.pku.edu.cn/dataset.xhtml?persistentId=doi:10.18170/DVN/XRV2WN.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Weifang Medical University. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

BW conceived and designed the study and contributed to materials/analysis tools. HD gathered and analyzed the data and wrote the paper. CD reviewed, edited, and approved the manuscript. All authors have read and approved the final manuscript.

FUNDING

The study was funded by the National Office for Philosophy and Social Sciences (Grant Nos. 18BRK013 and 17ZDA124). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

ACKNOWLEDGMENTS

The authors would like to thank all the participants**.** They also offer their gratitude for the data support provided by the Chinese Longitudinal Healthy Longevity Survey (CLHLS) research group.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.690645/full#supplementary-material

REFERENCES

1. Meng X, Qian N. The long run health and economic consequences of famine on survivors: evidence from China's great famine. *IZA Discuss Pap.* (2006) 11:1–67.
2. Habibi E, Sajedi F, Afzali HM, Hatamizadeh N, Shahshahanipour S, Glascoe FP. Early childhood development and iranian parents' knowledge: a qualitative study. *International J Prev Med.* (2017) 8:84. doi: 10.4103/ijpvm.IJPVM_159_17
3. Haddad L, Ross J, Oshaug A, Le T, Cogill B. *Report on the World Nutrition Situation: Nutrition for Improved Development Outcomes, 5th.* (2004). doi: 10.1111/j.1751-0813.1963.tb04208.x
4. Currie J, Almond D. Human capital development before age five. *Handb Labor Econ.* (2014) 11:1315–486. doi: 10.1016/S0169-7218(11)02413-0
5. Akbulut-Yuksel M. War during childhood: the long run effects of warfare on health. *J Health Econ.* (2017) 53:117–30. doi: 10.1016/j.jhealeco.2017.02.005
6. Neelsen S, Stratman T. Effects of prenatal and early life malnutrition: evidence from the Greek famine. *Soc Sci Electron Publ.* (2011) 3:479–88. doi: 10.1016/j.jhealeco.2011.03.001
7. Maccini S, Yang D. Under the weather: health, schooling, and economic consequences of early-life rainfall. *Am EconRev.* (2009) 3:1006–26. doi: 10.1257/aer.99.3.1006
8. Schellenberg JA, Victora CG, Mushi A, de Savigny D, Schellenberg D, Mshinda H, et al. Inequities among the very poor: health care for children in rural southern Tanzania. *Lancet.* (2003) 361:561–6. doi: 10.1016/S0140-6736(03)12515-9
9. Chen Y, Zhou LA. The long-term health and economic consequences of the 1959–1961 famine in China. *J Health Econ*. (2007) 26: 659–81. doi: 10.1016/j.jhealeco.2006.12.006
10. Cui H, Smith JP, Zhao Y. Early-life deprivation and health outcomes in adulthood: evidence from childhood hunger episodes of middle-aged and elderly Chinese. *J Dev Econ.* (2020) 143:1–50. doi: 10.1016/j.jdeveco.2019.102417

11. McGonagle K A, Schoeni R F. The panel study of income dynamics: overview and summary of scientific contributions after nearly 40 years. *Conference Longitudinal Social and Health Surveys in an International Perspective, Montreal*. (2006).
12. Lumey LH, Stein AD, Susser E. Prenatal famine, and adult health. *Ann Rev Public Health.* (2011) 32:237–62. doi: 10.1146/annurev-publhealth-031210-101230
13. Meng X, Qian N. The long-term consequences of famine on survivors: evidence from a unique natural experiment using China's great famine. *Natl Bur Econ Res.* (2009) 14917:1–34. doi: 10.3386/w14917
14. Almond D, Currie J. Killing me softly: the fetal origins hypothesis. *J Econ Perspect.* (2011) 3:153–72. doi: 10.1257/jep.25.3.153
15. Kim S, Fleisher B, Sun JY. The long term health effects of fetal malnutrition: evidence from the 1959–1961 China great leap forward famine. *Health Econ.* (2017) 10:1264–77. doi: 10.1002/hec.3397
16. Kesternich I, Siflinger B, Smith JP, Winter JK. The effects of World War II on economic and health outcomes across Europe. *Rev Econ Stat.* (2014) 1:103–18. doi: 10.1162/REST_a_00353
17. Bertoni M. Hungry today, unhappy tomorrow? Childhood hunger and subjective wellbeing later in life. *J Health Econ.* (2015) 40:40–53. doi: 10.1016/j.jhealeco.2014.12.006
18. Kesternich I, Siflinger B, Smith JP, Winter JK. Individual behavior as a pathway between early-life shocks and adult health: evidence from hunger episodes in post-war Germany. *Econ J.* (2015) 125:372–93. doi: 10.1111/ecoj.12281
19. Paola Z, Head J, Steptoe A. Behavioural risk factors and healthy life expectancy: evidence from two longitudinal studies of ageing in England and the US. *Sci Rep.* (2020) 10:2157–63. doi: 10.1038/s41598-020-63843-6
20. Pedersen J, Bjorner JB. Worklife expectancy in a cohort of Danish employees aged 55–65 years-comparing a multi-state Cox proportional hazard approach with conventional multi-state life tables. *BMC Public Health.* (2017) 1:879. doi: 10.1186/s12889-017-4890-7
21. Zhang WJ, Wen M. Estimation of disability level and time of the elderly in China: an analysis based on combined data. *Popul Stud.* (2015) 39:3–14.
22. Huang RT. On the concept of child and its age definition in forensic identification standard. *The Fifth National Forensic academic Exchange Conference*. China (1996).
23. Qin SS. An important analysis of children's mental health care in different ages. *Contemp Med.* (2014) 24:159.
24. Lievre A, Brouard N, Heathcote C. The estimation of health expectancy from longitudinal surveys. *Math Popul Stud.* (2003) 4:211–48. doi: 10.1080/713644739
25. Shi ZL, Wu ZM. In the early years, the long-term impact on health inequality: the history of life and the double-cumulative disadvantage. *Sociol Res.* (2018) 195:170–96.
26. Xu H, Li L, Zhang Z. Is natural experiment a cure? Re-examining the long-term health effects of China's 1959–1961 famine. *Soc Sci Med.* (2016) 148:110–22. doi: 10.1016/j.socscimed.2015.11.028
27. Doblhammer G, van den Berg GJ, Lumey LH. Long-term effects of famine on life expectancy: a re-analysis of the great finnish famine of 1866-1868. *Popul Stud.* (2011) 67:309–322. doi: 10.1080/00324728.2013.809140
28. Cole MA, Neumayer E. The impact of poor health on total factor productivity: an empirical investigation. *J Dev Stud.* (2006) 42:918–38. doi: 10.1080/00220380600774681
29. He Q, Yuan Y. The intertemporal income effect of health and nutritional status on Adulthood childhood. *Econ Rev.* (2014) 2:52–64.
30. Brown J, Michie S, Geraghty AWA, Yardley L, Gardner B, Shahab L, et al. Internet-based intervention for smoking cessation (StopAdvisor) in people with low and high socioeconomic status: a randomized controlled trial. *Lancet Respir Med.* (2014) 12:997–1006. doi: 10.1016/S2213-2600(14)70195-X
31. Yang JJ, Yoon HS, Lee SA, Choi JY, Song M, Han S, et al. Metabolic syndrome and sex-specific socio-economic disparities in childhood and adulthood: the Korea National Health and Nutrition Examination Surveys. *Diabet Med.* (2014) 11:1399–409. doi: 10.1111/dme.12525
32. Blackwell DL, Martinez ME, Gentleman JF, Sanmartin C, Berthelot JM. Socioeconomic status, and utilization of health care services in Canada and the United States: findings from a binational health survey. *Med Care.* (2009) 47:1136–46. doi: 10.1097/MLR.0b013e3181adcbe9
33. Sheehan CM. Education and health conditions among the currently incarcerated and the non-incarcerated populations. *Popul Res Policy Rev.* (2019) 1:73–93. doi: 10.1007/s11113-018-9496-y

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Public Health

DATA REPORT
published: 15 July 2021
doi: 10.3389/fpubh.2021.666844

A Map of the Initiatives That Harmonize Patient Cohorts Across the World

Ángel Rodríguez-Laso [1], *Laura Alejandra Rico-Uribe* [2], *Christine Kubiak* [3], *Josep Maria Haro* [4], *Leocadio Rodríguez-Mañas* [1,5,6] *and José Luis Ayuso* [2,7,8]*

[1] *Thematic Area for Frailty and Healthy Ageing of the Network of Biomedical Research Centers (CIBERFES), Instituto de Salud Carlos III, Madrid, Spain,* [2] *Centro de Investigación Biomédica en Red de Salud Mental, Instituto de Salud Carlos III, Madrid, Spain,* [3] *European Clinical Research Infrastructure Network (ECRIN-ERIC), Paris, France,* [4] *Parc Sanitari Sant Joan de Déu, Barcelona, Spain,* [5] *Biomedical Research Foundation, Hospital Universitario de Getafe, Madrid, Spain,* [6] *Geriatric Department, Hospital Universitario de Getafe, Madrid, Spain,* [7] *Department of Psychiatry, Universidad Autónoma de Madrid, Madrid, Spain,* [8] *Instituto de Investigación Sanitaria Princesa, Hospital Universitario de La Princesa, Madrid, Spain*

Keywords: clinical cohort, patient cohort, harmonization, integration, repository

OPEN ACCESS

Edited by:
Laura Beane Freeman,
National Cancer Institute (NCI),
United States

Reviewed by:
Yutong Samuel Cai,
University of Oxford, United Kingdom
Milagros Ruiz,
University College London,
United Kingdom

Correspondence:
José Luis Ayuso
joseluis.ayuso@uam.es

Specialty section:
This article was submitted to Life-Course Epidemiology and Social Inequalities in Health, a section of the journal Frontiers in Public Health

Received: *11 February 2021*
Accepted: *14 June 2021*
Published: *15 July 2021*

Citation:
Rodríguez-Laso Á, Rico-Uribe LA, Kubiak C, Haro JM, Rodríguez-Mañas L and Ayuso JL (2021) A Map of the Initiatives That Harmonize Patient Cohorts Across the World. Front. Public Health 9:666844. doi: 10.3389/fpubh.2021.666844

INTRODUCTION

Integration of cohort studies allows taking advantage of already collected information to increase the sample size to study uncommon exposures, rare diseases, less strong associations, or very restricted populations (personalized medicine). It also allows to carry out standardized analyses and avoid publication bias compared to the analysis of published data (1–5). Nevertheless, the growing energy spent in conducting cohort studies across the world in the last decades has not been paralleled by an effort to make them accessible to the scientific community and harmonize their data. This last limitation moved the European Commission to fund the SYNergies for Cohorts in Health: integrating the ROle of all Stakeholders (SYNCHROS) coordination and support action, endowed with almost €2 million[1] from 2019 to 2021. It aims "to establish a sustainable European strategy for the development of the next generation of integrated cohorts, thereby contributing to an international strategic agenda for enhanced coordination of cohorts globally, in order to address the practical, ethical, legal, and methodological challenge of optimizing the exploitation of current and future cohort data, toward the development of stratified and personalized medicine as well as facilitating health policy."

In order to achieve its objectives, the first activity proposed in SYNCHROS was to map the population, patient, and clinical trial cohort integration landscape. That would allow the project to have a first look at the challenges and tried solutions adopted by different groups, and, more importantly, it would provide a list of principal investigators of these initiatives who could be contacted for the process of developing the common strategy. This study reports the result of the mapping of the initiatives that integrate patient cohorts. The mapping of population cohorts will be reported elsewhere. The aim of the study was to obtain a non-exhaustive, but representative, list of these initiatives carried out in recent times in the world. To our knowledge, there is no other repository of integration initiatives of patient cohorts. Although excellent single cohort repositories exist, like the Maelstrom catalog, repositories of initiatives that integrate several patient cohorts could not be found.

This mapping will provide researchers with a useful tool to find initiatives on their areas of interest with whom they can share or analyze harmonized data.

[1] https://cordis.europa.eu/project/id/825884/es.

METHODS

The initiatives included in the mapping were obtained from three different sources:

1. Systematic searches, carried out in MEDLINE and the Maelstrom catalog.[2]
2. Suggestions of potential initiatives to be included in the mapping provided by partners of the SYNCHROS consortium.
3. References and links provided by the initiatives detected in the two previous sources.

The inclusion criteria were as follows:

a) initiatives that integrated patient, clinical, or disease cohorts;
b) individual patient meta-analysis and mega-analyses; and
c) at least one cohort included in the initiative having information about the sample at two or more points of time (at least two waves).

The exclusion criteria were as follows:

a) initiatives that only integrate population cohorts or clinical trials, without including patient cohorts;
b) initiatives published before the year 2000; and
c) initiatives that did not provide information in English.

Database Searches

MEDLINE Search

The process started with searches restricted to papers published in English from 2000 to 2019 using the terms selected by consensus among the SYNCHROS partners. Those terms which obtained fewer than 500 hits were retained, and the abstracts of the hits were reviewed to find new terms that were used in subsequent searches. In some cases, the term "cohort" was added to these searches to limit the number of hits.

The final search strategy used is given as follows:

(cohort OR "prospective study" OR "longitudinal study" OR "individual meta-analysis"[All Fields] OR "individual participant data meta-analysis"[All Fields] OR "individual patient data meta-analysis"[All Fields] OR "individual meta analysis"[All Fields] OR "individual participant data meta analysis"[All Fields] OR "individual patient data meta analysis"[All Fields] OR "meta analysis using individual"[All Fields] OR "meta-analysis using individual"[All Fields] OR "meta analysis of individual"[All Fields] OR "meta-analysis of individual"[All Fields] OR "mega-analysis"[All Fields] OR "mega analysis"[All Fields])

AND

("harmonization study" OR "integration study" OR "integration initiative" OR "integrated study" OR "merged cohort" OR "data pooling" OR "pooled sample" OR "combined data" OR "combining data" OR "harmonized data" OR "harmonised data" OR "harmonizing data" OR "data harmonization" OR "data harmonisation" OR "data sharing" OR "common database" OR "multiple cohorts" OR "multiple longitudinal studies" OR "international consortium" OR "collaborative effort").

AND

("2000/01/01"[Date - Publication]: "2019/07/31"[Date - Publication])

AND

English[Language]

AND

Humans[MeSH]

Maelstrom Catalog

The Maelstrom research catalog, supported by the Research Institute of the McGill University Health Centre, "contains comprehensive information about epidemiological research networks and studies, and the data they have collected. It also provides information about harmonized data generated by these research networks."

We looked for initiatives included in the "Networks" section of the catalog.

Selection of Initiatives

Initiatives that were obtained from the systematic searches and provided by the partners were evaluated against the inclusion and exclusion criteria by two different investigators. In case of a disagreement, a third reviewer was consulted.

Extraction of Information

The following information was extracted from each initiative: name of the initiative, principal investigator, partners, name of the institution responsible for the initiative, funding resources, contact person, information source, whether the research team is currently active, main objectives, criteria for the cohorts to be included in the initiative, type of harmonization (prospective/retrospective), number of cohorts included in the initiative (the total number and the number of harmonized cohorts), whether more cohorts are foreseen to be harmonized, number of participants (the total number and the number of participants with harmonized data), age range of the sample, threats to representativeness of the sample, maximum number of variables that have been harmonized, including those where harmonization was not possible for all the cohorts, setting of the harmonized cohorts (local-regional/national/international, including country of origin of the cohorts), and a brief description of the population considered by the initiative.

All this information was retrieved from the webpage and/or the scientific article that presented the initiative. Missing information was requested from the principal investigators of the projects, who were contacted initially by email and, if there was no answer, by phone call or by post.

ANALYSIS

Results of the identification process of the initiatives are presented in **Figure 1**.

Partners of the SYNCHROS project provided 39 initiatives. Of those, 28 were excluded, mainly because there was no data harmonization or because eligibility could not be ascertained due to unresponsiveness from the principal investigators. The

[2] https://www.maelstrom-research.org/maelstrom-catalogue.

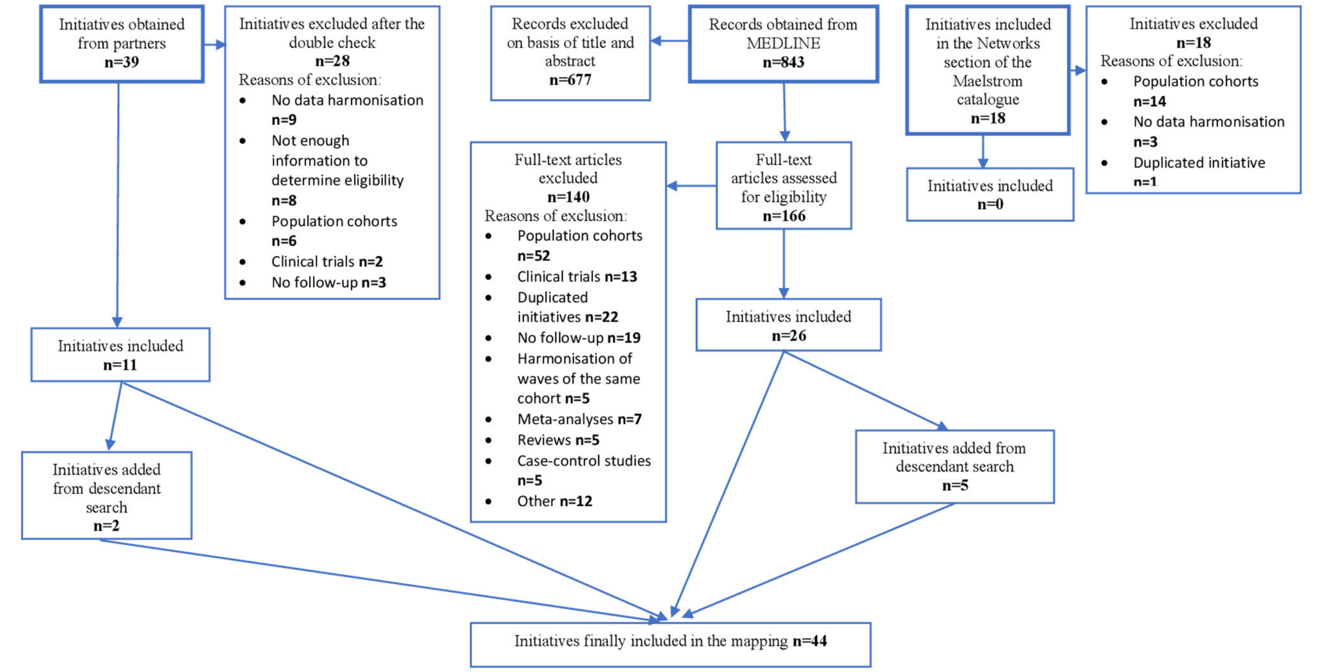

Initiatives obtained from partners n=39
Initiatives excluded after the double check n=28
Reasons of exclusion:
• No data harmonisation n=9
• Not enough information to determine eligibility n=8
• Population cohorts n=6
• Clinical trials n=2
• No follow-up n=3
Records excluded on basis of title and abstract n=677
Records obtained from MEDLINE n=843
Initiatives included in the Networks section of the Maelstrom catalogue n=18
Initiatives excluded n=18
Reasons of exclusion:
• Population cohorts n=14
• No data harmonisation n=3
• Duplicated initiative n=1
Full-text articles excluded n=140
Reasons of exclusion:
• Population cohorts n=52
• Clinical trials n=13
• Duplicated initiatives n=22
• No follow-up n=19
• Harmonisation of waves of the same cohort n=5
• Meta-analyses n=7
• Reviews n=5
• Case-control studies n=5
• Other n=12
Full-text articles assessed for eligibility n=166
Initiatives included n=0
Initiatives included n=26
Initiatives included n=11
Initiatives added from descendant search n=2
Initiatives added from descendant search n=5
Initiatives finally included in the mapping n=44

FIGURE 1 | Results of the search for harmonization initiatives of patient cohorts.

remaining 11 initiatives were selected. The descendent search from these initiatives provided two additional ones.

In the MEDLINE search, out of 843 hits obtained, 677 were excluded after reading their title and abstract. Of the remaining articles, 166 were read and, from those, 140 excluded. The main reasons for exclusion were that initiatives dealt only with population cohorts, that they had already been submitted by partners or already presented in another reference, or that the integration was only cross-sectional. In the end, 26 initiatives were selected. The reference list of these initiatives included five additional ones.

The search in the Maelstrom catalog only provided initiatives that harmonized population cohorts.

Overall, 44 initiatives were retrieved. They are presented in **Table 1**.

Table 1 shows a selection of the most relevant information obtained from each of the initiatives. Complete information can be found in the repository of the SYNCHROS project.[3] They are ordered by types of diseases covered (starting with those which consider several diseases) and by alphabetical order. Of the 44 initiatives found, no further information could be obtained from principal investigators in almost half (20) of them.

Eight initiatives (BIOMAP, CINECA, EHDEN, ESCAP-NET, HarmonicSS, HARMONY, Lifebrain, and ReCoDID) have recently started adding cohorts; 21 are led by active research teams; and 12 are adding, or considering adding, cohorts now. Nevertheless, there is plenty of missing information on the activity status of the initiatives.

In the selected initiatives, the most represented group of diseases is cancer (10 initiatives), followed by infectious diseases (8 initiatives, of which 5 focus on HIV) and cardiovascular disease (4 initiatives). There are five initiatives that have harmonized data from more than one type of disease. Other diseases and conditions producing a high burden in the high-income countries (6) are represented (dementia, osteoarthritis), but others included in this list (unipolar depressive disorders, alcohol use disorders, hearing loss, chronic obstructive pulmonary disease, diabetes mellitus, road traffic accidents) or poor-defined conditions with a well-defined impact on life-expectancy and quality of life (like back pain or functional deterioration) are missing. There is one initiative about a specific rare disease (Sjögren syndrome).

There is a sizable number of initiatives that have harmonized other types of cohorts in addition to patient cohorts. After Breast Cancer Pooling Project, BIOMAP, CLL-IPI, HARMONY, and the initiatives on obsessive-compulsive disorder and pulmonary embolism have harmonized at least one clinical trial cohort. CINECA, ESCAPE-NET, Lifebrain, and the project "Seasonal plasticity of cognition" have also harmonized population cohorts. BiomarCaRE and the National Cancer Institute Cohort Consortium have harmonized the three types of cohorts: patient, population and clinical trials cohorts.

Most of them (33) have an international scope, compared to seven national initiatives and one regional/local initiative. Two initiatives report that they include cohorts from across the world and eight initiatives incorporate cohorts from high- and low- and middle-income countries (LMIC); 30 (75%) initiatives

[3] https://www.synchros.es.

TABLE 1 | Initiatives that harmonize patient cohorts ordered by different categories of diseases (selected information).

Initiative	Types of cohorts	Region, Country were the cohorts were collected	Main objective	Number of cohorts with harmonized data	More cohorts foreseen to be harmonized?	Number of participants with harmonized data	Age range of the sample	No. of harmonized variables (Maximum)	Population
CINECA: Common Infrastructure for National Cohorts in Europe, Canada and Africa	Disease cohorts. Population cohorts	Africa, Canada and Europe	To develop a federated cloud enabled infrastructure to make population scale genomic and biomolecular data accessible across international borders, to accelerate research, and improve the health of individuals across continents	In progress	Possibly	In progress	Birth to old age	In progress	The dataset provides a diverse representation of studies in rare disease, common disease and national cohorts over time (longitudinal)
CNODES: the Canadian Network for Observational Drug Effect Studies	Disease cohorts	Canada, US and UK	Use collaborative, population-based approaches to obtain rapid answers to questions about drug safety and effectiveness	Depends on the research question	No	Depends on the research question	All ages	Depends on the research question	Population of Canada, UK and US which is prescribed or dispensed drugs
EHDEN: European Health Data and Evidence Network	Disease cohorts	All Horizon 2020 member states and associated countries	Harmonize in excess of 100 million anonymized health records to the OMOP common data model, supported by an ecosystem of certified SMEs, and technical architecture for a federated network		In progress		18	Considerable	European patients aged 18+
MIRACUM: Medical Informatics in Research and Care in University Medicine	Disease cohorts	Seven states of Germany	The spotlight is here on the data integration centers that will be embedded in the hospital IT-infrastructure and will facilitate the collection and exchange of data within the consortia university hospitals. Furthermore, we will elaborate a programme for strengthening medical informatics by extending the academic offer, including new professorships in the field of medical informatics, a novel, innovative master programme and personnel training. The MIRACUM partners have agreed to share data, based on interoperable data integration centers, develop common and interoperable tools and services, realize the power of such data and tools in innovative IT solutions, which shall enhance patient-centered collaborative research as well as clinical care processes, and finally to strengthen biomedical informatics in research, teaching and continued education	11	No information obtained	No information obtained	0 to the highest age of patients	No information obtained	Patients attended in hospitals of seven German states

(Continued)

TABLE 1 | Continued

Initiative	Types of cohorts	Region, Country were the cohorts were collected	Main objective	Number of cohorts with harmonized data	More cohorts foreseen to be harmonized?	Number of participants with harmonized data	Age range of the sample	No. of harmonized variables (Maximum)	Population
Sentinel initiative	Disease cohorts	US	Serve as a system to analyze and assess safety risks in FDA-approved drugs and medical products using electronic health data	17	No information obtained	310 million	All ages	No information obtained	Population of US which is prescribed or dispensed drugs
BiomarCaRE: Biomarker for Cardiovascular Risk Assessment across Europe	Disease cohorts. Population cohorts. Clinical trials	Australia, Europe, Israel, Latin America, New Zealand, South Africa, United States	Assess the value of established and emerging biomarkers for cardiovascular risk prediction	4	No information obtained	8.746	No information obtained	No information obtained	Patients with coronary heart disease or at risk of developing it
CADISP: Cervical Artery Dissection and Ischemic Patients	Disease cohorts	Western Europe and Turkey	International Consortium performing research on ischemic stroke in young and middle-aged adults and in particular on cervical artery dissection	No information obtained	No information obtained	No information obtained	No information obtained	No information obtained	Cervical artery dissection and ischemic stroke patients from some Western European countries and Turkey
Development and validation of the AMPREDICT model	Disease cohorts	US	The objective of this study was the development of AMPREDICT-Mobility, a tool to predict the probability of independence in either basic or advanced mobility 1 year after dysvascular major lower extremity amputation	2	No information obtained	200	No information obtained	38	Individuals undergoing their first major lower extremity amputation because of complications of peripheral artery disease or diabetes
ESCAPE-NET: European Sudden Cardiac Arrest network: toward Prevention, Education and NEw Treatment	Disease cohorts. Population cohorts	Czech Republic, Denmark, France, Italy, Sweden, The Netherlands	Aims to study: (1) risk factors and mechanisms for the occurrence of sudden cardiac arrest (SCA) in the population, and (2) risk factors and treatment strategies for survival after SCA on a European scale	No information obtained	Yes	No information obtained	No information obtained	No information obtained	Patients with sudden cardiac arrest
After Breast Cancer Pooling Project	Disease cohorts (one is based on the follow-up of a randomized clinical controlled trial)	China (Shanghai), US	Examine the role of physical activity, adiposity, dietary factors, supplement use, and quality of life in breast cancer prognosis	4	Yes	18.314	20–83	No information obtained	Breast cancer survivors (women). Cancers were diagnosed between 1976 and 2006

(Continued)

TABLE 1 | Continued

Initiative	Types of cohorts	Region, Country were the cohorts were collected	Main objective	Number of cohorts with harmonized data	More cohorts foreseen to be harmonized?	Number of participants with harmonized data	Age range of the sample	No. of harmonized variables (Maximum)	Population
B-CAST: Breast CAncer STratification	Disease cohorts	No information obtained	In B-CAST tools will be developed to allow precise identification of the individual risk of breast cancer, the subtype of cancer that is most likely to develop and the prognosis of that particular subtype	No information obtained	No information obtained	No information obtained	No information obtained	No information obtained	Patients with breast cancer
Collaborative Group on Epidemiological Studies of Ovarian Cancer	Disease cohorts	Worldwide	Study risk factors of oavarian cancer	58	No information obtained	31,000	No information obtained	No information obtained	Women with ovarian cancer
GENIE: Genomics Evidence Neoplasia Information Exchange	Disease cohorts	Canada, France, Netherlands, Spain, UK, USA	It is a multi-phase, multi-year, international data-sharing project that aims to catalyze precision cancer medicine	19	Yes	70,000	All ages	No information obtained	Cancer patients treated at multiple international institutions
HARMONY: European Public-Private Partnership for Big Data in Hematology	Disease cohorts. Clinical trials	All Europe	The HARMONY Alliance uses big data technologies to improve the treatment of seven hematologic malignancies	Acute Myeloid Leukemia: 5 patient cohorts. Multiple myeloma: 15 patient cohorts	In progress	11,664 (aims to harmonize between 75,000 and 100,000 anonymized hematologic patients by the end of the funding period)	All ages are considered	It depends on the specific research question	Patients with blood malignancies
International Collaboration of Epidemiological Studies of Cervical Cancer	Disease cohorts	Costa Rica, Denmark, Norway, Sweden, UK, US	Study the effects of hormonal contraceptive use and other factors on the risk of cervical cancer	9	No information obtained	2,109	No information obtained	No information obtained	Women with cervical cancer
MaGIC: Malignant Germ Cell International Consortium	Disease cohorts	No information obtained	Developing more effective treatments for germ cell tumors (GCT) through scientific inquiry	No information obtained	No information obtained	No information obtained	No information obtained	No information obtained	GCT patients all over the world
NCI: National Cancer Institute Cohort Consortium	Disease cohorts. Population cohorts. Clinical trials	Australia, Canada, New Zealand, USA	Foster communication among investigators leading cohort studies of cancer, promote collaborative research projects for topics not easily addressed in a single study and identify common challenges in cohort research and search for solutions	No information obtained	Yes	No information obtained	18+	No information obtained	Breast and colon family cancer patients and their families

(Continued)

TABLE 1 | Continued

Initiative	Types of cohorts	Region, Country were the cohorts were collected	Main objective	Number of cohorts with harmonized data	More cohorts foreseen to be harmonized?	Number of participants with harmonized data	Age range of the sample	No. of harmonized variables (Maximum)	Population
Second primary malignancies in thyroid cancer patients	Disease cohorts	France, Italy, Sweden	Evaluate the risk of second cancer and leukemia in patients with papillary or follicular thyroid cancer treated with radioiodine or external beam radiation therapy	3	No	6,841	7–80 (at time of diagnosis of thyroid cancer)	Around 10	Patients with papillary or follicular thyroid cancer
The International CLL-IPI working group	Disease cohorts. Clinical trials	France, Germany, Poland, UK, US	We established an international consortium with the aim to create an international prognostic index for chronic lymphocytic leukemia (CLL-IPI) that integrates the major prognostic parameters	2		1,254	No information obtained	18	Chronic lymphocytic leukemia patients
COASt: Clinical Outcomes in Arthroplasty Study	Disease cohorts	Europe	Describe whether body mass index is a clinically meaningful predictor of patient reported outcomes following primary total hip replacement (THR) surgery	4	No information obtained	4,413	No information obtained	24	Patients receiving primary THR for osteoarthritis
MARC-35: 35th Multicenter Airway Research Collaboration	Disease cohorts	US	Examine the association between the infectious etiology of a child's severe bronchiolitis and the level of serum 25-hydroxyvitamin D (25[OH]D) during severe bronchiolitis, with the severity of this illness, and the subsequent development of recurrent wheezing by age 3 years and combine these clinical and laboratory data to derive the wheezing index that will identify children at higher risk of developing recurrent wheezing by age 3 years	17	No	920	0–1	Thousands	Children age $<$ 1 year hospitalized with severe bronchiolitis
COSMIC: Cohort Studies of Memory in an International Consortium	Disease cohorts	The world	Harmonizing shared, non-identifiable data from cohort studies that longitudinally examine change in cognitive function and the development of dementia in older individuals (60+ years).	Data are harmonized on a project-by-project basis, and only subgroups of the member studies contribute to particular projects	Yes	Data are harmonized on a project-by-project basis, and only subgroups of the member studies contribute to particular projects	40–105	Harmonization is done on a project-by-project basis and the number of studies per project varies. For the largest project with 20 studies there are 16 harmonized variables	60+ years old individuals from 29 countries all over the world

(Continued)

TABLE 1 | Continued

Initiative	Types of cohorts	Region, Country were the cohorts were collected	Main objective	Number of cohorts with harmonized data	More cohorts foreseen to be harmonized?	Number of participants with harmonized data	Age range of the sample	No. of harmonized variables (Maximum)	Population
Lifebrain: Healthy minds 0–100 years: Optimizing the use of European brain imaging cohorts	Disease cohorts. Population cohorts	Western Europe	Maximize the exploitation of brain imaging cohorts by bringing together studies on how differences and changes in brain age relate to cognitive function and mental health	1 of anxiety and depression patients	No information obtained	2,981	No information obtained	No information obtained	Patients with anxiety or depression
Seasonal plasticity of cognition	Disease cohorts. Population cohorts	Canada, France, US	Test the hypotheses that season has a significant association with cognition, the odds of being diagnosed with mild cognitive impairment or dementia, cerebrospinal fluid Alzheimer disease biomarkers, and the expression of cognition-associated modules of coexpressed genes in the human brain	2	No information obtained	592	No information obtained	No information obtained	Alzheimer disease patients or patients with cognitive disorders visited in tertiary care clinics
HarmonicSS: HARMONIzation and integrative analysis of regional, national and international Cohorts on primary Sjögren's Syndrome (pSS) toward improved stratification, treatment and health policy making disease	Disease cohorts. Clinical trials	Europe, US	To bring together the largest well-characterized regional, national and international longitudinal cohorts of patients with Primary Sjögren's Syndrome (pSS) including those participating in clinical trials, and by taking into consideration the ethical, legal, privacy and intelectual propiety rights issues for sharing data from different countries, to semantically interlink and harmonize them into an integrative pSS cohort structure on the cloud	No information obtained	No information obtained	No information obtained	No information obtained	No information obtained	Cohorts and clinical trials of patients with Primary Sjögren's Syndrome
SABER: SAfety Assessment of Biologic ThERapy	Disease cohorts	US	Understanding the absolute and comparative risks of adverse events of biologic treatments for patients with autoimmune diseases	4	No information obtained	239,806	All ages	No information obtained	Patients with autoimmune diseases who had at least one dispensing of a biologic agent or comparison non-biologic regimen relevant to their autoimmune disease
Thousand Faces of Lupus	Disease cohorts	Canada	Evaluate factors affecting therapeutic approaches used in clinical practice for the management of systemic lupus erythematosus (SLE), in a multicenter cohort	10	No information obtained	1,497	No information obtained	No information obtained	Patients who meet American College of Rheumatology (ACR) criteria for Systemic Lupus Erythematosus

(Continued)

TABLE 1 | Continued

Initiative	Types of cohorts	Region, Country were the cohorts were collected	Main objective	Number of cohorts with harmonized data	More cohorts foreseen to be harmonized?	Number of participants with harmonized data	Age range of the sample	No. of harmonized variables (Maximum)	Population
Tumor Necrosis Factor alpha antagonist use and cancer in patients with rheumatoid arthritis	Disease cohorts	Canada and US	Estimate the association between treatment with biologic disease-modifying antirheumatic drugs (DMARDs) and development of cancer in patients with rheumatoid arthritis	3	No information obtained	8,458	65+	No information obtained	Rheumathoid arthritis patients who had been prescribed DMARDs or methotrexate
GEMRIC: Global ECT-MRI Research Collaboration	Disease cohorts	Japan, Western Europe and US. Currently approaching China	Creating a large database of multi-site imaging data and clinical/behavioral/physiological and metadata for analysis of the neural mechanisms and predictors of electroconvulsive therapy-related clinical response	15	Yes	345	19–86	More than a thousand because the initiative includes diagnostic imaging variables	Patients receiving electroconvulsive therapy
Predictors and moderators of cognitive and behavioral therapy outcomes for obsessive-compulsive dissorder	Disease cohorts. Clinical trials	Australia, Canada, Europe, and US	Identify potential factors that affect the outcome of cognitive and behavioral treatments of obsessive-compulsive disorders	8	No	359	18+ (very few over 65)	Around 20	Patients with obsessive-compulsive disorders
Antibiotic treatment and survival of nursing home patients with lower respiratory tract infection	Disease cohorts	The Netherlands and US	Assess the effects of different antibiotic treatment strategies on survival of elderly nursing home residents with lower respiratory tract infections in the United States and the Netherlands, where treatment approaches are quite different	2	No	1,221	70+	Around 40	Elderly nursing home residents with lower respiratory tract infections
ART-CC: Antiretroviral Therapy Cohort Collaboration	Disease cohorts	Western Europe and North America	Estimate prognosis of HIV-1 positive, treatment naïve patients initiating highly active antiretroviral therapy (ART)	No information obtained	No information obtained	No information obtained	No information obtained	No information obtained	HIV-1 positive, treatment naïve patients cohorts from Europe and North America

(Continued)

TABLE 1 | Continued

Initiative	Types of cohorts	Region, Country were the cohorts were collected	Main objective	Number of cohorts with harmonized data	More cohorts foreseen to be harmonized?	Number of participants with harmonized data	Age range of the sample	No. of harmonized variables (Maximum)	Population
COHERE: Collaboration of Observational HIV Epidemiological Research Europe	Disease cohorts	Western Europe and North America	Pool and harmonize existing longitudinal data on HIV-positive persons collected across Europe to answer key research questions that, in the era of potent combination antiretroviral therapy (cART), could not be addressed adequately by individual cohorts	No information obtained	No information obtained	No information obtained	No information obtained	Not reported	HIV-infected people residing in Europe
Early Antibiotic Treatment for Pediatric Febrile Urinary Tract Infection and Renal Scarring	Disease cohorts	US	Determine, in a well-characterized sample of children with febrile urinary track infections, whether delay in the initiation of antimicrobial therapy was associated with the occurrence and severity of renal scarring and to determine whether these associations persisted after adjusting for potential confounding factors	2	No	802	2–72 months	No information obtained	Children aged 2–72 months with a urinary tract infection producing fever
HAART and early mortality	Disease cohorts	Brazil and US	Compare the early mortality pattern and the causes of death among patients starting HAART in Brazil and the United States	2	No information obtained	1,774	No information obtained	10	HIV-infected patients
IeDEA: International epidemiology Databases to Evaluate AIDS	Disease cohorts	Africa, Asia-Pacific region, the Central/South America/Caribbean region, and North America	Collect and define key variables, harmonize data, and implement methodology to effectively pool data as a cost-effective means of generating large data sets to address the high priority research questions and streamline HIV/AIDS research	No information obtained	No information obtained	No information obtained	No information obtained	No information obtained	HIV/AIDS patients from Africa, the Asia-Pacific region, the Central/South America/Caribbean region, and North America
ReCoDID: Reconciliation of Cohort data in Infectious Diseases	Disease cohorts	No information obtained	Develop an equitable, accessible, and sustainable model for the storage, curation, and analyses of clinical-epidemiological and high-dimensional sample data collected by infectious disease cohorts in low-and-midle-income countries	No information obtained	In progress	No information obtained	No information obtained	No information obtained	Patients with infectious diseases
RESPOND: International Cohort Consortium of Infectious Disease	Disease cohorts	Australia, Georgia and Western Europe	Build an innovative, flexible and dynamic cohort consortium for the study of infectious diseases, including HIV and people at risk for HIV, as a generic structure for facilitating multi stakeholder involvement	No information obtained	No information obtained	No information obtained	No information obtained	No information obtained	People 18+ at high risk of acquiring HIV and people living with HIV and/or with other infectious diseases or across Europe, South America and Australia

(Continued)

TABLE 1 | Continued

Initiative	Types of cohorts	Region, Country were the cohorts were collected	Main objective	Number of cohorts with harmonized data	More cohorts foreseen to be harmonized?	Number of participants with harmonized data	Age range of the sample	No. of harmonized variables (Maximum)	Population
Adults Born Preterm International Collaboration	Disease cohorts	Australia, Canada, Finland, Netherlands, Northern Ireland, Norway, US	Our main aim was to identify factors that either increase or decrease risk of high blood pressure among adults born with very low birth weight	9	No information obtained	1,571 patients and 777 controls	No information obtained	No information obtained	Very low birth weight and very preterm babies who reach adulthood
Necrotizing enterocolitis (NEC) study	Disease cohorts	Austria and The Netherlands	The first aim of the study was to correlate the occurrence of a blood stream infection (BSI) during the early phase of necrotizing enterocolitis (NEC) with intestinal fatty acid-binding protein (I-FABP) levels, as a marker for loss of gut wall integrity owing to mucosal damage, and Interleukin (IL)-8 levels, as a biomarker for the pro-inflammatory cascade in NEC. The second aim of the study was to investigate the relation between the occurrence of a BSI and disease outcome	2	No information obtained	57	24–40 weeks	13	Patients with necrotizing enterocolitis
Recurrent leg venous ulcers study	Disease cohorts	Eastern Australia	Identify risk and protective factors for recurrence of venous leg ulcers	3	Yes	250	26–96	24	Patients with a healed leg ulcer of primarily venous etiology
MARS: Multicenter AVM Research Study	Disease cohorts	Scotland and US	Identify risk factors for intracranial hemorrhage in the natural history course of brain arteriovenous malformations	4	Yes	2,525	No information obtained	13	Patients with arteriovenous malformations
Pulmonary embolism presentation	Disease cohorts (one clinical trial)	Belgium, France and Switzerland	Compare clinical characteristics between women and men with suspected and confirmed pulmonary embolism (PE) and their impact on clinical probability prediction scores and on diagnostic work-up of PE, and to assess whether differences at presentation could account for the increased recurrence rate in men	3	No	3,414	18–98	Around 30	Patients with a clinical suspicion of pulmonary embolism
BIOMAP: Biomarkers in Atopic Dermatitis and Psoriasis	Disease cohorts. Clinical trials	No information obtained	Examine the causes and mechanisms of atopic dermatitis and psoriasis to enable optimal treatments and an individualized therapy scheme for each patient	No information obtained	In progress	No information obtained	No information obtained	No information obtained	Patients with atopic dermatitis and psoriasis

only include cohorts from high-income countries, and none harmonize data from LMIC countries alone.

Most initiatives are partnered with universities, hospitals, and research institutes. Governmental institutions take part in a few of them (9). The presence of patient associations and pharmaceutical companies as partners is anecdotal. The number of partners ranges between 2 and more than 100, with a median of 12. Three quarters comprise 20 partners or fewer.

Most initiatives have been or are funded by American (12) or European (10) institutions. Canadian funding comes third (4). The vast majority have received public funding alone (22). Five have received combined funding from public institutions and non-profit organizations. Private funding was provided in isolation to one initiative (RESPOND), combined with public funding to another one (EHDEN), and combined with non-profit funding to a third one (Tumor necrosis factor α antagonist use).

Their objectives may be classified into four general categories (some initiatives share more than one): determining the prognosis of subgroups of patients (14), providing a repository of patients (11), establishing the efficacy (6) or safety (4) of treatments, and exploring risk factors and biomarkers of diseases (10).

The median number of cohorts included in each initiative is 5, ranging from 1 (which also harmonizes population cohorts) to 58; three quarters include 17 cohorts or fewer. The number of individuals included varies wildly, from 57 to 310 million (Sentinel initiative). The median is 6,841. Eight out of 37 (21.6%) initiatives have harmonized fewer than 1,000 patients and the same proportion have harmonized 100,000 patients or more. Twenty-six have harmonized all or almost all the cohorts incorporated to the initiative, two (EHDEN and CINECA) are still in the process of harmonizing their cohorts and another two (CNODES and COSMIC) harmonize data on a project-by-project basis.

Eight initiatives included patients from all ages, eight included only adult patients, three included only children, and two included exclusively older people.

Of those which have declared the number of variables in their harmonized database, there are between 10 and more than 1,000 (median 24), with two out of 15 (13.3%) including more than 1,000 variables.

Four initiatives harmonized administrative databases. Thirty-three were retrospective vs. four prospective. The great majority do not report major threats to the representativity of their samples.

DATA AVAILABILITY STATEMENT

The dataset generated by this study can be found in the webpage of the SYNCHROS project https://www.synchros.es.

AUTHOR CONTRIBUTIONS

ÁR-L performed the database searches, extracted the information from the initiatives, and drafted the article. ÁR-L, LR-U, and CK evaluated the initiatives against inclusion and exclusion criteria. All authors made substantial contributions to the conception and design of the paper, analysis and interpretation of data, reviewed the manuscript, and read and approved its final version.

FUNDING

This work was supported by the European Union's H2020 Programme under grant agreement number 825884.

ACKNOWLEDGMENTS

The authors want to express their gratitude to all principal investigators and project managers who answered our request for further information on their harmonization initiatives.

REFERENCES

1. Blettner M, Sauerbrei W, Schlehofer B, Scheuchenpflug T, Friedenreich C. Traditional reviews, meta-analyses and pooled analyses in epidemiology. *Int J Epidemiol.* (1999) 28:1–9. doi: 10.1093/ije/28.1.1
2. Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, et al. The PhenX toolkit: get the most from your measures. *Am J Epidemiol.* (2011) 174:253–60. doi: 10.1093/aje/kwr193
3. Ioannidis JPA, Rosenberg PS, Goedert JJ, O'Brien TR. Commentary: meta-analysis of individual participants' data in genetic epidemiology. *Am J Epidemiol.* (2002) 156:204–10. doi: 10.1093/aje/kwf031
4. Lyman GH, Kuderer NM. The strengths and limitations of meta-analyses based on aggregate data. *BMC Med Res Methodol.* (2005) 5:14. doi: 10.1186/1471-2288-5-14
5. Thompson A. Thinking big: large-scale collaborative research in observational epidemiology. *Eur J Epidemiol.* (2009) 24:727–31. doi: 10.1007/s10654-009-9412-1
6. World Health Organization. *The Global Burden of Disease: 2004 Update.* Geneva, Switzerland: World Health Organization (2004).

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 26 July 2021
doi: 10.3389/fpubh.2021.711344

Adverse Childhood Experiences Among 28,047 Norwegian Adults From a General Population

Siri H. Haugland[1]*, *Anders Dovran*[1,2], *Ane U. Albaek*[1] *and Børge Sivertsen*[3,4,5]

[1] Department of Psychosocial Health, University of Agder, Grimstad, Norway, [2] Stine Sofies Foundation and Stine Sofie Centre, Grimstad, Norway, [3] Department of Health Promotion, Norwegian Institute of Public Health, Bergen, Norway, [4] Department of Research and Innovation, Helse Fonna HF, Haugesund, Norway, [5] Department of Mental Health, Norwegian University of Science and Technology, Trondheim, Norway

Aim: The purpose of this study was to estimate the prevalence of adverse childhood experiences (ACEs) among Norwegian adults from a general population and to identify potential associations with demographic and socioeconomic characteristics.

Methods: A randomly drawn sample ($N = 61,611$) from the public registry of inhabitants was invited to participate in the Norwegian Counties Public Health Survey. The present study was based on online responses from 28,047 adults ≥18 years (mean age: 46.9 years, SD = 16.03). Log-link binomial regression analyses were performed to examine associations between four measures of ACEs (family conflict, lack of adult support, bad memories, and difficult childhood) and demographic (age, gender, civil status, parental divorce) and socioeconomic characteristics (education level, perceived financial situation, and welfare benefits).

Results: Single individuals and those with parents that divorced during childhood were at elevated risk of all four ACEs. The risk varied to some degree between the sexes. The prevalence of ACEs declined with increasing age. We found a consistent social gradient that corresponded to the frequency of ACEs for all three socioeconomic characteristics investigated. The risks were highest for those in the lowest socioeconomic levels (RR: 1.53, 95% CI: 1.32–1.78 to RR: 4.95, CI: 4.27–5.74).

Conclusions: Public health strategies should direct more attention to the interplay between ACEs and socioeconomic factors. Welfare services should be sensitive to ACEs among their service recipients.

Keywords: adverse childhood experiences, family conflict, adult survivors of child adverse events, child abuse, socioeconomic factors

OPEN ACCESS

Edited by:
Orla Doyle,
University College Dublin, Ireland

Reviewed by:
Tor Åge Myklebust,
Cancer Registry of Norway, Norway
Tony Kuo,
UCLA Fielding School of Public Health, United States

***Correspondence:**
Siri H. Haugland
siri.h.haugland@uia.no

Specialty section:
This article was submitted to Life-Course Epidemiology and Social Inequalities in Health, a section of the journal Frontiers in Public Health

***Received:** 18 May 2021*
***Accepted:** 29 June 2021*
***Published:** 26 July 2021*

Citation:
Haugland SH, Dovran A, Albaek AU and Sivertsen B (2021) Adverse Childhood Experiences Among 28,047 Norwegian Adults From a General Population. Front. Public Health 9:711344. doi: 10.3389/fpubh.2021.711344

INTRODUCTION

Adverse childhood experiences (ACEs) are stressful and potentially traumatic events experienced by children before the age of 18 years. ACEs are relatively common (1, 2): a recent systematic review and meta-analysis found pooled prevalence of 23.5% Europeans that reported at least one ACE and 18.7% that reported two or more ACEs (3). The term "ACE" originated in the Adverse Childhood Experiences study, conducted in 1998. They grouped ACEs into three domains: abuse, neglect, and household dysfunction (4). Numerous studies have explored the consequences of ACEs, although, the concept has been defined differently among different studies (5). Typically, studies map the occurrence, number, and frequency of various types of adverse experiences (3). Other studies focused on the relevance of how participants rated the impact of their experiences (6). However, irrespective of how ACEs have been defined, the association between ACEs and reduced adult health and well-being has been confirmed repeatedly (4–8). Although, a large body of research has documented the connection between early adversities and adult health problems, fewer studies have explored factors that could influence and interact with this connection. One example is the complex relationship between ACEs and socioeconomic factors. ACEs appear to be highly socially patterned and individuals with low socioeconomic statuses report more ACEs (9). However, due to their occurrence early in life, it is likely that ACEs also impact socioeconomic outcomes in adulthood, such as educational attainment, employment, and income (10, 11). Although, Norway is considered an affluent country, we nevertheless display a social gradient in health and life expectancy (12, 13). The number of children living in low income families (below the poverty threshold) in Norway has tripled over the last two decades, and more people receive disability benefits here compared to other countries that are members of the Organization for Economic Co-operation and Development (14). In addition, Norway has regional differences in education, disability, and health-related measures. The current study was situated in a region characterized by low employment rates, a high distribution of work assessment allowance (unemployment benefit), and a large proportion of people that receive disability benefits. Moreover, among young adults (15–29 years) in this area, the frequency of seeking help for mental health problems was higher than the national average (14). Few studies have investigated the prevalence of ACEs and their association with socioeconomic and demographic factors in a Nordic context and in a general population of adults. However, this information is important to improve our understanding of the inequalities and determinants of health in Western societies. Epidemiological studies that identify high-risk groups are essential in developing policy and service-delivery systems directed toward reducing the negative consequences of ACEs (15).

The purpose of this study was to estimate the prevalence of adverse childhood experiences (ACEs) among Norwegian adults from a general population and to identify potential associations with demographic and socioeconomic characteristics.

METHODS

Study Design and Setting

A total of 61,611 inhabitants that resided in Agder county (Southern Norway), aged 18 or older, were invited to participate in an online questionnaire (Norwegian Counties Public Health Survey). The questions were related to health, well-being, childhood, living conditions, local environments, accidents, and injuries. All data were collected electronically through a web-based platform. Participants were selected randomly from the Norwegian Population Registry of inhabitants in Southern Norway; e-mails or telephone numbers were obtained from the contact registry from the Norwegian Agency for Public Management and eGovernment.

Instruments

Adverse Childhood Experiences

ACEs were assessed with these four questionnaire items:

1. *Did you experience a lot of arguing, turmoil, conflicts, or difficult communication in your childhood home?*
2. *Growing up, did you have a trusted adult from whom you could get support?*
3. *Do you struggle with bad memories from your childhood, due to loss, betrayal, neglect, violence, ill-treatment, or abuse?*
4. *When you think about your childhood/upbringing, how would you describe it?*

Response options for items 1–3 were: “not at all,” “to a very small degree,” “to a small degree,” “to a large degree,” and “to a very large degree.” The last three categories in item 1 were coded as a “dysfunctional family environment”; the first two categories in item 2 were coded as a “lack of trusted adult during childhood”; and the last three categories in item 3 were coded as a “struggle with bad memories.” Item 4 included the following response categories “very good,” “good,” “moderate,” “difficult,” and “very difficult.” The last two categories were coded as “perceived childhood as difficult.” Finally, one item assessed the participant’s experience with parental divorce, with the response options “no,” “yes, before I was 7 years old,” and “yes, when I was between 7 and 18 years old.”

These four ACE items were originally developed for a large Norwegian public health study (the HUNT study). Items 1–3 first appeared in the fourth wave of the study (HUNT4), and item 4 was included in both the third and fourth waves (HUNT3 and HUNT4). Before that, the ACE items were included in a pilot-testing of the HUNT4 questionnaire in the municipality of Selbu, where, 31 participants provided written comments to the questions. In addition, six participants were interviewed in detail by telephone. Of particular interest, the pilot study tested the comprehensibility of the questions and whether participants found them uncomfortable or invasive. No negative comments regarding the ACE items were received.

In a previous study, items 1, 3, and 4 were validated together in a short, Difficult Childhood questionnaire (DCQ). The discriminant and convergent validities of the tool were confirmed (16) in the same population that we analyzed in the current study.

Demographic Factors

Participant age and sex were obtained from the National Population Registry. Age was divided into 6 age groups (18–29, 30–39, 40–49, 50–59, 60–69, and 70+ years). All participants were also asked about their relationship status (coded as “single” vs. “married/partner” (including “girlfriend/boyfriend”).

Socioeconomic Variables

Three variables measured different aspects of socioeconomic status (SES). The educational level was collected by asking participants about the highest level of education completed.

The response categories were: "Low secondary/secondary modern/folk high school up to 10 years," "Vocational training/middle school/upper secondary/junior college-minimum 3 years," "University college/university <4 years," and "University college/university 4 years or more." In the current study, these categories were renamed "low," "medium," "high," and "highest" educational level, respectively. Economic capability was assessed with the following question: "How easy or difficult is it for you to make ends meet with your current income?" Response options ranged from 1 (very difficult) to 6 (very easy). For the purposes of the current study, these responses were revised to include three categories: "poor," "medium," "good." Participants were also asked whether their employment status included receiving disability pension/work assessment allowance or social assistance benefits. These categories were collapsed into two: "receiving welfare benefits" or "not receiving welfare benefits." As this last variable allowed the participants to tick off several answers, missing is not possible to estimate for this particular variable.

Ethics

This study was approved by the Norwegian Data Inspectorate and the Regional Committee for Medical and Health Research Ethics of South-East Norway (file number 162353/REK South-East-C), whose directives are based on the Helsinki Declaration. Written electronic consent was provided by all subjects included in the study. All data were stored and processed in compliance with The General Data Protection Regulation.

Statistical Analysis

All analyses were performed with IBM SPSS version 26 (SPSS Inc., Chicago, IL USA) for Windows. The overall distribution of ACEs, relative to demographic and socioeconomic factors were performed in cross-tables. Log-link binomial regression analyses were performed to examine associations between ACEs and demographic and socioeconomic factors. Rather than the more commonly used logistic regression model to obtain an odds ratio (OR), we used log-link binomial regressions, to obtain risk ratios (RR), and 95% confidence intervals (95% CIs). All analyses

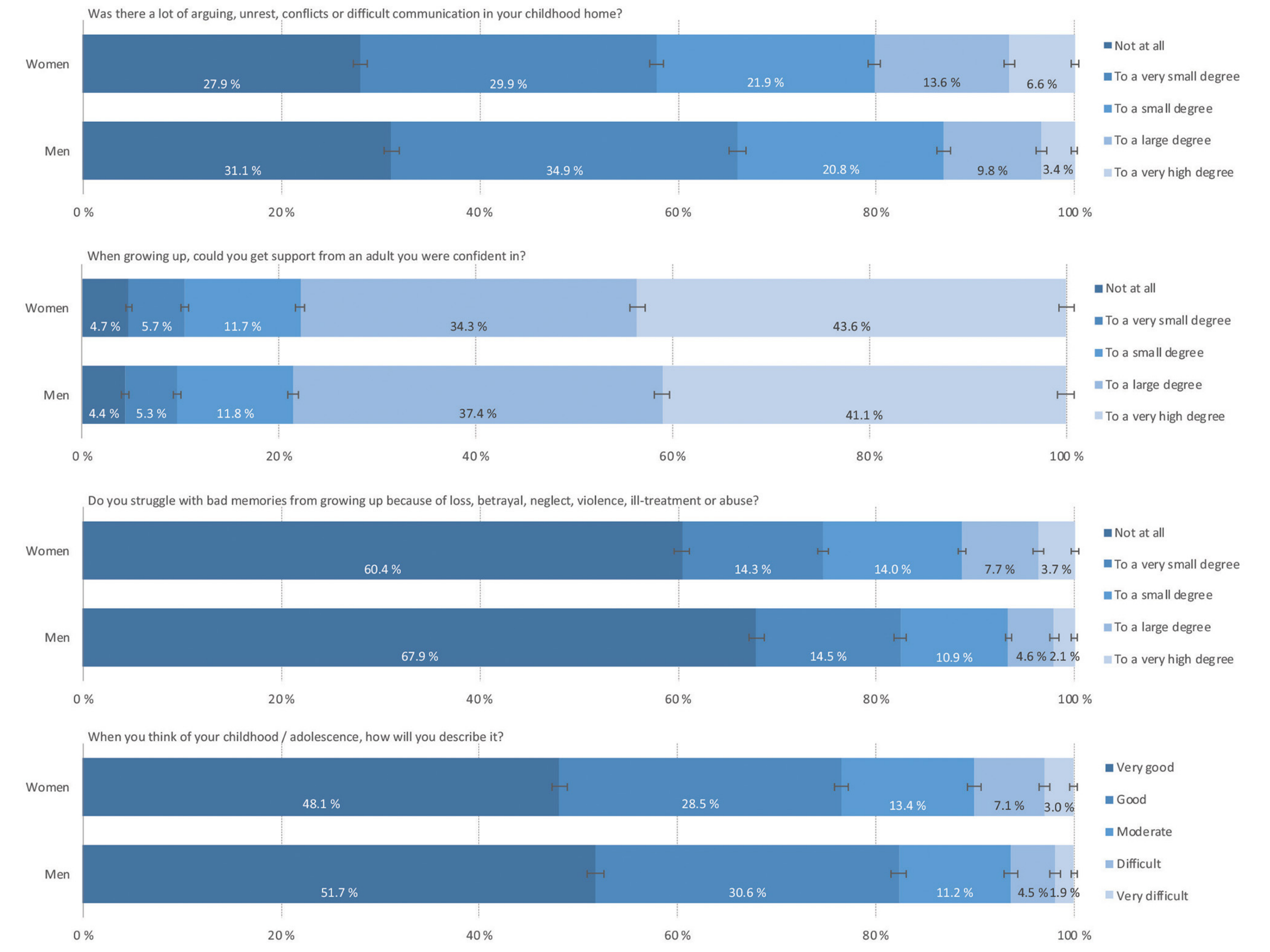


FIGURE 1 | The prevalence of ACEs among Norwegian adults from a general population, grouped by participant sex. Data retrieved from the Norwegian Counties Public Health Survey conducted in Agder, 2019.

TABLE 1 | Demographic and socioeconomic characteristics of adults with adverse childhood experiences; data from a Norwegian Counties Public Health Survey conducted in Agder, 2019 (*N* = 28,047).

Characteristic	Dysfunctional family environment *n* (%)	Lack of trusted adult during childhood *n* (%)	Struggle with bad memories *n* (%)	Perceives childhood as difficult *n* (%)
Sex				
Male	1,725 (13.2)	2,802 (21.5)	873 (6.7)	841 (6.4)
Female	3,008 (20.2)	3,290 (22.1)	1,696 (11.4)	1,505 (10.1)
Age group				
18–29 years	1,077 (20.7)	997 (19.1)	641 (12.3)	526 (10.1)
30–39 years	970 (21.6)	1,040 (23.2)	548 (12.2)	514 (11.4)
40–49 years	1,062 (19.2)	1,261 (22.8)	534 (9.6)	526 (9.5)
50–59 years	944 (16.9)	1,308 (23.5)	494 (8.8)	474 (8.5)
60–69 years	505 (10.7	1,000 (21.2)	256 (5.4)	225 (4.8)
70 years +	175 (7.3)	486 (20.2)	96 (4.0)	81 (3.3)
Parents divorced during childhood				
Yes, age <7 years	1,038 (44.0)	889 (37.7)	623 (26.4)	599 (25.3)
Yes, age 7–18 years	1,201 (43.4)	990 (35.8)	559 (20.2)	562 (20.3)
No	2,476 (10.9)	4,191 (18.4)	1,381 (6.1)	1,175 (5.2)
Marital status				
Single	1,291 (21.3)	1,626 (26.8)	807 (13.3)	760 (12.5)
Married/partner	3,432 (15.7)	4,451 (20.4)	1,757 (8.0)	1,582 (7.2)
Educational level				
Low	696 (20.9)	1,001 (30.3)	468 (14.1)	454 (13.6)
Medium	1,847 (16.7)	2,528 (22.9)	1,088 (9.8)	985 (8.9)
High	1,067 (16.4)	1,317 (20.3)	524 (8.1)	464 (7.1)
Highest	1,105 (15.8)	1,221 (17.5)	476 (6.8)	433 (6.2)
Financial difficulties				
Poor	768 (32.6)	923 (39.1)	610 (25.8)	548 (23.2)
Medium	812 (25.6)	962 (30.3)	493 (15.5)	432 (13.6)
No difficulty	2,978 (14.1)	3,923 (18.6)	1,360 (6.4)	1,267 (6.0)
Welfare Benefits				
Yes	1,006 (31.4)	1,180 (37.0)	819 (25.6)	702 (21.9)

were conducted separately for males and females. We also tested for interactions between sex and the other demographic and socioeconomic factors by entering the product of these variables in separate blocks. Missing data were handled with a listwise deletion method.

RESULTS

Of the 61,611 individuals invited, 28,047 completed the questionnaire, which yielded a response rate of 45.5%. The sample had a mean age of 46.9 years (SD = 16.03) and consisted of 13,122 (46.8%) males and 14,925 females (53.2%).

Figure 1 shows the prevalence of the different ACEs. More females (20.2%) than males (13.2%) reported frequent family conflicts. In contrast, the sexes were less different in the support perceived; approximately 10% reported a lack of support (small to very small degrees of support). Struggling with bad memories from childhood was reported by 11.4% of females and 6.7% of males. Similarly, 10.1% of females and 6.4% of males characterized their childhood as difficult or very difficult.

A comparison between the different age groups showed a declining trend, where older age groups reported less ACEs than younger age groups (data not shown). For all the ACE-related variables, females reported higher frequencies than males.

Table 1 shows the percentages of individuals that reported the four ACEs among different demographic and socioeconomic groups. Within the socioeconomic groups, the highest proportions of participants that reported ACEs were in the lower socioeconomic groups (low education levels, poor economic capability, and recipients of welfare benefits) and came from a background with parental divorce. **Figure 2** visualize the social gradient in the prevalence of ACE by education.

As detailed in **Table 2**, participants that reported that their parents had divorced during childhood had an elevated overall risk for all four ACEs, with RRs ranging from 1.85 to 5.62. We found that this association significantly interacted with sex; indeed, males generally showed stronger associations than females between having divorced parents and experiencing three of the four ACE outcomes: a dysfunctional family environment, struggling with bad memories, and perceiving childhood as difficult. The risk of ACEs declined with age; it was lowest among the oldest participants. We also found that single individuals (without partner) showed elevated risks for all four ACEs, compared to individuals with a partner (RRs ranged from 1.26

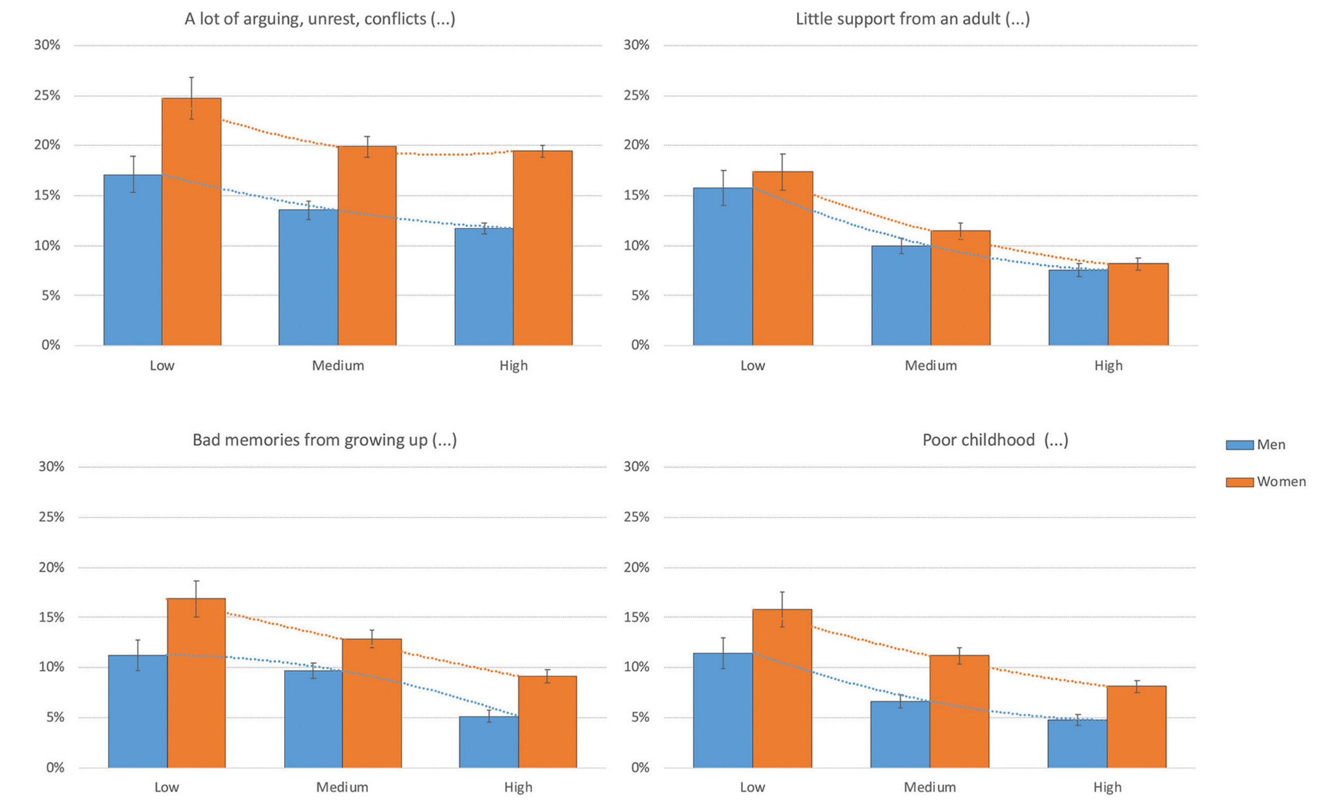


FIGURE 2 | The prevalence of ACEs among Norwegian adults from a general population, grouped by education. Data retrieved from the Norwegian Counties Public Health Survey conducted in Agder, 2019.

to 1.92). We also found that males generally had stronger associations than females between a single marital status and a dysfunctional family environment and perceiving childhood as difficult (**Table 2**).

The risk of ACEs was highest in disadvantaged subgroups (i.e., those with a low education level, poor economic capability, or recipients of welfare benefits), with RRs ranging from 1.28 (95% CI: 1.16–1.43) to 4.95 (95% CI: 4.27–5.74). Among participants with poor economic capability and participants that received welfare benefits, males were at significantly higher risk than females of struggling with bad memories. Similarly, among participants with low education and those that received welfare benefits, males were at significantly higher risk than females of perceiving childhood as difficult (**Table 2**).

DISCUSSION

Overall, our results showed that that the prevalence of ACEs (family conflict, lack of adult support, struggling with bad memories, and difficult childhood) in a large Norwegian adult sample drawn from the general population varied across demographic variables (i.e., age, gender, marital status, and a background of divorced parents). We also showed that exposure to childhood adversities was associated with low socioeconomic status in adulthood, including variables like low education levels, perceived financial difficulties, and receiving welfare benefits.

In our sample, the proportions of males and females that reported ACEs varied with age, where few of the oldest participants reported ACEs. Although, this result was consistent with results from previous studies (17), it may be somewhat surprising, because one might have expected that childhood conditions would have been worse among the oldest individuals. There are several possible explanations for this age-related decline in ACEs. First, the questions were retrospective, and therefore, they were susceptible to recall bias (18); this bias might have been more pronounced among the oldest participants. Second, studies have shown that ACEs were strongly related to multimorbidity (18) and premature mortality (17). This association may have introduced a selection bias, where the oldest individuals with high levels of ACEs might have been underrepresented. Furthermore, there may be differences between age cohorts in their understanding of what qualifies as a difficult childhood and their expectations of how childhood should be. Although, these are plausible explanations for our findings, another Norwegian study (18) evaluated one of our ACE items (a difficult childhood) and did not find any significant differences between age groups in the levels of self-reported childhood difficulties.

Single participants had a modestly increased risk of ACEs compared to those with a partner. Childhood adversities, such as family conflicts, might influence relationship aspects. For example, a study by Roth et al. (19) found that general

TABLE 2 | Risk of adverse childhood experiences and demographic and socioeconomic characteristics, among Norwegian adults from a general population; data from Norwegian Counties Public Health Survey conducted in Agder 2019 ($N = 28,047$).

Characteristic	Dysfunctional family environment				Lack of trusted adult during childhood				Struggle with bad memories				Perceives childhood as difficult			
	Males		Females		Males		Females		Males		Females		Males		Females	
	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI	RR	95% CI
Age group	Sex int. Wald (df) = 0.982(5), $p = 0.964$				Sex int. Wald (df) = 8.331(5), $p = 0.139$				Sex int. Wald (df) = 6.676(5), $p = 0.246$				Sex int. Wald (df) = 7.014(5), $p = 0.220$			
18–29 years	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–
30–39 years	1.04	0.907–1.19	1.06	0.967–1.16	1.18	1.05–1.34	1.23	1.11–1.36	1.14	0.944–1.38	0.942	0.828–1.07	1.24	1.01–1.51	1.09	0.950–1.26
40–49 years	0.918	0.804–0.1.08	0.963	0.879–1.05	1.15	1.02–1.29	1.23	1.18–1.36	0.791	0.648–0.964	0.814	0.716–0.926	0.877	0.716–1.07	1.01	0.879–1.16
50–59 years	0.824	0.72–0.943	0.849	0.77–93	1.23	1.10–1.38	1.22	1.11–1.34	0.783	0.643–0.953	0.725	0.633–0.829	0.848	0.693–1.03	0.873	0.754–1.01
60–69 years	0.527	0.449–0.618	0.547	0.483–0.620	1.17	1.04–1.32	1.04	0.939–1.16	0.473	0.374–0.598	0.462	0.388–0.550	0.468	0.365–0.599	0.509	0.421–0.617
70 years +	0.346	0.274–0.437	0.399	0.326–0.488	1.11	0.973–1.27	0.999	0.865–1.15	0.302	0.214–0.427	0.395	0.304–0.513	0.295	0.204–0.426	0.410	0.306–0.550
Parents divorced during childhood	Sex int. Wald (df) = 39.833(2), $p = 0.000$				Sex int. Wald (df) = 1.042(2), $p = 0.594$				Sex int. Wald (df) = 6.255(2), $p = 0.044$				Sex int. Wald (df) = 6.390(2), $p = 0.041$			
Yes, age <7 years	4.77	4.14–5.05	3.58	3.34–3.83	1.85	1.63–2.09	2.01	1.81–2.23	3.49	2.97–4.11	3.11	2.80–3.47	4.10	3.49–4.82	3.71	3.24–4.24
Yes, age 7–18 years	5.05	4.58–5.76	3.45	3.21–3.72	2.04	1.79–2.31	2.04	1.83–2.29	4.94	4.26–5.72	3.94	3.55–4.36	5.62	4.84–6.51	4.42	3.96–4.94
No	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–
Marital status	Sex int. Wald (df) = 4.77(1), $p = 0.029$				Sex int. Wald (df) = 0.04(1), $p = 0.307$				Sex int. Wald (df) = 0.007(1), $p = 0.931$				Sex int. Wald (df) = 4.01(1), $p = 0.045$			
Single	1.47	1.31–1.65	1.25	1.14–1.37	1.26	1.14–1.39	1.35	1.24–1.47	1.83	1.58–2.13	1.51	1.35–1.69	1.92	1.65–2.23	1.58	1.41–1.77
Married/partner	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–
Educational level	Sex int. Wald (df) = 5.741(3), $p = 0.125$				Sex int. Wald (df) = 0.689(3), $p = 0.876$				Sex int. Wald (df) = 39.833(3), $p = 0.225$				Sex int. Wald (df) = 7.406(3), $p = 0.006$			
Low	1.53	0.1.32–1.78	1.28	1.16–1.43	1.68	1.51–1.88	1.71	1.60–1.95	2.39	1.93–2.96	2.02	1.75–2.35	2.42	1.96–3.00	2.18	1.86–2.54
Medium	1.21	1.07–1.37	1.03	0.956–1.127	1.29	1.18–1.42	1.32	1.21–1.43	1.46	1.21–1.77	1.54	1.36–1.74	1.40	1.16–1.70	1.54	1.35–1.76
High	1.11	966–1.28	1.02	0.939–1.128	1.12	1.01–1.25	1.18	1.08–1.30	1.18	0.949–1.47	1.21	1.05–1.39	0.995	0.79–1.25	1.26	1.08–1.46
Highest	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–
Economic capabilities	Sex int. Wald (df) = 2.146(2), $p = 0.342$				Sex int. Wald (df) = 0.590(2), $p = 0.745$				Sex int. Wald (df) = 4.36(2), $p = 0.000$				Sex int. Wald (df) = 4.179(2), $p = 0.124$			
Poor	2.40	2.13–2.70	2.18	2.01–2.37	2.07	1.89–2.25	2.12	1.96–2.29	4.95	4.27–5.74	3.46	3.11–3.84	4.29	3.68–5.01	3.53	3.15–3.94
Medium	1.87	1.66–2.10	1.74	1.60–1.89	1.58	1.44–1.73	1.66	1.53–1.79	2.70	2.28–3.20	2.21	1.97–2.48	2.29	1.92–2.74	2.19	1.93–2.
Good	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–
Welfare benefits	Sex int. Wald (df) = 5.144(1), $p = 0.023$				Sex int. Wald (df) = 0.234(1), $p = 0.622$				Sex int. Wald (df) = 6.297(1), $p = 0.012$				Sex int. Wald (df) = 4.244(1), $p = 0.039$			
Yes	2.45	1.90–2.54	1.90	1.72–2.09	1.91	1.69–2.15	1.84	1.67–2.01	4.08	3.47–4.78	3.18	2.84–3.55	3.65	3.09–4.31	2.95	2.62–3.32
No	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–	1.00	–

confidence in sustainable romantic relationships was lower among individuals that were exposed to overt parental conflict during childhood.

Not surprisingly, individuals that experienced a parental divorce during childhood had a higher risk of ACEs than those with parents that did not divorce. Furthermore, adults that experienced a parental divorce during childhood were more prone to struggling with bad memories from their childhood due to loss, betrayal, neglect, violence, ill treatment, or abuse. This finding was consistent with other studies that revealed long-term associations between parental divorce and a wide range of different mental health problems (20).

Overall, our findings with socioeconomic measures revealed a strong, consistent social gradient that corresponded to the risks of all four ACEs. Among participants that reported financial difficulties, the risk of also struggling with bad memories from childhood was particularly increased among males, but the overall association was strong for both sexes. Similarly, participants with financial difficulties were at high risk of characterizing their childhood as difficult, although, this association was strongest among males. Among participants that received welfare benefits, the pattern was similar, with a pronounced increase in the risk of ACEs. The relationship between education level and ACEs also revealed a gradient; indeed, compared to participants with a high education level, participants with medium and low education levels were at increased risks of all ACEs. These associations were strongest among males with a low education level that characterized their childhood as difficult. The social gradient may have a multifactorial origin, although, it was not possible to investigate this hypothesis in the current study. Other studies have revealed that the prevalence of ACEs was highly socially patterned in childhood, which suggested that a low SES in childhood could be a determinant for ACEs (21). Therefore, a low SES in adulthood might represent a continuance of a low family SES during childhood. However, a previous prospective study (22) found that ACEs were associated with low educational attainment, even after adjusting for family socioeconomic factors.

Strengths and Limitations

A major strength of this study was the large sample drawn randomly from a general population, including participants that spanned a large age range. Many previous ACE-related studies focused on mapping the types and frequencies of ACEs. In contrast, the present study invited participants to report self-perceptions of the consequences and severity of ACEs, which may be more relevant in assessing experiences that cause detrimental effects on adult perceptions of quality of life (6, 16).

Although, the study is performed within a Norwegian context, we expect findings to be generalizable to other countries as well. Studies performed in other countries have also revealed a social gradient related to ACEs (9, 11). Moreover, as Norway has a well-functioning welfare system compared to many other countries, the social gradient may even be stronger in contexts outside Norway. The Norwegian health services system is mostly government operated and is freely available to all citizens for a minor self-share fee (maximum 240 € per year including medical supplies). Our social services include graded benefits for adults without employment; 1 year sickness benefit [100% of former wage (FW)], 1 year unemployment benefit (62,4% of FW), 3 year work clearance allowance (66% of FW), disability benefit (66% of FW or a minimum amount), and social benefits (cover necessities).

The limitations of this study include its retrospective design. Thus, recall bias might have increased the risk of measurement error (23). However, a comparison between prospective vs. retrospective reports of ACEs did not reveal any bias in our retrospective assessment (24). Another limitation was the cross-sectional study design; caution should be taken regarding potential causalities when interpreting our findings.

Policy and Practice Implications

Our study indicates that adults who think of their childhood as difficult often experience financial and employment problems. Adding this to our knowledge that adults with ACEs have increased risk for various health problems (5, 25) ACEs impose large human and economic costs on society. The relation between childhood adversity and lifelong well-being warrants fresh thinking on how to promote health and prevent structural inequities. As of now, the majority of societies' resources are allocated to adult health care and adult social services. We suggest that a redirection of resources toward prevention of ACEs, as well as protection and care for children experiencing adversity, will reduce overall human and economic costs. Moreover, interventions focused toward restoring inequities in SES to break intergenerational transmission of low SES and ACEs need to be explored.

CONCLUSION

This study showed a varied distribution of ACEs across demographic variables. In addition, a strong, consistent social gradient was revealed, which point to the necessity of increasing our awareness of the potential role that ACEs play in disturbing the life opportunities of children. This awareness should encourage political discourse to increase efforts to disrupt intergenerational patterns, where low SES and ACEs are transferred and upheld within families. The apparent interconnectivity between ACEs and SES calls for a more diverse set of preventive interventions directed toward both SES-related struggles and toward the prevention and treatment of ACEs, for both adults and children.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the data were provided by the NIPH, with permission. NIPH will make data available in a repository upon application. Requests to access the datasets should be directed to https://helsedata.no/en/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Regional Committee for Medical and Health Research Ethics of South-East-C, Norway (file number 162353).

The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SH, AA, BS, and AD: conceptualization and writing–original draft. BS, AD, and SH: formal analysis and methodology. BS: visualization. All authors contributed to the article and approved the submitted version.

ACKNOWLEDGMENTS

We acknowledge the Norwegian Institute of Public Health and Agder County for performing the data collection, and the inhabitants of Agder for their participation. We also wish to acknowledge the work of the HUNT Databank (HUNT Research Center, Norwegian University of Science and Technology), for developing the four items for measuring adverse childhood events.

REFERENCES

1. Merrick MT, Ports KA, Ford DC, Afifi TO, Gershoff ET, Grogan-Kaylor A. Unpacking the impact of adverse childhood experiences on adult mental health. *Child Abuse Negl.* (2017) 69:10–9. doi: 10.1016/j.chiabu.2017.03.016
2. WHO. *Investing in Children: the European Child Maltreatment Prevention Action Plan 2015–2020*. Copenhagen: World Health Organization (2014).
3. Bellis, MA, Hughes K, Ford K, Rodriguez G, Sethi, D, Passmore J. Life course health consequences and associated annual costs of adverse childhood experiences across Europe and North America: a systematic review and meta-analysis. *Lancet Public Health.* (2019) 4:e517–28. doi: 10.1016/S2468-2667(19)30145-8
4. Felitti VJ, Anda RF, Nordenberg D, Williamson DF, Spitz AM, Edwards V, et al. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults. The Adverse Childhood Experiences (ACE) study. *Am J Prev Med.* (1998) 14:245–58. doi: 10.1016/S0749-3797(98)00017-8
5. Hughes K, Bellis MA, Hardcastle KA, Sethi D, Butchart A, Mikton C, et al. The effect of multiple adverse childhood experiences on health: a systematic review and meta-analysis. *Lancet Public Health.* (2017) 2:e356–66. doi: 10.1016/S2468-2667(17)30118-4
6. LaNoue M, Graeber DA, Helitzer DL, Fawcett J. The relationship between self-reported adult impact of adverse childhood events and health-related quality of life. *J Community Med Health Educ.* (2013) 4:267. doi: 10.4172/2161-0711.1000267
7. Ferrara P, Guadagno C, Sbordone A, Amato M, Spina G, Perrone G, et al. Child abuse and neglect and its psycho-physical and social consequences: a review of the literature. *Curr Pediatr Rev.* (2016) 12:301–10. doi: 10.2174/1573396312666160914193357
8. Bellis MA, Hughes K, Leckenby N, Hardcastle KA, Perkins C, Lowey H. Measuring mortality and the burden of adult disease associated with adverse childhood experiences in England: a national survey. *J Public Health.* (2015) 37:445–54. doi: 10.1093/pubmed/fdu065
9. Metzler M, Merrick MT, Klevens J, Ports KA, Ford DC. Adverse childhood experiences and life opportunities: shifting the narrative. *Child Youth Serv Rev.* (2017) 72:141-9. doi: 10.1016/j.childyouth.2016.10.021
10. Gilbert, R, Widom, CS, Browne, K, Fergusson D, Webb E, Janson S. Burden and consequences of child maltreatment in high-income countries. *Lancet.* (2009) 373:68–81. doi: 10.1016/S0140-6736(08)61706-7
11. Hardcastle K, Bellis MA, Ford K, Hughes K, Garner J, Ramos Rodriguez G. Measuring the relationships between adverse childhood experiences and educational and employment success in England and Wales: findings from a retrospective study. *Public Health.* (2018) 165:106–16. doi: 10.1016/j.puhe.2018.09.014
12. Kinge JM, Modalsli JH, Øverland S, Gjessing HK, Tollånes MC, Knudsen AK, et al. Association of household income with life expectancy and cause-specific mortality in Norway, 2005-2015. *JAMA.* (2019) 321:1916–25. doi: 10.1001/jama.2019.4329
13. Mackenbach JP, Kunst AE, Cavelaars AE, Groenhof F, Geurts JJ. Socioeconomic inequalities in morbidity and mortality in western Europe. The EU Working Group on Socioeconomic Inequalities in Health. *Lancet.* (1997) 349:1655–9. doi: 10.1016/S0140-6736(96)07226-1
14. Agdertall AR. *(County Statistics). Aust- og Vest Agder: Aust-Agder og Vest-Agder Fylkeskommune (Aust-Agder and Vest-Agder County).* Kristiansand (2018).
15. Saunders, BE, Adams, ZW. Epidemiology of traumatic experiences in childhood. *Child Adolesc Psychiatr Clin N Am.* (2014) 23:167–84. doi: 10.1016/j.chc.2013.12.003
16. Vederhus, J-K, Timko, C, Haugland, SH. Adverse childhood experiences and impact on quality of life in adulthood: development and validation of a short difficult childhood questionnaire in a large population-based health survey. *Qual Life Res.* (2021) 30:1769–78. doi: 10.1007/s11136-021-02761-0
17. Brown DW, Anda RF, Tiemeier H, Felitti VJ, Edwards VJ, Croft JB, et al. Adverse childhood experiences and the risk of premature mortality. *Am J Prev Med.* (2009) 37:389–96. doi: 10.1016/j.amepre.2009.06.021
18. Tomasdottir MO, Sigurdsson JA, Petursson H, Kirkengen AL, Krokstad S, McEwen B, et al. Self reported childhood difficulties, adult multimorbidity and allostatic load. A cross-sectional analysis of the Norwegian HUNT study. *PLoS ONE.* (2015) 10:e0130591. doi: 10.1371/journal.pone.0130591
19. Roth K, Harkins D, Eng L. Parental conflict during divorce as an indicator of adjustment and future relationships: a retrospective sibling study. *J Divorce Remarriage.* (2014) 55:117–38. doi: 10.1080/10502556.2013.871951
20. Auersperg F, Vlasak T, Ponocny I, Barth A. Long-term effects of parental divorce on mental health - a meta-analysis. *J Psychiatr Res.* (2019) 119:107–15. doi: 10.1016/j.jpsychires.2019.09.011
21. Walsh D, McCartney G, Smith M, Armour G. Relationship between childhood socioeconomic position and adverse childhood experiences (ACEs): a systematic review. *J Epidemiol Community Health.* (2019) 73:1087–93. doi: 10.1136/jech-2019-212738
22. Houtepen LC, Heron J, Suderman MJ, Fraser A, Chittleborough CR, Howe LD. Associations of adverse childhood experiences with educational attainment and adolescent health and the role of family and socioeconomic factors: a prospective cohort study in the UK. *PLoS Med.* (2020) 17:e1003031 doi: 10.1371/journal.pmed.1003031
23. Hardt, J, Rutter, M. Validity of adult retrospective reports of adverse childhood experiences: review of the evidence. *J Child Psychol Psychiatry.* (2004) 45:260–73. doi: 10.1111/j.1469-7610.2004.00218.x
24. Hardt J, Vellaisamy P, Schoon I. Sequelae of prospective versus retrospective reports of adverse childhood experiences. *Psychol Rep.* (2010) 107:425–40. doi: 10.2466/02.04.09.10.16.21.PR0.107.5.425-440
25. Shonkoff JP, Slopen N, Williams DR. Early childhood adversity, toxic stress, and the impacts of racism on the foundations of health. *Ann Rev Public Health.* (2021) 42:115–34. doi: 10.1146/annurev-publhealth-090419-101940

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Public Health

ORIGINAL RESEARCH
published: 16 August 2021
doi: 10.3389/fpubh.2021.693830

A Bayesian Model to Analyze the Association of Rheumatoid Arthritis With Risk Factors and Their Interactions

Leon Lufkin[1], *Marko Budišić*[2], *Sumona Mondal*[2] *and Shantanu Sur*[3]*

[1] The Clarkson School, Clarkson University, Potsdam, NY, United States, [2] Department of Mathematics, Clarkson University, Potsdam, NY, United States, [3] Department of Biology, Clarkson University, Potsdam, NY, United States

Rheumatoid arthritis (RA) is a chronic autoimmune disorder that commonly manifests as destructive joint inflammation but also affects multiple other organ systems. The pathogenesis of RA is complex where a variety of factors including comorbidities, demographic, and socioeconomic variables are known to associate with RA and influence the progress of the disease. In this work, we used a Bayesian logistic regression model to quantitatively assess how these factors influence the risk of RA, individually and through their interactions. Using cross-sectional data from the National Health and Nutrition Examination Survey (NHANES), a set of 11 well-known RA risk factors such as age, gender, ethnicity, body mass index (BMI), and depression were selected to predict RA. We considered up to third-order interactions between the risk factors and implemented factor analysis of mixed data (FAMD) to account for both the continuous and categorical natures of these variables. The model was further optimized over the area under the receiver operating characteristic curve (AUC) using a genetic algorithm (GA) with the optimal predictive model having a smoothed AUC of 0.826 (95% CI: 0.801–0.850) on a validation dataset and 0.805 (95% CI: 0.781–0.829) on a holdout test dataset. Apart from corroborating the influence of individual risk factors on RA, our model identified a strong association of RA with multiple second- and third-order interactions, many of which involve age or BMI as one of the factors. This observation suggests a potential role of risk-factor interactions in RA disease mechanism. Furthermore, our findings on the contribution of RA risk factors and their interactions to disease prediction could be useful in developing strategies for early diagnosis of RA.

Keywords: rheumatoid arthritis, comorbidities, interactions, prediction, Bayesian, NHANES, genetic algorithm, factor analysis of mixed data

OPEN ACCESS

Edited by:
MinJae Lee,
University of Texas Southwestern
Medical Center, United States

Reviewed by:
David Hollar,
Mercer University School of Medicine,
United States
Rahim Alhamzawi,
University of Al-Qadisiyah, Iraq

****Correspondence:***
Shantanu Sur
ssur@clarkson.edu

Specialty section:
This article was submitted to
Life-Course Epidemiology and Social
Inequalities in Health,
a section of the journal
Frontiers in Public Health

Received: *12 April 2021*
Accepted: *19 July 2021*
Published: *16 August 2021*

Citation:
Lufkin L, Budišić M, Mondal S and
Sur S (2021) A Bayesian Model to
Analyze the Association of
Rheumatoid Arthritis With Risk Factors
and Their Interactions.
Front. Public Health 9:693830.
doi: 10.3389/fpubh.2021.693830

1. INTRODUCTION

Rheumatoid arthritis (RA) is a systemic autoimmune disorder of the joints and internal organs that affects 0.5–1.0% of the adult population worldwide (1, 2). It is a major cause of disability and is associated with an increased risk of premature death (3). The chronic and progressive nature of RA poses a significant financial burden, with the annual societal cost of RA estimated to be $19.3 billion in the United States alone (4). Despite its profound impact on society and the healthcare system, many aspects of this complex, multifactorial disease remain unknown. A variety of genetic, environmental, and behavioral risk factors have been identified for RA and its association with a number of comorbidities has been reported (5). Since current medicine does not offer a cure for RA, the major therapeutic goal is preventing flare-ups, inducing fast remission, and slowing down progressive changes such as irreversible joint deformity (6). Despite RA's demand for close and specialized medical supervision, the number of rheumatologists across the United States has been steadily decreasing. There were roughly 5,000 practicing rheumatologists in 2015, but this number is projected to decrease to 3,500 by the year 2025 (7). One promising approach to address this increasing disparity in the patient-to-rheumatologist ratio is the development of analytical tools to facilitate early diagnosis and predict disease progression, thus enabling better access to care and improving the plan for managing the disease.

RA has a strong connection to age and sex. Disease onset is most likely between 50 and 75 years of age (5, 8) and females are affected 2–3 times more than males (5). Race and ethnicity are also known to influence RA; for example, a lower rate of remission and increased disease activity are reported in African-Americans relative to whites (9). While the reason for such differences is not completely understood, the presence of a "shared epitope"(SE) that is highly correlated with RA severity and outcome is suggested to underlie the higher incidence of the disease in certain sub-populations (10, 11). Apart from demographic factors, several genetic, environmental, behavioral, and socioeconomic risk factors are identified for RA (5, 12). Increased RA incidence in the presence of a family history with 66% heritability observed among twins suggests a genetic link of RA (13). SE alleles within the major histocompatibility complex are shown to have the strongest association with RA, accounting for up to 40% of total genetic risk (12, 13). Environmental factors that can increase the risk of RA include certain infections such as *Porphyromonas gingivalis* bacteria and Epstein-Barr virus (EBV), where an inappropriate immune response to these microbial agents could trigger autoimmunity (14, 15). Additionally, air pollution and occupational exposure to silica have been reported to increase the risk of RA (16, 17). Multiple studies show a strong association of RA with history of tobacco smoking, and the risk of RA increases with the intensity of smoking (18–20). Lower socioeconomic status and less education pose a higher risk of developing the disease (21) as well as experiencing a poorer prognosis (22).

Comorbidities are widespread with RA and often contribute to worse health outcomes (23, 24). Consistent with the complex, systemic nature of RA, these comorbidities often also affect many systems in the body. Among them are widely prevalent chronic conditions such as cardiovascular disease (CVD) and diabetes, which increase the risk of mortality in RA patients (25, 26). Likewise, hypertension and depression increase the risk of disability (26). Gout, another disease of joints, has been found to have a higher association with RA (27). Additionally, RA interferes with the antinociceptive pathway, resulting in enhanced pain perception and leading to a greater risk of sleep problems (28, 29). Several of RA's comorbidities, such as obesity and depression, demonstrate a bidirectional association with RA, implying their presence elevates the risk of developing RA (30, 31). It is of great clinical interest for physicians and researchers to study the concurrent presence of high Body Mass Index (BMI), depression, and CVD in RA patients as it poses a unique clinical repertoire and has significant consequences on affected individuals. Therefore, careful consideration of comorbidities is important for clinicians working in rheumatology care.

Studies have aimed to predict the occurrence of common diseases like CVD to provide early diagnosis or risk assessment using data mining, machine learning algorithms, and mathematical modeling (32). While some studies have attempted to predict RA using a similar approach (33, 34), these studies were neither very selective in defining relevant factors for disease prediction nor did consider their interactions. Karlson et. al. (35) developed prediction models for RA from a combination of clinical and genetic predictors. The models considered age, sex, and smoking as clinical risk factors and studied eight human leukocyte antigen (HLA) and 14 single nucleotide polymorphism (SNP) alleles associated with seropositive RA as genetic risk factors. Models considering either clinical risk factors alone or both clinical and genetic risk factors were compared for discrimination ability using the receiver operating characteristic (ROC) curve. The models with clinical risk factors alone had areas under the ROC curve (AUC) of 0.566-0.626, while models considering both clinical and genetic risk factors had AUC of 0.660-0.752, indicating an improvement of discrimination ability following the inclusion of genetic risk factors. Chibnik et. al. (36) developed a weighted Genetic Risk Score (GRS) from 39 alleles associated with an increased risk of RA. After controlling for age and smoking, the authors used the Genetic Risk Score in a logistic regression to discriminate between non-RA and four phenotypes of RA in the NHS dataset. Their model predicted seronegative, seropositive, erosive and seropositive, and erosive RA with AUCs of 0.563, 0.654, 0.644, and 0.712, respectively. Several other studies (37–40) have performed similar predictive analyses using a combination of environmental and genetic risk factors to create models with good discrimination abilities. The best predictive model we are aware of (as measured by AUC) was developed by Scott, et. al. (41). In this study, the authors considered age, sex, and 25 human leukocyte antigens and 31 single nucleotide polymorphism alleles to develop a model with an AUC of 0.857 (95% CI: 0.804–0.910), indicating high discrimination ability.

While previous studies have demonstrated the feasibility of predicting RA from environmental and genetic information, patient genetic data are not readily available in a regular healthcare set-up, thus limiting their practical applicability. In this work, we aimed to develop a predictive model of RA using information commonly available in peripheral health centers or rural infrastructures, such as comorbidities, demographic, socioeconomic, and behavioral factors that are known to

Abbreviations: FAMD, factor analysis of mixed data; HDI, highest density interval; GA, genetic algorithm; SEC, socieconomic condition; IPR, income to poverty ratio; MA, mexican-american; OH, other hispanic; ONH, other non-hispanic; BP, systolic blood pressure; PHQ, patient health questionnaire.

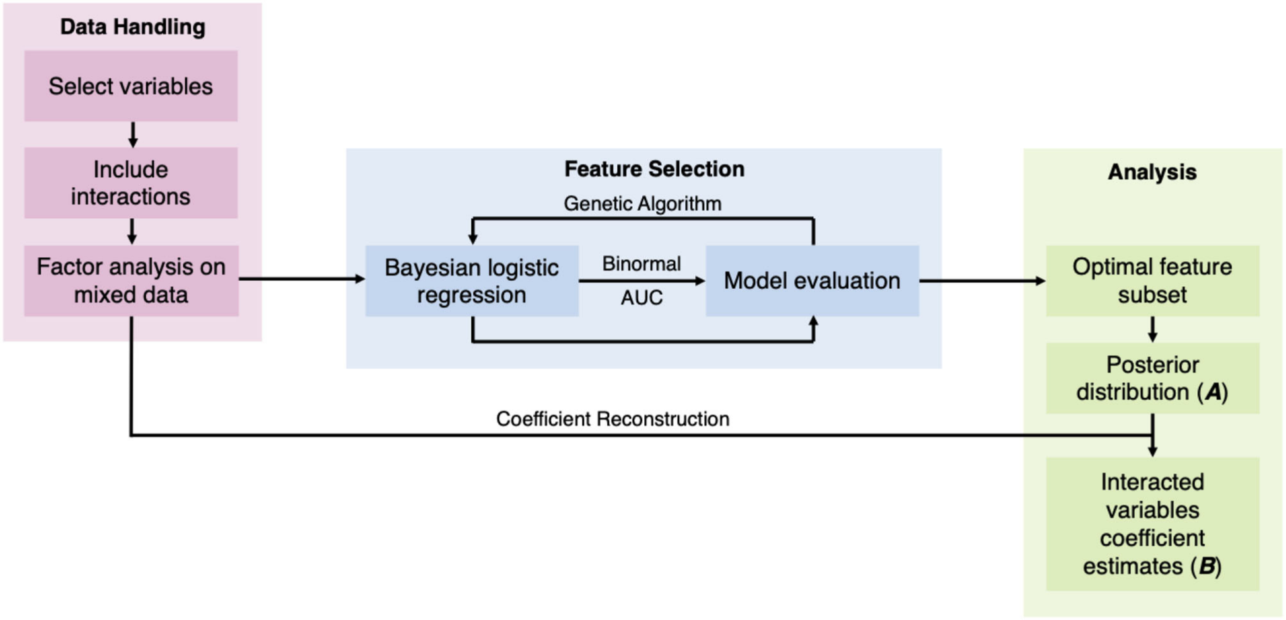

Data Handling
Select variables
Include interactions
Factor analysis on mixed data
Feature Selection
Genetic Algorithm
Bayesian logistic regression
Binormal
AUC
Model evaluation
Analysis
Optimal feature subset
Posterior distribution (*A*)
Coefficient Reconstruction
Interacted variables coefficient estimates (*B*)

FIGURE 1 | Study methods diagram.

associate with RA. We used Bayesian logistic regression to build our model and considered up to third-order interaction between the variables (**Figure 1**). Furthermore, to reduce the computational need without compromising predictive accuracy, we implemented FAMD and wrapper methods, which allowed the selection of the most important variables for the model.

2. MATERIALS AND METHODS

2.1. Description of Data and Preprocessing

Subjects in this study were participants in the National Health and Nutrition Examination Survey (NHANES)[1], a biannual survey designed to assess the health of the US population administered by the Centers for Disease Control and Prevention. NHANES offers freely accessible detailed health datasets on a sample drawn from the US that is representative of the national population. These datasets provide information on demographic variables, socioeconomic condition, survey questionnaires, and bio-specimen examinations. Participants are deidentified and represented by a unique sequence number in each dataset.

NHANES data cohorts from 2007 to 2016 were used in this study, providing an initial dataset with 48,484 participants (**Figure 2**). The survey protocol and data collection methods for the data were approved by the National Center for Health Statistics Research Ethics Review Board (protocol #2005-06 and protocol #2011-17). The Institutional Review Board (IRB) at the researchers' institution does not require an IRB approval or an exemption for the analysis of de-identified and publicly available NHANES data. NHANES uses a multistage, probabilistic sampling design to select participants and provides sample weights for variables to obtain a more accurate estimate of the nationally representative population. While the implementation of sample weights for complex survey data is straightforward for classical analysis, this is a challenging problem for a Bayesian model and is still an active area of research (42, 43). In our preliminary analysis with NHANES RA sample data, we did not find any substantial changes in the distribution of variables after sample weight adjustment and therefore we used the data in our model without further accounting for the sample weights.

Information on demographics, medical conditions, depression, body measures, blood pressure, diabetes, smoking habits, and sleep were obtained from each release cycle, giving a total of 11 variables. Data for gender, age, ethnicity, and socioeconomic condition were obtained from the demographics datasets. Socioeconomic condition was measured using the ratio of a participant's family's income to their poverty threshold (IPR). Participants 17 years old or younger were excluded from the analysis to prevent confounding effects from juvenile RA. Participants were divided into five categories according to their reported ethnicity: Mexican-American (MA), other Hispanic (OH), white, black, and other non-Hispanic (ONH). The ethnicity variable was coded into four new dummy variables using the white ethnicity as the reference category because it contained the largest number of participants. Self-reported diagnoses of RA and gout were obtained from the medical questionnaire dataset. Depression was measured using the nine-question Patient Health Questionnaire (PHQ) (44). Scores on each of the nine questions were manually summed to create a quasi-continuous variable for measuring depression. BMI for each participant was obtained from the body measures dataset as a continuous measurement of obesity. Systolic blood pressure

[1] https://www.cdc.gov/nchs/nhanes/

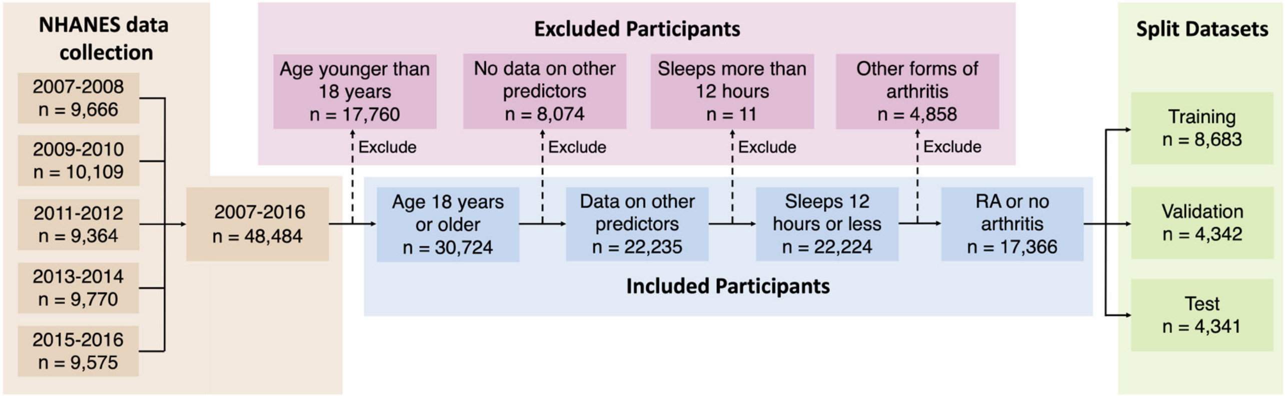


FIGURE 2 | Selection of study population.

(BP) was calculated from the average of four readings in the blood pressure dataset. Self-reported diagnosis of diabetes were used in this analysis. Borderline diabetes was not considered as diabetes. Participants were included in the smoking category if they indicated smoking of at least 100 cigarettes in their life on the smoking questionnaire. Nightly hours of sleep were recorded in 1-h increments with a maximum of 12 to accommodate for variations in NHANES data collection between 2007–2014 and 2015–2016.

Participants who responded "don't know," refused to respond, or had missing data for any variable were excluded from this study, retaining 17,366 participants who fulfilled the selection criteria (**Figure 2**). We created second- and third-order interactions between the independent variables by multiplying the initial variables together (except for sequence number and RA). Interactions created by squaring binary variables and multiplying mutually exclusive binary variables were removed from the dataset. New variables that represent an interaction between two or three initial variables are termed "interacted" variables. Quantitative variables were centered and scaled to have means of zero and standard deviations of one. This dataset was further divided into training, validation, and test datasets by randomly distributing to a 50–25–25% split for use in model building and validation.

2.2. Factor Analysis of Mixed Data

The added interacted variables are highly-correlated, posing a problem for regression analysis. Using factor analysis of mixed data (FAMD) (45) new uncorrelated synthetic variables were created, and data projected onto them. FAMD effectively performs principal component analysis (PCA) on quantitative variables and multiple correspondence analysis on qualitative variables. PCA takes in observations of correlated variables and constructs a change of coordinates such that the synthetic output variables are decorrelated. Similarly, multiple correspondence analysis takes in observations of nominal categorical variables and returns a set of decorrelated synthetic variables that represent the underlying structures in the original data. In both cases, the physical interpretability of the created variables is sacrificed to obtain favorable statistical properties, allowing efficient representation of data by a small set of uncorrelated variables.

In FAMD, a new synthetic variable v is created by maximizing the criterion

$$\sum_{k \in K_1} r^2(k, v) + \sum_{q \in K_2} \eta^2(q, v), \tag{1}$$

where K_1 are qualitative variables, K_2 are continuous variables, r^2 is Pearson's correlation statistic, and η^2 is the effect size measure from analysis of variance model (46). A complete disjunctive coding was performed on all qualitative variables. This created a pair of indicator variables corresponding to each state of every categorical variable in the dataset, all of which were already boolean variables. This process creates K_2 indicator variables that are only used in FAMD. The original categorical variables were kept as supplementary variables in the dataset (variables that are not used for calculating the synthetic variables but are projected onto them for interpretation), while all remaining quantitative and indicator variables are active variables (used for calculating the synthetic variables).

A decorrelated set of synthetic variables maximizing (Equation 1) can be computed using the singular value decomposition (SVD) of the data matrix $\boldsymbol{M}$, whose columns correspond variables and rows to observations, that is to the values of those variables for participants. SVD was performed on all active variables, amounting to calculating matrices $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, used to project the data onto orthogonal axes (synthetic variables). $\boldsymbol{U}$ is an orthogonal matrix used to calculate the projections of the participants onto the synthetic variables. $\boldsymbol{V}$ was used to find the projections of the active variables on the new synthetic variables. The projections of categorical variables onto the synthetic variables are determined from their indicator variables. $\boldsymbol{\Sigma}$ is a diagonal matrix containing the singular values, which are in turn square-roots of variance they explain in the dataset so that $\boldsymbol{\Sigma}^2$ is the (diagonal) covariance matrix of the synthetic (decorrelated) variables. Synthetic variables corresponding to variances less than one were omitted to maintain low intercorrelation after the validation and test datasets were projected onto them. Due to the properties of SVD, discarding low-variance synthetic variables is known to be the optimal approach, in the sense of Equation (1), to construction of reduced-order representation of the data. FAMD was performed using the package `FactoMineR` (47) in R 3.6.0.

2.3. Statistical Analysis

Bayesian logistic regression was used to predict RA in this study (48). A Bayesian approach was preferred over standard logistic regression because the former provides full posterior information as opposed to point-estimates by the later, and also allows one to incorporate prior information. The model being linear has also an advantage over the common supervised learning algorithms such as random forest by allowing an easier interpretation of predictor effects, important for this study.

Bayesian regression summarizes model coefficients and predictions with probability distributions. The results are frequently reported using the highest density interval (HDI), which is the smallest interval corresponding to a certain probability of the posterior distribution. Here 50% and 99% HDIs were included in the interval plots of the posterior distributions. Variables are ranked in the interval plots based on the posterior probabilities that their coefficients are greater or less than one (when transformed from log-odds to odds scale). If a coefficient's median is greater than one, the probability that it is greater than one is calculated, $\Pr(\beta > 1 \mid y)$. However, if a coefficient's median is less than one, the probability that it is less than one is used, $\Pr(\beta < 1 \mid y)$. Ties between coefficients with equal probabilities of being greater or less than one are broken using the absolute values of the medians of their posterior distributions. A Bayesian approach also allows us to specify prior information about model coefficients using a probability distribution (the prior distribution).

The posterior distribution for Bayesian logistic regression up to a normalizing factor is given by

$$p(\vec{\beta}|y_1, \ldots, y_N, \boldsymbol{X}) \propto p(\vec{\beta}) \prod_{i=1}^{N} p(y_i|\vec{\beta}, \boldsymbol{X}), \tag{2}$$

where $p(\vec{\beta})$ is the prior and $p(y_i|\vec{\beta}, \boldsymbol{X})$ the likelihood for each data point. The model uses a total of K predictors, combined using coefficients $\vec{\beta} = (\beta_k)_{k=1}^{K}$, and the added intercept term β_0. The data set $\boldsymbol{X}$ contains data points $\boldsymbol{X}_{i,k}$, where i indexes up to a total of N participants and k the predictors. Binary variables y_i indicate whether the i-th participant has RA (if so, $y_i = 1$, otherwise $y_i = 0$). Because we are performing Bayesian *logistic* regression, the distribution $p(y_i|\vec{\beta}, \boldsymbol{X})$ is the Bernoulli distribution

$$p(y_i|\vec{\beta}, \boldsymbol{X}) = \begin{cases} p & y_i = 1 \\ 1 - p & y_i = 0 \end{cases}, \tag{3}$$

where

$$p = F\left(\beta_0 + \sum_{k=1}^{K} \beta_k X_{i,k}\right) \tag{4}$$

is calculated using the standard logistic function $F(x) := [1 + \exp(-x)]^{-1}$. Each of the coefficients $\beta_0, \beta_1, \ldots$ is assigned a uniform prior, weighing all possible values equally. Although uniform densities supported on the entire real line are improper, i.e., they cannot have densities that integrate to one, such a choice of the prior still leads to a valid posterior and is standard in Bayesian analysis.

We implemented Bayesian logistic regression using `Stan` in `R 3.6.0` through the package `RStan` (49), which uses Hamiltonian Monte Carlo to sample the posterior distribution described by Equation (2). Markov chains were required to have potential scale reduction factors below 1.1 to indicate approximate convergence, imposing a stringent convergence requirement (50).

2.4. Predictive Performance and Feature Selection

A wrapper approach to feature selection was implemented in this study to identify the optimal subset of synthetic variables to predict RA. Feature selection is necessary to identify the most relevant predictors from a larger set, and such operation also improves the precision of estimated effects of the selected predictors. A wrapper approach (as opposed to a filter or embedded approach) uses the predictive performance of subsets of synthetic variables to identify the optimal subset. The predictive performance of the regression models in the genetic algorithm (GA, described below) was determined using the area under the receiver operating characteristic curve. Binormal smoothing of the ROC curve is implemented for its robustness in obtaining an unbiased estimate of the model's true discrimination ability (51). This assumes that the distributions of the predicted probabilities of response for the positive and negative cases can be described by a pair of normal distributions, y_1 and y_0, respectively:

$$y_1 \sim N(\mu_1, \sigma_1^2), \; y_0 \sim N(\mu_0, \sigma_0^2).$$

In this study, the binormally smoothed AUC is calculated using two parameters:

$$a = \frac{\mu_1 - \mu_0}{\sigma_1} \quad \text{and} \quad b = \frac{\sigma_0}{\sigma_1}.$$

The AUC is calculated as

$$\text{AUC} = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right), \tag{5}$$

where Φ is the standard normal cumulative distribution function. Estimates for a and b are obtained by linear regression to the equation

$$\Phi^{-1}(\text{TPR}) = a + b\Phi^{-1}(\text{FPR}), \tag{6}$$

where TPR and FPR represent the true positive and false positive rates across all thresholds of classification.

A radial sweep is used to generate confidence bands for the ROC curve to provide optimal coverage (52). Equation (6) is transformed to polar coordinates with center (FPR = 1, TPR = 0) in ROC space. r is calculated for values of θ in increments of

0.01 from zero to $\pi/2$. Confidence intervals (CIs) for the AUC and for values of r are found from 10,000 bootstrapped samples of the predicted probabilities used to generate the ROC curve.

We used a GA in this study to implement a wrapper approach to feature selection. The GA performs an optimization to find the best subset of synthetic variables for predictive performance according to the AUC Equation (5). The GA was parameterized to have a population size of 500 and run for 200 generations. The GA was seeded with variable subsets always containing the first seven synthetic variables and randomly containing the remaining 45 synthetic variables. All computation for the GA was performed using a server from the Clarkson Open Source Institute at Clarkson University with two Intel Xeon E5-2650 processors with 192 gigabytes of usable physical memory. Running the GA on this server took approximately 2 weeks.

Rank selection was used to determine which variable subsets would be selected for genetic transformation to create the next population. The probability that a subset x will be selected is given by

$$p(x) = \frac{1}{n}\left(\min + (\max - \min)\frac{\text{rank}(x)}{n-1}\right), \tag{7}$$

where n is the size of the population. min represents the expected number of times the subset with the poorest predictive ability is selected, while max represents the same for the subset with the best predictive ability, with the constraint that $\min + \max = 2$ is imposed (53). rank(x) gives the rank of the variable subset within the population such that the best subset has rank n. In this study, we set $\min = 0.7$ and $\max = 1.3$ to allow for substantial generational improvement while maintaining sufficient exploration of the search space.

Each variable subset had a probability of 0.8 to be selected for single-point crossover, which was used for its simplicity and performance in GAs (54). Each subset was also subject to a 0.1 probability of being randomly mutated. Elitism was implemented using 5% of the population to maintain high-quality solutions throughout the GA's search. The optimal subset was tested on a holdout set of data to assess for overfitting.

2.5. Coefficient Reconstruction

HDIs for the coefficients of the original variables are obtained from the posterior distributions of the optimal subset of synthetic variables. The optimal feature subset of size n was fit to the data using eight Markov chains each with 400 samples of the posterior distribution, creating a 3200-row by n-column matrix $\boldsymbol{A}$ of the probability distributions of the coefficients for synthetic variables in the logistic regression model. Columns corresponding to synthetic variables that were omitted were set to zero in $\boldsymbol{A}$. Probability distributions for the coefficients of the interacted variables $\boldsymbol{B}$ are calculated from $\boldsymbol{V}$ according to the equation below.

$$\boldsymbol{B} = \boldsymbol{A}\boldsymbol{V}^T \tag{8}$$

Estimates for the binary variables in the interacted dataset were obtained from the difference between the estimates for their indicator variables.

3. RESULTS

3.1. Variable Selection

Selection of risk factor variables to incorporate in our model for RA prediction was guided by their reported association with RA and data availability in the NHANES database. Although a large number of risk factors are reported to be associated with RA, in the present study we selected only a few well-known factors to better understand the contribution of their individual and interaction effects. These variables include disease comorbidities (diabetes, depression, high BMI, hypertension, and gout), demographic factors (gender and ethnicity), socioeconomic factors (IPR), and behavioral factors (smoking and sleep hours) (**Figure 3**). Other RA risk factors such as asthma or EBV infection were not included in the present analysis even though data for these variables are available in the NHANES database (15, 55). Consistent with the literature, the NHANES dataset demonstrated an association of these risk factors with RA (see **Figure 3** and **Table 1**), although the extent of the difference varied. For example, RA was found to be less common among males (41.4% of RA subjects) but the gender disparity was substantially smaller than reported by previous studies (**Figure 3A**) (5). This difference could be attributed to the survey-based diagnosis of RA, the data preprocessing procedure, and the inherent design of NHANES (see **Supplementary Table 1**). Subjects with RA were also more likely to suffer from diabetes, gout, high BMI, depression (measured by PHQ score), and high BP (**Figures 3A,C**). Risk of RA was found to increase with age, and it was more common among black ethnicity (56) but substantially less prevalent among the ONH population (**Figures 3B,C**). Behavioral factors such as smoking were observed more among RA subjects, while sleep has a less conspicuous impact even though it was reported previously (29). Interestingly, subjects with RA were found have a lower IPR, suggesting an association of RA with lower economic status.

A total of 11 risk factors were considered in our study, which generated 14 first-order variables including 4 binary variables obtained from dummy coding ethnicity, using the white population as the reference category. For model building and validation, the dataset was further divided into training, validation, and test categories (**Table 2**). The distribution of the variables were found to be nearly equivalent across each category, indicating an even split after data preprocessing. A slightly greater variation among the three datasets was observed for the RA group, which could be attributed to a substantially smaller number of individuals in this group than the control no arthritis group. In order to analyze second- and third-order interactions, we created 475 interaction variables from the 14 first-order variables, leading to a total of 489 variables.

3.2. Predictive Performance

To build our model, we first excluded the highly correlated variables from the total set of variables containing higher-order interactions. Since our data contained both categorical and continuous variables, we implemented FAMD to identify the correlated variables. A total of 52 synthetic variables with

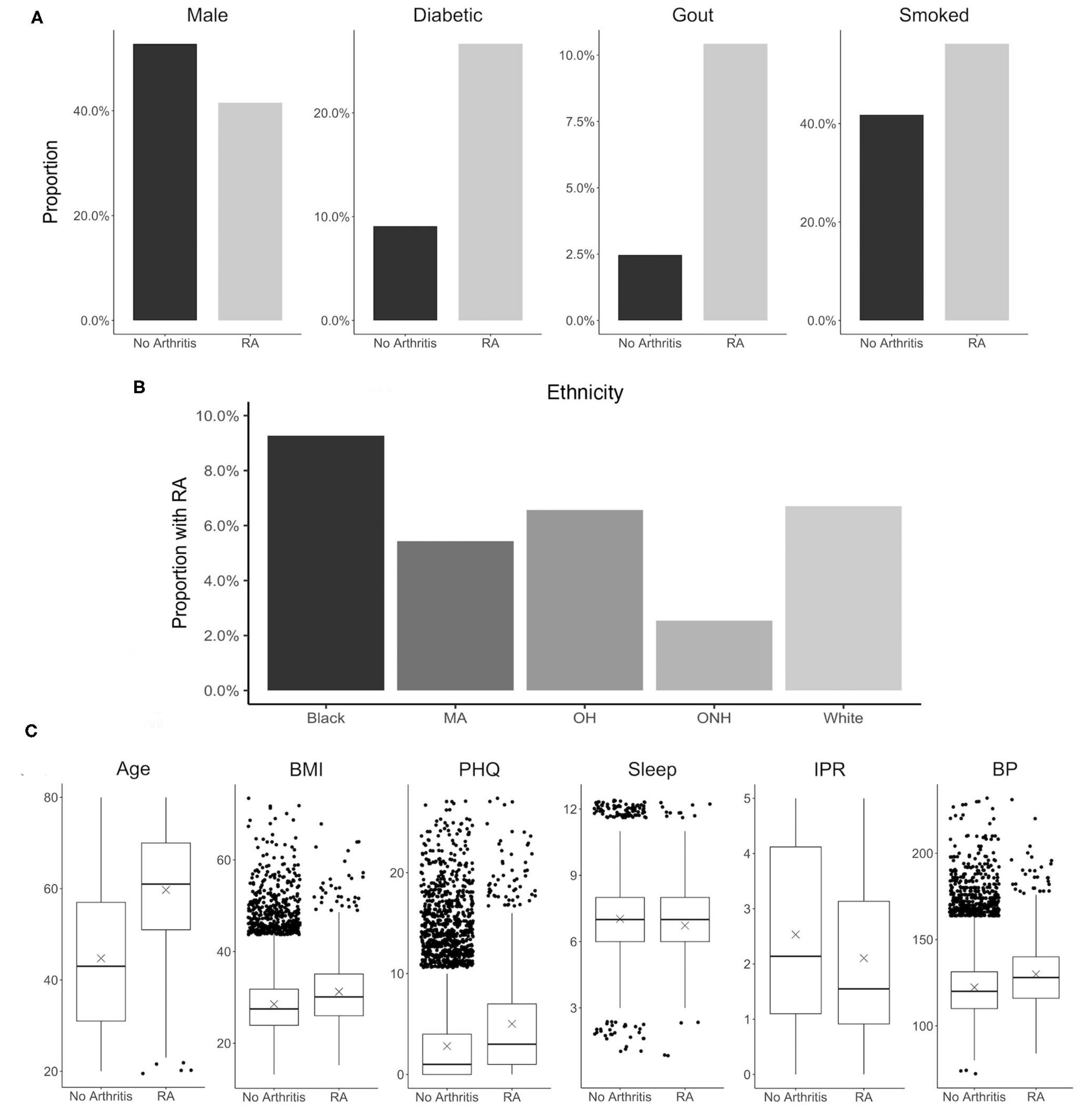


FIGURE 3 | Distribution of various RA risk factors in the study population. **(A)** Comparison between RA and no arthritis population for risk factors coded as binary variables. **(B)** Prevalence of RA among various ethnicity. **(C)** Comparison between RA and no arthritis population for risk factors coded as continuous variables. Units: Age, years; BMI, kg/m^2; PHQ, PHQ score; Sleep, hours; IPR, nondimensional; Systolic, mmHg. MA, Mexican-American; OH, Other Hispanic.

variances greater than one were obtained by FAMD that represented 92.3% of the variation in the training data. **Table 3** summarizes these synthetic variables according to the percentage of variance explained by each of them. A feature selection from these synthetic variables was further performed by a wrapper approach using GA. An optimal subset containing 33 of these synthetic variables was identified that provides the greatest discrimination ability. **Figure 4A** shows the progression of the GA's search to find the subset of synthetic variables that best predicts RA. 33 of the 52 total synthetic variables were selected through this process, which was able to predict RA with a smoothed AUC of 0.826 with 95% CI of 0.801–0.850

TABLE 1 | Summary characteristics of demographics and risk factors for RA and no arthritis (None) group in the study population.

	Total participants ($n = 17,366$)	
Prop.	**RA**	**None**
Male	473 (41.4%)	8,523 (52.5%)
Female	670 (58.6%)	7,700 (47.5%)
Gout	122 (10.7%)	406 (2.50%)
Diabetic	308 (26.9%)	1,480 (9.12%)
Smoked	643 (56.3%)	6,769 (41.7%)
MA	157 (13.7%)	2,652 (16.3%)
OH	120 (10.5%)	1,720 (10.6%)
Black	339 (29.7%)	3,313 (20.4%)
White	476 (41.6%)	6,608 (40.7%)
ONH	51 (4.465%)	1,930 (11.9%)
$\bar{x}$ (s)		
Age	59.8 (13.3)	44.8 (16.8)
BMI	31.3 (7.76)	28.6 (6.62)
PHQ	5.04 (5.44)	2.81 (3.91)
Sleep	6.72 (1.75)	7.02 (1.42)
IPR	2.09 (1.49)	2.53 (1.64)
BP	129 (19.8)	122 (17.5)

Counts and percentages are shown for discrete variables; sample means and standard deviations are shown for continuous variables. Units: Age, years; BMI, kg/m^2; PHQ, PHQ score; Sleep, hours; IPR, unitless; Systolic, mmHg. MA, Mexican-American; OH, Other Hispanic; ONH, Other Non-Hispanic.

(**Figure 4B**). The potential scale reduction factors ($\widehat{R}$) and estimated coefficients from the final regression model for these selected synthetic variables are shown in **Table 3**. For variables omitted through the feature selection process, the medians for posterior distributions of coefficients (β) were set to one and do not have $\widehat{R}$ values (**Table 3**). This subset of variables was also used on the test dataset to obtain a smoothed AUC of 0.805 (95% CI: 0.781–0.829), indicating high accuracy on external data and that the model was not overfitting to the training dataset during regression or the validation dataset during feature selection.

Interestingly, we find that even the first-order variables alone are highly predictive, with an AUC of 0.823, and that higher-order interactions yield only a small improvement of AUC to 0.826. Furthermore, our approach can generate a predictive accuracy higher than most previous works reported even when using a small set of first-order variables (see **Supplementary Table 2**) (35–37). For example, considering age and smoking alone can generate a model with an AUC of 0.748, and including sex further increased the AUC to 0.772. While these findings suggest the potential of model building from first-order variables alone, future studies are required to identify the set of variables that maximizes the predictive accuracy of the model.

3.3. Risk Factor Interactions

The subset of synthetic variables returned by the GA is not easily interpretable on its own. Each synthetic variable represents a latent variable that is a linear combination of the total pool of 489 variables. The posterior distribution of the synthetic variables

TABLE 2 | Breakdown of participants into training, validation, and test datasets with comparison of summary characteristics between RA and no arthritis (None) groups in each dataset.

Prop.	**Training (n = 8,683)**			**Validation (n = 4,342)**			**Test (n = 4,341)**		
	All	**RA**	**None**	**All**	**RA**	**None**	**All**	**RA**	**None**
Male	0.523	0.436	0.529	0.513	0.353	0.524	0.523	0.434	0.529
Gout	0.030	0.103	0.024	0.033	0.112	0.028	0.027	0.100	0.022
Diabetic	0.106	0.270	0.095	0.099	0.283	0.087	0.098	0.244	0.088
Smoked	0.428	0.555	0.419	0.419	0.543	0.410	0.435	0.595	0.424
MA	0.163	0.128	0.166	0.157	0.152	0.157	0.163	0.129	0.166
OH	0.105	0.097	0.106	0.114	0.123	0.114	0.102	0.111	0.101
Black	0.208	0.305	0.201	0.216	0.309	0.210	0.205	0.269	0.200
White	0.410	0.427	0.408	0.402	0.368	0.404	0.412	0.448	0.409
ONH	0.114	0.043	0.119	0.111	0.048	0.115	0.118	0.043	0.124
$\bar{x}$ (s)									
Age	45.6 (16.9)	59.9 (13.3)	44.6 (16.6)	45.9 (17.1)	59.5 (13.4)	45.0 (16.9)	45.8 (17.1)	59.6 (13.2)	44.8 (16.9)
BMI 28.7 (6.65)	31.4 (7.49)	28.5 (6.54)	28.9 (6.81)	31.6 (7.75)	28.7 (6.70)	28.6 (6.59)	30.6 (7.91)	28.4 (6.47)	
PHQ	2.89 (4.03)	4.80 (5.45)	2.76 (3.88)	3.01 (4.16)	5.37 (5.69)	2.84 (3.98)	3.03 (4.02)	5.11 (5.18)	2.88 (3.88)
Sleep	7.01 (1.44)	6.83 (1.68)	7.02 (1.42)	7.02 (1.47)	6.79 (1.76)	7.03 (1.45)	7.01 (1.44)	6.46 (1.81)	7.05 (1.40)
IPR	2.50 (1.64)	2.13 (1.50)	2.53 (1.64)	2.50 (1.63)	2.05 (1.50)	2.54 (1.63)	2.51 (1.65)	2.11 (1.48)	2.54 (1.65)
BP	123 (17.9)	130 (19.9)	122 (17.6)	123 (18.0)	130 (20.6)	122 (17.7)	123 (18.3)	129 (19.7)	122 (18.1)

Proportions are shown for binary variables. Sample means and standard deviations are reported for continuous variables. Units: Age, years; BMI, kg/m^2; PHQ, PHQ score; Sleep, hours; IPR, unitless; Systolic, mmHg. MA, Mexican-American; OH, other Hispanic; ONH, other non-Hispanic.

TABLE 3 | Percentage of variance explained, $\widehat{R}$, and medians for posterior distributions of coefficients for synthetic variables (β) returned by FAMD.

Var	% Exp	$\widehat{R}$	β	Var	% Exp	$\widehat{R}$	β	Var	% Exp	$\widehat{R}$	β	Var	% Exp	$\widehat{R}$	β
1	10.3	1.00	1.0963	14	1.60	1.00	0.9788	27	0.849	—	1	40	0.442	1.00	0.9086
2	7.30	1.00	1.0545	15	1.59	1.00	0.9239	28	0.821	—	1	41	0.433	—	1
3	6.65	1.00	0.9542	16	1.49	1.00	0.8932	29	0.802	1.00	1.1096	42	0.419	—	1
4	6.49	1.00	0.9640	17	1.24	—	1	30	0.740	1.00	1.0633	43	0.387	—	1
5	6.17	1.00	0.9967	18	1.21	1.00	0.9688	31	0.658	1.00	1.0092	44	0.367	—	1
6	5.96	1.00	0.9963	19	1.18	1.00	0.9598	32	0.630	1.00	0.9839	45	0.343	—	1
7	5.19	1.00	0.9836	20	1.15	1.00	1.0240	33	0.590	1.00	0.9812	46	0.336	—	1
8	4.26	1.00	1.0169	21	1.07	—	1	34	0.552	—	1	47	0.304	1.00	0.9828
9	3.31	1.00	0.8923	22	1.04	—	1	35	0.537	1.00	0.9777	48	0.281	—	1
10	2.91	1.00	1.0987	23	1.02	—	1	36	0.493	1.00	1.0614	49	0.277	1.00	0.9798
11	2.61	1.00	0.9484	24	0.931	—	1	37	0.491	—	1	50	0.262	1.00	1.0314
12	1.83	1.00	1.0846	25	0.903	1.00	1.0062	38	0.470	1.00	0.9688	51	0.248	1.00	0.8923
13	1.61	1.00	1.0362	26	0.872	—	1	39	0.466	—	1	52	0.230	—	1

Synthetic variables omitted through feature selection have β set to one and do not have $\widehat{R}$ values.

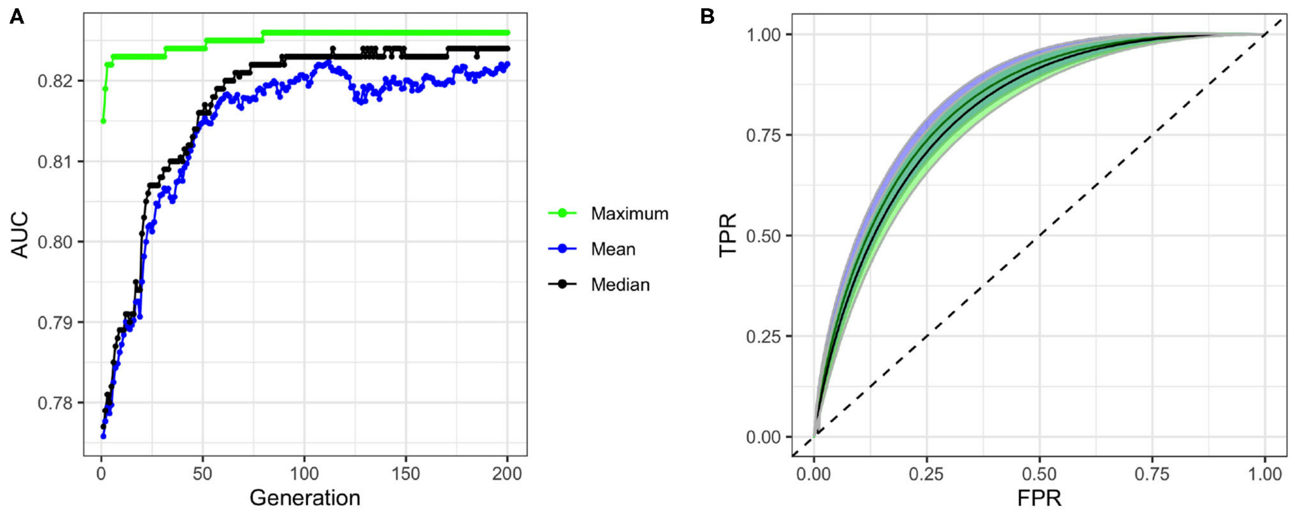


FIGURE 4 | Performance of genetic algorithm (GA) for feature selection. **(A)** Convergence of GA on an optimal subset of synthetic variables with maximum, mean, and median fitness values in each generation of the search. **(B)** ROC curves and confidence bands of optimal model predicting on validation datasets (confidence region shaded blue) and test datasets (confidence region shaded green). Dashed line represents the ROC curve of a model with no predictive ability, corresponding to an AUC of 0.5.

obtained through this process was used to construct HDIs for each of the 489 variables using Equation (8). Furthermore, to allow intuitive comparison across variable types and effect orders, the coefficient estimates were computed for the standardized versions of the variables (**Figure 5**). Thus, the variables with HDIs further away from 1.0 are more significant predictors of RA, while a narrower interval indicates a greater certainty about how a specific variable affects RA. The analysis aims to identify the effects of first-order variables and the influence of any second- and third-order interactions as illustrated in **Figure 5A**.

The prediction of RA in the test dataset by the first-order variables overall aligns well with the association of these variables to RA observed in **Figure 3**. Age, BMI, depression (PHQ score), diabetes, gout, and smoking are found to be positive predictors, while male gender and financial wellness (IPR) reduce the risk of having RA (**Figure 5B**). A clear influence of ethnicity is also observed: Risk of RA is higher among black population and lower among Mexican-American population when compared against white. Interestingly, sleep emerged as a strong negative predictor (the most influential first-order variable after age), even though an association of RA and sleep was not clearly observed in the data. In contrast, systolic BP played no effect on RA prediction as a first-order variable (HDI roughly symmetric about one), although RA subjects had a higher mean systolic BP than the control population. The key first order effects are summarized in **Table 4** and RA probabilities against the amplitude of variables

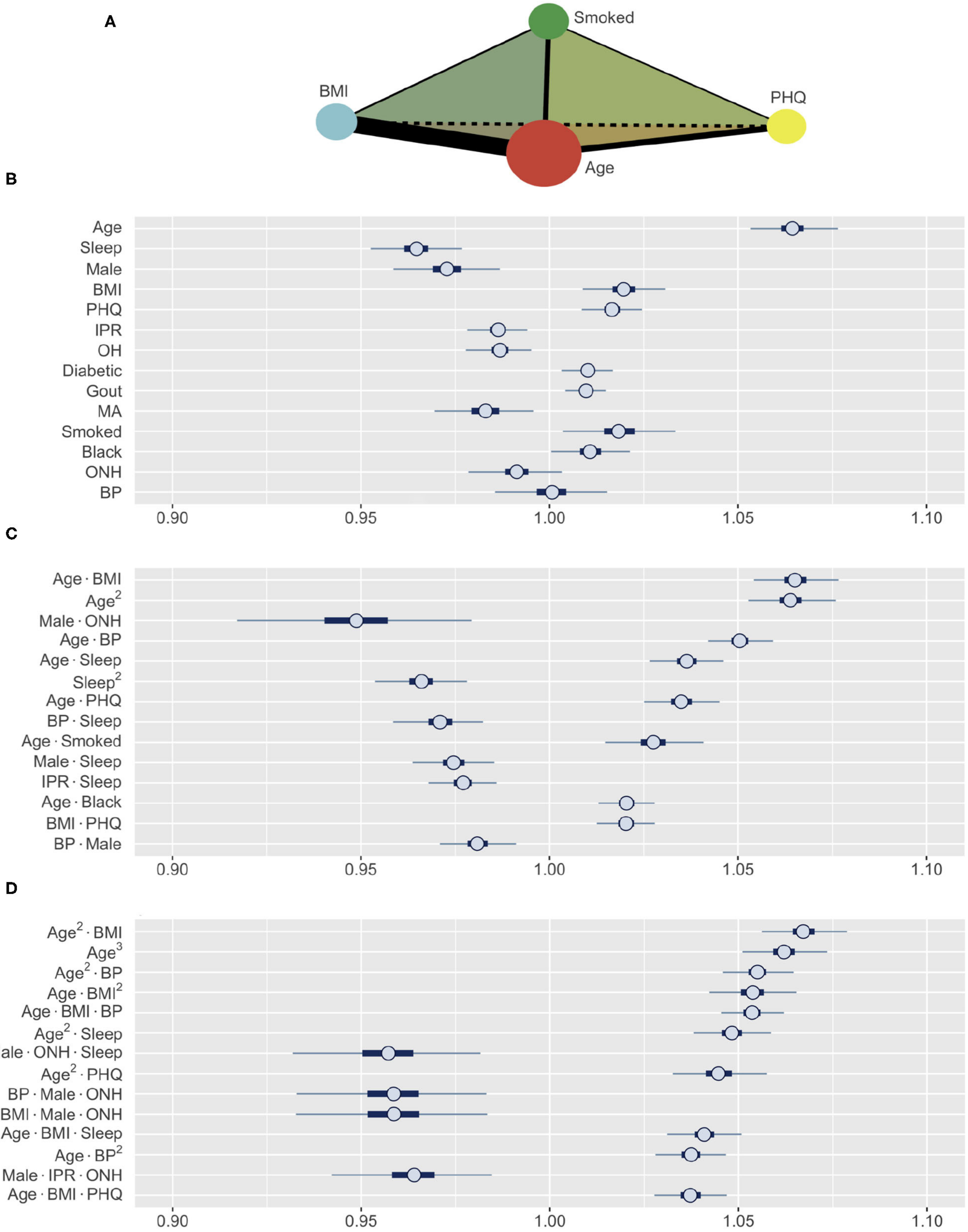


FIGURE 5 | (A) A schematic illustrating the interaction analysis between four risk factors. Size of vertices, thickness of edges, and color of faces denote the size of first-, second-, and third-order effects, respectively. **(B–D)** Posterior distributions of standardized coefficients for selected variables. HDIs for **(B)** all first-order variables, **(C)** most influential second-order, and **(D)** third-order variables (inner bounds, 50% HDIs; outer bounds, 99% HDIs). Horizontal scale represents the odds multipliers for risk of RA for one standard deviation increase in value of the variables. MA, Mexican-American; OH, Other Hispanic; ONH, Other Non-Hispanic.

TABLE 4 | Summary of key findings for (A) first-order, (B) second-order, and (C) third-order variables.

(A)	Risk factor	$\hat{\beta}$	Comments
	Age	1.0533–1.0765	• Aging increases risk for RA • Most influential first-order effect
	Sleep	0.9526–0.9768	• More sleep decreases risk of RA
	BMI	1.0088–1.0308	• Higher BMI increases risk of RA • Weaker first-order effect
	BP	0.9856–1.0153	• No direct effect on risk of RA
	ONH	0.9785–1.0032	• No direct effect on risk of RA

(B)	Risk factor	$\hat{\beta}$	Comments
	Age^2	1.0527–1.0752	• Effect of increased age on RA risk is greater at older ages
	Age·BMI	1.0535–1.0770	• Age and BMI have the strongest second-order interaction
	Male·ONH	0.9171–0.9793	• Being male and ONH ethnicity markedly reduces risk for RA
	Age·BP	1.0421–1.0592	• Interaction of age and BP increases risk of RA
	BP·Sleep	0.9585–0.9824	• Higher BP further lowers RA risk afforded by sleep

(C)	Risk factor	$\hat{\beta}$	Comments
	Age^2·BMI	1.0562–1.0787	• Strong third-order interaction between age and BMI
	Age^3	1.0511–1.0735	• Effect of increased age on RA risk is greater at older ages
	Age^2·BP	1.0459–1.0645	• Third-order interaction between age and high BP increases RA risk
	Age·BMI·BP	1.0454–1.0621	• High BP further increases RA risk from aging and high BMI
	Male·ONH·Sleep	0.9319–0.9816	• Sleep adds to lowering RA risk afforded by being male and ONH

Ninety-nine percent HDIs are shown for estimated regression coefficients on the odds scale.

are shown by marginal effects plots for a few representative variables in **Supplementary Figure 2**.

Apart from the effects of individual first-order variables, we were interested to identify any influence of higher-order interactions in RA prediction. **Figure 5C** enumerates the 14 most influential second-order variables observed in our study. Age turns out to not only be the strongest first-order predictor variable but also to have prominent second-order interactions with several other variables, including BMI, BP, depression, sleep, and smoking. The strongest second-order interaction effect was found between age and BMI (median: 1.0648, 99% HDI: 1.0535–1.0770), which is comparable to the influence of age (1.0642, 1.0533–1.0757) or three times the influence of BMI (1.0196, 1.0088–1.0306), considered individually (**Table 4**). Interestingly, the second-order effect of age (1.0636, 1.0527–1.0752) is similar in magnitude to its first-order effect, suggesting that the effect of age on RA risk increases with age. We also observed several second-order interactions to reduce the risk of RA. For example, the combination of ONH ethnicity with male gender strongly reduces the risk of having RA (0.9485, 0.9156–0.9797), even though ONH does not have a significant influence in lowering RA risk and male gender has a less prominent effect. This finding suggest the second-order interaction with male gender could underlie low RA prevalence observed among ONH ethnicity (**Figure 3B**). Sleep demonstrates an interesting interaction effect on RA. While increased sleep hours was found to lower the risk of RA, its second-order effect with age increased the risk significantly, suggesting an altered role of sleep on the body's immune system with aging.

Our model was also able to reveal the existence of strong third-order interactions. **Figure 5D** lists 14 most prominent third-order interactions where we find the frequent appearance of a few variables, with age and BMI being most common. Other factors involved in strong third-order interactions are gender, ONH ethnicity, sleep, depression, and BP. Similar to the second-order interactions, these third-order interactions are seen to either increase or decrease the risk of RA (**Figure 5D** and **Table 4**). In particular, for interactions posing high risk, we often observe age and BMI, either as a third-order variant of the interaction between these variables, or in combination with a third variable such as BP, sleep, or depression. By contrast, the coexistence of ONH ethnicity with male gender in a third-order interaction prominently reduces the risk of RA when associated with sleep, BMI, BP, or IPR as the third variable. Thus, variables such as sleep or BP, when involved in third-order interactions, can both increase or decrease the risk of RA, suggesting a complex interplay of underlying physiological mechanisms.

3.3.1. Range of Interactions: Age vs. BMI

The finding of several prominent second- and third-order interactions in our model further motivated us to investigate the range of interactions for an individual risk factor. In this direction, we focused on comparing age and BMI, two variables that demonstrated the strongest higher-order interaction (**Figure 6**). Our analysis shows that these two variables have very different interaction profiles. Age demonstrates strong second-order interactions with multiple comorbidities (BMI, BP, and depression), sleep, and smoking, all of which increase the risk of RA (**Figure 6A**). In contrast, second-order interaction effects to BMI are moderate to weak (except with age) and, depending on the interacting variable, increases or decreases the RA risk (**Figure 6B**). The third-order interactions for age and BMI follow a similar pattern as observed in the second-order interactions, except the combination of male and ONH ethnicity reduces RA risk (**Figures 6C,D**). We hypothesize that general changes in body physiology accompanied with aging cause other risk factors to have a greater impact on RA, resulting in these interaction effects. In contrast, high BMI potentially elicits specific influence in the pathophysiology of interacting risk factors, increasing or decreasing the magnitude of the effects. Together, these results confirm that the interactions of a risk factor with other risk factors are highly specific in nature and are dependent on the variables considered.

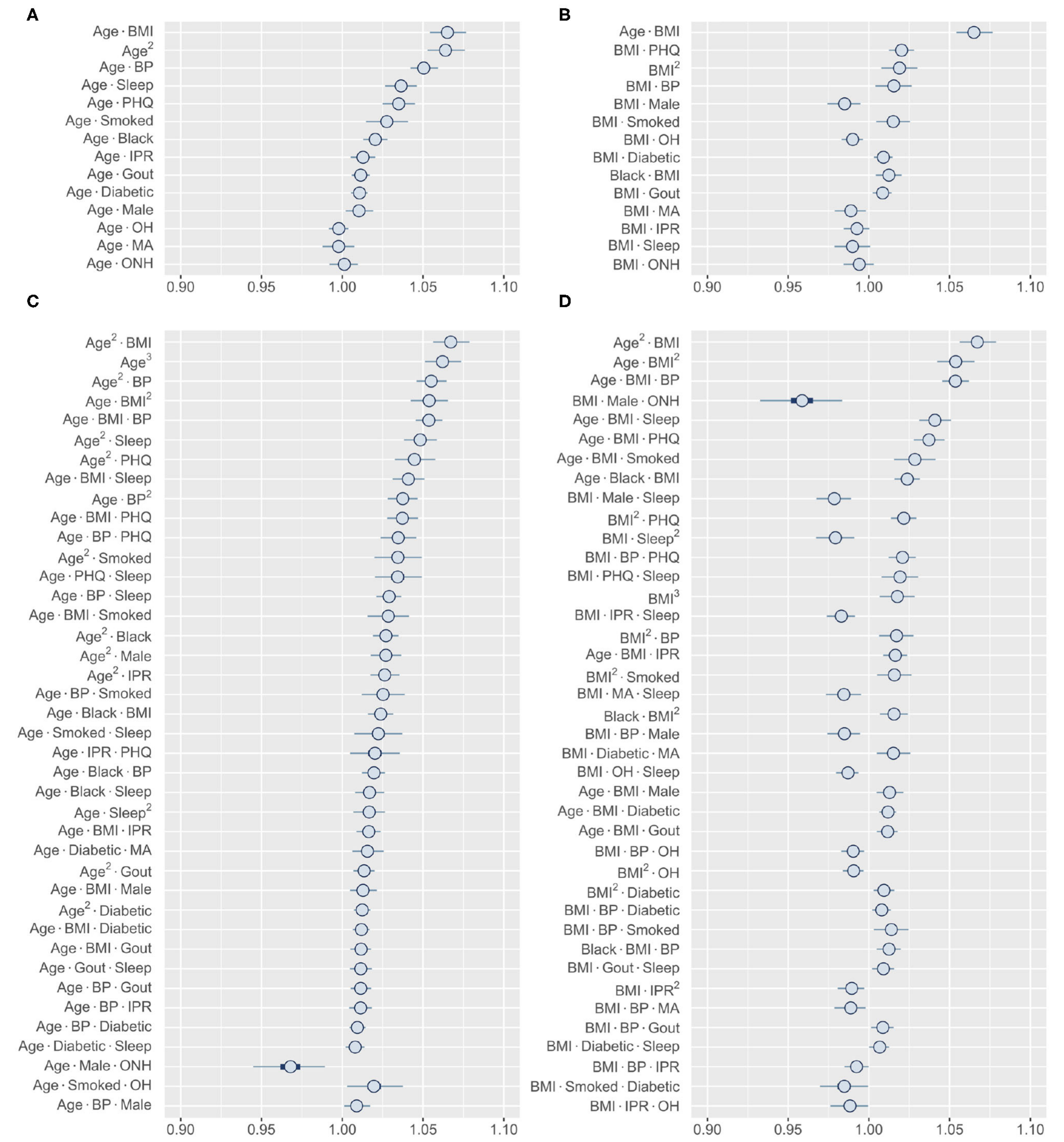


FIGURE 6 | Higher-order interactions of age and BMI. **(A,B)** Posterior distribution of standardized coefficients for all second-order interactions of age **(A)** and BMI **(B)**. **(C,D)** Posterior distribution of standardized coefficients for a selection of 40 third-order interactions involving age **(C)** and BMI **(D)**. Inner bounds and outer bounds represent 50% HDIs and 99% HDIs, respectively. MA, Mexican-American; OH, Other Hispanic; ONH, Other Non-Hispanic. Horizontal scale shows odds multipliers for risk of RA for a one standard deviation increase in value of variable.

3.3.2. Influence Through Interactions: BP and ONH Ethnicity

Finally, we wanted to explore the higher-order interactions for risk factors that did not show a significant first-order effect. Among all first-order effects, only BP and ONH category had 99% HDIs that contained one (**Figure 5B**). Identifying the most influential second-order interactions for BP or ONH category reveals that 12 out of the top 13 involve BP (**Figure 7A**). The

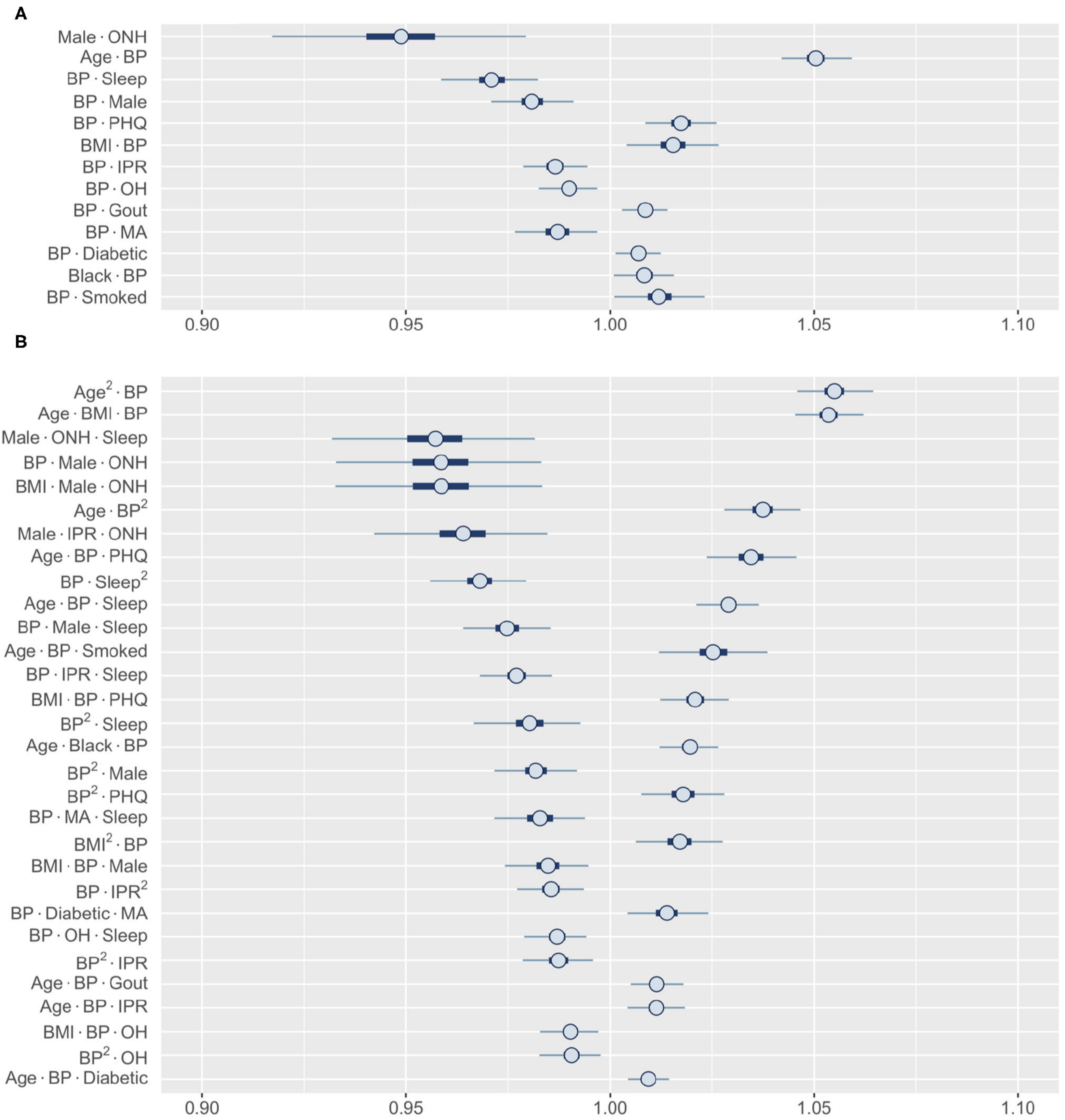


FIGURE 7 | Posterior distributions of most relevant interactions that include either BP or ONH ethnicity as one risk factor. **(A)** 13 most influential second-order interactions and **(B)** 30 most influential third-order interactions are shown. Inner bounds represent 50% HDIs and outer bounds represent 99% HDIs. Standardized coefficient estimates are shown. MA, Mexican-American; OH, Other Hispanic; ONH, Other Non-Hispanic. Horizontal scale shows odds multipliers for risk of RA for a one standard deviation increase in value of variable.

only interaction involving ONH category included in this list (it was also the strongest interaction) is with male gender, strongly lowering the risk for RA. In contrast, the posterior distribution of the interactions of BP indicate that the risk could both increase or decrease depending on the specific interaction. For example, the risk can increase from interaction with age, depression, and BMI, while sleep and male gender reduce the risk. Interestingly, we found the interaction effects of BP with individual risk factors to be similar to their first-order effects. Thus, high BP is expected to enhance the effect of an interaction between risk factors on RA risk. The third-order interactions corroborate well to the second-order interactions with BP occupying 29 of the top 30 interactions (**Figure 7B**). The effect follows the pattern demonstrated by the interaction between the other two factors. While hypertension is generally considered as a comorbidity of RA, there is a lack of consensus on the true association between RA and hypertension (57). Our finding that BP does not have a significant first-order effect but has prominent interaction effects with coexisting conditions, offers a potential explanation for the varying results reported in the literature.

4. DISCUSSION

In this work, we have developed a Bayesian regression model to characterize the risk of RA from common comorbidities, demographic, socioeconomic, and behavioral factors that are known to associate with RA. Apart from providing high predictive accuracy, our model is able to capture the effects of individual variables as well as the important higher-order interactions between them. Consistent with previous literature, known RA risk factors such as depression, high BMI, and smoking are also found to be predictors of RA in our model. Additionally, our model shows that age is not only a key predictor for RA, but also has strong interaction effects with several other variables; prominent among them are BMI, BP, depression, and smoking. Interestingly, some variables such as ONH ethnicity have weak influence as a single-order variable, but their combination with certain other variables (male gender in case of ONH ethnicity) could elicit a prominent higher-order interaction. The knowledge of these strong interactions will help to determine if a person is at a higher or lower risk of RA when both conditions coexist.

One of our primary objectives in this study was to identify and elucidate the effects of important higher-order interactions between risk factors in the prediction of RA. The main challenge in performing such a study comes from the exponential increase in the number of synthetic variables as more higher-order interactions are considered, correspondingly increasing the computational cost. This limitation led us to restrict our study to a maximum of third-order interactions. Our implementation of FAMD further reduced the number of predictor variables analyzed during regression, substantially lowering the requirement for computation. FAMD also allowed the consideration of both categorical and continuous risk factor variables in the model.

In our model, we used feature selection to select an optimal subset of synthetic variables. This step was introduced to not only improve the model's predictive ability but also to obtain a greater precision in determining the effect of risk factors on RA. When studying the manifold interactions between these risk factors, increased precision from feature selection helps to address increases in posterior variances resulting from dramatic increases in the number of variables being analyzed (see **Supplementary Figure 1**). We implemented a wrapper method for feature selection. However, there are alternative approaches, the most common being filter methods (58). Filter methods employ a ranking system to determine the most relevant variables before any prediction is performed (59), some examples of which include the Pearson correlation coefficient, Fisher score, and mutual information (58). Filter-based approaches generally perform faster than wrapper methods since they do not require the predictive model to be run simultaneously. However, because of this, they do not necessarily return the optimal subset of features for prediction (59). Additionally, some filter methods are prone to selecting redundant features (59), while wrapper methods find the optimal subset based on their performance in the predictive model and do not encounter this issue. Thus, employing a wrapper approach for feature selection allowed us to determine the most important subset of synthetic variables for prediction, and subsequently enabled more precise estimates of the effects of interactions between risk factors on RA. One downside of wrapper methods is that they are generally more computationally expensive than filter methods and implementation of techniques based on exhaustive searches can become computationally infeasible for large datasets (59). To overcome this limitation, we implement a wrapper approach using a GA, a type of evolutionary algorithm, and is capable of providing high-quality solutions with reasonable computational effort (60).

Although GA is a robust method for problems involving subset-selection over a large search space, there are alternatives, most notably the Least Absolute Shrinkage And Selection Operator (LASSO) method (61). The presented approach can be interpreted as a heuristic direct search for the best-fit solution using the minimum number of non-zero regression coefficients ("best subset selection"), or an ℓ^0-regularized optimization problem. The LASSO amounts to the relaxation to the best-fit solution with a minimum absolute-sum of regression coefficients, or an ℓ^1-regularized optimization problem. While the discussions about the trade-offs between true best-subset and relaxed best-subset (LASSO) methods are available in the literature [see (62) for an exhaustive list of references], a comparison on this specific problem should be performed in future studies.

Our rationale for using a Bayesian logistic regression model along with feature selection through GA is to achieve a balance between computational efficiency and information obtained. The use of Bayesian inference provides the advantage of getting full posterior information. When compared with decision-tree-based prediction models such as classification and regression trees (CART), logistic regression model allows for a better interpretation of the effects of the individual predictor variables. It also offers a substantial computational advantage when there are a large number of predictors as in the present work.

Existing RA models primarily use genetic, environmental, and behavioral risk factors as predictors (35–37, 41). Karlson et al. reported a logistic regression model that uses a weighted GRS representing the aggregated effects of HLAs and SNPs associated with RA, age, sex, and smoking to predict RA that achieved an AUC of 0.660–0.752, depending on the dataset used (35). Subsequent works using the same model framework but including updated or additional predictor variables such as GRS incorporating newly validated RA risk alleles, exposure to silica, alcohol intake, education, parity, and some of the major interactions between predictors exhibited a similar classification performance (36, 37). A different model using genetic risk factors and smoking data, and determining risk through computer simulation and confidence interval based risk categorization achieved a higher discrimination ability of seropositive RA from control with AUC of 0.837–0.857, although the model is evaluated for male gender alone (41). Although genetic risk factors are demonstrated to be important in RA prediction in these models, our model does not include them considering the potential applicability in peripheral and rural health infrastructures where such advanced genotyping will unlikely be available for patients. Instead, common comorbidities and demographic variables, such as ethnicity, were incorporated in our model as predictors. The promise of our model in predicting RA is demonstrated by a high predictive accuracy in comparison to previous studies, especially when only a smaller subset of first-order variables are considered (see **Supplementary Table 2**). We speculate that a conflation of RA with other forms of arthritis in NHANES datasets could prevent our model achieving substantially higher predictive abilities after incorporation of higher-order effects. This conflation potentially results from self-reported diagnosis of RA and other arthritis in NHANES, and is reflected by a higher proportion of RA in the population than expected from the existing literature (5) (**Supplementary Table 1**).

Our results suggest that our model could achieve high predictive accuracy from the first-order variables alone when an appropriate set of risk factors are selected. While model predictive performance might not improve significantly by incorporating higher-order interactions in such a scenario, identifying the strong interactions could provide important clinical insight. Furthermore, in situations where health resources are highly constrained with severely limited data availability, higher-order interactions could play a significant role in achieving a sufficient degree of predictive accuracy. Our model could also be applicable to predict other chronic diseases that multiple, potentially interacting, factors are known to be associated with.

Even though NHANES provides a rich dataset of risk factors associated with RA, one limitation of the study comes from the self-reported nature of RA diagnosis, which tend to inflate the numbers through false positive diagnosis of other form of arthritis (63). Although a meta-analyses inferred that self-reported diagnosis is sufficiently accurate for large-scale epidemiological studies (64), the model could be made more robust by future validation and optimization with patient data where more rigorous criteria for RA diagnosis, such as the one provided by the American College of Rheumatology, is used (65). The ability to implement sample weights in the model could also marginally improve the model performance. The second limitation comes from the cross-sectional nature of the NHANES data, where the old and new RA cases cannot be discriminated. Furthermore, the comorbidities, socioeconomic and behavioral risk factors coexisted with RA in this data, and thus it could not be temporally resolved whether RA appeared before or after the manifestation of these risk factors. This restricts our model's prediction results on the NHANES dataset to be better interpreted as correlation rather than causation, essentially identifying risk factors and risk interactions associated with RA. We expect the model accuracy to improve along with the ability to infer a causal relationship by training with longitudinal data where the diagnosis of RA can be studied against a population with existing risk factors. Furthermore, Bayesian logistic regression model assumes a simple linear relationship between the predictors and the log-odds of having RA, however, the relationship could be more complex in reality. Although consideration of higher order interactions partially addresses this limitation, a better understanding of the relationship between risk factors and RA could help to construct a more accurate model in the future.

In summary, we have developed a model to predict RA from comorbidities, demographic, socioeconomic, and behavioral risk factors. The model demonstrated a high predictive accuracy in comparison with other models reported in the literature. Moreover, our model was able to identify important second- and third-order interactions between the risk factors, which may have important clinical relevance and stimulate further research to understand the mechanisms underlying such interactions. Since the model prediction utilizes patient information commonly available in a regular healthcare set-up, it has the future potential for translation to the clinical setting.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://wwwn.cdc.gov/nchs/nhanes/.

AUTHOR CONTRIBUTIONS

LL, SM, and SS conceptualized the study. LL developed and executed the procedure. SM and MB verified the procedure. All authors contributed in writing the manuscript.

ACKNOWLEDGMENTS

We thank Daniel Fuller for his generous assistance and meaningful suggestions. We also thank the Clarkson Open Source Institute at Clarkson University for access to their servers. We especially thank Graham Northup for his continued help in resolving technical issues. A preprint of the manuscript is available at the medRxiv preprint server (doi: https://doi.org/10.1101/2020.07.09.20150326).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.693830/full#supplementary-material

REFERENCES

1. Hunter TM, Boytsov NN, Zhang X, Schroeder K, Michaud K, Araujo AB. Prevalence of rheumatoid arthritis in the United States adult population in healthcare claims databases, 2004–2014. *Rheumatol Int*. (2017) 37:1551–7. doi: 10.1007/s00296-017-3726-1
2. Otón T, Carmona L. The epidemiology of established rheumatoid arthritis. *Best Pract Res Clin Rheumatol*. (2019) 33:101477. doi: 10.1016/j.berh.2019.101477
3. Wolfe F, Mitchell DM, Sibley JT, Fries JF, Bloch DA, Williams CA, et al. The mortality of rheumatoid arthritis. *Arthritis Rheum*. (1994) 37:481–494. doi: 10.1002/art.1780370408
4. Birnbaum H, Pike C, Kaufman R, Maynchenko M, Kidolezi Y, Cifaldi M. Societal cost of rheumatoid arthritis patients in the US. *Curr Med Res Opin*. (2010) 26:77–90. doi: 10.1185/03007990903422307
5. Alamanos Y, Drosos AA. Epidemiology of adult rheumatoid arthritis. *Autoimmun Rev*. (2005) 4:130–6. doi: 10.1016/j.autrev.2004.09.002
6. Pincus T. Aggressive treatment of early rheumatoid arthritis to prevent joint damage. *Bull Rheum Dis*. (1998) 47:2.
7. Battafarano DF, Ditmyer M, Bolster MB, Fitzgerald JD, Deal C, Bass AR, et al. 2015 American College of rheumatology workforce study: supply and demand projections of adult rheumatology workforce, 2015–2030. *Arthritis Care Res*. (2018) 70:617–26. doi: 10.1002/acr.23518
8. Symmons D, Barrett E, Bankhead C, Scott D, Silman A. The incidence of rheumatoid arthritis in the United Kingdom: results from the norfolk arthritis register. *Rheumatology*. (1994) 33:735–9. doi: 10.1093/rheumatology/33.8.735
9. Greenberg JD, Spruill TM, Shan Y, Reed G, Kremer JM, Potter J, et al. Racial and ethnic disparities in disease activity in patients with rheumatoid arthritis. *Am J Med*. (2013) 126:1089–98. doi: 10.1016/j.amjmed.2013.09.002
10. Schiff B, Mizrachi Y, Orgad S, Yaron M, Gazit E. Association of HLA-Aw31 and HLA-DR1 with adult rheumatoid arthritis. *Ann Rheum Dis*. (1982) 41:403–4. doi: 10.1136/ard.41.4.403
11. Willkens RF, Nepom GT, Marks CR, Nettles JW, Nepom AS. Association of HLA-Dw16 with rheumatoid arthritis in Yakima Indians. Further evidence for the "shared epitope" hypothesis. *Arthritis Rheum*. (1991) 34:43–7. doi: 10.1002/art.1780340107
12. Deane KD, Demoruelle MK, Kelmenson LB, Kuhn KA, Norris JM, Holers VM. Genetic and environmental risk factors for rheumatoid arthritis. *Best Pract Res Clin Rheumatol*. (2017) 31:3–18. doi: 10.1016/j.berh.2017.08.003
13. van der Woude D, Houwing-Duistermaat JJ, Toes REM, Huizinga TWJ, Thomson W, Worthington J, et al. Quantitative heritability of anti-citrullinated protein antibody-positive and anti-citrullinated protein antibody-negative rheumatoid arthritis. *Arthritis Rheum*. (2009) 60:916–23. doi: 10.1002/art.24385
14. Arleevskaya MI, Kravtsova OA, Lemerle J, Renaudineau Y, Tsibulkin AP. How rheumatoid arthritis can result from provocation of the immune system by microorganisms and viruses. *Front Microbiol*. (2016) 7:1296. doi: 10.3389/fmicb.2016.01296
15. Balandraud N, Roudier J. Epstein-Barr virus and rheumatoid arthritis. *Joint Bone Spine*. (2018) 85:165–70. doi: 10.1016/j.jbspin.2017.04.011
16. Stolt P, Yahya A, Bengtsson C, Källberg H, Rönnelid J, Lundberg I, et al. Silica exposure among male current smokers is associated with a high risk of developing ACPA-positive rheumatoid arthritis. *Ann Rheum Dis*. (2010) 69:1072–6. doi: 10.1136/ard.2009.114694
17. Chang KH, Hsu CC, Muo CH, Hsu CY, Liu HC, Kao CH, et al. Air pollution exposure increases the risk of rheumatoid arthritis: a longitudinal and nationwide study. *Environ Int*. (2016) 94:495–9. doi: 10.1016/j.envint.2016.06.008
18. Heliövaara M, Aho K, Aromaa A, Knekt P, Reunanen A. Smoking and risk of rheumatoid arthritis. *J Rheumatol*. (1993) 20:1830–5.
19. Ksällberg H, Ding B, Padyukov L, Bengtsson C, Rönnelid J, Klareskog L, et al. Smoking is a major preventable risk factor for rheumatoid arthritis: estimations of risks after various exposures to cigarette smoke. *Ann Rheum Dis*. (2011) 70:508–11. doi: 10.1136/ard.2009.120899
20. Di Giuseppe D, Discacciati A, Orsini N, Wolk A. Cigarette smoking and risk of rheumatoid arthritis: a dose-response meta-analysis. *Arthritis Res Ther*. (2014) 16:R61. doi: 10.1186/ar4498
21. Bengtsson C, Nordmark B, Klareskog L, Lundberg I, Alfredsson L. Socioeconomic status and the risk of developing rheumatoid arthritis: results from the Swedish EIRA study. *Ann Rheum Dis*. (2005) 64:1588–1594. doi: 10.1136/ard.2004.031666
22. Markenson JA. Worldwide trends in the socioeconomic impact and long-term prognosis of rheumatoid arthritis. *Semin Arthritis Rheum*. (1991) 21:4–12. doi: 10.1016/0049-0172(91)90046-3
23. Gabriel SE, Crowson CS, O'Fallon WM. Mortality in rheumatoid arthritis: have we made an impact in 4 decades? *J Rheumatol*. (1999) 26:2529–33.
24. Dougados M. Comorbidities in rheumatoid arthritis. *Curr Opin Rheumatol*. (2016) 28:282–8. doi: 10.1097/BOR.0000000000000267
25. Solomon DH, Goodson NJ, Katz JN, Weinblatt ME, Avorn J, Setoguchi S, et al. Patterns of cardiovascular risk in rheumatoid arthritis. *Ann Rheum Dis*. (2006) 65:1608–12. doi: 10.1136/ard.2005.050377
26. Michaud K, Wolfe F. Comorbidities in rheumatoid arthritis. *Best Pract Res Clin Rheumatol*. (2007) 21:885–906. doi: 10.1016/j.berh.2007.06.002
27. Merdler-Rabinowicz R, Tiosano S, Comaneshter D, Cohen AD, Amital H. Comorbidity of gout and rheumatoid arthritis in a large population database. *Clin Rheumatol*. (2017) 36:657–60. doi: 10.1007/s10067-016-3477-5
28. Lee YC, Chibnik LB, Lu B, Wasan AD, Edwards RR, Fossel AH, et al. The relationship between disease activity, sleep, psychiatric distress and pain sensitivity in rheumatoid arthritis: a cross-sectional study. *Arthritis Res Ther*. (2009) 11:R160. doi: 10.1186/ar2842
29. Drewes AM, Svendsen L, Taagholt SJ, Bjerregård K, Nielsen KD, Hansen B. Sleep in rheumatoid arthritis: a comparison with healthy subjects and studies of sleep/wake interactions. *Br J Rheumatol*. (1998) 37:71–81. doi: 10.1093/rheumatology/37.1.71
30. Voigt LF, Koepsell TD, Nelson JL, Dugowson CE, Daling JR. Smoking, obesity, alcohol consumption, and the risk of rheumatoid arthritis. *Epidemiology*. (1994) 5: 525–32.
31. Lu MC, Guo HR, Lin MC, Livneh H, Lai NS, Tsai TY. Bidirectional associations between rheumatoid arthritis and depression: a nationwide longitudinal study. *Sci Rep*. (2016) 6:20647. doi: 10.1038/srep20647
32. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. In: *2008 IEEE/ACS International Conference on Computer Systems and Applications*. Doha: IEEE (2008). p. 108–15.
33. Chin CY, Hsieh SY, Tseng VS. eDRAM: effective early disease risk assessment with matrix factorization on a large-scale medical database: a case study on rheumatoid arthritis. *PLoS ONE*. (2018) 13:e0207579. doi: 10.1371/journal.pone.0207579
34. Shanmugam S, Preethi J. Improved feature selection and classification for rheumatoid arthritis disease using weighted decision tree approach (REACT). *J Supercomput*. (2019) 75:5507–19. doi: 10.1007/s11227-019-02800-1
35. Karlson EW, Chibnik LB, Kraft P, Cui J, Keenan BT, Ding B, et al. Cumulative association of 22 genetic variants with seropositive rheumatoid arthritis risk. *Ann Rheum Dis*. (2010) 69:1077–85. doi: 10.1136/ard.2009.120170
36. Chibnik LB, Keenan BT, Cui J, Liao KP, Costenbader KH, Plenge RM, et al. Genetic risk score predicting risk of rheumatoid arthritis phenotypes and age of symptom onset. *PLoS ONE*. (2011) 6:e24380. doi: 10.1371/journal.pone.0024380
37. Karlson EW, Ding B, Keenan BT, Liao K, Costenbader KH, Klareskog L, et al. Association of environmental and genetic factors and gene-environment interactions with risk of developing rheumatoid arthritis. *Arthritis Care Res*. (2013) 65:1147–56. doi: 10.1002/acr.22005
38. de Hair MJ, Landewé RB, van de Sande MG, van Schaardenburg D, van Baarsen LG, Gerlag DM, et al. Smoking and overweight determine the likelihood of developing rheumatoid arthritis. *Ann Rheum Dis*. (2013) 72:1654–8. doi: 10.1136/annrheumdis-2012-202254
39. Yarwood A, Han B, Raychaudhuri S, Bowes J, Lunt M, Pappas DA, et al. A weighted genetic risk score using all known susceptibility variants to estimate rheumatoid arthritis risk. *Ann Rheum Dis*. (2015) 74:170–6. doi: 10.1136/annrheumdis-2013-204133
40. Sparks JA, Chen CY, Jiang X, Askling J, Hiraki LT, Malspeis S, et al. Improved performance of epidemiologic and genetic risk models for rheumatoid arthritis serologic phenotypes using family history. *Ann Rheum Dis*. (2015) 74:1522–9. doi: 10.1136/annrheumdis-2013-205009
41. Scott IC, Seegobin SD, Steer S, Tan R, Forabosco P, Hinks A, et al. Predicting the risk of rheumatoid arthritis and its age of onset through

modelling genetic risk variants with smoking. *PLoS Genet*. (2013) 9:e1003808. doi: 10.1371/journal.pgen.1003808

42. Kang J. On bayesian inference with complex survey data. *Biom Biostat Int J*. (2016) 3:00076. doi: 10.15406/bbij.2016.03.00076
43. Lesón-Novelo LG, Savitsky TD. Fully Bayesian estimation under informative sampling. *Electr J Stat*. (2019) 13:1608–45. doi: 10.1214/19-EJS1538
44. Löwe B, Unützer J, Callahan CM, Perkins AJ, Kroenke K. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Med Care*. (2004) 42:1194–201. doi: 10.1097/00005650-200412000-00006
45. Pagès J. Analyse factorielle de données mixtes. *Rev Stat Appl*. (2004) 52:93–111. Available online at: http://www.numdam.org/item/RSA_2004__52_4_93_0/
46. Pagès J. Multiple Factor Analysis by Example Using R. Boca Raton, FL: Chapman and Hall/CRC (2014). p. 67–78.
47. Lê S, Josse J, Husson F. FactoMineR: a package for multivariate analysis. *J Stat Softw*. (2008) 25:1–18. doi: 10.18637/jss.v025.i01
48. Gelman A, Carlin JB, Stern HS, Bunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. 3rd Edn. Boca Raton, FL: Chapman and Hall/CRC (2013).
49. Stan Development Team. *RStan: the R interface to Stan.* (2019). R package version 2.19.2. Available online at: http://mc-stan.org/.
50. Brooks SP, Gelman A. General methods for monitoring convergence of iterative simulations. *J Comput Graph Stat*. (1998) 7:434–55. doi: 10.1080/10618600.1998.10474787
51. Hanley JA. The robustness of the" binormal" assumptions used in fitting ROC curves. *Med Decis Mak*. (1988) 8:197–203. doi: 10.1177/0272989X8800800308
52. Macskassy S, Provost F. Confidence bands for ROC curves: methods and an empirical study. In: *Proceedings of the First Workshop on ROC Analysis in AI.* (2004).
53. Collins RJ, Jefferson DR. *Selection in Massively Parallel Genetic Algorithms*. Los Angeles, CA: University of California (Los Angeles); Computer Science Department (1991).
54. Magalhaes-Mendes J. A comparative study of crossover operators for genetic algorithms to solve the job shop scheduling problem. *WSEAS Trans Comput* . (2013) 12:164–73. Available online at: http://www.wseas.us/journal/pdf/computers/2013/5705-156.pdf
55. Sheen YH, Rolfes MC, Wi CI, Crowson CS, Pendegraft RS, King KS, et al. Association of asthma with rheumatoid arthritis: a population-based case-control study. *J Allergy Clin Immunol*. (2018) 6:219–26. doi: 10.1016/j.jaip.2017.06.022
56. Molokhia M, McKeigue P. Risk for rheumatic disease in relation to ethnicity and admixture. *Arthritis Res Ther*. (2000) 2:115. doi: 10.1186/ar76
57. Panoulas VF, Metsios GS, Pace A, John H, Treharne G, Banks M, et al. Hypertension in rheumatoid arthritis. *Rheumatology*. (2008) 47:1286–298. doi: 10.1093/rheumatology/ken159
58. He X, Cai D, Niyogi P. Laplacian score for feature selection. In: *Advances in Neural Information Processing Systems.* (2006). p. 507–14.
59. Chandrashekar G, Sahin F. A survey on feature selection methods. *Comput Electr Eng*. (2014) 40:16–28. doi: 10.1016/j.compeleceng.2013.11.024
60. Goldenberg DE. *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading: MA: Addison Wesley (1989).
61. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc.* (1996) 58:267–88. doi: 10.1111/j.2517-6161.1996.tb02080.x
62. Hastie T, Tibshirani R, Wainwright M. *Statistical Learning with Sparsity: The Lasso and Generalizations. Monographs on Statistics and Applied Probability*. Boca Raton, FL: CRC Press; Taylor & Francis Group (2015).
63. O'Rourke JA, Ravichandran C, Howe YJ, Mullett JE, Keary CJ, Golas SB, et al. Accuracy of self-reported history of autoimmune disease: a pilot study. *PLoS ONE*. (2019) 14:e0216526. doi: 10.1371/journal.pone.0216526
64. Peeters GG, Alshurafa M, Schaap L, de Vet HC. Diagnostic accuracy of self-reported arthritis in the general adult population is acceptable. *J Clin Epidemiol*. (2015) 68:452–9. doi: 10.1016/j.jclinepi.2014.09.019
65. Kay J, Upchurch KS. ACR/EULAR 2010 rheumatoid arthritis classification criteria. *Rheumatology*. (2012) 51:vi5–vi9. doi: 10.1093/rheumatology/kes279

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers
in Public Health

ORIGINAL RESEARCH
published: 25 October 2021
doi: 10.3389/fpubh.2021.715956

Prevalence and Socio-Demographic Correlates of Body Weight Categories Among South African Women of Reproductive Age: A Cross-Sectional Study

Monica Ewomazino Akokuwebe† and Erhabor Sunday Idemudia*†*

Faculty of Humanities, North West University, Mafikeng, South Africa

Background: The shift in disease patterns has been connected with increased body weight burden, becoming a major public health concern in South Africa, as previous studies have assessed overweight or obesity among certain populations. However, little is known about bodyweight burden (underweight, overweight, and obesity) among women aged 15–49 years. Therefore, this study was conducted to identify the prevalence and its associated socio-demographic correlates of bodyweight categories among women of reproductive age in South Africa.

Methods: The present study used the South Africa Demographic Health Survey (2016 SADHS) data for 2016. A total of 3,263 women of reproductive age were included in the analysis. Both bivariable and multivariable logistics regressions were performed to determine the prevalence and socio-demographic correlates of bodyweight categories among women in South Africa. Thus, this study used the criteria of the WHO standard body mass index (BMI) cut-offs to classify bodyweight categories. The odds ratios (ORs) with 95% CIs were estimated for potential determinants included in the final model.

Results: The overall prevalence of body weight burden was 66.5%, with 4.9% underweight, 27.1% overweight, and 34.5% obese ($p < 0.05$). The identified factors associated with underweight among women of reproductive age were those from "other" population group [adjusted odds ratio (AOR) 2.65: 95% CI 1.40–5.00], rural residence (AOR 1.23: 95% CI 0.75–2.02), and Northern Cape Province (AOR 1.58: 95% CI 0.65–3.87). For overweight/obese, the main factors were those aged 45–49 years (AOR 10.73: 95% CI 7.41–15.52), tertiary education (AOR 1.41: 95% CI 0.97–2.03), and residing in Eastern Cape (AOR 1.27: 95% CI 0.82–1.99) and KwaZulu-Natal Provinces (AOR 1.20: 95% CI 0.78–1.84).

Conclusion: The findings presented in this study indicate the concurrence of underweight and overweight/obese among women aged 15–49 years in South Africa. Despite underweight prevalence being on the decline, yet overweight/obese is increasing over time. The health implication of body weight burden needs rapid and effective

OPEN ACCESS

Edited by:
Angelo d'Errico,
Azienda Sanitaria Locale TO3, Italy

Reviewed by:
Anand Krishnan,
All India Institute of Medical Sciences, India
Gobopamang Letamo,
University of Botswana, Botswana
Daniel Goon,
University of Fort Hare, South Africa

****Correspondence:***
Monica Ewomazino Akokuwebe
me.akokuwebenwupostdoctoral2021@gmail.com

†ORCID:
Monica Ewomazino Akokuwebe
orcid.org/0000-0001-8945-345X
Erhabor Sunday Idemudia
orcid.org/0000-0002-5522-7435

Specialty section:
This article was submitted to Life-Course Epidemiology and Social Inequalities in Health, a section of the journal Frontiers in Public Health

Received: *28 May 2021*
Accepted: *13 September 2021*
Published: *25 October 2021*

Citation:
Akokuwebe ME and Idemudia ES (2021) Prevalence and Socio-Demographic Correlates of Body Weight Categories Among South African Women of Reproductive Age: A Cross-Sectional Study. Front. Public Health 9:715956. doi: 10.3389/fpubh.2021.715956

interventions, focusing on factors such as rural, education, population group, older age 45–49 years, and Provinces (Northern Cape, Eastern Cape, and KwaZulu-Natal) – the high-risk groups identified herein are of most importance to curb the growing burden among South African women of reproductive age.

Keywords: body weight categories, nutrition transition, prevalence, South Africa, socio-demographic, women of reproductive age

BACKGROUND

The population dynamics of the rapid and major demographic transitions associated with socio-economic development and accompanied by an epidemiological transition have emerged in industrialised countries and across sub-Saharan African (SSA) countries. The emergence of demographic and epidemiological transitions has led to a decrease in widely known acute infectious diseases and a rising prevalence of non-communicable diseases (NCDs) and chronic degenerative diseases (1, 2). This increased burden of disease and the double burden of malnutrition (undernutrition and over-nutrition) in low-income countries have become a public health concern, which has received attention from both health and non-health experts. With increased knowledge of demographic changes and their impact on the nutrition of the general population, the shift in disease patterns towards diet- or nutrition-related NCDs has been linked with behavioural/lifestyles changes, diets, and environmental exposure. In effect, it has continued to cause a gradual shift in the age pattern of NCD mortality among younger persons (<60 years old) more than those in older age cohorts (>60 years old) (3, 4). However, the reasons for these increasing trends of the age pattern of NCDs are not completely understood.

Several studies have indicated a high prevalence of over-nutrition, which has increased by more than 33.0% (5, 6), contributing to a rapid rise in the NCD burdens in Africa (2, 7). The yearly contributions of each of the major NCD-related deaths include cardiovascular diseases account for 17 million deaths, cancers account for 7.6 million deaths, respiratory diseases account for 4.2 million deaths, and diabetes accounts for 1.3 million deaths (8). Other largely known risk factors shared by these four diseases include tobacco use, physical inactivity, harmful use of alcohol, and an unhealthy diet (9, 10). The occurrence of overweight and obesity has steadily been stated to have intensified, albeit with variation, in developed and developing countries across homogeneous and heterogeneous populations due to a prevalent "obesogenic" environment (11, 12). The contributing factors of obesity need to be better understood as the aetiology of obesity is complex (13, 14), despite the fact that globalisation and urbanisation are two of the key drivers of the malnutrition endemic in South Africa. In addition, to the physiological anatomy of individuals (15), there are behavioural/lifestyle determinants (16) along with economic (17) and environmental/socio-cultural factors (18). Thus, these factors either directly or indirectly have an influence on overweight and obesity progression among women in South Africa.

Since 1940, South Africa has been undergoing a nutrition transition, with an increase in the contribution of fats and a decrease in carbohydrates and fibre intake towards energy consumption (13, 14). Thus, nutrition transition was found to be faster among the black racial groups than the white and Indian/Asian racial groups (19, 20) and in urban rather than rural populations (21), driving an upsurge in obesity in the black population. Pertaining to provinces, for instance, Eastern Cape is the third largest province and the second poorest in South Africa, where 50% of households in the rural districts are food insecure (22). However, studies on the incidence of overweight and obesity in these rural areas are scarcely documented.

Replicating consistent research outcomes of socio-demographic and behavioural/lifestyles risk factors by body weight in developed nations, these and other studies have created indication of a connexion of body mass index (BMI) with gender and with behavioural risk factors: higher BMI among women than men, alcohol ingestion (positive relationship), tobacco use (negative relationship), physical exercise (higher level of physical body exercise associated with lower BMI), and place of residence (urban vs. rural living, with the former associated with higher values of BMI). Considering the observed trends in the BMI, distributions of risk factors (such as increasing urbanisation and average socioeconomic status, stable but high alcohol intake, reduction in tobacco smoking, and decrease physical body exercise) in the population have not been given full attention in non-medical studies (16, 23); yet there is an indication that BMI is connected with factors such socioeconomic status, which is also comparatively proven in demographic and population studies. However, unlike what has been observed in developed countries, there exists an assumption that observed variations in BMI and the prevalence of obesity in South Africa are at least partly driven by changes in the distribution of the above-mentioned risk factors.

Notably, these determinants appear to be remarkably different across culture, age, gender, and social class (24). In addition, there are gaps in knowledge regarding socio-cultural determinants of underweight and overweight/obesity, in particular, at the national level (13, 14). Thus, this study set out to explore the prevalence and socio-demographic correlates of the bodyweight of women in South Africa. Moreover, being underweight or overweight

Abbreviations: AIDS, Acquired Immunodeficiency Syndrome; AOR, Adjusted odds ratio; BMI, Body Mass Index; CI, confidence interval; DHS, Demographic Health Survey; DUs, Dwelling units; HIV, Human Immunodeficiency Virus; NHIS, National Health Insurance Scheme; NCDs, Non-communicable Diseases; OR, Odds ratio; PSUs, Primary sampling units; SDG, Sustainable Development Goal; SES, Socio-economic status; SSA, sub-Saharan Africa; UOR, Unadjusted odds ratio; WHO, World Health Organisation.

is likely to lead to adverse health outcomes throughout life (25), such as the increased risk of maternal disease and adverse pregnancy outcomes (26). The upsurge in the prevalence of obesity at its 2010 level and the worldwide NCD target of halting obesity by the year 2025 have been driven mostly by health concerns and the economic burden of an increasing BMI (27, 28). In addition, there is a realisation from the previous studies that there are scarce analyses on the trends of underweight (23), particularly for the male populations. Disparities in the levels of underweight and obesity across African regions, for example, in Southern African countries comprising Botswana, Namibia, South Africa, and Swaziland (now Eswatini), the obesity index is highest (13, 21, 29–32). In the year 2008, South Africa was rated with the highest BMI, with a median score at the population level approximated at 26.9 kg/m^2 among males (in contrast to a world average of 23.8 kg/m^2), and 29.5 kg/m^2 among females (in contrast to a world average of 24.1 kg/m^2), respectively.

It is important for respective nations to identify their predictable prevalence and changes in body weight (underweight, overweight, and obese) so that this can be used as a basis by governmental stakeholders to improve and enforce suitable intervention policies. At the moment, South Africa does not have ample evidence on the prevalence and its associated socio-demographic correlates of bodyweight categories (underweight, normal, overweight, and obese) among women of reproductive age. Consequently, the research questions under investigation in the current study are: what is the prevalence of each bodyweight category and what are the associated socio-demographic correlates of body weight of women of reproductive age? To avert a further health burden of unhealthy bodyweight categories in the general population of South Africa, it is important that the prevalence, change over time, and communal factors that are quantifiable and responsive to intervention should be identified. We utilised nationally representative data collected using similar methods and techniques to determine the prevalence and correlates of bodyweight categories associated with socio-demographic factors among women of reproductive age in South Africa. This study will, therefore, add valuable information in describing the prevalence of body weight associated with socio-demographic factors to identify immediate and effective interventions for South African women with problems of body weight.

METHODS

Data Source

This study utilised the data from the third round of the South Africa Demographic and Health Survey, conducted in 2016 (2016 SADHS). The 2016 SADHS is a nationally representative sample survey of 15,292 households; 8,514 eligible women in the age range 15–49 years were interviewed with a response rate of 86% (32). The primary objective of this survey was to provide updated and reliable information on marriage and sexual activity, fertility, fertility preferences, contraception, infant and child mortality, maternal healthcare, child health, nutrition of children, HIV/AIDS-related knowledge and behaviour, HIV prevalence, adult and pregnancy-related mortality, use of health services and prescribed medications, adult morbidity, adult nutrition, tobacco, alcohol, and codeine use among adults, women's empowerment, and domestic violence. A stratified two-stage design was utilised in sample selection comprising 750 primary sampling units (PSUs); 468 in urban, 224 in rural, and 58 from farm areas, from a list provided by Statistics South Africa (Stats SA). In stage one, PSUs were selected using probability proportional to PSU size. In stage two, 220 residential dwelling units (DUs) from each PSU were selected with an equal chance of systematic selection from the household listing. The sampling frame used for the 2016 SADHS was the 2011 South African Population and Housing Census. The 2016 SADHS covered age groups of 15–49 years, i.e., those of reproductive age, which made it possible to identify the prevalence of the outcome and explanatory variables associated with weight status. Since the data used in the survey represent only the participants sampled, the data were weighted to make it nationally representative of the participants aged 15–49 years. A comprehensive report of the sampling techniques is provided in the national report of 2016 SADHS (32). This study is based on 5,251 women of reproductive age (15–49 years) who had at least one live birth in the past 5 years preceding the survey. However, 1,988 women who did not respond to BMI questions and who were pregnant at the time of the survey were excluded from the analysis, and a total of 3,263 women were included in the final analysis.

Description of the Measurement of BMI and Its Classification

During the 2016 SADHS, field workers used the portable height/length board in measuring height in centimetres, which was later converted to metres, with restrictions to 1.0–2.7 m (32). Weights were measured using Seca 213 portable stadiometers and formed the boundary of 20–350 kg as advocated by the WHO (33, 34). Using these boundaries, 1.1% of the respondents were in exception in 2016. BMI is calculated with the metric system as follows:

$$\text{BMI} = \frac{\text{Person's weight (kg),}}{\text{Person'sheight (m)}^2}$$

(where kg is kilogrammes and m is metres).

Note that height is commonly measured in centimetres (cm), and height (cm) is divided by 100 to obtain height in metres (m). With the WHO recommendations, the standard weight status categories associated with BMI ranges for adult men and women are the same for all body types and ages. Thus, epidemiological studies have shown substantial risk in people with very high BMI, for instance, severe ($\geq$35 kg/m^2) or morbid ($\geq$40 kg/m^2) obesity (10).

Outcome Variables

The outcome variable for this study was BMI, which was dichotomized as underweight and overweight/obese, respectively. In view of this, binary outcomes with two possible values were constructed as the dependent variable for this study based on underweight vs. normal weight and overweight/obese vs. normal weight, respectively, based on the WHO standard

BMI cut-offs (33, 34). Thus, women who were underweight or overweight/obese were coded "1," and those with normal weight were coded "0." This categorisation was done to ensure large sample sizes for analyses and to obtain more robust binary logistic regression estimates (35–37). Women with BMI < 18.5 kg/m^2 were described as underweight, while those with BMI of 18.5–24.9 kg/m^2 were described as having normal body weight, those with BMI of ≥25 kg/m^2 were overweight, and those with BMI ≥ 30 kg/m^2 were obese.

Explanatory Variables

The selected socio-demographic predictor factors incorporated in the analysis are age (15–19, 20–24, 25–29, 30–34, 35–39, 40–44, and 45–49), marital status (never married and ever married), population group (African/Black and other), educational level (not educated and educated), place of residence (urban and rural), work status (not employed and employed), provinces (Western Cape, Eastern Cape, Northern Cape, Free State, KwaZulu-Natal, North West, Gauteng, Mpumalanga, and Limpopo), and wealth quintile (lowest, middle, and highest). However, the household wealth quintile is a substitute for household economic status and was assessed from possession of household assets, such as consumer items and dwelling characteristics. A score was created for each individual using principal component analysis and classified into five quintiles as Lowest (Poorest), Second (Poorer), Third (Middle), Fourth (Rich), and Highest (Richest) (32).

Statistical Analysis

The data were weighted using sample weights, and the weighted data were used to study the characteristics of the respondents, adjusted for the degree of differences of odds of selection, as the sample design involves more than one stage of selection. This further ensured that data were representative of the target population; in this case, women aged 15–49 years old. To identify the prevalence and socio-demographic correlates of body weight, statistical analyses were carried out at the univariate, bivariate, and multivariate levels. At the univariate level, frequencies and percentages were used to describe the study population and bar graphs was used to describe the prevalence of body weight categories among women (aged 15–49 years). The bivariate analyses were carried out to examine the nature of the association between body weights by selected socio-demographic characteristics. Also at the multivariate level, binary logistic regressions were employed to assess the socio-demographic determinants of body weight of women. The binary regression was calculated as the exponential function of the regression coefficient (e^{b1}) as the measure of the odds ratios (ORs) associated with the outcome and explanatory variables. The findings from the regression analysis were presented as an unadjusted odds ratio (UOR) and adjusted odds ratio (AOR), using 95% CIs and sample covariates (socio-demographic factors) were used to estimate the outcomes. All analyses performed were carried out using STATA version 12.1 (StataCorp LP, College Station, TX, USA).

Ethical Statement

The current analysis is based on the use of secondary datasets from the 2016 SADHS. The 2016 SADHS was conducted under the scientific and administrative supervision of Stats SA, in partnership with the South African Medical Research Council (SAMRC), which conducted the 2016 SADHS, at the request of the National Department of Health (NDoH). Stat SA performed an independent Ethics review of the 2016 SADHS protocol. The data collection procedures were also monitored and approved by the ICF Macro DHS programme team, Calverton, MA, USA. All individuals selected in the SADHS were provided with informed voluntary and written consent. Approval of the individual was sought, and only then was the interview conducted. The survey data collection took place from June 27, 2016 to November 4, 2016. The SADHS dataset is in the public domain and accessible upon a request granted from the Demography Health Survey (DHS) programme (http://www.measuredhs.com).

RESULTS

Socio-Demographic Characteristics of Study Respondents

Table 1 depicts the socio-demographic characteristics of the respondents. The majority of women were in the age cohorts of 15–29 years (50.6%), and 62.2% were never married. More than two-thirds of the women were African/Black, and most of the respondents had secondary education (78%). The majority of women were residing in urban areas, belonged to the unemployed category (70%), the lowest wealth quintile (43.2%), and were from the KwaZulu-Natal province of the country.

Percentage Prevalence of Bodyweight Categories Among Women Aged 15–49 Years

Figure 1 illustrates the four body weight categories of nutrition among women of reproductive age in South Africa. The bar chart reveals that a majority of the women aged 15–49 years were obese (34.5%) followed by those in the normal body weight category (34.3%) in South Africa (**Figure 1**).

Table 2 presents the Chi-square results of the body weight by socio-demographic of women aged 15–49 years in South Africa. Socio-demographic factors such as age, marital status, population group, education, work status, and province have significant relationships with underweight. The prevalence of underweight is 4.1%, and overall, underweight is more prevalent among women who are never married (5.7%). The proportion of underweight women among Black/African is lower than the other population groups (7.4%), and 4.3% of women with secondary education are underweight compared with women with no education/primary, and higher educational level (**Table 2**). Rural women (4.4%) are more likely to be underweight than their urban counterparts (3.9%). There is a significant association among women who are not employed. About 4.6% of women who are underweight are in the middle wealth quintile, and among women residing in Northern Cape Province, 8.8% are underweight (**Table 2**).

TABLE 1 | Socio-demographic characteristics of the weighted samples of women aged 15–49 years old in South Africa, 2016.

Variables	%	*n*
Age		
15–19	17.3	563
20–24	16.3	532
25–29	17.0	554
30–34	14.3	467
35–39	12.7	415
40–44	11.5	375
45–49	10.9	357
Marital status		
Never married	62.2	2,030
Ever married	37.8	1,233
Population group		
African/Black	89.2	2,911
Other	10.8	352
Educational level		
No education and primary	13.1	427
Secondary	77.9	2,542
Higher	9.0	294
Residence		
Urban	53.5	1,747
Rural	46.5	1,516
Work status		
Not employed	69.9	2,281
Employed	30.1	982
Wealth quintile		
Lowest	43.2	1,411
Middle	24.5	800
Highest	32.2	1,052
Provinces		
Western Cape	5.1	166
Eastern Cape	13.0	425
Northern Cape	9.1	297
Free State	10.4	338
KwaZulu-Natal	15.9	520
North West	10.6	346
Gauteng	9.2	299
Mpumalanga	13.3	434
Limpopo	13.4	438
Total, *n*	**100.0**	**3,263**

South Africa Demographic Health Survey (SADHS) (32).

In addition, **Table 2** presents the Chi-square results of overweight/obese by socio-demographics of women in South Africa. The bivariate analysis shows that age, marital status, education, place of residence, work status, and wealth quintile have a significant association with overweight/obese. From the table, the findings revealed that women aged 45–49 years were found to be more overweight/obese (83.5%) than other age cohorts; while 75% of those ever married were found to be overweight/obese, while 61.8% were overweight/obese among Black/African population group. Furthermore, 69.7% of women with higher education were overweight/obese. For women living in urban areas, 63.3% of them were overweight/obese, and 75.1% of the employed women were overweight/obese. Women in the highest wealth quintile were more overweight/obese, and women residing in Western Cape were found to be more overweight/obese (66.9%; **Table 2**).

Table 3 presents multivariate logistic regressions of the UOR and AOR on underweight and overweight/obese and their socio-demographic factors in reference to normal weight among women aged 15–49 years. Thus, compared with women aged 15–19 years, women within the age cohort of 20–24 years were 0.46 times less likely to be underweight. Women in the age cohort of 20–24 years were 2.76 times more likely to be overweight/obese compared to women who were 15–19 years old. Adjusting for other variables used in this study, the findings indicated that women who were 40–44 years old were 0.1 times less likely to be underweight compared to their counterparts who are in the age cohort of 15–19 years. The 40–44-year-old women were found to be 6.79 times more likely to be overweight/obese compared with those aged 15–19 years. In addition, the odds of being overweight/obese were significantly 12.83 and 10.73 times higher among women aged 45–49 years in the unadjusted and adjusted analyses (**Table 3**). Marital status has a definite positive influence on body weight, as marital status is associated with underweight and overweight/obesity in both unadjusted and adjusted analyses. These findings revealed that ever-married women were 0.27 times less likely to be underweight compared to never-married women. When controlling for other variables used in the study, the findings revealed that ever-married women were 0.39 times less likely to be underweight, compared with never-married women. Similarly, ever-married women were found to be 2.62 times more likely to be overweight/obese compared to never-married women.

The adjusted analysis showed that ever-married women were 1.55 times more likely to be overweight/obese compared with never-married women (**Table 3**). Similarly, women who belong to the "other" population group were 2.05 times more likely to be underweight compared with Black/African women. Adjusting for other variables used in the study, women in the "other" population group were 2.56 times more likely to be underweight compared with the Black/African population group. With regards to overweight/obese, women in the "other" population group were 0.94 times less likely to be overweight/obese compared with women in the Black/African population group (**Table 3**). The unadjusted logistic regression analysis showed that women from the "other" population group were 0.59 times less likely to be overweight/obese compared with women in the Black/African population group. Further, there is a significant relationship between education and overweight/obesity, as women with higher education were 1.54 times more likely to be overweight/obese compared with women with no education or primary education. Women with secondary education were 1.61 times more likely to be overweight/obese compared to women with no education or primary education.

Conversely, rural women were found to be 0.86 times less likely to be overweight/obese compared to their urban

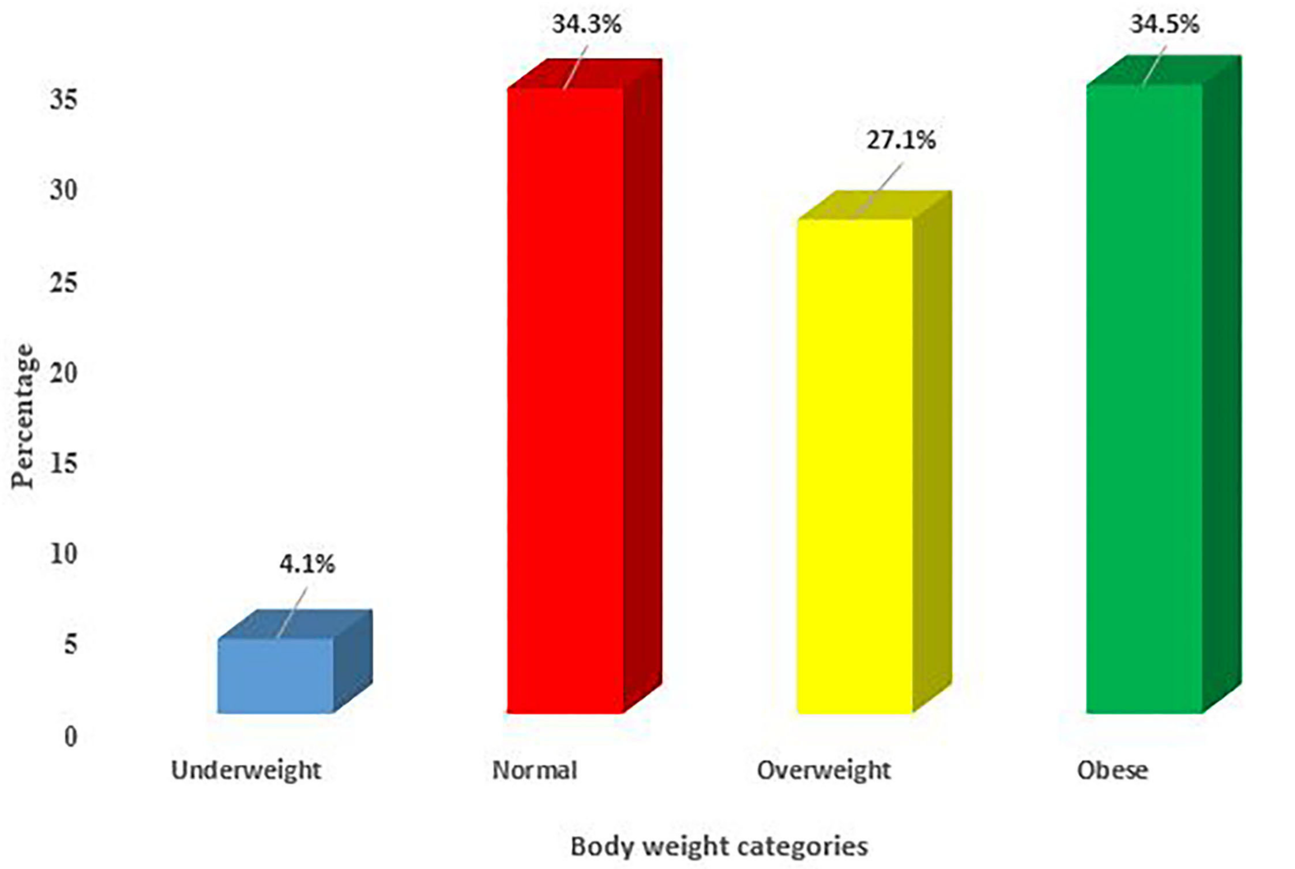


FIGURE 1 | Percentage prevalence of bodyweight categories among women aged 15–49 years who had at least one live birth in the past 5 years preceding the survey in South Africa, 2016.

counterparts, Employed women were 0.41 times less likely to be underweight compared to unemployed women in the unadjusted analysis (**Table 3**). This further indicated that employed women were 2.38 times more likely to be overweight/obese in the unadjusted analysis, and after adjusting for covariates, women who were employed were 1.33 times more likely to be overweight/obese compared to unemployed women (**Table 3**). As regards the wealth quintile, women in the middle and highest wealth quintiles were 1.38 and 1.60 times more likely to be overweight/obese, respectively. The adjusted regression analysis shows that women in the middle and rich wealth quintile were 1.48 and 1.80 times more likely to be overweight/obese, respectively. Furthermore, the province shows a statistical relationship with overweight/obesity in the unadjusted analysis. The findings show that women who reside in Northern Cape were more likely to be underweight compared to other provinces (UOR 2.18; 95% CI 0.92–5.14; AOR 1.58: 95% CI 0.65–3.87). Women residing in the Eastern Cape (AOR 1.27: 95% CI 0.82–1.99) and KwaZulu-Natal Provinces (AOR 1.20: 95% CI 0.78–1.84) were likely to be associated with overweight/obese body weight (**Table 3**).

DISCUSSION

In spite of several national and international efforts to improve the nutritional status of women of reproductive age, a substantial proportion of women are still plagued with underweight and overweight/obesity in South Africa. However, notable progress has been witnessed in the health delivery system along with the provision of the National Health Insurance Scheme (NHIS). South Africa needs to improve health awareness and public sensitisation about the health concerns of underweight and overweight/obesity to achieve the core targets of Sustainable Development Goal (SDG) 3.4 for the reduction of premature deaths from NCDs. The 2016 South Africa key indicator report revealed that the prevalence of overweight (27%) and obesity (41%) were highest among women in South Africa (32). This study established synchronicity of a 2-fold burden of underweight and overweight/obesity among women of reproductive age in South Africa, even though the prevalence of underweight was declining, yet overweight/obesity increased significantly in the period under study.

TABLE 2 | Chi-square associations of body weight by socio-demographic factors among weighted samples of women aged 15–49 years in South Africa, 2016.

Socio-demographic factors	2016					
	Underweight			Overweight and obese		
	No	Yes	*p*-value	No	Yes	*p*-value
Age			**0.000**			**0.000**
15–19	506 (89.9)	57 (10.1)		404 (71.8)	159 (28.2)	
20–24	506 (95.1)	26 (4.9)		255 (47.9)	277 (52.1)	
25–29	536 (96.7)	18 (3.3)		210 (37.9)	344 (62.1)	
30–34	455 (97.4)	12 (2.6)		134 (28.7)	333 (71.3)	
35–39	401 (96.6)	24 (3.7)		104 (25.1)	311 (74.9)	
40–44	371 (98.9)	4 (1.1)		87 (23.2)	288 (76.8)	
45–49	353 (98.9)	4 (1.1)		59 (16.5)	298 (83.5)	
Marital status			**0.000**			**0.000**
Never married	1,915 (94.3)	115 (5.7)		945 (46.6)	1,085 (53.4)	
Ever married	1,213 (98.4)	20 (1.6)		308 (25.0)	925 (75.0)	
Population group			**0.000**			**0.575**
Black/African	2,802 (96.3)	109 (3.7)		1,113 (38.2)	1,798 (61.8)	
Other	326 (92.6)	26 (7.4)		140 (39.8)	212 (60.2)	
Educational level			**0.000**			**0.010**
No education and primary	411 (96.3)	16 (3.7)		171 (40.1)	256 (59.9)	
Secondary	2,433 (95.7)	109 (4.3)		993 (39.1)	1,549 (60.9)	
Higher	284 (96.6)	10 (3.4)		89 (30.3)	205 (69.7)	
Residence			**0.451**			**0.031**
Urban	1,679 (96.1)	68 (3.9)		641 (36.7)	1,106 (63.3)	
Rural	1,449 (95.6)	67 (4.4)		612 (40.4)	904 (59.6)	
Work status			**0.000**			**0.000**
Not employed	2,167 (95.0)	114 (5.0)		1,008 (44.2)	1,273 (55.8)	
Employed	961 (97.9)	21 (2.1)		245 (24.9)	737 (75.1)	
Wealth quintile			**0.444**			**0.000**
Lowest	1.350 (95.7)	61 (4.3)		619 (43.9)	792 (56.1)	
Middle	763 (95.4)	37 (4.6)		289 (36.1)	511 (63.9)	
Highest	1,015 (96.5)	37 (3.5)		345 (32.8)	707 (67.2)	
Provinces			**0.000**			**0.116**
Western Cape	159 (95.8)	7 (4.2)		55 (33.1)	111 (66.9)	
Eastern Cape	416 (97.9)	9 (2.2)		145 (34.1)	280 (65.9)	
Northern Cape	271 (91.2)	26 (8.8)		130 (43.8)	167 (56.2)	
Free State	321 (95.0)	17 (5.0)		125 (37.0)	213 (63.0)	
KwaZulu-Natal	510 (98.1)	10 (1.9)		193 (37.1)	327 (62.9)	
North West	330 (95.4)	16 (4.6)		130 (37.6)	216 (62.4)	
Gauteng	292 (97.7)	7 (2.3)		115 (38.5)	184 (61.5)	
Mpumalanga	414 (95.4)	20 (4.6)		176 (40.6)	258 (59.4)	
Limpopo	415 (94.7)	23 (5.3)		184 (42.0)	254 (58.0)	

$p < 0.001$, $p < 0.01$, $p < 0.05$ is considered statistically significant (Chi-Square test).

The study also observed that the key socio-demographic correlates of underweight were being aged 20–24 years, never married, in the "other" population group, having secondary education, rural, not employed, in the middle wealth quintile, and residing in Northern Cape Province. Since the data represent women of reproductive age (15–49 years) in South Africa, the findings of the study can be generalised to the general population in that age group. The study finding showed that there was a low prevalence of underweight among rural women in South Africa. This result correlates with other findings in sub-Saharan Africa, where underweight prevalence has decreased significantly (38–40). However, a few studies have reported the increased prevalence of underweight in both rural and urban areas in countries, such as Madagascar, Mali, and Senegal (41–43). With regard to overweight/obesity, the main socio-demographic correlates were increased, such as age (44–48), married, tertiary

TABLE 3 | Odd Ratios for socio-demographic factors associated with underweight and overweight/obese among weighted samples of women aged 15–49 years who had at least one live birth in the 5 years preceding the survey in South Africa, 2016.

Socio-demographic factors	Underweight		Overweight/obese	
	UOR (95% CI)	**AOR (95% CI)**	**UOR (95% CI)**	**AOR (95% CI)**
Age				
15–19	RC	RC	RC	RC
20–24	0.46 (0.28–0.74)*	0.49 (0.30–0.82)*	2.76 (2.15–3.54)*	2.60 (2.01–3.36)*
25–29	0.30 (0.17–0.51)*	0.38 (0.211–068)*	4.16 (3.24–5.35)*	3.62 (2.77–4.73)*
30–34	0.23 (0.12–0.44)*	0.34 (0.17–0.69)*	6.31 (4.81–8.29)*	5.00 (3.71–6.73)*
35–39	0.31 (0.17–0.56)*	0.46 (0.24–0.89)*	7.60 (5.70–10.13)*	6.30 (4.60–8.62)*
40–44	0.10 (0.03–0.27)*	0.15 (0.05–0.43)*	8.41 (6.22–11.38)*	6.79 (4.86–9.49)*
45–49	0.10 (0.04–0.28)*	0.16 (0.05–0.47)*	12.83 (9.19–17.93)*	10.73 (7.41–15.52)*
Marital status				
Never married	RC	RC	RC	RC
Ever married	0.27 (0.17–0.44)*	0.39 (0.23–0.68)*	2.62 (2.24–3.06)*	1.55 (1.28–1.86)*
Population group				
Black/African	RC	RC	RC	RC
Other	2.05 (1.32–3.19)*	2.65 (1.40–5.00)*	0.94 (0.75–1.18)*	0.59 (0.43–0.80)*
Educational level				
No education and primary	RC	RC	RC	RC
Secondary	1.15 (0.67–1.97)	0.76 (0.43–1.36)	1.04 (0.84–1.28)	1.61 (1.26–2.06)*
Higher	0.90 (0.40–2.02)	1.06 (0.43–2.59)	1.54 (1.12–2.11)*	1.41 (0.97–2.03)
Residence				
Urban	RC	RC	RC	RC
Rural	1.14 (0.81–1.61)	1.23 (0.75–2.02)	0.86 (0.74–0.99)*	1.11 (0.90–1.36)
Work status				
Not employed	RC	RC	RC	RC
Employed	0.41 (0.26–0.67)*	0.72 (0.43–1.22)	2.38 (2.02–2.81)*	1.33 (1.11–1.61)*
Wealth quintile				
Lowest	RC	RC	RC	RC
Middle	1.07 (0.71–1.63)	0.94 (0.59–1.49)	1.38 (1.15–1.65)*	1.48 (1.20–1.82)*
Highest	0.81 (0.53–1.22)	0.66 (0.38–1.13)	1.60 (1.36–1.89)*	1.80 (1.43–2.26)*
Provinces				
Western Cape	RC	RC	RC	RC
Eastern Cape	0.49 (0.18–1.34)	0.43 (0.14–1.28)	0.96 (0.65–1.40)	1.27 (0.82–1.99)
Northern Cape	2.18 (0.92–5.14)	1.58 (0.65–3.87)	0.64 (0.43–0.95)*	0.80 (0.52–1.23)
Free State	1.20 (0.49–2.96)	1.46 (0.54–3.93)	0.84 (0.57–1.25)	0.82 (0.52–1.29)
KwaZulu-Natal	0.45 (0.17–1.19)	0.35 (0.12–1.03)	0.84 (0.58–1.21)	1.20 (0.78–1.84)
North West	1.10 (0.44–2.73)	1.20 (0.44–3.32)	0.82 (0.56–1.22)	0.85 (0.54–1.34)
Gauteng	0.54 (0.19–1.58)	0.70 (0.22–2.18)	0.79 (0.53–1.18)	0.76 (0.48–1.20)
Mpumalanga	1.10 (0.46–2.64)	1.00 (0.37–2.71)	0.73 (0.50–1.06)	0.85 (0.55–1.33)
Limpopo	1.26 (0.53–2.99)	1.04 (0.38–2.86)	0.68 (0.47–0.99)*	0.85 (0.54–1.34)

**Significant p-values: $p < 0.005$; $p < 0.001$; $p < 0.0001$ 95% Confidence intervals (CI); AOR, adjusted odds ratio; UOR, unadjusted odds ratio; RC, Reference Category; Adjustment variables of the multivariable models are age, marital status, educational level, residence, work status, wealth quintile, and provinces.*

education, rural, employed, high wealth quintile, and residence in the Eastern Cape and KwaZulu-Natal provinces. These findings are consistent with previous studies conducted (41–43).

The high prevalence of overweight/obesity among women aged 15–49 years in South Africa showed differences when compared to the prevalence trends found in studies conducted in Asian countries, where the contrasting trend is observed. Several Asian studies have demonstrated that adult females were more likely to be underweight than overweight/obese (44, 49). The trends of body weight categories in South Africa in 2016 indicated that the prevalence of underweight is decreasing over the years while that of overweight/obesity is increasing. It has been observed that nutrition transition, changes in lifestyles, rapid urbanisation, increasing incomes, and consumption of high-fat food coupled with lack of physical activity are the key causes of the overweight/obesity epidemic in sub-Saharan Africa

(31, 38), including South Africa (17, 43). For instance, a study conducted in Botswana reported that about 82% of people eat insufficient fruits and vegetables, 13% consume alcohol, 12% used tobacco, and about 52% of the respondents reported engaging in no or a low level of physical exercise (41, 42). In South Africa, about 59% of adults have reported consumption of fruits and vegetables while 49% reported that they consumed fruits only, and about 39% consume an unhealthy diet, and 8% and 5% engage in using tobacco products and risky drinking, respectively (14, 45). These findings are similar to evidence from countries, such as Nigeria (46), Ghana (47), Namibia (48), which are countries with an increasing prevalence of overweight/ obesity associated with poor lifestyles. These high prevalence rates of modifiable unhealthy lifestyles can aid in understanding the causes for increased rates of overweight/obesity in South Africa.

As in other parts of the world, it has been ascertained that high blood pressure and diabetes were more predominant among those who were overweight and obese in South Africa (12, 43). Risk factors, such as alcohol, tobacco use, physical inactivity, dietary intake, and sugary drinks, have been identified as high risks for overweight/obesity among the South African population; however, preventable behaviours could lead to its reduction (43). The study findings also demonstrated that women who are employed and are in the Black/African population group demonstrated a higher prevalence of overweight/obesity than women in the "other" population group, while women in the "other" population group were more likely to be associated with underweight than Black/Africans. These findings are similar to several extant studies conducted in South Africa (14, 43). Several scholars have shown a strong impact of socio-economic status on overweight/obesity, predominantly in women, causing disparities in their behaviours towards changes in their energy intake and expenditure, which, as a result, affect their body fat storage (50, 51). The racial disparity in overweight/obesity prevalence remained largely proportional for each respective educational level among Black/African women. The findings of this study reported that Black/African women in South Africa were more likely to be overweight/obese than women in the "other" population group. These outcomes are comparable to the study findings conducted in the United States of America, which illustrated racial trends associated with a higher prevalence of overweight/obesity among black women, such as African-American women (19, 20, 52).

Similarly, existing evidence from previous studies shows that South Africa has the highest proportion of people who were overweight and obese among Coloured (26%) and Black/African population groups (20%), with the majority being women (14, 35, 43). Hence, the advancement of social change, urbanisation, and ageing could be the possible reasons for the key drivers of the prevalence of overweight/obesity among Black/African women in South Africa. Thus, one of the preventive measures in reducing the health burden of overweight/obesity is by ensuring a greater focus on political will and regulation of the way in which products, such as sugar-sweetened drinks and other items, tobacco, and alcohol, have to be scrutinised in South Africa (14, 43). This study did not investigate the effects of these products and items on body weight, although they are generally believed to be linked with poor health outcomes. This study has identified several socio-demographic correlates of body weights of South African women of reproductive age. The findings from the multivariate analysis of this study have established that women were more likely to be overweight/obese with increased age (45–49 years), married, urban, with tertiary education, employed, in the highest wealth quintile, Black/African, and resident in the Eastern Cape and KwaZulu-Natal Provinces than underweight women. This finding is consistent with other studies conducted in India (44) and Nigeria (46).

A possible explanation for higher odds of women with higher education and high wealth quintile being overweight/obese might be due to lifestyles and dietary choices. Women with higher education may not associate their lifestyles with affluence, neglecting the health implication of overweight/obesity. In addition, women in the highest wealth quintile may be less physically active with better dietary choices, such as poor consumption of fruits and vegetables, a higher intake of highly caloric foods, and a poor routine of body exercise (40, 47). The 2016 SADHS report indicated that severe obesity increases with increasing wealth quintile for women. The multivariate analyses found out that rural women were more likely to be underweight compared to urban women. Urban women have quite a lot of advantages over their rural counterparts, such as higher levels of educational knowledge, greater awareness, employment, affluence, and easy access to health services, whereas rural women are often deprived of social and economic prospects (13, 46, 53). However, rural-urban has no significant relationship with body weight in the multivariate analysis. A study conducted in Tanzania reported a similar finding (54).

Our findings highlight the potential impacts of nutrition transition in rural areas as it requires urgent attention to fight against poverty, inequality, unemployment, and lack of basic social amenities. Our study has found that the two provinces having the highest levels of overweight/obesity were Eastern Cape, followed closely by KwaZulu-Natal, but this was not significantly associated in the adjusted model of overweight/obesity multivariate analyses. Contrary to the findings from the 2016 SADHS (32) and the General Household Survey (2016) (55), KwaZulu-Natal and Western Cape were the two provinces that had the highest overweight/obesity status (meanwhile, Western Cape was used in our study, as a reference category in the multivariate analysis). Although the explanations for these increasing trends are not completely understood, a few studies have reported that trends in overweight/obesity are not homogeneous across population strata, but they are defined by biological, socio-economic, and behavioural factors (9, 56–61). The identification of socio-economic and behavioural factors has an immediate prospect from a public health perspective, as these factors are potentially modifiable. In addition, the identification of biological factors is also of public health interest since this knowledge can help in targeting high-risk population groups more effectively, especially in grassroots communities, avoiding waste of resources associated with broad interventions. The presence of socio-economic inequalities in overweight/obesity prevalence is a well-established finding and has been previously confirmed among the South African population (10, 62).

Limitations of the Study

Since this study is based on cross-sectional data, no causal relationships can be deduced from this type of data. However, the ORs were used to determine how different socio-demographic factors are risk factors for either underweight or overweight/obesity. One of the major limitations is that some of the key modifiable factors associated with nutritional transition were excluded from the analysis because the datasets used did not cover behavioural subject themes, such as lifestyle variables. This study is constrained to socio-demographic factors and as such one cannot explain the behavioural aspects of underweight and overweight/obesity. Women who did not respond to BMI questions or who were pregnant at the time of the survey were excluded. In addition, only weighted datasets for women 15–49 years who had at least one live birth in the 5 years preceding the survey were included in this study analysis. Despite the above limitations, the present study uses nationally representative data on underweight and overweight/obesity in South Africa to present generalizable findings.

CONCLUSION

The present study shows that the prevalences of underweight and overweight/obesity were 4.1 and 61.6%, respectively, among women in South Africa. However, the 2016 SADHS key indicators revealed a prevalence of 3.0% for underweight and 68.0% for overweight/obesity among women in South Africa (32). Key socio-demographic factors connected with underweight included women who reside in rural areas and belong to "other" population groups, while the factors linked with overweight/obesity were increasing age, Black/African, higher educational attainment, and higher wealth quintile. Women who were formally employed were less likely to be underweight. Locally relevant policy and interventions should not only target improvement in the socio-economic status of South African women but should also focus on the education of women around the benefits of regular physical activity and healthly dietary choices. It is, therefore, important to take cognizance of those direct interventions, which are designed to tackle the health burden of underweight and overweight/obesity to alleviate health problems associated with nutrition transition. Further research is needed to unravel other factors accompanying underweight and overweight/obesity, which were not enclosed in the existing SADHS datasets. It is also crucial to explore the underlying behavioural factors for underweight and overweight/obesity, such as reasons for low dietary intake, excessive alcohol intake, and tobacco use.

DATA AVAILABILITY STATEMENT

The datasets analysed during the current study are available from the DHS Program: https://dhsprogram.com/data/available-datasets.cfm.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Ethical review and approval for procedures and questionnaires for standard DHS surveys are provided by ICF Institutional Review Board (IRB). Country-specific DHS survey protocols are reviewed by the ICF IRB and typically by an IRB in the host country. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

MA wrote, analysed the DHS datasets, interpreted the DHS data, and was a primary contributor in the manuscript writing. MA and EI revised the manuscript and agreed on the findings and the views expressed within. All authors read and approved the final manuscript.

FUNDING

MA received funding for this research from the North-West University Post-doctoral Research Fellowship (PDRF), Mafikeng, South Africa (Postdoc Grant No. NWU PDRF Fund NW. 1G01487). All funders have been credited and all grant numbers were correctly included in this section of Funding.

ACKNOWLEDGMENTS

The authors are grateful to the DHS Programme for granting access to these data. For their roles in editing and improving the statistical quality of this study, the authors acknowledge Mrs. Helen Thomas and Dr. Ololade Julius Baruwa. The authors also gratefully acknowledge the support of the North-West University Faculty of Humanities, Mafikeng, South Africa.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.715956/full#supplementary-material

REFERENCES

1. Amarya S, Singh K, Sabharwal M. *Ageing Process and Physiological Changes, Gerontology, Grazia D'Onofrio, Antonio Greco and Daniele Sancarlo*. London: IntechOpen (2018). doi: 10.5772/intech open.76249
2. Kyomuhendo C, Adeola R. Green and grey: nutritional lifestyle and healthful ageing in rural and urban areas of three sub-Saharan African countries. *Bus Strat Dev.* (2021) 4:22–33. doi: 10.1002/bsd2.153
3. Khorrami Z, Rezapour M, Etemad K, Yarahmadi S, Khodakarim S, Hezaveh AM, et al. The patterns of non-communicable disease

multimorbidity in Iran: a multilevel analysis. *Sci Rep*. (2020) 10:3034. doi: 10.1038/s41598-020-59668-y

4. Wang Y, Wang J. Modelling and prediction of global non-communicable diseases. *BMC Public Health*. (2020) 20:822. doi: 10.1186/s12889-020-08890-4
5. Abubakari AR, Lauder W, Agyemang C, Jones M, Kirk A, Bhopal RS. Prevalence and time trends in obesity among adult West African populations: a meta-analysis. *Obes Rev*. (2008) 94:297–311. doi: 10.1111/j.1467-789X.2007.00462.x
6. Letamo G. Dual burden of underweight and overweight/obesity among adults in Botswana: prevalence, trends and socio-demographic correlates: a cross-sectional survey. *BMJ Open*. (2020) 10:e038614. doi: 10.1136/bmjopen-2020-038614
7. Modjadji P. Socio-demographic determinants of overweight and obesity among mothers of primary school children living in a rural health and demographic surveillance system site, South Africa. *Open Public Health J. (2020)* 13:518–28. doi: 10.2174/1874944502013010518
8. Wandai ME, Aagaard-Hansen J, Manda SOM, Norris SA. Transitions between body mass index categories, South Africa. *Bull World Health Org*. (2020) 98:878–51. doi: 10.2471/BLT.20.255703
9. Sartorius B, Veerman LJ, Manyema M, Chola L, Hofman K. Determinants of obesity and associated population attributability, South Africa: empirical evidence from a national panel survey, 2008-2012. *PLoS ONE*. (2015) 10:e0130218. doi: 10.1371/journal.pone.0130218
10. Cois A, Day C. Obesity trends and risk factors in the South African adult population. *BMC Obesity*. (2015) 2:42. doi: 10.1186/s40608-015-0072-2
11. Balaskas P, Jackson ME, Blundell JE, Dalton M, Gibbons C, le Roux CW, et al. Aetiology of obesity in adults. In: Hankey C, Whelan K, editors. *Advanced Nutrition and Dietetics in Obesity*. Series: Advanced nutrition and dietetics (BDA). John Wiley & Sons, Ltd, New Jersey USA. 85–137. ISBN 9780470670767; doi: 10.1002/97811188579991.ch3
12. Gadde KM, Martin CK, Berthoud H-R, Heymsfield SB. Obesity: pathophysiology and management. *J Am Coll Cardiol*. (2018) 71:69–84. doi: 10.1016/j.jacc.2017.11.011
13. Kruger HS. Obesity among women: a complex setting. *South Afr J Clin Nutr*. (2018) 31:4–5.
14. Ndinda C, Ndhlovu TP, Juma P, Asiki G, Kyobutungi C. The evolution of non-communicable diseases policies in post-apartheid South Africa. *BMC Public Health*. (2018) 18:956. doi: 10.1186/s12889-018-5832-8
15. Jiménez EG. Obesity: ethiologic and pathophysiological analysis. *Endocrinol Nutr*. (2013) 60:17–24. doi: 10.1016/j.endoen.2013.01.005
16. Narciso J, José Silva A, Rodrigues V, Montéiro JM, Almeida A, Saavedra R, et al. Behavioral, contextual and biological factors associated with obesity during adolescence: a systematic review. *PLoS ONE*. 14:e0214941. doi: 10.1371/journal.pone.0214941
17. Micklesfield LK, Kagura J, Munthali R, Crowther NJ, Jaff N, Gradidge P, et al. Demographic, socio-economic and behavioural correlates of BMI in middle-aged black men and women from urban Johannesburg, South Africa. *Glob Health Action*. (2018) 11(Suppl. 2):1448250. doi: 10.1080/16549716.2018.1448250
18. Heymsfield SB, Wadden TA. Mechanisms, pathophysiology, and management of obesity. *N Engl J Med*. (2017) 376:254–66. doi: 10.1056/NEJMra1514009
19. Agyemang P, Powell-Wiley TM. Obesity and black women: special considerations related to Genesis and therapeutic approaches. *Curr Cardiovasc Risk Rep*. (2013) 7:378–86. doi: 10.1007/s12170-013-0328-7
20. Oraka CS, Faustino DM, Oliveira E, Teixeira JAM, Souza ASP de, Luiz CO. do. Race and obesity in the black female population: a scoping review. *Saude Soc*. (2020) 29:e191003. doi: 10.1590/s0104-12902020191003
21. Cheteni P, Khamfula Y, Mah G. Exploring food security and household dietary diversity in the Eastern Cape Province, South Africa. *Sustainability*. (2020) 12:1851. doi: 10.3390/su12051851
22. Macek P, Biskup M, Terek-Derszniak M, Krol H, Smok-Kalwat J, Stanislaw G, et al. Optimal cut-off values for anthropometric measures of obesity in screening for cardiometabolic disorders in adults. *Sci Rep*. (2020) 10:11253. doi: 10.1038/s41598-020-68265-y
23. Higgins V, Nazroo J, Brown M. Pathways to ethnic differences in obesity: the role of migration, culture and socio-economic position in the UK. *SSM-Popul Health*. (2019) 12:100716. doi: 10.1136/jech-2019-SSMabstracts.172
24. Peer N, Ganie YN. A weighty matter: identification and management of overweight and obesity. *S Afr Med J*. (2016) 106:7. doi: 10.7196/SAMJ.2016.v106i7.10946
25. Basu J, Jeketera CM, Basu D. Obesity and its outcomes among pregnant South African women. *Int J Gynaecol Obstet*. (2010) 110:101–4. doi: 10.1016/j.ijgo.2010.02.020
26. Pisa PT, Pisa NW. Economic growth and obesity in South African adults: an ecological analysis between 1994 and (2014). *Eur J Public Health*. (2016) 27:404–9. doi: 10.1093/eurpub/ckw119
27. Weir CB, Jan A. *BMI Classification Percentile and Cut off Points*. Treasure Island, FL: StatPearls Publishing (2021).
28. Keetile M, Navaneetham K, Letamo G, Bainame K, Rakgoasi SD, Gabaitiri L, et al. Socioeconomic and behavioural determinants of overweight/obesity among adults in Botswana: cross-sectional survey. *BMJ Open*. (2019) 9:12.
29. Zegeye B, Shibre G, Woldeamanuel GG. Time trends in socio-economic, urban-rural and regional disparities in prevalence of obesity among non-pregnant women in Lesotho: evidence from Lesotho demographic and health surveys (2004-2014). *BMC Public Health*. (2021) 21:537. doi: 10.1186/s12889-021-10571-9
30. Craig LS, Gage AJ, Thomas AM. Prevalence and predictors of hypertension in Namibia: a national-level cross-sectional study. *PLoS ONE*. (2018) 13:e0204344. doi: 10.1371/journal.pone.0204344
31. Neupane S, Prakash KC, Doku DT. Overweight and obesity among women: analysis of demographic and health survey data from 32 sub-Saharan African countries. *BMC Public Health*. (2015) 16:30. doi: 10.1186/s12889-016-2698-5
32. South Africa Demographic and Health Survey (SADHS). *South Africa Demographic and Health Survey Key Indicators Report (2016)*. Pretoria; Rockville, MD: National Department of Health (NDoH); Statistics South Africa (Stats SA); South African Medical Research Council (SAMRC); ICF. Available online at: https://www.google.com/search?q=38.%09South+Africa+Demographic+and+Health+Survey+%28SADHS%29+South+Africa+Demographic+and+Health+Survey+Key+Indicators+Report+2016. (accessed August 3, 2021).
33. World Health Organization (WHO). *WHO/Obesity and Overweight*. World Health Organization (2015). Available online at: http://www.who.int/mediacentre/factsheets/fs311/en/ (accessed August 4, 2021).
34. World Health Organization (WHO). Physical Status: the use and interpretation of anthropometry. Report of a WHO Expert Committee. *World Health Organ Tech Rep Ser*. (1995) 854:1–452.
35. Linaker CH, Angelo SD, Syddall HE, Harris EC, Cooper C, Walker-Bone K. Body mass index (BMI) and work ability in older workers: results from the Health and Employment after Fifty (HEAF) Prospective Cohort Study. *Int J Environ Res Public Health*. (2020) 17:1647. doi: 10.3390/ijerph17051647
36. Ohlsson C, Gidestrand E, Bellman J, Larsson C, Palsdottir V, Hagg D, et al. Increased weight loading reduces body weight and body fat in obese subjects-A proof of concept randomized clinical trial. *EClinicalMedicine*. (2020) 22:100338. doi: 10.1016/j.eclinm.2020.100338
37. Doku DT, Neupane S. Double burden of malnutrition: increasing overweight and obesity and stall underweight trends among Ghanaian, women. *BMC Public Health*. (2015) 15:670. doi: 10.1186/s12889-015-2033-6
38. Akombi BJ, Agho KE, Hall JJ, Wali N, Renzaho AMN, Merom D. Stunting, wasting and underweight in sub-Saharan Africa: a systematic review. *Int J Environ Res Public Health*. (2017) 14:863. doi: 10.3390/ijerph14080863
39. Moise IK, Kangmennaang J, Halwiindi H, Grigsby-Toussaint DS, Fuller DO. Increase in obesity among women of reproductive age in Zambia, 2002-2014. *J Women Health (Larchmt)*. (2109) 28:1679–87. doi: 10.1089/jwh.2018.7577
40. Mangemba NT, San Sebastian M. Societal risk factors for overweight and obesity in women in Zimbabwe: a cross-sectional study. *BMC Public Health*. (2020) 20:103. doi: 10.1186/s12889-020-8215-x
41. Letamo G. The prevalence of, and factors associated with, overweight and obesity in Botswana. *J Biosoc Sci*. (2011) 43:75–84. doi: 10.1017/S0021932010000519
42. Navaneetham K, Keetile M, Letamo G, Bainame K, Rakgoasi SD, Masupe T, Gabaitiri L, Molebatsi R, et al. *Chronic). Non-communicable Diseases in Botswana: a Study of Prevalence, Healthcare Utilization and Health*

Expenditure. Project Report, Office of Research and Development, University of Botswana, Gaborone. 2018

43. Maimela E, Alberts M, Bastiaens H, Fraeyman J, Meulemans H, Wens J, et al. Interventions for improving management of chronic non-communicable diseases in Dikgale, a rural area in Limpopo Province in South Africa. *BMC Health Serv Res*. (2018) 18:331. doi: 10.1186/s12913-018-3085-y
44. Al kibria GM, Swasey K, Hasan MZ, Sharmeen A, Day B. Prevalence and factors associated with underweight, overweight and obesity among women of reproductive age in India. *Glob Health Res Policy*. (2019) 4:24. doi: 10.1186/s41256-019-0117-z
45. Hofman K. Non-communicable diseases in South African: a challenge to economic development. *South Afr Med J*. (2014) 104:647. doi: 10.7196/SAMJ.8727
46. Adeloye D, Ige-Elegbede JO, Ezejimofor M, Owolabi EO, Ezeigwe N, Omoyele C, et al. Estimating the prevalence of overweight and obesity in Nigeria in 2020: a systematic review and meta-analysis. *Ann Med*. (2020) 53:495–507. doi: 10.1080/07853890.2021.1897665
47. Lartey ST, Magnussen CG, Si L, Boateng GO, de Graaff B, Biritwum RB, et al. Rapidly increasing prevalence of overweight and obesity in older Ghanaian adults from 2007-2015: Evidence from WHO-SAGE Waves 1 &2. *PLoS ONE*. (2019) 14:e0215045. doi: 10.1371/journal.pone.0215045
48. Haufiku D, Amukugo HJ. Prevalence and factors associated with obesity amongst employees of open-cast diamond mine in Namibia. *Int J Adv Nurs Stud*. (2015) 4:85. doi: 10.14419/ijans.v4i2.4906
49. Biswas T, Townsend N, Magalhaes RJS, Islam S, Hasan M, Mamun A. Current progress and future directions in the double burden of malnutrition among women in South and Southeast Asian countries. *Curr Dev Nutr*. (2019) 3:nzz026. doi: 10.1093/cdn/nzz026
50. Matos UR, Mesenburg MA, Victoria CG. Socioeconomic inequalities in the prevalence of underweight, overweight and obesity among women aged 20-49 in low- and middle-income countries. *Int J Obes*. (2020) 44:609–16. doi: 10.1038/s41366-019-0503-0
51. Hasan E, Khanam M, Shimul SN. Socio-economic inequalities in overweight and obesity among women of reproductive age in Bangladesh: a decomposition approach. *BMC Women Health*. (2020) 20:263. doi: 10.1186/s12905-020-01135-x
52. Jackson CL, Szklo M, Yeh H-C, Wang N-Y, Dray-Spira R, Thorpe R, et al. Black-White disparities in overweight and obesity Trends by educational attainment in the United States, 1977-2008. *J Obes*. (2013) 2013:140743. doi: 10.1155/2013/140743
53. Frayne B, Crush J. McLachlan. Urbanization, nutrition and development in Southern Africa cities. *Food Sec*. (2014) 6:101–12. doi: 10.1007/s12571-013-0325-1
54. Pinchoff J, Mills CW, Balk D. Urbanization and health: the effects of the built environment on chronic diseases risk factors among women in Tanzania. *PLoS ONE*. (2020) 15:e02418. doi: 10.1371/journal.pone.0241810
55. Statistics South Africa. *General Household Survey*, 2016. (P0318). Pretoria: Statistics South Africa. Available online at: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjG8eaLoerwAhUjQkEAHaWNDb4QFjAAegQIAxAD&url=https%3A%2F%2Fhttp://www.statssa.gov.za%2Fpublications%2FP0318%2FP03182016.pdf&usg=AOvVaw315sYUDIFUHrZWuZovOGY
56. Akokuwebe ME. "Healthy women, healthy world": a theoretical discourse of general health status of women in Nigeria. *West Afr J Archaeol (Revue Quest Afr Archeol)*. (2016) 46:87–111.
57. Akokuwebe ME, Okunola RA, Falayi SE. Youths and risky sexual behaviour: a KAP study on HIV/AIDS amongst University of Ibadan student. *Eur J Soc Sci*. (2015) 49:140–52.
58. Amusan L, Akokuwebe ME, Odularu G. Women development in agriculture as agency for fostering innovative agricultural financing in Nigeria. *Afr J Food Agric Nutr Dev*. (2021) 21:18332–52. doi: 10.18697/ajfand.102.19345
59. Akokuwebe ME, Clifford Odimegwu C. Socioeconomic determinants of knowledge of kidney disease among residents in Nigerian communities in Lagos State, Nigeria. *Oman Med J*. (2019) 34:444–55. doi: 10.5001/omj.2019.81
60. Akokuwebe ME, Clifford Odimegwu C, Omololu F. Prevalence, risk-inducing lifestyle, and perceived susceptibility to kidney diseases by gender among Nigerians residents in South Western Nigeria. *Afr Health Sci*. (2020) 20:860–70. doi: 10.4314/ahs.v20i2.40
61. Odularu G, Aluko OA, Odularu A, Akokuwebe M, Adedugbe A. Conclusion: Fostering nutrition security, climate adaptation and sustainable agriculture strategies amid Covid-19 pandemic. In: Odularu G, editors. *Nutrition, Sustainable Agriculture and Climate Change in Africa*. Cham: Palgrave Macmillan (2020). p. 175–82.
62. Goetjes E, Pavlova M, Hongoro C, Groot W. Socioeconomic inequalities and obesity in South Africa–a decomposition analysis. *Int J Environ Res Public Health*. (2021) 18:9181. doi: 10.3390/ijerph18179181

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Public Health

ORIGINAL RESEARCH
published: 28 October 2021
doi: 10.3389/fpubh.2021.742952

Causal Associations Between Educational Attainment and 14 Urological and Reproductive Health Outcomes: A Mendelian Randomization Study

Menghua Wang[1†], *Zhongyu Jian*[1,2†], *Xiaoshuai Gao*[1†], *Chi Yuan*[1], *Xi Jin*[1], *Hong Li*[1] *and Kunjie Wang*[1*]

[1] Department of Urology, Institute of Urology (Laboratory of Reconstructive Urology), West China Hospital, Sichuan University, Chengdu, China, [2] West China Biomedical Big Data Center, Sichuan University, Chengdu, China

OPEN ACCESS

Edited by:
Joao Sollari Lopes,
National Statistical Institute of Portugal, Portugal

Reviewed by:
Tor Åge Myklebust,
Cancer Registry of Norway, Norway
Vu Duy Kien,
Independent Researcher, Hanoi, Vietnam

***Correspondence:**
Kunjie Wang
wangkj@scu.edu.cn

[†] These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Life-Course Epidemiology and Social Inequalities in Health, a section of the journal Frontiers in Public Health

Received: *17 July 2021*
Accepted: *27 September 2021*
Published: *28 October 2021*

Citation:
Wang M, Jian Z, Gao X, Yuan C, Jin X, Li H and Wang K (2021) Causal Associations Between Educational Attainment and 14 Urological and Reproductive Health Outcomes: A Mendelian Randomization Study. Front. Public Health 9:742952. doi: 10.3389/fpubh.2021.742952

Background: The impact of educational attainment (EA) on multiple urological and reproductive health outcomes has been explored in observational studies. Here we used Mendelian randomization (MR) to investigate whether EA has causal effects on 14 urological and reproductive health outcomes.

Methods: We obtained summary statistics for EA and 14 urological and reproductive health outcomes from genome-wide association studies (GWAS). MR analyses were applied to explore the potential causal association between EA and them. Inverse variance weighted was the primary analytical method.

Results: Genetically predicted one standard deviation (SD) increase in EA was causally associated with a higher risk of prostate cancer [odds ratio (OR) 1.14, 95% confidence interval (CI) 1.05–1.25, $P = 0.003$] and a reduced risk of kidney stone (OR 0.73, 95% CI 0.62–0.87, $P < 0.001$) and cystitis (OR 0.76, 95% CI 0.67–0.86, $P < 0.001$) after Bonferroni correction. EA was also suggestively correlated with a lower risk of prostatitis (OR 0.76, 95% CI 0.59–0.98, $P = 0.037$) and incontinence (OR 0.64, 95% CI 0.47–0.87, $P = 0.004$). For the bioavailable testosterone levels and infertility, sex-specific associations were observed, with genetically determined increased EA being related to higher levels of testosterone in men (β 0.07, 95% CI 0.04–0.10, $P < 0.001$), lower levels of testosterone in women (β −0.13, 95% CI−0.16 to−0.11, $P < 0.001$), and a lower risk of infertility in women (OR 0.74, 95% CI 0.64–0.86, $P <0.001$) but was not related to male infertility (OR 0.79, 95% CI 0.52–1.20, $P = 0.269$) after Bonferroni correction. For bladder cancer, kidney cancer, testicular cancer, benign prostatic hyperplasia, and erectile dysfunction, no causal effects were observed.

Conclusions: EA plays a vital role in urological diseases, especially in non-oncological outcomes and reproductive health. These findings should be verified in further studies when GWAS data are sufficient.

Keywords: educational attainment, Mendelian randomization, urology, reproductive health, oncology

INTRODUCTION

It is well-established that educational attainment (EA) is an essential social determinant of health (1). A prior study reported that EA was correlated with many health outcomes, including adiposity, diabetes, and coronary artery diseases (2), suggesting the non-negligible role of EA in health.

In the field of urology and reproductive medicine, there have also been some observational studies that investigated the correlation between EA and health outcomes, namely, prostate cancer (3, 4), bladder cancer (5), kidney cancer (6), testicular cancer (7), kidney stone (8, 9), benign prostatic hyperplasia (BPH) (10, 11), prostatitis, cystitis, incontinence (12, 13), erectile dysfunction (ED) (14), male infertility (15), female infertility (15–17), and testosterone levels among males and females (18), showing that EA might play a vital role in urological and reproductive health. However, there are few relevant studies, and the results from prior studies were partially inconsistent. Additionally, existing observational studies are vulnerable to confounding factors and reverse causality.

Mendelian randomization (MR) is a genetic epidemiological method that applies genetic variants, such as single nucleotide polymorphisms (SNPs), to estimate the causal effect of an exposure (e.g., EA) on an outcome (e.g., kidney stone). Compared with conventional observational studies, this method is less vulnerable to confounding factors and reverse causation and has been widely used in current epidemiological studies (19).

Recently, a large-scale genome-wide association study (GWAS) identified genetic variants associated with EA (20), which provides high-quality genetic instruments for us to estimate the causal effects of EA on health outcomes. The genetic variants derived from this GWAS have already been used to evaluate the causal effects of EA on osteoarthritis (21) and diabetes (22).

As a result, in the current research, we used MR analysis to determine the causal effect of EA on the 14 urological and reproductive health outcomes mentioned above, to provide new insights into the role of EA in these health outcomes.

MATERIALS AND METHODS

We performed the current MR study based on the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) guideline (**Supplementary Table 1**). The overall study design of the current MR analysis is presented in **Supplementary Figure 1**.

Instrumental Variables Selection

We used SNPs that were identified to be correlated with EA from a GWAS performed by the Social Science Genetic Association Consortium (20). This GWAS was a meta-analysis of 71 cohort-level studies that enrolled 1,131,881 individuals of European ancestry. Education attainment was measured as the number of years of schooling that participants completed. Although there are differences in education systems for EA between cohorts, the International Standard Classification of Education system was applied to match education qualifications across the cohorts. Under the threshold of $P < 5 \times 10^{-8}$ and pairwise $r^2 < 0.1$, the GWAS identified 1,271 SNPs that are correlated with EA, which explained 11–13% of the variance. Among the 1,271 SNPs, the SNPs with potential linkage disequilibrium (pairwise $r^2 > 0.01$), those not found in the GWAS outcome datasets, and those that were palindromic with intermediate allele frequencies were excluded. Since the quality of the instrumental variables was essential for the MR study, we used the F statistics to evaluate the strength of the instrumental variables. Although we did not calculate the F statistics specifically in the current study, a prior study that investigated the association between EA and osteoarthritis using similar SNPs as our study reported a median F statistics of 45 (21), suggesting that the instrument strength was generally reliable. The SNP coefficients were per standard deviation (SD) units of years of schooling (SD = 4.2 years).

GWAS Data Sources for 14 Urological and Reproductive Outcomes

We extracted summary statistics for prostate cancer (79,148 cases and 61,106 controls) from the Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome (PRACTICAL) Consortium (23). The genetic variants for bioavailable testosterone levels were extracted from a gender-specific GWAS performed in the UK Biobank (178,782 men and 188,507 women) (24). The UK Biobank is a large-scale biomedical database and has been widely used in the field of health. Summary statistics for ED were obtained from another GWAS with 6,175 cases and 217,630 controls in total (25). We obtained the genetic variants for the remaining 10 outcomes, including bladder cancer (1,115 cases and 174,006 controls), kidney cancer (971 cases and 174,006 controls), testicular cancer (199 cases and 74,685 controls), kidney stone (4,969 cases and 213,445 controls), BPH (13,118 cases and 72,799 controls), prostatitis (1,859 cases and 72,799 controls), cystitis (8,081 cases and 195,140 controls), incontinence (1,357 cases and 202,910 controls), male (680 cases and 72,799 controls), and female (6,481 cases and 68,969 controls) infertility from the latest R5 release of the FinnGen project. The FinnGen project is an ongoing project combining the genotype data and digital health record data of Finnish individuals, which provides a high-quality database for researchers to explore genetic variation in diseases. Detailed information about the FinnGen project can be found at their official site (26). A description of the 14 urological and reproductive health outcomes, including data sources, sample size, and definitions, is presented in **Supplementary Table 2**.

Statistical Analysis

Inverse variance weighted (IVW) was the primary analytical method in our study, which could provide the most precise causal estimates (27). Additionally, we performed several sensitivity analyses to validate our findings, including MR-Egger, weighted median, and weighted mode. MR-Egger is a method that can provide estimates after the correction of pleiotropy (28). The weighted median method could generate reliable estimates even if up to 50% of weights come from invalid instruments (29). The weighted mode has natural robustness to outlying variants (30). We used the MR-Egger intercept to examine directional pleiotropy and Cochrane's Q-test to estimate heterogeneity. Since we included 14 urological and reproductive health outcomes, the significance threshold was $P < 0.0036$ (0.05/14) after Bonferroni correction. A $P < 0.05$, but above the threshold of Bonferroni correction significance, was considered a suggestive causal association. All the statistical analyses were conducted using R software.

RESULTS

A flow diagram for eligible SNPs selection for each of 14 outcomes was presented in **Figure 1**.

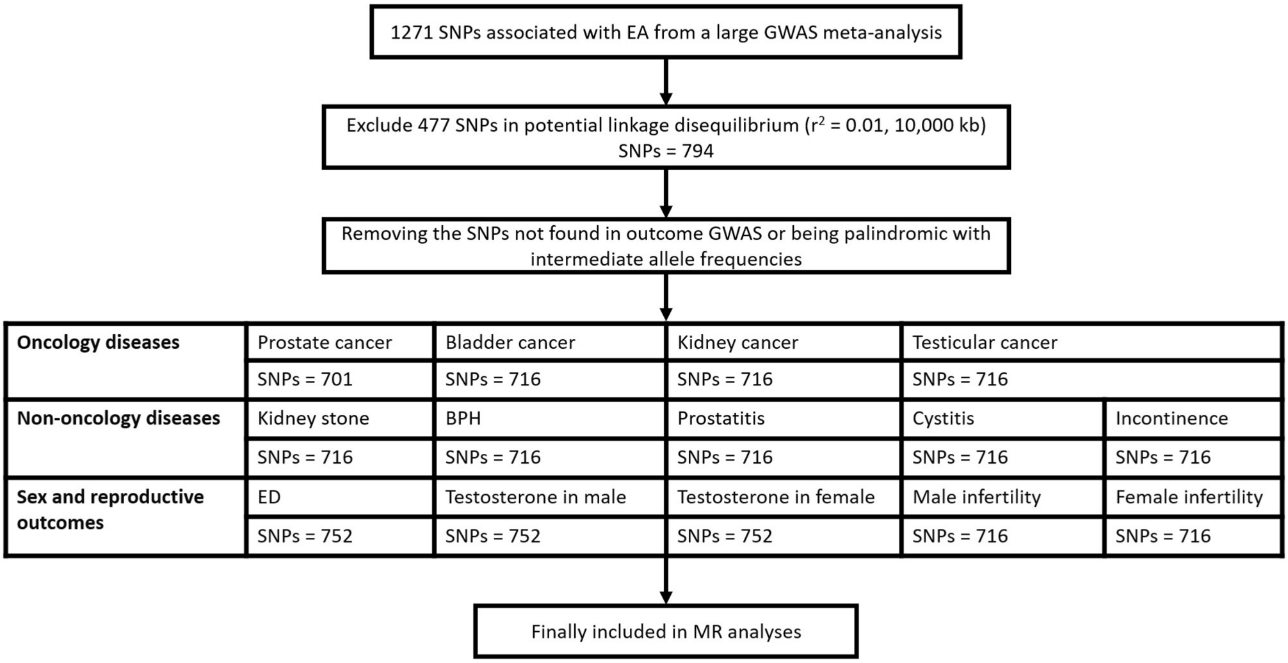
For the four oncological diseases, the primary analysis using IVW suggested that genetically predicted one SD increase in EA was causally correlated with a higher risk of prostate cancer [odds ratio (OR) 1.14, 95% confidence interval (CI) 1.05–1.25, $P = 0.003$], while no causal effect was observed for bladder cancer (OR 0.85, 95% CI 0.62–1.18, $P = 0.347$), kidney cancer (OR 0.73, 95% CI 0.52–1.04, $P = 0.080$), and testicular cancer (OR 1.55, 95% CI 0.71–3.38, $P = 0.270$) (**Figure 2**). However, not all the sensitivity analyses supported the causation between EA and prostate cancer (**Supplementary Table 3**).

In terms of the five non-oncological diseases, the results from IVW showed that genetically predicted one SD increase in EA was correlated with a decreased risk of kidney stone (OR 0.73, 95% CI 0.62–0.87, $P < 0.001$) and cystitis (OR 0.76, 95% CI 0.67–0.86, $P < 0.001$) after Bonferroni correction and suggestively correlated with a lower risk of prostatitis (OR 0.76, 95% CI 0.59–0.98, $P = 0.037$) and incontinence (OR 0.64, 95% CI 0.47–0.87, $P = 0.004$) (**Figure 2**). Most of the sensitivity analyses supported the causation between EA and them (**Supplementary Table 3**). For BPH (OR 0.92, 95% CI 0.81–1.05, $P = 0.233$), no causal relationship was found (**Figure 2**).

For the remaining five sexual and reproductive health outcomes, we found that genetically predicted one SD increase in EA was causally associated with a higher testosterone level in men (β 0.07, 95% CI 0.04–0.10, $P < 0.001$) and a lower level (β −0.13, 95% CI−0.16 to−0.11, $P < 0.001$) in women after Bonferroni correction (**Figure 3**). For infertility, the results from IVW estimates showed that genetically predicted one SD increase in EA was correlated with a lower risk of infertility in females (OR 0.74, 95% CI 0.64–0.86, $P < 0.001$), while no causal effect was observed in males (OR 0.79, 95% CI 0.52–1.20, $P = 0.269$). For ED (OR 1.00, 95% CI 0.86–1.15, $P = 0.961$), no causal relationship was observed (**Figure 2**). The detailed results of our MR study were presented in **Supplementary Table 3**.

DISCUSSION

In the current research, we investigated the causal effects of EA on 14 urological and reproductive health outcomes. Our findings suggested that genetically determined increased EA was correlated with a higher prostate cancer risk and a lower risk of kidney stone, prostatitis, cystitis, and incontinence. For the bioavailable testosterone levels and infertility, sex-specific associations were observed, with genetically determined increased EA being related to higher levels of testosterone in men and lower levels of testosterone in women, and correlated with a


1271 SNPs associated with EA from a large GWAS meta-analysis
Exclude 477 SNPs in potential linkage disequilibrium (r² = 0.01, 10,000 kb)
SNPs = 794
Removing the SNPs not found in outcome GWAS or being palindromic with intermediate allele frequencies
Finally included in MR analyses

Oncology diseases	Prostate cancer	Bladder cancer	Kidney cancer	Testicular cancer	
	SNPs = 701	SNPs = 716	SNPs = 716	SNPs = 716	
Non-oncology diseases	Kidney stone	BPH	Prostatitis	Cystitis	Incontinence
	SNPs = 716	SNPs = 716	SNPs = 716	SNPs = 716	SNPs = 716
Sex and reproductive outcomes	ED	Testosterone in male	Testosterone in female	Male infertility	Female infertility
	SNPs = 752	SNPs = 752	SNPs = 752	SNPs = 716	SNPs = 716

FIGURE 1 | Flow diagram for eligible SNPs selection for each of the 14 outcomes. SNP, single nucleotide polymorphism; EA, educational attainment; GWAS, genome wide association study; BPH, benign prostatic hyperplasia; ED, erectile dysfunction.

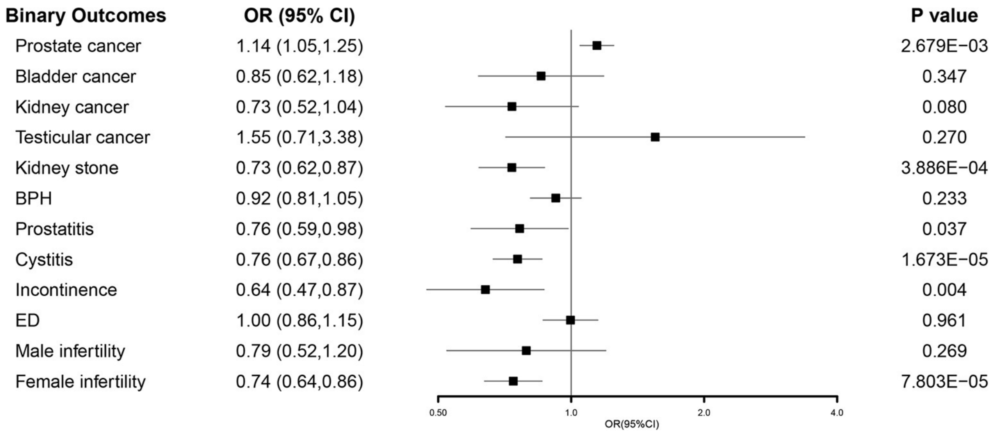


FIGURE 2 | Forest plot of the MR results between EA and 12 binary outcomes including prostate cancer, bladder cancer, kidney cancer, testicular cancer, kidney stone, BPH, prostatitis, cystitis, incontinence, ED, male infertility, and female infertility. MR, Mendelian randomization; EA, educational attainment; BPH, benign prostatic hyperplasia; ED, erectile dysfunction; OR, odds ratio; CI, confidence interval.

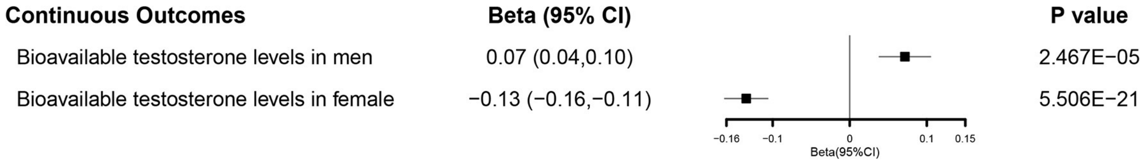


FIGURE 3 | Forest plot of the MR results between EA and two continuous outcomes including bioavailable testosterone levels in male and female. MR, Mendelian randomization; EA, educational attainment; CI, confidence interval.

decreased risk of infertility in women but was not related to male infertility. In terms of kidney cancer, bladder cancer, testicular cancer, BPH, and ED, no causal effects were observed.

Prostate cancer is a common malignancy worldwide. Prior observational studies have reported that populations with higher EA were at a higher risk of prostate cancer (3, 4), which was in accordance with our finding. Apart from the higher diagnostic activity among the well-educated population (4), another possible explanation for this correlation might be that people with higher EA commonly report higher fat consumption and less physical activity (31), thus increasing the risk of prostate cancer. However, there are few relevant studies, and the underlying mechanism still needs further research.

Kidney stone is another common urological disease. In a prior study, by analyzing the education levels and 24-h urine composition of 435 kidney stone patients, they found that a decreasing level of education was correlated with increased urine calcium, supersaturation of calcium oxalate, and supersaturation of calcium phosphate (9), thus appearing to increase kidney stone formation. However, a significant limitation of this study was that they only enrolled patients with stone formation. Thus, their results might not be generalizable to those without a history of nephrolithiasis. While in our MR analyses, using the GWAS data from 4,969 cases and 213,445 controls, our results were more reliable and generalizable to the general population. In addition to kidney stone incidence risk, EA has also been correlated with the degree of stone burden. In a retrospective study conducted by Bayne (8), after analyzing the socioeconomic and clinical data of 650 patients, they found that a lower education level was correlated with an increased stone burden >2 cm. One possible explanation for the association between EA and kidney stone might be *Oxalobacter formigenes*. Increasing evidence has revealed that *Oxalobacter formigenes* are essential in regulating oxalate homeostasis, with effects that inhibit calcium oxalate stone formation. Researchers found that education level, especially for education levels lower than high school, was associated with an abundance of *Oxalobacter formigenes* after analyzing over 8,000 American Gut Project fecal samples (32).

We also observed that increased EA was correlated with a lower risk of prostatitis and cystitis, which has rarely been reported before. Although the exact underlying mechanism was unknown, one possible explanation was that the population with higher EA are less likely to smoke and more likely to participate in physical activity and have better health habits (33), thus decreasing the risk of prostatitis and cystitis. Regarding incontinence, similar to previous observational studies (12, 13), we found that the population with higher EA were at a lower risk of incontinence. A possible reason could be that individuals with higher EA are inclined to pay more attention to their health and are more willing to take preventive measures to maintain their good health and decrease the risk of incontinence. In contrast, those with lower EA usually perform labor intensive work, which has been regarded as a risk factor for incontinence (34).

In terms of the sexual and reproductive health outcomes, sex-specific associations were observed. We observed that increased EA was causally related to higher levels of testosterone in men and lower levels of testosterone in women, which was partially consistent with prior findings (18). The presence of this association suggested that EA might affect the homeostatic setpoints by which typical hormone concentrations are maintained (18), but the reason for this sex-specific association needs further research. We also observed that increased EA was related to a lower risk of female infertility but was not related to male infertility. Previous studies on the correlation between EA and female infertility have yielded inconsistent results. Two studies (16, 17) reported that EA was inversely associated with female infertility, while one study reported a positive relationship (15). In the current research, we added new evidence to the inverse association between EA and female infertility. This might be because women with higher EA usually have healthier lifestyles and better curative care (16). For male infertility, no causal effect was observed, which was consistent with a prior observational study (15).

For bladder cancer, kidney cancer, and testicular cancer, similar to prior observational studies (5–7), we observed no causality in our study. The role of EA in determining BPH is inconsistent (10, 11), with one study that reported the population with higher EA were at a higher risk (11) and another reporting a lower risk (10). However, all these observational studies are prone to confounding factors, and the results from our MR study indicated that EA might not have a causal effect on BPH. In terms of ED, although a prior observational study reported that increased EA was correlated with a lower risk of ED (14), no causal relationship was observed in the current MR study.

Our study has several strengths. First, as far as we know, this is the first MR study to explore the causal association between EA and urological or reproductive health outcomes. Compared with other observational studies, our research is less vulnerable to confounding factors. Second, all the included individuals within the GWAS were of European-descent, making the potential bias from population stratification minimal. In addition, a total of 14 outcomes were analyzed in our study, which is comprehensive and informative. Nevertheless, our study could not avoid limitations. First, since the large percentage of individuals in the EA exposure GWAS is from the UK Biobank, we extracted GWAS data for 10 outcomes from the FinnGen project to avoid overlap as much as possible, but this also leads to a disadvantage since the FinnGen project is prepublication and the data quality might weaken slightly. However, quality control has already been applied to the FinnGen project, and detailed information can be found on their official site (26). Second, although directional pleiotropy was not detected in our study, heterogeneity was found for part of our results, leading to some potential biases. Third, EA might also correlate with some other factors, such as intelligence, income, testosterone levels (35), which might mediate the effects of EA on the 14 included urological and reproductive health outcomes. However, whether these factors play a mediating role between EA and these 14 outcomes was not included in the primary aim of our study and should be explored in future research.

CONCLUSIONS

Our findings indicated that genetically determined increased EA was correlated with a higher risk of prostate cancer and a lower risk of kidney stone, prostatitis, cystitis, and incontinence. For the bioavailable testosterone levels and infertility, sex-specific associations were observed, with genetically determined increased EA being related to higher levels of testosterone in men, decreased levels of testosterone in women and a lower risk of infertility in women but was not related to male infertility. In terms of kidney cancer, bladder cancer, testicular cancer, BPH, and ED, no causal effects were observed. All of these results indicate that EA plays a vital role in urological diseases, especially in non-oncological outcomes and reproductive health. Further research is needed to examine these findings.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study and can be accessed via the references we used.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

MW, ZJ, XG, and KW: conceptualization. MW, ZJ, and XG: methodology. CY and XJ: data curation. MW, ZJ, XG, and CY: writing—original draft preparation. XJ, HL, and KW: writing—review and editing. ZJ, HL, and KW: funding acquisition. All authors have read and agreed to the published version of the manuscript.

FUNDING

This research was funded by the 1·3·5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (ZYGD18011 and ZYJC18015) and Post-Doctor Research Project, West China Hospital, Sichuan University (2020HXBH016).

ACKNOWLEDGMENTS

We want to acknowledge the participants and investigators of the FinnGen study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.742952/full#supplementary-material

REFERENCES

1. Krueger PM, Dehry IA, Chang VW. The economic value of education for longer lives and reduced disability. *Milbank Q.* (2019) 97:48–73. doi: 10.1111/1468-0009.12372
2. Cao M, Cui B. Association of educational attainment with adiposity, type 2 diabetes, and coronary artery diseases: a mendelian randomization study. *Front Public Health.* (2020) 8:112. doi: 10.3389/fpubh.2020.00112
3. Vidarsdottir H, Gunnarsdottir HK, Olafsdottir EJ, Olafsdottir GH, Pukkala E, Tryggvadottir L. Cancer risk by education in Iceland; a census-based cohort study. *Acta Oncol.* (2008) 47:385–90. doi: 10.1080/02841860801888773
4. Pudrovska T, Anishkin A. Clarifying the positive association between education and prostate cancer: a Monte Carlo simulation approach. *J Appl Gerontol.* (2015) 34:293–316. doi: 10.1177/0733464812473798
5. Everatt R, Kuzmickiene I, Virviciute D, Tamosiunas A. Cigarette smoking, educational level and total and site-specific cancer: a cohort study in men in Lithuania. *Eur J Cancer Prev.* (2014) 23:579–86. doi: 10.1097/CEJ.0000000000000018
6. Mouw T, Koster A, Wright ME, Blank MM, Moore SC, Hollenbeck A, et al. Education and risk of cancer in a large cohort of men and women in the United States. *PLoS One.* (2008) 3:e3639. doi: 10.1371/journal.pone.0003639
7. Marsa K, Johnsen NF, Bidstrup PE, Johannesen-Henry CT, Friis S. Social inequality and incidence of and survival from male genital cancer in a population-based study in Denmark, 1994–2003. *Eur J Cancer.* (2008) 44:2018–29. doi: 10.1016/j.ejca.2008.06.012
8. Bayne DB, Usawachintachit M, Armas-Phan M, Tzou DT, Wiener S, Brown TT, et al. Influence of socioeconomic factors on stone burden at presentation to tertiary referral center: data from the registry for stones of the kidney and ureter. *Urology.* (2019) 131:57–63. doi: 10.1016/j.urology.2019.05.009
9. Eisner BH, Sheth S, Dretler SP, Herrick B, Pais VM. Effect of socioeconomic status on 24-hour urine composition in patients with nephrolithiasis. *Urology.* (2012) 80:43–7. doi: 10.1016/j.urology.2011.12.017
10. Park MB, Hyun DS, Song JM, Chung HC, Kwon SW, Kim SC, et al. Association between the symptoms of benign prostatic hyperplasia and social disparities: Does social capital promote prostate health? *Andrologia.* (2018) 50:e13125. doi: 10.1111/and.13125
11. Fowke JH, Murff HJ, Signorello LB, Lund L, Blot WJ. Race and socioeconomic status are independently associated with benign prostatic hyperplasia. *J Urol.* (2008) 180:2091–6. doi: 10.1016/j.juro.2008.07.059
12. Ge J, Yang P, Zhang Y, Li X, Wang Q, Lu Y. Prevalence and risk factors of urinary incontinence in Chinese women: a population-based study. *Asia Pac J Public Health.* (2015) 27:NP1118–31. doi: 10.1177/1010539511429370
13. Lin YF, Lin YC, Wu IC, Chang YH. Urinary incontinence and its association with socioeconomic status among middle-aged and older persons in Taiwan: a population-based study. *Geriatr Gerontol Int.* (2021) 21:245–53. doi: 10.1111/ggi.14115
14. Selvin E, Burnett AL, Platz EA. Prevalence and risk factors for erectile dysfunction in the US. *Am J Med.* (2007) 120:151–7. doi: 10.1016/j.amjmed.2006.06.010
15. Datta J, Palmer MJ, Tanton C, Gibson LJ, Jones KG, Macdowall W, et al. Prevalence of infertility and help seeking among 15 000 women and men. *Hum Reprod.* (2016) 31:2108–18. doi: 10.1093/humrep/dew123
16. Zhou Z, Zheng D, Wu H, Li R, Xu S, Kang Y, et al. Epidemiology of infertility in China: a population-based study. *BJOG.* (2018) 125:432–41. doi: 10.1111/1471-0528.14966
17. Safarinejad MR. Infertility among couples in a population-based study in Iran: prevalence and associated risk factors. *Int J Androl.* (2008) 31:303–14. doi: 10.1111/j.1365-2605.2007.00764.x
18. Bann D, Hardy R, Cooper R, Lashen H, Keevil B, Wu FC, et al. Socioeconomic conditions across life related to multiple measures of the endocrine system in older adults: Longitudinal findings from a British birth cohort study. *Soc Sci Med.* (2015) 147:190–9. doi: 10.1016/j.socscimed.2015.11.001
19. Minica CC, Boomsma DI, Dolan CV, de Geus E, Neale MC. Empirical comparisons of multiple Mendelian randomization approaches in the presence of assortative mating. *Int J Epidemiol.* (2020) 49:1185–93. doi: 10.1093/ije/dyaa013
20. Lee JJ, Wedow R, Okbay A, Kong E, Maghzian O, Zacher M, et al. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 11 million individuals. *Nat Genet.* (2018) 50:1112–21. doi: 10.1038/s41588-018-0147-3
21. Gill D, Karhunen V, Malik R, Dichgans M, Sofat N. Cardiometabolic traits mediating the effect of education on osteoarthritis risk: a Mendelian randomization study. *Osteoarthritis Cartilage.* (2021) 29:365–71. doi: 10.1016/j.joca.2020.12.015
22. Liang J, Cai H, Liang G, Liu Z, Fang L, Zhu B, et al. Educational attainment protects against type 2 diabetes independently of cognitive performance: a Mendelian randomization study. *Acta Diabetol.* (2021) 58:567–74. doi: 10.1007/s00592-020-01647-w
23. Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, et al. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. *Nat Genet.* (2018) 50:928–36. doi: 10.1038/s41588-018-0142-8
24. Ruth KS, Day FR, Tyrrell J, Thompson DJ, Wood AR, Mahajan A, et al. Using human genetics to understand the disease impacts of testosterone in men and women. *Nat Med.* (2020) 26:252–8. doi: 10.1038/s41591-020-0751-5
25. Bovijn J, Jackson L, Censin J, Chen CY, Laisk T, Laber S, et al. GWAS identifies risk locus for erectile dysfunction and implicates hypothalamic neurobiology and diabetes in etiology. *Am J Hum Genet.* (2019) 104:157–63. doi: 10.1016/j.ajhg.2018.11.004
26. FinnGen. *Documentation of R5 Release* (2021). Available online at: https://finngen.gitbook.io/documentation/
27. Yuan S, Larsson SC. Assessing causal associations of obesity and diabetes with kidney stones using Mendelian randomization analysis. *Mol Genet Metab.* (2021) S1096-7192(21)00774-5. doi: 10.1016/j.ymgme.2021.08.010
28. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol.* (2015) 44:512–25. doi: 10.1093/ije/dyv080
29. Hartwig FP, Davey Smith G, Bowden J. Robust inference in summary data Mendelian randomization *via* the zero modal pleiotropy assumption. *Int J Epidemiol.* (2017) 46:1985–98. doi: 10.1093/ije/dyx102
30. Fu Y, Xu F, Jiang L, Miao Z, Liang X, Yang J, et al. Circulating vitamin C concentration and risk of cancers: a Mendelian randomization study. *BMC Med.* (2021) 19:171. doi: 10.1186/s12916-021-02041-1
31. Allen L, Williams J, Townsend N, Mikkelsen B, Roberts N, Foster C, et al. Socioeconomic status and non-communicable disease behavioural risk factors in low-income and lower-middle-income countries: a systematic review. *Lancet Glob Health.* (2017) 5:e277–89. doi: 10.1016/S2214-109X(17)30058-X
32. Liu M, Koh H, Kurtz ZD, Battaglia T, PeBenito A, Li H, et al. Oxalobacter formigenes-associated host features and microbial community structures examined using the American Gut Project. *Microbiome.* (2017) 5:108. doi: 10.1186/s40168-017-0316-0
33. Lawrence EM. Why do college graduates behave more healthfully than those who are less educated? *J Health Soc Behav.* (2017) 58:291–306. doi: 10.1177/0022146517715671
34. Walker GJ, Gunasekera P. Pelvic organ prolapse and incontinence in developing countries: review of prevalence and risk factors. *Int Urogynecol J.* (2011) 22:127–35. doi: 10.1007/s00192-010-1215-0

35. Davies NM, Hill WD, Anderson EL, Sanderson E, Deary IJ, Davey Smith G. Multivariable two-sample Mendelian randomization estimates of the effects of intelligence and education on health. *Elife.* (2019) 8:43990. doi: 10.7554/eLife.43990

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Public Health

ORIGINAL RESEARCH
published: 06 December 2021
doi: 10.3389/fpubh.2021.743003

Differential Impact of COVID-19 Risk Factors on Ethnicities in the United States

Prashant Athavale [1]†, Vijay Kumar [1]†, Jeremy Clark [1], Sumona Mondal [1] and Shantanu Sur [2]**

[1] Department of Mathematics, Clarkson University, Potsdam, NY, United States, [2] Department of Biology, Clarkson University, Potsdam, NY, United States

OPEN ACCESS

Edited by:
Jie Hu,
The Ohio State University,
United States

Reviewed by:
Samuel Manda,
South African Medical Research Council, South Africa
Shivani Sharma,
University of Hertfordshire,
United Kingdom

Correspondence:
Sumona Mondal
smondal@clarkson.edu
Shantanu Sur
ssur@clarkson.edu

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Life-Course Epidemiology and Social Inequalities in Health, a section of the journal Frontiers in Public Health

Received: *17 July 2021*
Accepted: *10 November 2021*
Published: *06 December 2021*

Citation:
Athavale P, Kumar V, Clark J, Mondal S and Sur S (2021) Differential Impact of COVID-19 Risk Factors on Ethnicities in the United States. Front. Public Health 9:743003. doi: 10.3389/fpubh.2021.743003

The coronavirus disease (COVID-19) has revealed existing health inequalities in racial and ethnic minority groups in the US. This work investigates and quantifies the non-uniform effects of geographical location and other known risk factors on various ethnic groups during the COVID-19 pandemic at a national level. To quantify the geographical impact on various ethnic groups, we grouped all the states of the US. into four different regions (Northeast, Midwest, South, and West) and considered Non-Hispanic White (NHW), Non-Hispanic Black (NHB), Hispanic, Non-Hispanic Asian (NHA) as ethnic groups of our interest. Our analysis showed that infection and mortality among NHB and Hispanics are considerably higher than NHW. In particular, the COVID-19 infection rate in the Hispanic community was significantly higher than their population share, a phenomenon we observed across all regions in the US but is most prominent in the West. To gauge the differential impact of comorbidities on different ethnicities, we performed cross-sectional regression analyses of statewide data for COVID-19 infection and mortality for each ethnic group using advanced age, poverty, obesity, hypertension, cardiovascular disease, and diabetes as risk factors. After removing the risk factors causing multicollinearity, poverty emerged as one of the independent risk factors in explaining mortality rates in NHW, NHB, and Hispanic communities. Moreover, for NHW and NHB groups, we found that obesity encapsulated the effect of several other comorbidities such as advanced age, hypertension, and cardiovascular disease. At the same time, advanced age was the most robust predictor of mortality in the Hispanic group. Our study quantifies the unique impact of various risk factors on different ethnic groups, explaining the ethnicity-specific differences observed in the COVID-19 pandemic. The findings could provide insight into focused public health strategies and interventions.

Keywords: COVID-19, infection, mortality, ethnicity, Hispanic, risk factors, diabetes

1. INTRODUCTION

Numerous researchers have found various comorbidities and other risk factors affecting the spread and prognosis of coronavirus disease (COVID-19). Recent work by many researchers has also demonstrated that the COVID-19 pandemic has affected marginalized ethnicities more severely. We thus hypothesize that the risk factors for COVID-19 must have affected different ethnic groups in a distinctive manner. In this paper, we aim to quantify the differential effect of risk factors on different ethnicities.

1.1. COVID-19 and Ethnicity

The public health crisis created by the COVID-19 has uncovered the historical inequalities (1–4) between ethnic groups in certain countries, in particular in the UK and US, which are countries with ethnically diverse populations. These observations and consistent fatal outcomes in the minority ethnic groups (5, 6) have led to speculations about why patients from these groups are susceptible to infections, followed by severe complications. These trends could be due to different rates of COVID-19 infections, underlying health conditions, living conditions including housing density, having jobs as essential workers, access to health care, quality of care, and a mixture of multiple factors among these groups. The United States national data (7) from states and municipalities reports disproportionate COVID-19 infections, hospitalizations, and deaths among minority ethnic groups. Dobin and Dobin (8) showed that the infection rate is 4-fold for the Black and Hispanic population in selected counties in New York state. Moore et al. (9) observe a disproportionate number of COVID-19 cases among underrepresented racial/ethnic groups in the United States. Adhikari et al. (10) show that the racial and ethnic disparities in COVID-19 infections and deaths existed beyond those explained by income inequality.

1.2. Effects of Geographical Location on COVID-19

The impact of COVID-19 varies widely across countries and even within a country or a region. For example, Sun et al. (11) showed a negative correlation between the number provincial COVID-19 cases and latitude, as well as altitude. Breen and Ermisch (12) use spatial autoregressive regression to show that the relation of COVID-19 mortality to social composition of geographical areas in England is distinct than that of non-COVID mortality. A number of factors including societal awareness and culture, public health measures, healthcare infrastructure, and more recently vaccination coverage are known underlie the variation for COVID-19 infection rates and adverse health outcome (13).

Although multiple studies have confirmed that black and Hispanic populations in the US are more vulnerable to COVID-19, to our knowledge, no data is available if they are equally susceptible across geographical locations. Stephens-Davidowitz (14) uses the search data from Google to show that there exists a wide variation in racism in the US within the 50 states. Thus, we surmise that the impact of the COVID-19 on various ethnicities may not be uniform in all regions of the US. Hence, we are interested in understanding if a geographical location plays a part in the variation of COVID-19 impact on minorities.

1.3. Comorbidities for COVID-19

Emerging evidence highlights that comorbid conditions such as obesity, cardiovascular disease (CVD), and type 2 diabetes are directly linked to the severity of the COVID-19 disease (15–17). A meta-analysis including 76,993 patients with COVID-19 showed diabetes, CVD, smoking, malignancy, chronic kidney disease, hypertension, chronic obstructive pulmonary disease (COPD) are associated with poor prognosis (18). This conclusion was further supported by Richardson et al. (19), and Sun et al. (20). Using logistic regression (21) show that obesity was a risk factor for the severity of the COVID-19 disease. Furthermore, in a retrospective cohort study, Busetto et al. (22) conclude that despite their young age, overweight patients were more likely to need assisted ventilation and access to intensive care units than patients with normal weight. The connection between obesity and pulmonary function is well-established, e.g., Sharp et al. (23) observe that obese patients have significantly decreased total respiratory compliance. Moreover, Li et al. (24) find that reduction in functional residual capacity and diffusion impairment are the most common abnormalities in obese patients. Yan et al. (18) show that diabetic patients experienced more mortality than non-diabetic patients. Finally, just as in the case of the SARS epidemic (25), COVID-19 has disproportionately affected the older population (26). In fact, in the US, 92% of the COVID-19 recorded deaths till June, 2020 are in the age group 55 years and above (27). In summary, the main comorbidities for COVID-19 include obesity, diabetes, advanced age, hypertension, and cardiovascular disease.

However, these risk factors affect different ethnicities differently. For example, Paeratakul et al. (28) find that among obese individuals, the prevalence of hypertension was higher in NHB subjects than other groups. Sturm and Hattori (29) observe that the prevalence of obesity is about double among NHB than among Hispanics or NHW. Kuzawa and Sweet (30) note that NHB suffer from a disproportionate burden of CVD relative to NHW. Thus, we are motivated to understand whether or not these comorbidities affect different ethnicities differently.

1.4. Impact of Poverty on COVID-19 Prognosis

Patel et al. (31) note that economically disadvantaged people are vulnerable to COVID-19 due to a combination of factors. A time-series analysis conducted by Elgar et al. (32) reveals that income inequality is associated with a higher number of deaths due to COVID-19 in 84 countries. In particular, in the US, the states with higher income inequality experienced a higher rate of infection as well as the number of COVID-19 related deaths (33). This pattern could be because the comorbidities associated with COVID-19 are linked to poverty.

A longitudinal study involving 600,662 adults from Taiwan's National Healthcare Insurance database indicates that diabetes incidence is associated with poverty (34). This finding is particularly notable since the subjects from this study had access to universal healthcare. However, the subjects were from a ethnically homogeneous population. Thus, we are motivated to investigate the differential role of poverty among various races.

1.5. Objectives

For this study, we choose Non-Hispanic white (NHW), Non-Hispanic Black (NHB), Hispanic, and Non-Hispanic Asian (NHA) as four ethnic groups. The risk factors we choose to focus on in this work are advanced age, obesity, cardiovascular disease, diabetes, hypertension, and poverty. We aim to investigate the following in this study:

1. Does a geographical location have different impact on COVID-19 infection and mortality rates for different ethnicities?
2. Do various COVID-19 risk factors have a different effect on different ethnicities?

To this end, we collected COVID-19 related infections and mortality data from various publicly available sources, as described in section 2.1. To investigate the geographical variation in the impact of COVID-19, we seek to quantify the difference between the distributions of infection rates, mortality rates, and populations of various ethnic groups in four geographical regions of the US (35). We describe our approach for this analysis in section 2.4.1. Finally, in section 2.4.3 we construct robust linear models with infection and mortality rates of various ethnicities as response variables.

2. MATERIALS AND METHODS

2.1. Selection of Variables and Data Sources

Given the discussion in the sections 1.3, 1.1, and 1.4, we focus on the following factors in this work: obesity, diabetes, poverty, advanced age, hypertension, and cardiovascular diseases. We included the following ethnic groups in our work: NHW, NHB, Hispanic, and NHA. We excluded the groups American Indian/Alaska Native, and Native Hawaiian/Other Pacific Islander from this work due to lack of reliable and consistent data sets in the US, cf. (36–38). We used the American Community Survey (39) census database to collect the following information for each state: NHW, NHB, Hispanic, and NHA population and their respective percentage contribution to the population of each state. To investigate the geographical effect on COVID-19 prognosis, we used the classification of the US into the following categories: Northeast, South, Midwest, and West (35). COVID-19 infection and death counts between January 21, 2020, to September 30, 2020 from KFF Covid-19 data (40) were obtained. The time window roughly corresponds to the first pandemic wave experienced in the US. For each ethnicity **E**, and for a specific state **S** the variable `Relative Infection %` is defined as follow:

$$\texttt{Relative Infection \%} := \frac{\text{COVID-19 positive subjects of ethnicity } \mathbf{E} \text{ in state } \mathbf{S}}{\text{Total number of COVID-19 positive subjects in state } \mathbf{S}} \times 100.$$

We note that the choice of "relative" infection percentage as a response variable is deliberate. This choice allows us to directly compare this number with the population share of that ethnicity in the region. Similarly, for each ethnicity **E**, and for a specific state **S** we define the variable `Relative Mortality %` as follow:

$$\texttt{Relative Mortality \%} := \frac{\text{COVID-19 related deaths for ethnicity } \mathbf{E} \text{ in state } \mathbf{S}}{\text{Total number of COVID-19 deaths in state } \mathbf{S}} \times 100.$$

In our work, `Relative Infections %` and `Relative Mortality %` were considered as the response variables. For brevity, we write infection rate instead of `Relative Infections %`, and so on. The use of relative percentages allows a direct comparison with the population percentages of that ethnicity. For example, in a state with a 5% NHB population, relative mortality of 15% in the NHB community indicates disproportionately large mortality compared to the NHB population. The use of "relative" percentage is independent of the population of the state itself. The use of this measure also allows us to compare the impact on a certain ethnic group in two states with similar proportion of the minority population. As a concrete example, when we consider the states of California and Texas, both have a similar percentage of the Hispanic population, 39.5 and 40%, respectively. However, the relative mortality percentages for the Hispanic group in California and Texas are 48.3 and 56.1%, respectively. We collected the data on the percentage of people with age 60 or more in each state and ethnicity is obtained from the CDC dataset (41). Race and state-wise data were obtained from adults who reported being told by a health professional that they have diabetes (excluding prediabetes and gestational diabetes) using the America's Health Rankings (42). We used the body mass index (BMI) as a measure of obesity following (43) and define obesity as a condition of having a BMI of 30.0 or higher. The dataset (44) were used to obtain the obesity data from each state and for the races NHW, NHB, Hispanic, NHA. We acquired the percentage of adults whom a health care professional informed that they had a coronary heart disease, or myocardial infarction, or a stroke from AHR CVD data (45). This was gathered for each state and ethnicity of interest. We obtained the race and state-wise data on adults who reported being informed by a health professional that they have high blood pressure from AHR HBP data (46). The US Census Bureau defines the "poverty threshold" for a family with two adults and one child as $20,578 in 2019. We extracted the data from KFF Poverty data (47) on poverty defined by the "poverty threshold." We obtained this data for each state and ethnicity of interest.

For each state, and each of the four ethnicity of interest (NHW, NHB, Hispanic, and NHA) we defined the variables: `Age60+`, `BMI30+` (a measure of obesity), `CVD`, `Diabetes`, `HBP`, `Poverty`. For a state **S**, and an ethnicity **E** we defined the relative percentage of people with age 60 or over `Age60+` as follows:

$$\texttt{Relative Age60+ \%} := \frac{\text{Number of people of ethnicity } \mathbf{E} \text{ with age over 60 in state } \mathbf{S}}{\text{Total number of people with age over 60 in state } \mathbf{S}} \times 100.$$

We use the variable name `Age60+` instead of `Relative Age60+ %` for conciseness, and so on. We define the relative percentage variables `Obesity`, `CVD`, `Diabetes`, `HBP`, and `Poverty` in a similar manner.

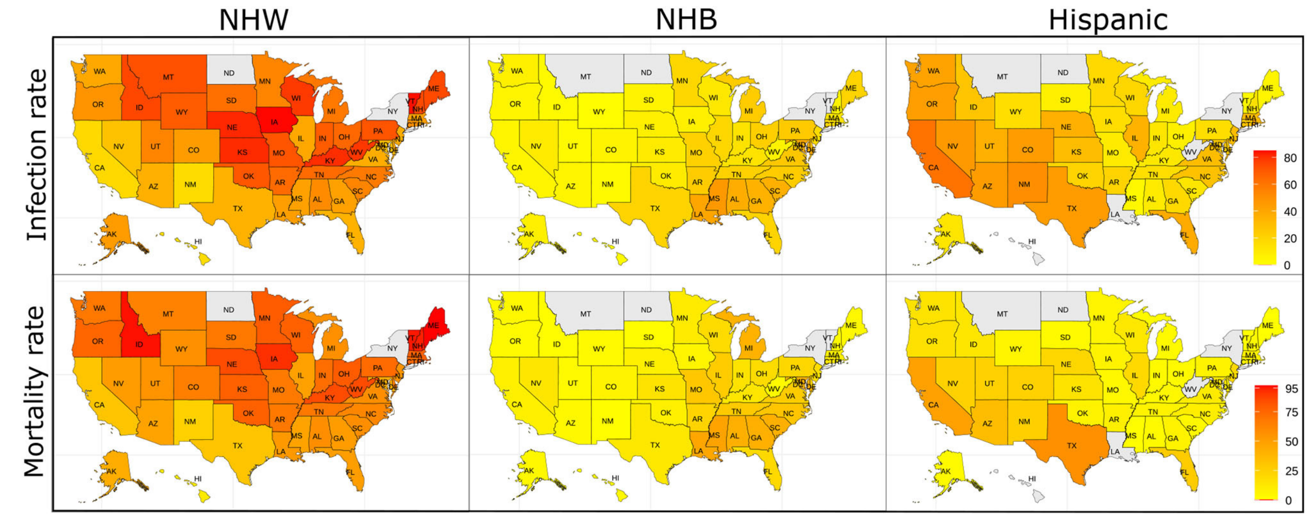


FIGURE 1 | The relative infection % and relative mortality % amongst the NHW, NHB, and Hispanic groups across the US. The states with no color indicate that the ethnicity-wise data on relative infection % and relative mortality % were not available in those states.

2.2. A Note on Unavailability of Data From Some States

We encountered a few irregularities during our data collection process in the format the data was made available by various states (36–38). For example, New York state does not provide ethnicity-wise COVID-19 infection and mortality data. Similarly, infection data for NHW communities only are available for North Dakota. Data availability of specific variables for different ethnicities also varied across states. Data for all variables for the NHW group could be obtained from 48 states whereas for the NHB and Hispanic communities such data were available from only 38 and 33 states, respectively. In our analysis, we included the states for which data for all the variables are available. Hatcher et al. (37) find that only 23 state in the US have complete data for American Indian and Alaska Native Persons.

2.3. Description of Data

The data for this study are state-level demographics based on four ethnic groups. We depict the relative infection % and relative mortality % for NHW, NHB, and Hispanic group in the map in **Figure 1**. As described in section 2.2 some states do not make the ethnicity-wise data public. The states with no color in the **Figure 1** indicate that the ethnicity-wise infection and mortality data was not available in those states. We calculated state-wise descriptive statistics for the relative infection and mortality percentages and population comparing each ethnic group. We performed a descriptive analysis to explore the region-specific, state-wise characteristics of for the relative infection % and mortality % and population by calculating their medians, first and third quartiles, and presented in **Figure 2**.

2.4. Analytical Approach

2.4.1. Quantifying the Regional Variability of COVID-19 on Various Races

The infection and mortality rates for various ethnic groups are disproportionate to their share of the population in the US (7, 8). We aim to understand this phenomenon and its severity across various regions in the US. To this effect, we employed the Kruskal-Wallis (KW) test (48), a non-parametric equivalent of the one-way analysis of variance. Since the test does not identify the groups that differ in their distributions, we followed it with Dunn's multiple comparisons test (49) for cases for which the KW test yielded statistically significant results. We used the combination of KW test, and Dunn's comparison test for the groups NHW, NHB, Hispanic, and NHA separately for all four regions of the US, as well as the whole country.

2.4.2. Correlation Analyses

In order to quantify the association between the impact of COVID-19 on the ethnic groups and the risk factors across the country, we consider each state, for which the data are available, as a data point. We computed the pairwise Pearson's correlation coefficients between various risk factors for the racial groups NHW, NHB, and Hispanic, along with their 2-tailed statistical significance values. We summarized the comparisons between the variables in correlation matrices.

2.4.3. Constructing Robust Linear Models With Infection and Mortality Rates as Response Variables

In order to elucidate the role of the explanatory variables on a specific aspect of the COVID-19 burden linear models are employed. For these linear models, we considered each state as a data point. From the **Figure 2**, we observe that the rate of infection and mortality in the NHA are consistently lower when compared to their population. Thus, we consider building

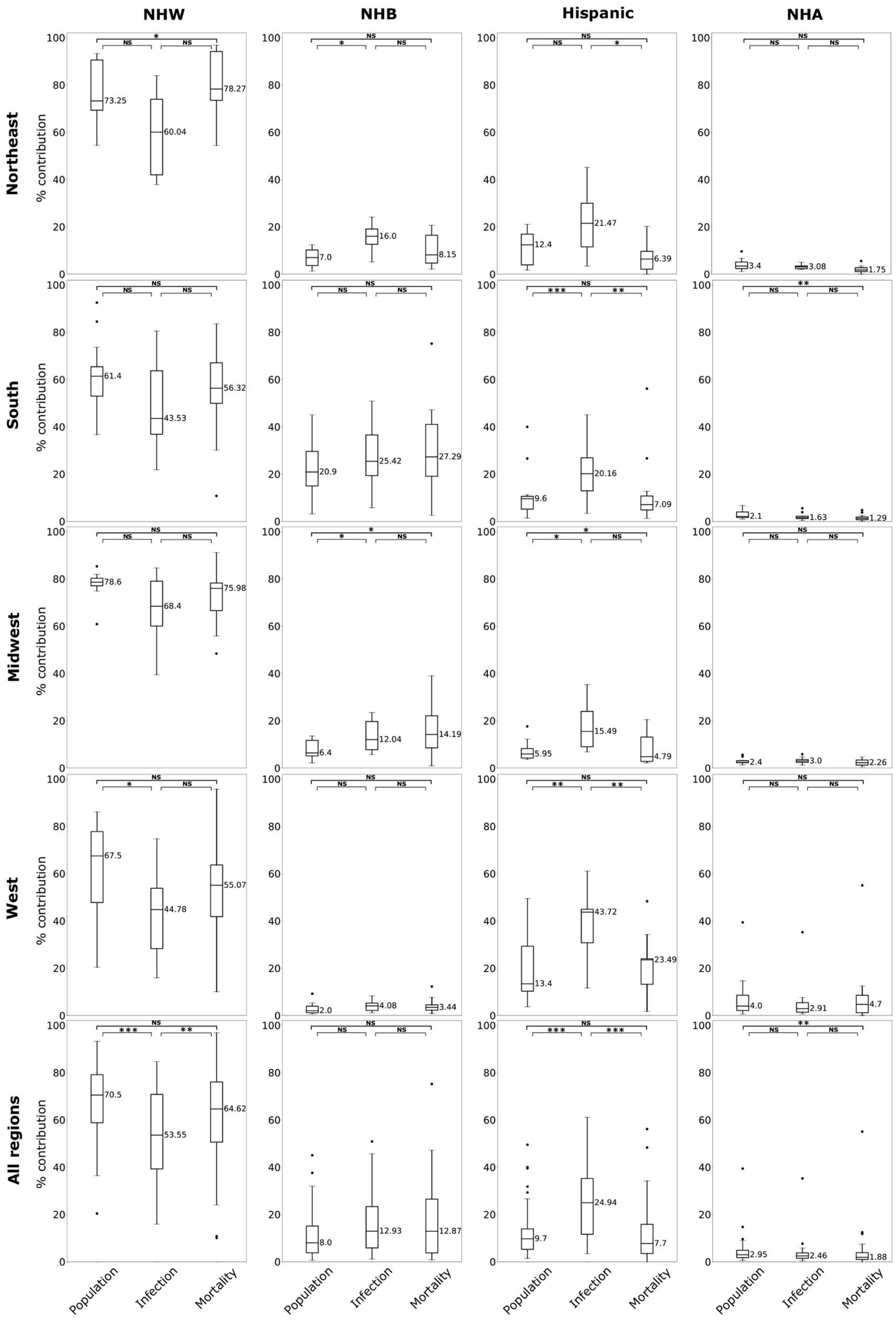


FIGURE 2 | Box plots of population, relative infection %, and relative mortality % in each of four US regions, and combining all regions for NHW, NHB, Hispanic and NHA groups. Horizontal bars represent medians. "*" significance at $p < 0.1$, "**" significance at $p < 0.05$, "***" significance at $p < 0.01$, NS, not significant (Kruskal-Wallis tests followed by Dunn's tests).

linear models for NHW, NHB, and Hispanic groups only to elucidate the contributions of risk factors considered in this study. However, infection and mortality rates, the response variables for our model, showed skewness in their distributions. Since logarithmic (log) transformation of data is one of the most commonly used techniques to conform to normality (50), we implemented it on infection and mortality rates. The log transformation was effective in correcting the skewness and introduce normality (**Supplementary Figures S1, S2**). As log transformation of infection and mortality rates improved their normality behavior, we used log transformed form of these variables exclusively for model construction. Thus, when we refer to infection rate and mortality rate in the context of linear models they denote log transformed infection rate and mortality rate, respectively. For NHW, NHB, and Hispanic groups, we built our preliminary linear models with infection rate, and mortality rate as our response variables and the risk factors defined in section 1.3, i.e., advanced age `Age60+`, `BMI30+` (a measure of obesity), `CVD`, `Diabetes`, `HBP`, `Poverty` as the explanatory variables.

However, conditions of advanced age, cardiovascular disease, diabetes, obesity, and hypertension are interrelated. This interrelation can also be observed from the correlation **Tables 3–5**. Multicollinearity among the explanatory variables can lead to unstable and unreliable estimates of regression coefficients (51). We used the variance inflation factor (VIF) to assess the multicollinearity between the explanatory variables (52). Following Kutner et al. (53, p. 409) an upper cut-off value of VIF for explanatory variables is set as 10 to minimize the contribution of multicollinearity in our model. Starting from the preliminary model for ethnicity of interest, we propose the procedure outlined below to construct our final model:

1. Compute the VIF for each explanatory variable in the model. If all the VIFs are less than 10, we declare this to be the final linear model.
2. If an explanatory variable has a VIF of more than 10, we remove the explanatory variable with the largest VIF. If there are more than one explanatory variables with VIF within 5% of the maximum VIF, we remove the variable that leads to a model with the highest adjusted R^2.
3. We construct the linear model with the remaining explanatory variables. After the removal of a variable, it is possible to include more data points in our model. For example, after removing the variable `Diabetes`, we could include states for which data on diabetes was not available.
4. Go to Step 2.4.3.

After constructing the linear models, we checked the normality of the residuals of the regression models with Lilliefors normality test (54).

2.5. Geographically Weighted Regression

Linear regression yields stationary and global regression coefficients. However, it is conceivable that these coefficients might have local variability. To find the geographical variability in the coefficients, we employed the geographically weighted regression (GWR) (55). Rather than producing global regression results, GWR yields "local" regression coefficients in terms of geographically varying functions. For our analysis, we used the infection rate and mortality rates as response variables and the variables obtained from section 2.4.3 as the explanatory variables.

2.6. Coding Language and Libraries Used

For our coding, we used R language (version 4.0.0), along with the following libraries in our coding: readxl, dplyr, tidyr, FSA, ggplot2, car, qqplotr, nortest, pwr, spgwr, sp, sf, rgdal, rgeos, tmap, tmaptools.

3. RESULTS

3.1. Regional Variation of COVID-19 Impact on Various Ethnicities

The boxplots in **Figure 2** summarize the relative impact of COVID-19 on various ethnicities across the four regions of the US and all regions as an aggregate. In **Figure 2** we present various descriptive statistics of the population, infection, and mortality rates for the NHW, NHB, Hispanic, and the NHA groups across various regions. As noted in the section 2.2, not all the states are included in the analyses. Thus, the statistics shown in this plot do not correspond closely to those of the whole country. We describe the Kruskal-Wallis test results in **Table 1**. We see in **Table 1** that the KW test for NHW is statistically significant in the Northeast and the West with $p < 0.1$. For the NHB group, the KW test is significant in the Northeast and the Midwest with $p < 0.1$. The KW test was statistically significant for the Hispanic group in "all four regions," with $p < 0.1$ in the Northeast; with $p < 0.05$ in Midwest and the West; and $p < 0.01$ in the South. The NHA data yielded significant results with the KW test only in the South with $p < 0.05$.

When we considered all four regions in the US together, the KW test was statistically significant for all ethnicities with $p < 0.01$ for NHW and Hispanic communities. The KW test was significant for the NHB and NHA when all regions were combined with $p < 0.1$.

We followed the significant KW tests with Dunn's multiple comparison test to identify factors differing in their distributions. We depict the results from the Dunn's test in **Table 2**. In particular, we obtained statistically significant results (with $p < 0.05$) in the South, Midwest, and West for the Hispanic population between the pairs 'infection & mortality rates' and "infection rate and population share."

3.2. Results of the Correlation Analyses

The KW test provides evidence of geographical impact on various ethnicities. In this section we provide the results of correlation analysis between other risk factors. In **Table 3** we see the Pearson correlations between the variables along with the 2-tailed significance values for the NHW group. The same statistics are provided in **Tables 4**, **5** for the NHB and Hispanic communities respectively. All the variables are strongly ($p < 0.01$) and positively correlated with every other variable for all ethnicities, with poverty being the sole exception. To be precise, for the NHW group, poverty is positively correlated

TABLE 1 | Non-parametric Kruskal-Wallis test for each ethnic group's relative infection %, relative mortality %, and population percentages per region as groups and their significance levels.

	NHW		NHB		Hispanic		NHA	
	Statistic	***p*−value**	**Statistic**	***p*−value**	**Statistic**	***p*−value**	**Statistic**	***p*−value**
Northeast	5.38*	0.068	5.43*	0.066	5.21*	0.074	3.81	0.148
South	4.37	0.112	2.95	0.228	10.96***	0.004	7.77**	0.020
Midwest	3.88	0.137	5.83*	0.054	8.66**	0.013	2.40	0.302
West	4.88*	0.087	2.02	0.364	8.83**	0.012	0.60	0.741
All regions	13.03***	0.002	4.83*	0.089	26.18***	<0.01	5.38*	0.067

“” significance at p < 0.1, “**” significance at p < 0.05, “***” significance at p < 0.01.*

TABLE 2 | *Post-hoc* analysis using Dunn Test and their significance levels.

	Infection%-Mortality%		Infection%-Population		Mortality%-Population	
Ethnic group-region	**Statistic**	***p*−value**	**Statistic**	***p*−value**	**Statistic**	***p*−value**
NHW-Northeast	−2.30*	0.064	−1.41	0.314	0.88	0.377
NHW-South	–	–	–	–	–	–
NHW-Midwest	–	–	–	–	–	–
NHW-West	−1.09	0.275	−2.21*	0.081	−1.12	0.527
NHW-All	−2.29**	0.043	−3.55***	0.001	−1.26	0.207
NHB-Northeast	1.29	0.392	2.32*	0.060	1.03	0.301
NHB-South	–	–	–	–	–	–
NHB-Midwest	−0.19	0.842	1.09*	0.094	2.18*	0.087
NHB-West	–	–	–	–	–	–
NHB-All	0.37	0.711	2.06	0.117	1.69	0.181
Hispanic-Northeast	2.27*	0.068	1.31	0.375	−1.02	0.305
Hispanic-South	2.88**	0.011	2.87***	0.008	−0.10	0.917
Hispanic-Midwest	2.67**	0.022	2.42**	0.031	−0.31	0.753
Hispanic-West	2.33**	0.039	2.78**	0.015	0.36	0.718
Hispanic-All	4.60***	<0.01	4.28***	<0.01	−0.47	0.638
NHA-Northeast	–	–	–	–	–	–
NHA-South	1.22	0.221	−1.55	0.239	−2.78**	0.016
NHA-Midwest	–	–	–	–	–	–
NHA-West	–	–	–	–	–	–
NHA-All	1.39	0.325	−1.30	0.190	−2.7**	0.020

“” significance at p < 0.1, “**” significance at p < 0.05, “***” significance at p < 0.01.*

with the infection rate ($r = 0.35, p < 0.05$), and diabetes ($r = 0.29, p < 0.05$). Poverty is either uncorrelated or weakly correlated with other variables in this study for all ethnicities.

3.3. Results of the Linear Models With Infection and Mortality Rates as Response Variables for Each Ethnic Group

In **Table 6** we depict the linear models with infection rate and mortality rate as response variables for NHW. The first column shows preliminary models along with the VIFs for each explanatory variable. The second column depicts the final model obtained via the maximum VIF elimination algorithm described in section 2.4.3. The preliminary model with infection rates in the NHW community as the response variable accounts for 83% [$R^2 = 0.83, R^2_{adj} = 0.80, F_{(6,41)} = 32.87, p < 0.01$] of the variability in the infection rates for NHW population. The final model for the NHW infection rates consists of obesity, diabetes, and poverty as the only explanatory variables. This model accounts for 82% [$R^2 = 0.82, R^2_{adj} = 0.80, F_{(3,44)} = 65.06, p < 0.01$] of the variability in the NHW infection rates. The final NHW infection model and preliminary model both use 48 states.

The preliminary model with NHW mortality rates as the response variable accounts for 88% [$R^2 = 0.88, R^2_{adj} = 0.87, F_{(6,411)} = 51.95, p < 0.01$] of the variability in the NHW mortality rates. The final model for the NHW mortality also consists of obesity, diabetes, and poverty as the only explanatory

TABLE 3 | Pearson correlations for NHW ethnic group between variables used in the study and their significance levels.

		Infections%	Mortality%	Poverty	Age60+	Diabetes	BMI30+	HBP	CVD
Infection%	Pearson correlations	1							
	Sig. (2-tailed)								
Mortality%	Pearson correlations	0.82***	1						
	Sig. (2-tailed)	<0.01							
Poverty	Pearson correlations	0.35**	0.17	1					
	Sig. (2-tailed)	0.014	0.230						
Age60+	Pearson correlations	0.82***	0.91***	0.16	1				
	Sig. (2-tailed)	<0.01	<0.01	0.267					
Diabetes	Pearson correlations	0.78***	0.83***	0.29**	0.94***	1			
	Sig. (2-tailed)	<0.01	<0.01	0.042	<0.01				
BMI30+	Pearson correlations	0.88***	0.91***	0.21	0.96***	0.90***	1		
	Sig. (2-tailed)	<0.01	<0.01	0.141	<0.01	<0.01			
HBP	Pearson correlations	0.86***	0.91***	0.20	0.98***	0.93***	0.99***	1	
	Sig. (2-tailed)	<0.01	<0.01	0.178	<0.01	<0.01	<0.01		
CVD	Pearson correlations	0.84***	0.91***	0.21	0.98***	0.93***	0.97***	0.98***	1
	Sig. (2-tailed)	<0.01	<0.01	0.148	<0.01	<0.01	<0.01	<0.01	

"" significance at $p < 0.1$, "**" significance at $p < 0.05$, "***" significance at $p < 0.01$.*

TABLE 4 | Pearson correlations for NHB ethnic group between variables used in the study and their significance levels.

		Infection%	Mortality%	Poverty	Age60+	Diabetes	BMI30+	HBP	CVD
Infection%	Pearson correlations	1							
	Sig. (2-tailed)								
Mortality%	Pearson correlations	0.93***	1						
	Sig. (2-tailed)	<0.01							
Poverty	Pearson correlations	−0.03	−0.07	1					
	Sig. (2-tailed)	0.833	0.677						
Age60+	Pearson correlations	0.90***	0.95***	−0.10	1				
	Sig. (2-tailed)	<0.01	<0.01	0.550					
Diabetes	Pearson correlations	0.92***	0.95***	−0.084	0.99***	1			
	Sig. (2-tailed)	<0.01	<0.01	0.610	<0.01				
BMI30+	Pearson correlations	0.94***	0.96***	−0.09	0.99***	0.99***	1		
	Sig. (2-tailed)	<0.01	<0.01	0.592	<0.01	<0.01			
HBP	Pearson correlations	0.94***	0.96***	−0.09	0.99***	0.99***	0.99***	1	
	Sig. (2-tailed)	<0.01	<0.01	0.592	<0.01	<0.01	<0.01		
CVD	Pearson correlations	0.86***	0.92***	0.07	0.99***	0.98***	0.96***	0.97***	1
	Sig. (2-tailed)	<0.01	<0.01	0.691	<0.01	<0.01	<0.01	<0.01	

"" significance at $p < 0.1$, "**" significance at $p < 0.05$, "***" significance at $p < 0.01$.*

variables. This NHW mortality final model accounts for 7% [$R^2 = 0.78, R^2_{adj} = 0.77, F_{(3,44)} = 53.48, p < 0.01$] of the variability in the NHW mortality.

In **Table 7** we depict the linear models with infection rate and mortality rate as response variables for NHB. The preliminary model with infection rates in the NHB group as the response variable accounts for 81% [$R^2 = 0.81, R^2_{adj} = 0.77, F_{(6,31)} = 17.51, p < 0.01$] of the variability in the infection rates for NHB population. The final model for the NHW infection rates consists of obesity and poverty as the only explanatory variables. This model accounts for 66% [$R^2 = 0.66, R^2_{adj} = 0.64, F_{(2,37)} = 37.89, p < 0.01$] of the variability in the NHB infection rates. Note that the final NHB infection model uses 40 states instead of 38 states in the preliminary model. This discrepancy is because of the unavailability of data for the NHB community for all the explanatory variables, as discussed in section 2.4.3. The preliminary model with NHB mortality rates as the response variable accounts for 77% [$R^2 = 0.77, R^2_{adj} = 0.73, F_{(6,31)} = 17.51, p < 0.01$] of the variability in the NHB mortality rates. The final model for the NHB mortality also consists of only obesity

TABLE 5 | Pearson correlations for Hispanic ethnic group between variables used in the study and their significance levels.

		Infection%	Mortality%	Poverty	Age60+	Diabetes	BMI30+	HBP	CVD
Infection%	Pearson correlations	1							
	Sig. (2-tailed)								
Mortality%	Pearson correlations	0.84***	1						
	Sig. (2-tailed)	<0.01							
Poverty	Pearson correlations	−0.16	−0.14	1					
	Sig. (2-tailed)	0.306	0.349						
Age60+	Pearson correlations	0.74***	0.82***	−0.07	1				
	Sig. (2-tailed)	<0.01	<0.01	0.632					
Diabetes	Pearson correlations	0.83***	0.85***	−0.04	0.96***	1			
	Sig. (2-tailed)	<0.01	<0.01	0.793	<0.01				
BMI30+	Pearson correlations	0.85***	0.87***	−0.16	0.97***	0.99***	1		
	Sig. (2-tailed)	<0.01	<0.01	0.319	<0.01	<0.01			
HBP	Pearson correlations	0.80***	0.81***	−0.11	0.98***	0.98***	0.98***	1	
	Sig. (2-tailed)	<0.01	<0.01	0.482	<0.01	<0.01	<0.01		
CVD	Pearson correlations	0.75***	0.82***	−0.01	0.98***	0.97***	0.98***	0.98***	1
	Sig. (2-tailed)	<0.01	<0.01	0.950	<0.01	<0.01	<0.01	<0.01	

"" significance at $p < 0.1$, "**" significance at $p < 0.05$, "***" significance at $p < 0.01$.*

TABLE 6 | Linear models with infection and mortality rates as response variables for NHW.

Linear regression models withs infection rate in NHW population as a response variable						
	Preliminary model			Final model		
	β (95% *CI*)	*Pr* (> *t*)	VIF	*β* (95% *CI*)	*Pr* (> |*t*|)	VIF
(Intercept)	2.21 (1.53, 2.89)	<0.01***		2.46 (2.16, 2.74)	<0.01***	
BMI30+	0.04 (0.01, 0.06)	<0.01***	59.51	0.02 (0.02, 0.03)	<0.01***	5.35
Diabetes	−0.002 (−0.01, 0.01)	0.956	9.67	−0.004 (−0.01, 0.03)	0.318	5.47
Poverty	0.02 (−0.00, 0.05)	0.067*	1.20	0.02 (−0.00, 0.05)	0.082*	1.09
Age60+	0.20 (−0.01, 0.05)	0.787	58.57			
HBP	−0.03 (−0.06, 0.01)	0.725	104.88			
CVD	−0.01 (−0.03, 0.01)	0.309	39.07			
	$R^2 = 0.83, R^2_{adj} = 0.80, n = 48$ $F_{(6,41)} = 32.87, p < 0.01$***			$R^2 = 0.82, R^2_{adj} = 0.80, n = 48$ $F_{(3,44)} = 65.06, p < 0.01$***		

Linear regression models with mortality rate in NHW population as a response variable						
	Preliminary model			Final model		
	β (95% *CI*)	*Pr* (> *t*)	VIF	*β* (95% *CI*)	*Pr* (> |*t*|)	VIF
(Intercept)	1.58 (1.03, 2.12)	<0.01***		2.59 (2.27, 2.91)	<0.01**	
BMI30+	0.003 (−1.65, 0.02)	0.785	59.51	0.86 (0.01, 0.02)	0.001***	5.35
Diabetes	−0.002 (−1.08, 0.01)	0.643	9.67	0.01 (0.00, 0.02)	0.039**	5.47
Poverty	0.01 (−1.38, 0.03)	0.512	1.20	−0.00 (−1.44, 0.98)	0.955	1.09
Age60+	0.04 (1.18, 0.06)	0.005***	58.57			
HBP	−0.03 (−5.40, 0.00)	0.070*	104.88			
CVD	0.02 (1.57, 0.04)	<0.050**	39.07			
	$R^2 = 0.88, R^2_{adj} = 0.87, n = 48$ $F_{(6,41)} = 51.95, p < 0.01$***			$R^2 = 0.78, R^2_{adj} = 0.77, n = 48$ $F_{(3,44)} = 53.48, p < 0.01$***		

The first column shows preliminary model along with the VIFs for each explanatory variable. The second column indicates the model obtained using maximum VIF elimination algorithm. "" significance at $p < 0.1$, "**" significance at $p < 0.05$, and "***" significance at $p < 0.01$.*

TABLE 7 | Linear models with infection and mortality rates as response variables for NHB.

Linear regression models with infection rate in NHB population as a response variable						
	Preliminary model			**Final model**		
	β(95% *CI*)	***Pr*(> *t*)**	**VIF**	**β(95% *CI*)**	***Pr*(> |*t*|)**	**VIF**
`(Intercept)`	1.10 (0.42, 1.79)	0.002**		2.12 (1.42, 2.84)	<0.01***	
`BMI30+`	0.04 (−0.04, 0.13)	0.309	97.03	0.05 (0.04, 0.06)	<0.01***	1.00
`Poverty`	0.02 (−0.01, 0.05)	0.169	1.10	−0.01 (−0.04, 0.02)	0.39	1.00
`Diabetes`	0.02 (−0.06, 0.10)	0.695	101.10			
`Age60+`	−0.20 (−0.34, −0.05)	0.009***	164.01			
`HBP`	0.16 (0.06, 0.26)	0.003***	127.51			
`CVD`	−0.02 (−0.09, 0.05)	0.514	58.27			
	$R^2 = 0.81, R^2_{adj} = 0.77, n = 38$			$R^2 = 0.66, R^2_{adj} = 0.64, n = 40$		
	$F_{(6,31)} = 17.51, p < 0.01$***			$F_{(2,37)} = 37.89, p < 0.01$***		

Linear regression models with mortality rate in NHB population as a response variable						
	Preliminary model			**Final model**		
	β (95% *CI*)	***Pr* (> *t*)**	**VIF**	**β (95% *CI*)**	***Pr* (> |*t*|)**	**VIF**
`(Intercept)`	1.26 (0.36, 2.17)	0.008***		2.16 (1.32, 3.01)	<0.01***	
`BMI30+`	0.10 (−0.01, 0.22)	0.073*	97.03	0.06 (0.04, 0.07)	<0.01***	1.00
`Poverty`	0.00 (−0.04, 0.04)	0.921	1.10	−0.02 (−0.05, 0.12)	0.191	1.00
`Diabetes`	−0.04 (−0.15, 0.06)	439	101.10			
`Age60+`	−0.25 (−0.45, −0.06)	0.011**	164.01			
`HBP`	0.17 (−0.03, 0.30)	0.017**	127.51			
`CVD`	0.03 (−0.06, 0.12)	0.521	58.27			
	$R^2 = 0.77, R^2_{adj} = 0.73, n = 38$			$R^2 = 0.67, R^2_{adj} = 0.65, n = 40$		
	$F_{(6,31)} = 17.51, p < 0.01$***			$F_{(2,37)} = 199.9, p < 0.01$***		

The first column shows preliminary model along with the VIFs for each explanatory variable. The second column indicates the model obtained using maximum VIF elimination algorithm. "" significance at $p < 0.1$, "**" significance at $p < 0.05$, and "***" significance at $p < 0.01$.*

and poverty as the explanatory variables. This NHB mortality final model accounts for 67% $[R^2 = 0.67, R^2_{adj} = 0.65, F_{(2,37)} = 199.9, p < 0.01]$ of the variability in the NHB mortality rates.

The **Table 8** depicts the linear models with infection rate and mortality rate as response variables for the Hispanic group. The preliminary model for infection rates in the Hispanic community accounts for 67% $[R^2 = 0.67, R^2_{adj} = 0.60, F_{(6,26)} = 8.97, p < 0.01]$ of the variability in the infection rates for the Hispanic population. The final model for the Hispanic infection rates consists of diabetes and poverty as the explanatory variables. This model accounts for 51% $[R^2 = 0.51, R^2_{adj} = 0.48, F_{(6,25)} = 9.98, p < 0.01]$ of the variability in the Hispanic infection rates. The preliminary model with mortality rates among the Hispanic community as the response variable accounts for 71% $[R^2 = 0.71, R^2_{adj} = 0.63, F_{(6,25)} = 9.98, p < 0.01]$ of the variability in the Hispanic mortality rates. The final model for Hispanic mortality consists of advanced age and poverty as the explanatory variables. Note that advanced age is the most significant explanatory variable in the final mortality model in the Hispanic group, whereas having diabetes was the most significant variable predicting infection in the Hispanic community. This final model accounts for 55% $[R^2 = 0.55, R^2_{adj} = 0.53, F_{(2,39)} = 23.93, p < 0.01]$ of the variability in the Hispanic mortality rates. We note that the final model for the Hispanic mortality includes 42 states, whereas the preliminary model has only 32 states due to lack of data availability.

Adjusted R^2 value for the regression model for NHW mortality was much higher (0.77) in comparison to NHB (0.65) and Hispanic (0.53). However, all six models showed statistical significance and satisfied normality tests for the residual values. Indeed, the Lilliefors normality test applied to the residuals obtained from each of these models revealed that the residuals were normally distributed with $p > 0.001$. The histograms, and the QQ plots for the residuals are provided in **Figure 3**.

3.4. Results From the Geographically Weighted Regression

The geographically weighted regression yields coefficients for each risk factor for every state. We show the state-wise coefficients for the most significant explanatory variable for each ethnicity in **Figure 4**. Empty spaces for states in **Figure 4** indicate that ethnicity-wise data was not available for those states for the

TABLE 8 | Linear models with infection and mortality rates as response variables for Hispanic.

Linear regression models with infection rate in Hispanic population as a response variable						
	Preliminary model			**Final model**		
	β(95% *CI*)	***Pr*(> *t*)**	**VIF**	**β(95% *CI*)**	***Pr*(> |*t*|)**	**VIF**
(Intercept)	2.32 (1.56, 3.083)	<0.01***		2.34 (1.63, 3.04)	<0.01***	
Diabetes	−0.03 (−0.11, 0.05)	0.439	44.72	0.04 (0.03, 0.052)	< 0.01***	1.00
Poverty	0.00 (−0.04, 0.04)	0.931	1.04	0.01 (−0.02, 0.05)	0.489	1.00
BMI30+	0.15 (0.05, 0.24)	0.003***	74.89			
Age60+	−0.83 (−0.20, 0.04)	0.162	37.24			
HBP	0.02 (−0.10, 0.13)	0.783	61.21			
CVD	−0.08 (−0.18, 0.02)	0.113	40.27			
	$R^2 = 0.67, R^2_{adj} = 0.60, n = 33$ $F_{(6,26)} = 8.97, p < 0.01$***			$R^2 = 0.51, R^2_{adj} = 0.48, n = 41$ $F_{(2,38)} = 19.52, p < 0.01$***		
Linear regression models with mortality rate in Hispanic population as a response variable						
	Preliminary model			**Final model**		
	β (95% *CI*)	***Pr* (> *t*)**	**VIF**	**β (95% *CI*)**	***Pr* (> |*t*|)**	**VIF**
(Intercept)	2.07 (−0.96, 3.18)	0.001***		2.51 (1.60, 3.42)	< 0.01***	
Age60+	−0.05 (−0.20, 0.10)	0.504	37.24	0.09 (0.06, 0.12)	< 0.01***	1.00
Poverty	−0.04 (−0.10, 0.01)	0.138	1.04	−0.047 (−0.09, −0.00)	0.043**	1.00
Diabetes	−0.04 (−0.14, 0.06)	0.398	44.72			
BMI30+	151 (0.03, 0.27)	0.002**	74.89			
HBP	−0.02 (−0.16, 0.13)	0.828	61.21			
CVD	−0.04 (−0.17, 0.08)	0.487	40.27			
	$R^2 = 0.71, R^2_{adj} = 0.63, n = 32$ $F_{(6,25)} = 9.98, p < 0.01$***			$R^2 = 0.55, R^2_{adj} = 0.53, n = 42$ $F_{(2,39)} = 23.93, p < 0.01$***		

The first column shows preliminary model along with the VIFs for each explanatory variable. The second column indicates the model obtained using maximum VIF elimination algorithm. "" significance at $p < 0.1$, "**" significance at $p < 0.05$, and "***" significance at $p < 0.01$.*

corresponding risk factors. The *p*-values for the most significant explanatory variable and state-wise R^2 values obtained from the GWR model are included in **Supplementary Figures S4, S5**.

4. DISCUSSION

Our analysis of the nationwide data revealed that geographical location, and other COVID-19 risk factors affect different ethnicities in a dissimilar way. We observed that the disparate burden of the pandemic was most prominent on the NHB and Hispanic communities. This observation is supported by Anyane-Yeboa et al. (56) and Escoba et al. (57) other studies. In particular, the rate of infection was exceptionally high for the Hispanic community compared to their population share. Discordant impact on NHB and Hispanic populations has been reported by Centers for Disease Control (58) and studied using data from metropolitan cities and combining selected states, but the nationwide study is limited. In our work, this effect was observed in the four US regions separately and also when all the states' data was aggregated. When considered the four regions individually, we found that the excessive infection rate in the Hispanic community was most prominent in the South region. However, compared to the Hispanic group's infection rates, their mortality rates were statistically lower in all regions of the US. This apparent discrepancy could be because the Hispanic community is the youngest of the four ethnic groups considered in our study (59). The infection rate of NHB population was higher compared to their population share in the Midwest, and the Northeast than other regions.

The correlation analysis confirmed that the COVID-19 related risk factors such as advanced age, cardiovascular disease, diabetes, hypertension, and obesity are highly interrelated. This finding is consistent with numerous studies. For example, Mokdad et al. (60) show that obesity (BMI≥ 30) was significantly associated with diabetes, hypertension, high cholesterol, asthma, and arthritis. Wilson and Kannel (61) conclude that obesity and diabetes are associated with atherogenic risk factors. Abdullah et al. (62) also conclude that obesity is associated with type 2 diabetes. We also found that "within" an ethnic group, poverty was uncorrelated or weakly correlated with infections and mortality for all three ethnic groups, implying that poverty

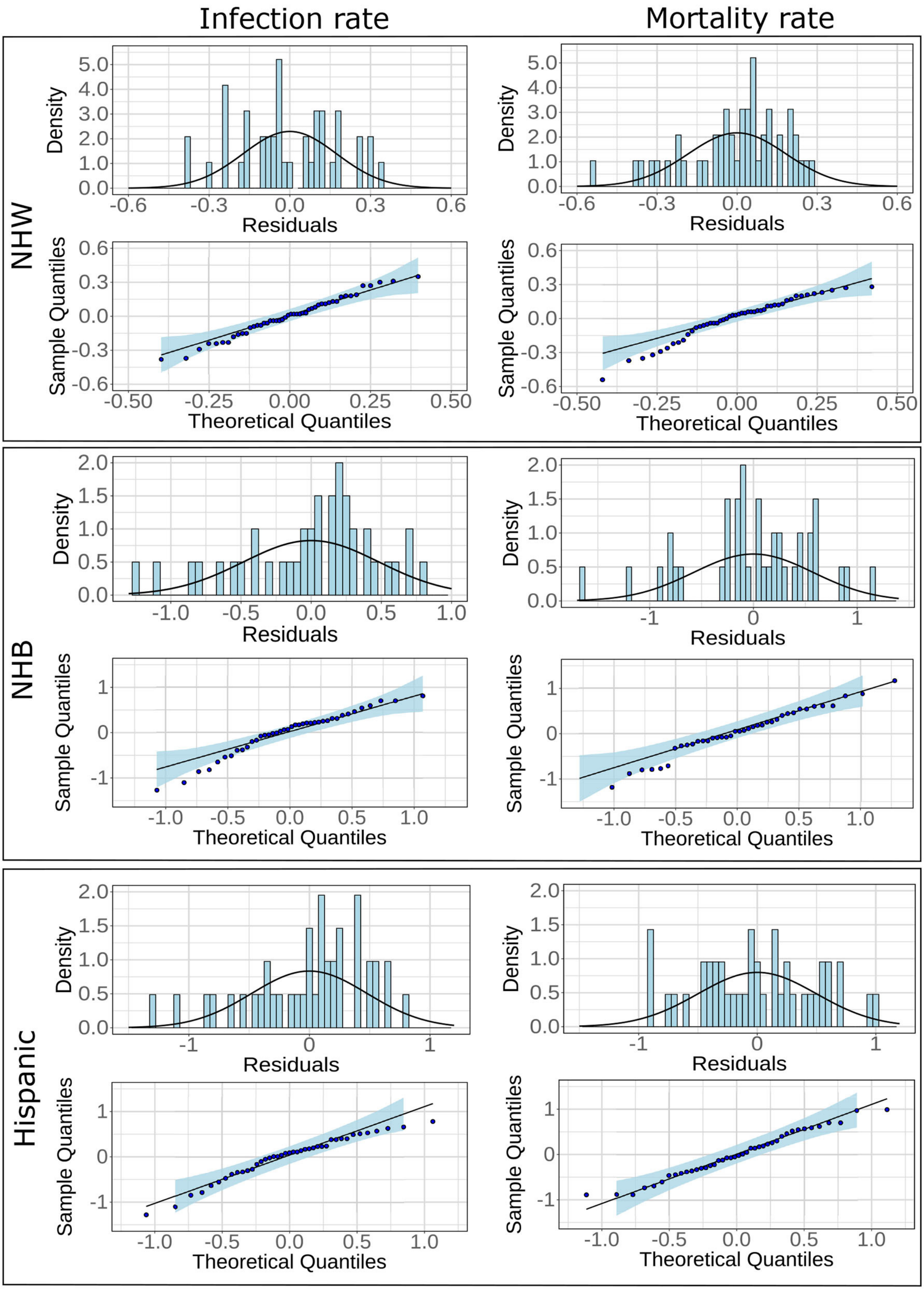


FIGURE 3 | Histograms and fitted normal curves along with QQ plots of the residuals for linear regression for NHW, NHB, and Hispanics groups. The linear models are based on infections and mortality rates as response variables.

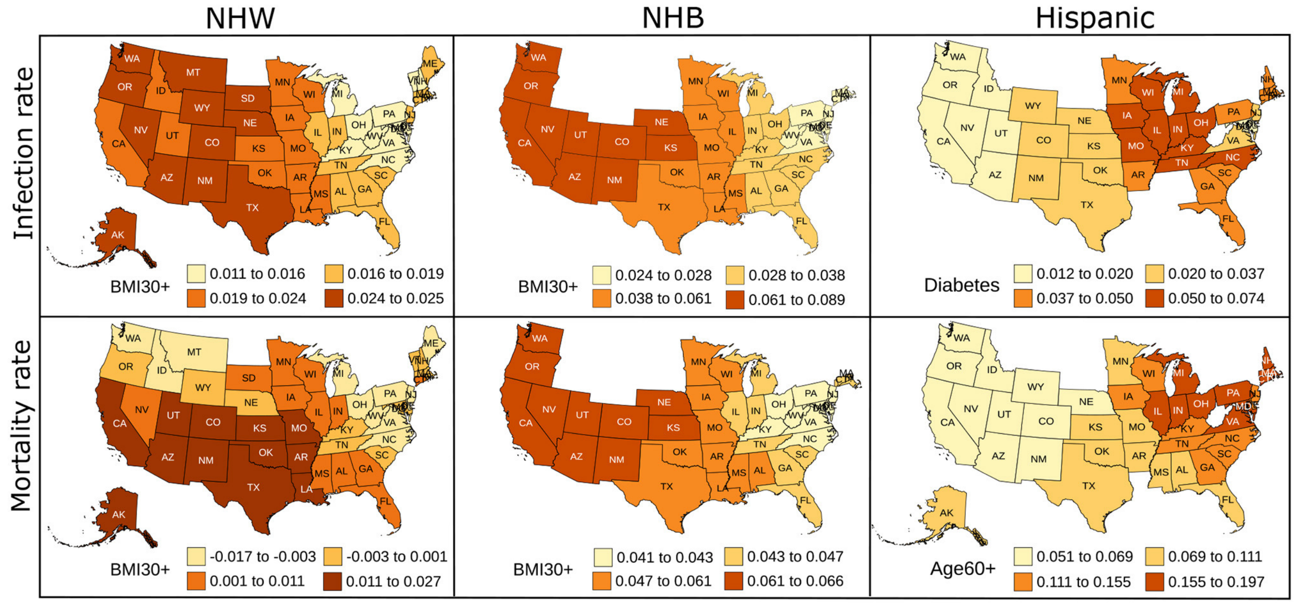


FIGURE 4 | Map showing the coefficient values for the most significant variables for the GWR models constructed on predictors used in linear regression for NHW, NHB, and Hispanics data. The states with no color indicate that the ethnicity-wise data on the corresponding risk factors was not available in those states.

is an "independent" risk factor for COVID-19. This finding is supported by Elgar et al. (32) and Oronce et al. (33) which we discussed in section 1.4.

After eliminating variables with high multicollinearity, we formulated robust and parsimonious linear models for NHW, NHB, and Hispanic populations. The linear models described in section 3.3 reveal that "obesity" encapsulates many other co-dependent risk factors for the infection and mortality in NHW and NHB groups. This finding is expected in light of numerous studies (61, 63). Obesity and diabetes are well-established risk factors for COVID-19. In these two conditions, adipose tissue is compromised, which can directly or indirectly get involved in interaction with SARS-CoV-2, the pathogen responsible for COVID-19 disease (64). Thus, it is not surprising that obesity highly influences the regression models for NHW and NHB with death rate as a response variable. However, the degree of influence of obesity on infection rates and mortality is noteworthy, with obesity emerging as the most significant factor contributing to the infection rates and mortality for the NHW and NHB groups.

The Hispanic community markedly differs from NHW, and NHB with respect to the results of the linear models. Diabetes was the most significant factor for infection rate in Hispanics, while advanced age emerged as most significant for mortality. The effect of advanced age on Hispanic mortality could be also due to the relatively younger, and thus working-age, population of Hispanics (59) in the US.

The regression models indicate a strong association of poverty with a high infection rate, followed by death for all ethnic groups studied. This finding is in agreement with several studies focusing on the association of low socioeconomic status, which increases the exposure to COVID-19 (31, 65). People with low socioeconomic status avail healthcare services at an advanced stage of illness, thus experience a worse prognosis. The disease burden associated with obesity is linked to socioeconomic status and race (28). Ethnic minorities and populations with low socioeconomic status have been disproportionately affected in previous pandemics (5, 6, 8). Evidence from the COVID-19 pandemic is not an exception to the above fact. To this end, public health strategies to control the current and future pandemics need to take these ethnicity-specific effects into account to mitigate the spread and severity of the disease.

The linear regression furnishes global and static coefficients for the explanatory variables. However, the geographically weighted regression gives coefficients that are geographically varying. We see in **Figure 4** the variability in the coefficients of the GWR. We note that the neighboring states seem to have similar coefficients, indicating similarity in the risk factors in nearby states. Obesity is the most prominent risk factor amongst the NHW and NHB populations, and diabetes and advanced age seem to be more influential in the Hispanic community. The GWR results for the Hispanic group show more variability than the NHW group, which could be due to the higher percentage of people of Hispanic origin in southern and western states. The local R^2 map in **Supplementary Figure S5** also indicates that the GWR model fits the NHW and NHB groups better than the Hispanic group. We plan to explore the geographical variation of the risk factors in more detail in future work.

4.1. Limitations

As discussed in section 2.2 and noted by other researchers (36, 38) there is a lack of consistency and availability of COVID-19 related data. Our study does not include data from the state of New York, since the state does not make the ethnicity wise data available. Moreover, the data we use is state wise statistics of the various risk factors. However, we note that the observations made using such data is consistent with those made by other researchers.

4.2. Practical Implications of the Study

Although racial and ethnic disparities in COVID-19 infections and mortality are becoming increasingly clear from several studies based on available data, drivers of these disparate outcomes remain less understood at a national level. Our models, based on the nationwide data, indicate that "obesity" effectively encapsulates the effect of other co-dependent factors for NHW, and NHB populations (section 3.3). The link between COVID-19 infection severity and obesity is noted by Watanabe et al. (66) even in the early stages of the pandemic. Similarly, during the H1N1 pandemic of 2009 (67, 68) observed that obesity was associated with higher mortality.

Another implication from our work is that the Hispanic community is more susceptible to the COVID-19 infection. This observation is valid throughout the US. This situation could be remedied via public policy changes and awareness of the issue. The disproportionate impact of COVID-19 on the minority population is largely attributed to existing socioeconomic inequities. The low-income minority population are often compelled to work in an environment with higher risk of disease exposure, live in a crowded accommodation, and lack adequate access to healthcare. The government support to low-income families in the form of the CARES Act, Consolidated Appropriations Act, 2021, Department of Treasury US (69) and the American Rescue Plan Act of 2021 (70) are critical but might not be sufficient to fully mitigate observed disparity in infection and mortality rates. Our analysis indicates that certain subpopulations of the minority population are at higher risk of COVID-19 infection and mortality. Identifying these vulnerable subpopulations, such as Hispanics with diabetes or age over 60 years, and prioritizing additional attention to these populations could enable a more efficient allocation and utilization of resources. Increased effort toward educating and raising awareness on COVID-19 and associated risk factors could also be an effective method to develop community resilience. One potential avenue to improve awareness on COVID-19 will be through recruiting volunteers to educate the vulnerable population. For example, "Philly counts" (71), a program supported by the Philadelphia Department of Public Health, initially created for Census 2020, currently helps direct community engagement efforts for the COVID-19 vaccine. Extending similar initiatives to populations with major risk factors such as obesity could result in a major beneficial impact on overall COVID-19 burden.

5. CONCLUSION

Several researchers have concluded that several health conditions, poverty, and geographical location affect the COVID-19 prognosis. Studies have shown that the COVID-19 pandemic has impacted some minorities in the US more severely than other groups. Our work focused on quantifying this distinct effect of various COVID-19 risk factors on different ethnicities in the US during the first pandemic wave.

To this effect, we included Non-Hispanic White, Non-Hispanic Black, Hispanic, Non-Hispanic Asians. Our work has revealed differences in the way the COVID-19 pandemic affected various ethnic groups. We observed that the infection rates in the Hispanic population were disproportionately larger than the share of their population across all regions of the US. This effect was most prominent in the South region. The NHA populations consistently had lower infection rates and mortality rates compared to their population. Furthermore, we studied the following risk factors in this work: advanced age, obesity, cardiovascular diseases, diabetes, hypertension, and poverty for NHW, NHB, and Hispanic populations. We aimed to quantify the different effects of these risk factors on various ethnicities. To this end, we constructed linear models with infection and mortality rates as the response variables. We eliminated variables causing multicollinearity from our models, leading to robust linear models. Our models indicate that "obesity" parsimoniously describes the impact of other co-dependent comorbidities for NHW and NHB populations (section 3.3). However, for the infection rates in the Hispanic group, the factor leading to the robust linear model was the prevalence of diabetes. On the other hand, advanced age was more significant for COVID-19 related mortality for the Hispanic community. We also established "poverty" as an independent risk factor for infection and mortality amongst the three ethnicities: NHW, NHB, and Hispanics. The findings in this study quantified ethnicity-specific effects of COVID-19 risk factors, which we hope could be mollified with public policy interventions and community engagement.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/prashantva/Covid-19-Ethnicity.

AUTHOR CONTRIBUTIONS

PA: writing—review, conceptualization, editing, investigation, and analysis. JC: data collection. VK: data curation, formal analysis, visualization, and coding. SM: writing—original draft, methodology, formal analysis, and project administration. SS: supervision, conceptualization, validation, and editing. All authors contributed to the article and approved the submitted version.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.743003/full#supplementary-material

REFERENCES

1. Chartier K, Caetano R. Ethnicity and health disparities in alcohol research. *Alcohol Res*. (2010) 33:152.
2. Barr DA. *Health Disparities in the United States: Social Class, Race, Ethnicity, and Health*. Baltimore, MD: JHU Press (2014).
3. Dressler WW, Oths KS, Gravlee CC. Race and ethnicity in public health research: models to explain health disparities. *Annu Rev Anthropol*. (2005) 34:231–52. doi: 10.1146/annurev.anthro.34.081804.120505
4. Braun L. Race, ethnicity, and health: can genetics explain disparities? *Perspect Biol Med*. (2002) 45:159–74. doi: 10.1353/pbm.2002.0023
5. Khunti K, Platt L, Routen A, Abbasi K. Covid-19 and ethnic minorities: an urgent agenda for overdue action. *Br Med J Publish Group*. (2020) doi: 10.1136/bmj.m2503
6. Kabarriti R, Brodin NP, Maron MI, Guha C, Kalnicki S, Garg MK, et al. Association of race and ethnicity with comorbidities and survival among patients with COVID-19 at an urban medical center in New York. *JAMA Netw Open*. (2020) 3:e2019795–e2019795. doi: 10.1001/jamanetworkopen.2020.19795
7. Owen WF, Carmona R, Pomeroy C. Failing another national stress test on health disparities. *JAMA*. (2020) 323:1905–6. doi: 10.1001/jama.2020.6547
8. Dobin D, Dobin A. Racial/ethnic and socioeconomic disparities of Covid-19 attacks rates in Suffolk County communities. *arXiv[Preprint]*. (2020). *arXiv:2004.12175.*
9. Moore JT, Ricaldi JN, Rose CE, Fuld J, Parise M, Kang GJ, et al. Disparities in incidence of COVID-19 among underrepresented racial/ethnic groups in counties identified as hotspots during June 5-18, 2020–22 states, February-June (2020). *Morbid Mortal Wkly Rep*. (2020) 69:1122. doi: 10.15585/mmwr.mm6933e1
10. Adhikari S, Pantaleo NP, Feldman JM, Ogedegbe O, Thorpe L, Troxel AB. Assessment of community-level disparities in coronavirus disease 2019 (COVID-19) infections and deaths in large US metropolitan areas. *JAMA Netw Open*. (2020) 3:e2016938. doi: 10.1001/jamanetworkopen.2020.16938
11. Sun Z, Zhang H, Yang Y, Wan H, Wang Y. Impacts of geographic factors and population density on the COVID-19 spreading under the lockdown policies of China. *Scie Total Environ*. (2020) 746:141347. doi: 10.1016/j.scitotenv.2020.141347
12. Breen R, Ermisch J. The distributional impact of COVID-19: Geographic variation in mortality in England. *Demogr Res*. (2021) 44:397–414. doi: 10.4054/DemRes.2021.44.17
13. Desmet K, Wacziarg R. JUE Insight: Understanding spatial variation in COVID-19 across the United States. *J Urban Econ*. (2021) 103332. doi: 10.1016/j.jue.2021.103332
14. Stephens-Davidowitz S. "The cost of racial animus on a black candidate: evidence using Google search data". *J Public Econ*. (2014). 118:26–40. doi: 10.1016/j.jpubeco.2014.04.010
15. Drucker DJ. Coronavirus infections and type 2 diabetes–shared pathways with therapeutic implications. *Endocr Rev*. (2020) 41:457–70. doi: 10.1210/endrev/bnaa011
16. Muniyappa R, Gubbi S. COVID-19 pandemic, coronaviruses, and diabetes mellitus. *Am J Physiol Endocrinol Metab*. (2020) 318:E736–41. doi: 10.1152/ajpendo.00124.2020
17. Orioli L, Hermans MP, Thissen JP, Maiter D, Vandeleene B, Yombi JC. COVID-19 in diabetic patients: related risks and specifics of management. In: *Annales d'endocrinologie. Vol. 81*. Elsevier (2020). p. 101–9. doi: 10.1016/j.ando.2020.05.001
18. Yan Y, Yang Y, Wang F, Ren H, Zhang S, Shi X, et al. Clinical characteristics and outcomes of patients with severe covid-19 with diabetes. *BMJ Open Diabetes Res Care*. (2020) 8:e001343. doi: 10.1136/bmjdrc-2020-001343
19. Richardson S, Hirsch JS, Narasimhan M, Crawford JM, McGinn T, Davidson KW, et al. Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *JAMA*. (2020) 323:2052–9. doi: 10.1001/jama.2020.6775
20. Sun C, Zhang X, Dai Y, Xu X, Zhao J. Clinical analysis of 150 cases of 2019 novel coronavirus infection in Nanyang City, Henan Province. *Zhonghua jie he he hu xi za zhi= Zhonghua Jiehe he Huxi Zazhi= Chin J Tuberculosis Respiratory Dis*. (2020) 43:E042. doi: 10.3760/cma.j.cn112147-20200224-00168
21. Gao F, Zheng KI, Wang XB, Sun QF, Pan KH, Wang TY, et al. Obesity is a risk factor for greater COVID-19 severity. *Diabetes Care*. (2020) 43:e72–4. doi: 10.2337/dc20-0682
22. Busetto L, Bettini S, Fabris R, Serra R, Dal Pra C, Maffei P, et al. "Obesity and COVID-19: an Italian snapshot". *Obesity*. (2020) 28:1600–5. doi: 10.1002/oby.22918
23. Sharp JT, Henry JP, Sweany SK, Meadows WR, Pietras RJ. The total work of breathing in normal and obese men. *J Clin Invest*. (1964) 43:728–39. doi: 10.1172/JCI104957
24. Li AM, Chan D, Wong E, Yin J, Nelson EAS, Fok TF. The effects of obesity on pulmonary function. *Arch Dis Child*. (2003) 88:361–3. doi: 10.1136/adc.88.4.361
25. Anderson RM, Fraser C, Ghani AC, Donnelly CA, Riley S, Ferguson NM, et al. Epidemiology, transmission dynamics and control of SARS: the 2002&#x2013;2003 epidemic. *Philos Trans R Soc Lond B Biol Sci*. (2004) 359:1091–105. doi: 10.1098/rstb.2004.1490
26. Ma C, Gu J, Hou P, Zhang L, Bai Y, Guo Z, et al. "Incidence, clinical characteristics and prognostic factor of patients with COVID-19: a systematic review and meta-analysis". *medRxiv*. (2020) doi: 10.1101/2020.03.17.20037572
27. CDC dataset. *Provisional COVID-19 Deaths by Sex and Age*. (2020). Available online at: https://data.cdc.gov/NCHS/Provisional-COVID-19-Deaths-by-Sex-and-Age/9bhg-hcku
28. Paeratakul S, Lovejoy JC, Ryan DH, Bray GA. The relation of gender, race and socioeconomic status to obesity and obesity comorbidities in a sample of US adults. *Int J Obes*. (2002) 26:1205–10. doi: 10.1038/sj.ijo.0802026
29. Sturm R, Hattori A. Morbid obesity rates continue to rise rapidly in the United States. *Int J Obes*. (2013) 37:889–91. doi: 10.1038/ijo.2012.159
30. Kuzawa CW, Sweet E. Epigenetics and the embodiment of race: developmental origins of US racial disparities in cardiovascular health. *Am J Hum Biol*. (2009) 21:2–15. doi: 10.1002/ajhb.20822
31. Patel J, Nielsen F, Badiani A, Assi S, Unadkat V, Patel B, et al. Poverty, inequality and COVID-19: the forgotten vulnerable. *Public Health*. (2020) 183:110. doi: 10.1016/j.puhe.2020.05.006
32. Elgar FJ, Stefaniak A, Wohl MJA. The trouble with trust: time-series analysis of social capital, income inequality, and COVID-19 deaths in 84 countries. *Soc Sci Med*. (2020) 263:113365. doi: 10.1016/j.socscimed.2020.113365
33. Oronce CIA, Scannell CA, Kawachi I, Tsugawa Y. Association between state-level income inequality and COVID-19 cases and mortality in the USA. *J Gen Intern Med*. (2020) 35:2791–3. doi: 10.1007/s11606-020-05971-3
34. Hsu CC, Lee CH, Wahlqvist ML, Huang HL, Chang HY, Chen L, et al. Poverty increases type 2 diabetes incidence and inequality of care despite universal health coverage. *Diabetes Care*. (2012) 35:2286–92. doi: 10.2337/dc11-2052
35. Census Bureau US. *2010 Census Regions and Divisions of the United States*. (2018). Available online at: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
36. Carroll SR, Akee R, Chung P, Cormack D, Kukutai T, Lovett R, et al. Indigenous peoples' data during COVID-19: from external to internal. *Front Sociol*. (2021) 6:62. doi: 10.3389/fsoc.2021.617895
37. Hatcher SM, Agnew-Brune C, Anderson M, Zambrano LD, Rose CE, Jim MA, et al. COVID-19 among American Indian and Alaska native persons–23 states, January 31-July 3, (2020). *Morbid Mortal Wkly Rep*. (2020) 69:1166. doi: 10.15585/mmwr.mm6934e1
38. Talagadadeevi T, Iyer S, Saurer D. An analysis of limitations for demographic data reporting on State COVID-19 dashboards. (2020). Available online at: https://covid19dataproject.org/wp-content/uploads/2020/09/9_27-vertical-demographic-poster_good.pdf
39. KFF Race data. *Population Distribution by Race/Ethnicity*. (2020). Available online at: https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/
40. KFF Covid-19 data. *Population Distribution by Race/Ethnicity*. (2020). Available online at: https://github.com/KFFData/COVID-19-Data
41. ACL dataset. *Administration for Community Living*. (2018). Available online at https://p-agid-2021-wapp1.azurewebsites.net/StateProfiles/

42. AHR Diabetes data. *Public Health Impact: Diabetes.* (2019). Available online at: https://www.americashealthrankings.org/explore/annual/measure/Diabetes/state/ALL?edition-year=2019
43. National Institute of Health NH, Lung, Institute B, of Diabetes NI, Digestive, (US) KD. *Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults-The Evidence Report. Vol. 6.* Bethesda, MD: National Heart, Lung, and Blood Institute. (1998).
44. AHR Obesity data. *Public Health Impact: Obesity.* (2020). Available online at: https://www.americashealthrankings.org/explore/annual/measure/Obesity/state/ALL
45. AHR CVD data. *Public Health Impact: Cardiovascular Diseases.* (2019). Available online at: https://www.americashealthrankings.org/explore/annual/measure/Obesity/state/ALL
46. AHR HBP data. *Public Health Impact: High Blood Pressure.* (2019). Available online at: https://www.americashealthrankings.org/explore/annual/measure/Hypertension/state/ALL
47. KFF Poverty data. *Poverty Rate by Race/Ethnicity.* (2019). Available online at: https://www.kff.org/other/state-indicator/poverty-rate-by-raceethnicity
48. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc.* (1952) 47:583–21. doi: 10.1080/01621459.1952.10483441
49. Dunn OJ. Multiple comparisons among means. *J Am Stat Assoc.* (1961) 56:52–64. doi: 10.1080/01621459.1961.10482090
50. Bland JM, Altman DG. Statistics notes. Logarithms. *BMJ.* (1996) 312:700. doi: 10.1136/bmj.312.7032.700
51. Belsley DA, Kuh E, Welsch RE. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* Hoboken, NJ: John Wiley & Sons;. (2004).
52. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W. *Applied Linear Statistical Models.* 4th ed. Chicago, IL: IRWIN (1996).
53. Kutner MH, Nachtsheim CJ, Neter J. *Applied Linear Regression Models.* New York, NY: McGraw Hill (2004).
54. Lilliefors HW. On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *J Am Stat Assoc.* (1967) 62:399–402. doi: 10.1080/01621459.1967.10482916
55. Brunsdon C, Fotheringham S, Charlton M. Geographically weighted regression. *J R Stat Soc D.* (1998) 47:431–43. doi: 10.1111/1467-9884.00145
56. Anyane-Yeboa A, Sato T, Sakuraba A. Racial disparities in COVID-19 deaths reveal harsh truths about structural inequality in America. *J Intern Med.* (2020) 288:479–80. doi: 10.1111/joim.13117
57. Escobar GJ, Adams AS, Liu VX, Soltesz L, Chen YFI, Parodi SM, et al. Racial disparities in COVID-19 testing and outcomes: retrospective cohort study in an integrated health system. *Ann Intern Med.* (2021) 174:786–93. doi: 10.7326/M20-6979
58. for Disease Control C, Prevention, et al. COVID-19 hospitalization and death by race/ethnicity. (2020). Available online at: https://www.cdc.gov/coronavirus/2019-ncov/covid-data/investigations-discovery/hospitalization-death-by-race-ethnicity.html (accessed September 9, 2021).
59. Patten E. *The Nation's Latino Population Is Defined by Its Youth.* (2016). Available online at: https://www.pewresearch.org/hispanic/2016/04/20/the-nations-latino-population-is-defined-by-its-youth/
60. Mokdad AH, Ford ES, Bowman BA, Dietz WH, Vinicor F, Bales VS, et al. Prevalence of obesity, diabetes, and obesity-related health risk factors, 2001. *JAMA.* (2003) 289:76–9. doi: 10.1001/jama.289.1.76
61. Wilson PW, Kannel WB. Obesity, diabetes, and risk of cardiovascular disease in the elderly. *Am J Geriatr Cardiol.* (2002) 11:119–24. doi: 10.1111/j.1076-7460.2002.00998.x
62. Abdullah A, Peeters A, de Courten M, Stoelwinder J. The magnitude of association between overweight and obesity and the risk of diabetes: a meta-analysis of prospective cohort studies. *Diabetes Res Clin Pract.* (2010) 89:309–19. doi: 10.1016/j.diabres.2010.04.012
63. Dietz W, Santos-Burgoa C. Obesity and its implications for COVID-19 mortality. *Obesity.* (2020) 28:1005–5. doi: 10.1002/oby.22818
64. Kruglikov IL, Shah M, Scherer PE. Obesity and diabetes as comorbidities for COVID-19: underlying mechanisms and the role of viral-bacterial interactions. *Elife.* (2020) 9:e61330. doi: 10.7554/eLife.61330
65. Yechezkel M, Weiss A, Rejwan I, Shahmoon E, Ben-Gal S, Yamin D. Human mobility and poverty as key drivers of COVID-19 transmission and control. *BMC Public Health.* (2021) 21:1–13. doi: 10.1186/s12889-021-10561-x
66. Watanabe M, Risi R, Tuccinardi D, Baquero CJ, Manfrini S, Gnessi L. Obesity and SARS-CoV-2: a population to safeguard. *Diabetes Metab Res Rev.* (2020) 36:e3325. doi: 10.1002/dmrr.3325
67. Morgan OW, Bramley A, Fowlkes A, Freedman DS, Taylor TH, Gargiullo P, et al. Morbid obesity as a risk factor for hospitalization and death due to 2009 pandemic influenza A (H1N1) disease. *PLoS ONE.* (2010) 5:e9694. doi: 10.1371/journal.pone.0009694
68. Louie JK, Acosta M, Samuel MC, Schechter R, Vugia DJ, Harriman K, et al. A novel risk factor for a novel virus: obesity and 2009. Pandemic Influenza A (H1N1). *Clin Infect Dis.* (2011) 52:301–12. doi: 10.1093/cid/ciq152
69. Department of Treasury US. *About the CARES Act and the Consolidated Appropriations Act.* (2020). Available online at: https://home.treasury.gov/policy-issues/coronavirus/about-the-cares-act
70. Department of Treasury US. *FACT SHEET: The American Rescue Plan Will Deliver Immediate Economic Relief to Families.* (2021). Available online at: https://home.treasury.gov/policy-issues/coronavirus/about-the-cares-act
71. City of Philladelphia. *Philly Counts: Empowering Residents to Take Action and be Agents of Change in Their Communities.* (2021). Available online at: https://www.phila.gov/programs/philly-counts-2020/

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

SYSTEMATIC REVIEW
published: 24 December 2021
doi: 10.3389/fpubh.2021.784958

Genetic Risk Factors for Alzheimer's Disease in Racial/Ethnic Minority Populations in the U.S.: A Scoping Review

Lindsey Rubin[1], *Lucy A. Ingram*[1]*, *Nicholas V. Resciniti*[2], *Brianna Ashford-Carroll*[1], *Katherine Henrietta Leith*[1], *Aubrey Rose*[3], *Stephanie Ureña*[1], *Quentin McCollum*[4] and *Daniela B. Friedman*[1]

[1] Department of Health Promotion, Education, and Behavior, University of South Carolina, Columbia, SC, United States, [2] Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, United States, [3] School of Medicine, University of South Carolina, Columbia, SC, United States, [4] College of Social Work, University of South Carolina, Columbia, SC, United States

Objectives: As the United States (U.S.) population rapidly ages, the incidence of Alzheimer's Disease and Related Dementias (ADRDs) is rising, with racial/ethnic minorities affected at disproportionate rates. Much research has been undertaken to test, sequence, and analyze genetic risk factors for ADRDs in Caucasian populations, but comparatively little has been done with racial/ethnic minority populations. We conducted a scoping review to examine the nature and extent of the research that has been published about the genetic factors of ADRDs among racial/ethnic minorities in the U.S.

Design: Using an established scoping review methodological framework, we searched electronic databases for articles describing peer-reviewed empirical studies or Genome-Wide Association Studies that had been published 2005–2018 and focused on ADRD-related genes or genetic factors among underrepresented racial/ethnic minority population in the U.S.

Results: Sixty-six articles met the inclusion criteria for full text review. Well-established ADRD genetic risk factors for Caucasian populations including *APOE*, *APP*, *PSEN1,* and *PSEN2* have not been studied to the same degree in minority U.S. populations. Compared to the amount of research that has been conducted with Caucasian populations in the U.S., racial/ethnic minority communities are underrepresented.

Conclusion: Given the projected growth of the aging population and incidence of ADRDs, particularly among racial/ethnic minorities, increased focus on this important segment of the population is warranted. Our review can aid researchers in developing fundamental research questions to determine the role that ADRD risk genes play in the heavier burden of ADRDs in racial/ethnic minority populations.

Keywords: genetic risk factors, Alzheimer's disease, race, ethnicity, minority, review

OPEN ACCESS

Edited by:
Tushar Trivedi,
Regional Medical Center,
United States

Reviewed by:
Xiangzhu Zhu,
Vanderbilt University, United States
Marcos Vasconcelos Pais,
University of São Paulo, Brazil

****Correspondence:***
Lucy A. Ingram
lannang@sc.edu

Specialty section:
This article was submitted to
Life-Course Epidemiology and Social
Inequalities in Health,
a section of the journal
Frontiers in Public Health

Received: *29 September 2021*
Accepted: *23 November 2021*
Published: *24 December 2021*

Citation:
Rubin L, Ingram LA, Resciniti NV, Ashford-Carroll B, Leith KH, Rose A, Ureña S, McCollum Q and Friedman DB (2021) Genetic Risk Factors for Alzheimer's Disease in Racial/Ethnic Minority Populations in the U.S.: A Scoping Review. Front. Public Health 9:784958. doi: 10.3389/fpubh.2021.784958

INTRODUCTION

As the United States (U.S.) population rapidly ages, the incidence of Alzheimer's Disease and related dementias (ADRD) is on the rise (1, 2). Alzheimer's Disease (AD) is the sixth leading cause of death in the U.S. and the fifth leading cause of death for those age 65 years and older (1, 2). In the U.S., 5.7 million people are living with AD, which is projected to grow to 13.9 million adults (3.3% of the population) by 2060 (2). Although the primary risk factor for ADRD is age, race, and ethnicity are also associated with ADRD (2–4).

The U.S. population is becoming more racially and ethnically diverse, with Census projections showing that the country will be a "majority–minority" nation by 2050. That is, racial/ethnic minorities will comprise more than 50% of the population by this date (5). African Americans are twice as likely as Non-Hispanic Whites to have AD, while Hispanics are 1.5 times as likely to have AD compared to their Non-Hispanic White counterparts (1). Also by 2050 in the U.S., it is estimated that the proportion of racial/ethnic minorities who suffer from AD will double in size compared to current figures (6). Regarding rates of diagnoses, in particular, African Americans are diagnosed later in the course of ADRD than White patients. Quinones et al. (7) suggest that this is likely due to cultural factors and normalization of ADRD symptoms as part of the usual aging process. There are also noted disparities in cognitive decline and impairment with racial/ethnic minorities suffering greater cognitive decline after ADRD diagnosis compared to other groups (8–10), potentially related to socio-economic resources, such as education quality, development of cognitive reserve, financial means, and early midlife stressors (7). Racial/ethnic health disparities in the U.S. proliferate, are multilayered, and are rooted in a variety of structural and historical inequalities that continue to disproportionately burden racial/ethnic minorities. These disparities underscore the need to examine the factors underlying ADRD in racial/ethnic minority U.S. populations.

As our population ages and the size of our minority populations increase in the U.S., understanding the burden of ADRD on our aging populations can aid in providing insight into the most appropriate and effective public health actions. For example, to provide the best care and community support for aging minority populations, it is valuable to understand any patterns of genetic risk factors to address comorbid disease management, environmental, and socio-economic factors that may affect ADRD prevention, diagnosis, and progression. Similarly, more precise knowledge of differences in prevalence of ADRD in minority populations is useful for policy planning when allocating resources, ensuring access to care, and improving quality of care (11).

Much research has been undertaken to test, sequence, and analyze genetic risk factors for ADRD in White populations, but comparatively little has been done with racial/ethnic minority populations (12). In fact, examining genetic factors in heath disparities research has sometimes led to intense controversy (13, 14), oftentimes for concern of the racialization of medicine, misuse of pharmacogenomics, and racial biology (15–17). In studies that have explored ADRD genetic risk factors in minorities, the study sizes have been relatively small, making the conclusions about genetic associations less powerful. Some data appears to show differences in the genetic etiologies between Caucasians and African Americans, especially relating to the *APOE* gene, which needs to be explored further (11, 18, 19).

There are multiple types of AD classified by age at onset and method of inheritance. The two main categories from a genetic perspective are Early Onset Alzheimer's Disease (EOAD) and Late Onset Alzheimer's Disease (LOAD). According to the National Institute on Aging website, EOAD is also referred to as Familial Alzheimer's Disease and follows an autosomal dominant inheritance pattern, meaning that only one allele from either parent is required to cause disease. EOAD is caused by mutations in three genetic loci, *APP, PSEN1,* and *PSEN2* (20–22). Late Onset Alzheimer's Disease, which is also referred to as Sporadic Alzheimer's Disease, is polygenic, meaning that multiple genes along with environmental factors contribute to the risk of AD, age of onset, and severity of disease (20, 21). *APOE* is one of the most well-established genetic risk factors for LOAD and has implications for risk of other types of AD (20, 21).

The *APOE* e4 allele is a strong risk factor for Sporadic or Late-Onset Familial AD, with the degree of risk increased with two copies of the allele (homozygous e4/e4), but possession of an e4 allele is not in itself necessary to produce AD or sufficient alone to cause the disease (23). Homozygosity in genetics refers to an individual with two copies of the same allele at a particular genetic loci or gene, while heterozygosity refers to the presence of two different alleles at a loci or gene (24). The effects of homozygosity and heterozygosity for the e4 allele has been studied extensively in European American populations, with homozygotes having a 12 times increased risk of LOAD, and heterozygotes having a 2–3 times increased risk of LOAD (18, 19). In African American populations and Hispanic populations, e4 heterozygosity or homozygosity does not correlate with increased risk of AD, indicating that other genetic and environmental factors are responsible for the increased incidence and prevalence of AD in these populations (25–27).

Examining genetic risk factors for ADRDs in minority populations can deepen our understanding of the interaction between biological or genetic factors and socio-ecological determinants of health. It also has the potential to aid in preventive care and early diagnosis for these populations with greater incidence of ADRDs (28). To better understand the risk profile of racial/ethnic minorities who are impacted by ADRD, research should be conducted to comprehend the disease mechanism in these populations, including influential genetic risk factors. If advances in genomic medicine continue to be valid, reliable, and promising, racial/ethnic minorities should be afforded the opportunities to participate in research at similar rates as their White counterparts (13). Other systematic reviews have been conducted in this general subject area. These reviews have had a more segmented focus, with some examining one specific gene and others focusing on a specific population (29, 30). Additional scoping or systematic reviews were focused on a single type of ADRD, such as Lewy Body Dementia or LOAD (31, 32). To explore this gap in the literature, we conducted a scoping review to examine the nature and extent of research that has been published about the genetic factors of ADRDs among racial/ethnic minorities in the U.S.

METHODS

Search Strategy and Selection Criteria

Our study protocol was developed using the established and peer-reviewed scoping review methodological framework and updated based on prior ADRD-focused scoping reviews (33–38). Scoping reviews are a useful format used to explore fields of study not already well-explored or defined. A scoping review is a "preliminary assessment of the potential size and scope of available research literature. It aims to identify the nature and extent of research evidence" [(39), p. 31]. Scoping reviews can be utilized for a variety of research purposes including discovering the scope of existing research in a field of study, in order to identify gaps in the literature for future study. Scoping reviews can also be used to explore the need for a systematic review and the potential value of a systematic review (34, 38).

The databases used to conduct the search were PubMed, CINAHL, and Science Direct. We chose to limit the search to those articles published from 2005 to 2019, as 2005 is when next generation DNA sequencing was available, allowing for more extensive genetic studies with larger sample sizes (40). We conducted a search within the databases using a combination of three concepts: (1) ADRD Genes, (2) Populations and Minority Groups, and (3) ADRDs. The search used a combination of terms from the three concepts to find articles relevant to our research questions. Specific ADRD candidate gene terms were chosen by recent data from Genome Wide Association Studies (GWAS) (41, 42). Some included terms were: *APOE*, beta Amyloid Protein Precursor, *CD2AP*, Genetic Predisposition to Disease, *PSEN1, PSEN2, STM2, APP, TREM2,* African American, Alaska Native, Arabs, Asian American, Ethnic Groups, Hispanic American, Native American, Jews, Minority Groups, Alzheimer's Disease, Dementia, Lewy Bodies, Lewy Body Disease. Inclusion criteria for the review were (1) articles published after January 1, 2005, (3) available in English, (3) peer reviewed empirical studies or Genome-Wide Association Studies (GWAS) (4) that focus on or include an underrepresented minority population in the U.S., (5) that focus on ADRDs, and (6) that focus on ADRD-related genes or genetic factors.

Data Extraction and Synthesis

The study selection process included three interrelated steps: Title/abstract reviews, full-article reviews, and reviewers' examination of reference lists from full articles to identify articles for possible inclusion (43). First, five out of nine of our team members were randomly assigned to review the 1,134 article titles and abstracts in Covidence systematic review online software, with each abstract randomly assigned to two reviewers. Two team members were designated as arbitrators for review discrepancies. When a discrepancy occurred between reviewers (e.g., one "Yes, include in the review" and one "No, do not include in the review"), the designated team members arbitrated the discrepancy. When both randomly assigned reviewers marked an abstract as "Yes" for inclusion, Covidence automatically moved it into the full article review list. Once all titles and abstracts were reviewed twice and all discrepancies arbitrated, the research team then performed a complete review of the resulting 115 articles. Seven team members were randomly assigned a set of articles for full review and the same inclusion and exclusion criteria were used. A data abstraction tool was developed to facilitate review of the full articles and to abstract relevant data. The tool included 21 questions to aid in summarizing the key characteristics of each article. Discrepancies on final article selection and data extraction were then arbitrated by two team members with consultation with the rest of the research team. Once all full articles had been determined, the abstracted data were converted to a Microsoft Excel file for management.

RESULTS

Studies Identified

From the searches in all three databases there were a total of 1,891 articles and 14 additional articles identified from reference lists, for a total of 1,905. We removed 771 duplicates, for a total 1,134 articles for the abstract review stage. During the title abstract review we excluded 1,019 articles due to the following reasons: published outside of the date range, article not available in English, dissertation, metanalysis, systematic review, scoping review, not focused on ADRD, not focused on minority U.S. population, not focused on ADRD genetic factors. After title abstract review, 115 articles remained for full text review. An additional 49 articles were excluded during the full-text review stage if the criteria were not met through examination of the full article. The full text review resulted in 66 included articles (see **Figure 1**).

Populations and Genes Examined

Tables 1, **2** present the general characteristics of the studies included in the full-text review. **Table 3** presents a detailed listing of the characteristics of the articles that were included in the full-text review. Among the resulting 66 studies, most of the studies ($n = 41$, 62%) were focused on African Americans as the population of interest followed by those focusing on the Hispanic population ($n = 28$, 42%). Asian American populations were examined in seven out of the 66 studies (11%), and Native American/Alaska Natives populations were included in only one study (1.5%) (**Table 1**).

There were many different study designs represented in our results. The most common study design was a case control study design, with 18 included articles using this design. The next most frequently found study design was cross-sectional with 15 included studies in this category. There were nine GWAS which is expected because candidate risk genes for ADRD in minority populations have not been fully established. There were five longitudinal studies in the results and two case studies. Lastly, there were five studies that could not be classified into one of these categories (**Table 1**).

Many different types of ADRDs were represented in our search results. The most frequently examined type of AD in our results was LOAD ($n = 26$, 40%), followed by AD ($n = 12$, 18%) and EOAD ($n = 7$, 11%). Vascular Dementia was the focus of four articles out of the total 66 results ($n = 4$, 6%). Both Mild Cognitive Impairment (MCI) and Cognitive Decline were examined in two

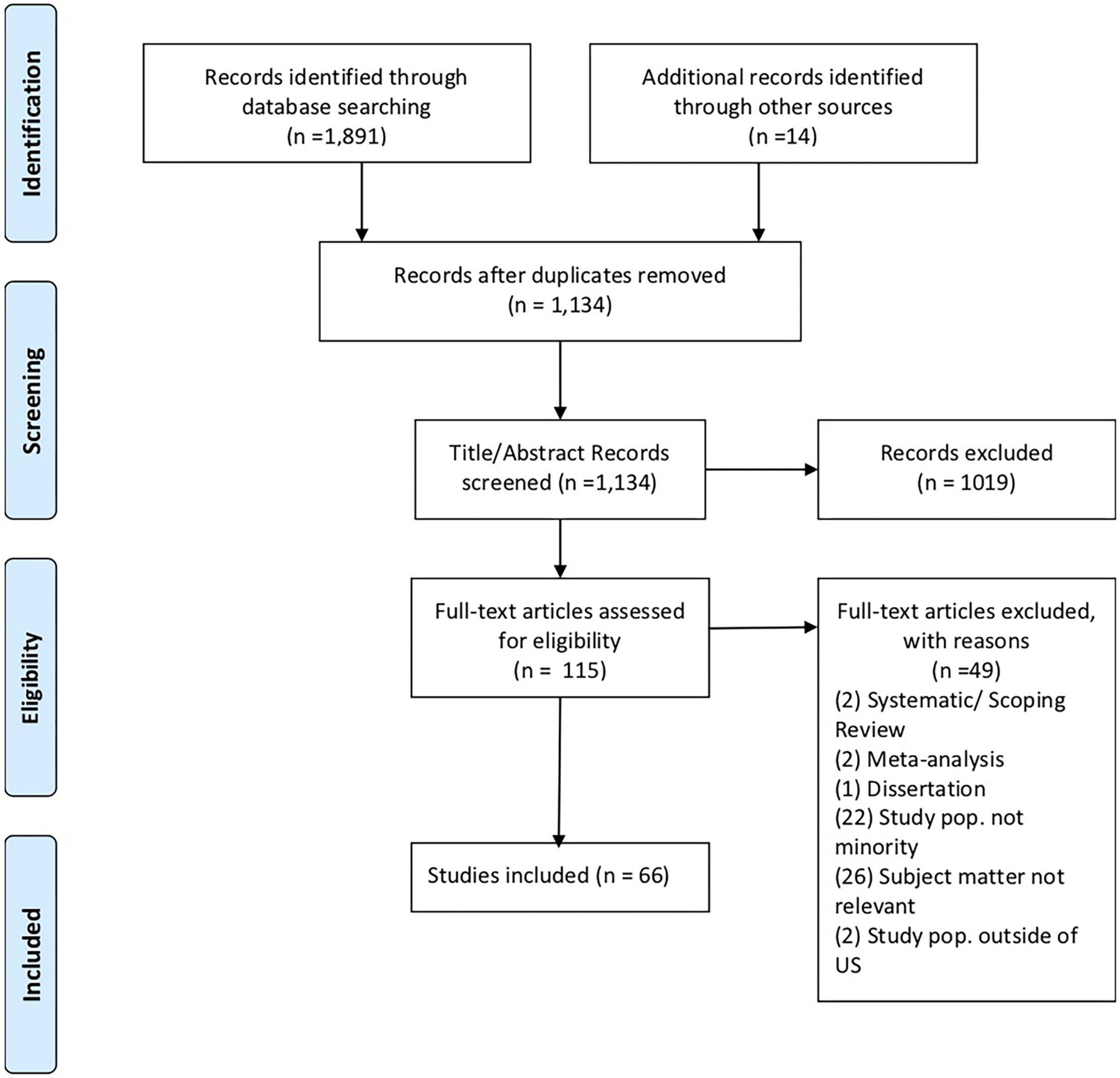

Identification
Screening
Eligibility
Included
Records identified through database searching (n =1,891)
Additional records identified through other sources (n =14)
Records after duplicates removed (n = 1,134)
Title/Abstract Records screened (n =1,134)
Records excluded (n = 1019)
Full-text articles assessed for eligibility (n = 115)
Full-text articles excluded, with reasons (n =49)
(2) Systematic/ Scoping Review
(2) Meta-analysis
(1) Dissertation
(22) Study pop. not minority
(26) Subject matter not relevant
(2) Study pop. outside of US
Studies included (n = 66)

FIGURE 1 | PRISMA chart. Source: Moher et al. (43).

articles each ($n = 2$, 3%). Lewy Body Dementia was the subject of one article ($n = 1$, 1.5%). Lastly, there were 20 articles that did not specify a particular ADRD designation ($n = 20$, 30%) (**Table 2**).

In terms of specific ADRD risk genes, *APOE* was examined in most studies, with 44 out of 66 included studies examining this genetic risk factor. Other potential ADRD risk genes that were examined by multiple studies included *ABCA7, CLU, CR1, PICALM, APP, PSEN1, SORL1* and *AKAP9, APP,* and *PSEN1* are well-established genetic risk factors for EOAD, but in total, they were examined in only eight out of 66 included studies (**Table 2**).

DISCUSSION

Our findings provide an overview of the published literature examining the association between genetic factors and ADRD risk among racial/ethnic minorities in the U.S. These findings help to illuminate knowledge gaps and suggest whether further study should be undertaken to assess more comprehensively the role that ADRD genes play in AD rates and disease outcomes for minority populations.

Regarding the extent of the genes examined in the studies that we found, *APOE* was examined in most studies, with 44

TABLE 1 | Characteristics of studies included in the full-text review ($N = 66$).

Characteristic	Number	Percentage (%)
Publication year		
2005–2006	7	10.6
2007–2008	6	9.1
2009–2010	5	7.6
2011–2012	9	13.6
2013–2014	15	22.7
2015–2016	9	13.6
2017–2018	15	22.7
Race/Ethnicity[a]		
African American	41	62.1
Hispanic American	28	42.4
Asian American	7	10.6
Native American/Alaska Native	1	1.5
Sample size		
0–100	3	4.5
101–500	11	16.7
501–1,000	10	15.2
1,001–1,500	11	16.7
1,501–2,000	7	10.6
2,001–2,500	4	6.1
2,501–3,000	2	3.0
3,001–3,500	2	3.0
3,501–4,000	0	0.0
4,001–4,500	1	1.5
4,501 or more	15	22.7
Type of study		
Case-control	18	27.3
Cross-sectional	15	22.7
Cohort	12	18.2
Genome Wide Association Study (GWAS)	9	13.6
Longitudinal	5	7.6
Other	5	7.6
Case report/case study	2	3.0

[a]Some articles included multiple races/ethnicities in the study sample.

TABLE 2 | Type of ADRD and risk genes identified in full-text review articles ($N = 66$).

Characteristic	Number	Percentage (%)
Type of ADRD[a]		
Lewy Body Dementia	1	1.5
Mild Cognitive Impairment	2	3.0
Cognitive Decline	2	3.0
Vascular Dementia	4	6.1
Early onset AD (EOAD)	7	10.6
Alzheimer's Disease	13	19.7
Type of ADRD not specified	20	30.3
Late Onset AD (LOAD)	26	39.4
ADRD risk genes identified[b]		
PSEN2	1	1.5
AKAP9	2	3.0
GRIN3B	2	3.0
SORL1	2	3.0
CR1	3	4.5
APP	4	6.1
PSEN1	4	6.1
ABCA7	6	9.1
CLU	6	9.1
PICALM	7	10.6
APOE	43	65.2
Other	42	63.6

[a]Some articles examined more than one type of ADRD. [b]Some articles included multiple risk genes.

out of 66 included studies examining this genetic risk factor. This corresponds with extant ADRD genetic risk factor research findings in general, as *APOE* is the most well-established genetic risk factor for Sporadic or LOAD (23). We found that well-established ADRD genetic risk factors for Caucasian populations including *APOE, APP, PSEN1,* and *PSEN2* have not been studied to the same degree in minority U.S. populations. The *APOE* genotype has been shown to be less predictive of ADRD risk in African American, Asian American, Hispanic American, and Native American populations (26, 27, 29, 98). Other genetic risk factors may play a larger role in ADRD genetic risk in these populations, with potential candidates including genes with various functions such as *ABCA7, CLU, CR1, PICALM, SORL1, AKAP9,* and *TREM2* (26, 27, 29, 98, 109). These genes were noted in our review, however with far less frequency than *APOE.* Preliminary findings indicate that there may be a more complex polygenic profile of ADRD genetic risk in these populations, and this has potential implications for the possible polygenic nature of ADRD risk in all populations (27, 59, 87).

In comparison to the amount of research that has been conducted on Caucasians in the U.S., we found that some minority communities were vastly underrepresented in the research, namely Hispanics, Native Americans, and Asian Americans. Though the number of studies on ADRD genetic risk factors in minority populations has increased over time, especially for certain populations such as African Americans, more comprehensive studies with large sample sizes should be performed to establish key genetic risk factors for these populations as well (27, 109–112). Among the studies in our review, sample size for non-GWAS studies started as low as $N = 19$ for a case report design. As the sample size increases and more diverse persons are included, additional, more statistically sound conclusions can be made about the associations between genetic expression and disease outcome.

Additionally, comparative studies with both minority and majority population group samples would be useful in examining genetic risk factors, as well as the effects of environment and other factors. Studies exploring genetic risk factors in these populations is warranted to determine the role that both genes and environmental factors play in increased ADRD risks in these populations. A larger, systematic review of existing literature on genetic risk factors for minority U.S. populations would be

TABLE 3 | Detailed listing of studies included in the full-text review.

Author and year	Study design	URM group	Data source	Sample size*	Type of ADRD	Gene(s) included
Akomolafe et al. (44)	Case-control	African American	MIRAGE Study	511 cases, 679 controls*	EOAD, LOAD	*NOS3, APOE*
Arnold et al. (45)	Cohort	Puerto Rican	Original data	283	EOAD, LOAD	*PSEN1*
Beeri et al. (46)	Longitudinal, cohort	African American	ACCORD-MIND Study	466	Cognitive Decline	*HP*
Borenstein et al. (47)	Prospective, cohort	Japanese American	The Kame Project	1,859	Alzheimer's disease	*APOE*
Borenstein et al. (48)	Prospective, cohort	Japanese American	The Kame Project	1,859	Vascular Dementia and Alzheimer's Disease	*APOE*
Bressler et al. (49)	GWAS	African Americans	The ARIC Study	10,359*	LOAD	*APOE, ABCA7, BIN1, CD2AP, CDS33, CELF1, EPHA1, MS4A4E, NME8, PICALM, PKT2B, ZCWPW1,*
Campos et al. (50)	Case-control	Hispanic Americans, Amerindians	Original data	56 cases, 56 controls*	Alzheimer's Disease	*APOE*
Carrion-Baralt et al. (51)	Cohort	Puerto Ricans	Original data	87	Alzheimer's Disease	*APOE*
Conway et al. (52)	Case-control, targeted sequencing	African Americans	Mayo Clinic	5,924 cases, 5,173 controls*	EOAD, LOAD, Lewy Body Dementia	*ABI3, APOE, PLCG2*
Cukier et al. (53)	Case-control	African Americans, Caribbean Hispanics	HIHG and ADGC data sets	149 cases, 137 controls*	LOAD	*ABC1, ABCA7*
Desai et al. (54)	Case-control	African Americans	ADRC data set	1,059 cases, 716 controls*	LOAD	*BDNF*
Edwards-Lee et al. (55)	Family study	African Americans	Original data	7	EOAD (autosomal dominant)	*APP, PS1, MAPT*
Erlich et al. (56)	Case-control study	African Americans	MIRAGE Study	520 cases, 677 controls*	Alzheimer's Disease	*PON1, PON2, PON3*
Fitten et al. (57)	Cross-sectional study	Hispanic Americans	ADRC data set, OVMC data set	290*	Alzheimer's Disease, Vascular Dementia	*APOE*
Ghani et al. (58)	Case-control, GWAS	Hispanic Americans	Washington Heights-Inwood Columbia Aging Project, Estudio Familiar de Influencia Genetica de Alzheimer Study	547 cases, 542 controls*	LOAD	*APOE, CLU, PICALM, BIN1*
Gonzalez et al. (59)	Cohort study	Hispanic Americans	The Hispanic Community Health Study/Study of Latinos (HCHS/SOL)	10,887*	Alzheimer's Disease	*APOE*
Harwood et al. (60)	Cross-sectional study	African Americans, Hispanic Americans	Original data	685	Alzheimer's Disease	*APOE*
He et al. (61)	Cross-sectional study	African Americans, Hispanic Americans	Original data	439	Mild Cognitive Impairment (MCI)	*APOE*
Hendrie et al. (62)	Case-control study	African Americans	Original data	221 cases, 218 controls	MCI, Dementia, Alzheimer's Disease	*APOE*
Hohman et al. (63)	Case-control, GWAS	African Americans	ADGC	1,840 cases, 3,804 controls	LOAD	*APOE, STM2, ABCA7, CR1, PICALM, BIN1, EPHA1, CD33, SLC24A4, GRIN3B, FERMT2, MS4A6A*
Janicki et al. (64)	Cohort study	African Americans, Hispanic Americans	Washington Heights Inwood Columbia Aging Project (WHICAP)	1,686*	Alzheimer's Disease	*APOE, CYP19*

(Continued)

TABLE 3 | Continued

Author and year	Study design	URM group	Data source	Sample size*	Type of ADRD	Gene(s) included
Jin et al. (65)	Case-control	African Americans	Knight-ADRC + NIA-LOAD, Mayo Clinic, Indiana University, WHICAP, Emory University	906 cases, 2,487 controls*	LOAD	*TREM2*
Janicki et al. (66)	Prospective Cohort study	African Americans, Hispanic Americans	WHICAP	1,686*	Alzheimer's Disease	*ESR1*
Kim et al. (67)	Longitudinal prospective community-based study	African Americans	IIDP	1,858*	AD, Dementia	*CD2AP, CBS, DTWD2, DYNC111, JRKL-AS1, BIRC8, HCY*
Kuller et al. (68)	Longitudinal cohort study	African Americans	Pittsburgh Cardiovascular Health Study	532	LOAD	*APOE*
Kunkle et al. (69)	Case-control study, GWAS	African Americans	HIHG/CWRU, NIMH Genetic Studies of Alzheimer's Disease Cohort, NCRAD/NIA-LOAD, African American Alzheimer's Disease Genomics Coalition (AAADGC)	2,762 cases, 2,812 controls*	LOAD	*ABCA7*
Kwon et al. (70)	Cohort study	African Americans, Hispanic Americans	Original data	1,309*	LOAD	*APOE*
Lee et al. (71)	Nested Case-control study, prospective	African Americans, Hispanic Americans	Original data	296 cases, 428 controls*	AD	*SORL1*
Lee et al. (72)	Family-based cohort study, GWAS	Caribbean Hispanic	Original data	1,161 individuals from 209 families	Familial LOAD	*APOE, PSEN1,* 5q15, 7q36.3, 14q32.12, 17q25.1, 17p13
Lee (73)	Family-based case-control and unrelated case-control study, GWAS	Caribbean Hispanics	ADRC	693 cases, 442 controls*	LOAD	*APOE,* 12p13
Lee (74)	Nested case-control GWAS	Caribbean Hispanics	WHICAP and EFIGA datasets	549 cases, 544 controls	LOAD	*CLU, PICALM, BIN1, PSEN1, GHITM, C10orf99, PCDH21, LRT2, LRT1, RGR, DGKB, HPCAL1, ODC1*
Lee (75)	Family-based cohort study	Caribbean Hispanics	WHICAP and EFIGA datasets	2,888	EOAD, LOAD	*PSEN1, SNX25, PDLIM3, SORBS2, SH3RF3, NPHP1*
Livney (76)	Cross-sectional study	African American, Hispanic Americans	Original data	1,341	AD	*APOE*
Logue (77)	Case-control study	African Americans	MIRAGE, GenerAAtions, ADNI, GenADA, NIA-LOAD, FHS	3,568 cases, 6,205 controls*		*APOE, PVRL2, CLU, PICALM, BIN1, EPHA1, MS4A, ABCA7, and CD33, TOMM40*
Logue et al. (78)	Case-control	African Americans	MIRAGE Study, GenerAAtions Study	422 cases, 394 controls	EOAD, LOAD	*AKAP9, APOE, BIN1, CLU, CR1, PICALM, MS4A6E, CD2AP, CD33, ABCA7, EPHA1, SORL1, ACE, PSEN1, PSEN2, APP*

(Continued)

TABLE 3 | Continued

Author and year	Study design	URM group	Data source	Sample size*	Type of ADRD	Gene(s) included
Logue et al. (79)	Case-control	African Americans	MIRAGE Study, GenerAAtions Study, National Cell Repository for Alzheimer's Disease (NCRAD), Ibadan/Indianapolis (INDY) Study	489 cases, 472 controls	LOAD	*ABCA7, AKAP9, KIAA0196, KANSL1, CNN2, TRIM2*
Marden et al. (80)	Cohort	African Americans	Health and Retirement Study (HRS)	7,690*	AD and Dementia	*APOE, BIN1, CLU, ABCA7, CR1, PICALM, MS4A6A, CD33, MS4A4E, CD2AP*
Marden et al. (81)	Cohort	African Americans	HRS	8,253*	AD	*APOE, CLU, CR1, PICALM*
McAninch et al. (82)	Cohort	African Americans	Original data	12,348*	AD	*DIO2*
Melville et al. (83)	Case-control	African Americans	MIRAGE Study, ADNI Study	1,146 cases, 956 controls*	AD, MCI	*APOE, PICALM, F5/SELP, LHFP, GCFC2, SYNPR,* TTC27
Mez et al. (84)	Case-control	African Americans	ADGC, GenerAAtions, MIRAGE, CHAP	1,825 cases, 3,784 controls	LOAD	*APOE, ABCA7, COBL, SLC10A2*
Mount et al. (85)	Cross-sectional, retrospective	African Americans	ADCR	65	LOAD	*APOE*
Murrell et al. (86)	Cohort	African Americans	Original data	480	LOAD	*APOE*
N'Songo et al. (87)	Cohort	African Americans	Original data	198 cases, 350 controls	EOAD	*APP, PSEN1, PSEN2*
O'Bryant et al. (88)	Cohort	Mexican Americans	Project FRONTIER, TARCC	1,628	MCI	*APOE*
O'Bryant et al. (89)	Cohort, CBPR	Mexican Americans	Project FRONTIER, TARCC	1,069*	MCI, AD	*APOE*
Olarte et al. (90)	Population-based, case series	Hispanics	HCFA	680	Sporadic and familial AD	*APOE*
Pedraza et al. (91)	Case-control	African Americans	Mayo Clinic Alzheimer's Disease Research Center Data, Mayo Clinic Study of Aging, Mayo Clinic LOAD-GWAS	476 cases, 2,443 controls*	LOAD	*CLU, CR1, PICALM*
Peila et al. (92)	Nested case-control	Japanese-Americans	Honolulu-Asia Aging Study (HAAS), Honolulu Heart Program (HHP)	283 cases, 573 controls	AD, Vascular Dementia	*APOE, TGF-β1*
Petrovich et al. (93)	Longitudinal, cohort	Japanese-Americans	The Honolulu-Asia Aging Study	375	AD	*APOE*
Qian et al. (94)	Prospective, cohort	Latinos	NACC, Rotterdam Study, Framingham Heart Study, and Sacramento Area Latino Study (SALSA)	16,844*	AD	*APOE*
Rajabli et al. (95)	Case-control	African Americans, Hispanic Americans	HGDP (Human Genome Diversity Project)	1,986 cases, 3,899 controls*	LOAD	*APOE*
Reitz et al. (96)	Case-control	African Americans and Caribbean Hispanics	Toronto dataset, NIA-LOAD, MIRAGE Caucasian dataset, MIRAGE African American dataset, Miami Caucasian, Caribbean Hispanic dataset	2,809 cases, 3,482 controls	AD	*SORCS1, APP, Aβ, SORL1*

(Continued)

TABLE 3 | Continued

Author and year	Study design	URM group	Data source	Sample size*	Type of ADRD	Gene(s) included
Reitz et al. (97)	Case-control	Caribbean Hispanics	DMS-IV, NINCDS-ADRDA	160 cases, 294 controls	LOAD	*IDE, KIF1, HHEX*
Reitz et al. (98)	Case-control	African Americans	CHAP, MARS/CORE, UM/VU	1,968 cases, 3,928 controls	LOAD	*ABCA7, APOE*
Rippon et al. (99)	Family-based cohort study	Latinos	NINDCS-ADRDA	1,498	Familial AD	*APOE*
Roses et al. (100)	Cohort	African Americans, Japanese Americans	Bryan ADRC Database/Repository, Coriell Cell Repositories	447*	LOAD	*TOMM40, APOE*
Sacyzynsky et al. (101)	Cohort	Japanese-Americans	The Honolulu Heart Program, Cooperative Lipoprotein Study	929	Dementia	*APOE*
Sawyer et al. (102)	Prospective cohort	African Americans	Duke EPESE Study	2,076*	Cognitive decline (CD)	*APOE*
Simino et al. (103)	Cohort	African Americans	CHARGE, the NHLBI Exome Sequencing Project	1,414*	AD	*Amyloid*-β, KLKB1, F12, PLIN2, ITPRIP
Tosto et al. (104)	Cohort	Caribbean Hispanics	NIA-LOAD, EFIGA	8,116*	LOAD	*APOE* ε4
Vardarajan et al. (105)	Case-control	African Americans	ADGC	8,309 cases, 7,366 controls*	AD	*APP, KIAA1033, SNX1, SORL1, SNX3, RAB7A*
Vardarajan et al. (106)	Family and cohort-based genetic association study	Caribbean Hispanics	Original data	464 familial subjects—(350 affected, 114 unaffected), 498 unrelated controls	LOAD	*SORL1*
Weiner et al. (107)	Case-control	Choctaw Indians	Original data (Choctaw Indians) and UT Southwestern Alzheimer's Disease Center (ADC)	78 cases, 39 controls*	AD	*APOE*
Yu et al. (108)	Longitudinal, cohort	African Americans	Religious Orders Study (ROS), Rush Memory and Aging Project (MAP), Minority Aging Research Study (MARS)	2,388*	AD	*APOE, TOMM40*

**Article included multiple races/ethnicities in the study sample.*

an appropriate next step in better understanding the existing study landscape with intentions toward implementing GWAS and meta-analyses for diverse U.S. populations.

Knowledge gaps in the disease mechanism among racial/ethnic minority populations is a critical indicator of inequities in genetics and genomics research in these communities, as well as a lack of equity in the health care system for these groups (112). Advancements in genetic medicine and genomic research proliferate, unfortunately not at the same rate for all persons. The impact that disproportionate expansion, innovation, and progress in the field can have on health disparities is significant (12, 112). With that in mind, it is also important to acknowledge that while genetic inquiry is crucial to understanding the disparities present in ADRD, it is not the sole risk factor. Other factors such as environment and socio-environmental context, are implicated in the distribution of racial health disparities, and in fact, the complex interplay of all these factors contribute to many disease outcomes (12, 113, 114).

Of additional consideration as an important implication of this research, particularly for minority populations, is the potential of stigma related to ADRD diagnosis. Some groups have been found to consider dementia as a normal part of aging (115), while others may find shame in an AD diagnosis or the need to keep such health information private (116–118). We highlight these studies as further evidence of the need to focus research in racially and ethnically diverse communities. Furthermore, we acknowledge that such research should consider both quantitative and qualitative approaches.

This study is not without limitations. First, while we conducted a systematic and structured process for the scoping review, we did not evaluate the quality of the evidence presented or the authors' research methods as part of this review. Second, some studies more clearly identified the characteristics of interest for our review than others, and as such, some of the data presented was left to the interpretation of the authorship team. Third, we acknowledge that there is limited generalizability of our findings to research that has been conducted in the U.S. among racial/ethnic minorities. That said, we find that an important strength of this review is in identifying the knowledge gaps in examining and understanding the genetic factors associated with ADRD among racial/ethnic minority populations, which is of growing disease and economic burden in the U.S.

CONCLUSION

Based our findings, we recommend that additional studies be undertaken to map out and more deeply explore ADRD genetic risk factors among racial/ethnic minority populations in the U.S. at levels comparable to non-minority populations. An increased number of larger scale studies of racially/ethnically diverse persons can aid researchers in making more powerful conclusions about genetic associations in ADRD among populations most affected. Examining genetic risk factors for ADRDs in minority populations can deepen our understanding of the interaction between biological or genetic factors and socio-ecological determinants of health. Furthermore, understanding the role of genetic predisposing factors has the potential to increase preventive health measures and screening, which could lead to reduced time to diagnosis and improved ADRD disease management. Lastly, ethical concerns about the impact that this knowledge of genetic risk factors may have on the health and well-being of individuals must be addressed as we continue to obtain more data on these genetic factors. As our population ages and the size of our minority populations increase in the U.S., understanding the burden of ADRD on our aging populations can aid in providing insight into the most appropriate and effective public health actions.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

LR conceptualized the study, was a scoping reviewer, and contributed to the manuscript narrative. LI and DF were scoping reviewers, contributed to the manuscript narrative, and helped to edit the manuscript. NR was a scoping reviewer and contributed to the manuscript narrative. BA-C, AR, KL, SU, and QM were scoping reviewers and helped to edit the manuscript. All authors contributed to the article and approved the submitted version.

REFERENCES

1. *Alzheimer's Association Report: 2020 Alzheimer's Disease Facts and Figures*. *Alzheimers Dement*. (2020). 16:391–460. doi: 10.1002/alz.12068
2. Hebert LE, Weuve J, Scherr PA, Evans DA. Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology*. (2013). 80:1778–83. doi: 10.1212/WNL.0b013e31828726f5
3. Matthews KA, Xu W, Gaglioti AH, Holt JB, Croft JB, Mack D, et al. Racial and Ethnic Estimates of Alzheimer's Disease and Related Dementias in the United States (2015-2060) in Adults Aged ≥ 65 years. *Alzheimers Dement.* (2019) 15:17–24. doi: 10.1016/j.jalz.2018.06.3063
4. Chen C, Zissimopoulos JM. Racial and ethnic differences in trends in dementia prevalence and risk factors in the United States. *Alzheimers Dement.* (2018) 4:510–20. doi: 10.1016/j.trci.2018.08.009
5. Steenland K, Goldstein FC, Levey A, Wharton W. A meta-analysis of Alzheimer's disease incidence and prevalence comparing African-Americans and Caucasians. *J Alzheimers Dis.* (2016) 50:71–6. doi: 10.3233/JAD-150778
6. Colby SL, Ortman JM. Projections of the size and composition of the U.S. population: 2014 to 2060. Population estimates and projections: current population reports. Washington, DC: US Census Bureau (2015). Available online at: http://www.census.gov/content/dam/Census/library/publications/2015/demo/p25-1143.pdf (accessed March 22, 2021).
7. Shin J, Doraiswamy PM. Underrepresentation of African Americans in Alzheimer's trials: a call for affirmative action. *Front Aging Neurosci.* (2016) 8:123. doi: 10.3389/fnagi.2016.00123
8. Quiñones AR, Kaye J, Allore HG, Botoseneanu A, Thielke SM. An agenda for addressing multimorbidity and racial and ethnic disparities in Alzheimer's disease and related dementia. *Am J Alzheimers Dis Other Demen.* (2020) 35:153331752096087. doi: 10.1177/1533317520960874
9. Peterson RL, Fain MJ, Butler AE, Ehiri JE, Carvajal SC. The role of social and behavioral risk factors in explaining racial disparities in age-related cognitive impairment: a structured narrative review. *Aging, Neuropsychol Cogn.* (2019) 27:173–96. doi: 10.1080/13825585.2019.1598539
10. Weuve J, Barnes LL, Mendes de Leon CF, Rajan KB, Beck T, Aggarwal NT, et al. Cognitive aging in black and white Americans. *Epidemiology.* (2018) 29:151–9. doi: 10.1097/EDE.0000000000000747
11. Zuelsdorff M, Okonkwo OC, Norton D, Barnes LL, Graham KL, Clark LR, et al. Stressful life events and racial disparities in cognition among middle-aged and older adults. *J Alzheimers Dis.* (2020) 73:671–82. doi: 10.3233/JAD-190439
12. Froehlich TE, Bogardus ST Jr, Inouye SK. Dementia and race: are there differences between African Americans and Caucasians?. *J Am Geriatr Soc.* (2001) 49:477–84. doi: 10.1046/j.1532-5415.2001.49096.x
13. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet.* (2019) 51:584–91. doi: 10.1038/s41588-019-0379-x
14. Moonesinghe R, Jones W, Honoré PA, Truman BI, Graham G. Genomic medicine and racial/ethnic health disparities: promises, perils, and the challenges for health care and public health policy. *Ethn Dis.* (2009) 19:473–8.
15. Hunt LM, Megyesi MS. Genes, race and research ethics: who's minding the store?. *J Med Ethics*. (2008) 34:495–500. doi: 10.1136/jme.2007.021295
16. Lee SS. Pharmacogenomics and the challenge of health disparities. *Publ Health Genomics.* (2009) 12:170–9. doi: 10.1159/000189630
17. Rotimi CN. Are medical and nonmedical uses of large-scale genomic markers conflating genetics and 'race'?. *Nat Genet*. (2004) 36(11 Suppl):S43–7. doi: 10.1038/ng1439
18. Scharff DP, Mathews KJ, Jackson P, Hoffsuemmer J, Martin E, Edwards D. More than Tuskegee: understanding mistrust about research participation. *J Health Care Poor Underserved.* (2010) 21:879–97. doi: 10.1353/hpu.0.0323
19. Berg CN, Sinha N, Gluck MA. The effects of APOE and ABCA7 on cognitive function and Alzheimer's disease risk in African Americans: a focused mini review. *Front Hum Neurosci.* (2019) 13:387. doi: 10.3389/fnhum.2019.00387
20. Potter H, Wisniewski T. Apolipoprotein E: essential catalyst of the Alzheimer amyloid cascade. *Int J Alzheimers Dis.* (2012) 2012:489428. doi: 10.1155/2012/489428
21. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet.* (2007) 39:17–23. doi: 10.1038/ng1934
22. Bertram L, Tanzi RE. Alzheimer disease risk genes: 29 and counting. *Nat Rev Neurol.* (2019) 15:191–2. doi: 10.1038/s41582-019-0158-4
23. Cacace R, Sleegers K, Van Broeckhoven C. Molecular genetics of early-onset Alzheimer's disease revisited. *Alzheimers Dement.* (2016) 12:733–48. doi: 10.1016/j.jalz.2016.01.012

24. Roberts JS, Patterson AK, Uhlmann WR. Genetic testing for neurodegenerative diseases: Ethical and health communication challenges. *Neurobiol Dis.* (2020) 141:104871. doi: 10.1016/j.nbd.2020.104871
25. Elston RC, Satagopan JM, Sun S. Genetic terminology. *Methods Mol Biol.* (2012) 850:1–9. doi: 10.1007/978-1-61779-555-8_1
26. Graff-Radford NR, Besser LM, Crook JE, Kukull WA, Dickson DW. Neuropathologic differences by race from the National Alzheimer's Coordinating Center. *Alzheimers Dement.* (2016) 12:669–77. doi: 10.1016/j.jalz.2016.03.004
27. Heffernan AL, Chidgey C, Peng P, Masters CL, Roberts BR. The neurobiology and age-related prevalence of the ε4 allele of apolipoprotein E in Alzheimer's Disease Cohorts. *J Mol Neurosci.* (2016) 60:316–24. doi: 10.1007/s12031-016-0804-x
28. Mez J, Marden JR, Mukherjee S, Brewster P, Hamilton JL, Gilsanz P, et al. P2-076: Alzheimer's disease genetic risk variants beyond Apoe ε4 predict mortality in the adult changes in thought (ACT) study. *Alzheimers Dement.* (2016) 12:P637–8. doi: 10.1016/j.jalz.2016.06.1281
29. Berkowitz CL, Mosconi L, Rahman A, Scheyer O, Hristov H, Isaacson RS. Clinical application of APOE in Alzheimer's prevention: a precision medicine approach. *J Prev Alzheimers Dis.* (2018) 5:245–52. doi: 10.14283/jpad.2018.35
30. Tang M-X, Stern Y, Marder K, Bell K, Gurland B, Lantigua R, et al. The APOE-epsilon4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics. *JAMA.* (1998) 279:751–5. doi: 10.1001/jama.279.10.751
31. Huang M, Wang D, Xu Z, Xu Y, Xu X, Ma Y, et al. Lack of genetic association between TREM2 and Alzheimer's disease in East Asian population: a systematic review and meta-analysis. *Am J Alzheimers Dis Other Dement.* (2015) 30:541–6. doi: 10.1177/1533317515577128
32. Huq AJ, Fransquet P, Laws SM, Ryan J, Sebra R, Masters CL, et al. Genetic resilience to Alzheimer's disease in APOE ε4 homozygotes: a systematic review. *Alzheimers Dement.* (2019) 15:1612–23. doi: 10.1016/j.jalz.2019.05.011
33. Sanghvi H, Singh R, Morrin H, Rajkumar AP. Systematic review of genetic association studies in people with Lewy Body Dementia. *Int J Geriatr Psychiatry.* (2020) 35:436–48. doi: 10.1002/gps.5260
34. Andrews SJ, McFall GP, Booth A, Dixon RA, Anstey KJ. Association of Alzheimer's disease genetic risk loci with cognitive performance and decline: a systematic review. *J Alzheimers Dis.* (2019) 69:1109–36. doi: 10.3233/JAD-190342
35. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol.* (2005) 8:19–32. doi: 10.1080/1364557032000119616
36. Peters MD, Godfrey CM, Khalil H, McInerney P, Parker D, Soares CB. Guidance for conducting systematic scoping reviews. *Int J Evid Based Healthc.* (2015) 13:141–6. doi: 10.1097/XEB.0000000000000050
37. Friedman DB, Becofsky K, Anderson LA, Bryant LL, Hunter RH, Ivey SL, et al. Public perceptions about risk and protective factors for cognitive health and impairment: a review of the literature. *Int Psychogeriatr.* (2015) 27:1263–75. doi: 10.1017/S1041610214002877
38. Resciniti NV, Tang W, Tabassum M, Pearson JL, Spencer SM, Lohman MC, et al. Knowledge evaluation instruments for dementia caregiver education programs: a scoping review. *Geriatr Gerontol Int.* (2020) 20:397–413. doi: 10.1111/ggi.13901
39. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. Prisma Extension for Scoping Reviews (PRISMA-SCR): checklist and explanation. *Ann Intern Med.* (2018) 169:467–73. doi: 10.7326/M18-0850
40. Grant MJ, Booth A. A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Info Libr J.* (2009) 26:91–108. doi: 10.1111/j.1471-1842.2009.00848.x
41. Ureña S, Ingram LA, Leith K, Lohman MC, Resciniti N, Rubin L, et al. Mentorship and training to increase diversity of researchers and practitioners in the field of aging and Alzheimer's disease: a scoping review of program characteristics. *J Aging Health.* (2021) 33:48–62. doi: 10.1177/0898264320953345
42. Tanzi RE, Bertram L. New frontiers in Alzheimer's disease genetics. *Neuron.* (2001) 32:181–4. doi: 10.1016/S0896-6273(01)00476-7
43. Kilpinen H, Barrett JC. How next-generation sequencing is transforming complex disease genetics. *Trends Genet.* (2013) 29:23–30. doi: 10.1016/j.tig.2012.10.001
44. Akomolafe A, Lunetta KL, Elrich PM, Cupples, LA, Baldwin CT, Huyck M, et al. Genetic association between endothelial nitric oxide synthase and Alzheimer disease. *Clin Genet.* (2006) 70:49–56. doi: 10.1111/j.1399-0004.2006.00638.x
45. Arnold SE, Vega IE, Karlawish JH, Wolk DA, Nunez J, Negron M, et al. Frequency and clinicopathological characteristics of presenilin 1 Gly206Ala mutation in Puerto Rican Hispanics with dementia. *J Alzheimers Dis.* (2013) 33:1089–95. doi: 10.3233/JAD-2012-121570
46. Beeri MS, Lin HM, Sano M, Springer RR, Liu X, Bendlin BB, et al. Association of the haptoglobin gene polymorphism with cognitive function and decline in elderly african american adults with type 2 diabetes: findings from the action to control cardiovascular risk in diabetes-memory in diabetes (accord-mind) study. *JAMA Netw Open.* (2018) 7:e184458. doi: 10.1001/jamanetworkopen.2018.4458
47. Borenstein AR, Wu Y, Mortimer JA, Schellenberg GD, McCormick WC, Bowen JD, et al. Developmental and vascular risk factors for Alzheimer's disease. *Neurobiol Aging.* (2005) 26:325–34. doi: 10.1016/j.neurobiolaging.2004.04.010
48. Borenstein AR, Wu Y, Bowen JD, McCormick WC, Uomoto J, McCurry SM, et al. Incidence rates of dementia, Alzheimer disease, and vascular dementia in the Japanese American population in Seattle, WA: the Kame Project. *Alzheimer Dis Assoc Disord.* (2014) 28:23–9. doi: 10.1097/WAD.0b013e3182a2e32f
49. Bressler J, Mosley TH, Penman A, Gottesman RF, Windham BG, Knopman DS,et al. Genetic variants associated with risk of Alzheimer's disease contribute to cognitive change in midlife: The atherosclerosis risk in communities study. *Am J Med Genet B Neuropsychiatr Genet.* (2017) 174:269–82. doi: 10.1002/ajmg.b.32509
50. Campos M, Edlan S, Peavy G. An exploratory study of APOE-*ε*4 genotype and risk of Alzheimer's disease in Mexican Hispanics. *J Am Geriatr Soc.* (2013) 61:1038-y40. doi: 10.1111/jgs.12292
51. Carrión-Baralt J, Meléndez-Cabrero J, Rodríguez-Ubiñas H, Schmeidler J, Beei M, Angelo G, et al. Impact of APOE *ε*4 on the cognitive performance of a sample of non-demented Puerto Rican Nonagenarians. *J Alzheimer Dis.* (2009) 18:533–40. doi: 10.3233/JAD-2009-1160
52. Conway O, Carrasquillo M, Wang X, Bredenberg J, Reedy J, Strickland S, et al. ABI3 and PLCG2 missense variants as risk factors for neurodegenerative diseases in Caucasians and African Americans. *Mol. Neurodegener.* (2018) 13. doi: 10.1186/s13024-018-0289-x
53. Cukier H, Kunkle B, Vardarajan B, Rolati S, Hamilton-Nelson K, Kohli M, et al. ABCA7 frameshift deletion associated with Alzheimer disease in African Americans. *Neurol. Genet.* (2016) 2:34–8. doi: 10.1212/NXG.0000000000000079
54. Desai P, DeKosky S, Kamboh I. Genetic variation in the cholesterol 24-hydroxylase (CYP46) gene and the risk of Alzheimer's disease. *Neurosci. Lett.* (2002) 328:9–12. doi: 10.1016/s0304-3940(02)00443-3
55. Edwards-Lee T, Ringman JM, Chung J, Werner J, Morgan A, Hyslop G, et al. An African American family with early-onset Alzheimer disease and an APP (T714I) mutation. *Neurology.* (2005) 64:23. doi: 10.1212/01.WNL.0000149761.70566.3E
56. Elrich P, Lunetta K, Cupples A, Huyck M, Green R, Baldwin C, et al. Polymorphisms in the PON gene cluster are associated with Alzheimer disease. *Hum Mol Genet.* (2006) 15:77–85. doi: 10.1093/hmg/ddi428
57. Fitten J, Ortiz F, Fairbanks L, Bartzokis G, Lu P, Klein E, et al. Younger age of dementia diagnosis in a Hispanic population in southern California. *Int J Geriatr.* (2014) 29:586–93. doi: 10.1002/gps.4040
58. Ghani M, Sato C, Lee J, Reitz C, Moreno D, Mayeux R, et al. Evidence of recessive Alzheimer disease loci in a Caribbean Hispanic data set: Genome-wide survey of runs of homozygosity. *JAMA Neurol.* (2013) 70:1261–7. doi: 10.1001/jamaneurol.2013.3545
59. González HM, Tarraf W, Jian X, Vasquez PM, Kaplan R, Thyagarajan B, et al. Apolipoprotein E genotypes among diverse middle-aged and older latinos: study of latinos-investigation of neurocognitive aging results (HCHS/SOL). *Sci Rep.* (2018) 8:17578. doi: 10.1038/s41598-018-35573-3

60. Harwood DG, Kalechstein A, Barker WW, et al. The effect of alcohol and tobacco consumption, and apolipoprotein E genotype, on the age of onset in Alzheimer's disease. *Int J Geriatr Psychiatry*. (2010) 25:511–8. doi: 10.1002/gps.2372
61. He J, Farias S, Martinez O, Reed B, Mungas D, Decarli C. Differences in brain volume, hippocampal volume, cerebrovascular risk factors, and apolipoprotein E4 among mild cognitive impairment subtypes. *Arch Neurol*. (2009) 66:1393–9. doi: 10.1001/archneurol.2009.252
62. Hendrie HC, Murrell J, Baiyewu O, Lane KA, Purnell C, Ogunniyi A, et al. APOE ε4 and the risk for Alzheimer disease and cognitive decline in African Americans and Yoruba. *Int Psychogeriatr*. (2014) 26:977–85. doi: 10.1017/S1041610214000167
63. Hohman TJ, Cooke–Bailey JN, Reitz C, Jun G, Naj A, Beecham GW, et al. Global and local ancestry in African-Americans: Implications for Alzheimer's disease risk. *Alzheimers Dement*. (2016) 12:233–43. doi: 10.1016/j.jalz.2015.02.012
64. Janicki SC, Park N, Cheng R, Schupf N, Clark LN, Lee JH. Aromatase variants modify risk for Alzheimer's disease in a multiethnic female cohort. *Dement Geriatr Cogn Disord*. (2013) 35:340–6. doi: 10.1159/000343074
65. Janicki SC, Park N, Cheng R, Clark LN, Lee JH, Schupf N. Estrogen receptor α variants affect age at onset of Alzheimer's disease in a multiethnic female cohort. *Dement Geriatr Cogn Disord*. (2014) 38:200–13. doi: 10.1159/000355559
66. Jin SC, Carrasquillo MM, Benitez BA, Tara S, Carrell D, Patel D, et al. TREM2 is associated with increased risk for Alzheimer's disease in African Americans. *Mol Neurodegener*. (2015) 10:19. doi: 10.1186/s13024-015-0016-9
67. Kim S, Nho K, Ramanan VK, Lai D, Foroud TM, Lane K, et al. Genetic Influences on Plasma Homocysteine Levels in African Americans and Yoruba Nigerians. *J Alzheimers Dis*. (2016) 49:991–1003. doi: 10.3233/JAD-150651
68. Kuller LH, Lopez OL, Becker JT, Chang Y, Newman AB. Risk of dementia and death in the long-term follow-up of the Pittsburgh Cardiovascular Health Study-Cognition Study. *Alzheimers Dement*. (2016) 12:170–83. doi: 10.1016/j.jalz.2015.08.165
69. Kunkle BW, Carney RM, Kohli MA, Naj AC, Nelson KLH, Whitehead PL, et al. Targeted sequencing of ABCA7 identifies splicing, stop-gain and intronic risk variants for Alzheimer disease. *Neurosci Lett*. (2017) 649:124–9. doi: 10.1016/j.neulet.2017.04.014
70. Kwon OD, Khaleeq A, Chan W, Pavlik VN, Doody RS. Apolipoprotein E polymorphism and age at onset of Alzheimer's disease in a quadriethnic sample. *Dement Geriatr Cogn Disord*. (2010) 30:486–91. doi: 10.1159/000322368
71. Lee JH, Cheng R, Schupf N, Manly J, Lantigua R, Stern Y, et al. The association between genetic variants in SORL1 and Alzheimer disease in an urban, multiethnic, community-based cohort. *Arch Neurol*. (2007) 64:501–6. doi: 10.1001/archneur.64.4.501
72. Lee JH, Cheng R, Rogaeva E, Meng Y, Stern Y, Santana V, et al. Further examination of the candidate genes in chromosome 12p13 locus for late-onset Alzheimer disease. *Neurogenetics*. (2008) 9:2. doi: 10.1007/s10048-008-0122-8
73. Lee JH, Barral S, Cheng R, Chacon I, Santana V, Williamson J, et al. Age-at-onset linkage analysis in Caribbean Hispanics with familial late-onset Alzheimer's disease. *Neurogenetics*. (2008) 9:127–38. doi: 10.1007/s10048-007-0103-3
74. Lee JH, Cheng R, Barral S, Reitz C, Medrano M, Lantigua R, et al. Identification of novel loci for Alzheimer disease and replication of CLU, PICALM, and BIN1 in Caribbean Hispanic individuals. *Arch Neurol*. (2011) 68:320–8. doi: 10.1001/archneurol.2010.292
75. Lee JH, Cheng R, Vardarajan B, Lantigua R, Dumeyeer DR, Ortmann W, et al. Genetic Modifiers of Age at Onset in Carriers of the G206A Mutation in PSEN1 With Familial Alzheimer Disease Among Caribbean Hispanics. *JAMA Neurol*. (2015) 72:1043–51. doi: 10.1001/jamaneurol.2015.1424
76. Livney MG, Clark CM, Karlawish JH, Cartmell S, Negron M, Nunez J, et al. Ethnoracial differences in the clinical characteristics of Alzheimer's disease at initial presentation at an urban Alzheimer's disease center. *Am J Geriatr Psychiatry*. (2011) 19:430–9. doi: 10.1097/JGP.0b013e3181f7d881
77. Logue MW, Schu M, Vardarajan BN, Buros J, Green RC, Go RCP, et al. A comprehensive genetic association study of Alzheimer disease in African Americans. *Arch Neurol*. (2011) 68:1569–79. doi: 10.1001/archneurol.2011.646
78. Logue MW, Schu M, Vardarajan BN, Farell J, Bennett DA, Buxbaum JD, et al. Two rare AKAP9 variants are associated with Alzheimer's disease in African Americans. *Alzheimers Dement*. (2014) 10:609–18. doi: 10.1016/j.jalz.2014.06.010
79. Logue MW, Lancour D, Farrell J, Simhina I, Fallin MD, Lunetta KL, et al. Targeted Sequencing of Alzheimer Disease Genes in African Americans Implicates Novel Risk Variants. *Front Neurosci*. (2018) 12:592. doi: 10.3389/fnins.2018.00592
80. Marden JR, Walter S, Tchetgen Tchetgen EJ, Kawachi I, Glymour MM. Validation of a polygenic risk score for dementia in black and white individuals. *Brain Behav*. (2014) 4:687–97. doi: 10.1002/brb3.248
81. Marden JR, Mayeda ER, Walter S, Vivot A, Tchetgan EJT, Kawachi I, et al. Using an Alzheimer Disease Polygenic Risk Score to Predict Memory Decline in Black and White Americans Over 14 Years of Follow-up. *Alzheimer Dis Assoc Disord*. (2016) 30:195–202. doi: 10.1097/WAD.0000000000000137
82. McAninch E, Rajan K, Evans D, Jo S, Chaker L, Peeters R, et al. A common DIO2 polymorphism and Alzheimer disease dementia in African and European Americans. *J Clin Endocrinol Metab*. (2018) 103:505–30. doi: 10.1210/jc.2017-01196
83. Melville S, Buros J, Parrado A, Vardarajan B, Logue M, Shen L, et al. Multiple loci influencinghippocampal degeneration identified by genome scan. *Ann Neurol*. (2012) 72:108–20. doi: 10.1002/ana.23644
84. Mez J, Chung J, Jun G, Kriegel J, Bourias A, Sherva R, Logue M, et al. Two novel loci, COBL andSLC10A2, for Alzheimer's disease in African Americans. Alzheimer's Dement. (2017) 13:45-81. doi: 10.1016/j.jalz.2016.09.002
85. Mount D, Ashley A, Lah J, Levey A, Goldstein F. Is ApoE ε4 Associated with Cognitive Functioning in African Americans Diagnosed with Alzheimer Disease? An Exploratory Study. *South Med J*. (2009) 102:945–8. doi: 10.1097/SMJ.0b013e3181b21b82
86. Murrell JR, Price B, Lane KA, Baiyewu O, Gureje O, Ogunnieyi A, et al. Association of apolipoprotein E genotype and Alzheimer disease in African Americans. *Arch Neurol*. (2006) 63:431–4. doi: 10.1001/archneur.63.3.431
87. N'Songo A, Carrasquillo MM, Wang X, Burgess JD, Nguyen T, Asmann YW, et al. African American exome sequencing identifies potential risk variants at Alzheimer disease loci. *Neurol Genet*. (2017) 3:e141. doi: 10.1212/NXG.0000000000000141
88. O'Bryant SE, Johnson L, Balldin V, Edwards M, Barbar R, Williams B, et al. Characterization of Mexican Americans with mild cognitive impairment and Alzheimer's disease. *J Alzheimers Dis*. (2013) 33:373–9. doi: 10.3233/JAD-2012-121420
89. O'Bryant SE, Johnson L, Reisch J, Edwards M, Hall J, Barbar R, et al. Risk factors for mild cognitive impairment among Mexican Americans. *Alzheimers Dement*. (2013) 9:622–31. doi: 10.1016/j.jalz.2012.12.007
90. Olarte L, Schupf N, Lee JH, Tang MX, Santana V, Williamson J, et al. Apolipoprotein E epsilon4 and age at onset of sporadic and familial Alzheimer disease in Caribbean Hispanics. *Arch Neurol*. (2006) 63:1586–90. doi: 10.1001/archneur.63.11.1586
91. Pedraza O, Allen M, Jennette K, Carrasquilo M, Crook J, Serie D, et al. Evaluation of memory endophenotypes for association with CLU, CR1, and PICALM variants in black and white subjects. *Alzheimers Dement*. (2014) 10:205–13. doi: 10.1016/j.jalz.2013.01.016
92. Peila R, Yucesoy B, White LR, Johnson V, Kashon ML, Wu K, et al. A TGF-beta1 polymorphism association with dementia and neuropathologies: the HAAS. *Neurobiol Aging*. (2007) 28:1367–73. doi: 10.1016/j.neurobiolaging.2006.06.004
93. Petrovitch H, Ross GW, He Q, Lock JU, Markesbery W, Davis D, et al. Characterization of Japanese-American men with a single neocortical AD lesion type. *Neurobiol Aging*. (2008) 29:1448–55. doi: 10.1016/j.neurobiolaging.2007.03.026
94. Qian J, Wolters FJ, Beiser A, Haan M, Ikram MA, Karlawish J, et al. APOE-related risk of mild cognitive impairment and dementia for prevention trials: An analysis of four cohorts. *PLoS Med*. (2017) 14:e1002254. doi: 10.1371/journal.pmed.1002254

95. Rajabli F, Feliciano BE, Celis K, Nelson KLH, Whitehead PL, Adams LD, et al. Ancestral origin of ApoE $\varepsilon 4$ Alzheimer disease risk in Puerto Rican and African American populations. *PLoS Genet.* (2018) 14:e1007791. doi: 10.1371/journal.pgen.1007791
96. Reitz C, Tokuhiro S, Clark LN, Conrad C, Vonsattel JP, Hazrati LN, et al. SORCS1 alters amyloid precursor protein processing and variants may increase Alzheimer's disease risk. *Ann Neurol.* (2011) 69:47–64. doi: 10.1002/ana.22308
97. Reitz C, Cheng R, Schupf N, Lee JH, Mehta PD, Rogaeva E, et al. Association between variants in IDE-KIF11-HHEX and plasma amyloid β levels. *Neurobiol Aging.* (2012) 199:e13–7. doi: 10.1016/j.neurobiolaging.2010.07.005
98. Moher D, Liberati A, Tetzlaff J, Altman DG, for the PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ.* (2009) 339:b2535. doi: 10.1136/bmj.b2535
99. Rippon GA, Tang MX, Lee JH, Lantigua R, Medrano M, Mayeux R. Familial Alzheimer disease in Latinos: interaction between APOE, stroke, and estrogen replacement. *Neurology.* (2006) 66:35–40. doi: 10.1212/01.wnl.0000191300.38571.3e
100. Roses AD, Lutz MW, Saunders AM, Goldgaber D, Saul R, et al. African-American TOMM40'523-APOE haplotypes are admixture of West African and Caucasian alleles. *Alzheimers Dement.* (2014) 10:592–601. doi: 10.1016/j.jalz.2014.06.009
101. Saczynski JS, White L, Peila RL, Rodriguez BL, Launer LJ. The relation between apolipoprotein A-I and dementia: the Honolulu-Asia aging study. *Am J Epidemiol.* (2007) 165:91–9. doi: 10.1093/aje/kwm027
102. Sawyer K, Sachs-Ericsson N, Preacher KJ, Blazer DG. Racial differences in the influence of the APOE epsilon 4 allele on cognitive decline in a sample of community-dwelling older adults. *Gerontology.* (2009) 55:105–20. doi: 10.1159/000137666
103. Simino J, Wang Z, Bressler J, Chouraki V, Yang Q, Younkin SG,et al. Whole exome sequence-based association analyses of plasma amyloid-β in African and European Americans; the Atherosclerosis Risk in Communities-Neurocognitive Study. *PLoS One.* (2017) 12:e0180046. doi: 10.1371/journal.pone.0180046
104. Tosto G, Bird TD, Tsuang D, Bennett DA, Boeve BF, Crushaga C, et al. Polygenic risk scores in familial Alzheimer disease. *Neurology.* (2017) 88:1180–86. doi: 10.1212/WNL.0000000000003734
105. Vardarajan BN, Bruesegem SY, Harbour ME, et al. Identification of Alzheimer disease-associated variants in genes that regulate retromer function. *Neurobiol Aging.* (2012) 33:e15–e30. doi: 10.1016/j.neurobiolaging.2012.04.020
106. Vardarajan BN, Zhang Y, Lee JH, Cheng R, Bohm C, Ghani M, et al. Coding mutations in SORL1 and Alzheimer disease. *Ann Neurol.* (2015) 77:215–27. doi: 10.1002/ana.24305
107. Weiner MF, Hynan LS, Rossetti H, Womack KB, Rosenberg RN, Gong YH, et al. The relationship of cardiovascular risk factors to Alzheimer disease in Choctaw Indians. *Am J Geriatr Psychiatry.* (2011) 19:423–9. doi: 10.1097/JGP.0b013e3181e89a46
108. Yu L, Lutz MW, Wilson RS, Burns DK, Roses AD, Saunders AM, et al. APOE $\varepsilon 4$-TOMM40 '523 haplotypes and the risk of Alzheimer's disease in older Caucasian and African Americans. *PLoS One.* (2017) 12:e7–e9. doi: 10.1371/journal.pone.0180356
109. Reitz C, Jun G, Naj A, Rajbhandary R, Vardarajan BN, Wang L-S, et al. Variants in the ATP-binding cassette transporter (ABCA7), apolipoprotein E $\epsilon 4$, and the risk of late-onset Alzheimer disease in African Americans. *JAMA.* (2013) 309:1483–92. doi: 10.1001/jama.2013.2973
110. Barnes LL, Leurgans S, Aggarwal NT, Shah RC, Arvanitakis Z, James BD, et al. Mixed pathology is more likely in black than white decedents with Alzheimer dementia. *Neurology.* (2015) 85:528–34. doi: 10.1212/WNL.0000000000001834
111. Kunkle BW, Schmidt M, Klein H-U, Naj AC, Hamilton-Nelson KL, Larson EB, et al. Novel Alzheimer disease risk loci and pathways in African American individuals using the african genome resources panel: a meta-analysis. *JAMA Neurol.* (2021) 78:102–13. doi: 10.1001/jamaneurol.2020.3536
112. Nussbaum RL. Genome-wide association studies, Alzheimer disease, and understudied populations. *JAMA.* (2013) 309:1527–8. doi: 10.1001/jama.2013.3507
113. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature.* (2019) 570:514–8. doi: 10.1038/s41586-019-1310-4
114. McGinnis JM, Williams-Russo P, Knickman JR. The case for more active policy attention to health promotion. *Health Aff (Millwood).* (2002) 21:78–93. doi: 10.1377/hlthaff.21.2.78
115. Adler NE, Newman K. Socioeconomic disparities in health: pathways and policies. *Health Aff (Millwood).* (2002) 21:60–76. doi: 10.1377/hlthaff.21.2.60
116. Mahoney DF, Cloutterbuck J, Neary S, Zhan L. African American, Chinese, and Latino family caregivers' impressions of the onset and diagnosis of dementia: cross-cultural similarities and differences. *Gerontologist.* (2005) 45:783–92. doi: 10.1093/geront/45.6.783
117. Jang Y, Kim G, Chiriboga D. Knowledge of Alzheimer's disease, feelings of shame, and awareness of services among Korean American elders. *J Aging Health.* (2010) 22:419–33. doi: 10.1177/0898264309360672
118. Liu D, Hinton L, Tran C, Hinton D, Barker JC. Reexamining the relationships among dementia, stigma, and aging in immigrant Chinese and Vietnamese family caregivers. *J Cross Cult Gerontol.* (2008) 23:283–99. doi: 10.1007/s10823-008-9075-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Public Health

METHODS
published: 07 January 2022
doi: 10.3389/fpubh.2021.757942

G-Computation to Causal Mediation Analysis With Sequential Multiple Mediators—Investigating the Vulnerable Time Window of HBV Activity for the Mechanism of HCV Induced Hepatocellular Carcinoma

An-Shun Tai[1], *Yen-Tsung Huang*[2], *Hwai-I Yang*[3], *Lauren V. Lan*[1,4] *and Sheng-Hsuan Lin*[1]*

[1] Institute of Statistics, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, [2] Institute of Statistical Science, Academia Sinica, Taipei, Taiwan, [3] Genomics Research Center, Academia Sinica, Taipei, Taiwan, [4] Department of Biostatistics, Johns Hopkins University, Baltimore, MD, United States

OPEN ACCESS

Edited by:
MinJae Lee,
University of Texas Southwestern
Medical Center, United States

Reviewed by:
Matthew Valente,
Florida International University,
United States
Roch A. Nianogo,
UCLA School of Public Health,
United States

***Correspondence:**
Sheng-Hsuan Lin
shenglin@stat.nctu.edu.tw

Specialty section:
This article was submitted to
Life-Course Epidemiology and Social
Inequalities in Health,
a section of the journal
Frontiers in Public Health

Received: 13 August 2021
Accepted: 30 November 2021
Published: 07 January 2022

Citation:
Tai A-S, Huang Y-T, Yang H-I, Lan LV and Lin S-H (2022) G-Computation to Causal Mediation Analysis With Sequential Multiple Mediators—Investigating the Vulnerable Time Window of HBV Activity for the Mechanism of HCV Induced Hepatocellular Carcinoma. Front. Public Health 9:757942. doi: 10.3389/fpubh.2021.757942

Regression-based approaches are widely used in causal mediation analysis. The presence of multiple mediators, however, increases the complexity and difficulty of mediation analysis. In such cases, regression-based approaches cannot efficiently address estimation issues. Hence, a flexible approach to mediation analysis is needed. Therefore, we developed a method for using g-computation algorithm to conduct causal mediation analysis in the presence of multiple ordered mediators. Compared to regression-based approaches, the proposed simulation-based approach increases flexibility in the choice of models and increases the range of the outcome scale. The Taiwanese Cohort Study dataset was used to evaluate the efficacy of the proposed approach for investigating the mediating role of early and late HBV viral load in the effect of HCV infection on hepatocellular carcinoma (HCC) in HBV seropositive patients ($n =$ 2,878; HCV carrier $n = 123$). Our results indicated that early HBV viral load had a negative mediating role in HCV-induced HCC. Additionally, early exposure to a low HBV viral load affected HCC through a lag effect on HCC incidence [OR = 0.873, 95% CI = (0.853, 0.893)], and the effect of early exposure to a low HBV viral load on HCC incidence was slightly larger than that of a persistently low viral load on HCC incidence [OR = 0.918, 95% CI = (0.896, 0.941)].

Keywords: causal inference, mechanism investigation, mediation analysis, path-specific effect, multiple mediators

INTRODUCTION

Epidemiology studies and other health-related studies often investigate the overall effect of a certain risk factor or exposure on health-related outcomes. Confirmation of such effects facilitates further elucidation of possible biological mechanisms. Path analysis and mediation analysis are often used to investigate causal mechanisms because they can decompose these effects into several pathways according to the involvement of various mediators of interest (1). Mediation analysis aims to assess how exposure affects the outcome of interest through mediators and sheds deep insight into the underlying mechanism of the relationship between the exposure and outcome. Causal mediation analysis, a branch of mediation analysis, explicitly defines the causal effects of interest based on a counterfactual (potential) outcome model (2–4). The counterfactual model denotes the hypothetical outcome (here, it indicates the "counterfactual level" of a certain variable of interest) an individual would have, under a hypothetical condition when the same individual had received a particular intervention on previous variables. It is called "counterfactual" because this individual might not have received this intervention in real world. Since causal mediation analysis accounts for non-linearity of outcomes and interactions between exposure and mediator, it expands the use of mediation analysis to more general conditions (2, 5–7). Additionally, in scenarios involving a single fixed exposure and a single mediator, several techniques have been proposed to account for various outcome scales, including dichotomous variables (8), time-to-event variables (9–12), and many others (13, 14).

In multiple mediator settings, i.e., settings involving more than one mediator, however, mediation analysis is often challenging. One example is the extreme complexity of decomposing the effects of hepatitis C virus (HCV) infection on hepatocellular carcinoma (HCC) in the presence of hepatitis B virus (HBV) activity, which was the motivation for this study (15, 16). **Figure 1** shows that the mediation analysis assumed causal relationships among HCV infection status, HBV viral load at baseline, HBV viral load at follow up, and HCC status. Baseline HBV viral load activity was used to represent the current status of HBV activity; baseline HCV infection status was used to represent relatively long-term HCV infection status. That is, HCV infection status was assumed to precede HBV viral load, which was considered a reasonable assumption. The role of HBV viral activity in this mechanism in HBV sero-positive patients at baseline and during follow up was investigated by using mediation analysis to decompose the effects into four paths (**Figure 2**). Effects in each of the four paths (i.e., the path-specific effects, PSEs) can be categorized as (1) paths only through change in early HBV viral load (PSE_1); (2) paths only through change in late HBV viral load (PSE_2); (3) paths through change in early HBV viral load that further impacts late HBV viral load (PSE_{12}); and (4) paths not through change in early or late HBV viral load (PSE_0). Decomposition of the overall effect into four PSEs facilitates understanding of the role of HBV viral activity and when the role of HBV viral activity is critical. These data can then be used to reduce the HCC incidence in patients with dual virus infection.

Before conducting mediation analysis in this case, the two settings must be differentiated according to the relationships between mediators. In the first setting, mediators are independent of each other conditioned on all previous covariates, including baseline confounders and the exposure. In this setting, which is also referred to as "parallel" or "non-ordered" multiple mediators, the motivating example is rational only if early HBV viral load does not affect late HBV viral load. The standard causal mediation analysis framework for a single mediator is easily extended to this setting by performing a sequential mediation analysis of each mediator. Notably, methods have also been developed for simultaneous analysis of parallel mediators (17, 18). Apparently, however, the above parallel setting does not fit our motivating example since early HBV viral activity would surely affect viral activity at follow up. In the case of early HBV viral activity, the alternative setting, "ordered" or "sequential" multiple mediators, is reasonable. Unfortunately, effect decomposition in this setting is infeasible since some PSEs cannot be identified by empirical data without additional strong assumptions (15, 19–21). For example, to identify full PSEs in the presence of two ordered mediators, the assumption of independence between two counterfactuals of the mediator is proposed for identification (21). This independence assumption is extremely strong and unrealistic. Without further assumptions, only partial effect decomposition, which evaluates the cumulative PSEs, can be achieved.

Specifically, only PSE_2, PSE_0, and the sum of PSE_1 and PSE_{12} are identifiable. However, PSE_1 and PSE_{12} cannot be further distinguished, even without time-varying and unmeasured

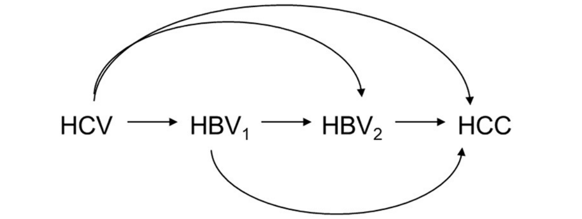


FIGURE 1 | Causal relationship among HCV infection status (HCV), HBV viral load at baseline (HBV_1), HBV viral load at follow-up (HBV_2), and HCC status (HCC).

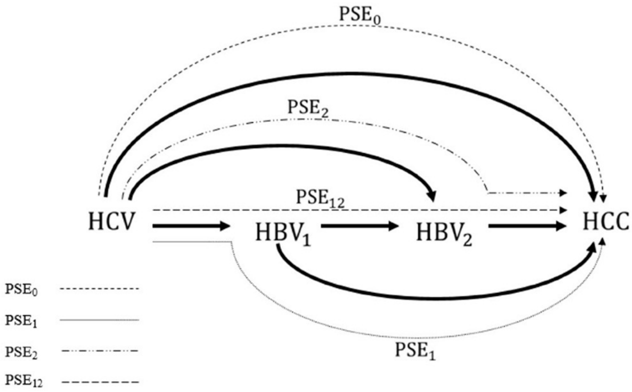


FIGURE 2 | Four path-specific effects (PSEs), as well as four interventional PSEs, to be decomposed from the overall effect of HCV infection on the incidence of HCC. PSE_1: the path through the HBV_1 only; PSE_2: the path through the HBV_2 only, PSE_{12}: the path through HBV_1 which further impacts HBV_2; and PSE_0: the path not through HBV_1 or HBV_2. PSE, path-specific effect; HCV, hepatitis C virus; HBV, hepatitis B virus; HCC, hepatocellular carcinoma.

baseline confounders. Two strategies for resolving this problem are possible. First, the overall effect can be decomposed into the three components described above (16, 22, 23). We can further pool all ordered mediators as a single mediator, and decompose the total effect into effect either through or not through this pooled mediator (18, 20). The second approach is to measure the upper and lower bounds of PSE through sensitivity analysis under causal framework (24, 25). However, point estimate of PSE still cannot be obtained through this method (21). Previously, Lin and VanderWeele proposed an interventional approach to estimate analogs of PSEs under no-unmeasured-confounding assumptions with a regression-based approach (26). The concepts of the interventional approach and PSE were also adopted by VanderWeele, Vansteelandt, and Robins (20) for mediation analysis with a single mediator in the presence of an exposure-induced mediator-outcome confounder. Note that their work only derives the direct effect, the sum of two PSEs passed through the mediator, and the indirect effect, the sum of two PSEs without passing through the mediator. Meanwhile, Vansteelandt and Daniel also proposed a new interventional approach, which has no assumption of structure among mediators, for deriving PSEs (27), but different from Lin and VanderWeele's method, they still cannot distinguish PSE_{12} from the other PSEs. A limitation of Lin and VanderWeele's method is that the link function of outcome model has to be linear or log-linear, and that it cannot be adapted for a non-linear or generalized linear models. Moreover, unlike the analysis of overall effect, the analytical solutions for all PSEs estimates vary substantially in different models even when the linear function of outcome model is linear or log-linear. Therefore, the software of the regression-based approach can only be applied to few model choices.

To remedy this research gap, we adopted the simulation-based approach based on g-computation algorithm to provide a flexible computational algorithm for the estimation of causal mediation analysis. g-computation algorithm was first introduced by Robins in 1986 to estimate the causal effect of a time-varying exposure in the presence of time-varying confounders that are affected by exposure (3). Recently, the simulation-based approach has been widely used for standard causal mediation analysis (27–34). These methods usually involve using maximum likelihood estimation (MLE) to fit a set of parametric models and then using g-computation algorithm and bootstrapping methods to generate point and interval estimates, respectively. This simulation-based approach provides the flexibility to choose models and variables without considering an analytic form. This approach also obtains more stable and efficient estimates compared to weighted approach (14, 31, 35). Therefore, simulation-based approach is useful for investigating mechanisms when the outcome variable does not fit the requirements of a linear regression model. Therefore, this study used this approach to develop a method of performing mediation analysis in scenarios involving two ordered multiple mediators. The proposed method was then used investigate the mechanisms through which HCV induces HCC through HBV activity.

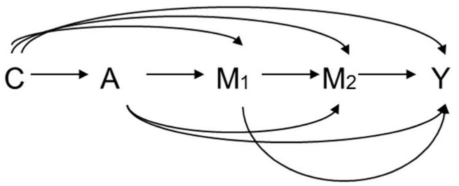


FIGURE 3 | Relation among exposure A, two ordering mediators M_1 and M_2, outcome Y, and covariates C. A: exposure, M_1: the first mediator, M_2: the second mediator, Y: outcome, C: covariates.

MATERIALS AND METHODS

Data Description of the REVEAL-HBV Study

This study was motivated by the Risk Evaluation of Viral Load Elevation and Associated Liver Disease/Cancer–Hepatitis B Virus (REVEAL-HBV) study (36). The details of the REVEAL-HBV study design and participant enrollment have been illustrated in literatures (36–39). 23,820 Taiwanese residents aged 30–65 years were recruited from 1991 to 1992. Among the participants, 2,878 were HBV-positive, of which 188 developed HCC during the follow-up period. Written informed consent for interview questionnaires, health examinations, biospecimen collection, and data linkage of health status with death certification profiles and National Cancer Registry were obtained. Blood samples collected at enrollment were examined for seromarkers and viral load of HBV and HCV. Newly diagnosed HCC was recorded using computerized data linkage with National Cancer Registry and death certification systems.

Notation, Definition, and Effect Decomposition for Dichotomous Outcome

Let A denote the exposure, Y a dichotomous outcome, M_1 the first mediator, M_2 the second mediator, and C a set of baseline covariates. For example, A is HCV infection status, Y is an HCC event before the end of follow up, M_1 is early HBV viral load, and M_2 is late HBV viral load. Let A =1 and A = 0 denote two hypothetical levels of exposure: HCV infection and non-infection, respectively. **Figure 3** graphically illustrates the causal relationships among A, Y, M_1, M_2, and C based on substantive prior knowledge. **Figure 4** is the case of more than two mediators as well as time-varying mediator-outcome confounders, which are affected by exposure. For simplicity, however, we assume the absence of time-varying confounders, and we assume the presence of only two ordered mediators of interest.

Counterfactual outcome models are used to define four PSEs corresponding the four paths in **Figure 2** based on causal theory (2–6, 19, 40). For the individual i, $Y_i(a)$ denotes the counterfactual level of Y_i if this individual had received an intervention on exposure A as level a. Similarly, $M_{2i}(a, m_1)$ denotes the counterfactual level of M_{2i} if this individual had received an intervention on exposure A as level a and on the first mediator M_{1i} as level m_1. Here, the notation can be simplified by removing the subscript i.

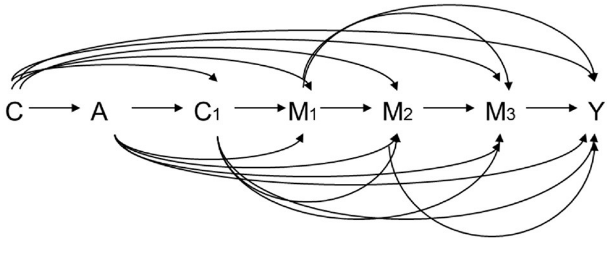


FIGURE 4 | Relation among exposure A, three ordering mediators M_1, M_2, and M_3, outcome Y, baseline covariates C, and time-varying mediator-outcome confounder C_1. A: exposure, M_1: the first mediator, M_2: the second mediator, Y: outcome, C: baseline covariates, C_1: time-varying mediator-outcome confounder.

For a dichotomous outcome, the total effect may be expressed on risk difference (RD), risk ratio (RR), or odds ratio (OR) scale. Although software used to perform simulation-based approaches provides the results of all scales, the OR scale is used throughout this discussion since OR is the most frequently used scale for dichotomous outcomes. The total effect in OR scale, $\mathrm{OR}_{\mathrm{TE}}(1,0)$, is defined as Odds(Y(1))/Odds(Y(0)), where Odds(B) is defined as $\Pr(B = 1)/\Pr(B = 0)$ for any dichotomous variable B [e.g. Y, Y(1), or Y(0)]. The definitions of RD and RR scales are detailed in **Appendix A**.

When investigating a mechanism with two mediators M_1 and M_2 of interest, the total effect ($\mathrm{OR}_{\mathrm{TE}}$) can be decomposed into four PSEs: path not through M_1 or M_2; path through M_1 only; path through M_2 only; and path through M_1 and then through M_2; these four PSEs are expressed in OR scale as $\mathrm{OR}_{\mathrm{PSE0}}$, $\mathrm{OR}_{\mathrm{PSE1}}$, $\mathrm{OR}_{\mathrm{PSE2}}$, and $\mathrm{OR}_{\mathrm{PSE12}}$, respectively, and are defined as follows:

$$\begin{aligned}
\mathrm{OR}_{\mathrm{PSE0}} &= \Phi(1,0,0,0)/\Phi(0,0,0,0)\\
\mathrm{OR}_{\mathrm{PSE1}} &= \Phi(1,1,0,0)/\Phi(1,0,0,0)\\
\mathrm{OR}_{\mathrm{PSE2}} &= \Phi(1,1,1,0)/\Phi(1,1,0,0)\\
\mathrm{OR}_{\mathrm{PSE12}} &= \Phi(1,1,1,1)/\Phi(1,1,1,0) \qquad (1)
\end{aligned}$$

where $\Phi(a_1,a_2,a_3,a_4)$ is defined as $\mathrm{Odds}(Y(a_1,M_1(a_2),M_2(a_3,M_1(a_4))))$. Here, $Y(a_1, M_1(a_2),M_2(a_3,M_1(a_4)))$ denotes the counterfactual value of outcome Y if the exposure is set to a_1, the first mediator is set to $M_1(a_2)$, and the second mediator is set to $M_2(a_3,M_1(a_4))$ (or the counterfactual value of M_2 if exposure is set to a_3 and first mediator is set to $M_1(a_4)$). The $\mathrm{OR}_{\mathrm{TE}}$ is the product of four PSEs in OR scale, which can be expressed as

$$\mathrm{OR}_{\mathrm{TE}} = OR_{\mathrm{PSE0}}\times\mathrm{OR}_{\mathrm{PSE1}}\times\mathrm{OR}_{\mathrm{PSE2}}\times\mathrm{OR}_{\mathrm{PSE12}} \qquad (2)$$

While Equation (1) gives a definition of four PSEs decomposed from TE, the decomposition of TE is not unique. For example, $\mathrm{OR}_{\mathrm{PSE0}} = \Phi(1,1,1,1)/\Phi(0,1,1,1)$, $\mathrm{OR}_{\mathrm{PSE1}} = \Phi(0,1,1,1)/\Phi(0,0,1,1)$, $\mathrm{OR}_{\mathrm{PSE2}} = \Phi(0,0,1,1)/\Phi(0,0,0,1)$, and $\mathrm{OR}_{\mathrm{PSE12}} = \Phi(0,0,0,1)/\Phi(0,0,0,0)$ are alternative decomposition of TE. For two sequential mediators, 24 possible decompositions have been provided in the previous study (21). This work primarily focuses on the decomposition type defined in Equation (1). The following identification and estimation are valid no matter which decomposition is used.

Interventional Approach to Identification

The $\Phi(a_1,a_2,a_3,a_4)$ can be non-parametrically identified only when a_2 is equal to a_4. Consequently, only $\mathrm{OR}_{\mathrm{PSE0}}$, $\mathrm{OR}_{\mathrm{PSE2}}$ and the sum of $\mathrm{OR}_{\mathrm{PSE1}}$ and $\mathrm{OR}_{\mathrm{PSE12}}$ are identified by empirical data. Here, we introduce an interventional approach: instead of defining the four paths as four traditional PSEs, the four paths are defined as four interventional path-specific effects (iPSEs). In an earlier work, these paths were referred to as randomly interventional analogs of PSEs (26). The advantage of the interventional approach is that all iPSEs can be non-parametrically identified under the assumption of no unmeasured confounding factors. (26). In OR scale, the paths are denoted $\mathrm{OR}_{\mathrm{iPSE0}}$, $\mathrm{OR}_{\mathrm{iPSE1}}$, $\mathrm{OR}_{\mathrm{iPSE2}}$, and $\mathrm{OR}_{\mathrm{iPSE12}}$ and are defined as follows:

$$\begin{aligned}
\mathrm{OR}_{\mathrm{iPSE0}} &= \Psi(1,0,0,0)/\Psi(0,0,0,0)\\
\mathrm{OR}_{\mathrm{iPSE1}} &= \Psi(1,1,0,0)/\Psi(1,0,0,0)\\
\mathrm{OR}_{\mathrm{iPSE2}} &= \Psi(1,1,1,0)/\Psi(1,1,0,0)\\
\mathrm{OR}_{\mathrm{iPSE12}} &= \Psi(1,1,1,1)/\Psi(1,1,1,0) \qquad (3)
\end{aligned}$$

where $\Psi(a_1,a_2,a_3,a_4)$ is defined as $\mathrm{Odds}(Y(a_1,G_1(a_2),G_2(a_3,G_1(a_4))))$. Here, we set the exposure as a_1, the first mediator as $G_1(a_2)$, and the second mediator as $G_2(a_3,G_1(a_4))$. For any value of a and m, $G_1(a)$ is the random draw of $M_1(a)$, and $G_2(a,m_1)$ is the random draw of $M_2(a,m_1)$. In this setting, $Y(a_1,G_1(a_2),G_2(a_3,G_1(a_4)))$ denotes the counterfactual value of outcome Y . Consequently, $G_2(a_3,G_1(a_4))$ is the random draw of $M_2(a_3,G_1(a_4))$ while $G_1(a_4)$ is the random draw of $M_1(a_4)$. As in the conventional definition, the interventional definition for each path replaces the counterfactual level of each mediator with its random draw. We further define the product of four $\mathrm{OR}_{\mathrm{iPSE}}$ as the interventional total effect (iTE), which can be expressed in OR scale as the following equation:

$$\mathrm{OR}_{\mathrm{iTE}} = \mathrm{OR}_{\mathrm{iPSE0}}\times\mathrm{OR}_{\mathrm{iPSE1}}\times\mathrm{OR}_{\mathrm{iPSE2}}\times\mathrm{OR}_{\mathrm{iPSE12}} \qquad (4)$$

The $\mathrm{OR}_{\mathrm{iTE}}$ are very similar to the standard $\mathrm{OR}_{\mathrm{TE}}$ but not identical (14, 35). Therefore, as in the effect decomposition of $\mathrm{OR}_{\mathrm{TE}}$, the interventional decomposition can be viewed as its analog. The interpretations obtained when using iTE and iPSE, which are defined according to the stochastic interventions, differ from those of TE and PSE. These interpretations might be the best interpretations for a mechanism investigation as only the upper and lower bounds on PSE can be identified by empirical data even without time-varying confounders. Since iPSEs are PSEs analogs, iPSEs can still capture pathways. For example, $\mathrm{OR}_{\mathrm{iPSE12}}$ is non-zero only under the following conditions: (1) the change in the exposure affects the distribution of the first mediator; (2) the change in the first mediator affects the distribution of the second mediator; and (3) the change in the second mediator affects the distribution of the outcome. In extremely pathological settings, iPSEs may fail to represent the effects obtained by traditional PSEs. One example is the case of no overlap among individuals in whom the exposure affects the first mediator, individuals in whom the first mediator affects the second mediator, and individuals in whom the second mediator affects the outcome. In this case, $\mathrm{OR}_{\mathrm{iPSE12}}$ is non-zero while $\mathrm{OR}_{\mathrm{PSE12}}$ is under null. In contrast, in the case of complete overlap among all of these individuals (i.e., in the case of complete overlap among individuals in whom the exposure affects the first mediator, individuals in whom the first mediator affects the second mediator, and individuals in whom the second mediator affects the outcome) $\mathrm{OR}_{\mathrm{iPSE12}}$ is biased toward null. Further research on this topic in needed to elucidate the de*via*tion between PSE and its interventional version in different scenarios and to extend the applications of our method.

To identify $\Psi(\mathrm{a}_1,\mathrm{a}_2,\mathrm{a}_3,\mathrm{a}_4)$ and to identify $\mathrm{OR}_{\mathrm{iPSE}}$ and $\mathrm{OR}_{\mathrm{iTE}}$, four no-unmeasured-confounding assumptions are required:

Assumption (1) no-unmeasured-confounding between the relationships of exposure and outcome

Assumption (2) no-unmeasured-confounding between the relationships of mediators and outcome;

Assumption (3) no-unmeasured-confounding between the relationships of exposure and mediators;

Assumption (4) no-unmeasured-confounding between the relationships of two mediators.

Assumptions (1) to (4) are essentially used to avoid confounding bias in estimating iPSEs. It is worthy to note that a further cross-world assumption of no exposure-induced mediator-outcome confounder is commonly made in the conventional approaches of mediation analysis (9, 15, 21) but is unnecessary to the interventional approach. Using random draw permits that iPSEs are identifiable even when an exposure-induced mediator-outcome confounder presents. Here, we consider the case without an exposure-induced mediator-outcome confounder for identification. The identification result can be straightforwardly extended to the case where mediator-outcome confounders are affected by exposure directly. Under assumptions (1) to (4), $\mathrm{OR}_{\mathrm{iPSE}}$ and $\mathrm{OR}_{\mathrm{iTE}}$ are identified as follows:

$$\begin{aligned}
\mathrm{OR}_{\mathrm{iTE}} &= V(1,1,1,1)/V(0,0,0,0)\\
\mathrm{OR}_{\mathrm{iPSE0}} &= V(1,0,0,0)/V(0,0,0,0)\\
\mathrm{OR}_{\mathrm{iPSE1}} &= V(1,1,0,0)/V(1,0,0,0)\\
\mathrm{OR}_{\mathrm{iPSE2}} &= V(1,1,1,0)/V(1,1,0,0)\\
\mathrm{OR}_{\mathrm{iPSE12}} &= V(1,1,1,1)/V(1,1,1,0)
\end{aligned} \tag{5}$$

where $V(a_1,a_2,a_3,a_4)$ is defined as $\frac{Q(a_1,a_2,a_3,a_4)}{(1-Q(a_1,a_2,a_3,a_4)\ )}$

and

$$\begin{aligned}
&Q(a_1,a_2,a_3,a_4) =\\
&\sum\nolimits_c \sum\nolimits_{m_2,m_1} \Pr[Y=1|C=c,\ A=a_1, M_1=m_1, M_2=m_2]\\
&\Pr(M_1=m_1|C=c,\ A=a_2) \times \sum m_1^{'}\\
&\Pr\left(M_2=m_2\middle|C=c, A=\ a_3, M_1=m_1'\right)\\
&\Pr\left(M_1=m_1'\middle|C=c, A=\ a_4\right)\Pr(C=c\ )
\end{aligned} \tag{6.1}$$

If both M_1 and M_2 are continuous variables, (6.1) are replaced by integrals (6.2):

$$\begin{aligned}
&Q(a_1,a_2,a_3,a_4) =\\
&\int_c \int_{m_2,m_1} \{\Pr[Y=1|C=c, A=a_1, M_1=m_1, M_2=m_2]\\
&dF_{M_1|C,A}(M_1=m_1|C=c,\ A=a_2)\} \times\\
&\int_{m_1'} \{dF_{M_2|C,A,M_1}\left(M_2=m_2\middle|C=c,\ A=a_3, M_1=m_1'\right)\\
&dF_{M_1|C,A}\left(M_1=m_1'\middle|C=c,\ A=a_4\right)\} dF_C(c\ )
\end{aligned} \tag{6.2}$$

A previous work provide the proof for a generalized case in the presence of time-varying confounders (26). **Appendix A** defines iPSEs in RD and RR scales.

A logistic regression or other non-linear model can be used to estimate the conditional probability of outcome. Without assuming a rare disease (conditional probability of outcome $< 10\%$), $\mathrm{Q}(\mathrm{a}_1,\mathrm{a}_2,\mathrm{a}_3,\mathrm{a}_4)$ cannot be adequately approximated by a closed form. Consequently, a regression-based method is inapplicable, which was our motivation for developing the proposed simulation-based approach. In the simulation-based approach, the g-computation algorithm for iPSE is used for point estimation, and bootstrapping procedures are used for interval estimation. Since it does not consider the existence of the analytic form for all estimations, the simulation-based approach provides flexibility in the selection of statistical models.

Simulation-Based Approach for Estimation

In the proposed simulation-based approach, we use g-computation algorithm for iPSE point estimation and bootstrapping procedures for interval estimation. First, we build parametric models for the outcome and two mediators. For example, if two mediators are continuous variables and the outcome is a binary variable, three regression models are built:

$$\begin{aligned}
&logit(\Pr(Y=1|A=a,\ M_1=m_1,\ M_2=m_2,\ C=c))\\
&=\theta_0+\theta_1 a+\theta_2 m_1+\theta_3 m_2+\theta_c c
\end{aligned} \tag{7.1}$$

$$\begin{aligned}
&E(M_2|A=a,\ M_1=m_1,\ C=c)\\
&=\beta_0+\beta_1 a+\beta_2 m_1+\beta_c c
\end{aligned} \tag{7.2}$$

$$E(M_1|A=a,\ C=c)=\gamma_0+\gamma_1 a+\gamma_c c \tag{7.3}$$

The simulation-based approach allows for flexible selection of statistical models. Without considering the existence of the analytic form for all estimation, we can use any link function such as complementary log or probit function. Quadratic term or even log transformation or exposure and an interaction term between the exposure and the first mediator in model (7.1) can be included:

$$\begin{aligned} clog\left(-log\left(1-Pr\left(Y=1|A=a,\ M_1=m_1,\ M_2=m_2,\ C=c\right)\right)\right) \\ =\theta_0+\theta_1 a+\theta_{1s}a^2+\theta_{1l}\log\left(a\right) \\ +\theta_2 m_1+\theta_{12}am_1+\theta_3 m_2+\theta_c c \end{aligned} \tag{8}$$

After building parametric models for two mediators and outcome, we fit these models and obtain MLEs for all parameters. Based on all MLEs, we simulate the point estimations Q(1,1,1,1), Q(1,1,1,0), Q(1,1,0,0), Q(1,0,0,0), and Q(0,0,0,0) based on equation (6), as well as four OR$_{\mathrm{iPSE}}$ and OR$_{\mathrm{iTE}}$ based on the definition in (5). We generate confidence intervals by bootstrapping for the PSE inference as follows.

(step 1) Construct a regression model for conditional distribution $M_1,\ M_2,\ and\ Y$.

(1a) Construct a regression model for M_1 on A and all confounders.

(1b) Construct a regression model for M_2 on M_1, A and all confounders.

(1c) Construct a regression model for Y on M_2, M_1, A and all confounders.

For example, we can construct models using the following procedure as models (7.1)–(7.3):

$$M_1=\theta_{1,0}+\theta_{1,a}A+\tilde{\theta}_{1,c}\tilde{C}+\varepsilon_1$$

$$M_2=\theta_{2,0}+\theta_{2,a}A+\theta_{2,1}M_1+\tilde{\theta}_{2,c}\tilde{C}+\varepsilon_2$$

$$u_y=\left[1+\exp\left(-\left(\theta_{y,0}+\theta_{y,a}A+\theta_{y,1}M_1+\theta_{y,2}M_2+\tilde{\theta}_{y,c}\tilde{C}\right)\right)\right]^{-1}$$

$$\tilde{C}=\left(C_1,C_2,\ldots,C_{n_c}\right)^T$$

$$\tilde{\theta}_{1,c}=\left(\theta_{1,c_1},\ \theta_{1,c_2},\ldots,\ \theta_{1,c_{n_c}}\right)$$

$$\tilde{\theta}_{2,c}=\left(\theta_{2,c_1},\ \theta_{2,c_2},\ldots,\ \theta_{2,c_{n_c}}\right)$$

$$\tilde{\theta}_{y,c}=\left(\theta_{y,c_1},\ \theta_{y,c_2},\ldots,\theta_{y,c_{n_c}}\right)$$

$$Y\sim\ Bernoulli\left(\mu_y\right),\varepsilon_1\sim\ normal(0,\ \sigma_1^2),\varepsilon_2\sim\ normal(0,\ \sigma_2^2)$$

(step 2) Fit models with real data to obtain MLE for all parameters, i.e.

$$\hat{\theta}_{1,0},\hat{\theta}_{1,a},\hat{\tilde{\theta}}_{1,c},\ \hat{\theta}_{2,0},\hat{\theta}_{2,a},\hat{\theta}_{2,1},\hat{\tilde{\theta}}_{2,c},\ \hat{\theta}_{y,0},\hat{\theta}_{y,a},\hat{\theta}_{y,1},\hat{\theta}_{y,2},\hat{\tilde{\theta}}_{y,c},\ \ \hat{\sigma}_1^2,\ and\ \hat{\sigma}_2^2.$$

(step 3) Conduct g-computation algorithm using MLE and bootstrap.

(3a) Randomly sample the confounders $\tilde{C}$ with replacement and intervene the exposure A as 1. Use models built in Step 1 and MLEs in Step 2 to generate M_1 [denoted as $G_1\left(1\ \right)$].

(3b) Randomly sample the confounders $\tilde{C}$ with replacement, and intervene the exposure A as 0. Use models built in Step 1 and MLEs in Step 2 to generate M_1 [denoted as $G_1\left(0\ \right)$].

(3c) Randomly sample the confounders $\tilde{C}$, $G_1\left(1\right)$ with replacement, and intervene the exposure A as 1 and M_1 as $G_1\left(1\right)$. Use models built in Step 1 and the MLEs in Step 2 to generate M_2 [denoted as $G_2\left(1,\ G_1\left(1\right)\ \right)$].

(3d) Randomly sample the confounders $\tilde{C}$, $G_1\left(0\right)$ with replacement, and intervene the exposure A as 1 and M_1 as $G_1\left(0\right)$. Then use models from Step 1 and MLEs in Step 2 to generate M_2 [denoted as $G_2\left(1,\ G_1\left(0\right)\ \right)$].

(3e) Randomly sample the confounders $\tilde{C}$, $G_1\left(0\right)$ with replacement, and intervene the exposure A as 0 and M_1 as $G_1\left(0\right)$. Use models constructed in Step 1 and MLEs from Step 2 to generate M_2 [denoted as $G_2\left(0,\ G_1\left(0\right)\ \right)$].

(3f) Randomly sample the confounders $\tilde{C}$, $G_1\left(1\right)$, $G_2\left(1,\ G_1\left(1\right)\right)$ with replacement, and intervene the exposure A as 1, M_1 as $G_1\left(1\right)$, and M_2 as $G_2\left(1,\ G_1\left(1\right)\right)$. Use models built in Step 1 and MLEs from Step 2 to generate Y [denoted as $Y\left(1,\ G_1\left(1\right),\ G_2\left(1,\ G_1\left(1\right)\right)\ \right)$].

(3g) Randomly sample the confounders $\tilde{C}$, $G_1\left(1\right)$, $G_2\left(1,\ G_1\left(0\right)\right)$ with replacement, and intervene the exposure A as 1, M_1 as $G_1\left(1\right)$, and M_2 as $G_2\left(1,\ G_1\left(0\right)\right)$. Use models built in Step 1 and MLEs from Step 2 to generate Y [denoted as $Y\left(1,\ G_1\left(1\right),\ G_2\left(1,\ G_1\left(0\right)\right)\ \right)$].

(3h) Randomly sample the confounders $\tilde{C}$, $G_1\left(1\right)$, $M_2\left(0,\ G_1\left(0\right)\right)$ with replacement, and intervene the exposure A as 1, M_1 as $G_1\left(1\right)$, and M_2 as $G_2\left(0,\ G_1\left(0\right)\right)$. Use models built in Step 1 and MLEs from Step 2 to generate Y [denoted as $Y\left(1,\ G_1\left(1\right),\ G_2\left(0,\ G_1\left(0\right)\right)\ \right)$].

(3i) Randomly sample the confounders $\tilde{C}$, $G_1\left(0\right)$, $G_2\left(0,\ G_1\left(0\right)\right)$ with replacement, and intervene the exposure A as 1, M_1 as $G_1\left(0\right)$, and M_2 as $G_2\left(0,\ G_1\left(0\right)\right)$. Use models built in Step 1 and MLEs from Step 2 to generate Y [denoted as $Y\left(1,\ G_1\left(0\right),\ G_2\left(0,\ G_1\left(0\right)\right)\ \right)$].

(3j) Randomly sample the confounders $\tilde{C}$, $G_1\left(0\right)$, $G_2\left(0,\ G_1\left(0\right)\right)$ with replacement, and intervene the exposure A as 0, M_1 as $G_1\left(0\right)$, and M_2 as $G_2\left(0,\ G_1\left(0\right)\right)$. Use models built in Step 1 and MLEs from Step 2 to generate Y [denoted as $Y\left(0,\ G_1\left(0\right),\ G_2\left(0,\ G_1\left(0\right)\right)\ \right)$].

(3k) Compute the means $Y\left(a_1,\ G_1\left(a_2\right),\ G_2\left(a_3,\ G_1\left(a_4\right)\right)\right)$, $for\ i\ =\ 1,\ 2,\ 3,\ 4,\ and\ a_i\ \in\ \{0,\ 1\}$, which is the g-computation algorithm approximation estimation of $Q\left(a_1,a_2,a_3,a_4,\right)$. Based on formulae (5), we can obtain the point estimations of iTE and the four iPSEs in the OR scale.

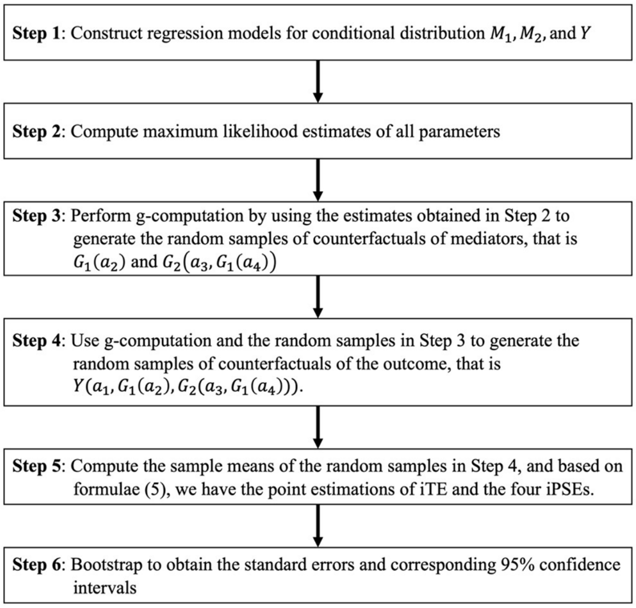
(3l) Bootstrap to obtain the standard errors and corresponding 95% confidence intervals. An R package for this analysis can be downloaded free from webpage http://shenglin.blog.nctu.edu.tw/methodology/, or see the **Supplementary Material**.

A flow chart for the proposed simulation-based approach is provided in **Figure 5**. In the approach above, randomly sampling the confounders can be replaced by just using the observed confounders if the sample size is large enough. For a small sample size, the technique of sampling the confounders with a sufficiently large sampling size could improve the stability of the g-computation algorithm approximation. The proposed estimation algorithm in Step 1 demonstrates how to construct regression models for mediators and the outcome with main effects. In practice use, the specifications of these regression models are flexible and are allowed to include any interaction effect. We evaluated the performance of the proposed method *via* a simulation study. The detail of simulation settings is provided in **Appendix B**, and the result is shown in Section Simulation Study. In Section Simulation Study, we show the operating characteristics of the new proposed estimators and compare them with traditional linear SEM estimators. Additionally, we add mediator interactions into the outcome model for evaluating the characteristics of traditional SEM under model misspecification. We evaluate the two methods by calculating the bias, the empirical standard errors (ESEs), estimated standard errors (SSEs), and coverage rates (COVs). ESE is calculated by the sample standard de*via*tion of estimates over simulations, and SSE is computed by averaging the standard error estimated by bootstrap resampling for each replication. ESEs and SSEs from the bootstrap procedure agree closely for the estimators of iPSEs, implying that the bootstrap procedure provides valid inference. Coverage rate is a proportion of the time that the 95% confidence interval obtained by bootstrap covers the true value of the parameter. In the simulation study, COVs were calculated by using 1,000 replications. If all assumptions we used in the approach are satisfied, COVs should be close to 95%. By contrast, if any assumptions are not met, COVs would be biased.

RESULTS

Simulation Study

A simulation study is conducted in **Appendix B** to show the properties of the proposed estimators and compare them with traditional linear SEM estimators. The corresponding simulation code is provided in **Appendix C**. Results are shown in **Appendix Tables 1** and **2**. Without mediator interaction (i.e. $\theta_{y,3} = 0$), both iPSE and SEM methods have small biases. The ESE and SSE values are similar in both methods. iPSE produced slightly larger ESE and SSE values than the SEM method. The coverage rates of both methods are approximately 0.95. When there exists interaction between mediators (i.e. $\theta_{y,3} = 1, \ 2, \ 3$), the biases for SEM method increase while the coverage rates approach zero with the exception of $iPSE_0$ because the SEM estimate for PSE_0 is still unbiased under this scenario. The iPSE method yielded small bias, and the coverage rate was remained approximately 95%.



Step 1: Construct regression models for conditional distribution M_1, M_2, and Y

Step 2: Compute maximum likelihood estimates of all parameters

Step 3: Perform g-computation by using the estimates obtained in Step 2 to generate the random samples of counterfactuals of mediators, that is $G_1(a_2)$ and $G_2(a_3, G_1(a_4))$

Step 4: Use g-computation and the random samples in Step 3 to generate the random samples of counterfactuals of the outcome, that is $Y(a_1, G_1(a_2), G_2(a_3, G_1(a_4)))$.

Step 5: Compute the sample means of the random samples in Step 4, and based on formulae (5), we have the point estimations of iTE and the four iPSEs.

Step 6: Bootstrap to obtain the standard errors and corresponding 95% confidence intervals

FIGURE 5 | Flow chart for the proposed simulation-based approach. M_1, M_2, and Y represent the first mediator, second mediator, and outcome, respectively.

Application to Taiwanese REVEAL-HBV Study

The performance of the proposed method was tested in the Taiwanese REVEAL-HBV dataset. Specifically, the method was used to investigate the role of HBV viral load in different time windows as a mediating mechanism in HCV-induced HCC. Here, the outcome was HCC status at the end of follow up, and the exposure of interest was HCV status at enrollment. Mediators M_1 and M_2 were HBV viral load at baseline and at follow up, respectively. Baseline confounders included gender, age, smoking status, and ALT level. All analyses were performed in R 3.4.1.

Path-specific effects were estimated using g-computation algorithm (number = 100,000) and bootstrap (resampling size = 1,000). The overall OR of HCV to HCC was 3.122 [95 % CI = (3.108, 3.226)]. For the four paths, the OR of HCV to HCC was 3.910 [95 % CI =(3.785, 4.035)] without mediation by (i.e., without change in) HBV viral load(iPSE$_0$) ; 0.873 (95 % CI = (0.853, 0.893) with mediation by baseline but not late HBV viral load (iPSE$_1$) ; 0.994 [95 % CI =(0.971, 1.018)] with mediation by late but not baseline HBV viral load (iPSE$_2$); and 0.918 [95 % CI = (0.896, 0.941)] with mediation by both baseline and late HBV viral load (iPSE$_{12}$). Note that a high OR for PSE$_0$ implies that HBV viral load change conceals the detrimental effect of HCV on HCC. **Table 1** lists the above results along with RD and RR scales.

DISCUSSION

Three common approaches to causal mediation analysis include regression-based method, weighting method, and simulation-based method. Since the simulation-based estimation is an

TABLE 1 | Total interventional effect of HCV infection on HCC incidence: four interventional path-specific effects with HBV viral load at baseline (M_1), HBV viral load at follow-up (M_2) as mediators in scales for risk difference, risk ratio, and odds ratio.

	Estimate	SE	95 % CI (lower bound)	95% CI (upper bound)
Risk difference				
Total effect	0.096	0.001	0.092	0.099
not *via* M_1 or M_2	0.127	0.001	0.123	0.130
via M_1 only	−0.019	0.001	−0.022	−0.015
via M_2 only	0.000	0.001	−0.003	0.002
via M_1 then *via* M_2	−0.011	0.001	−0.014	−0.007
Risk ratio				
Total effect	2.805	0.044	2.718	2.891
not *via* M_1 or M_2	3.385	0.050	3.285	3.484
via M_1 only	0.893	0.008	0.876	0.911
via M_2 only	0.995	0.010	0.975	1.015
via M_1 then *via* M_2	0.930	0.009	0.911	0.950
Odds ratio				
Total effect	3.122	0.053	3.018	3.226
not *via* M_1 or M_2	3.910	0.063	3.785	4.035
via M_1 only	0.873	0.010	0.853	0.893
via M_2 only	0.994	0.012	0.971	1.018
via M_1 then *via* M_2	0.918	0.011	0.896	0.941

HCV, hepatitis C virus; HBV, hepatitis B virus; HCC, hepatocellular carcinoma; SE, standard error; CI, confidence interval.

approximation of the MLE, it is asymptotically efficient provided all regression models are correctly specified. Contrarily to the regression-based method, the weighting estimation cannot achieve the efficiency bound even if the parametric assumptions for the weights are correct. Here, our approach is more flexible as it allows incorporation of non-linear, polynomial or cross-product interaction terms. Even though OR is the outcome scale of interest here, our method also allows for other non-linear outcome scales.

In some applications, portion mediated (PM) is a measure of interest to assess the proportion of the effect of the exposure mediated by the mediators. On the risk difference scale for continuous outcomes, PM for each mediation path is defined as a ratio of the corresponding iPSE to iTE. For a dichotomous outcome, a odds ratio scale is adopted to define iPSEs, and PMs would be defined on the log odds scale (8). Regardless of on the risk difference scale or the log odds scale, reporting PMs, however, is generally meaningful only if all of iPSEs are in the same direction (e.g., all positive or all negative). As the illustrative example of the Taiwanese REVEAL-HBV dataset, the effects corresponding to the paths involving HBV were negative while other effects were positive. In such a case, PM would not be an appropriate measure to reveal the extent to which mediators affect the causal effect.

There are several noteworthy limitations. Like all simulation-based methods, this approach is computationally intensive. Suppose the time of g-computation algorithm is similar to that of the regression-based method, the computation time would be five-hundred-fold if we constructed confidence intervals by 500 bootstrap repetitions. Note that our approach may be particularly prone to bias due to model misspecification. However, this drawback can be resolved by including quadratic terms for continuous independent variables in regression models and increasing model flexibility. Moreover, the assumptions of no unmeasured confounders may be violated and hard to check. Longitudinal datasets are mostly used to investigate the causal relationship between the exposure and outcome variables. Since mediation analysis or path analysis is usually the secondary analysis of longitudinal datasets, where we mainly focus on exploration of exposure-outcome relationship instead of mediator-outcome, mediator-exposure, or mediator-mediator relationships when collecting confounding variables. We could include application of sensitivity analysis techniques to address violations of these assumptions in future research. Furthermore, estimation of the simulation-based method is unstable when the sample size is small in relation to the complexity of the models, though this is not an issue here because the sample size in Taiwanese HCC cohort is relatively large. It is also worthy to note that a less complicated model is preferred for generating more stable estimations despite flexible model choices in the software.

CONCLUSION

HCC ranks sixth in cancer incidence and third in cancer mortality and is a major social burden for all nations (41). Currently, there are about 170 million HCV and 350 million HBV infected cases in the world (42). Our proposed method partially separates the mechanism of HBV and HCV infections on the incidence of HCC. Although HBV and HCV have been confirmed as two etiologic factors for HCC and classified as human carcinogens by the International Agency for Research on Cancer (43), their biological mechanisms remains elusive. Previous studies have shown that HBV and HCV have subadditive interaction on HCC incidence (44–46), and that HCV may suppress the expression and duplication of HBV (47–51). These studies provide evidence that HBV viral activity change may mask the effect of HCV on the HCC risk. In addition, a previous study showed that the early HBV viral activity is an important factor in the development of HCC (15, 16). However, due to the restriction of traditional methods, differentiation of the effects of early HBV viral activity on HCC risk through or not through late HBV viral activity remained difficult. In this study, we utilized the interventional approach to show that both pathways are statistically significant. This result implies that, though the increased HCC caused by HCV infection is not solely through the late HBV viral load ($iPSE_2$), both early and late viral load play important roles in the mechanism. Consequently, the decreasing HBV viral load in both time-points can partially prevent the HCC.

Categorical outcomes such as dichotomous or time-to-event outcomes are common, especially epidemiology and health-related fields. Although the iPSE can be identified non-parametrically, the existing regression-based method does not have a closed form (i.e., analytic solution) for non-linear outcome without the rare disease assumption. With our approach, we can ensure that the effect decomposition is applicable for non-linear outcome even without the rare disease assumption. Finally, in our study only allow measurement taken at the end of study as the outcome. It is also important to develop methods for settings with multiple mediators. This can be done by incorporating time-to-event outcome with survival models such as Cox proportional hazard model or accelerated failure time model.

In conclusion, our approach is powerful and versatile for settings with multiple mediators where the traditional PSE is not identified. Furthermore, we facilitate application for mechanism investigation in more complicated settings in epidemiology and health science.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available at http://doi.org/10.1001/jama.295.1.65.

AUTHOR CONTRIBUTIONS

A-ST: conceptualization, formal analysis, software, visualization, methodology, and writing—original draft. Y-TH: validation and writing—review and editing. H-IY: data curation, validation, and writing—review and editing. LL: writing—review and editing. S-HL: conceptualization, data curation, funding acquisition, investigation, methodology, project administration, resources, supervision, writing—original draft, and writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

A-ST and S-HL are funded by 109-2636-B-009 -001 (Ministry of Science and Technology, Taiwan). Y-TH is funded by AS-CDA-108-M03 (Academia Sinica) and 108-2118-M-001-013-MY5 (Ministry of Science and Technology, Taiwan).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.757942/full#supplementary-material

REFERENCES

1. MacKinnon DP. Introduction to statistical mediation analysis. New York, NY: Routledge. (2008).
2. Pearl J. Causal inference in statistics: An overview. *Stat Surv.* (2009) 3:96–146. doi: 10.1214/09-SS057
3. Robins J, A. new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling.* (1986) 7:1393–512. doi: 10.1016/0270-0255(86)90088-6
4. Rubin DB. Formal mode of statistical inference for causal effects. *J Stat Plan Inference.* (1990) 25:279–92. doi: 10.1016/0378-3758(90)90077-8
5. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. (1992) 143–55. doi: 10.1097/00001648-199203000-00013
6. Pearl J. Direct and indirect effects. Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence. San Francisco, CA, USA: Morgan kaufmann publishers Inc. (2001) p. 411–420.
7. VanderWeele T. Explanation in Causal Inference: Methods for Mediation and Interaction. New York, NY: Oxford University Press. (2015).
8. VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. *Am J Epidemiol.* (2010) 172:1339–48. doi: 10.1093/aje/kwq332
9. Lange T, Hansen JV. Direct and indirect effects in a survival context. *Epidemiology*. (2011) 22:575–81. doi: 10.1097/EDE.0b013e31821c680c
10. Martinussen T, Vansteelandt S, Gerster M. Hjelmborg JvB. Estimation of direct effects for survival data by using the Aalen additive hazards model. *J Royal Stat Soc.* (2011) 73:773–88. doi: 10.1111/j.1467-9868.2011.00782.x
11. Tchetgen Tchetgen EJ. On causal mediation analysis with a survival outcome. *Int J Biostat*. (2011) 7:1–38. doi: 10.2202/1557-4679.1351
12. VanderWeele TJ. Causal mediation analysis with survival data. *Epidemiology (Cambridge, Mass).* (2011) 22:582. doi: 10.1097/EDE.0b013e31821db37e
13. Valeri L, VanderWeele TJ. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with SAS and SPSS macros. *Psychol Methods.* (2013) 18:137. doi: 10.1037/a0031034
14. Lin SH, Young J, Logan R, Tchetgen Tchetgen EJ, VanderWeele TJ. Parametric mediational g-formula approach to mediation analysis with time-varying exposures, mediators, and confounders. *Epidemiology.* (2017) 28:266–74. doi: 10.1097/EDE.0000000000000609
15. Huang Y-T, Yang H-I. Causal mediation analysis of survival outcome with multiple mediators. *Epidemiology.* (2017) 28:370–8. doi: 10.1097/EDE.0000000000000651
16. Huang Y-T, Yang H-I, Liu J, Lee M-H, Freeman JR, Chen C-J. Mediation analysis of hepatitis B and C in relation to hepatocellular carcinoma risk. *Epidemiology.* (2016) 27:14–20. doi: 10.1097/EDE.0000000000000390
17. Taguri M, Featherstone J, Cheng J. Causal mediation analysis with multiple causally non-ordered mediators. *Stat Methods Med Res.* (2015) 27:3–19. doi: 10.1177/0962280215615899
18. VanderWeele TJ, Vansteelandt S. Mediation analysis with multiple mediators. *Epidemiol Method*. (2014) 2:95–115. doi: 10.1515/em-2012-0010
19. Avin C, Shpitser I, Pearl J. Identifiability of path-specific effects. Los Angeles, CA: Department of Statistics, UCLA. (2005).
20. VanderWeele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. *Epidemiology.* (2014) 25:300–6. doi: 10.1097/EDE.0000000000000034
21. Daniel R, De Stavola B, Cousens S, Vansteelandt S. Causal mediation analysis with multiple mediators. *Biometrics.* (2015) 71:1–14. doi: 10.1111/biom.12248
22. Huang YT, Cai T. Mediation analysis for survival data using semiparametric probit models. *Biometrics.* (2015). doi: 10.1111/biom.12445
23. Huang Y-T, Yang H-I, Liu J, Lee M-H, Freeman JR, Chen C-J. Mediation analysis of hepatitis b and c in relation to hepatocellular carcinoma risk. *Epidemiology (Cambridge, Mass).* (2015) 27:14–20.
24. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology (Cambridge, Mass).* (2010) 21:540. doi: 10.1097/EDE.0b013e3181df191c
25. VanderWeele TJ. Unmeasured confounding and hazard scales: sensitivity analysis for total, direct and indirect effects. *Eur J Epidemiol.* (2013) 28:113–7. doi: 10.1007/s10654-013-9770-6
26. Lin S-H, VanderWeele T. Interventional approach for path-specific effects. *J Causal Inference*. (2017) 5. doi: 10.1515/jci-2015-0027

27. Vansteelandt S, Daniel RM. Interventional effects for mediation analysis with multiple mediators. *Epidemiology (Cambridge, Mass).* (2017) 28:258. doi: 10.1097/EDE.0000000000000596
28. Imai K, Keele L, Tingley D, A. general approach to causal mediation analysis. *Psychol Methods.* (2010) 15:309. doi: 10.1037/a0020761
29. Imai K, Keele L, Tingley D, Yamamoto T. Causal mediation analysis using R. Advances in social science research using R. Springer. (2010). p. 196:129–154. doi: 10.1007/978-1-4419-1764-5_8
30. Tingley D, Yamamoto T, Hirose K, Keele L, Imai K. Mediation: R package for causal mediation analysis. *J Stat Softw.* (2014) 59:1–38. doi: 10.18637/jss.v059.i05
31. Wang A, Arah OA. G-computation demonstration in causal mediation analysis. *Eur J Epidemiol.* (2015) 30:1119–27. doi: 10.1007/s10654-015-0100-z
32. Taubman SL, Robins JM, Mittleman MA, Hernán MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol.* (2009) 38:1599–611. doi: 10.1093/ije/dyp192
33. Hernan JMRaMA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice G, Verbeke G, Molenberghs G, editor. *Longitudinal Data Analysis.* Boca Raton, FL: Chapman & Hall/CRC. (2009).
34. Westreich D CS, Young JG, Palella F, Tien PC, Kingsley L, Gange SJ, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Stat Med.* (2012). doi: 10.1002/sim.5316
35. Lin SH, Young JG, Logan R, VanderWeele TJ. Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. *Stat Med.* (2017) 36:4153–66. doi: 10.1002/sim.7426
36. Chen C-J, Yang H-I, Su J, Jen C-L, You S-L, Lu S-N, et al. Risk of hepatocellular carcinoma across a biological gradient of serum hepatitis B virus DNA level. *JAMA.* (2006) 295:65–73. doi: 10.1001/jama.295.1.65
37. Chen CL, Yang HI, Yang WS, Liu CJ, Chen PJ, You SL, et al. Metabolic factors and risk of hepatocellular carcinoma by chronic hepatitis B/C infection: a follow-up study in Taiwan. *Gastroenterology.* (2008) 135:111–21. doi: 10.1053/j.gastro.2008.03.073
38. Iloeje UH, Yang HI, Jen CL, Su J, Wang LY, You SL, et al. Risk and predictors of mortality associated with chronic hepatitis B infection. *Clin Gastroenterol Hepatol.* (2007) 5:921–31. doi: 10.1016/j.cgh.2007.06.015
39. Lee M-H, Yang H-I, Lu S-N, Jen C-L, Yeh S-H, Liu C-J, et al. Hepatitis C virus seromarkers and subsequent risk of hepatocellular carcinoma: long-term predictors from a community-based cohort study. *Journal of Clinical Oncology.* (2010) 28:4587–93. doi: 10.1200/JCO.2010.2.1500
40. Hernán M, A. definition of causal effect for epidemiological research. *J Epidemiol Community Health.* (2004) 58:265–71. doi: 10.1136/jech.2002.006361
41. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin.* (2005) 55:74–108. doi: 10.3322/canjclin.55.2.74
42. Lauer GM, Walker BD. Hepatitis C virus infection. *N Engl J Med.* (2001) 345:41–52. doi: 10.1056/NEJM200107053450107
43. WHO. *IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. Vol. 59.* International Agency for Research on Cancer, & World Health Organization. (1994). Available online at: https://monographs.iarc.who.int/wp-content/uploads/2018/06/mono93.pdf
44. Huang Y-T, Jen C-L, Yang H-I, Lee M-H, Su J, Lu S-N, et al. Lifetime risk and sex difference of hepatocellular carcinoma among patients with chronic hepatitis B and C. *J Clin Oncol.* (2011) 29:3643–50. doi: 10.1200/JCO.2011.36.2335
45. Kuper H, Tzonou A, Kaklamani E, Hadziyannis S, Tasopoulos N, Lagiou P, et al. Hepatitis B and C viruses in the etiology of hepatocellular carcinoma; a study in Greece using third-generation assays. *Cancer Causes Control.* (2000) 11:171–5. doi: 10.1023/A:1008951901148
46. Sun C-A, Wu D-M, Lin C-C, Lu S-N, You S-L, Wang L-Y, et al. Incidence and cofactors of hepatitis C virus-related hepatocellular carcinoma: a prospective study of 12,008 men in Taiwan. *Am J Epidemiol.* (2003) 157:674–82. doi: 10.1093/aje/kwg041
47. Tsiquaye K, Tovey G, Kessler H, Hu S, Lu XZ, Zuckerman A, et al. Non-A, non-b hepatitis in persistent carriers of hepatitis b virus. *J Med Virol.* (1983) 11:179–89. doi: 10.1002/jmv.1890110302
48. Liaw YF. Role of hepatitis C virus in dual and triple hepatitis virus infection. *Hepatology.* (1995) 22:1101–8. doi: 10.1002/hep.1840220413
49. Koike K, Yotsuyanagi H, Moriya K, Kurokawa K, Yasuda K, Lino S, et al. Dominant replication of either virus in dual infection with hepatitis viruses B and C. *J Med Virol.* (1995) 45:236–9. doi: 10.1002/jmv.1890450222
50. Shih CM, Lo SJ, Miyamura T, Chen SY, Lee Y. Suppression of hepatitis B virus expression and replication by hepatitis C virus core protein in HuH-7 cells. *J Virol.* (1993) 67:5823–32. doi: 10.1128/jvi.67.10.5823-5832.1993
51. Schüttler CG, Fiedler N, Schmidt K, Repp R, Gerlich WH, Schaefer S. Suppression of hepatitis B virus enhancer 1 and 2 by hepatitis C virus core protein. *J Hepatol.* (2002) 37:855–62. doi: 10.1016/S0168-8278(02)00296-9

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership