MEASURING AND ANALYSING SOCIAL DETERMINANTS OF HEALTH IN THE ERA OF BIG DATA

EDITED BY: Yi Guo, Jiang Bian and Fei Wang

PUBLISHED IN: Frontiers in Public Health

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-424-2
DOI 10.3389/978-2-88976-424-2

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

MEASURING AND ANALYSING SOCIAL DETERMINANTS OF HEALTH IN THE ERA OF BIG DATA

Topic Editors:
Yi Guo, University of Florida, United States
Jiang Bian, University of Florida, United States
Fei Wang, Cornell University, United States

Citation: Guo, Y., Bian, J., Wang, F., eds. (2022). Measuring and Analysing Social Determinants of Health in the Era of Big Data. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-424-2

Table of Contents

frontiers | Frontiers in Public Health

EDITORIAL
published: 27 May 2022
doi: 10.3389/fpubh.2022.902942

Editorial: Measuring and Analysing Social Determinants of Health in the Era of Big Data

Yi Guo[1]*, *Jiang Bian*[1]* and *Fei Wang*[2]*

[1] *Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, United States,* [2] *Department of Population Health Sciences, Weill Cornell Medical College, New York, NY, United States*

Keywords: electronic health records, natural language processing, environmental determinants, outcomes, intervention

Editorial on the Research Topic

Measuring and Analysing Social Determinants of Health in the Era of Big Data

A large and rapidly growing body of literature has provided convicting evidence on the significant role of social determinants of health (SDoH) in affecting human health, wellbeing, and quality of life (1). The World Health Organization defines SDoH as the non-medical factors that describe the conditions in which people are born, grow, live, work, and age (2). These factors include social and environmental circumstances such as education, income, housing, transportation, food access, diet, physical activity, discrimination, neighborhood safety, and many more. SDoH are one of the major contributors to the widespread health disparities and health inequities. It is estimated that SDoH are responsible for up to 40 percent of all preventable deaths in the United States (US), yet better medical care only accounts for a much smaller 10–15 percent (3). All evidence suggests that efforts to improve health need to shift from focusing on clinical factors to considering SDoH as key drivers of health outcomes.

Recognizing the importance of SDoH in shaping human health, a committee established by the US. Institute of Medicine (now the National Academy of Medicine) recommended ways of capturing 12 standardized measures from 11 SDoH domains in electronic health records (EHRs) in 2014 (4). Since then, healthcare systems began to explore ways to capture and integrate SDoH data within patients' EHRs. For example, Kaiser Permanente Northwest developed a set of EHR-based data collection tools to facilitate SDoH documentation using the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) social diagnostic codes (Z codes) (5), which are intended to document patients' social, economic, occupational, and psychosocial circumstances. However, despite the increasing efforts to collect SDoH, they are infrequently captured in EHRs. Recent studies have shown that the Z codes are rarely used by clinicians in clinical documentations (6, 7).

Considering the importance of SDoH and the lack of SDoH data in EHRs, the editors proposed this Research Topic to provide a forum for cutting-edge research on the development and application of novel methods for measuring and analyzing SDoH in health outcomes research. For example, much information on SDoH is captured in unstructured EHR fields as free-text narratives. Yu et al. developed a natural language processing (NLP) pipeline that can extract 15 categories of SDoH from clinical narratives using a transformer-based model [i.e., Bidirectional Encoder Representations from Transformers (BERT)]. These SDoH included gender, race, ethnicity, smoking, employment, education, alcohol use, substance use, marital status, occupation, language, physical activity, transportation, financial constraint, and social cohesion. Using EHRs from a large health system in the United States, the authors obtained about 1.8 million clinical notes from over

Edited and reviewed by:
Mike Conway,
The University of Utah, United States

****Correspondence:***
Yi Guo
yiguo@ufl.edu
Jiang Bian
bianjiang@ufl.edu
Fei Wang
few2001@med.cornell.edu

Specialty section:
This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

Received: *23 March 2022*
Accepted: *13 May 2022*
Published: *27 May 2022*

Citation:
Guo Y, Bian J and Wang F (2022)
Editorial: Measuring and Analysing
Social Determinants of Health in the
Era of Big Data.
Front. Public Health 10:902942.
doi: 10.3389/fpubh.2022.902942

10 thousand lung cancer patients and examined the frequencies of these SDoH in each category. Hatef et al. evaluated a text mining approach (i.e., pattern matching with regular expressions) in identifying phrases related to 5 categories of housing issues using EHRs from a large multispecialty medical group in the United States. Collaborating with SDoH experts, the research team reviewed existing literature and coding standards, and developed phrases addressing each housing issue and pattern-matching algorithms. Using data in 2.5 million clinical notes from 20 thousand patients, the authors found that, compared to manual annotation, the regular expression approach had a high level (> 94%) of precision at the phrase, note, and patient levels across different housing issues, although the recall level was relatively low.

Four articles in this Research Topic reported results from empirical analyses of the impact of SDoH in diverse research areas, including urbanization (Fang et al.), treatment adherence (Daabek et al.), liver cancer survival (Wu et al.), and happiness of rural residents (Xu and Ge). Another four review or opinion articles discussed the importance of collecting SDoH, identified areas of improvement, and proposed action plans, in the areas of sexual minority health (Wu et al.), patient segmentation (Rezaeiahari), physiatry (i.e., physical medicine and rehabilitation) (Conic et al.), and pediatric cancer (Reeves et al.). The remaining article simulated time-to-event data under various missing mechanisms (e.g., missing not at random) and assessed the performance of machine learning missing data imputation techniques based on the Cox proportional hazard model (Guo et al.). Given the poor documentation of SDoH in EHRs, methods for handling missing data are much needed in health outcomes research.

Overall, although SDoH are important factors driving health outcomes, they are poorly documented in EHRs. Even when SDoH are documented in EHRs, they are buried in unstructured clinical narratives and thus not readily accessible for downstream studies of health outcomes. As a result, current clinical research mainly studies the impacts of clinical factors (e.g., disease history, medical treatment) on health outcomes (e.g., prognosis, survival), ignoring the perhaps more important SDoH (e.g., financial constraint, housing issues) as contributing factors. Advances in research methods such as NLP provides new opportunities for identifying and capturing SDoH. However, what is really needed is targeted (and tailored) SDoH collection supported by EHR-based data collection tools, rather than using the non-specific Z codes. First of all, not all SDoH factors are equally important across the continuum of patient care or research areas. For example, insurance and geographic location (e.g., rural vs. urban residency) are more important for acute care settings, whereas social support and living conditions are more important for post-acute care and outpatient settings. A deep understanding of important SDoH in different phases of patient care is required for designing effective interventions that aim to improve health outcomes. Further, different representations or measures of the same SDoH factor are likely needed in different research areas. For example, physical activity can be measured using self-report survey instruments or wearable devices, depending on the desired level of data granularity. There needs to be clear guidelines, by type of disease and patient care, that outline which standardized SDoH measures to collect and how they should be integrated with patient EHRs.

Another more realistic solution to the lack of SDoH in EHRs is linking SDoH from other data sources. SDoH data are widely available in many local and national data sources such as Bureau of Labor Statistics (e.g., employment), Department of Education (e.g., literacy), Census Bureau (e.g., food insecurity, poverty), Bureau of Justice Statistics (e.g., Incarceration), Environmental Protection Agency (e.g., environmental factors) and many more. Linking these SDoH data longitudinally to patient EHR data at either the individual- or contextual level is the next essential step in advancing health outcomes research.

AUTHOR CONTRIBUTIONS

YG drafted the manuscript. JB and FW reviewed and revised the manuscript. All authors approved the final version of manuscript.

FUNDING

YG and JB were funded in part by the National Institutes of Health (NIH), National Cancer Institute (NCI) grants 5R01CA246418-02, 3R01CA246418-02S1, 1R21CA245858-01A1, 3R21CA245858-01A1S1, and 1R21CA253394-01A1, NIH National Institute on Aging (NIA) grant 5R21AG068717-02, and Centers for Disease Control and Prevention (CDC) grant U18DP006512.

REFERENCES

1. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. *Public Health Rep*. (2014) 129(Suppl. 2):19–31. doi: 10.1177/00333549141291S206
2. World Health Organization. *Social Determinants of Health*. Available online at: https://www.who.int/health-topics/social-determinants-of-health (accessed Mar 19, 2022).
3. Danaei G, Ding EL, Mozaffarian D, Taylor B, Rehm J, Murray CJL, et al. The preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. *PLoS Med*. (2009) 6:e1000058. doi: 10.1371/journal.pmed.100005
4. Committee on the Recommended Social and Behavioral Domains and Measures for Electronic Health Records, Board on Population Health and Public Health Practice, Institute of Medicine. *Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2*. Washington, DC: National Academies Press (US) (2015).
5. Nicole L Friedman MPB. Toward addressing social determinants of health: a health care system strategy. *Perm J*. (2018) 22:18-095. doi: 10.7812/TPP/18-095
6. Guo Y, Chen Z, Xu K, George TJ, Wu Y, Hogan W, et al. International Classification of Diseases, Tenth Revision, Clinical Modification social determinants of health codes are poorly used in electronic health records. *Medicine*. (2020) 99:e23818. doi: 10.1097/MD.0000000000023818

7. Truong HP, Luke AA, Hammond G, Wadhera RK, Reidhead M, Joynt Maddox KE. Utilization of social determinants of health ICD-10 Z-codes among hospitalized patients in the United States, 2016-2017. *Med Care*. (2020) 58:1037–43. doi: 10.1097/MLR.0000000000001418

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: 07 June 2021
doi: 10.3389/fpubh.2021.677910

Spatial Characterization of Urban Vitality and the Association With Various Street Network Metrics From the Multi-Scalar Perspective

Chuanglin Fang[1], *Sanwei He*[1,2]* *and Lei Wang*[3,4]

[1] Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, [2] School of Public Administration, Zhongnan University of Economics and Law, Wuhan, China, [3] Key Laboratory of Watershed Geographic Sciences, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, China, [4] Department of Planning and Environmental Management, University of Manchester, Manchester, United Kingdom

OPEN ACCESS

Edited by:
Yi Guo,
University of Florida, United States

Reviewed by:
Hui Hu,
University of Florida, United States
Qingsong He,
Huazhong University of Science and Technology, China

***Correspondence:**
Sanwei He
sanwei.87@163.com

Specialty section:
This article was submitted to Environmental health and Exposome, a section of the journal Frontiers in Public Health

Received: *08 March 2021*
Accepted: *22 April 2021*
Published: *07 June 2021*

Citation:
Fang C, He S and Wang L (2021) Spatial Characterization of Urban Vitality and the Association With Various Street Network Metrics From the Multi-Scalar Perspective. Front. Public Health 9:677910. doi: 10.3389/fpubh.2021.677910

In the context of rapid urbanization in developing countries, the spatial organization of cities has been progressively restructured over the past decades. However, little has been done to understand how the physical expansion affected the reorganization of socioeconomic spaces in cities. This study explores the association between various street network metrics and urban vitality and how it changes across different scales using geographic big data through a case study of Wuhan, China. Urban vitality is characterized by four components: concentration, accessibility, livability, and diversity. The new technique of spatial design network analysis (sDNA) is employed to characterize street network metrics, including connectivity, closeness, betweenness, severance, and efficiency, with 16 localized network variables. Furthermore, the stratified spatial heterogeneity between street network metrics at multiple scales and the four components of urban vitality is investigated using the Geodetector tool. First, concentration, accessibility, and diversity decline with distance from the urban center, whereas livability has a fluctuating upward trend with distance from the urban core. Second, the correlation between street network characteristics and urban vitality is sensitive to different spatial scales. Third, connectivity explains the largest amount of the variance in urban vitality (over 40%), while both betweenness and closeness explain roughly 28% of urban vitality. Efficiency and severance contribute 22 and 10% to the spatial heterogeneity of urban vitality, respectively. The study sheds light on the mechanisms between street configurations and urban vitality from the multi-scalar perspective. Some implications are provided for the improvement of the streets' urban vitality.

Keywords: urban vitality, spatial design network analysis, spatial scales, big data, Wuhan

INTRODUCTION

The world is experiencing rapid urbanization, especially in developing countries (1). By 2050, two thirds of the world's population will reside in cities. Although urbanization visibly improves standards of living, it also carries risks, such as shrinking or ghost cities[1], which are associated with considerably low urban vitality (2–5). As an important proxy for the sustainability of urban growth (6), urban vitality measures the attractiveness and competitiveness of a city. As a new source of urban competitiveness, urban vitality helps a city gain comparative advantages and thus produce sustainable economic growth and endure regional innovation (7, 8). Understanding urban vitality is essential to urban health monitoring, compact urban development, innovative urban growth, and people-oriented urbanization.

The concept of urban vitality revolves around people's satisfaction with all aspects of urban life (9). It is difficult to capture the extensive meanings of urban vitality in specific measures. Many scholars have tried to measure urban vitality from different aspects. Yue et al. (10) measured urban vitality by the dimensions of built environment, human activities, and human–environment interaction. Pugalis (11) subdivided urban space into economy, culture, and society, classifying urban vitality as economic, cultural, and social vitality. A quantitative model of the determinants of urban vitality can be an informative tool to achieve sustainability in spatial planning and urban design (12, 13). Following the conceptual framework proposed by Jacobs (14) and Gehl (15), many scholars define and quantify urban vitality in terms of multiple facets (e.g., density, accessibility, and diversity), multiple spatial scales (e.g., *jiedao* units, community neighborhoods, and street blocks), and multiple temporal horizons (e.g., night time, twilight, and early morning) (16, 17). The microscale analysis of communities or street blocks provides a granular and comprehensive perspective of the spatiotemporal dynamics of urban vitality in cities (18). Moreover, the advent of geographic big data such as location-based social media, points of interest (POIs), mobile Internet data, and other web crawler data allows researchers to capture spatial dynamics of urban vitality more accurately and at finer scales.

The association between urban form metrics and urban vitality has been widely recognized in the literature (19). Empirical studies of both developed and developing countries have examined the inherent vital nature in urban areas and provided different conceptualizations of urban vitality in various territorial contexts. Jacobs (14) proposed that land-use mixture, block size, age of buildings, density, accessibility, and border vacuums are the major components of measures of urban vitality in the United States. Using his framework, Delclòs-Alió and Miralles-Guasch (16) also interpreted urban vitality in Barcelona, Spain. Sung et al. (20) found an association between a diverse physical environment and walking activity on the streets in Seoul, South Korea. In addition, Long and Huang (21) found significant and positive influences of urban design variables such as land-use mix, road intersection density, and accessibility to various facilities on economic vitality for the 286 largest cities in China. Similarly, Yue et al. (10) discovered that urban vitality measurement is closely linked with built environment (e.g., buildings, blocks, and land types), human activities (e.g., concentration of residents, employees, and tourists), and human–environment interaction (e.g., infrastructure, road network, natural vacuums, and artificial segregations) in Shanghai, China. In the context of vitality debates worldwide and the prospects of future urban sustainability, it is emergent for modern planners to deeply analyze and realign the urban spatial structure to promote the necessary interactions and provide a sufficient physical environment for the socioeconomic dynamics. This study aims to better understand the spatial organization of cities from the multi-scalar perspective of street network design and how urban vitality is related to their distributions.

Previous studies have widely acknowledged that street configurations have significant effects on physical activities in streets and thus urban vitality (22). Focusing on the streets of Cypriot towns, Jalaladdini and Oktay (23) argued that good connections to the street, the harbor, or the historic quarter and proximity to important magnets are essential to understanding the issue of vitality in urban public spaces. Another study taking Dutch towns and new Chinese towns as examples reported that spatial configurations of a street network, in terms of topological, geometrical, and metric distances, directly determine economic vitality and encourage vibrant street life (24). Using network centrality indices, Kang (25) examined the effects of street network configurations on walking mobility in Seoul, Korea. While the existing literature has measured various aspects in terms of typology, geometrics, and network connectivity, the discussion has rarely considered the scale effect when measuring the street configurations. Notably, the metrics of the street layout can be quantified differently within the neighborhood environment as defined by various buffer widths (26). Hajrasouliha and Yin (27) argued that the gridiron street patterns with small or large blocks have various effects on pedestrian volumes, and the street networks with small block sizes help pedestrians to understand their surroundings better. He et al. (28) proved that the street configurations under walking or driving modes should be differently measured, which has an influence on the distribution of leisure entertainment facilities. Thus, it is critical to measure the street configurations from a multi-scale perspective and compare how urban vitality is associated with street configurations at different geographic scales.

In addition to the topological, geometric, and distance features of the street layout, the network metaphor has a long tradition in the analysis of urban planning and transportation (29, 30). More recently, the centrality assessment model and space syntax analysis have been used to evaluate the structural properties of street networks in an urban system. Street centrality indices representing closeness, betweenness, and straightness

[1] In the era of globalization, the phenomenon of cities growing slowly or declining has become a significant international political and economic issue. A shrinking city can be characterized by population loss, financial crisis, employment decline, and other social problems.

capture the skeleton of the urban system; these factors shape economic activities and land-use intensity (31, 32). Space syntax focuses on topological distance within a network and offers an effective tool to measure street connectivity (33). However, these techniques fail to capture the challenges of physical severance and network efficiency, especially the navigating difficulties and psychological barriers of pedestrians (28, 34). One of these measurements is directly mirrored in spatial design network analysis (sDNA), which incorporates six important features (density, connectivity, closeness, betweenness, severance, and efficiency) that are hypothesized to affect urban vitality in an urban system.

As a representative developing country, China has been undergoing rapid urbanization and economic growth. Driven by both industrialization and urbanization, numerous satellite towns, industrial zones, commercial centers, and residential areas have emerged at the urban fringe and pose significant challenges to sustainable urban development (35). Meanwhile, the opposite phenomenon of "shrinking cities" in China has been a catalyst for the proposal of the "National New-type Urbanization Plan (2014–2020)," which specifically stresses people-oriented urbanization and human well-being (36). Urban vitality was initially studied as an important path to new-type urbanization in North America. Until recently, empirical evidence in developing countries has indicated that the dynamic process and determinants of urban vitality can be different from the case studies in the United States or Europe due to their dissimilar urban morphology and spatial planning (37–39). Therefore, this study aims to enrich the existing empirical studies by taking an inland city of China as an example and formulating policy implications to enhance urban vitality and promote urban health.

This study contributes to the literature in the following three ways. First, due to the inconclusive evidence in the literature, this study aims to examine the inherent vital nature within a neighborhood community in an inland city of China using geographic big data. Second, as existing studies have rarely considered the scale effect when measuring street configurations, this study adopts a multi-scale perspective incorporating street networks to examine how urban vitality is associated with spatial network layouts at different geographic scales. Third, the study applies a newly developed technique called sDNA to capture the navigating difficulties and psychological barriers faced by pedestrians, which are hypothesized to affect urban vitality.

TABLE 1 | The description of street network configurations.

Metric	Name (abbrev.)	Description
Connectivity	Connectivity in radius (CONN)	The total number of link ends connected at each junction
	Junctions in radius (JNC)	The number of junctions in the radius
Closeness	Mean Euclidean distance in radius (MED)	The mean length between an origin and all destinations in the radius
	Network quantity penalized by distance in radius Euclidean (NQPDE)	The mean length of network weight is divided by network quantity in the radius
	Angular distance in radius (ANGD)	The total angular curvature on all links in the radius
Betweenness	Betweenness Euclidean (BTE)	The number of geodesic paths that pass through a vertex
	Two-phase betweenness Euclidean (TPBTE)	The sum of geodesics that pass through a link, weighted by the proportion of network quantity
	Two-phase Destination Euclidean (TPD)	The proportion of origin weight received by each destination in the two phase betweenness model
Severance	Mean crow flight distance in radius (MCF)	The mean of the crow flight distance between each origin and all links in the radius
	Diversion ratio in radius Euclidean (DIVE)	The mean ratio of geodesic length to crow flight distance over all links in the radius
	Mean geodesic length in radius Euclidean (MGLE)	The mean length in Euclidean metric of all geodesics in the radius
Efficiency	Convex hull area (HULLA)	The area of the convex hull covered by the network in the radius
	Convex hull perimeter (HULLP)	The perimeter of the convex hull covered by the network in the radius
	Convex hull maximum (crow flight) radius (HULLR)	The distance from the origin to the point where the convex hull has its greatest radius
	Convex hull bearing of maximum radius (HULLB)	The direction of the projected grid for HULLR
	Convex hull shape index (HULLSI)	The perimeter of the convex hull divided by the area of the convex hull

CONCEPTUAL FRAMEWORK

The association between street network configurations and urban vitality must be measured at multiple scales to assess the design-oriented characteristics of street networks that most influence street activities and then urban vitality. All networks are composed of nodes and links. Within street networks, nodes represent junctions or intersections between streets. As a specific type of spatial network, nodes always have distinct geographic locations, and links always have a physical shape (34). From the perspective of sDNA, the street network configurations—connectivity, closeness, betweenness, severance, and efficiency—as seen in **Table 1**, are hypothesized to affect urban vitality in an urban system.

The connectivity aspect refers to the quantity of link ends connected at each junction and the number of junctions within the user-defined radius. More connected street networks tend to have many short links, numerous intersections, and minimal dead ends. Good street connectivity means providing various routes from residential neighborhoods to destinations such as schools and shops by walking or driving. The benefits of better connectivity include improved ease of mobility through the road network, high reachability from an origin to the desired destination, less traffic congestion, and a safer street environment (40). Street connectivity is significantly correlated with active transportation, which helps to create more walkable and livable communities (41). However, how street connectivity is associated with urban vitality remains to be studied.

The closeness aspect refers to the mean Euclidean distance, the network quantity penalized by distance, and the angular distance within the radius. As a form of network centrality, closeness measures the difficulty of navigating to all possible destinations from each link in the radii. This study emphasizes the significance of angular distance (distance measured in terms of angular change) and prefers to use the shortest angular paths instead of Euclidean distance. In reality, pedestrians and drivers tend to follow straight roads and angular geodesics because they are easier to remember and tend to be faster on average (42). The accessible locations for pedestrians and drivers are always linked with diversified land-use and a convenient street layout (43). The closeness measures can reflect accessibility and flow potential, which is conducive to creating a vibrant neighborhood.

The betweenness aspect measures "through-movement" on spatial networks. The conventional betweenness indicator is based on the idea that a vertex is central if it lies on the shortest path between other vertices. Previous studies have proven that betweenness largely determines the location of retail shops and services in the urban area (44). Higher betweenness is always associated with high housing prices and rent, traffic flow, population density, and the flow of commuting in street networks (45). This study proposes two new betweenness indices—two-phase betweenness and two-phase destination. The former calculates the betweenness weighted by the network quantity, while the latter measures that weighted by both the origin and destination. In a sense, the standard betweenness can be seen as an opportunity model, whereas the new betweenness indicators correspond to transport models with trip generation and distribution phases. The betweenness concept measures the complex flow of commuting and can produce vibrant economic activities.

The severance aspect belongs to network detour analysis and mainly measures the degree the network deviates each other from the most direct path. By comparing straight-line distance to actual network distance, the severance measures can effectively reflect the twistedness of the localized network (28). In this study, the severance indices include mean crow flight distance, diversion ratio, and mean geodesic length in the radius, which proxy the physical and psychological separation in the network. High severance represents the more cognitive difficulties of pedestrians or drivers when navigating the road network. The severance can reflect the network characteristics in more detail and may indicate unfavorable locations for vibrant commercial activities.

The efficiency aspect belongs to network shape analysis and refers to the form of the overall spatial footprint within the user-defined network. It measures the overall efficiency of the network in a sophisticated manner by considering the shape of links and the spatial structure of link connections (34). High efficiency represents high frequent navigation through the network on foot or by automobile. While the road design in the urban center does not provide an efficient walking environment for pedestrian interaction, the major roads connecting the urban core and urban fringe always have low efficiency for driving (28). An efficient network tends to create more opportunities for daily interaction such as relaxation and leisure activities.

The key component of measuring street network metrics is to choose a spatial scale of interest. This defines how much of the "surrounding network" we consider when computing statistics for each individual link. Corresponding to urban vitality, the spatial scale possibly refers to the surrounding environment within a specific network radius where most daily activities take place. The network radius can be defined by the distance from either walking or driving (34). The multi-scalar perspective compares how each street network metric is spatially associated with urban vitality over five different scales of interest (500, 1,000, 1,500, 2,000, and 2,500 m). There is no single optimal scale for assessing urban vitality and designing street networks. The existing literature has rarely compared how different urban forms are associated with social interaction, physical activities, and land-use diversity. Thus, the spatial explicit analysis between the multi-scalar street network metrics (e.g., density, connectivity, closeness, betweenness, severance, and efficiency) and urban vitality yields important information for human-scale street design and land-use planning.

METHODS

Description of the Study Area

At a longitude of $11341^{'} \sim 11505^{'}$ and a latitude of $2958^{'} \sim 3122^{'}$, Wuhan is situated in the eastern part of Hubei and is the largest megacity in central China. In 2018, the permanent population of Wuhan amounted to 11.08 million people, and the gross domestic product (GDP) for the city was 1.48 trillion RMB. Per capita GDP in Wuhan (135,136 yuan) ranked first among cities in central China. Primary, secondary, and tertiary industries comprise 2.4, 43.0, and 54.6% of the economy, respectively, with the service industry being dominant. The multimodal street network in Wuhan, known as a national bus city in China, is highly developed and includes roadways, subways, railways, waterways, and greenways. The local government has emphasized transit-oriented development to promote smart urban growth, inject vitality into old towns, and create a more sustainable transport system.

The central city of Wuhan is divided by the Yangtze River and the Han River into seven districts. There are 89 *jiedaos* and over 1,076 community neighborhoods[2] in central Wuhan. Most physical activities of human life take place in the public spaces of community neighborhoods. As the center of human

[2] In China, the hierarchical administrative structure is provinces/prefecture-level cities/counties/towns. In each prefecture-level city, there are many city districts in charge of *jiedao*-level units. The community neighborhood is the lower governance unit of *jiedao*s.

activities, the community neighborhood provides an important space for people to live, rest, socialize, produce, and work. Thus, a fine spatial scale is necessary to capture the details of street configurations and the patterns of residents' social interaction in daily life.

There are several historic community neighborhoods in the inner city of Wuhan. As the old residential quarters, these neighborhoods face difficulties associated with aging municipal public infrastructure and a lack of public services. Since 2016, the local government has focused on urban redevelopment and advanced some reform policies to reconstruct these areas. Meanwhile, because of traffic congestion and poor living conditions in the urban core, more people have chosen to live in the suburbs, which feature spacious housing and better community environments. Revitalizing old towns and strengthening vitality in the suburbs are important issues for the local government. This study employs a spatial network analysis tool to characterize the distribution of urban vitality and investigate how street network designs can create a vital city.

Data Sources

Demographic Data

The 6th National Census survey in 2010 provides precise demographic data aggregated to the administrative boundaries (provinces-prefectures-counties/urban districts-townships/*jiedao*s). As the most detailed demographic data can only be obtained by the public at the *jiedao* level, the demographic data at the scale of community neighborhood are missing. The WorldPop dataset provides gridded population maps with 100-m spatial resolution for each country in the dataset. However, Ye et al. (46) find that the WorldPop dataset underestimates the population in urban areas and overestimates it in rural areas of the Chinese mainland. Thus, this paper utilizes the improved population images developed by Ye et al. (46), with higher accuracy than WorldPop for China by integrating remotely sensed and POI data within a random forest model.

Points of Interest Data

POIs refer to all geographic entities that can be abstracted into points. Each POI observation contains information such as latitude, longitude, names, and addresses. The POI data in this study are taken from the Baidu map (http://map.baidu.com/). There are 66,161 POIs within the study area in 2016, including the following: shopping malls, petrol stations, restaurants, tourist attractions, banks, parks, chess and card rooms, theaters, karaokes, stadiums, bars, hospitals, hotels, bus stations, universities, ATMs, and government agencies. These POI data can be categorized into nine classes: financial service facilities, research and educational facilities, cultural facilities, health service facilities, leisure and recreational facilities, commuting facilities, government agencies, catering services, and lodging services. The POI data reflect the venues of social activities and identify the functional diversity in the urban core.

Additional Geographic Big Data

This study utilizes additional geographic information about housing prices and age to measure the livability of community neighborhoods. This information is collected from *Fangtianxia*, the largest real estate transaction platform in China (http://fang.com). The kriging interpolation method[3] is applied to obtain the spatial pattern of housing prices and age with a spatial resolution of 100 m in Wuhan. Building density is an important indicator that reflects the concentration aspect of urban vitality. WorldView-2 image with a high spatial resolution of 0.6 m is utilized to draw the building bases in the study area. Subsequently, it is overlaid with the boundary of each community neighborhood to obtain the building density of each spatial unit. ArcGIS 10.2 is employed to calculate the Euclidean distance between each community neighborhood and all bus stops and subway stations to measure accessibility.

Method

Quantification of Urban Vitality

Urban vitality is measured with four major components: concentration, accessibility, livability, and diversity. A dense concentration of people, buildings, and social activities is the most basic condition for ensuring that an urban area is vital. The concentration component can be evaluated by three variables: population density, building density, and POI density. According to Jacob (14), a vibrant city also requires high accessibility on foot and by public transport, contrary to car-dependent urban planning. The accessibility component can be quantified in terms of distances to bus stops and subway stations. If a certain balance between new and aged buildings is maintained, diversity from both a land-use and social perspective can be strengthened (47). A precarious housing market and high house prices are likely to cause residential segregation and social polarization. The livability component can be assessed by housing prices and age. Highly diversified neighborhoods and streets shape urban vitality mainly by increasing social interactions. The diversity component is evaluated by land-use diversity. The formulas for calculating concentration, accessibility, livability, and diversity are as follows:

$$Con_i = f(popd_i,\ bd_i,\ POID_i) \tag{1}$$

$$Acc_i = f(dis\_bs_i, dis\_sw_i) \tag{2}$$

$$Liv_i = f(hage_i,\ hpr_i) \tag{3}$$

$$Div_i = f\left(landuse_i\right),\ landuse_i = -\sum_{i=1}^{n}(p_i \times 1n\, p_i) \tag{4}$$

In the final equation, p_i is the proportion of the i^{th} POI type among the total number of POI records, and n is the total number of POI types. The weights of all variables are determined by the entropy method.

Spatial Stratified Heterogeneity Analysis

Spatial heterogeneity characterizes the local variance of spatial dependence (48). Spatial stratified heterogeneity, which stresses

[3] The popular interpolation methods include inverse distance weighted (IDW) and kriging. The authors use these two methods to calculate the pattern of housing prices and age. By comparing the model performances, the results produced by kriging conform to the spatial trend of housing prices and age with lower prediction errors over the map.

the between-strata variance more than the within-strata variance, is used in many fields of natural and social sciences. The geographic detector developed by Wang et al. (49) is a new statistical technique for detecting spatial differentiation and investigating the factors that drive this spatial phenomenon. This tool consists of four functions: factor, interaction, risk, and ecological detectors. This study primarily employs the factor detector to examine the spatial stratified heterogeneity of the dependent variable Y (urban vitality) and the determinant power of independent variables X (various street network metrics). The contribution of the covariate X to the spatial heterogeneity of urban vitality can be formulated by the q statistic as follows:

$$q = 1 - \frac{SSW}{SST} \tag{5}$$

$$SSW = \sum_{h=i}^{L} N_h \sigma_h^2 \tag{6}$$

$$SST = N\sigma^2 \tag{7}$$

where SSW refers to the within sum of squares, and SST is the total sum of squares. $h = 1, 2, \ldots, L$ is the strata of covariate X; N_h and N are respectively the number of spatial units in strata h and the whole area; σ_h^2 and σ^2 refer to the variance of Y in strata h and the whole area, respectively.

$q \in [0, 1]$ indicates that X contributes $100 \times q\%$ to the pattern of Y. A large value of q indicates the large contribution of X to Y. If q is equal to 0, then no association of X and Y is shown. If q is equal to 1, the distribution of Y can be completely explained by X.

RESULT ANALYSES

Spatial Characterization of Urban Vitality

Figure 1 demonstrates the spatial distribution of urban vitality, showing the community neighborhoods with the highest value of urban vitality. The high values of urban vitality are mainly clustered in the downtown area, whereas the uptown area is characterized by low urban vitality. The community neighborhood with the highest value of urban vitality is located near the city government, namely, the Jianghan walking street, a famous century-old commercial street that features shopping, entertainment, tourism, and culture. In the grouping analysis, urban vitality is categorized into four classes in **Figure 2**: high vitality, moderate vitality, low vitality, and non-vital areas. High-vitality areas account for 7% of all community neighborhoods in Wuhan. These communities mainly correspond to the traditional city centers, which are characterized by dense population, pedestrian-friendly street networks, and diversified built environments. The moderate-vitality category takes up around 38% of all neighborhoods in the study area. These communities are regarded as the transition buffer between high- and low-vitality areas, and they are expected to have a certain level of vibrant street life. Respectively, 43 and 12% of community neighborhoods are categorized as low-vitality and non-vital areas. These areas are mainly located in the disadvantaged urban fringe near agrarian, natural, or industrial lands.

The left column in **Figure 2** demonstrates the spatial patterns of concentration, accessibility, livability, and diversity. Overall, both concentration and accessibility decline as distance from the urban center increases, while livability and diversity show different patterns. A core-periphery pattern can be observed in urban concentration. The high values of concentration are mainly found in historical neighborhoods, which meet the requirements of Jacobs (14) of a dense distribution of people, buildings, and streets. Accessibility is primarily explained by a centric-periphery logic in which the downtown is equipped with good transport infrastructure. Areas near shopping malls, big city hospitals, municipal governments, office buildings, and schools have good accessibility. Conversely, while urban suburbs with low housing prices and good residential environments show good livability, the urban center has the majority of old residential buildings without elevators, sports facilities, or sufficient open spaces. However, as some successful urban renewal projects have recently been implemented in the downtown, livability within the urban core is expected to improve in the future. It is rather difficult to depict the pattern of diversity clearly. The neighborhoods near the Yangtze River and the Han River are less diversified as they are constrained by urban planning and aquatic environment protection.

The right column in **Figure 2** depicts how concentration, accessibility, livability, and diversity change with the distance to the city center. As the key component of urban vitality in Jacobs' arguments, concentration decreases rapidly as the distance to the city center increases, despite an upturn in the interval between 18 and 21 km, where some new towns labeled "high-tech zones" or "ecological cities" are built. Characterized by convenient transportation, pleasant living environments, and more employment opportunities, these new towns are becoming the new urban center and attracting several people to live and work there. Accessibility slowly declines when the distance from the city center is <15 km. Subsequently, accessibility drops rapidly despite slight increases in remote urban new towns. However, livability increases sharply when the distance to the city center increases to 9 km, indicating the low livability in the downtown. Livability fluctuates slightly in areas between the second and third ring roads, increases sharply in the urban suburbs, and subsequently decreases rapidly. It seems that the urban fringe is rather uniformly unlivable, indicating the heterogeneous space of livability in the urban fringe. Diversity demonstrates an upward trend within the buffer of 4.5 km and subsequently levels off within the third ring road. Diversity decreases dramatically after the distance of 15 km despite a slight upturn in the new town.

The Correlation Between Street Network Metrics and Urban Vitality

Spatial network analysis is conducted to create statistics that describe the multi-scalar configuration of street networks. Contrary to techniques such as street centrality, space syntax, and accessibility analyses, sDNA comprehensively describes

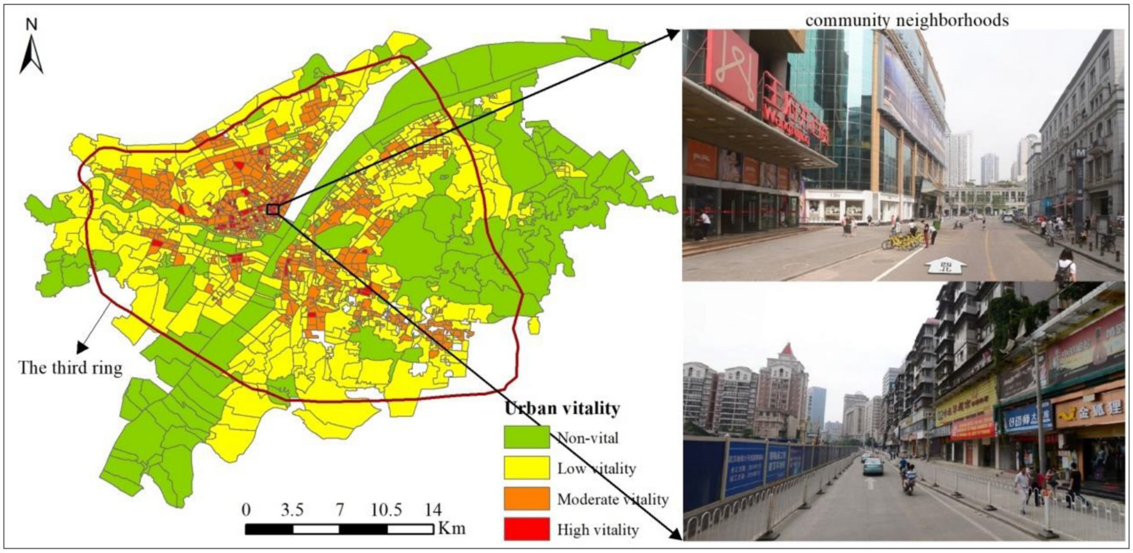


FIGURE 1 | The spatial distribution of urban vitality and community neighborhoods with the highest value of urban vitality (Red denotes high vitality; orange denotes moderate vitality; yellow denotes low vitality; and green denotes non-vital areas).

network features, including centrality, network shapes, and the navigability of areas, at user-defined network scales. A key component of sDNA is the standardization of network links, which avoids the modifiable areal unit problem (MAUP) by dividing the network into individual links. Another key component of sDNA is the specification of a scale of interest. To detect the influence of different scales on urban vitality, this study chooses five spatial scales–500, 1,000, 1,500, 2,000, and 2,500 m—providing a wide range from sensible walking distances to driving distances.

A total of 16 variables representing density, connectivity, closeness, betweenness, severance, and efficiency are computed using the sDNA software developed by Cardiff University (https://www.cardiff.ac.uk/). These 16 localized network measures, shown in **Table 2**, are assumed to affect urban vitality in Wuhan. The Pearson correlations between these network variables and the neighborhood urban vitality are calculated at each of the five spatial scales to determine how these network characteristics are associated with urban vitality and at which scale the association is most evident.

Density captures certain built environment features, such as the density of jobs and homes. The hypotheses behind these two variables are an optimum built environment density for urban vitality. The best correlation coefficients between dense built environment and urban vitality amount to 0.68 for LINK and 0.69 for LEN, both at the spatial scale of 1,000 m. The scale of 1,000 m implies the built environment within the walking distance. More density of street links is highly linked with higher urban vitality, possibly denoting that people prefer to walk within the 1,000-m neighborhood and thus produce more socioeconomic activities when navigating the street networks.

Connectivity measures how well streets are linked with others and the density of intersections. Neighborhoods with high street connectivity tend to have streets with several short links, numerous intersections, and few dead ends, which facilitate physical activities such as walking and cycling. The best correlation coefficients between connectivity variables and urban vitality reach 0.65 at the spatial scale of 1,500 m, which corresponds to the cycling distance. The best correlation between connectivity and urban vitality at the scale of 1,500 m represents that well-connected streets facilitate cycling within the neighborhood and then encourage more physical activities.

Closeness reflects the level of accessibility and reachability between origins and destinations. Previous studies emphasize the shortest Euclidean path to measure closeness; however, sDNA utilizes angular analysis to reflect the cognitive difficulty of navigation. The shortest angular path can reflect the subtleties in the network layout. According to the best correlation coefficients, network quantity penalized by distance in radius Euclidean (NQPDE) and angular distance in radius (ANGD) are more related with urban vitality at the spatial scale of 2,500 m than at other scales. The scale of 2,500 m corresponds to the driving distance. The high correlation between closeness and urban vitality at the scale of 2,500 m signifies that a street network design with less angular distance is conducive to driving and can encourage more diversified activities in the neighborhood.

Betweenness measures how street networks are populated with entities when traveling from origins to destinations in the user-specified radius. It involves all possible trips that pass through the link in the radius and can effectively reflect the flow volume through walking or traffic to the destination. Contrary to the normal betweenness model, the two-phase betweenness model considers a fixed amount of weight in each origin or destination according to the quantity of visits per link. The best correlation coefficients between these three variables and urban vitality are 0.60 for betweenness Euclidean (BTE) in the

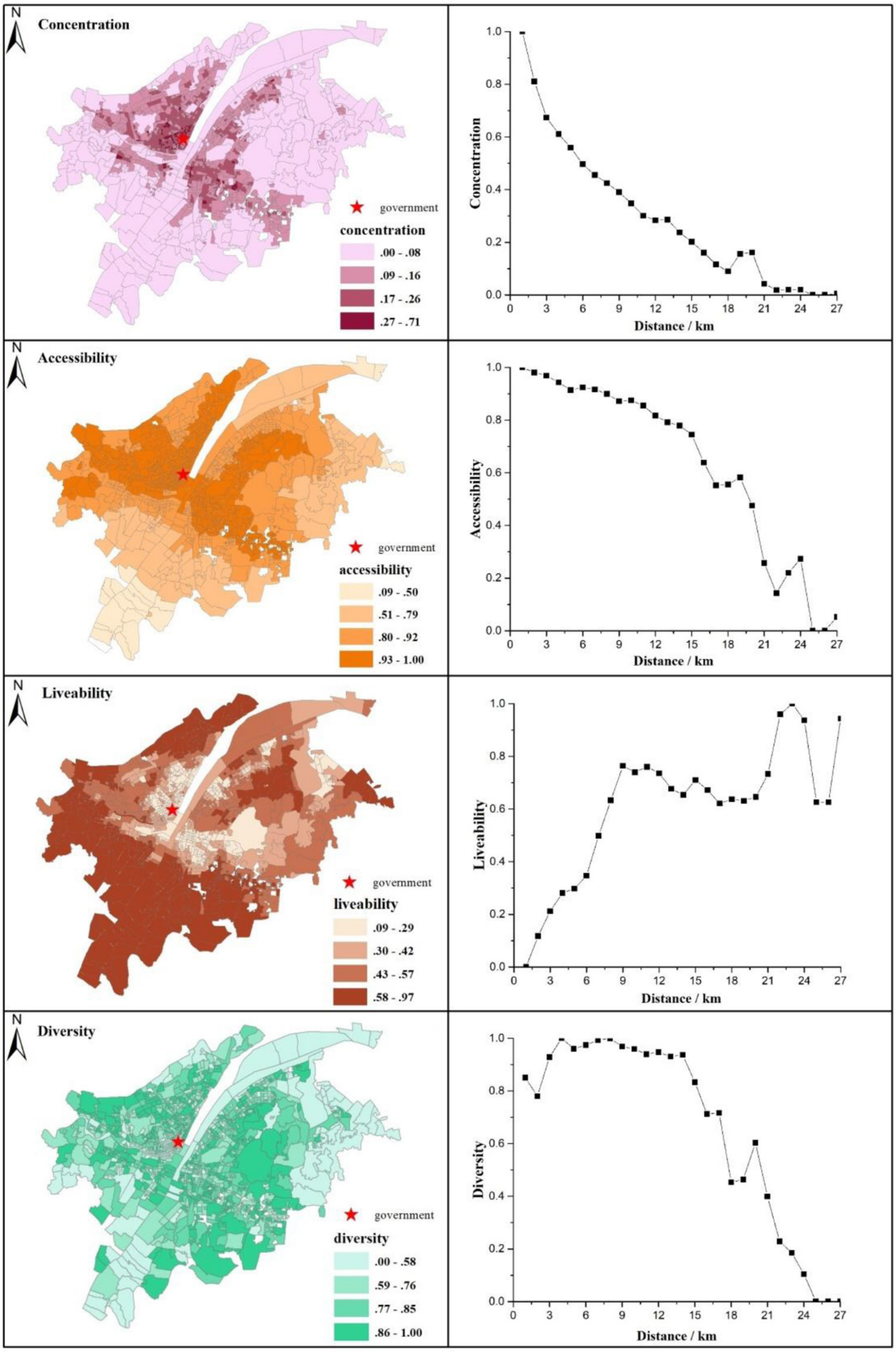


FIGURE 2 | Spatial pattern of concentration, accessibility, livability, and diversity (the left column) along with concentration, accessibility, livability, and diversity in concentric rings with the distance to the city center (the right column) in Wuhan.

radius of 1,500 m, 0.53 for two-phase betweenness Euclidean (TPBTE) in the radius of 1,000 m, and 0.29 for two-phase destination Euclidean (TPD) in the radius of 1,500 m. The highest correlation between betweenness metrics and urban vitality at the scale of 1,000 or 1,500 m demonstrates that the street design facilitating walking or cycling can encourage

TABLE 2 | The correlation between street network metrics and urban vitality.

Metric	Name (abbrev.)	Best correlation and radius	Metric	Name (abbrev.)	Best correlation and radius
Connectivity	CONN	+0.65, 1,500 m	Severance	MCF	+0.23, 2,500 m
	JNC	+0.65, 1,500 m		DIVE	−0.34, 2,500 m
Closeness	MED	+0.28, 500 m		MGLE	+0.29, 500 m
	NQPDE	+0.66, 2,500 m	Efficiency	HULLA	+0.61, 500 m
	ANGD	+0.51, 2,500 m		HULLP	+0.53, 1,000 m
Betweenness	BTE	+0.60, 1,500 m		HULLR	+0.26, 2,500 m
	TPBTE	+0.53, 1,000 m		HULLB	+0.05, 1,000 m
	TPD	+0.29, 1,500 m		HULLSI	+0.07, 1,500 m

mixed land-use and social interaction, which help to create a vibrant neighborhood.

Severance measures the opposite metrics of connectivity in network detour analysis and primarily reflects how the street network deviates from the most direct path, which proxies the navigating difficulties of pedestrians or vehicles by measuring the extent to which the local network is twisted. Diversion ratio in radius Euclidean (DIVE) is negatively correlated with urban vitality, and the best correlation coefficient is −0.34 at the spatial scale of 2,500 m. The negative correlation between street severance and urban vitality at the scale of 2,500 m represents that street network detours increase navigating difficulties of vehicles and are not conducive to vibrant commercial activities.

Efficiency measures the ease of navigation in network shape analysis considering the shape of links, the arrangement of links, and the number of connections. Convex hull area (HULLA) and convex hull perimeter (HULLP) are more associated with urban vitality than other variables, and the best correlation coefficient is 0.61 for HULLA at the spatial scale of 500 m and 0.53 for HULLP at the spatial scale of 1,000 m. The best association between efficiency metrics and urban vitality at the scale of 500 or 1,000 m denotes that the walking-friendly street shape is important for pedestrian interactions and creating a vital environment.

Spatial Stratified Heterogeneity Between the Multi-Scalar Network Metrics and Urban Vitality

Table 3 demonstrates the *q* statistics between urban vitality and spatial network metrics at multiple scales. Overall, the density metric has the largest explanatory power (∼46%) for the distribution of urban vitality at all scales. Dense street networks can increase people's contact opportunities, promoting economic and social activities. In the Chinese context, the downtown area is usually characterized by narrow streets and densely distributed networks, leading to high urban vitality. The two density metrics—LINK and LEN—have explained 46% of urban vitality at all spatial scales according to **Table 2**. Over increasing distances, the influences of both LINK and LEN on urban vitality increase, peak at the spatial scale of 1,000 m, and then monotonically decrease.

Connectivity is the second most important factor influencing the spatial heterogeneity of urban vitality. Well-connected street networks, which facilitate residents' physical activities in the neighborhood and enhance visual contact between people on the street, explain 44% of urban vitality. The mean q statistics of both connectivity in radius (CONN) and junctions in radius (JNC) reach 0.445 and 0.444, respectively. Approximately 44% of urban vitality is attributed to the number of street links and junctions. The explanatory power of these two variables on urban vitality has a slight upward trend and then decreases with the highest point at the spatial scale of 1,500 m.

Betweenness has maximum explanatory power for urban vitality of 46.7% at the spatial scale of 1,500 m. TPBTE, which reflects the quantity of visits per street link, is also non-negligible with the mean contribution of 30.1% to the pattern of urban vitality. Overall, the explanatory power of BTE demonstrates an inverted U shape over increasing spatial scales, whereas TPBTE reveals an N-shape pattern. TPD, which reflects the total flow to the destination, only explains 9.5% of urban vitality on average, with a slightly upward trend when spatial scale increases.

Among the metrics of closeness, NQPDE, which reflects network quantity and accessibility, contributes the most to the pattern of urban vitality, with the mean *q* statistic as high as 0.432. On average, ANGD explains 27.2% of urban vitality, indicating the importance of the angular analysis and cognitive difficulty during navigation. Mean Euclidean distance in radius (MED) explains the least variation in urban vitality with a mean *q* statistic of 0.109. A comparative analysis of these three variables illustrates that the conventional measure of street centrality by Euclidean distance is limited in its ability to reflect the subtleties in the network layout. The explanatory power of both NQPDE and ANGD increases as the distance interval increases.

Efficiency can explain 22% of the variation in urban vitality, and the communities with better spatial arrangements of street networks tend to be more vibrant. As a representative indicator reflecting network efficiency, HULLA can explain ∼40% of urban vitality, although this figure decreases slightly over the incremental spatial scales. The explanatory power of HULLP demonstrates an upward trend from 30.6% at the spatial scale of 500 m to 34.0% at the spatial scale of 2,500 m. Convex hull shape index (HULLSI) contributes 29.4% to the heterogeneity of urban vitality, peaking at the spatial scale of 1,000 m. Therefore, the form of the overall spatial footprint of the network shapes the navigating efficiency of pedestrians and vehicles. The convex hull with a regular and straight-line shape can diversify the built environment and enhance urban vitality.

TABLE 3 | *q* statistics between urban vitality and spatial network metrics at multiple scales.

Metrics	Variables	500 m	1,000 m	1,500 m	2,000 m	2,500 m	Trend	Mean
Connectivity	CONN	0.428	0.465	0.467	0.447	0.420		0.445
	JNC	0.427	0.465	0.469	0.444	0.415		0.444
Closeness	MED	0.148	0.110	0.098	0.074	0.117		0.109
	NQPDE	0.329	0.435	0.468	0.459	0.471		0.432
	ANGD	0.122	0.259	0.312	0.329	0.337		0.272
Betweenness	BTE	0.382	0.453	0.467	0.434	0.428		0.433
	TPBTE	0.307	0.317	0.322	0.269	0.288		0.301
	TPD	0.057	0.089	0.109	0.109	0.111		0.095
Severance	MCF	0.057	0.101	0.105	0.081	0.095		0.088
	DIVE	0.082	0.064	0.054	0.077	0.166		0.089
	MGLE	0.140	0.113	0.096	0.073	0.118		0.108
Efficiency	HULLA	0.382	0.388	0.367	0.365	0.350		0.371
	HULLP	0.306	0.329	0.325	0.335	0.340		0.327
	HULLR	0.073	0.079	0.095	0.109	0.121		0.095
	HULLB	0.022	0.034	0.036	0.041	0.027		0.032
	HULLSI	0.256	0.326	0.311	0.298	0.280		0.294

Severance, which reflects the navigating difficulties of pedestrians or vehicles, only explains 10% of the spatial heterogeneity of urban vitality. This illustrates the negative effect of network twistedness on creating a vibrant city. As spatial scale increases, the contribution of both DIVE and mean geodesic length in radius Euclidean (MGLE) demonstrates a U-shape pattern with the maximum value at the scale of 2,000 m. The explanatory power of severance metrics on urban vitality tends to increase rapidly because residents prefer simple routes when they choose to drive.

Therefore, their stratified spatial associations between various street network metrics and urban vitality are sensitive to spatial scales ranging from walking to driving distance. When people navigate street networks under different transport modes (e.g., walking, cycling, driving, etc.), the corresponding street network metrics within the walking or driving distance are different at multiple spatial scales. Thus, there is no single optimal scale for assessing urban vitality and designing street networks (50). Each scale from walking to driving distance enables different types of analysis and assessment about the vitality-led urban street design facilitating walking or driving. The spatial explicit analysis between the multi-scalar street network metrics and urban vitality yields important information for human-scale street design and land-use planning.

Policy Implications

Interest in promoting a healthy, vibrant, and interactive neighborhood is a worldwide issue for various stakeholders in urban development. Good neighborhoods tend to have ideal environments that encourage walking, bicycling, and a sense of community, making them more spirited and livable (51). Examining the street configurations using sDNA reveals that conventional street centrality indices, such as accessibility and betweenness, cannot effectively guide the design of a good street network in real estate development. Two promising variables, namely, severance and efficiency, may provide some new design elements for streets. The geometric features of street networks, such as detours, shapes, and angular curvature, are important, as they influence people's subjective cognition when driving or walking. The irregular and complicated design of streets will multiply residents' navigating difficulties and threaten their psychological safety, causing people to stay indoors and making the community lifeless. Therefore, urban planners and real estate developers should design streets that benefit the physical and emotional health of children, seniors, and indeed every resident who plays a part in creating a truly safe and healthy neighborhood.

More strategies of vitality-related urban design should be encouraged to vitalize the traditional neighborhoods in the urban center and build new neighborhoods in the suburbs. These design elements are important to inform planners to create walkable, bike-friendly streets that are connected adequately to provide more walking routes. Streets that are less twisted and have regular shapes can make people feel psychologically safe, thus encouraging outdoor activities and enhancing personal interaction. At least one local main street with a straight-line shape in the community should be designed for pedestrians to meet, make friends, and share information, thus strengthening neighborhood bonds. Real estate developers should allocate space for recreational facilities and children's playgrounds on the local main street for residents to enjoy. Intersections should have regulations on driving speeds. Considering the aforementioned aspects, urban planners, real estate developers, policy makers, and non-profit representatives should devise appropriate street design guidelines for creating a healthy neighborhood.

Urban planners can better identify vitality by considering spatial dynamics in their assessments. In our study, notably, the downtown in the Chinese city of Wuhan is more vibrant than other areas. The street design in the downtown makes residents feel safe and comfortable while walking, creating a healthy, interactive neighborhood, while the less vital neighborhoods are mainly situated in the urban suburbs. The street design for the uptown area encourages people to drive, and blocks are often longer than 2,000 ft, which is less pedestrian-friendly. Street network design is an important way to strengthen urban vitality in the suburbs. A geospatial view can help urban planners and real estate developers target areas in which street design can be improved.

DISCUSSION

Possible Mechanisms

It is interesting to explore the reason for the curious relation between geometry and urban vitality. Spatial network analysis provides extensive information about the metrics of street networks, as well as conventional accessibility and reachability. The correlation coefficients and the q statistics between various spatial network metrics and urban vitality from the multi-scalar perspective shed light on the details of their precise causal mechanisms.

Some metrics, such as density, connectivity, and betweenness calculated by spatial network analysis, are not new. These conventional street configurations provide an objective view of the location advantages of various places. Neighborhoods with densely distributed roads, high street connectivity, and many intermediary streets tend to attract more commercial and service activities, creating more employment. From a consumer's perspective, these network-based centrality metrics reflect how convenient access is to various services or facilities. From a vendor's perspective, central streets with high volumes of pedestrians or vehicles can provide larger market potential and more economic opportunities. Therefore, concentration, accessibility, livability, and diversity are closely linked with these centrality-based network measures.

Spatial network analysis provides a new view of the geometry of street networks, including detours, shape, and angular curvature, which are closely associated with people's subjective cognition. These geometric measures reflect the cognitive difficulties residents experience while walking or driving. Spatial network analysis includes a novel measure of closeness, ANGD, which calculates the distance in terms of angular changes, such as corners on links and turns at junctions. Residents who live in neighborhoods with high ANGD encounter more navigating difficulties en route to destinations, including traffic lights at crossroads and the need to make more turns. The severance and efficiency metrics are also new in spatial network analysis. When navigating a street network with high severance, the twisted streets make pedestrians or drivers feel psychologically insecure, decreasing the traffic flow and ultimately weakening urban vitality in the neighborhood. The efficiency metrics that consider the network shape directly represent the intrinsic navigability by foot. Efficient street networks are easily navigated by pedestrians, increasing residents' contact opportunities and making the local communities more active. The other possible causal mechanism of efficiency metrics on urban vitality is that high values of HULLA, HULLP, and HULLSI may indicate a long, straight pedestrian route in the local community. Such a road would be convenient in a small district and provide opportunities for people to interact with each other, ultimately making the community more lively.

The multi-scalar perspective is important for spatial network analysis to characterize the street configurations under different traveling scenarios. Defining a scale of interest is the key component of spatial network analysis. In the analysis of urban vitality, the scale tends to match different traveling scenarios from walking distances (up to 1,500 m or less) to driving distances. **Table 3** shows that the metrics of density, connectivity, betweenness, and efficiency under the walking mode have more explanatory power over urban vitality. The densely distributed, well-connected, and efficient streets provide a walking-friendly environment for residents to strengthen communications and ultimately create a lively neighborhood. However, metrics such as angular curvature and severance are more important for understanding travel by car. People tend to drive when they must travel on streets with more angular changes and twistedness. In the Chinese context, the spatial design of street networks in downtown areas facilitates people's ability to walk around, which explains the high urban vitality.

Strengths and Limitations

Prior studies have widely confirmed the associations between street centrality and land-use intensity, the location of economic activities, and social cohesion (52). However, spatial network analysis can better measure the details of network geometry, such as network shape, detours, and angular changes. These geometric details of street networks are related to the cognition difficulties experienced by pedestrians or drivers and whether they feel comfortable or safe psychologically. Thus, severance and efficiency are two promising parameters to provide a comprehensive view of the street network geometry. The other strength of spatial network analysis is the multi-scalar measure of street network characteristics, which allows the modeling of different navigating scenarios from walking to driving. The multi-scalar perspective allows urban planners and policy makers to design pedestrian streets or roads in ways that strengthen urban vitality.

One limitation of this study is its failure to consider transportation capacity and multiple transport modes like railways, subways, and highways. Although long, straight roads probably bring more traffic flows, the incorporation of traffic variables into spatial network analysis tends to enhance the relationship between street configurations and urban vitality. The other limitation is the lack of clarity of the causal mechanisms linking each street network metric to urban vitality. Although this study has provided insightful views about the possible causal mechanisms, a further detailed investigation is required to create vibrant neighborhoods by designing walkable and drivable streets.

CONCLUSIONS

This study has explored the influence of spatial network layouts on urban vitality using geographic big data for Wuhan, an inland city in China. Concentration, accessibility, livability, and diversity are four major components that characterize urban vitality in Wuhan. The new technique of sDNA was employed to measure street configurations, including density, connectivity, closeness, betweenness, severance, and efficiency, from a multi-scalar perspective. Furthermore, the stratified spatial heterogeneity between street network metrics and urban vitality was investigated using the Geodetector tool. The following conclusions can be drawn.

First, the areas with the highest levels of urban vitality are clustered in the downtown area, whereas the uptown area is characterized by low urban vitality. Concentration, accessibility, and livability demonstrate a declining trend in concentric rings, whereas livability reveals a fluctuated upward trend. Second, 16 variables representing connectivity, closeness, betweenness, severance, and efficiency are computed. The correlation between these network characteristics and urban vitality is sensitive to different spatial scales. Third, the influence of street network layouts on urban vitality varies at multiple scales. Overall, connectivity has the largest explanatory power for urban vitality, amounting to over 40%, whereas betweenness and closeness have similar explanatory power of ∼28%. Efficiency and severance contribute 22 and 10% to the spatial heterogeneity of urban vitality, respectively.

These conclusions shed light on the mechanisms between street configurations and urban vitality from the multi-scalar perspective. In the future, more strategies of vitality-based urban design should be encouraged to revitalize traditional downtown neighborhoods and build new neighborhoods in the uptown area. Multiple stakeholders, such as urban planners, real estate developers, policy makers, and non-profit representatives, should collaborate to devise effective street design guidelines for creating healthy, vibrant neighborhoods.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

CF contributed to paper writing and spatial data analysis. SH contributed to data analysis and paper revision. LW contributed to result analysis and map designing. All authors contributed to the article and approved the submitted version.

FUNDING

This research was funded by the National Natural Science Foundation of China (41601162), General Financial Grant from the China Postdoctoral Science Foundation (2016M601112), and Natural Science Foundation of Jiangsu Province, China (BK20200109).

REFERENCES

1. Cobbinah PB, Erdiaw-Kwasie MO, Amoateng P. Rethinking sustainable development within the framework of poverty and urbanisation in developing countries. *Environ Dev*. (2015) 13:18–32. doi: 10.1016/j.envdev.2014.11.001
2. Schilling J, Logan J. Greening the rust belt: a green infrastructure model for right sizing America's shrinking cities. *J Am Plann Assoc*. (2008) 74:451–66. doi: 10.1080/01944360802354956
3. Großmann K, Bontje M, Haase A, Mykhnenko V. Shrinking cities: notes for the further research agenda. *Cities*. (2013) 35:221–5. doi: 10.1016/j.cities.2013.07.007
4. Long Y, Wu K. Shrinking cities in a rapidly urbanizing China. *Environ Plann*. (2016) 48:220–2. doi: 10.1177/0308518X15621631
5. He SY, Lee J, Zhou T, Wu D. Shrinking cities and resource-based economy: the economic restructuring in China's mining cities. *Cities*. (2017) 60:75–83. doi: 10.1016/j.cities.2016.07.009
6. Freire M, Stren R. *The Challenge of Urban Government: Policies and Practices*. The World Bank (2001). doi: 10.1596/0-8213-4738-1
7. Barton H. Land use planning and health and well-being. *Land Use Policy*. (2009) 26:S115–23. doi: 10.1016/j.landusepol.2009.09.008
8. Lopes MN, Camanho AS. Public green space use and consequences on urban vitality: an assessment of European cities. *Soc Indicators Res*. (2013) 113:751–67. doi: 10.1007/s11205-012-0106-9
9. Lynch K. *Good City Form*. MIT press (1984).
10. Yue W, Chen Y, Zhang Q, Liu Y. Spatial explicit assessment of urban vitality using multi-source data: a case of Shanghai, China. *Sustainability*. (2019) 11:638. doi: 10.3390/su11030638
11. Pugalis L. The culture and economics of urban public space design: public and professional perceptions. *Urban Design Int*. (2009) 14:215–30. doi: 10.1057/udi.2009.23
12. Montgomery J. Making a city: urbanity, vitality and urban design. *J Urban Design*. (1998) 3:93–116. doi: 10.1080/13574809808724418
13. Anderson S, Allen J, Browne M. Urban logistics—-how can it meet policy makers' sustainability objectives? *J Transport Geogr*. (2005) 13:71–81. doi: 10.1016/j.jtrangeo.2004.11.002
14. Jacobs J. *The Death and Life of Great American Cities*. Vintage (1961).
15. Gehl J. *Life Between Buildings: Using Public Space*. Copenhagen: Danish Architectural Press (1971).

16. Delclòs-Alió X, Miralles-Guasch C. Looking at Barcelona through Jane Jacobs's eyes: mapping the basic conditions for urban vitality in a Mediterranean conurbation. *Land Use Policy*. (2018) 75:505–17. doi: 10.1016/j.landusepol.2018.04.026
17. Xia CA, Yeh GO, Zhang A. Analyzing spatial relationships between urban land use intensity and urban vitality at street block level: a case study of five Chinese megacities. *Landscape Urban Plann.* (2020) 193:103669. doi: 10.1016/j.landurbplan.2019.103669
18. Kim YL. Data-driven approach to characterize urban vitality: how spatiotemporal context dynamically defines Seoul's nighttime. *Int J Geogr Informat Sci.* (2019) 34:1–22. doi: 10.1080/13658816.2019.1694680
19. Lu S, Huang Y, Shi C, Yang X. Exploring the associations between Urban Form and neighborhood vibrancy: a case study of Chengdu, China. *ISPRS Int J Geo-Informat.* (2019) 8:165. doi: 10.3390/ijgi8040165
20. Sung H-G, Go D-H, Choi GC. Evidence of Jacobs's street life in the great Seoul city: identifying the association of physical environment with walking activity on streets. *Cities*. (2013) 35:164–73. doi: 10.1016/j.cities.2013.07.010
21. Long Y, Huang C. Does block size matter? The impact of urban design on economic vitality for Chinese cities. *Environ Plann.* (2019) 46:406–22. doi: 10.1177/2399808317715640
22. Mitra R, Buliung RN. Built environment correlates of active school transportation: neighborhood and the modifiable areal unit problem. *J Transport Geogr.* (2012) 20:51–61. doi: 10.1016/j.jtrangeo.2011.07.009
23. Jalaladdini S, Oktay D. Urban public spaces and vitality: a socio-spatial analysis in the streets of Cypriot towns. *Proc Soc Behav Sci.* (2012) 35:664–74. doi: 10.1016/j.sbspro.2012.02.135
24. Ye Y, Van Nes A. Measuring urban maturation processes in Dutch and Chinese new towns: combining street network configuration with building density and degree of land use diversification through GIS. *J Space Syntax*. (2013) 4:18–37.
25. Kang C-D. Measuring the effects of street network configurations on walking in Seoul, Korea. *Cities*. (2017) 71:30–40. doi: 10.1016/j.cities.2017.07.005
26. Oliver LN, Schuurman N, Hall WA. Comparing circular and network buffers to examine the influence of land use on walking for leisure and errands. *Int J Health Geogr*. (2007) 6:41. doi: 10.1186/1476-072X-6-41
27. Hajrasouliha A, Yin LJUS. The impact of street network connectivity on pedestrian. *Urban Stud.* (2015) 52:2483–97. doi: 10.1177/0042098014544763
28. He S, Yu S, Wei P, Fang C. A spatial design network analysis of street networks and the locations of leisure entertainment activities: a case study of Wuhan, China. *Sustain Cities Soc.* (2019) 44:880–7. doi: 10.1016/j.scs.2018.11.007
29. Porta S, Crucitti P, Latora V. The network analysis of urban streets: a primal approach. *Environ Plann.* (2006) 33:705–25. doi: 10.1068/b32045
30. Lyu G, Bertolini L, Pfeffer K. Developing a TOD typology for Beijing metro station areas. *J Transport Geogr.* (2016) 55:40–50. doi: 10.1016/j.jtrangeo.2016.07.002
31. Wang F, Antipova A, Porta S. Street centrality and land use intensity in Baton Rouge, Louisiana. *J Transport Geogr.* (2011) 19:285–93. doi: 10.1016/j.jtrangeo.2010.01.004
32. Porta S, Latora V, Wang F, Rueda S, Strano E, Scellato S, et al. Street centrality and the location of economic activities in Barcelona. *Urban Stud.* (2012) 49:1471–88. doi: 10.1177/0042098011422570
33. Koohsari MJ, Oka K, Owen N, Sugiyama T. Natural movement: a space syntax theory linking urban form and function with walking for transport. *Health Place*. (2019) 58:102072. doi: 10.1016/j.healthplace.2019.01.002
34. Cooper CH, Fone DL, Chiaradia JA. Measuring the impact of spatial network layout on community social cohesion: a cross-sectional study. *Int J Health Geogr.* (2014) 13:11. doi: 10.1186/1476-072X-13-11
35. Yang DF, Yin CZ, Long Y. Urbanization and sustainability in China: an analysis based on the urbanization Kuznets-curve. *Plann Theory*. (2013) 12:391–405. doi: 10.1177/1473095213485558
36. Chen M, Liu W, Lu D, Chen H, Ye C. Progress of China's new-type urbanization construction since 2014: a preliminary assessment. *Cities*. (2018) 78:180–93. doi: 10.1016/j.cities.2018.02.012
37. Chen T, Hui E, Lang CM, Tao WL. People, recreational facility and physical activity: new-type urbanization planning for the healthy communities in China. *Habitat Int.* (2016) 58:12–22. doi: 10.1016/j.habitatint.2016.09.001
38. He Q, He W, Song Y, Wu J, Yin C, Mou Y. The impact of urban growth patterns on urban vitality in newly built-up areas based on an association rules analysis using geographical 'big data'. *Land Use Policy*. (2018) 78:726–38. doi: 10.1016/j.landusepol.2018.07.020
39. Zeng C, Song Y, He Q, Shen F. Spatially explicit assessment on urban vitality: case studies in Chicago and Wuhan. *Sustain Cities Soc.* (2018) 40:296–306. doi: 10.1016/j.scs.2018.04.021
40. Wei YD, Xiao WY, Medina R, Tian G. *Effects of Neighborhood Environment, Safety, and Urban Amenities on Origins and Destinations of Walking Behavior*. Urban Geography (2019) 1–21. doi: 10.1080/02723638.2019.1699731
41. Tewahade S, Li K, Goldstein RB, Haynie D, Iannotti RJ, Simons-Morton B. Association between the built environment and active transportation among U.S. adolescents. *J Transport Health*. (2019) 15:100629. doi: 10.1016/j.jth.2019.100629
42. Cooper CHV. Spatial localization of closeness and betweenness measures: a self-contradictory but useful form of network analysis. *Int J Geogr Inform Sci.* (2015) 29:1293–309. doi: 10.1080/13658816.2015.1018834
43. Kang C-D. The effects of spatial accessibility and centrality to land use on walking in Seoul, Korea. *Cities*. (2015) 46:94–103. doi: 10.1016/j.cities.2015.05.006
44. Porta S, Strano E, Iacoviello V, Messora R, Latora V, Cardillo A, et al. Street centrality and densities of retail and services in Bologna, Italy. *Environ Plann.* (2009) 36:450–65. doi: 10.1068/b34098
45. Barthélemy M, Flammini A. Co-evolution of density and topology in a simple model of city formation. *Netw Spatial Econom.* (2009) 9:401–25. doi: 10.1007/s11067-008-9068-5
46. Ye T, Zhao N, Yang X, Ouyang Z, Liu X, Chen Q, et al. Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. *Sci Total Environ.* (2019) 658:936–46. doi: 10.1016/j.scitotenv.2018.12.276
47. King K. Jane Jacobs and 'the need for aged buildings': neighbourhood historical development pace and community social relations. *Urban Stud.* (2013) 50:2407–24. doi: 10.1177/0042098013477698
48. Jiang B. Geospatial analysis requires a different way of thinking: the problem of spatial heterogeneity. *Geo J.* (2015) 80:1–13. doi: 10.1007/s10708-014-9537-y
49. Wang JF, Li XH, Christakos G, Liao YL, Zhang T, Gu X, et al. Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun Region, China. *Int J Geogr Inform Sci.* (2010) 24:107–27. doi: 10.1080/13658810802443457
50. Verburg PH, Veldkamp AJL. Projecting land use transitions at forest fringes in the Philippines at two spatial scales. *Landscape Ecol.* (2004) 19:77–98. doi: 10.1023/B:LAND.0000018370.57457.58
51. Burden D, Wallwork M, Sides K, Trias R, Rue H. *Street Design Guidelines for Healthy Neighborhoods, Center for Livable Communities*. Sacramento: Calif (1999).
52. Wang S, Yu D, Kwan M-P, Zheng L, Miao H, Li Y. The impacts of road network density on motor vehicle travel: an empirical study of Chinese cities based on network theory. *Transportation Res Part A.* (2020) 132:144–56. doi: 10.1016/j.tra.2019.11.012

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.2021.680054

The Optimal Machine Learning-Based Missing Data Imputation for the Cox Proportional Hazard Model

Chao-Yu Guo [1,2], Ying-Chen Yang [1,2] and Yi-Hau Chen [3]*

[1] Institute of Public Health, School of Medicine, National Yang-Ming University, Taipei, Taiwan, [2] Institute of Public Health, School of Medicine, National Yang Ming Chiao Tung University, Hsinchu, Taiwan, [3] Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

An adequate imputation of missing data would significantly preserve the statistical power and avoid erroneous conclusions. In the era of big data, machine learning is a great tool to infer the missing values. The root means square error (RMSE) and the proportion of falsely classified entries (PFC) are two standard statistics to evaluate imputation accuracy. However, the Cox proportional hazards model using various types requires deliberate study, and the validity under different missing mechanisms is unknown. In this research, we propose supervised and unsupervised imputations and examine four machine learning-based imputation strategies. We conducted a simulation study under various scenarios with several parameters, such as sample size, missing rate, and different missing mechanisms. The results revealed the type-I errors according to different imputation techniques in the survival data. The simulation results show that the non-parametric "missForest" based on the unsupervised imputation is the only robust method without inflated type-I errors under all missing mechanisms. In contrast, other methods are not valid to test when the missing pattern is informative. Statistical analysis, which is improperly conducted, with missing data may lead to erroneous conclusions. This research provides a clear guideline for a valid survival analysis using the Cox proportional hazard model with machine learning-based imputations.

Keywords: machine learning, k-nearest neighbors imputation, random forest imputation, survival data simulation, cox proportional hazard model

OPEN ACCESS

Edited by:
Yi Guo,
University of Florida, United States

Reviewed by:
Hui Hu,
University of Florida, United States
Yan Zhang,
Bristol Myers Squibb, United States

****Correspondence:***
Chao-Yu Guo
cyguo@ym.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 March 2021*
Accepted: *04 June 2021*
Published: *05 July 2021*

Citation:
Guo C-Y, Yang Y-C and Chen Y-H (2021) The Optimal Machine Learning-Based Missing Data Imputation for the Cox Proportional Hazard Model. Front. Public Health 9:680054. doi: 10.3389/fpubh.2021.680054

BACKGROUND

Before statistical analysis, data management plays a crucial role and missing data occur frequently. If there are too many missing values, excluding the missing data from the analysis is not ideal since the loss of information is substantial. In addition to the reduced power, missing data may introduce potential biases or an unsolvable issue in statistical modeling. There are three significant missingness mechanisms (1). They are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Under MCAR, one could simply exclude the missing data from the analysis. However, it may introduce bias if the missing pattern is MAR or MNAR.

To preserve statistical power, one should conduct missing data imputation techniques before the analysis. The single imputation is a simple way that substitutes the mean, mode, or median for the missing data. Unfortunately, this intuitive concept may not capture the variability in the study sample and underestimate the variance, which reduces the correlation between variables or introduces a bias in the inference of the population distribution (2).

The U.S. Census Bureau developed the hot-deck imputation to investigate the missing value of current population income (3), a non-parametric imputation based on Euclidean distance (4). A new way of finding the donor is the random hot-deck, cold-deck, or sequential hot-deck imputation (5). The imputation does not require strong assumptions about the distribution, and it is applied to different types of variables. However, the primary issue is the assumption of MCAR.

The multiple imputations perform better than the simple imputation, but it still requires the assumption of MCAR or MAR (6) based on the multivariate imputation by chained equations or the Markov chain Monte Carlo.

The k-nearest neighbors (KNN) is a simple discriminatory analysis (7). Algorithms of the KNN were studied, and the minimum probability of error was pointed out (8). The KNN also implemented the direct gradient analysis (9). The concept of the training and testing sets using the KNN was further proposed (10). Later, the iterative KNN imputation based on the gray relational analysis was carried out (11). Regarding the truncated data, a previous work developed the KNN-truncated imputation to deal with the chemical compound (12).

The randomness of a decision tree could enhance predictive accuracy (13), and a random forest is a powerful tool for classification problems (14). The missing data imputation is the "rfImpute" function of the "randomForest" package. We denoted it as RF_{prx} in the simulation study. It is based on the proximity matrix to update the imputation of the missing values. For continuous predictors, the imputed value is the weighted average of the non-missing observations, where the weights are the proximities.

The "missForest" imputation is non-parametric missing value imputation using the random forest (15). We denoted it as RF_{mf} in the simulations. The fast unified random forests for survival, regression, and classification (RF-SRC) solved the problem when estimating the missing data with out-of-bag errors (16). This method not only applies to classification problems and the regression model but also fits the survival analysis. The random forest on-the-fly is the missing data imputation of RF-SRC. We denoted it as RF_{otf} in the simulation study. Despite the promising development of missing data imputation, none of the strategies further examined the validity of imputed data using the Cox proportional hazard model. In this research, four machine learning-based imputation strategies were compared, including the KNN, RF_{prx}, RF_{mf}, and RF_{otf}.

In this research, we define supervised and unsupervised missing data imputation as the following. The supervised imputation techniques refer to methods that included the outcome variable as predictors to infer the missing data. In contrast, the unsupervised missing imputation is the one that excludes the outcome of interest in the process. The impact of various missing mechanisms, including MCAR, MAR, and MNAR, would be carefully examined under numerous scenarios. In addition to the conventional approach that evaluated the root mean square error (RMSE) or the proportion of falsely classified entries (PFC) of imputed values, we further analyzed the whole imputed data by the Cox proportional hazard model. Type-I errors of the Cox model using imputed data reveal how the imputation technique performs in the survival analysis. If the Type-I error is over 5% of the nominal level, then, the method is invalid.

METHODS

We want to assess how machine learning-based missing data imputation techniques perform in the survival analysis. The Cox proportional hazard model would incorporate the imputed data, and the results under various scenarios demonstrate overall type-I error. Therefore, the first step is to simulate the survival data under the null hypothesis, including the time to the event, censoring status, six continuous, and four categorical predictive variables. It is noted that the 10 predictors denoted as $x_1, x_2, \cdots, x_{10}$ are uncorrelated, and they are not associated (independent) with the two outcome variables. One of the outcome variables, t, denotes the time to the event, and e denotes the censoring status. Note that if "e $=$ 1," then the subject has an event, and it also means that the individual is not censored. Thus, "e $=$ 0" identifies the censored subject. Each of the four categorical predictors $(x_1, x_2, x_3, x_{4)}$ follows a binomial distribution with $p = 0.5$. Each of the six continuous predictors $(x_5, x_6, \cdots, x_{10)}$ follow a normal distribution with the mean of zero and SD of one. The censoring status "e" follows a uniform distribution between zero and one, representing the random censoring. The time to the event "t" follows the exponential distribution with $\lambda = 0.5$. We employed four categorical and six continuous unrelated variables to assess the validity of various methods. The reason is that if under such a simplistic scenario, a strategy could not yield a valid estimate or result, it is unrealistic that the method would be valid under a more complicated structure.

The second step is to assign missing values for the two predictors. One of the predictors (x_1) is categorical, and the other one (x_5) is continuous. Each scenario simulated 1,000 repetitions. Parameters included the sample size (100, 250, 500, and 1,000), the overall missing rate (10, 20, and 30%), and missing mechanisms (MCAR, MAR, and MNAR). Hence, we carefully examined a total of 36 scenarios for the 4 imputation strategies. It is noted that within each overall missing rate, the weights of missingness are 0.2 (x_1 is missing), 0.4 (x_5 is missing), and 0.4 (both x_1 and x_1 are missing), respectively.

Two statistics evaluate the four machine learning-based imputation methods, including the RMSE for the continuous variable and the PFC for the categorical variable. This research

Abbreviations: RMSE, root means square error; PFC, proportion of falsely classified entries; MCAR, missing completely at random; MAR, missing at random; MNAR, missing not at random; KNN, k-nearest neighbors; RF-SRC, Random Forests for Survival, Regression, and Classification; SVM, support vector machine; XGBoost, Extreme Gradient Boosting Machine; ANN, artificial neural network.

examines the performance of imputed data in survival analysis based on the overall type-I error of the Cox model.

Root mean square error is a measure used to measure the difference between the imputed value and the actual value for continuous outcomes. A smaller RMSE indicates a smaller prediction error. The equation is $RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(y_i - \hat{y}_i\right)^2}$, where m represents the number of missing values, y_i is the actual value, and $\hat{y}_i$ is the imputed value.

The PFC is used to determine the imputation situation of category variables. The PFC equation is given by, $PFC = \frac{\sum count(y_i \neq \hat{y}_i)}{\sum count(y_i)}$. The denominator is the number of missing values of the categorical variable, and the numerator is the number of imputed values that are not identical to the actual values. The PFC ranges from 0 to 1, and the smaller the value means better imputation.

In addition to the RMSE and PFC, this research further examines the type I error of the Cox model using the imputed data. The likelihood ratio test derives the type-I error. In this way, the type-I error could reveal the impact of imputation on the correlation structure between the predictors and the two survival outcome variables. Finally, we recorded the computation speed that tells the practicality of the different strategies. In this research, we selected machine learning-based imputation strategies that may or may not be suitable for the survival data. In addition, we considered models that any researcher could implement effortlessly. Thus, the KNN and random forest were selected.

The programming language used in this study is R language, version 3.6.1 [(17). R: A language and environment for statistical computing. R Foundation for Statistical Computing]. The **Supplementary Materials** of the R code (user_utility.r and main.r) listed packages used to simulate the study samples, missing mechanisms, and the imputation methods. The "VIM" package implemented the KNN. The three packages "randomForest," "randomForestSRC," and "missForest" are the random forest-based imputation methods. We clarified some notations as to the following: RF_{prxt} included time to the event as the continuous outcome to generate the proximity matrix. RF_{prxe} treated the censoring status as the categorical outcome and calculated the proximity matrix. RF_{mf} excluded both time to the event and the censoring status in the imputation procedure. RF_{mfy} included both time to the event and the censoring status as two more predictors when inferring the missing values in the dataset. RF_{otf} is designed for survival analysis; thus, it included both time to the event and the censoring status when inferring the missing values. In summary, RF_{mf} is an unsupervised imputation, RF_{prxt} and RF_{prxe} are partially supervised imputation methods, and KNN, RF_{mfy}, and RF_{otf} are the three supervised imputation techniques.

RESULTS

The simulations were conducted under the null hypothesis that 10 predictors and the survival outcome are independent. However, the missing mechanism, MNAR, altered the correlation structure that introduced the dependence between the complete data and imputed values. When the model failed to adjust for the condition, the independent variables and the outcome are correlated under MAR. Therefore, this study has 36 scenarios, and each presents a comparison between four methods. **Figure 1** displays the distribution of the PFC using 500 samples. **Supplementary Materials** displayed results based on different sample sizes and missing rates that yielded a similar pattern. The RF_{otf} has the best performance in the absence of MCAR and MAR. The RF_{prxe} has the best performance in the absence of MNAR, but the accuracy is ∼0.5, which means that the predictive accuracy is not satisfying.

Figure 2 shows that the RMSE evaluates the imputation accuracy of a continuous variable and the distributions of the RMSE using 500 samples. The results with different sample sizes and missing rates yielded a similar pattern (refer to **Supplementary Materials**). KNN performs the best in each scenario, but the differences between the KNN and other random forest-based imputation methods are not discernable. The size of RMSE is approximately one SD. When the missing rate is higher, the gap of RMSE among the four methods will be smaller. The RMSE decreased from 0.08 to 0.01. The RF_{otf} is similar to the KNN. The higher the missing rate is, the higher the PFC and RMSE, which means that the higher missing rate decreases the imputation accuracy.

The KNN consistently performs better in RMSE, but the superiority is minor. It is not easy to identify the most prominent method in **Figure 2**. Regarding the PFC, the best performer is the RF_{otf}, because the two outcome variables, "time to the event, t" and "censoring status, e," are incorporated in the random survival forest. **Table 1** summarizes the best performer of the RMSE and PFC under different scenarios.

Type-I error of the Cox proportional hazard model under different situations further evaluated the overall performance of each imputation strategy (**Table 2**). This step is crucial since the comparisons between the PFC and RMSE after imputation could not warrant a valid Cox regression analysis. There are some scenarios where the results of RF_{otf} and RF_{mf} are very close, but the RF_{otf} is consistently larger than the RF_{mf}. In conclusion, the overall performance of the RF_{mf} method is the best, a non-parametric and unsupervised imputation method that excludes the two survival outcome variables (t and e). It is noted that the RF_{prxe} and RF_{prxt} have much inflated type-I error, since this type of imputation considers only one dependent variable (time to the event or censoring status) when constructing the proximity matrix. However, the simulation study was based on survival data with two outcome variables. Therefore, including one of the two survival outcome variables will result in an inflated type-I error. We highly recommended avoiding the "rfImpute" function in survival data. The KNN imputation also demonstrated inflated type-I errors and should not be used for survival analysis.

The RF_{otf} includes both times to the event and censoring status in the random survival forest to impute missing values under MCAR or assumption of MAR. Thus, the RF_{otf} is valid for survival data, and the type-I error behaves well under MCAR and MAR. However, when the missing pattern is MNAR, the RF_{otf} showed an inflated type-I error.

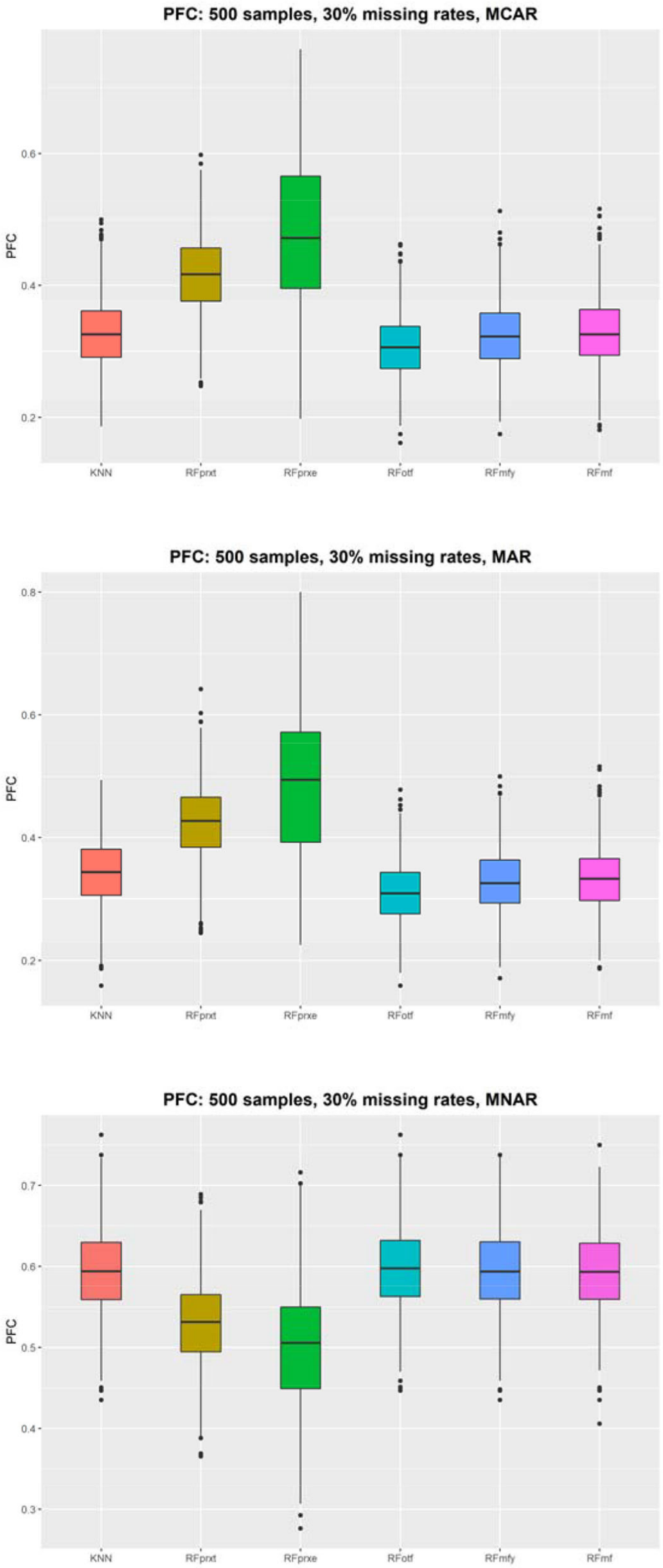
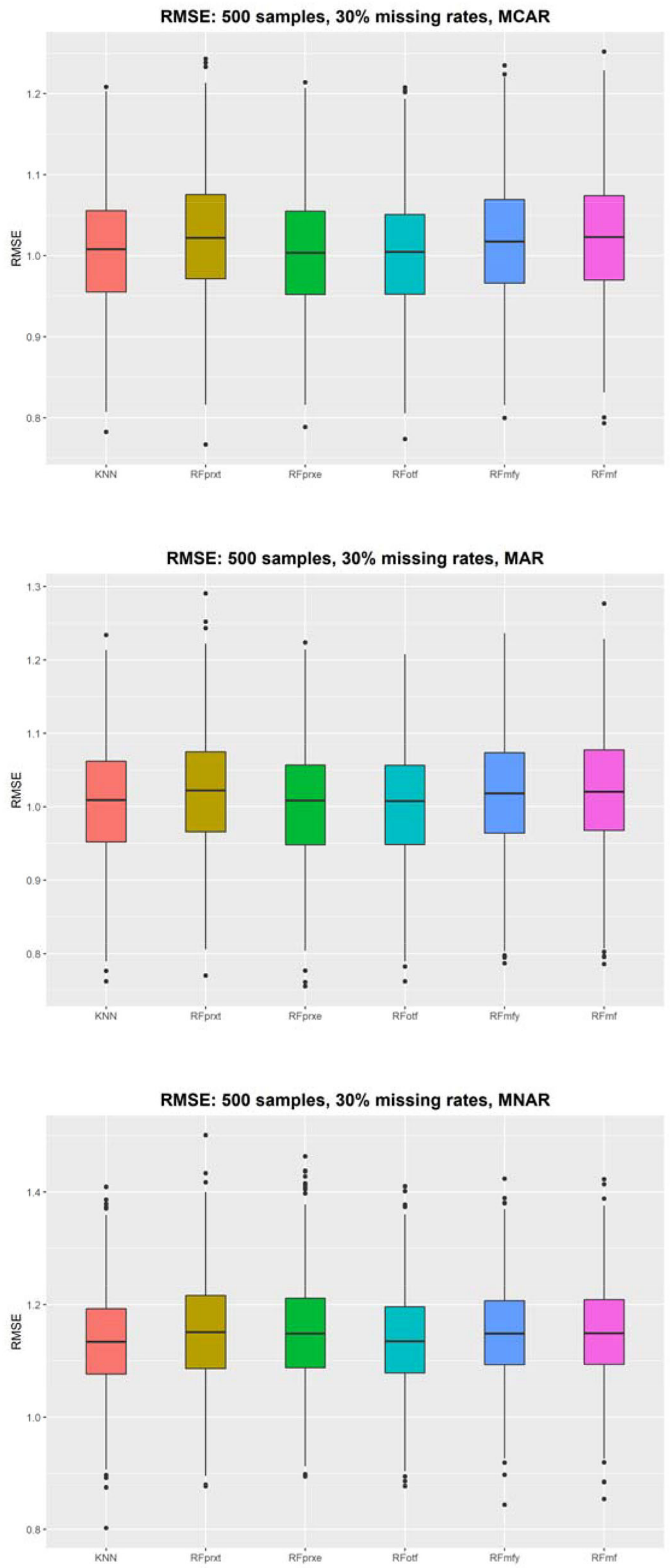


FIGURE 1 | The proportion of falsely classified (PFC) using 500 subjects.


RMSE: 500 samples, 30% missing rates, MCAR

RMSE: 500 samples, 30% missing rates, MAR

RMSE: 500 samples, 30% missing rates, MNAR

FIGURE 2 | The root means square error (RMSE) using 500 subjects.

TABLE 1 | The best performer for proportion of falsely classified (PFC) and root means square error (RMSE).

n	Missing rate	Missing pattern	PFC	RMSE
100	0.1	MCAR	RF$_{otf}$ (0.3062)	KNN (0.8781)
100	0.1	MAR	RF$_{otf}$ (0.3147)	KNN (0.8744)
100	0.1	MNAR	RF$_{prxe}$ (0.4983)	KNN (1.0584)
100	0.2	MCAR	RF$_{otf}$ (0.3071)	KNN (0.9522)
100	0.2	MAR	RF$_{otf}$ (0.3184)	KNN (0.9519)
100	0.2	MNAR	RF$_{prxe}$ (0.4862)	KNN (1.1101)
100	0.3	MCAR	RF$_{otf}$ (0.3049)	KNN (0.9871)
100	0.3	MAR	RF$_{otf}$ (0.3058)	KNN (0.9942)
100	0.3	MNAR	RF$_{prxe}$ (0.4998)	KNN (1.0973)
250	0.1	MCAR	RF$_{otf}$ (0.396)	KNN (0.9603)
250	0.1	MAR	RF$_{otf}$ (0.3069)	KNN (0.9669)
250	0.1	MNAR	RF$_{prxe}$ (0.4987)	KNN (1.1751)
250	0.2	MCAR	RF$_{otf}$ (0.3022)	KNN (0.9977)
250	0.2	MAR	RF$_{otf}$ (0.3079)	KNN (0.9938)
250	0.2	MNAR	RF$_{prxe}$ (0.4999)	KNN (1.1489)
250	0.3	MCAR	RF$_{otf}$ (0.3107)	KNN (1.0083)
250	0.3	MAR	RF$_{otf}$ (0.3143)	KNN (1.0099)
250	0.3	MNAR	RF$_{prxe}$ (0.4971)	KNN (1.1278)
500	0.1	MCAR	RF$_{otf}$ (0.3114)	KNN (0.98)
500	0.1	MAR	RF$_{otf}$ (0.3057)	KNN (0.9852)
500	0.1	MNAR	RF$_{prxe}$ (0.5072)	KNN (1.2091)
500	0.2	MCAR	RF$_{otf}$ (0.3069)	KNN (1.0045)
500	0.2	MAR	RF$_{otf}$ (0.307)	KNN (1.0044)
500	0.2	MNAR	RF$_{prxe}$ (0.5046)	KNN (1.1735)
500	0.3	MCAR	RF$_{otf}$ (0.3073)	KNN (1.01)
500	0.3	MAR	RF$_{otf}$ (0.3087)	KNN (1.0098)
500	0.3	MNAR	RF$_{prxe}$ (0.502)	KNN (1.1367)
1,000	0.1	MCAR	RF$_{otf}$ (0.3093)	KNN (0.9958)
1,000	0.1	MAR	RF$_{otf}$ (0.3067)	KNN (0.9963)
1,000	0.1	MNAR	RF$_{prxe}$ (0.5272)	KNN (1.2208)
1,000	0.2	MCAR	RF$_{otf}$ (0.3074)	KNN (1.0034)
1,000	0.2	MAR	RF$_{otf}$ (0.3102)	KNN (1.0056)
1,000	0.2	MNAR	RF$_{prxe}$ (0.5274)	KNN (1.1766)
1,000	0.3	MCAR	RF$_{otf}$ (0.3089)	KNN (1.0081)
1,000	0.3	MAR	RF$_{otf}$ (0.31)	KNN (1.0102)
1,000	0.3	MNAR	RF$_{prxe}$ (0.5142)	KNN (1.1385)

The type-I error of the RF_{mf} is lower than RF_{mfy}, which also included time to the event and censoring status as the predictors. This phenomenon is probably due to the missing mechanism of the MNAR, and the conditional missingness introduced correlation between the observed predictors and the two survival outcome variables, time to the event and censoring.

For a small sample study, estimators, in general, have a large variance. As a result, the missing data imputed by RF_{mf} also showed an inflated type-I error in the Cox model. When the sample size increases, the type-I error decreases and approaches the significance level of 0.05.

Finally, the run time is also studied. In each scenario, the fastest method is the KNN, followed by RF_{otf}, and the rest of the methods are similar. When the sample size is 100, the simulation time of each method is within 0.5 s, where the KNN and RF_{otf} only spend 0.1 s. When the missing rate is higher and the missing mechanism is MNAR, the run time of two methods is almost identical.

When the sample size is 250, the run time of KNN and RF_{otf} is <0.5 s, but the other methods take 1–2 s. If the sample size is 500, the KNN only requires 0.5 s, RF_{otf} takes 1.5 s, and the rest methods take about 5–6 s. When the number of samples is 1,000, the KNN takes about 1 s, RF_{otf} takes about 5 s, and the other methods take about 15–23 s. The greater the sample size, the more significant difference in the run time between various methods.

In summary, according to the type-I error of the Cox model, the RF_{mf} strategy that excludes the two survival outcome

TABLE 2 | The type-I error of the Cox model.

Sample	Missing rate	Missing mechanism	Complete data	KNN	RF_{prxt}	RF_{prxe}	RF_{otf}	RF_{mfy}	RF_{mf}
100	0.1	MCAR	0.079	0.088	0.085	0.092	0.082	0.088	0.08
100	0.1	MAR	0.084	0.095	0.094	0.098	0.09	0.094	0.085
100	0.1	MNAR	0.087	0.093	0.095	0.093	0.09	0.091	0.089
100	0.2	MCAR	0.091	0.092	0.102	0.113	0.081	0.095	0.079
100	0.2	MAR	0.076	0.083	0.089	0.105	0.073	0.085	0.073
100	0.2	MNAR	0.075	0.08	0.09	0.106	0.075	0.082	0.074
100	0.3	MCAR	0.085	0.101	0.12	0.134	0.094	0.114	0.086
100	0.3	MAR	0.072	0.092	0.114	0.13	0.087	0.107	0.078
100	0.3	MNAR	0.085	0.105	0.114	0.136	0.092	0.104	0.086
250	0.1	MCAR	0.055	0.057	0.063	0.069	0.051	0.059	0.051
250	0.1	MAR	0.06	0.066	0.079	0.089	0.064	0.072	0.064
250	0.1	MNAR	0.054	0.059	0.072	0.083	0.054	0.058	0.053
250	0.2	MCAR	0.052	0.065	0.082	0.095	0.055	0.068	0.055
250	0.2	MAR	0.062	0.078	0.093	0.117	0.069	0.087	0.067
250	0.2	MNAR	0.054	0.072	0.087	0.115	0.059	0.068	0.056
250	0.3	MCAR	0.069	0.094	0.125	0.169	0.075	0.097	0.07
250	0.3	MAR	0.059	0.09	0.126	0.153	0.073	0.09	0.064
250	0.3	MNAR	0.07	0.083	0.126	0.165	0.069	0.089	0.059
500	0.1	MCAR	0.061	0.057	0.073	0.082	0.058	0.062	0.057
500	0.1	MAR	0.051	0.059	0.062	0.077	0.055	0.061	0.055
500	0.1	MNAR	0.068	0.066	0.08	0.096	0.069	0.071	0.066
500	0.2	MCAR	0.05	0.065	0.105	0.141	0.055	0.062	0.053
500	0.2	MAR	0.056	0.069	0.113	0.141	0.055	0.067	0.055
500	0.2	MNAR	0.063	0.072	0.102	0.147	0.064	0.067	0.061
500	0.3	MCAR	0.046	0.065	0.137	0.201	0.053	0.074	0.046
500	0.3	MAR	0.047	0.078	0.154	0.224	0.058	0.079	0.054
500	0.3	MNAR	0.057	0.068	0.131	0.204	0.061	0.071	0.056
1,000	0.1	MCAR	0.053	0.059	0.074	0.094	0.055	0.057	0.053
1,000	0.1	MAR	0.042	0.049	0.077	0.1	0.044	0.048	0.044
1,000	0.1	MNAR	0.047	0.054	0.07	0.081	0.052	0.056	0.051
1,000	0.2	MCAR	0.043	0.054	0.121	0.173	0.048	0.057	0.043
1,000	0.2	MAR	0.048	0.057	0.127	0.191	0.052	0.06	0.051
1,000	0.2	MNAR	0.053	0.065	0.151	0.217	0.06	0.069	0.055
1,000	0.3	MCAR	0.047	0.061	0.21	0.32	0.047	0.07	0.046
1,000	0.3	MAR	0.043	0.066	0.237	0.362	0.049	0.069	0.049
1,000	0.3	MNAR	0.045	0.07	0.217	0.351	0.061	0.071	0.055

variables in the imputation procedure is the optimal method. However, if the run time is the only concern, the random forest on-the-fly imputation is better.

DISCUSSIONS

This research examined four machine learning-based imputation methods, including the KNN, and three strategies based on the random forest. We proposed the concepts of supervised and unsupervised imputations. Although the RMSE and PFC are similar for the four machine learning-based imputation strategies, type-I error of the Cox model could be inflated dramatically for the different methods under MNAR. Hence, the validity of the Cox model using imputed data changes dramatically under different settings. The simulation results showed that the RF_{mf} performs the best even under the most challenging situation, MNAR. Therefore, this strategy would be valid under all types of missing mechanisms.

One of the most significant advantages of machine learning is that it is suitable for high-dimensional data, time-series data, or complex data interactions. Although this study focuses on survival data with 10 predictive variables, the concept of supervised or unsupervised imputation and the structure of predictors could be easily extended for different study designs.

Finally, the simulation study is the null hypothesis of the Cox model, where the predictive and survival outcomes are independent. Therefore, the type-I error is the essential tool when comparing performances of the four imputation strategies.

Power study is not meaningful since the only valid imputation method is the RF_{mf}. In addition, the rest strategies revealed inflated type-I errors under MNAR.

The R code implemented in the simulations is freely available. We have included the code as **Supplementary Materials**. The file "USER_UTILITY" is the first program that generates the study samples, missing mechanisms, and imputation strategies. The second file, "MAIN," generates all statistical results and figures. Researchers could quickly adopt supervised and unsupervised imputations for the four methods by using the two R codes for future applications.

Limitations

In machine learning, there are many methods for prediction and classification, such as the support vector machine (SVM) (18), extreme gradient boosting machine (XGBoost) (19), and artificial neural network (ANN) (20). In future studies, these methods may also develop novel imputation strategies. Therefore, we did not include the three methods in this research. We simulated the four dichotomous and six continuous predictors as independent variables. A high correlation among them may cause more bias in type-I errors. The categorical predictors could have more levels in simulations, but we expect that the comparisons and patterns between the methods studied in this research are likely to be similar.

This machine learning-based research revealed a robust missing data imputation strategy for survival analysis under various missing mechanisms. The non-parametric "missForest" imputation (RF_{mf}), that excludes the survival time and censoring status from the imputation scheme, could provide valid results using the Cox proportional hazard model under the impact of MCAR, MAR, and MNAR.

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

C-YG proposed the research concept, supervised the project, and wrote the manuscript. Y-CY conducted the analysis and prepared figures and tables. Y-HC jointly supervised the research and provided a scholarship for Y-CY's work. All authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.680054/full#supplementary-material

REFERENCES

1. Little RJ, Rubin DB. *Statistical Analysis with Missing Data*. Manhattan, NY: John Wiley and Sons (2019). doi: 10.1002/9781119482260
2. Hair JF, Black WC, Babin BJ, Anderson RE, Tatham RL. *Multivariate Data Analysis*. Upper Saddle River, NJ: Prentice Hall.
3. Ono M, Miller HP. Income nonresponses in the current population survey. In: *Proceedings of the Social Statistics Section*. Bethesda, MD: American Statistical Association (1969). p. 277–88.
4. Ford BL. An overview of hot-deck procedures. *Incom Data Sample Surv.* (1983) 2:185–207.
5. Andridge RR, Little RJ. A review of hot deck imputation for survey non-response. *Int Stat Rev.* (2010) 78:40–64. doi: 10.1111/j.1751-5823.2010.00103.x
6. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc.* (1996) 91:473–89. doi: 10.1080/01621459.1996.10476908
7. Fix E. *Discriminatory Analysis: Nonparametric Discrimination, Consistency Properties*. Randolph Field, TX: USAF School of Aviation Medicine (1951). doi: 10.1037/e471672008-001
8. Cover T, Hart P. Nearest neighbor pattern classification. *IEEE Trans Inform Theor.* (1967) 13:21–7. doi: 10.1109/TIT.1967.1053964
9. Ohmann JL, Gregory MJ. Predictive mapping of forest composition and structure with direct gradient analysis and nearest-neighbor imputation in coastal Oregon, U.S.A. *Can J For Res.* (2002) 32:725–41. doi: 10.1139/x02-011
10. Peterson LE. K-nearest neighbor. *Scholarpedia.* (2009) 4:1883. doi: 10.4249/scholarpedia,.1883
11. Zhu M, Cheng X. Iterative KNN imputation based on GRA for missing values in TPLMS. In: *Proceedings of the 2015 4th International Conference on Computer Science and Network Technology (ICCSNT)*. Harbin: IEEE (2015). doi: 10.1109/ICCSNT.2015.7490714
12. Shah JS, Rai SN, DeFilippis AP, Hill BG, Bhatnagar A, Brock GN. Distribution based nearest neighbor imputation for truncated high dimensional data with applications to pre-clinical and clinical metabolomics studies. *BMC Bioinformatics.* (2017) 18:114. doi: 10.1186/s12859-017-1547-6
13. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell.* (1998) 20:832–44. doi: 10.1109/34.709601
14. Breiman L. Random forests. *Mach Learn.* (2001) 45:5–32. doi: 10.1023/A:1010933404324
15. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics.* (2012) 28:112–8. doi: 10.1093/bioinformatics/btr597
16. Tang F, Ishwaran H. Random forest missing data algorithms. *Statist Analy Data Mining ASA Data Sci J.* (2017) 10:363–77. doi: 10.1002/sam.11348
17. R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing V, Austria. Available online at: http://www.R-project.org/
18. Mitchell TM. *Machine Learning*. New York, NY: McGraw-Hill (1997).
19. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA (2016). p. 785–94. doi: 10.1145/2939672.2939785
20. Hassoun MH. *Fundamentals of Artificial Neural Networks.* Cambridge, MA: United States: MIT Press (1995).

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

OPINION
published: 15 July 2021
doi: 10.3389/fpubh.2021.716754

Moving Beyond Simple Risk Prediction: Segmenting Patient Populations Using Consumer Data

*Mandana Rezaeiahari**

Department of Health Policy and Management, University of Arkansas for Medical Sciences, Little Rock, AR, United States

Keywords: consumer marketing, patient segmentation, population health, social determinants of health, risk stratification

OPEN ACCESS

Edited by:
Yi Guo,
University of Florida, United States

Reviewed by:
Hui Shao,
University of Florida, United States
Yanmin Zhu,
Harvard Medical School,
United States

****Correspondence:***
Mandana Rezaeiahari
mrezaeiahari@uams.edu

Specialty section:
This article was submitted to
Digital Public Health,
a section of the journal
Frontiers in Public Health

***Received:** 29 May 2021*
***Accepted:** 24 June 2021*
***Published:** 15 July 2021*

Citation:
Rezaeiahari M (2021) Moving Beyond Simple Risk Prediction: Segmenting Patient Populations Using Consumer Data. Front. Public Health 9:716754. doi: 10.3389/fpubh.2021.716754

INTRODUCTION

There has been growing interest among health systems in population health (1, 2). Population health aims to improve the overall health of a population across the full continuum of care by more targeted, effective and coordinated health services (3). Given the rising trend of aging population and chronic disease burden, managing population health becomes more important for health systems trying to control cost (4).

In order to improve outcomes and efficiency, health systems need to customize care and interventions based on identified risks and costs (5). One of the systematic approaches in the literature for targeting interventions to subgroups of patients with different needs is population segmentation or risk-stratification (referred as *patient segmentation* in the remainder of this paper). Population segmentation that divides a population into groups with related service needs is an important foundation for effective and sustainable care delivery (6–8). Segmentation divides patients into distinct groups with specific needs, characteristics or behaviors and allows for health services to be organized around patients with similar needs (7). Patient segmentation models are becoming essential element of healthcare management due to the increase in the number of programs that incentivize value-based care (9).

Although patient segmentation models can help design interventions targeting subgroups of patients, they are often based on International Classification of Diseases (ICD) codes found in electronic health records (EHRs) and/or insurance claims data and lack important social risk factors that are essential for designing interventions. World Health Organization defines social determinants of health (SDOH) as the conditions in which people are born, grow, work, live, and age (10). These factors include economic policies and systems, development agendas, social norms, social policies and political systems (10). There are numerous studies demonstrating social factors acting as powerful determinants on multiple health outcomes including coronary heart disease (11), breast cancer (12), childhood obesity (13) and end-stage renal failure (14). Literature suggests that high utilizers of healthcare resources among Medicaid and uninsured population often have multiple chronic conditions (15, 16) and programs targeting this population collectively argue that social risk factors including but not limited to language, health literacy, unemployment, substance abuse and housing are important drivers of healthcare utilization (17, 18).

Most of the current patient segmentation models use administrative billing data because insurance claims data provides a nearly complete view of patients' interactions with health care delivery system; therefore, it is a reliable source to extract utilization outcomes (19). Majority of the EHRs on the other hand contain data from clinical encounters occurring between individuals and providers within a single health system and hence miss out of network events (19). On the positive side of EHRs is that they offer more extensive data including family history, lab results, vital signs and symptoms which could help improve the population segmentation model (20). One drawback of reliance on insurance claims data and EHRs is that they miss social and behavioral factors that complicate care (21). Although, there is a subset of ICD-10-CM codes, the Z codes, for documenting SDOH in EHRs, these codes are underutilized (22, 23). As such, SDOH Z codes may not reflect the actual burden of social needs experienced by patients. To address this gap, this paper presents the complementary benefit of consumer data when it is linked to EHRs or insurance claims data. The consumer marketing data include individual-level SDOH (including income, education, lifestyle variables, language spoken, household size, smoking status, life events, shopping activity) that are not available in the insurance claims data or majority of EHR data. The combined data provides 360-degree view of patients and can help predict the risk of repeat emergency room visits or hospital admissions (24). Inclusion of SDOH is essential to improve population health as medical interventions without addressing social determinants are not sustainable and effective. This unprecedented view into the lives of patients has significant potential to improve upon segmentation approaches relying exclusively on health plan or EHR data that lack measures or even decent proxies for fitness, diet and other SDOH which can profoundly alter the course of chronic diseases. A number of commercial companies provide marketing data that is well-utilized by organizations that subscribe to their services. Experian's ConsumerView^{SM} U.S. database is one of the world's largest consumer database on more than 300 million individuals and 126 million households (25). ConsumerView^{SM} U.S. database is compiled from hundreds of resources. For example, property and mortgage data are compiled from public records and county deeds while lifestyle and interest data are compiled from consumers who have completed self-reported surveys (25). Marketing companies match and mange patient identity across the healthcare ecosystems enabling the linkage of datasets across channels and silos (26, 27). According to Acxiom, two-thirds of hospitals actively use or want third-party consumer and lifestyle data to improve patient care (24).

CURRENT PATIENT SEGMENTATION MODELS

There are two major approaches for conducting population segmentation in the literature. Expert-driven approaches are informed by expert consensus while data-driven approaches use statistical analysis such as clustering to segment a population (28). John Hopkins Adjusted Clinical Group (ACG) system and the Clinical Risk Group (CRG) system by 3M Health Information Systems are examples of expert-driven approaches (29, 30). The ACG system assigns each diagnosis code to one or more of 32 diagnosis groups referred to as Aggregated Diagnosis Groups (ADG). Both ACG and CRG system use diagnostic codes to classify patients into over 200 mutually exclusive risk groups (28). ADGs are assigned based on five features of conditions: duration, severity, diagnostic certainty, type of etiology and expected need for specialty care. The 3M CRG system assigns an individual five-digit classification code with first digit representing the core health status group, second through the fourth digit representing the base 3M CRG and the fifth digit identifying the severity-of-illness level (30). One drawback of expert-driven approaches is that they subjectively segment populations and no specific standards are set to derive the number of segments. Data-driven approaches generate evidence-based insights of population health status based on patient healthcare data to support policy decisions (1). There have been multiple studies using data-driven approaches to segment populations (19, 31, 32). Zhang et al. (33) developed a patient taxonomy with ten categories to divide high-cost Medicare Fee-For-Service patients. They found high-cost patients were most likely to have multiple chronic conditions, serious mental illness, serious medical illness and frailty (33). Low et al. (28) used cluster analysis and healthcare utilization data from electronic medical records to develop five segments of population (28). Concurrent with patient segmentation models developed by researchers, many predictive models based on SDOH have been developed by health payers and analytics companies. Most often these models are proprietary hence not available for review and scrutiny (34). For instance, a non-profit health insurance company used consumer data to develop a segmentation model to make informed adjustments to its Medicare marketing efforts (35).

DISCUSSION

As medical care is only responsible for 15 to 20% of preventable mortality in the US (36) and due to the increasing impact of social factors on health, it is now time to leverage data analytics to start to understand SDOH and its impact on health and design more social centered care coordination interventions (37). A recent critical review of patient segmentation models shows a lack of comprehensive models that integrates data from multiple sources, with a majority of the models limited to administrative billing data alone (21).

Healthcare organizations and payers should strive to link their traditional resources including EHRs and insurance claims data to consumer marketing data. Through this linkage, they can then apply advanced analytics to get tangible results that can be acted upon to improve quality of care and health outcomes. Specifically, more data driven approaches are needed to utilize available data to assess whether distinct patient subgroups might exist within population. For example, cluster analysis may be used to determine if individual level SDOH (based on consumer marketing data) and insurance claims, together can represent social, medical and behavioral health conditions to form specific relevant subgroup of patients. The proposed patient segmentation framework will facilitate healthcare resource planning and development of interventions to improve the healthcare delivery for each segment. This approach to segmentation will demonstrate heterogeneity in population groups with respect to age, morbidity, lifestyle, setting in which care was mostly used, etc. Therefore, depending on the patterns of utilization of care, complexity level of patients and lifestyle segmentation, various models of care will be needed. For instance, for "young or middle age and healthy" segment that focus little on preventive care and are fans of fast food, the most important approaches may be disease prevention, health education and robust primary care, working with non-healthcare partners such as employers, community-based disease education in order to maintain the health status and promote healthy behavior. Patients with stable but chronic condition that are more interested in adopting technology, can instead benefit more from supportive self-management such as home-based self-monitoring tools to promote health empowerment. Patients with complex chronic conditions that are not managed well and live in neighborhoods with low levels of food access may require more multidisciplinary medical and social care coordination.

Some other examples of the opportunities as a result of linking consumer data to insurance claims and/or EHRs (not limited to patient segmentation) include reduction of obesity through increasing the relevance and effectiveness of weight loss engagement strategies by using consumer lifestyle segmentation variables including diet attitudes and motivations, gauging the receptivity of patients to different outreach channels (automated voice, live agent calls and text messages) using the digital media preference, age and education level, identifying food insecure households/individuals using frequency and dollars spend in food category particularly for individuals living in low-income and low food-access neighborhoods.

Unlike other traditional SDOH data sources that are only available at the county and/or zip code level, such as Area Health Resource Files (38), US Census County Business Patterns (39), and County Health Rankings (40), consumer data is available at the individual or household level. County level social data, although useful, only represent a profile of the community and does not reliably represent the profile of the individual patient. For example, research has shown that poverty is strongly associated with an increase in risk of dying, but simply living in a high-poverty area is not (41).

Despite the important opportunity that the consumer marketing data brings to healthcare, major concerns still exist about privacy of consumers. Linking consumer marketing data to EHRs and/or insurance claims data may increase informational risk (i.e., HIPAA violations), if strict data deidentification standards are not in place and/or data protections are applied inconsistently across various entities which collect, share and use the data (42). As such, any use cases of consumer data must be HIPAA compliant to ensure protection of "individually identifiable health information" (i.e., protected health information) (43). Some of the best practices to ensure compliance are safe sourcing (working with the source compilers of consumer data to ensure compliance), safe storage (reviewing and updating data privacy policies to control access), appropriate/ethical use of data (marketing data should never be used to deny access to anyone or result in health disparities) (44).

Other challenges of using consumer data include reproducibility and analytical challenges. Predictive models developed by the private sector are not shared publicly, therefore cannot be replicated by other researchers to ensure accuracy, validity and potential model bias (34). Additionally, researchers should be cautious when selecting the analytical approaches when it comes to the inclusion of marketing data to predict health outcomes. Highly flexible machine learning algorithms may select features (e.g., reality TV show from consumer interest data) to predict mortality which may not be clinically reasonable.

Despite the challenges discussed above, consumer marketing data may open up opportunities to health researchers to understand how individual level SDOH manifest throughout a person's life. Future patient segmentation models that incorporate SDOH from consumer marketing data have the potential to improve health and reduce health disparities by ensuring that the right patients will be intervened at the right time.

AUTHOR CONTRIBUTIONS

MR devised the idea, performed literature review, and wrote the manuscript.

REFERENCES

1. Vuik SI, Mayer E, Darzi A. A quantitative evidence base for population health: applying utilization-based cluster analysis to segment a patient population. *Popul Health Metr.* (2016). 14:44. doi: 10.1186/s12963-016-0115-z
2. Yan S, Kwan YH, Tan CS, Thumboo J, Low LL. A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Med Res Methodol.* (2018) 18:121. doi: 10.1186/s12874-018-0584-9
3. Felt-Lisk, S, Higgins T. *Exploring the Promise of Population Health Management Programs to Improve Health*. Washington, DC: Mathematica Policy Research (2011).
4. Nnoaham KE, Cann KF. Can cluster analyses of linked healthcare data identify unique population segments in a general practice-registered population? *BMC Public Health.* (2020) 20:798. doi: 10.21203/rs.2.12272/v2
5. *Value Transformation Framework Action Guide*. (2019). Available online at: https://www.nachc.org/wp-content/uploads/2019/03/Risk-Stratification-Action-Guide-Mar-2019.pdf (accessed May 28, 2021).
6. Chong JL, Matchar DB. Benefits of population segmentation analysis for developing health policy to promote patient-centred care. *Ann Acad Med Singap.* (2017) 46:287–9.
7. Vuik SI, Mayer EK, Darzi A. Patient segmentation analysis offers significant benefits for integrated care and support. *Health Aff.* (2016) 35:769–75. doi: 10.1377/hlthaff.2015.1311
8. Lynn J, Straube BM, Bell KM, Jencks SF, Kambic RT. Using population segmentation to provide better health care for all: the "bridges to health" model. *Milbank Q.* (2007) 85:185–208. doi: 10.1111/j.1468-0009.2007.00483.x
9. *Early Adopters of the Accountable Care Model: A Field Report on Improvements in Health Care Delivery. Commonwealth Fund*. Available online at: https://www.commonwealthfund.org/publications/fund-reports/2013/mar/early-adopters-accountable-care-model-field-report-improvements (accessed May 18, 2021).
10. *Social determinants of health*. Available online at: https://www.who.int/health-topics/social-determinants-of-health#tab=tab_1 (accessed June 21, 2021).
11. Kim D. The associations between US state and local social spending, income inequality, and individual all-cause and cause-specific mortality: the National Longitudinal Mortality Study. *Prev Med.* (2016) 84:62–8. doi: 10.1016/j.ypmed.2015.11.013
12. Shariff-Marco S, Yang J, John EM, Kurian AW, Cheng I, Leung R, et al. Intersection of race/ethnicity and socioeconomic status in mortality after breast cancer. *J Commun Health.* (2015) 40:1287–99. doi: 10.1007/s10900-015-0052-y

13. Flood TL, Zhao YQ, Tomayko EJ, Tandias A, Carrel AL, Hanrahan LP. Electronic health records and community health surveillance of childhood obesity. *Am J Prev Med.* (2015) 48:234–40. doi: 10.1016/j.amepre.2014.10.020
14. Hill KE, Gleadle JM, Pulvirenti M, McNaughton DA. The social determinants of health for people with type 1 diabetes that progress to end-stage renal disease. *Health Expect.* (2015) 18:2513–21. doi: 10.1111/hex.12220
15. Johnson TL, Rinehart DJ, Durfee J, Brewer D, Batal H, Blum J, et al. For many patients who use large amounts of health care services, the need is intense yet temporary. *Health Aff.* (2015) 34:1312–9. doi: 10.1377/hlthaff.2014.1186
16. *Faces of Medicaid III: Refining the Portrait of People with Multiple Chronic Conditions—Center for Health Care Strategies*. Available online at: https://www.chcs.org/resource/the-faces-of-medicaid-iii-refining-the-portrait-of-people-with-multiple-chronic-conditions/ (accessed May 14, 2021).
17. *Caring for High-Need, High-Cost Patients: What Makes for a Successful Care Management Program? Commonwealth Fund*. Available online at: https://www.commonwealthfund.org/publications/issue-briefs/2014/aug/caring-high-need-high-cost-patients-what-makes-successful-care (accessed May 14, 2021).
18. *Strategies to Reduce Costs and Improve Care for High-Utilizing Medicaid Patients: Reflections on Pioneering Programs—Center for Health Care Strategies*. Available online at: https://www.chcs.org/resource/strategies-to-reduce-costs-and-improve-care-for-high-utilizing-medicaid-patients-reflections-on-pioneering-programs/ (accessed May 14, 2021).
19. Kharrazi H, Chi W, Chang HY, Richards TM, Gallagher JM, Knudson SM, et al. Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient electronic health records versus administrative claims. *Med Care.* (2017) 55:789–96. doi: 10.1097/MLR.0000000000000754
20. Wilson J, Bock A. *The benefit of using both claims data and electronic medical record data in health care analysis*. (2012). Available online at: https://www.optum.com/content/dam/optum/resources/whitePapers/Benefits-of-using-both-claims-and-EMR-data-in-HC-analysis-WhitePaper-ACS.pdf (accessed June 22, 2021).
21. Jeffery AD, Hewner S, Pruinelli L, Lekan D, Lee M, Gao G, et al. Risk prediction and segmentation models used in the United States for assessing risk in whole populations: a critical literature review with implications for nurses' role in population health management. *JAMIA Open.* (2019) 2:205–14. doi: 10.1093/jamiaopen/ooy053
22. Truong HP, Luke AA, Hammond G, Wadhera RK, Reidhead M, Joynt Maddox KE. Utilization of social determinants of health ICD-10 Z-codes among hospitalized patients in the United States, 2016–2017. *Med Care.* (2020) 58:1037–43. doi: 10.1097/MLR.0000000000001418
23. Guo Y, Chen Z, Xu K, George TJ, Wu Y, Hogan W, et al. International classification of diseases, tenth revision, clinical modification social determinants of health codes are poorly used in electronic health records. (2020) 99:e23818. doi: 10.1097/MD.0000000000023818
24. Acxiom. *The Power of Consumer and Lifestyle Data Iin Healthcare*. Available online at: https://www.acxiom.com/resources/infographic-the-power-of-consumer-and-lifestyle-data-in-healthcare/ (accessed June 22, 2021).
25. Experian. *Experian audience lookbook*. Available online at: https://www.experian.com/content/dam/marketing/na/assets/ems/marketing-services/documents/product-sheets/audience-lookbook.pdf (accessed June 19, 2021).
26. *Healthcare Marketing—Predictive Analytics, Database Solutions, Strategy*. Available online at: https://www.acxiom.com/healthcare/ (accessed June 22, 2021).
27. ConsumerView SM. *Tap Into the Power of the World's Largest Consumer Database* (2018).
28. Low LL, Yan S, Kwan YH, Tan CS, Thumboo J. Assessing the validity of a data driven segmentation approach: a 4 year longitudinal study of healthcare utilization and mortality. *PLoS ONE.* (2018) 13:e0195243. doi: 10.1371/journal.pone.0195243
29. The Johns Hopkins ACG® System. *Excerpt from Version 11.0 Technical Reference Guide*. The Johns Hopkins ACG® System (2014).
30. *3MTM Clinical Risk Groups: Measuring risk, managing care*. (2016). Available online at: https://multimedia.3m.com/mws/media/765833O/3m-crgs-measuring-risk-managing-care-white-paper.pdf (accessed June 22, 2021).
31. Rinehart DJ, Oronce C, Durfee MJ, Ranby KW, Batal HA, Hanratty R, et al. Identifying subgroups of adult superutilizers in an urban safety-net system using latent class analysis. *Med Care*. (2018). 56:e1–9. doi: 10.1097/MLR.0000000000000628
32. Murphy SME, Castro HK, Sylvia M. Predictive modeling in practice : improving the participant identification process for care management programs using condition-specific cut points. *Popul Health Manag.* (2011) 14:205–10. doi: 10.1089/pop.2010.0005
33. Zhang Y, Grinspan Z, Khullar D, Unruh MA, Shenkman E, Cohen A, et al. Developing an actionable patient taxonomy to understand and characterize high-cost Medicare patients. *Healthcare.* (2020) 8:100406. doi: 10.1016/j.hjdsi.2019.100406
34. Tan M, Hatef E, Taghipour D, Vyas K, Kharrazi H, Gottlieb L, et al. Including social and behavioral determinants in predictive models: trends, challenges, and opportunities. *JMIR Med Inform.* (2020) 8:e18084. doi: 10.2196/18084
35. Simpson M, Genovese A. *Leveraging consumer data to grow medicare market share*. (2016). Available online at: https://info.carrothealth.com/hubfs/Brochures%20and%20Whitepapers/Carrot%20Health%20-%20Leveraging%20Consumer%20Data%20to%20Grow%20Medicare%20Market%20Share.pdf?__hstc=122733652.515b5d9ff7a33417378b5a218fdca83f.1567383407805.1567386846670.1567398132639.3&__hssc=122733652.1.1567398132639 (accessed May 24, 2021).
36. McGinnis JM, Williams-Russo P, Knickman JR. The case for more active policy attention to health promotion. *Health Aff.* (2002) 21:78–93. doi: 10.1377/hlthaff.21.2.78
37. Mackenbach JP. The contribution of medical care to mortality decline: mcKeown revisited. *J Clin Epidemiol.* (1996) 49:1207–13. doi: 10.1016/S0895-4356(96)00200-4
38. *Area Health Resources Files*. Available online at: https://data.hrsa.gov/topics/health-workforce/ahrf (accessed June 22, 2021).
39. *Bureau UC. County Business Patterns (CBP)*. Available online at: https://www.census.gov/programs-surveys/cbp.html (accessed June 22, 2021).
40. *County Health Rankings & Roadmaps*. Available online at: https://www.countyhealthrankings.org/ (accessed June 22, 2021).
41. Holt-Lunstad J, Smith TB, Layton JB. Social relationships and mortality risk: a meta-analytic review. *PLoS Med.* (2010) 7:e1000316. doi: 10.4016/19865.01
42. Rahimzadeh V. A policy and practice review of consumer protections and their application to hospital-sourced data aggregation and analytics by third-party companies. *Front Big Data.* (2021) 3:44. doi: 10.3389/fdata.2020.603044
43. *Enhance Healthcare Analytics with Consumer Data*. (2019). Available online at: https://marketing.acxiom.com/US-Enhance-Healthcare-eb-main2.html?&utm_source=website&utm_medium=owned&utm_campaign=EnhancedHCeB (acessed October 3, 2021).
44. *The 3 keys to compliance for healthcare marketing data—Healthcare Blog*. Available online at: https://www.experian.com/blogs/healthcare/2019/05/the-3-keys-to-compliance-for-healthcare-marketing-data/ (accessed May 18, 2021).

Conflict of Interest: The author declares 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: 17 August 2021
doi: 10.3389/fpubh.2021.713313

Impact of Healthcare Non-Take-Up on Adherence to Long-Term Positive Airway Pressure Therapy

*Najeh Daabek [1,2], Renaud Tamisier [1,3], Alison Foote [4], Hélèna Revil [5], Marie Joyeux-Jaure [1,2,3], Jean-Louis Pépin [1,3], Sébastien Bailly [1,3] and Jean-Christian Borel [1,2]**

[1] HP2 Laboratory, INSERM U1042, University Grenoble Alpes, Grenoble, France, [2] AGIR à dom. Homecare Charity, Meylan, France, [3] EFCR Laboratory, Grenoble Alpes University Hospital, Grenoble, France, [4] Research Division, Grenoble Alpes University Hospital, Grenoble, France, [5] Social Sciences Research – PACTE Laboratory, CNRS UMR 5194, University Grenoble Alpes, Grenoble, France

Background: The effectiveness of positive airway pressure therapies (PAP) is contingent on treatment adherence. We hypothesized that forgoing healthcare may be a determinant of adherence to PAP therapy.

Research Question: The objectives were: (i) to assess the impact of forgoing healthcare on adherence to PAP in patients with Chronic Respiratory Failure (CRF) and patients with Obstructive Sleep Apnea Syndrome (OSAS); (ii) to compare forgoing healthcare patterns in these two chronic conditions.

Study design and methods: Prospective cohort of patients with OSAS or CRF, treated with PAP therapies at home for at least 12 months. At inclusion, patients were asked to fill-in questionnaires investigating (i) healthcare forgone, (ii) deprivation (EPICES score), (iii) socio-professional and familial status. Characteristics at inclusion were extracted from medical records. PAP adherence was collected from the device's built-in time counters. Multivariable logistic regression models were used to assess the associations between healthcare forgone and the risk of being non-adherent to CPAP treatment.

Results: Among 298 patients included (294 analyzed); 33.7% reported forgoing healthcare. Deprivation (EPICES score $>$ 30) was independently associated with the risk of non-adherence (OR = 3.57, 95%CI [1.12; 11.37]). Forgoing healthcare had an additional effect on the risk of non-adherence among deprived patients (OR = 7.74, 95%CI [2.59; 23.12]). OSAS patients mainly forwent healthcare for financial reasons (49% vs. 12.5% in CRF group), whereas CRF patients forwent healthcare due to lack of mobility (25%, vs. 5.9 % in OSAS group).

Interpretation: Forgoing healthcare contributes to the risk of PAP non-adherence particularly among deprived patients. Measures tailored to tackle forgoing healthcare may improve the overall quality of care in PAP therapies.

Clinical Trial Registration: The study protocol was registered in ClinicalTrials.gov, identifier: NCT03591250.

Keywords: CPAP, non-invasive ventilation, PAP therapy, healthcare non take up, adherence—compliance—persistence

OPEN ACCESS

Edited by:
Yi Guo,
University of Florida, United States

Reviewed by:
Xue Han,
Bristol Myers Squibb, United States
Zhaoyi Chen,
University of Florida, United States

****Correspondence:***
Jean-Christian Borel
j.borel@agiradom.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: *22 May 2021*
Accepted: *21 July 2021*
Published: *17 August 2021*

Citation:
Daabek N, Tamisier R, Foote A, Revil H, Joyeux-Jaure M, Pépin J-L, Bailly S and Borel J-C (2021) Impact of Healthcare Non-Take-Up on Adherence to Long-Term Positive Airway Pressure Therapy. Front. Public Health 9:713313. doi: 10.3389/fpubh.2021.713313

INTRODUCTION

Sleep breathing disorders, particularly obstructive sleep apnea syndrome (OSAS), nocturnal alveolar hypoventilation and at worst chronic respiratory failure (CRF) are associated with incapacitating symptoms affecting quality of life, and poor long term outcomes including cardio-vascular events and early mortality (1–3). Since the early 80s', non-invasive positive airway pressure therapies [Continuous Positive Airway Pressure (CPAP) and Non-Invasive Ventilation (NIV)] have been the first-line treatments for OSAS and CRF (4–7).

The effectiveness of positive airway pressure (PAP) therapies is however contingent on treatment adherence, and a significant proportion of non-adherence and high therapy termination rates are observed (8, 9). Despite continuous technological innovations, adherence to PAP therapies has plateaued over the last 20 years (10), suggesting that adherence is dependent on patients' personal characteristics such as their marital or social status (11–13), their perception of treatment efficacy (14), any benefits experienced (15), and their priorities regarding personal lifestyle (16–18).

Poor adherence to PAP therapies might reflect societal vulnerability, deprivation and non-prioritization of personal health. A comprehensive and holistic way of investigating and understanding health-related behaviors is to study reasons individuals forgo healthcare and to estimate the prevalence of this attitude. The concept of forgoing healthcare corresponds to societal, health-system contexts or personal conducts and/or beliefs leading individuals to forgo or postpone self-identified healthcare needs to which they have rights. This concept allows us to understand the relationship that people have with the healthcare system and to apprehend the influence of individual and collective factors on health related behavior. A large part of research on the forgoing healthcare phenomenon has focused on underprivileged populations who forgo healthcare primarily for financial reasons (19, 20).

However, multiple reasons for forgoing healthcare are also reported by individuals without financial constraints. These include lack of time owing to the burden of professional or personal life, lassitude or negligence, and inadequate transport with long distances between their residence and care facilities. In addition, some studies on the concept of forgoing healthcare show that not all people are exposed in the same way to this phenomenon. Depending on their sex, family and/or professional situation, or their level of multidimensional deprivation, the pattern of forgoing healthcare varies (21). Furthermore, qualitative social science studies indicate that individuals can forgo care related to a particular chronic condition but seek treatment for other conditions and vice versa (22).

Therefore, assessing influence of socioeconomics factors like deprivation and healthcare non-take up on specific populations like OSAS and CRF patients is an essential step to personalization and optimization of the healthcare delivery. In addition, unlike oral treatments, PAP therapies required for OSA and CRF patients have the advantage of a long-term objective assessment of treatment adherence (thanks to telemonitoring). These respiratory pathologies represent therefore an ideal disease model for designing and testing multifactorial interventions to promote treatment adherence. Moreover, OSA and CRF subgroups have well-known differences in clinical presentations and socio-economic status that could generate different profiles for health care renunciation.

In this study, we hypothesized that forgoing healthcare may be a significant determinant of PAP-therapies adherence. As, clinical presentation and socioeconomic status is dissimilar between OSAS and CRF populations, we decided to evaluate and compare the prevalence of forgoing healthcare (related or not to their respiratory disease) in two populations, OSAS and CRF patients, both on long-term home PAP treatment. We compared the ways in which individuals forwent healthcare and the reasons.

MATERIALS AND METHODS

Study Design

The present study was a prospective monocentric cohort study (Department of Pulmonology, Grenoble Alpes University Hospital). Ethical approval was obtained from the French Ethics Committee "Ile de France II" and the study protocol was registered in ClinicalTrials.gov (NCT03591250). The study was conducted between June 2018 and November 2019. Each participant provided written informed consent before inclusion in the study.

Study Participants

During a routine medical follow-up consultation, patients meeting the following inclusion criteria were asked to participate (**Supplementary Figure 1**):

- Age above 18 years
- Affiliated to the French social security system or a beneficiary of this system
- A diagnosis of OSAS or CRF
- Treated with CPAP or NIV for at least 12 months
- Routinely followed by the same homecare provider (AGIR à Dom, Meylan, France)
- Able to fill in the study questionnaires.

Study Objectives

Our primary objective was the impact of forgoing healthcare on adherence to PAP therapy. The secondary objective was a comparison of forgoing healthcare patterns between patients with CRF and patients with OSAS.

Data Collection and Procedures

Assessment of Healthcare Forgone

Participants were asked to fill-in the "healthcare non-take up" questionnaire during their routine medical follow-up

Abbreviations: AHI, Apnea hypopnea index; BMI, Body mass index; CPAP, Continuous Positive Airway Pressure; CRF, Chronic respiratory failure; CSS, Complémentaire Santé Solidaire (state-subsidized complementary (top-up) insurance); EPICES, Evaluation de la précarité et des inégalités de santé dans les Centres d'examens de santé (Assessment of precariousness and health inequalities in health examination centers); ESS, Epworth sleepiness score; IQR, Interquartile range; NIV, Non invasive ventilation; OR, Odds Ratio; OSAS, Obstructive Sleep Apnea Syndrome; PAP, Positive Airway Pressure; SD, Standard Deviation.

consultation in the Department of Pulmonology, Grenoble Alpes University Hospital. This questionnaire was originally developed by Dr. Revil's group at the PACTES laboratory (Grenoble-Alpes University, France); and previously used by us in a study of 164,092 public sector health insurance beneficiaries in France (23). Briefly, the questionnaire is structured into three sections and refers to healthcare forgone in the 12 months preceding the study inclusion consultation (**Supplementary Figure 2**):

i. Healthcare forgone: After the key question "Have you forgone or put-off healthcare on one or more occasions in the last 12 months (yes/no)," those answering "yes" were asked about the type(s) of healthcare forgone and their reasons, how long they had been forgoing or putting-off healthcare and their perception of their current state of health.

ii. Healthcare insurance: This section focused on whether participants had complementary, top-up health insurance [through a private company or the state-subsidized "Complémentaire Santé Solidaire" (CSS)]; and if not, the reasons why. They were also asked whether they benefited from 100% cover by the state system due to a long-term chronic condition (e.g., Type I diabetes).

> *Briefly, France has a two-tier system of health insurance: a compulsory primary health insurance scheme and complementary/top-up health insurance schemes. In the compulsory scheme, contributions are proportional to income and reimbursement of care is a fixed percentage of the total cost of care. The rate of reimbursement is set by the state and depends on the type of care. Complementary schemes are essentially private insurance policies which reimburse almost all the remaining healthcare costs not covered by the compulsory scheme. However, for people on low incomes, a means-tested top-up scheme is provided by the state; this "Complémentaire Santé Solidaire" (CSS) is free of charge. Finally, the compulsory French state scheme covers 100% of health expenses related to 29 severe chronic diseases including diabetes, chronic respiratory failure, cancer, cystic fibrosis etc. The list of eligible conditions is set by the public health code.*

iii. Standard of living and deprivation: Socio-professional and familial status were collected. Material and social deprivation were investigated using the 11 item EPICES questionnaire (24, 25). An individual score was calculated for each participant, by adding each question coefficient to the intercept whenever the answer is "yes." According to EPICES a score of $\geq$ 30 indicates deprivation.

Clinical Data and Other Socio-Demographics

Characteristics at inclusion, including age, sex, anthropometrics, main etiologies of respiratory disorders, and hospitalization in the year before inclusion were extracted from the participants' medical records. Data related to NIV or CPAP: date of treatment initiation, PAP adherence in the year following inclusion in the study (objectively measured from the device's built-in time counters and reported every 6 months) and type of mask, were collected from the homecare provider's database.

Sample Size

Based on the hypothesis of a 25% prevalence of forgoing healthcare in the population (23), and allowing for 10% dropout, the enrollment of 300 participants (150 patients treated with CPAP; 150 treated with NIV) would allow 80% power to detect a difference of 1.5 ± 3 (SD) hours/night in CPAP/NIV adherence between patients who forwent healthcare and those who did not.

Statistical Analysis

Descriptive statistics are presented as medians [IQR] for quantitative variables and frequencies (%) for qualitative variables. Chi square tests and non-parametric Mann-Whitney tests were used to compare qualitative and quantitative variables, respectively, between groups (OSAS vs. CRF).

A simple imputation method was used in cases with little missing data (<2%) (26). Otherwise, multiple imputation with fully conditional specification was performed (27).

Primary Outcome Analysis

Average PAP therapy use was defined as the mean of the measures collected from the devices in the year following inclusion. Normality of mean PAP use was assessed both graphically and using the Shapiro-Wilk test, and was not accepted. Thus, data were dichotomized using a threshold of 4 h/night and the adherence to PAP therapy was defined as: adherent for an average of ≥4 h/night, and otherwise non-adherent.

To identify whether forgoing healthcare impacted adherence to PAP therapies, univariable logistic modeling was performed. Covariates were chosen *a priori* based on factors that might impact PAP-therapy adherence, and included sex (28), age, BMI, family and socio-professional status (11, 29) complementary healthcare insurance, reimbursement rates, healthcare forgone, degree of deprivation (EPICES score) (30), PAP-therapy duration (years), number of hospitalizations, and etiology. Variables with a $p < 0.25$ were then introduced into a multivariable model. Given that we were not looking for a predictive model (therefore no evaluation based on performance) but an explanatory model, and in order to take into account the potential confounding factors, a stepwise descending selection was used for the final model selection.

Given the collinearity between forgoing healthcare and deprivation, a four-modality categorical variable was used in the model: (1) healthcare forgone and no deprivation, (2) no healthcare forgone and deprivation, (3) healthcare forgone and deprivation, (4) no deprivation and no healthcare forgone.

Secondary Outcome Analysis

A comparison of the pattern of forgoing healthcare between patients with OSAS and those with CRF was conducted using a Chi-square test for qualitative variables and a non-parametric Wilcoxon test for the quantitative variables.

RESULTS

The study flow chart is shown in **Supplementary Figure 3**. Of 298 patients included in the study and who responded to the healthcare non-take-up questionnaire, four patients had no objective measure of their PAP adherence and were excluded from the analysis. None of the patients refused to fill-in the questionnaire.

Table 1 shows the main characteristics of the study population. Participants were predominantly male (64.3%) and obese (30.8 [25.4; 35.4] kg/m^2). Large proportions of the study cohort were living as a couple (67.1%) and/or retired (61.8%). All participants with CRF were prescribed treatment with NIV and 93.7% of OSAS patients were prescribed CPAP at night. Median adherence to PAP therapy was high (7.3 h [5.4; 8.8]) with only 12.8% non-adherent patients, i.e., under the 4 h/night threshold. There was no difference in treatment adherence between CRF and OSAS patients.

Over a third of the population (33.7%) declared forgoing at least one item of healthcare in the 12 months preceding inclusion and 53.4 % were considered to be deprived (EPICES score $>$ 30) (**Table 2**). Patients with CRF were more often covered by health insurance at a rate of 100% than patients with sleep apnea (88.8% vs. 39.7%, respectively, $p < 0.01$) and were thus more often exempted from expenses related to their chronic illness (88.8 vs. 39.1, respectively; $p < 0.01$).

Concerning the Impact of forgoing healthcare on adherence to PAP therapies: univariable analysis between adherence to PAP therapies and the different variables of interest are presented in **Supplementary Table 1**. In multivariable analysis, deprivation (EPICES score $>$30) was independently associated with the risk of being non-adherent (OR = 3.57, 95%CI [1.12; 11.37], p = 0.031). We were not able to demonstrate an independent association between healthcare non take up and PAP therapy adherence, however forgoing healthcare had an additional effect on the risk of non-adherence among patients experiencing deprivation (OR = 7.74, 95%CI: [2.59; 23.12], $p < 0.001$) (**Table 3**).

Longer time since PAP-therapy initiation was significantly associated with a lower probability of being non-adherent (OR = 0.88, 95%CI: [0.81; 0.96], p-value: 0.002). Patients who had one or more hospitalization in the year preceding inclusion were less likely to be non-adherent compared to those with no hospitalization at all (OR = 0.40, 95%CI: [0.17; 0.96], p-value: 0.04) (**Table 3**).

Concerning the patterns of forgoing healthcare between OSAS and CRF, the four most frequent types of healthcare foregone

TABLE 1 | General and clinical characteristics of patients (*N* = 294).

Variable	Items	Whole population (*N* = 294)	OSAS (*N* = 158, 53.74%)	CRF (*N* = 136, 46.26 %)	*p*-value*	Missing
BMI (Kg/m^2)		30.8 [25.4; 35.4]	31.1 [26.6; 34.8]	29.4 [23.3; 37.3]	0.40	7
Sex	M	189 (64.3)	126 (79.7)	63 (46.3)	<0.01	0
Age	≤60 years	83 (28.2)	43 (27.2)	40 (29.4)	0.41	0
	]60;70]	109 (37.1)	64 (40.5)	45 (33.1)		
	>70 years	102 (34.7)	51 (32.3)	51 (37.5)		
Family situation	Couples	160 (55.4)	104 (67.1)	56 (41.8)	<0.01	5
	Alone	129 (44.6)	51 (32.9)	78 (58.2)		
Socio-professional status	Working	66 (22.8)	48 (30.8)	18 (13.4)	<0.01	4
	Retired or Unemployed	224 (77.2)	108 (69.2)	116 (86.6)		
PAP Therapy	CPAP	148 (50.3)	148 (93.7)	0 (0)	<0.01	0
	NIV	146 (49.7)	10 (6.3)	136 (100)		
Indication for PAP therapy	OSAS	157 (53.4)	157 (99.4)	0 (0)	<0.01	0
	COPD	26 (8.8)	0 (0)	26 (19.1)		
	Neuromuscular pathology	23 (7.8)	0 (0)	23 (16.9)		
	OHS	32 (10.9)	1 (0.6)	31 (22.8)		
	Chest well disorder and others	56 (19)	0 (0)	56 (41.2)		
Delay since PAP therapy initiation (years)		7.2 [2.2; 12.2]	7.9 [2.9; 13]	5.2 [1.7; 10.9]	0.02	0
Delay since primary diagnosis (years)		8.5 [4.6; 12.9]	8.4 [3.4; 12.5]	10.8 [6.4; 14.7]	0.08	137
Number of hospitalizations in the year preceding inclusion		0 [0; 2]	0 [0; 0]	2 [0; 4]	<0.01	0
% of patients with PAP adherence >4 h/night	Yes	232 (87.2)	134 (87.6)	98 (86.7)	0.84	28
Average PAP-therapy adherence (h/night)		7.3 [5.4; 8.8]	6.8 [5.3; 8]	8.2 [5.8; 10.4]	<0.01	28

Values in Numbers (%) or median [IQR].

**Comparaison of CRF and OSAS groups (p-value were calculated using Chi square tests and non-parametric Mann-Whitney tests).*

BMI, Body mass index; COPD, Chronic obstructive pulmonary disease; CPAP, Continuous positive airway pressure; CRF, Chronic respiratory failure; NIV, Non-invasive ventilation; OHS, Obesity hypoventilation syndrome; OSAS, Sleep apnea syndrome; PAP, Positive airway pressure.

TABLE 2 | Access to care, healthcare coverage, and deprivation ($N = 294$).

Variable	Items	All population ($N = 294$)	OSAS ($N = 158$, 53.74%)	CRF ($N = 136$, 46.26 %)	*p*-value	Miss
Access to care						
Health care forgo	Yes	99 (33.7)	51 (32.3)	48 (35.3)	0.59	0
Healthcare coverage						
Coverage ratio	Partial (60%)	109 (37.6)	94 (60.3)	15 (11.2)	<0.01	4
	100%	181 (62.4)	62 (39.7)	119 (88.8)		
100% coverage due to total disability or long-term illness	Yes	180 (62.1)	61 (39.1)	119 (88.8)	<0.01	4
Complementary (top-up) health insurance	None or state-subsidized "Complémentaire Santé Solidaire (CSC)"	46 (15.8)	18 (11.4)	28 (20.9)	0.03	2
	Private	246 (84.2)	140 (88.6)	106 (79.1)	.	.
Reasons for not having top-up insurance						
- "I don't see the point"	Yes	2 (14.3)	2 (50)	0 (0)	0.02	0
- "I am 100% covered by Health Insurance and I don't think I need any additional"	Yes	7 (50)	1 (25)	6 (60)	0.24	0
Deprivation						
EPICES score		31.4 [15.4; 47.3]	23.1 [10.1; 42.6]	37.3 [26.6; 50]	<0.01	0
Patients with EPICES score >30	N (%)	157 (53.4)	69 (43.7)	88 (64.7)	<0.01	0
Deprivation and healthcare non-take up						
No deprivation and no healthcare non-take up		111 (37.8)	40 (29.41)	71 (44.94)	<0.01	0
Deprivation and no healthcare non-take up		84 (28.6)	48 (35.29)	36 (22.78)		
No deprivation and healthcare non-take up		26 (8.8)	8 (5.88)	18 (11.39)		
Deprivation and healthcare non-take up		73 (24.8)	40 (29.41)	33 (20.89)		

Values in Numbers (%) or median [IQR].
CRF, Chronic respiratory failure; OSAS, Sleep apnea syndrome.

TABLE 3 | Multivariable association between predictors and the probability of being non-compliant ($N = 266$).

Variable	Items	OR	95% CI	*p*-value
Deprivation and healthcare non-take up	Deprivation and no healthcare non-take up	3.57	[1.12; 11.37]	0.0311
	Deprivation and healthcare non-take up	2.01	[0.35; 11.68]	0.4332
	Deprivation and healthcare non-take up	7.74	[2.59; 23.12]	0.0002
	Deprivation and no healthcare non-take up (reference)			
Time since initiation of PAP-Therapy		0.88	[0.81; 0.95]	0.0021
Hospitalizations in the previous year	≥1 hospitalization	0.40	[0.17; 0.96]	0.0395
	No hospitalization			

An univariable analysis was conducted to select the variablesto be included in the multivariable analysis, which were: The reimbursement rate, PAP-therapy duration (years), number of hospitalizations in the previous year and the interaction between healthcare non-take up and deprivation.

were consultations with specialists (51.5%), purchase of medical equipment (35.4%), consultations with primary care physicians (30.3%) and dental care (28.3%) (**Supplementary Table 2**).

Although the rate of forgoing healthcare was not different between OSAS and CRF (respectively 32.3% vs. 35.3%, $p = 0.59$; **Table 3**), the reasons for forgoing care were significantly different. For patients with OSAS it was mainly for financial reasons (49% vs. 12.5% in CRF group, $p < 0.01$), whereas patients with CRF forwent healthcare due to lack of mobility (25% vs. 5.9% in sleep apnea group, $p = 0 < 0.01$). **Figure 1** shows the types of healthcare forgone and reasons. **Figure 2** links types and reasons. In patients with CRF (2b), the lack of mobility was strongly linked to forgoing specialist consultations. In contrast, mainly financial reasons were given by the OSAS group (2a) (**Figures 1**, **2**, and **Supplementary Table 2**).

DISCUSSION

This study investigated the relationship between patterns of non-uptake of healthcare and PAP therapy adherence in two distinctive populations, OSAS and CRF. The rate of forgoing healthcare was higher (33.7%) than that reported in the general

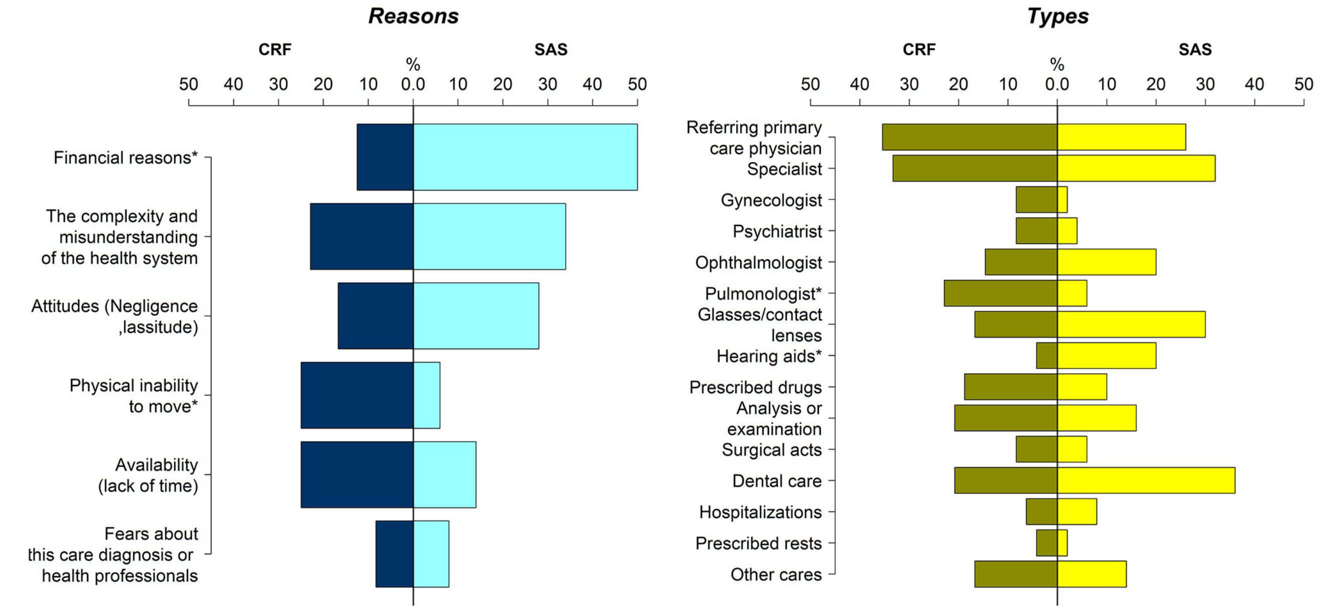


FIGURE 1 | Differences in the pattern of healthcare non-take up between OSAS and CRF patients. The rate of each type of healthcare type forgone and the reasons of the non-take up are presented as percentages (%).

French population (25.4%) (23). Deprivation and foregoing healthcare exert a synergistic effect, increasing the risk of being non-adherent to PAP therapies. The picture was different in OSAS and CRF patients reflecting the functioning of the French healthcare system.

As identified in previous studies (13, 31), our results show a significant association between the level of multidimensional deprivation and PAP adherence. The novelty of our findings is to demonstrate that the combination of deprivation and forgoing healthcare is associated with a nearly 8-fold higher risk of being non-adherent. This reflects the complexity and multi-dimensionality of PAP adherence issues and the need for more transdisciplinary approaches to understand how these several social factors interact (32). The known determinants of low adherence are poorly informative explaining the 4 to 25% of variance in PAP adherence (33). There is a need to include a systematic assessment of deprivation and healthcare non-take up using appropriate questionnaires at the time of PAP therapy initiation. The consideration of societal topics should be better addressed in the education of sleep and respiratory physicians. Additionally, studies are needed to investigate the impact of health policy interventions on PAP adherence, as has been done for medications in vulnerable populations (34).

The inclusion of patients with a variety of respiratory diseases requiring PAP therapies is another originality of our study. The subsets of patients with OSAS and CFR had different patterns of types and reasons for forgoing healthcare. This reflects both different socio-economic circumstances and different health insurance coverage for the respective underlying disease. OSAS patients, with public system coverage limited to 60%, declared forgoing healthcare mainly for financial reasons whereas CRF patients (100% public coverage) explained forgoing healthcare mainly due to their lack of mobility.

For individuals with CRF the total reimbursement of healthcare costs by the French state, potentially makes it possible to totally eliminate financial barriers to healthcare access. However, the physical and psychological disabilities of CRF have repercussions leading to a deterioration in quality of life and loss of autonomy (difficulty to move about and/or the need of assistance) (35). This underlines the need for tailored solutions with an extension of public coverage to a subset of OSAS cases and greater use of telemedicine to preserve the continuity of care for CRF patients (36, 37).

In OSAS patients having only partial (60%) cover, a clear renunciation of dental and ophthalmic care was found. This has recently been addressed in France by the implementation of universal full reimbursement ("Rest à charge 0" [Zero cost to patient]) of basic dental care and glasses.

While our study is unique, it also has limitations. The main one being that we included patients treated with PAP therapies for at least 1 year whereas the mean duration of PAP treatment exceeds 8 years. This restricted the subgroup of non-adherent patients and potentially the power of the study to demonstrate an even greater effect of health care non-take-up on adherence. Further studies are needed to investigate the impact of health care non-take-up on initial PAP refusal and early PAP termination. Secondly, the present results did not consider comorbidities and polypharmacy that may be associated with healthcare non-take-up and PAP adherence (38).

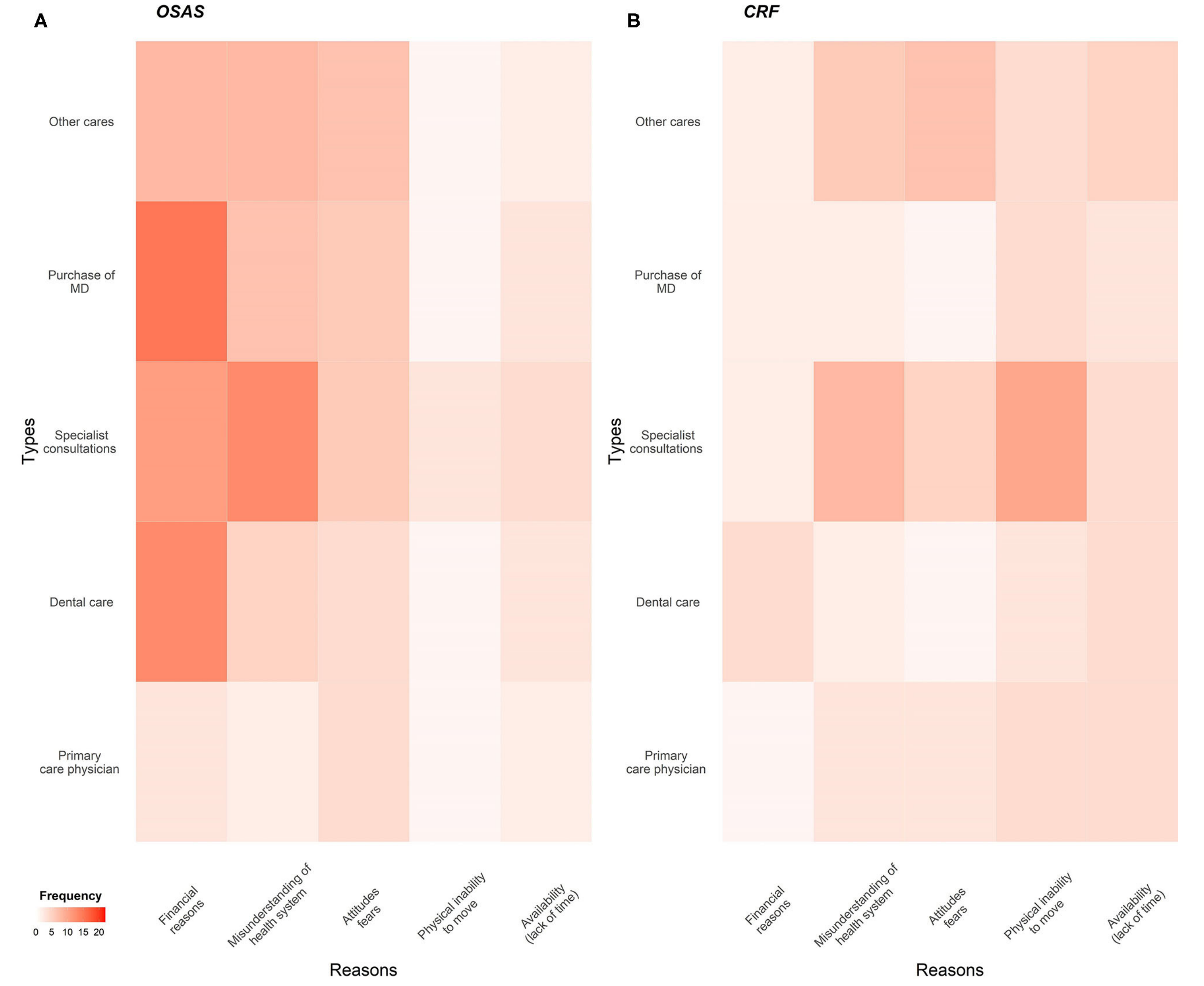


FIGURE 2 | Heatmap displaying the types and reasons of healthcare non-take up. **(A)** OSAS patients; **(B)** CRF patients. MD, medical device.

Finally, our study allowed to compare the healthcare non-take-up profile between the CFR and OSAS population using an explanatory exploratory approach. The aim was to obtain assumption for further research and not to provide conclusions based on a study which was not designed for this purpose. In conclusion, our study provides unique data indicating how the quality of care in PAP therapies could be improved and the design of interventional studies tailored to types and reasons for forgoing healthcare.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by French Ethics Committee Ile de France II. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

J-LP, MJ-J, SB and J-CB designed the study. ND collected the data. ND and SB carried out the statistical analyses and produced the figures. ND, J-CB, SB, AF, RT, and HR interpreted the data. ND, J-CB, SB and AF wrote the manuscript. RT, HR, MJ-J and J-LP revised the manuscript. All authors approved the version to be submitted for publication and took responsibility for the integrity of the work as a whole.

FUNDING

J-LP, SB, ND, HR, MJ-J, and RT were supported by the French National Research Agency in the framework of the Investissements d'avenir program (ANR-15-IDEX-02) and the e-health and integrated care and trajectories medicine and MIAI artificial intelligence Chairs of excellence from the Grenoble Alpes University Foundation (ANR-19-P3IA-0003). This work has been partially supported by the Mutualia private insurance company. AF was employed by Grenoble Alpes University hospital.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.713313/full#supplementary-material

REFERENCES

1. Lévy P, Kohler M, McNicholas WT, Barbé F, McEvoy RD, Somers VK, et al. Obstructive sleep apnoea syndrome. *Nat Rev Dis Primers.* (2015) 1:15015. doi: 10.1038/nrdp.2015.24
2. Borel J-C, Burel B, Tamisier R, Dias-Domingos S, Baguet J-P, Levy P, et al. Comorbidities and mortality in hypercapnic obese under domiciliary noninvasive ventilation. *PLoS ONE.* (2013) 8:e52006. doi: 10.1371/journal.pone.0052006
3. Adler D, Bailly S, Benmerad M. Clinical presentation and comorbidities of obstructive sleep apnea-COPD overlap syndrome. *PLoS ONE.* (2020) 15:e0235331. doi: 10.1371/journal.pone.0235331
4. Sullivan ColinE, Berthon-Jones M, Issa FaiqG, Eves L. Reversal of obstructive sleep apnoea by continuous positive airway pressure applied through the nares. *Lancet.* (1981) 317:862–5. doi: 10.1016/S0140-6736(81)92140-1
5. Wimms AJ, Kelly JL, Turnbull CD. Continuous positive airway pressure versus standard care for the treatment of people with mild obstructive sleep apnoea (MERGE): a multicentre, randomised controlled trial. *Lancet Respir Med.* (2020) 8:349–58. doi: 10.1016/S2213-2600(19)30402-3
6. McMillan A, Bratton DJ, Faria R. Continuous positive airway pressure in older people with obstructive sleep apnoea syndrome (PREDICT): a 12-month, multicentre, randomised trial. *Lancet Respir Med.* (2014) 2:804–12. doi: 10.1016/S2213-2600(14)70172-9
7. Masa JF, Mokhlesi B, Benítez I. Long-term clinical effectiveness of continuous positive airway pressure therapy versus non-invasive ventilation therapy in patients with obesity hypoventilation syndrome: a multicentre, open-label, randomised controlled trial. *Lancet.* (2019) 393:1721–32. doi: 10.1016/S0140-6736(18)32978-7
8. McEvoy RD, Antic NA, Heeley E, Luo Y, Ou Q, Zhang X, et al. CPAP for prevention of cardiovascular events in obstructive sleep apnea. *N Engl J Med.* (2016) 375:919–31. doi: 10.1056/NEJMoa1606599
9. Weaver TE, Grunstein RR. Adherence to continuous positive airway pressure therapy: the challenge to effective treatment. *Proc Am Thorac Soc.* (2008) 5:173–8. doi: 10.1513/pats.200708-119MG
10. Rotenberg BW, Murariu D, Pang KP. Trends in CPAP adherence over twenty years of data collection: a flattened curve. *J Otolaryngol Head Neck Surg.* (2016) 45:1–9. doi: 10.1186/s40463-016-0156-0
11. Gentina T, Bailly S, Jounieaux F, Verkindre C, Broussier P-M, Guffroy D, et al. Marital quality, partner's engagement and continuous positive airway pressure adherence in obstructive sleep apnea. *Sleep Med.* (2019) 55:56–61. doi: 10.1016/j.sleep.2018.12.009
12. Mendelson M, Gentina T, Gentina E, Tamisier R, Pépin J-L, Bailly S. Multidimensional evaluation of Continuous Positive Airway Pressure (CPAP) treatment for sleep apnea in different clusters of couples. *JCM.* (2020) 9:1658. doi: 10.3390/jcm9061658
13. Billings ME, Auckley D, Benca R, Foldvary-Schaefer N, Iber C, Redline S, et al. Race and residential socioeconomics as predictors of CPAP adherence. *Sleep.* (2011) 34:1653–8. doi: 10.5665/sleep.1428
14. Ando H, Williams C, Angus RM, Thornton EW, Chakrabarti B, Cousins R, et al. Why don't they accept non-invasive ventilation? Insight into the interpersonal perspectives of patients with motor neurone disease. *Br J Health Psychol.* (2015) 20:341–59. doi: 10.1111/bjhp.12104
15. Baron KG, Berg CA, Czajkowski LA, Smith TW, Gunn HE, Jones CR. Self-efficacy contributes to individual differences in subjective improvements using CPAP. *Sleep Breath.* (2011) 15:599–606. doi: 10.1007/s11325-010-0409-5
16. Villar I, Izuel M, Carrizo S, Vicente E, Marin JM. Medication adherence and persistence in severe obstructive sleep apnea. *Sleep.* (2009) 32:623–8. doi: 10.1093/sleep/32.5.623
17. Thornton CS, Tsai WH, Santana MJ, Penz ED, Flemons WW, Fraser KL, et al. Effects of wait times on treatment adherence and clinical outcomes in patients with severe sleep-disordered breathing: a secondary analysis of a noninferiority randomized clinical trial. *JAMA Netw Open.* (2020) 3:e203088. doi: 10.1001/jamanetworkopen.2020.3088
18. Platt AB, Kuna ST, Field SH, Chen Z, Gupta R, Roche DF, et al. Adherence to sleep apnea therapy and use of lipid-lowering drugs. *Chest.* (2010) 137:102–8. doi: 10.1378/chest.09-0842
19. Shi L, Stevens GD. Vulnerability and unmet health care needs: the influence of multiple risk factors. *J Gen Intern Med.* (2005) 20:148–54. doi: 10.1111/j.1525-1497.2005.40136.x
20. Lucevic A, Péntek M, Kringos D, Klazinga N, Gulácsi L, Brito Fernandes Ó, et al. Unmet medical needs in ambulatory care in Hungary: forgone visits and medications from a representative population survey. *Eur J Health Econ.* (2019) 20:71–8. doi: 10.1007/s10198-019-01063-0
21. Bazin F, Parizot I, Chauvin P. Déterminants psychosociaux du renoncement aux soins pour raisons financières dans cinq zones urbaines sensibles de la Région parisienne en 2001 [Psychosocial determinants of cessation of care for financial reasons in five sensitive urban areas of the Paris region in 2001. *Sci Soc Santé.* (2006) 24:11–32. doi: 10.3917/sss.243.0011
22. Revil H. Identifier les facteurs explicatifs du renoncement aux soins pour appréhender les différentes dimensions de l'accessibilité sanitaire [Identification of the factors explaining the renunciation of care to understand the different dimensions of health accessibility. *Regards.* (2018) 53:29–41. doi: 10.3917/regar.053.0029
23. Revil H, Daabek N, Bailly S. Synthèse descriptive des données du baromètre du renoncement aux soins (brs) [descriptive analysis of the healthcare non take-up barometer (brs).]. *Métropôle Odenore*. (2019) 41. Available online at: https://collectifhandicap54.files.wordpress.com/2019/06/synthese_analyses_descriptives_brs_-_v2_-_odenore_-_hp2.pdf
24. Bihan H, Laurent S, Sass C. Association among individual deprivation, glycemic control, and diabetes complications: the EPICES score. *Diabetes Care.* (2005) 28:2680–5. doi: 10.2337/diacare.28.11.2680
25. Labbe E, Blanquet M, Gerbaud L. A new reliable index to measure individual deprivation: the EPICES score. *Eur J Public Health.* (2015) 25:604–9. doi: 10.1093/eurpub/cku231
26. Grzymala-Busse JW, Hu M. A comparison of several approaches to missing attribute values in data mining. In: Ziarko W, Yao Y, éditors. *Rough Sets and Current Trends in Computing*. Berlin; Heidelberg: Springer Berlin Heidelberg (2001) p. 378–85.
27. Zhao Y. Statistical inference for missing data mechanisms. *Stat Med.* (2020) 39:4325–33. doi: 10.1002/sim.8727
28. Nadal N, Batlle J, Barbé F. Predictors of CPAP compliance in different clinical settings: primary care versus sleep unit. *Sleep Breath.* (2018) 22:157–63. doi: 10.1007/s11325-017-1549-7
29. Gagnadoux F, Le Vaillant M, Goupil F. Influence of marital status and employment status on long-term adherence with continuous

positive airway pressure in sleep apnea patients. *PLoS ONE.* (2011) 6:e22503. doi: 10.1371/journal.pone.0022503

30. Mehrtash M, Bakker JP, Ayas N. Predictors of continuous positive airway pressure adherence in patients with obstructive sleep apnea. *Lung.* (2019) 197:115–21. doi: 10.1007/s00408-018-00193-1
31. Bakker JP, O'Keeffe KM, Neill AM, Campbell AJ. Ethnic disparities in CPAP adherence in New Zealand: effects of socioeconomic status, health literacy and self-efficacy. *Sleep.* (2011) 34:1595–603. doi: 10.5665/sleep.1404
32. Bakker JP, Weaver TE, Parthasarathy S, Aloia MS. Adherence to CPAP: what should we be aiming for, and how can we get there? *Chest.* (2019) 155:1272–87. doi: 10.1016/j.chest.2019.01.012
33. Engleman HM, Wild MR. Improving CPAP use by patients with the sleep apnoea/hypopnoea syndrome (SAHS). *Sleep Med Rev*. (2003) 7:81–99. doi: 10.1053/smrv.2001.0197
34. Klein K, Bernachea MP, Irribarren S, Gibbons L, Chirico C, Rubinstein F. Evaluation of a social protection policy on tuberculosis treatment outcomes: a prospective cohort study. *PLOS Med.* (2019) 16:e1002788. doi: 10.1371/journal.pmed.1002788
35. Borel J-C, Borel A-L, Monneret D, Tamisier R, Levy P, Pepin J-L. Obesity hypoventilation syndrome: from sleep-disordered breathing to systemic comorbidities and the need to offer combined treatment strategies: obesity hypoventilation syndrome. *Respirology.* (2012) 17:601–10. doi: 10.1111/j.1440-1843.2011.02106.x
36. Duiverman ML, Vonk JM, Bladder G, van Melle JP, Nieuwenhuis J, Hazenberg A, et al. Home initiation of chronic non-invasive ventilation in COPD patients with chronic hypercapnic respiratory failure: a randomised controlled trial. *Thorax.* (2020) 75:244–52. doi: 10.1136/thoraxjnl-2019-213303
37. Barbosa MT, Sousa CS, Morais-Almeida M, Simões MJ, Mendes P. Telemedicine in COPD: an overview by topics. *COPD J Chron Obstruct Pulm Dis.* (2020) 17:601–17. doi: 10.1080/15412555.2020.1815182
38. Catho H, Guigard S, Toffart A-C, Frey G, Chollier T, Brichon P-Y, et al. What are the barriers to the completion of a home-based rehabilitation programme for patients awaiting surgery for lung cancer: a prospective observational study. *BMJ Open*. (2021) 11:e041907. doi: 10.1136/bmjopen-2020-041907

Conflict of Interest: J-CB and ND are employees of AGIR a dom. Homecare charity.

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.

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

SYSTEMATIC REVIEW
published: 24 August 2021
doi: 10.3389/fpubh.2021.710575

The Holistic Health Status of Chinese Homosexual and Bisexual Adults: A Scoping Review

Chanchan Wu, Edmond Pui Hang Choi and Pui Hing Chau*

School of Nursing, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China

Background: Same-sex marriage is currently not legalized in China, despite the considerably large number of homosexual and bisexual Chinese populations. At the same time, their holistic health status remains unclear. This is the first scoping review conducted to comprehensively examine all the available literature and map existing evidence on the holistic health of homosexual and bisexual Chinese.

Methods: This scoping review used the framework of Arksey and O'Malley and followed the Preferred Reporting Items for Systematic Review and Meta-Analysis extension for scoping reviews (PRISMA-ScR). A comprehensive search strategy was carried out across 20 English (EN) and Chinese (both traditional and simplified) electronic databases from January 1, 2001, to May 31, 2020. Two reviewers conducted the reference screening and study selection independently and consulted a third senior reviewer whenever a consensus must be achieved. Data extraction was conducted using a structured data form based on the Cochrane template, after which a narrative synthesis of the findings was performed.

Results: A total of 2,879 references were included in the final analysis, with 2,478 research articles, 167 reviews, and 234 theses. Regarding the study populations, the vast majority of studies centered on men only (96.46%), especially men who have sex with men (MSM). Only 1.32% of the studies targeted female sexual minorities. The geographical distribution of all research sites was uneven, with most of them being conducted in mainland China (95.96%), followed by Hong Kong (2.05%), Taiwan (2.02%), and Macau (0.06%). Regarding the specific study focus in terms of the health domain, around half of the studies (45.93%) focused on sexual health only, and an additional quarter of the studies (24.15%) investigated both sexual health and social well-being. Meanwhile, the studies focusing on mental health only accounted for approximately 15% of the total.

Conclusions: This scoping review revealed that previous research focused more on male than female sexual minorities, on disease-centered surveys than person-centered interventions, and investigations on negative health conditions than positive

Edited by:
Yi Guo,
University of Florida, United States

Reviewed by:
David Candon,
Nottingham Trent University,
United Kingdom
Dongyu Zhang,
University of Florida, United States

****Correspondence:***
Edmond Pui Hang Choi
h0714919@connect.hku.hk

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: *16 May 2021*
Accepted: *26 July 2021*
Published: *24 August 2021*

Citation:
Wu C, Choi EPH and Chau PH (2021)
The Holistic Health Status of Chinese
Homosexual and Bisexual Adults: A
Scoping Review.
Front. Public Health 9:710575.
doi: 10.3389/fpubh.2021.710575

health promotion. Therefore, investigations centered on the female sexual minorities and corresponding person-centered interventions are highly needed.

Review Registration: The protocol of this review has been registered within Open Science Framework (https://osf.io/82r7z) on April 27, 2020.

Keywords: bisexual, Chinese, health, homosexual, men who have men with men

INTRODUCTION

Homosexuality and bisexuality have long existed worldwide, but the recognition of same-sex marriage in many countries has only gradually occurred in recent years. In comparison, under the heavy influence of Confucianism, it has always been a traditional obligation of Chinese adults to bring offspring to the family. Thus, homosexuality is widely rejected by the Chinese and is considered not only a threat to the family but also a threat to society. For nearly 20 years (1979–1997), sex between men was considered illegal and criminalized as "hooliganism" (sodomy) in China until it was eliminated in 1997 (1). In 2001, homosexuality was no longer classified as a pathology under the Chinese Classification of Mental Disorders (2), marking a historical turning point in the progress of homosexuality in China. Despite such developments, in contemporary China, the mainland government still recognizes neither legal same-sex marriage nor civil unions, and the situations in Hong Kong and Macau are similar. In contrast, same-sex marriage has been legalized in Taiwan since 2019, even though the law was enacted outside the Civil Code (3). In general, homosexual and bisexual Chinese from the above Cross-Straits Four-Regions have experienced similar cultural and policy backgrounds in the past two decades. Thus, research on this population is of historical significance in such an era.

Compared to some Western countries where homosexuals could either cohabit or enter legal same-sex marriages when available (4), most Chinese homosexuals can only choose to either stay single or develop hidden relationships "in the closet," and even fewer bisexual Chinese choose to disclose their sexual orientation (5). In recent decades, as research on these populations has gradually increased, some widely used behavioral concepts have been proposed by researchers to describe similar population groups regardless of their sexual identity (6–8), namely, men who have sex with men (MSM) and women who have sex with women (WSW). MSM/WSW populations may not only be involved in homosexual behaviors but also in bisexual behaviors. For instance, approximately 40% of MSM acknowledged being men who have sex with both men and women (MSMW) according to a national Chinese survey (9), while more than a quarter of the MSM claimed their sexual orientation to be bisexual (10). These indicate that research should not only focus on gays and lesbians from the perspective of sexual orientation but also on MSM and WSW from the perspective of sexual behavior.

According to the widely used definition of "health" introduced by the WHO, "*Health is a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity"* (11). This definition explains the concept of "*holistic health*," which is a broad conceptualization of health that encompasses various dimensions, including complete physical health, mental health, and social well-being. Notably, sexual health is the most prevalent health domain in studies targeting homosexual or bisexual Chinese. Specifically, such research has always focused on certain diseases, such as AIDS and the related topic of HIV prevention, or other Sexually Transmitted Infections (STI) and unsafe sexual behaviors. Specifically, the overall national prevalence of HIV among MSM was 5.7% from 2001 to 2018 (12), while that for syphilis for the same period was 11.8% (13). Thus far, only a few studies on Chinese WSW have been conducted compared to studies on MSM. According to the only domestic study investigating 224 Beijing WSW, 15.8% of this population were infected with gonorrhea though no HIV-positive cases were detected (14). Furthermore, about half of the WSW reported bleeding during or after sex, and many of them reported that they had experienced engaging in different kinds of high-risk sexual behaviors (15). All of these findings indicate that their worrying sexual health concerns may require further attention.

Although sexual minorities in China still face many significant psychosocial difficulties that are yet to be addressed, the most common of which is long-standing social discrimination or stigma based on sexual orientation (16), there are relatively a few studies on mental health and social well-being in this population compared to their counterparts in Western countries. In particular, both lesbians and gays in China reported feeling stressed and helpless in the face of expectations from society and their parents (17, 18). Moreover, both gay and bisexual Chinese men reported having suffered internalized homophobia (19, 20), which was found to be positively correlated with loneliness and negatively correlated with lower self-evaluation (19). At the same time, most psychosocial studies targeting sexual minorities were carried out in MSM populations from a behavioral perspective, with both qualitative and quantitative studies indicating that

Abbreviations: WHO, World Health Organization; MSM, Men/Males who have Sex with Men/Males; WSW, Women who have Sex with Women; MSMW, Men who have Sex with both Men and Women; HIV, Human Immunodeficiency Virus; AIDS, Acquired Immune Deficiency Syndrome; STI, Sexually Transmitted Infections; JBI, Joanna Briggs Institute; PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement; PRISMA-ScR, Preferred Reporting Items for Systematic Review and Meta-Analysis Extension for Scoping Reviews; PRISMA-P, Preferred Reporting Items for Systematic Reviews and Meta-analysis Protocols; PCC, Population Concept Context; MeSH, Medical Subject Headings; CMeSH, Chinese translation of Medical Subject Headings; CDC, Center for Disease Control and prevention; MU, Medical University; EN, English; CN, Chinese; SC, Simplified Chinese; TC, Traditional Chinese; NGO, Non-Governmental Organization; LGBTQ, Lesbian, Gay, Bisexual, Transgender or Queer people; GDP, Gross Domestic Product; CNY, Chinese Yuan.

MSM often experienced homosexual stigma (21, 22) and HIV-related stigma (23). Furthermore, MSM in China reported significantly higher levels of internalized homophobia compared with those from outside China (24), with a mean score of 2.04 vs. 1.77, as measured by the 4-point Likert Internalized Homophobia Scale. Chinese MSM also reported experiencing a high prevalence of other mental health issues, such as loneliness (35.5%) (25), moderate-to-severe symptoms of depression (26.8–50.9%) (25–27), and anxiety (26.0–36.4%) (26, 27). Even worse are the high rates of suicide ideation and suicidal behavior. A study reported that the specific suicide ideation rates were 31% among gay and bisexual men in Taiwan (28) and 26% among MSM in nine cities of mainland China (29). Meanwhile, over 12% of MSM actually attempted suicide (29), which is several times higher than that of the normal adult men (30). All these indicate that many Chinese homosexuals and bisexuals suffer from poor mental health, which could lead to self-loathing and negative effects on their self-identity (31). These mental health issues could also have further negative effects on their sexual health and/or social well-being (32).

Currently, in China, there are no census data on either the homosexual or bisexual population. This has caused concerns, given that China is the most populous country in the world, which means that the number of sexual minorities could be proportionately higher compared with those in other countries. Furthermore, with the wide use of the Internet and increasing social tolerance, sexual minority groups are no longer as hidden as before; hence, their social and health needs should be understood and addressed. Nevertheless, prevailing public attitudes toward homosexuality remain negative. For example, over half (58.4%) of the MSM in Hong Kong reported experiencing public discrimination (33), similar to the situation in mainland China (34).

To date, there have been studies on homosexual and bisexual Chinese, especially within the MSM population. These studies mainly centered on STI-/HIV-related prevalence or prevention attempts (12, 35–43), or focused on human rights and the legalization of same-sex marriage, as conducted in the fields of sociology, anthropology, law, and psychology (44–46). However, other aspects of health and well-being have yet to be fully investigated. At the same time, there are even fewer studies on female sexual minorities compared with male ones (45), thus highlighting the need for further academic attention.

In summary, the current health-related research targeting Chinese homosexual and bisexual adults seem to be unbalanced from the perspective of the gender population and health domains, indicating the essential need for further scientific review evidence. So far, there is no systematically reviewed evidence available or ongoing review either in English (EN) or Chinese on the holistic health of homosexual and bisexual people within the Chinese context. In addition, the current evidence is difficult to summarize due to variations in the types of studies and the less precisely defined subjects and research variables.

In relation to the above, a scoping review, which is a type of systematic review, can be used to comprehensively map the known information about a topic based on the available information and then identify the potential gaps in the literature, thus facilitating an assessment of the state of knowledge about the specific topic (47–50). In 2018, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Statement was extended to Scoping Reviews (PRISMA-ScR) (51). Therefore, the current review was conducted as a systematic scoping review, following the PRISMA-ScR checklist (**Supplementary Material 1**). This review aims to comprehensively examine the literature to explore the breadth of current knowledge relating to the holistic health of homosexual and bisexual Chinese, identify potential knowledge gaps, and then inform future in-depth research on how to improve the health of this particular population.

METHODS

This study used the scoping review framework developed by Arksey and O'Malley (2005) (47) and further updated as recommended by the Joanna Briggs Institute (JBI) (50). This review was also conducted in accordance with a priori protocol currently under review (52), including five stages: 1) identification of the review questions, 2) identification of relevant studies, 3) study selection, 4) data extraction, and 5) summarization and reporting of the results.

Stage 1: Identification of the Review Questions

The "PCC" mnemonic, representing Population-Concept-Context, is recommended by the JBI (53) as a guide to construct clear research questions for a scoping review. Correspondingly, in this review, "Population" refers to all Chinese homosexual and bisexual adults living in mainland China, Hong Kong, Macau, or Taiwan. Chinese sexual minorities who were born or living abroad were excluded due to policy and cultural differences. Furthermore, in this review, "Concept" refers to holistic health, a broad conceptualization of health defined by the WHO (11) that encompasses varied dimensions of health, including the complete physical, mental, and social well-being of an individual. This review, therefore, targeted all these health-related aspects. Finally, "Context" refers to the locations of study settings. Thus, following the overarching review question: "What is the holistic health status of Chinese homosexual and bisexual adults?" some detailed review questions are as follows: 1) "What health-related variables have been investigated about homosexual and bisexual Chinese?" 2) "What types of research have been conducted and which disciplines were most involved in carrying out studies targeting this population?" and 3) "What are the differences among the sample populations in terms of sexual orientation (between homosexuals and bisexuals) and gender (between male and female minorities)?"

Stage 2: Identification of Relevant Studies and Search Strategy

The eligibility criteria for this scoping review using the PCC framework are shown in the a priori protocol. In addition to homosexual and bisexual people from the perspective of sexual orientation, MSM/MSMW and WSW groups were also included in terms of behavioral categories. Regarding the concept of holistic health, after reviewing health-related definitions (54–58), this review included both negative and positive variables related to mental health, physical (sexual) health, and social well-being. In terms of context, this review included studies conducted in all regions of China, including mainland cities, Hong Kong, Macau, and Taiwan. Regarding the inclusion criteria of the study types, all original studies using qualitative, quantitative, or mixed methods, reviews and published dissertations were included. Meanwhile, study protocols or blogs, book chapters, conference abstracts, research letters, editorial notes or commentaries were excluded.

For a comprehensive literature search, research articles, reviews, and theses published in 2001 or later were searched. This is because homosexuality has no longer been regarded as a mental illness in China since 2001 (2). All relevant databases in both EN and CN languages related to health care, psychology, and social science were searched. Specifically, 20 databases were searched, including 12 EN language databases (PubMed, Web of Science, CINAHL Plus, ScienceDirect, Social Work Abstracts, APA PsycInfo, etc.), four Simplified Chinese (SC) databases (China National Knowledge Infrastructure-CNKI, China Biological Medicine Database-SinoMed, etc.), and four traditional Chinese (TC) databases (Index to Taiwan periodical literature system, National Digital Library of Theses, Dissertation in Taiwan, etc.).

The search strategy for this review was adapted from the Peer Review of Electronic Search Strategies (PRESS) Evidence-Based Checklist (59). Handsearching was also used as a supplementary method, although unpublished documents were excluded from this review due to limited resources. Pilot searching was conducted in both EN (PubMed) and CN (SinoMed) language databases before the formal search, with the aim of identifying all relevant keywords or subject headings before finalizing the search strategy. The Medical Subject Headings (MeSH terms) and corresponding Chinese translation MeSH terms (CMeSH terms) are summarized in the **Supplemental Material 2**. The final search of the above mentioned databases was conducted throughout May 2020 and updated on May 31, 2020. Some updated studies were further reviewed on a monthly basis by checking the available email alerts (**Supplementary Material 3**).

Stage 3: Study Selection

After searching, the identified records were exported to and managed by EndNote X9 (for EN and TC literature) and NoteExpress (for SC literature). Both software programs could automatically identify duplicate records. Then, two reviewers independently performed the study selection, which included title screening, abstract screening, and full-text screening according to the JBI guidelines (51, 53). Specifically, the title and abstract screening were carried out simultaneously referring to the PCC criterion, and the number of excluded references in each step was recorded and compared. In case of inconsistencies, the two reviewers discussed until data consistency was achieved. Afterward, the full texts of all potentially eligible references were retrieved for further screening, and disagreements on the study selection were resolved by a discussion between the two main reviewers and consultation with a third senior researcher. Finally, those references that were excluded during full-text screening were recorded following specific exclusion reasons, in line with the PCC framework (**Supplementary Material 4**).

Stage 4: Data Extraction

Initially, a data extraction template was developed based on the Cochrane Data Extraction Template (60), after confirming all essential variables and key information to extract. Then, after conducting the pilot extraction using both qualitative and quantitative eligible studies, a revised and detailed version of the data extraction form was used. The data extraction was also conducted independently by two reviewers with regular discussion, and then continuously updated in an iterative manner. Specifically, the data extracted included specific details about the "Reference Characteristics," "Study Characteristics," and also "PCC-related Information," including authors and affiliations, year of publication, study design, population and sample size, study settings, health domains, and corresponding findings.

Stage 5: Summarization and Reporting of the Results

All retrieved information from the data extraction form was documented in Microsoft Excel, and narrative synthesis was used. The quantitative findings were descriptively summarized in the form of tables using frequencies and percentages. Next, the summarization of qualitative evidence including the collaboration network and co-word analysis was conducted *via* social network analysis by UCINET (61) and then visualized using the NetDraw program (62). The critical appraisal process is not necessary for a scoping review (48, 49), and it is also not feasible to evaluate the quality of each included reference (**Supplementary Material 5**). Nevertheless, this review tried to perform a quality evaluation of the journals in which all the included articles were published (**Supplementary Material 6**).

RESULTS

The final database search from January 1, 2001, to May 31, 2020, yielded a total of 14,811 references. After removing 4,227 duplicates, 10,584 references remained for further screening. Specifically, 5,655 were excluded after title screening, and 1,328 were excluded after abstract screening as they did not meet the PCC eligibility criteria. Of the 3,601 remaining records for full-text screening, 727 were excluded due to specific reasons (**Supplementary Material 4**). Finally, a total of 2,879 references incorporating an additional five references obtained through the manual search were included in the final analysis (**Supplementary Material 5**), including 2,645 articles and 234 theses. From the perspective of the publishing language, there were 708 EN references, 2,151 SC records, and 20 TC records. The study screening and selection were conducted following the PRISMA flow diagram (63). The detailed process and specific results are presented in **Figure 1**.

Time Trends of Publications and Corresponding Studies

Among all the 2,879 included references, the overall number of publications gradually increased and then stabilized in the

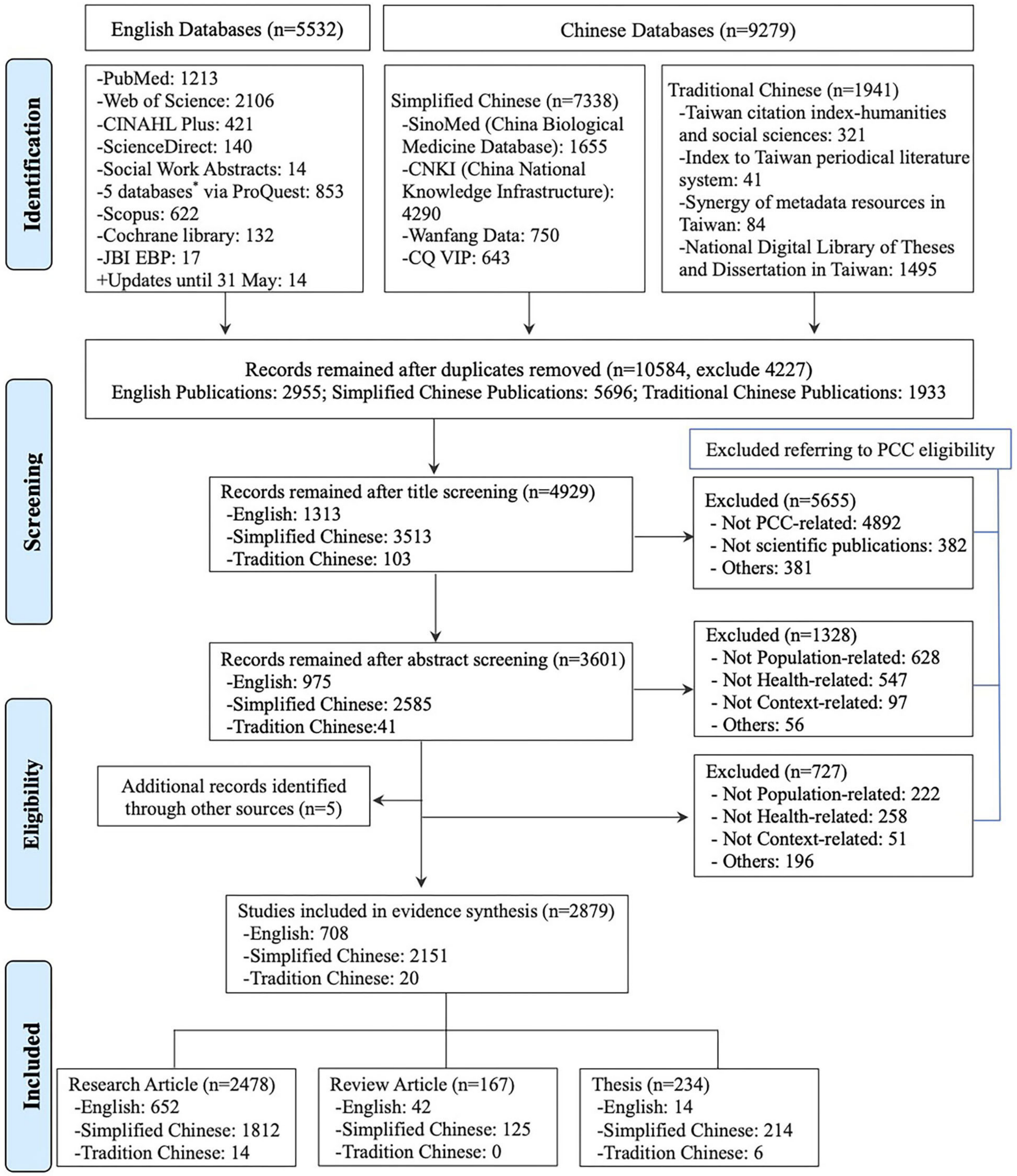

Identification
Screening
Eligibility
Included
English Databases (n=5532)
Chinese Databases (n=9279)
-PubMed: 1213
-Web of Science: 2106
-CINAHL Plus: 421
-ScienceDirect: 140
-Social Work Abstracts: 14
-5 databases* via ProQuest: 853
-Scopus: 622
-Cochrane library: 132
-JBI EBP: 17
+Updates until 31 May: 14
Simplified Chinese (n=7338)
-SinoMed (China Biological Medicine Database): 1655
-CNKI (China National Knowledge Infrastructure): 4290
-Wanfang Data: 750
-CQ VIP: 643
Traditional Chinese (n=1941)
-Taiwan citation index-humanities and social sciences: 321
-Index to Taiwan periodical literature system: 41
-Synergy of metadata resources in Taiwan: 84
-National Digital Library of Theses and Dissertation in Taiwan: 1495
Records remained after duplicates removed (n=10584, exclude 4227)
English Publications: 2955; Simplified Chinese Publications: 5696; Traditional Chinese Publications: 1933
Excluded referring to PCC eligibility
Records remained after title screening (n=4929)
-English: 1313
-Simplified Chinese: 3513
-Tradition Chinese: 103
Excluded (n=5655)
- Not PCC-related: 4892
- Not scientific publications: 382
- Others: 381
Records remained after abstract screening (n=3601)
-English: 975
-Simplified Chinese: 2585
-Tradition Chinese:41
Excluded (n=1328)
- Not Population-related: 628
- Not Health-related: 547
- Not Context-related: 97
- Others: 56
Additional records identified through other sources (n=5)
Excluded (n=727)
- Not Population-related: 222
- Not Health-related: 258
- Not Context-related: 51
- Others: 196
Studies included in evidence synthesis (n=2879)
-English: 708
-Simplified Chinese: 2151
-Tradition Chinese: 20
Research Article (n=2478)
-English: 652
-Simplified Chinese: 1812
-Tradition Chinese: 14
Review Article (n=167)
-English: 42
-Simplified Chinese: 125
-Tradition Chinese: 0
Thesis (n=234)
-English: 14
-Simplified Chinese: 214
-Tradition Chinese: 6

FIGURE 1 | Search results and study selection process referring to Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA). *: ProQuest searching of the 5 databases: ProQuest Dissertations & Theses A&I; ProQuest Dissertations & Theses Global; APA PsycInfo; Sociological Abstracts; Social Services Abstracts.

past two decades (blue line in **Figure 2**). As this study only included references published before May 2020, the data in 2020 were dropped from both **Figures 2**, **3**, so as not to distract the trends. Regarding the trend in the number of studies started each year over time, it can be seen that the research conducted before 2010 increased year by year, while the following decade (2011–2019) showed a significant downward trend (orange line in **Figure 2**). Review articles first appeared

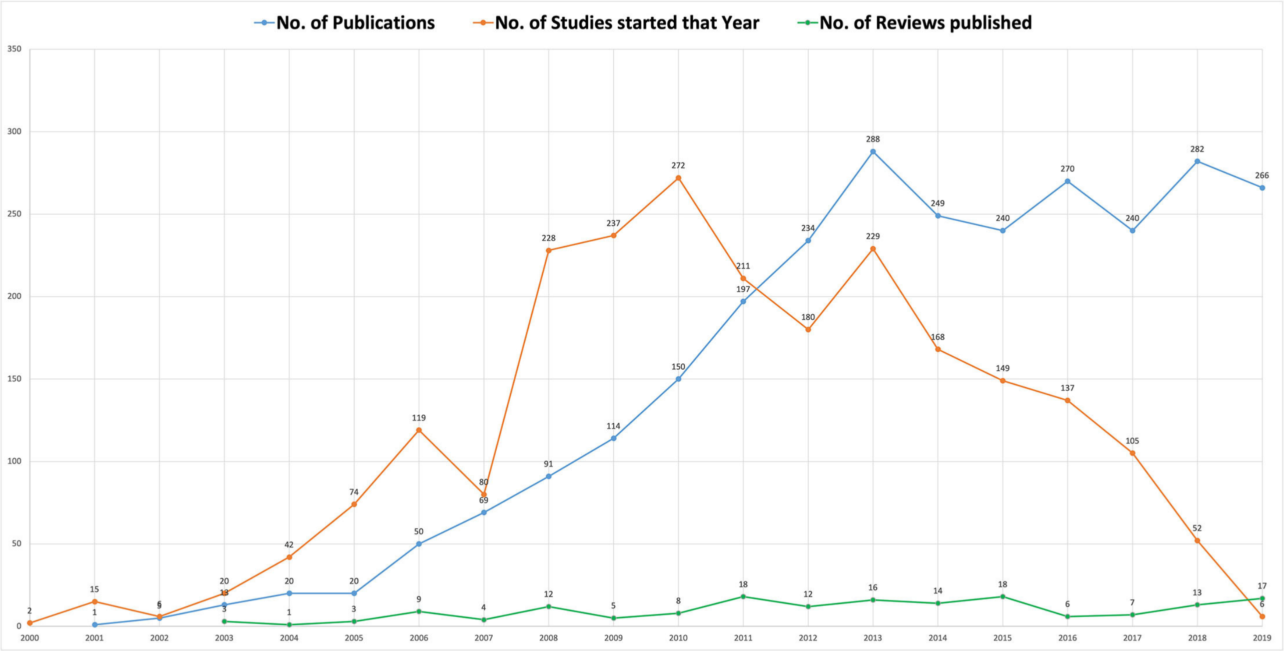
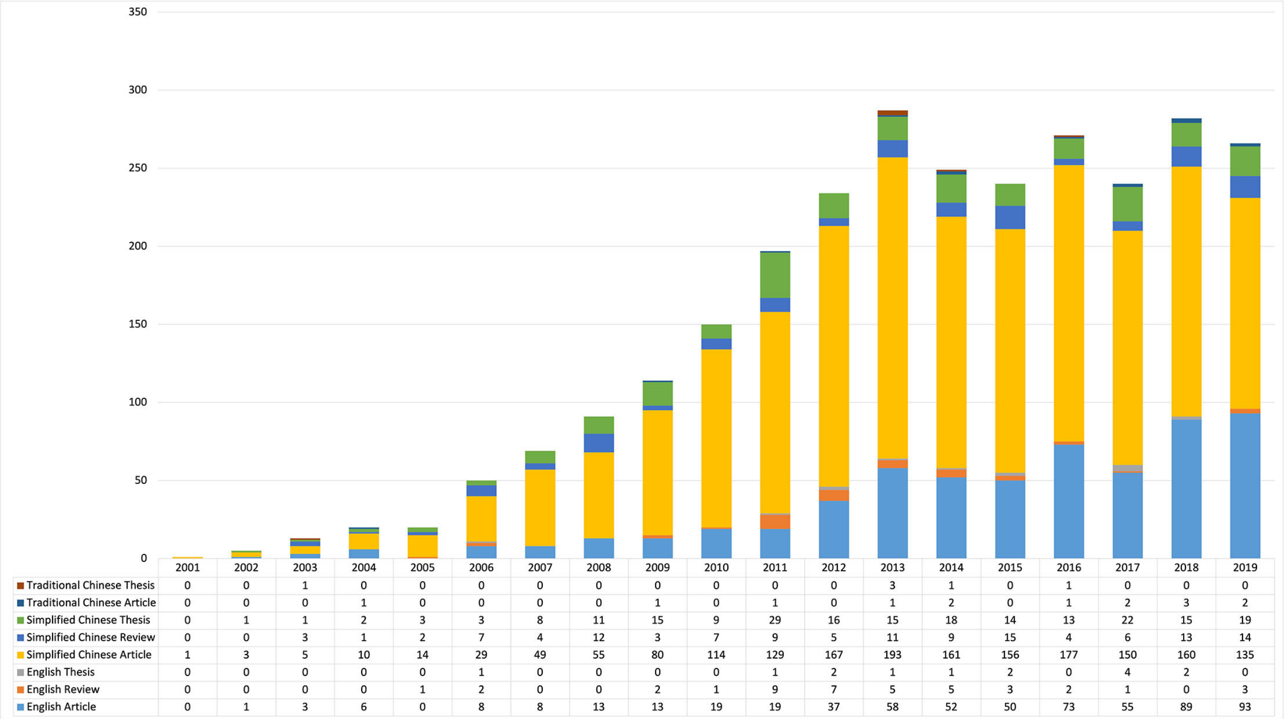


FIGURE 2 | Time trends of publications, corresponding studies, and reviews. Since this study only included references published before May 2020, the data in 2020 was dropped as it would be distracting the trends.



	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Traditional Chinese Thesis	0	0	1	0	0	0	0	0	0	0	0	0	3	1	0	1	0	0	0
Traditional Chinese Article	0	0	0	1	0	0	0	0	1	0	1	0	1	2	0	1	2	3	2
Simplified Chinese Thesis	0	1	1	2	3	3	8	11	15	9	29	16	15	18	14	13	22	15	19
Simplified Chinese Review	0	0	3	1	2	7	4	12	3	7	9	5	11	9	15	4	6	13	14
Simplified Chinese Article	1	3	5	10	14	29	49	55	80	114	129	167	193	161	156	177	150	160	135
English Thesis	0	0	0	0	0	1	0	0	0	0	1	2	1	1	2	0	4	2	0
English Review	0	0	0	0	1	2	0	0	2	1	9	7	5	5	3	2	1	0	3
English Article	0	1	3	6	0	8	8	13	13	19	19	37	58	52	50	73	55	89	93

FIGURE 3 | Publication distribution by study types per year (2001–2019). Since this study only included references published before May 2020, the data in 2020 was dropped as it would be distracting the trends.

in 2003, and around 10 reviews were conducted every year thereafter without an obvious trend of increase or decrease (green line in **Figure 2**).

Among the 2,879 included publications, 74.71% (2,151) were written in SC and 86.07% (2,478) were research articles. **Figure 3** shows the distribution of study types in both EN and CN publications. Among the 234 theses, some incorporated two or more independent studies (cross-sectional survey and following intervention study); thus a total of 276 identical studies were carried out.

Author Affiliations and Funding Information

Regarding the distribution of author affiliations, the majority of affiliations (89.41%) were located in mainland China, of which more than half (54.74%) were affiliated with the Center for Disease Control and Prevention (CDC), and around a third (33.53%) were affiliated with universities, mainly medical universities (MUs). As for institutional cooperation, over half of the studies (55.24%) were conducted by a single institution; and approximately one-third of the research (30.40%) was carried out through domestic cooperation, which was more

TABLE 1 | Institutional characteristics of 2,645 articles and 234 theses.

Article institutions ($n = 2,645$)	No. (%)	Thesis institutions ($n = 234$)	No. (%)
First author affiliation		**Degree**	
CDC	1,448 (54.74)	Master thesis	200 (85.47)
University	887 (33.53)	PhD thesis	34 (14.53)
Hospital	256 (9.68)	**Author affiliation**	
Scientific Institution	29 (1.10)	Mainland China	214 (91.45)
Health authority	10 (0.38)	Hong Kong	10 (4.27)
Others	15 (0.57)	Taiwan	6 (2.56)
First author affiliation location		Overseas	4 (1.71)
Mainland China	2365 (89.41)	**Universities**	
Overseas	144 (5.44)	中国疾病预防控制中心(China CDC)	28 (11.97)
Hong Kong	85 (3.21)	安徽医科大学(Anhui MU)	24 (10.26)
Taiwan	48 (1.81)	重庆医科大学(Chongqing MU)	15 (6.41)
Macau	2 (0.08)	山东大学(Shandong University)	11 (4.70)
Not available	1 (0.04)	The Chinese University of Hong Kong	10 (4.27)
First author profession[a]		第三军医大学(TMMU)	9 (3.85)
Public health	430 (63.70)	天津医科大学(Tianjin MU)	9 (3.85)
Psychology	43 (6.37)	河北医科大学(Hebei MU)	8 (3.42)
Global health	26 (3.85)	青岛大学(Qingdao University)	8 (3.42)
Nursing	18 (2.67)	吉林大学(Jilin University)	7 (2.99)
Sociology	18 (2.67)	昆明医科大学(Kunming MU)	6 (2.56)
Medicine	11 (1.63)	新疆医科大学(Xinjiang MU)	6 (2.56)
Psychiatry	10 (1.48)	中国医科大学(China MU)	6 (2.56)
Not available	15 (2.22)	山西医科大学(Shanxi MU)	5 (2.14)
Others	104 (15.41)	Others	82 (35.04)
If involved nursing profession		**Profession/Major**	
Yes	92 (3.48)	Epidemiology	112 (47.86)
No	2553 (96.52)	Public health	53 (22.65)
First author is nursing profession		Social Medicine	13 (5.56)
Yes	55 (2.08)	Psychology	10 (4.27)
No	2590 (97.92)	Dermatology	8 (3.42)
Institution cooperation		Sociology	6 (2.56)
International cooperation	380 (14.37)	Nursing	4 (1.71)
Domestic cooperation	804 (30.40)	Maternal, Child and Adolescent Health	4 (1.71)
Single institution	1461 (55.24)	Others	24 (10.25)
Funding sources[b]		**Funding sources**[b]	
Not funded/ Not available	1092 (37.93)	Global Funding	107 (3.72)
Mainland China	1056 (36.68)	Hong Kong	53 (1.84)
China and foreign funding	343 (11.91)	Taiwan	31 (1.08)
America	130 (4.52)	Funding from other organizations	67 (2.33)

[a] *Among 675 studies published in English (EN) (652 research papers and 23 studies from 14 theses).* [b] *Among all 2,879 publications.*

common in CN articles than in EN publications (29.38 vs. 1.02%, respectively). In contrast, the proportion of international cooperation among authors in EN publications exceeded that in CN publications (12.40 vs. 1.97%, respectively). Regarding the funding information, 36.68% (1,056/2,879), 1.84% (53), and 1.08% (31) of studies obtained funding from mainland China, Hong Kong, and Taiwan, respectively, while over one-third (37.93%) of studies were not funded. Furthermore, around 23% of the studies were funded by foreign organizations or global sources (**Table 1**), thereby revealing the global nature of these related studies.

Characteristics of the Included Reviews

Review articles accounted for the smallest proportion (167/2,645, 6.31%) of all articles, among which 25.15% (42/167) were published in EN and 74.85% (125/167) were published in SC. Most of the reviews centered only on the male population, especially MSM (151, 90.42%), while only four reviews focused on female sexual minorities (**Table 2**). These findings indicate the insufficient attention given to this population, thus highlighting the need for further research in this group. Notably, there was just one scoping review among all the included reviews, which summarized only the HIV prevalence and corresponding prevention intervention programs for MSM and transgender populations (64). Overall, the included reviews mostly focused on sexual health (86.83%), such as the incidence or prevalence of HIV/STI and related treatment. In comparison, less attention was given to mental health (7.19%) and social well-being (2.40%).

TABLE 2 | Characteristics of the included reviews.

Target population	English reviews	Chinese reviews	Total	%
Male only	41	113	154	92.22%
MSM	40	111	151	90.42%
Gay	/	2	2	1.20%
MSMW	1	/	1	0.60%
Female only	1	3	4	2.40%
Lesbian	1	1	2	1.20%
WSW	/	2	2	1.20%
Both gender	0	9	9	5.39%
Gay & Lesbian	/	6	6	3.59%
MSM & WSW	/	3	3	1.80%
Clinical people or not				
HIV-infected	1	4	5	2.99%
Non patient	41	121	162	97.01%
Review type				
General literature review	11	111	122	73.05%
Systematic Review (SR)	6	2	8	4.79%
SR and Meta-analysis	24	12	36	21.56%
Scoping review	1		1	0.60%
Review focus				
Sexual health	40	105	145	86.83%
Mental health	1	11	12	7.19%
Social well-being	1	3	4	2.40%
Overall review or others	/	6	6	3.59%

Characteristics of the Included Studies

There are 2,754 studies in all the included publications, including 2,478 research articles (1,812 SC, 14 TC, and 652 EN publications) and 276 studies from 234 theses. The majority of the studies (2,677, 97.20%) used the originally collected data, while the remaining ones were conducted as secondary analyses using previous data. Of all the original studies, nearly half (1,324, 48.46%) had sample sizes between 101 and 500; and 23.53% had sample sizes larger than 1,000, most of which involved series of cross-sectional surveys. Given that there were a certain number of published CN studies with unclear descriptions of the methodologies used, the detailed characteristics of the study design and sampling methods used data from the 675 published EN studies (652 research articles and 23 studies from 14 theses). Specifically, most of the studies (75.41%) were conducted as cross-sectional investigations, followed by qualitative studies (54, 8.00%); less than 1% were mixed-method studies (**Table 3**). Around a quarter of the studies used multiple sampling methods, followed by the Internet- and venue-based sampling methods.

Population Information

Among all the 2,879 publications, the vast majority of the research, whether original studies or reviews, focused on MSM (2,667, 92.64%). Specifically, in the 675 EN publications with detailed population descriptions, most average ages ranged between 20 and 35 years. There was just one study that targeted bisexual people only (65), while over 40 studies involved homosexuals only. In addition, in all studies that indicated specific sexual orientation, the proportion of homosexuals was greatly higher than that of bisexuals (**Table 4**). Nearly two-thirds of the studies collected data on the marital status of the sample populations (around 10–30% were married). Thus, the marital and living conditions of homosexuals and bisexuals, along with their spouses, are worthy of further exploration.

Context Information

Among all 2,712 research references (2,478 research articles and 234 theses), 2,704 provided clear setting information, including 3,310 study sites (both single and multiple sites), although the geographical distribution of all study sites was uneven. Specifically, when analyzing all study sites from publications in both languages, 95.96% of the studies were conducted in mainland China, followed by Hong Kong (2.05%), and Taiwan (2.02%); meanwhile, only two nationwide studies (66, 67) involved Macau (0.06%), and no original single-site study was conducted in Macau. Furthermore, when analyzing study sites in EN publications, similarly, the majority of studies (83.41%) were conducted in mainland China, followed by Hong Kong (8.89%) and Taiwan (6.52%). However, when considering the geographical area or population size, the results showed that more studies were conducted per unit area or population in Hong Kong, compared with Taiwan, the mainland, and Macau (in descending order). In all the research conducted in mainland China, study sites were mostly located in the east (841, 25.41%) and southwest (742, 22.42%). Three places

TABLE 3 | Study and sampling characteristics of included EN studies.

	Mixed method study	Qualitative study	Quantitative study	Total
Sampling method				
Multiple methods	1 (0.15)	12 (1.78)	160 (23.7)	173 (25.63)
Internet-based sampling	1 (0.15)	3 (0.44)	112 (16.59)	116 (17.19)
Venue-based sampling	/	5 (0.74)	97 (14.37)	102 (15.11)
Convenience sampling	/	4 (0.59)	82 (12.15)	86 (12.74)
Snowball sampling	2 (0.30)	9 (1.33)	65 (9.63)	76 (11.26)
Respondent Driven Sampling	2 (0.30)	/	62 (9.19)	64 (9.48)
Purposive sampling	/	18 (2.67)	6 (0.89)	24 (3.56)
Not available	/	/	13 (1.93)	13 (1.93)
Others	/	3 (0.44)	7 (1.04)	10 (1.48)
Random sampling	/	/	9 (1.33)	9 (1.33)
Time-Location Sampling	/	/	2 (0.30)	2 (0.30)
Modality of recruitment				
Offline	3 (0.44)	41 (6.07)	294 (43.56)	338 (50.07)
Mixed	2 (0.30)	8 (1.19)	172 (25.48)	182 (26.96)
Online	1 (0.15)	5 (0.74)	135 (20.00)	141 (20.89)
Not clear	/	/	14 (2.07)	14 (2.07)
Sample size				
1–50	/	49 (7.26)	1 (0.15)	50 (7.41)
51–100	/	5 (0.74)	14 (2.07)	19 (2.81)
101–200	1 (0.15)	/	39 (5.78)	40 (5.93)
201–300	1 (0.15)	/	68 (10.07)	69 (10.22)
301–400	/	/	87 (12.89)	87 (12.89)
401–500	/	/	97 (14.37)	97 (14.37)
501–600	/	/	67 (9.93)	67 (9.93)
601–700	/	/	29 (4.30)	29 (4.30)
701–800	/	/	15 (2.22)	15 (2.22)
801–900	/	/	20 (2.96)	20 (2.96)
901–1000	2 (0.30)	/	14 (2.07)	16 (2.37)
>1000	2 (0.30)	/	164 (24.3)	166 (24.59)
Data collection person				
Researcher	1 (0.15)	52 (7.70)	252 (37.33)	305 (45.19)
Participant	1 (0.15)	/	259 (38.37)	260 (38.52)
Researcher or Participants	4 (0.59)	2 (0.30)	75 (11.11)	81 (12.00)
Not clear	/	/	29 (4.30)	29 (4.30)
Data collection method				
Questionnaires	1 (0.15)	1 (0.15)	388 (57.48)	390 (57.78)
Laboratory tests & questionnaire	/	/	213 (31.56)	213 (31.56)
Interviews	/	43 (6.37)	2 (0.30)	45 (6.67)
Laboratory tests	/	/	10 (1.48)	10 (1.48)
Questionnaires & interviews	5 (0.74)	/	/	5 (0.74)
Focus Group Discussion (FGD)	/	5 (0.74)	/	5 (0.74)
Interviews & FGD	/	4 (0.59)	/	4 (0.59)
Others/ Multiple approaches	1 (0.15)	1 (0.15)	1 (0.15)	3 (0.45)
Total	*6 (0.89)*	*54 (8)*	*615 (91.11)*	*675 (100)*

TABLE 4 | Population characteristics of all included references.

Population characteristics	No. (%)	Characteristics	No. (%)
Target population[a]		**Gender[a]**	
MSM	2,667 (92.64)	Male only	2,777 (96.46)
Gay	66 (2.29)	Both gender	60 (2.08)
Gay & Lesbian	33 (1.15)	Female only	38 (1.32)
Homosexual & bisexual Male	28 (0.97)	Other (include transgender)	4 (0.14)
General homosexual & bisexual	21 (0.73)	**Whether patients or not[a]**	
Lesbian	19 (0.66)	Non patients	2,626 (91.21)
WSW	15 (0.52)	HIV-infected people	241 (8.37)
MSMW	10 (0.35)	Others (Syphilis)	12 (0.42)
Bisexual Male	6 (0.21)	**Orientation[b]**	
MSM and WSW	6 (0.21)	All homosexual	40 (5.93)
Homosexual & bisexual Female	4 (0.14)	All bisexual	1 (0.15)
Others (LGBTQ+)	4 (0.14)	**Homosexual rate**	
Mean years of age[b]		<40.0%	9 (1.34)
<20.0	1 (0.15)	40.0–49.9%	26 (3.85)
20.0~29.9	226 (33.48)	50.0–59.9%	50 (7.41)
30.0~39.9	100 (14.81)	60.0–69.9%	82 (12.15)
≥40.0	4 (0.89)	70.0–79.9%	119 (17.63)
NA	342 (50.67)	≥80.0%	80 (11.85)
Marriage rate (ever)[b]		NA	268 (39.70)
0–9.9%	73 (10.81)	**Bisexual rate**	
10.0–19.9%	183 (27.11)	<20.0%	82 (12.15)
20.0–29.9%	104 (15.41)	20.0–29.9%	107 (15.85)
30.0–39.9%	47 (6.96)	30.0–39.9%	48 (7.11)
40.0–59.9%	30 (4.44)	40.0–49.9%	21 (3.11)
60.0–88.6%	12 (1.78)	50.0–59.9%	6 (0.89)
All married	3 (0.44)	≥60.0%	2 (0.30)
NA	223 (33.04)	NA	368 (54.52)

[a]Data from all 2,879 publications; [b]Data from 675 studies published in EN only.

that had the highest number of studies were Sichuan Province (383, 11.57%), Guangdong Province (316, 9.55%), and Beijing Municipality (286, 8.64%).

In addition, this review attempted to provide an economic description of the study sites in terms of gross domestic product (GDP). The results showed that all study sites can be categorized into three groups, "Economically developed areas" [per capita GDP exceeding CNY¥100,000, Chinese Yuan (CNY)], "Economically moderate areas" (per capita GDP between CNY¥50,000 and CNY¥100,000), and "Economically underdeveloped areas" (per capita GDP less than CNY¥50,000). Most of the studies were conducted in economically moderate or developed areas (57.43 vs. 32.57%, respectively), while only 10% were carried out in economically underdeveloped areas, indicating the insufficient attention given to relatively poor places. Regarding the specific study settings, over half of the studies (63.85%) were conducted offline [usually in CDCs, some gay communities, or through non-governmental organizations

TABLE 5 | Context characteristics of all included references.

Context characteristics	No. (%)	Context characteristics	No. (%)
Single or multi-sites		**Study conducted setting[a]**	
Single site	2,251 (83.00)	***Offline***	
Multi-sites	453 (16.70)	Offline-CDC	92 (13.63)
Not clear	8 (0.29)	Offline-Clinic	66 (9.78)
Geographical division of all sites[b]		Offline-Gay community	81 (12.00)
Hong Kong	68 (2.05)	Offline-NGO	33 (4.89)
Taiwan	67 (2.02)	Offline-Not clear	158 (23.41)
Macau	2 (0.06)	Offline-University	1 (0.15)
Mainland	*3,173 (95.86)*	***Online***	
East China	841 (25.41)	Online-Application (App)	5 (0.74)
Southwest China	742 (22.42)	Online-Internet	100 (14.81)
South China	445 (13.44)	Online-Internet & App	27 (4.00)
North China	436 (13.17)	Online-Not clear	3 (0.44)
Central China	269 (8.13)	Online-Telephone	10 (1.48)
Northeast China	252 (7.61)	***Online & Offline***	54 (8.00)
Northwest China	188 (5.68)	Not mentioned	45 (6.67)
Geographical division of English publications[a]		**Site division according to economic status[c]**	
Mainland	563 (83.41)	Economically developed areas	1,078 (32.57)
Hong Kong	60 (8.89)	Economically moderate areas	1,901 (57.43)
Taiwan	44 (6.52)	Economically underdeveloped areas	331 (10.00)
Multi-sites	8 (1.19)		

[a] Data from 675 studies published in EN only. [b] Data from all studies mentioning the study sites (3,310 sites in total). [c] According to the Gross Domestic Product (GDP) per capita, "Economically developed areas" include Hong Kong, Macau, Taiwan, Beijing, Shanghai, Jiangsu, Fujian, Tianjin, and Zhejiang; "Economically moderate areas" include Guangdong, Chongqing, Hubei, Shandong, Inner Mongolia, Shaanxi, Anhui, Hunan, Liaoning, Sichuan, Jiangxi, Henan, Hainan, Ningxia, Xinjiang, Xizang, Yunnan, Qinghai, Jilin, and Shanxi; and "Economically underdeveloped areas" include Hebei, Guizhou, Guangxi, Heilongjiang, and Gansu.

(NGOs)], while around one-fifth of the studies (14.81%) were conducted online *via* Internet websites (**Table 5**).

Concept Information

Among all keywords that appeared in the included publications, "MSM" had the highest frequency as population variables in both CN and EN publications, followed by some sexual health-related concepts (e.g., HIV, STI, and sexual behaviors) and concepts related to mental health (e.g., stigma and depression), as shown in **Supplementary Material 7**. However, the concept of "social well-being" or other positive health-related concepts did not appear frequently. Co-word analysis was conducted to map the relationships among all keywords using UCINET and NetDraw (61), in which a higher number of co-occurrences indicated a closer relationship between the two keywords, as shown in **Figure 4**. The colors and lines were automatically generated after K-core analysis, specifically, the larger the number corresponding to the color, the higher the frequency of co-occurrence; the thicker the line, the stronger the degree of co-occurrence. Correspondingly, the results showed that "MSM" was the keyword with the highest frequency (1,908 times), co-occurring most frequently with other keywords. This was followed by "HIV" and "AIDS." In addition, HIV-/STI- and sexual health-related words were also high-frequency keywords. At the same time, keywords, such as "University students" appeared, indicating that young people were gaining increasing research attention. However, the frequency of the keyword "Lesbian" was very low, and that of "Bisexual" was even lower. Overall, male- and sexual health-related keywords appeared more frequently than female- and other health-related variables.

In addition, this review extracted the specific foci of 675 published EN studies, which had clear descriptions of the study variables they used. Among them, around half (45.93%) focused on sexual health only, and an additional quarter (24.15%) investigated both sexual health and social well-being. In comparison, the total number of studies focused on mental health accounted for only about 15%, whether it was the only variable in the research or one of the variables used (**Table 6**). Although the number of studies conducted for different genders varied greatly (627 vs. 16), the comparison revealed that male-related research focused more on sexual health (HIV/STI), while female-related research was relatively more concerned with mental health (25.00 vs. 3.35%). Furthermore, four studies investigated breast-related health among female sexual minorities (68–71), all of which were conducted in Taiwan.

This review also extracted all the health-related variables and corresponding measurement tools mentioned in the 675 studies published in EN. In comparison, less than a quarter of the studies clearly stated that they used validated scales (19.48%) or self-developed tools (4.73%), most of which reported their reliability or validity. Specifically, the most studied variables were still sexual health-related, such as sexual behavior and HIV screening (**Table 7**). The word clouds (**Supplementary Material 7**) also showed comparisons of variables investigated among the 16 female-specific studies and 627 male-specific studies.

DISCUSSION

To the best of our knowledge, this is the first scoping review conducted to synthesize the holistic health of homosexual and bisexual populations in mainland China, Hong Kong, Macau, and Taiwan. So far, it is also the most comprehensive systematic scoping review, with the highest number of included literature published in both EN and CN. Specifically, this review included all scientific literature published from January 2001 to May 2020, since homosexuality was removed from the list of mental illnesses in China in 2001 (2). The trends of all publications showed that research on related minorities increased significantly year by year, reaching a peak in 2013, and then gradually stabilized in the following years with some fluctuations. However, when observing the trend of new studies carried out every year, the trend of a gradual decrease in research over the past decade (2011–2020) is evident, indicating that research attention on this population is gradually decreasing. Considering the reality that

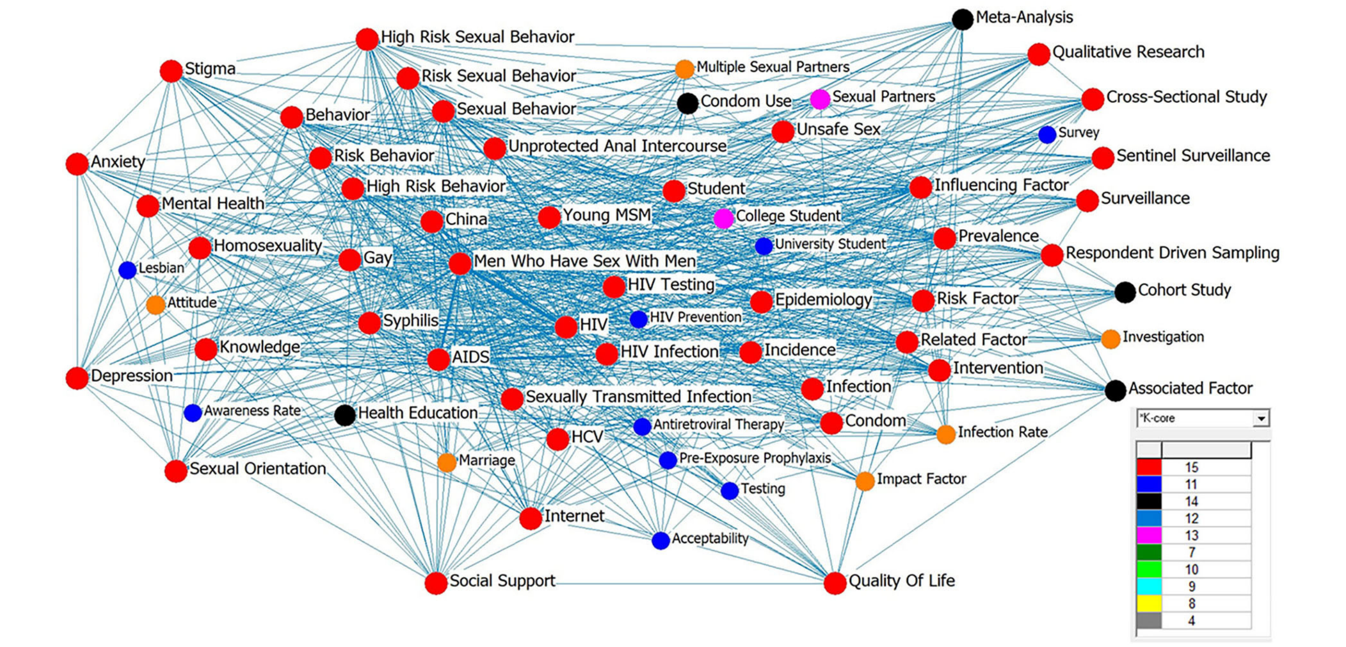


FIGURE 4 | Co-word analysis of all English (EN) keywords and translation of Chinese ones. The colors and lines were automatically generated after K-core analysis, specifically, the larger the number corresponding to the color, the higher the frequency of co-occurrence; the thicker the line, the stronger the degree of co-occurrence.

more and more sexual minorities are choosing to "come out" (72, 73), thus decreasing trend is a cause for concern and warrants further academic attention.

In response to the core review question in line with the PCC framework, this review found that there are far more studies on homosexuals than on bisexuals and other sexual minorities. In particular, the attention given to the male population was much higher than that to female groups (96.46 vs. 1.32%, respectively), especially in the most extensive MSM-related publications (92.64%). Regarding specific health-related concepts, there were far more studies on sexual health than investigations of mental health or social well-being (**Table 7**). In particular, the HIV-/STI-centered research was more common, with the concerns of scholars gradually evolving from HIV/STI prevalence (41, 74–76), treatment (77, 78), and corresponding adherence (79, 80) toward various prevention approaches, including both professional-led (81, 82) and self-testing/self-care programs (83–85). However, despite the prevalence of disease-centric research, less attention was paid to positive health concepts, such as sexual satisfaction, with only one survey conducted in the male group (86) while no relevant investigations for female minorities. All the above findings suggest that more research on female sexual minorities is needed in the future, whether research on mental health or the positive aspects of sexual health.

Specifically, there was an increased prevalence of HIV, syphilis, and gonorrhea among MSM (12, 87, 88). Although HIV was not detected in the female population in the existing studies, other diseases were relatively common, such as gonorrhea, chlamydia, and bacterial vaginosis (14), all indicating a worrying trend in sexual health in these sexual minorities. In terms of mental health, homophobia was universally experienced by gay men (19, 20), along with helplessness and other stressful feelings, which were also common among general homosexuals (17, 18). These may have further negative effects on their well-being directly or indirectly (32). Regarding social well-being, there were a certain number of investigations on the quality of life of general MSM (89, 90), MSMW (91), and HIV-infected MSM (92–94); and their perceived or received social support (95–99). These findings suggest that social support, which could serve as a protective factor against mental issues, should be further

TABLE 6 | Study focus and gender differences in 675 studies published in EN.

Study focus	Female (% in 16)	Male (% in 627)	Both gender (% in 32)	Total (% in 675)
Sexual health	2 (12.50)	307 (48.96)	1 (3.13)	310 (45.93)
Sexual health, Social well-being	/	160 (25.52)	2 (6.25)	163 (24.15)
Mental health, Social well-being	3 (18.75)	36 (5.74)	10 (31.25)	50 (7.41)
Sexual health, Mental health	/	43 (6.86)	3 (9.38)	46 (6.81)
Social well-being	3 (18.75)	33 (5.26)	8 (25.00)	44 (6.52)
Mental health	4 (25.00)	21 (3.35)	7 (21.88)	32 (4.74)
Holistic health	/	25 (3.99)	/	25 (3.70)
Breast health	4 (25.00)	/	/	3 (0.44)
Other physical health	/	2 (0.32)	1 (3.13)	3 (0.44)
Total	16	627	32	675

TABLE 7 | Specific health-related variables and corresponding measurements.

Concept characteristics		No. of studies	Percentage
Specific variables (≥30)	**Health domain**		
General sexual behavior	Sexual health	218	32.30%
HIV screening	Sexual health	107	15.85%
Condom use	Sexual health	94	13.93%
HIV testing behavior	Sexual health	78	11.56%
Syphilis screening	Sexual health	69	10.22%
Drug use	Social well-being	65	9.63%
Depression	Mental health	64	9.48%
HIV/AIDS knowledge	Sexual health	58	8.59%
HIV prevalence	Sexual health	54	8.00%
Risky behavior[b]	Sexual health	44	6.52%
Social support	Social well-being	33	4.89%
Syphilis prevalence	Sexual health	31	4.59%
Unprotected anal intercourse	Sexual health	31	4.59%
Anxiety	Mental health	30	4.44%
Measurements of all specific variables			
Not mentioned		1399	60.15%
Scales/Questionnaires[a]		453	19.48%
Laboratory tests		340	14.62%
Self-developed scales[a]		110	4.73%
Not available		24	1.03%
Validation of measurements[a]			
Yes		468	83.13%
No		95	16.87%

[a] Validation of those measurement tools explicitly mentioned in the study. [b] "Risky behavior" refers to different types of high-risk behavior, such as multiple sex partners, one-night stands, alcohol abuse, or other uncategorized behaviors that may cause health hazards.

enhanced. While there were no such studies targeting female sexual minorities, similar studies among this relatively unnoticed population should be conducted to investigate and improve their quality of life.

In addition, this review provided some other novel discoveries that have not yet been reported. First, the comparison of the cooperation between author institutions in CN and EN language journals showed greater international cooperation when publishing EN articles and greater domestic cooperation when publishing CN ones. This finding can be attributed to factors such as the language advantage and increasing international exchanges and cooperation of scholars. Second, the distribution of the profession of authors showed that the most common academic major was public health (63.7%), followed by psychology (6.4%), global health (3.9%), and nursing (2.7%). This indicates that scholars in the field of public health have relatively more experience, insights, and contributions to research on the health of sexual minorities. This finding also serves as a reminder for researchers in other disciplines to take the initiative to carry out relevant research. Third, regarding the dynamic development of sampling methods, with the increased visibility of these minority populations and the popularization of the Internet, sampling methods have gradually transitioned from using a venue-based approach to using multiple methods. Thus, more recent surveys now rely on the Internet or smartphone applications. Correspondingly, many new investigation approaches and more innovative intervention projects have been promoted, such as crowdsourcing interventions to promote HIV/STI prevention (100, 101) and the use of Internet popular opinion leaders (iPOL) interventions that rely on online peer support (102). This finding suggests that academics should carry out research with a developmental and more person-centered perspective to better promote health outcomes in these populations.

This review has some implications for relevant stakeholders regarding the holistic health improvement of homosexual and bisexual Chinese. First, for health service providers, especially medical staff in the field of public health, their gender, and sexuality literacy should be improved, as they are the direct providers of health care. Therefore, it is recommended that health professionals learn relevant knowledge so as to provide more diverse and professional advice. In addition, health-related institutions need to be as open and diversified as possible to provide more health services and increase health care accessibility, allowing more sexual minorities to actively seek help to a greater extent. Second, studies have shown that community engagement can help with health improvement by enhancing knowledge dissemination and facilitating testing (22, 103), thus highlighting the importance of a diverse and supportive community environment. Therefore, for social service providers, it is recommended that they proactively reach out to help and provide more social and psychological information to eliminate homophobia, biphobia, and other forms of discrimination. At the same time, all service providers should consciously participate in multidisciplinary cooperation to better provide person-centered holistic health care that can empower all sexual minorities to take pride in their own health and encourage them to maintain healthy behaviors. Finally, compared to studies on various general populations under the heteronormativity context, there are very few studies on sexual minorities. Thus, there is an urgent need for researchers or scholars from different disciplines to conduct research from different perspectives, with the common goal of helping everyone express their true selves and lead healthy lives.

This review has several limitations that should be addressed. First, despite the inclusion of numerous references in both languages, it is still difficult for a single review to accurately conclude the holistic health status of homosexual and bisexual Chinese from a comprehensive perspective. Due to the uneven quality of the included literature (as described in **Supplementary Material 6**), this review only summarized the objective demographic data (population and context information) from all the publications. While the summarization of concept information and study characteristics was based on the data extracted from EN publications only, most of which were peer-reviewed. Second, although this review attempted to conduct quality assessments of all included journals, nearly a quarter of the included articles were published in relatively low-quality journals, thereby leading to a potential limitation. Third, this review only included studies conducted in mainland China, Hong Kong, Macau, and Taiwan, while studies on overseas Chinese were excluded due to different social ideologies and policy backgrounds. This could mean that the findings were less representative of all homosexual and bisexual Chinese worldwide. Further comparison studies among the Chinese in various regions can be carried out in the future. Finally, this review involved a large number of references; hence, the findings might not be sufficiently population-specific or health-specific. Thus, more precise reviews or more targeted studies are needed in the future.

CONCLUSIONS

This scoping review summarized all the literature on the holistic health of homosexual and bisexual Chinese from 2001 to 2020. Existing evidence showed that previous research focused more on male than female sexual minorities, on HIV-/STI-centered surveys than person-centered investigations, on CDC-led interventions than crowdsourcing programs, and negative health conditions than positive health promotion. Thus, more investigations centered on the female population and following person-centered interventions are highly needed, along with the implementation of health promotion programs for this population group. Furthermore, projects that increase community engagement should be carried out among different communities, whether in communities based on gender or sexual orientation. Finally, researchers with more multidisciplinary backgrounds need to participate in the research on sexual minorities. In particular, there should be greater involvement of professionals from different disciplines to deliver adequate health care and social care services for Chinese homosexuals and bisexuals.

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

CW and EPHC conceptualized this review and designed the research questions. CW prepared and drafted the manuscript. CW and EPHC performed the initial screening of the articles and were involved in the development of the data extraction form. CW piloted the data extraction form and conducted the data extraction with EPHC independently. CW carried out data analysis under the guidance of EPHC. EPHC and PHC are the guarantors and have contributed to the critical revision of the manuscript. All the authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

We would like to thank Ms. Kitty Wai Ying Choi for her contribution to the pilot data searching and study selection and the development of the data extraction form.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.710575/full#supplementary-material

REFERENCES

1. Tanner HM. The Offense of Hooliganism and The moral dimension of china's pursuit of modernity, 1979–1996. *Twentieth-Century China.* (2000) 26:1–40. doi: 10.1179/tcc.2000.26.1.1
2. Chen YF. Chinese classification of mental disorders (CCMD-3): towards integration in international classification. *Psychopathology.* (2002) 35:171–5. doi: 10.1159/000065140
3. Yen CF, Ko NY, Huang YT, Chen MH, Lin IH, Lu WH. Preference about laws for the legal recognition of same-sex relationships in taiwanese people before and after same-sex marriage referenda: a facebook survey study. *Int J Environ Res Public Health.* (2020) 17:2000. doi: 10.3390/ijerph17062000
4. Esteve A, Lesthaeghe R, Lopez-Gay A. The latin American cohabitation boom, 1970-2007. *Popul Dev Rev.* (2012) 38:55–81. doi: 10.1111/j.1728-4457.2012.00472.x
5. Zhao Y, Ma Y, Chen R, Li F, Qin X, Hu Z. Non-disclosure of sexual orientation to parents associated with sexual risk behaviors among gay and bisexual MSM in China. *AIDS Behav.* (2016) 20:193–203. doi: 10.1007/s10461-015-1135-6
6. Bellhouse C, Walker S, Fairley CK, Chow EP, Bilardi JE. Getting the terminology right in sexual health research: the importance of accurately classifying fuck buddies among men who have sex with men. *Sex Transm Infect.* (2018) 94:487–9. doi: 10.1136/sextrans-2016-053000
7. Liu C. "Red is not the only color of a rainbow": the making and resistance of the "MSM" subject among gay men in China. *Soc Sci Med.* (2020) 252:112947. doi: 10.1016/j.socscimed.2020.112947
8. Bailey JV, Farquhar C, Owen C, Mangtani P. Sexually transmitted infections in women who have sex with women. *Sex Transm Infect.* (2004) 80:244–6. doi: 10.1136/sti.2003.007641
9. Liao M, Kang D, Jiang B, Tao X, Qian Y, Wang T, et al. Bisexual behavior and infection with HIV and syphilis among men who have sex with men along the east coast of China. *AIDS Patient Care STDS.* (2011) 25:683–91. doi: 10.1089/apc.2010.0371
10. Liao M, Wang M, Shen X, Huang P, Yang X, Hao L, et al. Bisexual behaviors, HIV knowledge, and stigmatizing/discriminatory attitudes among men who have sex with men. *PLoS ONE.* (2015) 10:e0130866. doi: 10.1371/journal.pone.0130866
11. World Health Organization. *What is the WHO definition of health?* (1948). Available online at: https://www.who.int/news-room/fact-sheets/detail/mental-health-strengthening-our-response
12. Dong MJ, Peng B, Liu ZF, Ye QN, Liu H, Lu XL, et al. The prevalence of HIV among MSM in China: a large-scale systematic analysis. *BMC Infect Dis.* (2019) 19:1000. doi: 10.1186/s12879-019-4559-1
13. Wu ZY, Xu J, Liu EW, Mao YR, Xiao Y, Sun XH, et al. HIV and syphilis prevalence among men who have sex with men: a cross-sectional survey of 61 cities in China. *Clin Infect Dis.* (2013) 57:298–309. doi: 10.1093/cid/cit210

14. Wang XF, Norris JL, Liu YJ, Reilly KH, Wang N. Health-related attitudes and risk factors for sexually transmitted infections of Chinese women who have sex with women. *Chin Med J.* (2012) 125:2819–25. doi: 10.3760/cma.j.issn.0366-6999.2012.16.004
15. Wang X, Norris JL, Liu Y, Vermund SH, Qian HZ, Han L, et al. Risk behaviors for reproductive tract infection in women who have sex with women in Beijing, China. *PLoS ONE.* (2012) 7:e40114. doi: 10.1371/journal.pone.0040114
16. Choi KH, Steward WT, Miege P, Hudes E, Gregorich SE. Sexual stigma, coping styles, and psychological distress: a longitudinal study of men who have sex with men in Beijing, China. *Arch Sex Behav.* (2016) 45:1483–91. doi: 10.1007/s10508-015-0640-z
17. Ren Z, Howe CQ, Zhang W. Maintaining "mianzi" and "lizi": Understanding the reasons for formality marriages between gay men and lesbians in China. *Transcult Psychiatry.* (2019) 56:213–32. doi: 10.1177/1363461518799517
18. Shi X, Xu W, Zheng Y. Heterosexual marital intention: effects of internalized homophobia, homosexual identity, perceived family support, and disclosure among Chinese gay and bisexual men. *J Homosex.* (2018) 67:452–67. doi: 10.1080/00918369.2018.1547558
19. Ren Z, Hood RW Jr. Internalized homophobia scale for gay chinese men: conceptualization, factor structure, reliability, and associations with hypothesized correlates. *Am J Mens Health.* (2018) 12:1297–306. doi: 10.1177/1557988318768603
20. Xu W, Zheng L, Xu Y, Zheng Y. Internalized homophobia, mental health, sexual behaviors, and outness of gay/bisexual men from Southwest China. *Int J Equity Health.* (2017) 16:36. doi: 10.1186/s12939-017-0530-1
21. Ding C, Chen X, Wang W, Yu B, Yang H, Li X, et al. Sexual minority stigma, sexual orientation concealment, social support and depressive symptoms among men who have sex with men in China: a moderated mediation modeling analysis. *AIDS Behav.* (2020) 24:8–17. doi: 10.1007/s10461-019-02713-3
22. Zhu Y, Liu J, Chen Y, Zhang R, Qu B. The relation between mental health, homosexual stigma, childhood abuse, community engagement, and unprotected anal intercourse among MSM in China. *Sci Rep.* (2018) 8:3984. doi: 10.1038/s41598-018-22403-9
23. Liao M, Kang D, Tao X, Bouey JH, Aliyu MH, Qian Y, et al. Alcohol use, stigmatizing/discriminatory attitudes, and hiv high-risk sexual behaviors among men who have sex with men in China. *BioMed Res Int.* (2014) 2014:143738. doi: 10.1155/2014/143738
24. Pyun T, Santos GM, Arreola S, Do T, Hebert P, Beck J, et al. Internalized homophobia and reduced HIV testing among men who have sex with men in China. *Asia Pac J Public Health.* (2014) 26:118–25. doi: 10.1177/1010539514524434
25. Su X, Zhou AN Li J, Shi LE, Huan X, Yan H, et al. Depression, loneliness, and sexual risk-taking among HIV-negative/unknown men who have sex with men in China. *Arch Sex Behav.* (2018) 47:1959–68. doi: 10.1007/s10508-017-1061-y
26. Liu JX, Zhong XN, Lu Z, Peng B, Zhang Y, Liang H, et al. Anxiety and depression associated with anal sexual practices among HIV-negative men who have sex with men in western China. *Int J Environ Res Public Health.* (2020) 17:464. doi: 10.3390/ijerph17020464
27. Hu Y, Zhong XN, Peng B, Zhang Y, Liang H, Dai JH, et al. Comparison of depression and anxiety between HIV-negative men who have sex with men and women (MSMW) and men who have sex with men only (MSMO): a cross-sectional study in Western China. *BMJ Open.* (2019) 9:e023498. doi: 10.1136/bmjopen-2018-023498
28. Wang PW, Ko NY, Hsiao RC, Chen MH, Lin HC, Yen CF. Suicidality among gay and bisexual men in taiwan: its relationships with sexuality and gender role characteristics, homophobic bullying victimization, and social support. *Suicide Life Threat Behav.* (2019) 49:466–77. doi: 10.1111/sltb.12451
29. Chen H, Li Y, Wang L, Zhang B. Causes of suicidal behaviors in men who have sex with men in China: a national questionnaire survey. *BMC Public Health.* (2015) 15:91. doi: 10.1186/s12889-015-1436-8
30. Mu H, Li Y, Liu L, Na J, Yu L, Bi X, et al. Prevalence and risk factors for lifetime suicide ideation, plan and attempt in Chinese men who have sex with men. *BMC Psychiatry.* (2016) 16:10. doi: 10.1186/s12888-016-0830-9
31. Higgins DJ. Differences between previously married and never married 'gay' men: family background, childhood experiences and current attitudes. *J Homosex.* (2004) 48:19–41. doi: 10.1300/J082v48n01_02
32. Sun S, Pachankis JE Li X, Operario D. Addressing Minority Stress and Mental Health among Men Who Have Sex with Men (MSM) in China. *Curr HIV/AIDS Rep.* (2020) 17:35–62. doi: 10.1007/s11904-019-00479-w
33. Gu J, Lau JT, Wang Z, Wu AM, Tan X. Perceived empathy of service providers mediates the association between perceived discrimination and behavioral intention to take up HIV antibody testing again among men who have sex with men. *PLoS ONE.* (2015) 10:e0117376. doi: 10.1371/journal.pone.0117376
34. Choi K, Hudes ES, Steward WT. Social discrimination, concurrent sexual partnerships, and HIV risk among men who have sex with men in Shanghai, China. *AIDS Behav.* (2008) 12:S71–7. doi: 10.1007/s10461-008-9394-0
35. Bai X, Xu J, Yang J, Yang B, Yu M, Gao Y, et al. HIV prevalence and high-risk sexual behaviours among MSM repeat and first-time testers in China: implications for HIV prevention. *J Int AIDS Soc.* (2014) 17:18848. doi: 10.7448/IAS.17.1.18848
36. Bien CH, Muessig KE, Lee R, Lo EJ, Yang LG, Yang B, et al. HIV and syphilis testing preferences among men who have sex with men in South China: a qualitative analysis to inform sexual health services. *PLoS ONE.* (2015) 10:e0124161. doi: 10.1371/journal.pone.0124161
37. Cai WD, Zhao J, Zhao JK, Raymond HF, Feng YJ, Liu J, et al. HIV prevalence and related risk factors among male sex workers in Shenzhen, China: results from a time-location sampling survey. *Sex Transm Infect.* (2010) 86:15–20. doi: 10.1136/sti.2009.037440
38. Chen L, Yang J, Ma Q, Pan X. Prevalence of active syphilis infection and risk factors among HIV-positive MSM in Zhejiang, China in 2015: a cross-sectional study. *Int J Environ Res Public Health.* (2019) 16:1507. doi: 10.3390/ijerph16091507
39. Chen Q, Sun Y, Sun W, Hao M, Li G, Su X, et al. Trends of HIV incidence and prevalence among men who have sex with men in Beijing, China: Nine consecutive cross-sectional surveys, 2008-2016. *PLoS ONE.* (2018) 13:e0201953. doi: 10.1371/journal.pone.0201953
40. Chen Y, Tang W, Chen L, Shi L, Liu X, Xu J, et al. Changing epidemic of HIV and syphilis among resident and migrant men who have sex with men in Jiangsu, China. *Sci Rep.* (2017) 7:9478. doi: 10.1038/s41598-017-08671-x
41. Chow EP, Lau JT, Zhuang X, Zhang X, Wang Y, Zhang L, et al. prevalence trends, risky behaviours, and governmental and community responses to the epidemic among men who have sex with men in China. *Biomed Res Int.* (2014) 2014:607261. doi: 10.1155/2014/607261
42. Chow EP, Wilson DP, Zhang L, HIV. and syphilis co-infection increasing among men who have sex with men in China: a systematic review and meta-analysis. *PLoS ONE.* (2011) 6:e22768. doi: 10.1371/journal.pone.0022768
43. Duan Y, Zhang H, Wang J, Wei S, Yu F, She M. Community-based peer intervention to reduce HIV risk among men who have sex with men in Sichuan province, China. *Aids Educ Prev.* (2013) 25:38–48. doi: 10.1521/aeap.2013.25.1.38
44. Hu X, Wang Y, LGB. identity among young Chinese: the influence of traditional culture. *J Homosex.* (2013) 60:667–84. doi: 10.1080/00918369.2013.773815
45. Li Y. *Risk Perception, Mental health, Risk Behaviors of Homosexuals and its Media Influence Factors* [Master of Philosophy]. Guangzhou: Jinan University. (2017).
46. Li H. The mental health status of homosexual people and its influencing factors. *Sci Soc Psychol.* (2010) 25:80–5. Avaiable online at: http://qikan.cqvip.com/Qikan/Article/Detail?id=34144610
47. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. *Int J Soc Res Methodol.* (2005) 8:19–32. doi: 10.1080/1364557032000119616
48. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. *Implement Sci.* (2010) 5:69. doi: 10.1186/1748-5908-5-69
49. 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

50. Peters M, Godfrey C, McInerney P, Munn Z, Tricco A, Khalil H. *Chapter 11: Scoping Reviews (2020 version)*. The Joanna Briggs Institute. (2020). Available online at: https://reviewersmanual.joannabriggs.org/
51. 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
52. Wu C, Choi EPH, Chau PH, Choi KWY. The holistic health status of chinese homosexual and bisexual adults: a scoping review protocol. *Jmir Res Protoc.* (2021). doi: 10.2196/preprints.30870
53. Joanna Briggs Institute. *Joanna Briggs Institute Reviewer's Manual*. 4th Edition: The Joanna Briggs Institute. (2020). Available online at: https://reviewersmanual.joannabriggs.org/.
54. World Health Organization. Special initiative for mental health (2019-2023): *Universal Health Coverage for Mental Health*. 2019.
55. Jaspal R. *The Social Psychology of Gay Men*. Cham: Palgrave Macmillan UK. (2019). doi: 10.1007/978-3-030-27057-5
56. Edwards WM, Coleman E. Defining sexual health: a descriptive overview. *Arch Sex Behav.* (2004) 33:189–95. doi: 10.1023/B:ASEB.0000026619.95734.d5
57. Giami A. Sexual health: the emergence, development, and diversity of a concept. *Annu Rev Sex Res.* (2002) 13:1–35.
58. World Health Organization. *Definitions of Sexual Health.* (2006). Available online at: https://www.who.int/health-topics/sexual-health#tab=tab_2.
59. McGowan J, Sampson M, Salzwedel DM, Cogo E, Foerster V, Lefebvre C, et al. Peer review of electronic search strategies: 2015 guideline statement. *J Clin Epidemiol.* (2016) 75:40–6. doi: 10.1016/j.jclinepi.2016.01.021
60. Rebecca R, Horey D, Oliver S, McKenzie J, Prictor M, Santesso N, et al. *Cochrane Consumers and Communication Protocol Text and Additional Guidance for Review Authors*. Melbourne: La Trobe University. (2019). Available online at: https://latrobe.figshare.com/articles/Standard_Protocol_Text_and_Additional_Guidance_for_Review_Authors/6692969.
61. Borgatti SP, Everett MG, Freeman LC. UCINET VI for windows: software for social network analysis. (2002).
62. Borgatti SP. *NetDraw Software for Network Visualization*. Lexington, KY: Analytic Technologies. (2002).
63. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *J Clin Epidemiol.* (2009) 62:1006–12. doi: 10.1016/j.jclinepi.2009.06.005
64. Tang S, Tang W, Meyers K, Chan P, Chen Z, Tucker JD, et al. Epidemiology and responses among men who have sex with men and transgender individuals in China: a scoping review. *BMC Infect Dis.* (2016) 16:588. doi: 10.1186/s12879-016-1904-5
65. Hou C-N, Lu H-Y. Online networks as a venue for social support: a qualitative study of married bisexual men in Taiwan. *J Homosex.* (2013) 60:1280–96. doi: 10.1080/00918369.2013.806170
66. Li Y, Johnson BD, Jenkins-Guarnieri MA. Sexual identity development and subjective well-being among Chinese lesbians. *Int Perspect Psychol Res Prac Consultation.* (2013) 2:242–54. doi: 10.1037/a0033752
67. Lin K. *Study on the Current Situation and Countermeasure of STD/AIDS Related Health Intervention of Homosexual Website in China*. Shanghai: Fudan University. (2008).
68. Liu PL, Yeo TED. Breast health, risk factors, and cancer screening among lesbian, bisexual, and queer/questioning women in China. *Health Care Women Int.* (2019) 40:1–15. doi: 10.1080/07399332.2019.1571062
69. Wang Y-C, Griffiths J, Grande G. Factors associated with Taiwanese lesbians' breast health-care behavior and intentions: qualitative interview findings. *Women Health.* (2017) 57:962–75. doi: 10.1080/03630242.2016.1222331
70. Wang Y-C, Griffiths J, Grande G. The influence of gender identities on body image and breast health among sexual minority women in Taiwan: implications for healthcare practices. *Sex Roles J Res.* (2018) 78:242–54. doi: 10.1007/s11199-017-0793-4
71. Wang YC, Chang SR, Miao NF. The role of butch versus femme identity in body image and breast health among lesbians in taiwan: results of an online survey. *J Nurs Sch.* (2020) 52:174–82. doi: 10.1111/jnu.12544
72. Wei C, Liu W. Coming out in Mainland China: a national survey of LGBTQ students. *J LGBT Youth.* (2019) 16:192–219. doi: 10.1080/19361653.2019.1565795
73. Mak WWS, Ng AC, Mo PKH, Chong ESK. Coming out among lesbians, gays, and bisexual individuals in Hong Kong: application of the theory of planned behavior and the moderating role of attitudinal ambivalence. *Sex Role J Res.* (2010) 63:189–200. doi: 10.1007/s11199-010-9778-2
74. Meng X, Zou H, Beck J, Xu Y, Zhang X, Miao X, et al. Trends in HIV prevalence among men who have sex with men in China 2003-09: a systematic review and meta-analysis. *Sex Health.* (2013) 10:211–9. doi: 10.1071/SH12093
75. Wang X, Lan G, Shen Z, Vermund SH, Zhu Q, Chen Y, et al. HIV and syphilis prevalence trends among men who have sex with men in Guangxi, China: yearly cross-sectional surveys, 2008-2012. *BMC Infec Dis*. (2014) 14:367. doi: 10.1186/1471-2334-14-367
76. Wei S, Zhang H, Wang J, Song D, Duan Y, Yu F, et al. HIV and syphilis prevalence and associated factors among young men who have sex with men in 4 cities in China. *AIDS Behav.* (2013) 17:1151–8. doi: 10.1007/s10461-011-0110-0
77. Yang X. Prevalence and factors of antiretroviral therapy usage among HIV-positive men who have sex with men under a new treatment policy in China [Ph.D.]. *Ann Arbor: The Chinese University of Hong Kong (Hong Kong)* (2018).
78. Zhou W, Zhao M, Wang X, Schilling RF, Zhou S, Qiu HY, et al. Treatment adherence and health outcomes in MSM with HIV/AIDS: patients enrolled in "one-stop" and standard care clinics in Wuhan China. *PLoS ONE.* (2014) 9:e113736. doi: 10.1371/journal.pone.0113736
79. Qu D, Zhong X, Xiao G, Dai J, Liang H, Huang A. Adherence to pre-exposure prophylaxis among men who have sex with men: a prospective cohort study. *Int J Infect Dis*. (2018) 75:52–9. doi: 10.1016/j.ijid.2018.08.006
80. Hu Y. Zhong X-n, Peng B, Zhang Y, Liang H, Dai J-h, et al. Associations between perceived barriers and benefits of using HIV pre-exposure prophylaxis and medication adherence among men who have sex with men in Western China. *BMC Infect Dis.* (2018) 18:1–8. doi: 10.1186/s12879-018-3497-7
81. Ye S, Xiao Y, Jin C, Cassell H, Blevins M, Sun J, et al. Effectiveness of integrated HIV prevention interventions among Chinese men who have sex with men: evaluation of a 16-city public health program. *PLoS ONE.* (2012) 7:e50873. doi: 10.1371/journal.pone.0050873
82. Zheng L, Zheng Y. Efficacy of human immunodeficiency virus prevention interventions among men who have sex with men in China: a meta-analysis. *Sex Transm Dis.* (2012) 39:886–93. doi: 10.1097/OLQ.0b013e31826ae85e
83. Ren XL, Wu ZY Mi GD, McGoogan J, Rou KM, Zhao Y. Uptake of HIV self-testing among men who have sex with men in Beijing, China: a cross-sectional study. *Biomed Environ Sci.* (2017) 30:407–17. doi: 10.1186/s40249-017-0326-y
84. Zhu X, Zhang W, Operario D, Zhao Y, Shi A, Zhang Z, et al. Effects of a mobile health intervention to promote HIV self-testing with MSM in China: a randomized controlled trial. *AIDS Behav.* (2019) 23:3129–39. doi: 10.1007/s10461-019-02452-5
85. Wong HT. Hoo, Tam HY, Chan DP, Chung, Lee SS. Usage and acceptability of HIV self-testing in men who have sex with men in Hong Kong. *AIDS Behav.* (2015) 19:505–15. doi: 10.1007/s10461-014-0881-1
86. Zheng L, Zheng Y. Sexual satisfaction in Chinese gay and bisexual men: relationship to negative sexual minority identity and sexual role preference. *Sex Relationship Ther.* (2017) 32:75–88. doi: 10.1080/14681994.2016.1200027
87. Chen G, Cao Y, Yao Y, Li M, Tang W, Li J, et al. Syphilis incidence among men who have sex with men in China: results from a meta-analysis. *Int J STD AIDS.* (2017) 28:170–8. doi: 10.1177/0956462416638224
88. Yang LG, Zhang XH, Zhao PZ, Chen ZY, Ke WJ, Ren XQ, et al. Gonorrhea and chlamydia prevalence in different anatomical sites among men who have sex with men: a cross-sectional study in Guangzhou, China. *BMC Infect Dis.* (2018) 18:675. doi: 10.1186/s12879-018-3579-6
89. Liu J, Qu B, Zhu Y, Hu B. The influence of social support on quality of life of men who have sex with men in China: a preliminary study. *PLoS ONE.* (2015) 10:e0127644. doi: 10.1371/journal.pone.0127644

90. Yaxin Z, Jie L, Bo Q, Bingxue H, Yang Z, Zhu Y, et al. Relationship between quality of life and unprotected anal intercourse among Chinese men who have sex with men: a cross-sectional study. *BMC Public Health*. (2016) 16:1–6. doi: 10.1186/s12889-016-3076-z
91. Chen JP, Han MM, Liao ZJ Dai ZZ, Liu L, Chen H, et al. HIV-related behaviors, social support and health-related quality of life among men who have sex with men and women (MSMW): a cross-sectional study in Chongqing, China. *PLoS One*. (2015) 10:e0118651. doi: 10.1371/journal.pone.0118651
92. Song B, Yan C, Lin Y, Wang F, Wang L. Health-related quality of life in HIV-infected men who have sex with men in China: a cross-sectional study. *Med Sci Monit*. (2016) 22:2859–70. doi: 10.12659/MSM.897017
93. Zhang P, Gao J, Wang Y, Sun Q, Sun X. Effect of chronic disease self-management program on the quality of life of HIV-infected men who have sex with men: an empirical study in Shanghai, China. *Int J Health Plann Manage*. (2019) 34:1055–64. doi: 10.1002/hpm.2874
94. Shao B, Song B, Feng S, Lin Y, Du J, Shao H, et al. The relationship of social support, mental health, and health-related quality of life in human immunodeficiency virus-positive men who have sex with men: From the analysis of canonical correlation and structural equation model: a cross-sectional study. *Medicine (Baltimore)*. (2018) 97:e11652. doi: 10.1097/MD.0000000000011652
95. Du M, Zhao J, Zhang J, Lau JTF, Mo PKH Li J. Depression and social support mediate the effect of HIV self-stigma on condom use intentions among Chinese HIV-infected men who have sex with men. *AIDS Care*. (2018) 30:1197–206. doi: 10.1080/09540121.2018.1487916
96. Li J, Mo PK, Wu AM, Lau JT. Roles of self-stigma, social support, and positive and negative affects as determinants of depressive symptoms among HIV infected men who have sex with men in China. *AIDS Behav*. (2017) 21:261–73. doi: 10.1007/s10461-016-1321-1
97. Mo PKH, Chen X, Lam EHK Li J, Kahler CW, Lau JTF. The moderating role of social support on the relationship between anxiety, stigma, and intention to use illicit drugs among HIV-positive men who have sex with men. *AIDS Behav*. (2020) 24:55–64. doi: 10.1007/s10461-019-02719-x
98. Yan H, Li X, Li J, Wang W, Yang Y, Yao X, et al. Association between perceived HIV stigma, social support, resilience, self-esteem, and depressive symptoms among HIV-positive men who have sex with men (MSM) in Nanjing, China. *AIDS Care*. (2019) 31:1069–76. doi: 10.1080/09540121.2019.1601677
99. Wang C-C, Lin H, Chen M-H, Ko N-Y, Chang Y-P, Lin IM, et al. Effects of traditional and cyber homophobic bullying in childhood on depression, anxiety, and physical pain in emerging adulthood and the moderating effects of social support among gay and bisexual men in Taiwan. *Neuropsychiatr Dis Treat*. (2018) 14:9. doi: 10.2147/NDT.S164579
100. Tang W, Han L, Best J, Zhang Y, Mollan K, Kim J, et al. Crowdsourcing HIV test promotion videos: a noninferiority randomized controlled trial in China. *Clinical Infect Dis*. (2016) 62:1436–42. doi: 10.1093/cid/ciw171
101. Tang W, Wei C, Cao B, Wu D, Li KT, Lu H, et al. Crowdsourcing to expand HIV testing among men who have sex with men in China: a closed cohort stepped wedge cluster randomized controlled trial. *PLoS Med*. (2018) 15:e1002645. doi: 10.1371/journal.pmed.1002645
102. Ko N-Y, Hsieh C-H, Wang M-C, Lee C, Chen C-L, Chung A-C, et al. Effects of internet popular opinion leaders (iPOL) among internet-using men who have sex with men. *J Med Internet Res*. (2013) 15:e40. doi: 10.2196/jmir.2264
103. Zhang TP, Liu C, Han L, Tang W, Mao J, Wong T, et al. Community engagement in sexual health and uptake of HIV testing and syphilis testing among MSM in China: a cross-sectional online survey. *J Int AIDS Soc*. (2017) 20:21372. doi: 10.7448/IAS.20.01/21372

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: 27 August 2021
doi: 10.3389/fpubh.2021.697501

Measuring the Value of a Practical Text Mining Approach to Identify Patients With Housing Issues in the Free-Text Notes in Electronic Health Record: Findings of a Retrospective Cohort Study

Elham Hatef[1]*, *Gurmehar Singh Deol*[1], *Masoud Rouhizadeh*[2], *Ashley Li*[3], *Katyusha Eibensteiner*[4], *Craig B. Monsen*[4], *Roman Bratslaver*[4], *Margaret Senese*[4] and *Hadi Kharrazi*[1]

[1] Center for Population Health IT, Johns Hopkins School of Public Health, Baltimore, MD, United States, [2] The Institute for Clinical and Translational Research, Johns Hopkins School of Medicine, Baltimore, MD, United States, [3] Department of Biomedical Engineering, Johns Hopkins Whiting School of Engineering, Baltimore, MD, United States, [4] Atrius Health, Newton, MA, United States

Introduction: Despite the growing efforts to standardize coding for social determinants of health (SDOH), they are infrequently captured in electronic health records (EHRs). Most SDOH variables are still captured in the unstructured fields (i.e., free-text) of EHRs. In this study we attempt to evaluate a practical text mining approach (i.e., advanced pattern matching techniques) in identifying phrases referring to housing issues, an important SDOH domain affecting value-based healthcare providers, using EHR of a large multispecialty medical group in the New England region, United States. To present how this approach would help the health systems to address the SDOH challenges of their patients we assess the demographic and clinical characteristics of patients with and without housing issues and briefly look into the patterns of healthcare utilization among the study population and for those with and without housing challenges.

Methods: We identified five categories of housing issues [i.e., homelessness current (HC), homelessness history (HH), homelessness addressed (HA), housing instability (HI), and building quality (BQ)] and developed several phrases addressing each one through collaboration with SDOH experts, consulting the literature, and reviewing existing coding standards. We developed pattern-matching algorithms (i.e., advanced regular expressions), and then applied them in the selected EHR. We assessed the text mining approach for recall (sensitivity) and precision (positive predictive value) after comparing the identified phrases with manually annotated free-text for different housing issues.

Results: The study dataset included EHR structured data for a total of 20,342 patients and 2,564,344 free-text clinical notes. The mean (SD) age in the study population was 75.96 (7.51). Additionally, 58.78% of the cohort were female. BQ and HI were the most frequent housing issues documented in EHR free-text notes and HH was the

OPEN ACCESS

Edited by:
Yi Guo,
University of Florida, United States

Reviewed by:
Xiaolin Xu,
Zhejiang University, China
Andreia Leite,
New University of Lisbon, Portugal
Xi Yang,
University of Florida, United States

***Correspondence:**
Elham Hatef
ehatef1@jhu.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: *19 April 2021*
Accepted: *28 July 2021*
Published: *27 August 2021*

Citation:
Hatef E, Singh Deol G, Rouhizadeh M, Li A, Eibensteiner K, Monsen CB, Bratslaver R, Senese M and Kharrazi H (2021) Measuring the Value of a Practical Text Mining Approach to Identify Patients With Housing Issues in the Free-Text Notes in Electronic Health Record: Findings of a Retrospective Cohort Study. Front. Public Health 9:697501. doi: 10.3389/fpubh.2021.697501

least frequent one. The regular expression methodology, when compared to manual annotation, had a high level of precision (positive predictive value) at phrase, note, and patient levels (96.36, 95.00, and 94.44%, respectively) across different categories of housing issues, but the recall (sensitivity) rate was relatively low (30.11, 32.20, and 41.46%, respectively).

Conclusion: Results of this study can be used to advance the research in this domain, to assess the potential value of EHR's free-text in identifying patients with a high risk of housing issues, to improve patient care and outcomes, and to eventually mitigate socioeconomic disparities across individuals and communities.

Keywords: electronic health record, free-text, clinical notes, housing, natural language processing, social determinants of health

INTRODUCTION

The adoption of electronic health records (EHRs) among U.S. hospitals and outpatient facilities has dramatically increased over the last decade (1, 2). Meaningful Use criteria (3, 4), the main driver of increased EHR adoption (5), has incentivized a higher capture rate of demographic and clinical information (6). Moreover, clinical informaticians and health information technology (HIT) experts have started to assess and optimize the documentation and collection of social determinants of health (SDOH) in EHRs for specific subpopulations of patients (7–12); however, SDOH documentation is still an uncommon practice in EHRs (13).

Despite the growing effort to standardize coding for SDOH concepts (14) such as Logical Observation Identifiers Names and Codes (LOINC) (15), SDOH variables are infrequently captured in EHR's structured fields and are often limited to certain SDOH types within specific clinical conditions (e.g., child abuse within the pediatric population; smoking cessation in primary care) (16, 17). However, SDOH challenges may be discussed with healthcare providers during visits and recorded in EHRs as free-text notes (i.e., providers' notes). Most SDOH variables are still captured in the unstructured fields of EHRs such as admission or clinical progress notes (14). For example, lack of social support among older adults is mentioned considerably more in geriatric notes compared to coded EHR data or other structured data sources such as insurance claims (7, 18).

While the HIT challenges exist, collecting SDOH information and implementing SDOH-specific interventions on a patient-level has become a priority for value-based care settings operating under specific organizational structures such as accountable care organizations or patient-centered medical homes (19, 20). Various factors have played a role in increasing the priority of SDOH collection among value-based settings. Some payers have started to mandate the collection of SDOH variables using survey instruments [e.g., Center for Medicare and Medicaid Innovation's Comprehensive Primary Care Plus (21) and some Medicaid (22) and private plans (23) among contracted value-based providers]. Additionally, certain states have recently introduced SDOH-derived variables to adjust the global budgets of their contracted health providers (24) [e.g., neighborhood stress index in Massachusetts' Medicaid program (20)].

Despite the incentives of value-based health systems to collect patient-level SDOH, operational challenges in rolling out large-scale SDOH surveys have limited such efforts on a population level (23, 25). Thus, the EHR free-text notes might provide a more complete or accurate accounting of SDOH challenges; however, traditional approaches for review and abstraction of patient information from medical record notes are laborious, expensive, and slow. Recent developments in text mining and natural language processing (NLP) of digitized text allow for reliable, low-cost, and rapid extraction of information from EHRs (7, 8, 18). Developing NLP algorithms that could function in different healthcare systems would improve the generalizability and application of such methods in extracting social needs from the EHR's free text. Thus, EHR text mining methods can be integrated within value-based operations to improve the identification of patient populations with SDOH challenges.

This study attempts to evaluate a practical text mining approach (i.e., advanced pattern matching techniques using regular expressions; RegEx) in identifying phrases referring to housing challenges, an important SDOH domain affecting value-based healthcare providers, using EHR of a large multispecialty medical group in New England region, United States. To present how this approach would help the health systems to address the SDOH challenges of their patients we assess the demographic and clinical characteristics of patients with and without housing issues and briefly look into the patterns of healthcare utilization among the study population and for those with and without housing challenges. The development of generalizable text mining methodologies with promising performance will help to identify social needs of patients for research purposes and to enhance the value of EHRs for population health management of at-risk patients across different health systems.

METHODS

Data Source

We used de-identified EHR data from a large multispecialty medical group from New England, United States. We utilized data on a cohort of members who received health insurance coverage between 2011 and 2013 (based on data availability and agreement with the medical group about data access) and were assigned to this medical group as their primary source of medical care from this health plan. We extracted both structured and unstructured EHR data. Structured EHR data included age, gender, ICD-9 diagnosis codes in different settings, and the number of visits to the emergency department (ED), in-patient (IP) visits (hospitalization), or outpatient (OP) clinic visits. Unstructured data included free-text provider notes for all patients who had at least one note between the years 2011 and 2013. We did not have any limitations in selecting the provider notes and only excluded lab results and radiology and pathology reports. We explored the use of text mining techniques (i.e., pattern matching using RegEx) to determine housing challenges in the unstructured data. The institutional review board at Johns Hopkins Bloomberg School of Public Health reviewed and approved this project. Written informed consent from the participants of the study was not required by local legislation and national guidelines.

Identifying SDOH Challenges

The data custodian identified housing issues as a growing source of challenge in their population. To address this need, the research team reviewed published articles in peer-reviewed journals, using PubMed as the preferred database. After reviewing the available evidence on housing challenges with high-impact on healthcare utilization and outcomes and consulting the subject matter experts we decided to determine five categories of housing challenges. The categories included: homelessness current (HC), homelessness history (HH), homelessness addressed (HA), housing instability (HI), and building quality (BQ). **Figure 1** presents each selected category, how we defined each category, and the type of phrases associated with each one in the EHR.

Homelessness was split into three distinct categories due to different operational interventions (clinical and social) addressing each category. For example, referring a patient to a homeless shelter does not apply if the patient only has a history of homelessness but is not currently homeless. Also homelessness status of a patient may change or a patient may have two or more homelessness statuses. We addressed this issue by reporting the housing challenges at a note level, when each encounter of homelessness was counted separately, and at a patient-level when a patient was considered homeless if they had at least one encounter of homelessness. We did not report the longitudinal change in the homelessness status for each patient. The HC and HH status was linked to the specific encounter when they were documented in the EHR and were reported both at a note level and patient level.

Generating Phrases for Each Housing Category

To identify notes containing housing issues, we used hand-crafted linguistic patterns that a team of experts developed. We first reviewed ICD-10, Current Procedural Terminology (CPT), LOINC codes, Systematized Nomenclature of Medicine (SNOMED) terminologies (14), and the description of housing issues in public health surveys and instruments [e.g., American Community Survey (26), American Housing Survey (27), The Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE) (28), and the Accountable Health Communities tool from the Center for Medicare and Medicaid Innovation (21)]. We also reviewed phrases derived from a literature review of other studies and the results of a manual annotation process from a past study (7). To craft the linguistic patterns the expert team developed a comprehensive list of all available codes and specific content areas for each selected housing domain and matched them across different coding systems. **Supplementary Table 1** presents examples of available codes and phrases for different categories of housing issues.

The expert team developed phrases based on aspects of the housing issues addressed in the codes, terminologies, and surveys. We further refined those phrases to address potential overlap with clinical phrases as well as learning from the underlying EHR's free-text manual tagging process. We categorized the refined phrases into green, yellow, and red phrases in multiple iterations. Green phrases indicated an active housing challenge referring to the existence of the housing issue during the encounter. Yellow phrases indicated a potential risk for a housing issue but were not conclusive. Red phrases were factors not necessarily correlated with a housing challenge. We only assessed the presence of the green phrases in free-text notes.

Development of the Regular Expression Patterns

We intended to develop a text mining approach that could be used in a healthcare system with minimal effort and no need for advanced computational capacity, hence we used the RegEx (i.e., pattern matching) as our text mining approach. We developed multi-level RegEx patterns using green phrases for each housing category. We then developed a custom web-based application and a backend Structured Query Language (SQL) database to automate the execution of the RegEx patterns, to provide advanced RegEx functionality (e.g., negation, context detection), and for storing/preparing the results for further analysis.

Development of the Training and Validation Dataset

The training dataset included 2,564,344 free-text clinical notes in the EHR of 20, 2017 patients. To develop the validation dataset we selected a sample of 100 patients based on the ICD-9 codes indicating a possible housing issue in their EHR structured data (20 patients for each category of the housing categories). We randomly selected 20 additional patients from the rest of the population who did not have any ICD-9 codes indicating a housing issue in their structured EHR (a total of 120 patients for the validation dataset).

Our SDOH expert (EH) trained two annotators to review and independently tag phrases describing any housing issues in the free-text EHR notes for the selected sample of 120 patients. We further customized an open-source application to pre-highlight keywords referring to housing challenges in the EHR free-text notes of the patients. The annotators initially annotated 3 test patients to assess inter-rater reliability and were consequently further trained to ensure higher agreement levels. Each annotator

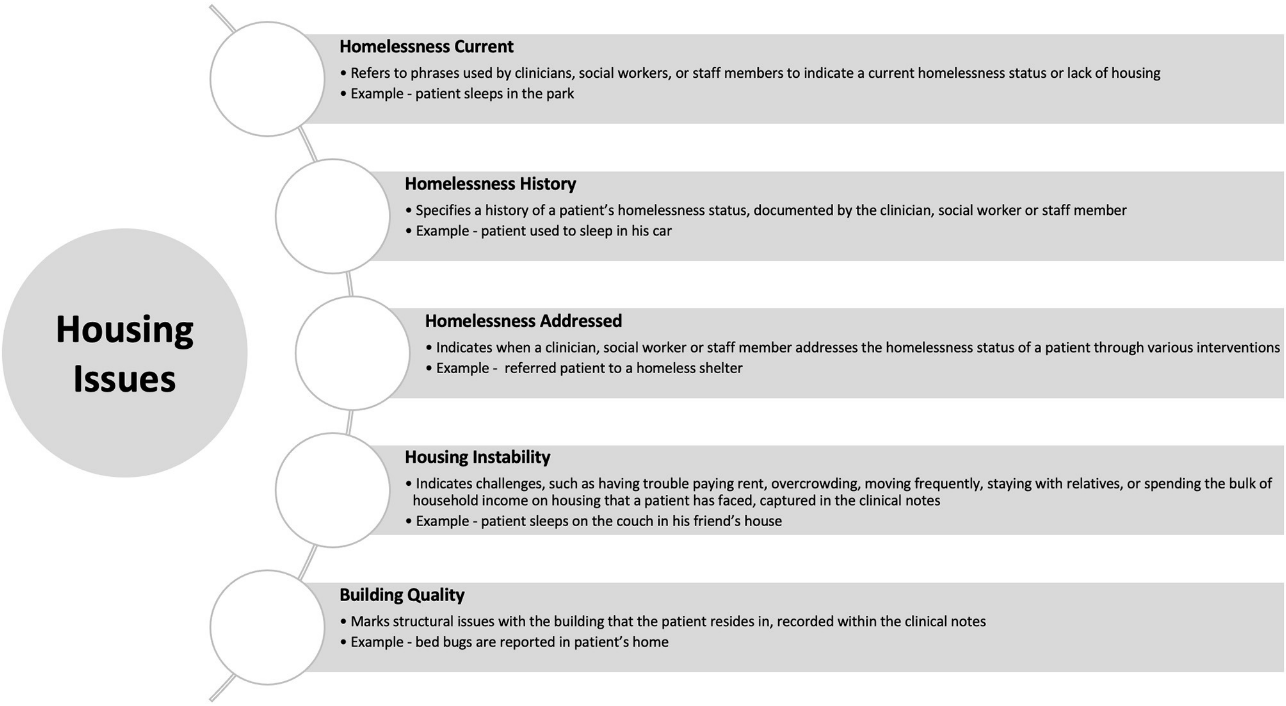

Housing Issues
Homelessness Current
• Refers to phrases used by clinicians, social workers, or staff members to indicate a current homelessness status or lack of housing
• Example - patient sleeps in the park
Homelessness History
• Specifies a history of a patient's homelessness status, documented by the clinician, social worker or staff member
• Example - patient used to sleep in his car
Homelessness Addressed
• Indicates when a clinician, social worker or staff member addresses the homelessness status of a patient through various interventions
• Example - referred patient to a homeless shelter
Housing Instability
• Indicates challenges, such as having trouble paying rent, overcrowding, moving frequently, staying with relatives, or spending the bulk of household income on housing that a patient has faced, captured in the clinical notes
• Example - patient sleeps on the couch in his friend's house
Building Quality
• Marks structural issues with the building that the patient resides in, recorded within the clinical notes
• Example - bed bugs are reported in patient's home

FIGURE 1 | Selected categories of housing challenges, definition of each category, and type of phrases associated with each one in the electronic health record.

manually annotated all EHR records for half of the sample patients using in-house built-in functions of the customized open-source application. A third annotator then reviewed all annotated phrases for potential false positive (FP) cases across all 120 patients.

Assessing the Performance of the Text Mining Approach

We used two different techniques to assess the performance of the RegEx text mining approach. First, we randomly selected and manually assessed 100 phrases per category of housing challenges identified by the RegEx techniques and documented the true positive (TP) and FP instances. Second, we compared the RegEx results against the manually annotated sample of 120 patients. The following sections provide more details of the two approaches.

Assessment #1: True Positive and False Positive Rates Among Random Patients

We first iteratively pruned the raw results of the RegEx technique to reduce potential high FP RegEx patterns. After finalizing the fine-tuning of the RegEx patterns, we extracted 100 random phrases per category of housing challenges from the pruned RegEx results and performed a phrase level assessment to calculate TP and FP rates. **Supplementary Table 2** includes sample phrases found by the RegEx technique. The table lists TP findings (i.e., the RegEx found a correct housing challenge) and FPs (i.e., the RegEx found a phrase that was not a housing challenge – falsely identified as positive) for each housing category (i.e., except homeless history, as RegEx did not find any matches). Some categories did not result in 100 patients hence this assessment was limited to the maximum number of phrases identified by the RegEx pattern technique (e.g., HC only returned 65 phrases hence we assessed 65 phrases for this category). A total of 372 patients were assessed by this methodology across all housing categories. We defined precision as TP/(TP+FP), representing the positive predictive value in the text mining field. This approach did not provide false negative (FN) rates [i.e., missed recall (sensitivity) rate calculations] but offered a larger sample of patients identified by the RegEx patterns (i.e., max 100 phrases times 5 categories).

Assessment #2: Recall (Sensitivity) and Precision (Positive Predictive Value) of the RegEx Model

The second approach, a common evaluation approach in the text mining domain, provided both recall (sensitivity) and precision (positive predictive value) measures for the RegEx technique – as it generated TP, FP, and FN rates – but was limited to 120 sample patients whose EHR records were manually annotated for housing issues. We defined TP as cases where RegEx matched the annotators' tagging (i.e., matching the housing categories) and FP as cases where RegEx found an incorrect phrase that was not annotated by the annotators. FN included phrases that the annotators deemed relevant, but RegEx did not mark them as a housing issue. We calculated TP, FP, and FN at three levels of phrase, note, and patient. We did not use true negative (TN) cases in the assessment due to the large text not being identified or annotated by either method (i.e., RegEx or annotators).

TABLE 1 | Demographic and clinical characteristics of study population categorized by housing issues[a].

	All patients	No housing issues[b]	Homelessness[b]	Housing instability[b]	Building quality[b]
Patient Count	20,342	19,919	125	160	162
Age – mean (SD)					
	75.96 (7.51)	75.90 (7.49)	78.40 (7.86)	78.78 (7.78)	77.98 (7.95)
Gender – female %					
	58.78	58.62	69.6	70.62	61.11
Comorbidity index – mean (SD)					
Charlson[c]	1.66 (1.65)	1.64 (1.64)	2.50 (1.20)	2.69 (2.04)	2.53 (1.20)
Elixhauser[d]	3.84 (2.71)	3.81 (2.69)	5.82 (3.27)	5.91 (3.15)	5.34 (3.20)
Charlson weighted	2.47 (2.72)	2.45 (2.71)	3.56 (3.10)	3.84 (3.32)	3.61(3.13)
Elix weighted AHRQ[e]	5.21 (10.41)	5.14 (10.35)	8.81 (12.15)	8.37 (13.42)	8.74 (12.57)
Elix weighted VW[f]	5.92 (8.55)	5.85 (8.50)	9.57 (9.84)	9.36 (10.16)	9.62 (10.43)
Utilization markers – patient count (%)					
Emergency department	7,103 (34.92)	6,854 (34.45)	78 (62.40)	101 (63.13)	87 (53.70)
Inpatient	4,145 (20.38)	3,969 (19.95)	67 (53.60)	76 (47.50)	48 (29.63)
Outpatient	10,637 (52.29)	10,325 (51.90)	100 (80.00)	125 (78.13)	108 (66.67)

Dx, diagnosis; SD, standard deviation.

[a] *Patients with mentions of any domains of housing issues in their free-text note or those with relevant ICD-9 codes were identified as patients with housing issues.*

[b] *Some patients had multiple housing challenges. Therefore, the sum of figures in the columns for Homelessness, Housing Instability, and Building Quality is higher than the actual number of patients with housing challenges ("All patients – No Housing Issues" column).*

Due to the large sample size, the differences in the demographic and clinical characteristics between patients without housing issues (column 3) and those with different categories of housing issues (columns 4–6) were statistically significant.

[c] *Charlson score is a weighted index that is predictive of the risk of death within 1 year of hospitalization for patients with specific comorbid conditions.*

[d] *Elixhauser score is calculated based on a method of categorizing comorbidities using diagnosis codes found in clinical data, which is predictive of hospital readmission and in-hospital mortality.*

[e] *A version of the Elixhauser score developed by the Agency for Healthcare Research and Quality (AHRQ) (32).*

[f] *A version of the Elixhauser score developed by van Walraven et al. (33).*

We defined recall as TP/(TP+FN) representing the sensitivity concept in the text mining domain and precision as TP/(TP+FP) representing positive predictive value. Due to the lack of TN results in the text mining field, we did not report specificity. We used the basic R function (the R version: 3.5.1) to calculate the recall (sensitivity) and precision (positive predictive value).

Clinical Characteristics and Healthcare Utilization

We assessed the impact of housing issues on healthcare utilization including inpatient, ED, and outpatient visits. We defined (1) the inpatient visits as the acute care inpatient hospitalization stays, regardless of cause excluding pregnancy and delivery, newborns, and injury, (2) ED visits as those that were not the precursors to subsequent observation stays and inpatient hospital stays in the same period, and (3) the outpatient visits as the instances where patients received ambulatory care in outpatient settings. To describe a patient's health status, we assigned each ICD diagnosis code to one or more of 32 diagnosis groups referred to as Aggregated Diagnosis Groups (ADGs) (29) (see **Supplementary Table 3** for more details) and also grouped over 8,600 diagnoses into condition categories. We also calculated the Charlson Comorbidity (30) Index, a weighted index to predict the risk of death within 1 year of hospitalization for patients with specific comorbid conditions. Additionally, we calculated the Elixhauser Comorbidity Index (31), a method of categorizing comorbidities of patients based on ICD diagnosis codes found in administrative data. The ADG, Charlson, and Elixhauser scores were used to measure the burden of chronic conditions and comorbidities in our analysis.

RESULTS

The study data included EHR structured data for a total of 20,342 patients and 2,564,344 free-text clinical notes. The mean age in the study population was 75.96 (SD: 7.51). Additionally, 58.78% of the cohort were female. **Table 1** presents the demographic and clinical characteristics of the total study population and those with and without housing issues. Patients with housing issues were older (mean ages of 78.40, 78.78, and 77.98 years for homeless patients, and those with housing instability and building quality issues) than those with no housing issues (mean age of 75.9 years). Patients with housing issues were more female (69.60%, 70.62%, 61.11% for homeless patients, and those with housing instability and building quality issues) than those with no housing issues (58.62%).

Clinical Characteristics and Healthcare Utilization

Table 1 also presents the results of descriptive analyses for patients with housing issues, those with no housing issues, and the general population. Patients with housing issues were sicker and had higher comorbidity scores than the overall population and those with no housing issues. They also utilized

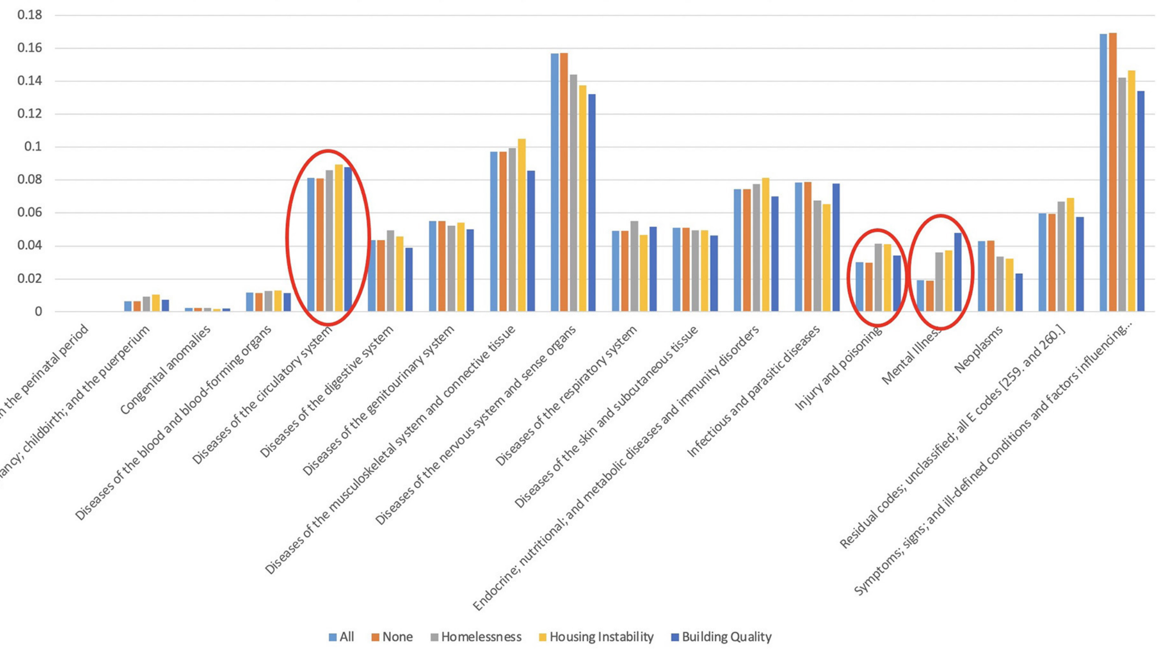


FIGURE 2 | Distribution of housing issues for patients with different clinical conditions in the study population. The groups identified by the red circles have statistically significant differences.

the healthcare services more often. For instance, 62.40% of patients with homelessness, 63.13% of patients with housing instability, and 53.70% of patients with building quality issues had an ED visit during the study period. The ED utilization was at 34.45% for those without housing issues. Among other notable findings was the high number of outpatient visits among patients with housing issues and particularly those with homelessness (80% of patients with homelessness had outpatient visits during the study period). **Figure 2** presents the distribution of housing issues for different clinical conditions. For example, a higher frequency of housing issues was noted among those with a mental illness.

Findings of the RegEx Text Mining Technique

Table 2 depicts the total number of phrases, notes, and patients found for each housing category using the RegEx text mining methodology. The RegEx text mining identified 526 unique phrases, 494 (0.02%) unique notes, and 369 (1.82%) unique patients with housing issues. We did not define the phrase-level denominator hence phrase percentages could not be calculated. Considering the FN rate, we estimated ∼890 (4.40%), unique patients, with any housing issues in our study population. Several patients had more than one housing issue documented in their free-text notes. **Table 3** shows the overlap of housing categories among notes and patients. For example, 21 patients in the HA category also had other housing issues; 9 of them had housing instability and 7 had building quality issues.

Assessing Performance of RegEx Technique

Table 4 presents the results of the performance assessment of the RegEx technique using 100 randomly selected phrases (Assessment #1). Housing instability had the highest precision (positive predictive value) rate of 89%. **Table 5** presents the performance assessment of the RegEx technique at the phrase, note, and patient level using manual annotation (Assessment #2). The RegEx technique had a high level of precision (positive predictive value) at all levels (96.36, 95.00, and 94.44%, respectively) but the recall (sensitivity) rate was relatively low (30.11, 32.20, and 41.46%, respectively).

DISCUSSION

Value-based healthcare systems are increasingly at stake to address the underlying SDOH challenges of the population they serve (24). However, SDOH variables are commonly captured in EHR's free-text which makes the use of this information challenging in operational settings (14). Furthermore, healthcare providers are facing operational challenges in rolling out population-level surveys to collect individual-level SDOH information from their patients (23). Hence, text mining approaches that reveal SDOH factors within EHR's free-text can be helpful to identify patients with SDOH challenges and to implement targeted interventions for patients with such challenges.

EHR data is also gradually playing an instrumental role in the population health management efforts of value-based providers (34, 35). Compared to and in the absence of insurance claims,

TABLE 2 | Total number of cases identified by the RegEx text mining technique[a].

	Phrases[a]		Notes[b]		Patients[c]		Total patients[d]	
Housing categories	**No.**	**%**	**No.**	**%**	**No.**	**%**	**No.**	**%**
Homelessness								
Homelessness current	65	NA	60	0.002	47	0.2325	113	0.5607
Homelessness history	7	NA	7	0.000	4	0.0198	10	0.0477
Homelessness addressed	104	NA	101	0.004	76	0.3759	183	0.9066
Housing instability								
	176	NA	172	0.007	125	0.6183	301	1.4912
Building quality								
	174	NA	165	0.006	146	0.7222	352	1.7417
Total[e]								
Unique	526	NA	494	0.019	369	1.8252	890	4.4019

[a] *Total number of phrases identified in the EHR during the study period describing each category of housing issues. The phrase-level denominator was not defined hence the phrase percentage could not be calculated.*

[b] *Total number of notes (and % of notes) in the EHR during the study period with mentions of housing challenges. The denominator included the total number of notes in the EHR during the study period.*

[c] *Total number of patients (and % of patients) in the EHR during the study period with mentions of housing challenges. The denominator included the total number of patients in the EHR during the study period.*

[d] *Total number (and % of patients) with housing issues after considering estimated false-negative rates, assuming a 41.46% recall (sensitivity) rate for patient-level RegEx analysis (see **Table 4**).*

[e] *Unique number of phrases, notes, and patients with mentions of housing challenges in the EHR during the study period. The phrase-level denominator was not defined hence the phrase percentage could not be calculated. Some notes contained more than one housing issue and some patients reported more than one housing challenge. Therefore, the numbers are different than the sum of all categories together.*

RegEx, regular expressions.

TABLE 3 | Total number of housing issue overlaps identified by the regex text mining technique.

	Notes					Patients				
Category	**HC**	**HH**	**HA**	**HI**	**BQ**	**HC**	**HH**	**HA**	**HI**	**BQ**
Homelessness current		0	2	0	0		0	5	4	4
Homelessness history	0		0	0	0	0		0	0	0
Homelessness addressed	2	0		8	1	5	0		9	7
Housing instability	0	0	8		1	4	0	9		6
Building quality	0	0	1	1		4	0	7	6	
Total number	2	0	11	9	2	13	0	21	19	17
Total %[a]	3.33	0	10.89	5.23	1.21	27.66	0	27.63	15.2	11.64

[a] *% of notes and patients with housing issues overlaps. The denominator is the total number of notes and patients with each category of housing issues (see **Table 2** for total numbers in each category).*

HC, homelessness current; HH, homelessness history; HA, homeless addressed; HI, housing instability; BQ, building quality; RegEx, regular expressions.

EHRs provide additional data types that can be utilized for risk stratification efforts (34, 36–39). EHR-derived SDOH data, such as housing challenges, can potentially help to improve these risk stratification efforts, although certain challenges such as potential immaturity of EHR's functionality across providers (40–42). SDOH data quality issues (43), and the need for complex text mining methods to extract SDOH from EHR's free-text should be addressed (7, 44). Moreover, as population health management efforts are gradually aligning clinical outcomes with public health goals (45–48), identifying SDOH factors of high-risk patients will be key in addressing underlying disparities within populations residing in states with statewide population-level global budgets such as Massachusetts (20) and Maryland (49, 50). Value-based providers may also utilize non-EHR data sources to access SDOH information (e.g., health information exchange) (51).

Therefore, the development of text mining approaches that could help extraction of SDOH information from EHR of a healthcare system regularly and could be generalizable across different healthcare systems would provide an operational solution to using this arguably largest source of SDOH information in the healthcare system. In this study, we exercised this approach by utilizing a pragmatic text-mining methodology (i.e., RegEx) and identified various phrases in EHR's free-text that reflected five categories of housing issues (i.e., three categories of homelessness, housing instability, and building quality). Our RegEx algorithm identified 369 unique patients (1.82% of the

TABLE 4 | Performance assessment of the RegEx text mining technique using 100 random phrases.

Category	Phrase level assessment		
	True positive[a]	False positive[a]	Precision (positive predictive value) %
Homelessness current	37	28	56.92
Homelessness history	0	7	0.00
Homelessness addressed	66	34	66.00
Housing instability	89	11	89.00
Building quality	58	42	58.00

[a]Number of phrases in each category of housing issues.
RegEx, regular expressions.

TABLE 5 | Performance assessment of the RegEx text mining technique using manual annotation.

Measure	Assessment level		
	Phrase	Note	Patient
True positive[a]	53	38	17
False positive[a]	2	2	1
False negative[a]	123	80	24
Recall (sensitivity) %	30.11	32.20	41.46
Precision (positive predictive value)%	96.36	95.00	94.44

[a]Number of phrases, notes, and patients in each category.
RegEx, regular expressions.

study population) with housing issues. Considering the 41.46% recall (sensitivity) of the RegEx patterns among the 120 manually annotated patients, total unique patients with housing issues after adding the estimated FNs were calculated at 890 (∼4.40% of the study population). In other words, the study results showed that potentially 1 in 20 patients in our study population had a housing issue.

Furthermore, to present how this text mining approach would help the health systems to address the SDOH challenges of their patients we assessed the demographic and clinical characteristics of patients with and without housing issues and briefly look into the patterns of healthcare utilization among the study population and for those with and without housing challenges. In our study population patients with housing issues were older (mean age of ∼78 years across three categories of housing issues and ∼76 years among those with no housing issues), had a higher number of comorbidities (e.g., Charlson Comorbidity Index of ∼2.5 across three categories of housing issues and ∼1.6 among those with no housing issues), and utilized the healthcare services more often (e.g., ∼54–63% ED utilization among patients with housing issues vs. ∼34% among those with no housing issues). This information would help care managers, care coordinators, and social workers to tailor specific social interventions and/or conducting referrals to community-based social services organizations (52, 53). Clinicians can also utilize such information to explore the underlying housing issues at the point of care, and population health experts might use this information to better predict utilization rates associated with such patient population (54).

We provided a comprehensive approach to the performance assessment of our RegEx technique. We first assessed the performance by selecting 100 random phrases from each category of housing issues. This approach showed a precision (positive predictive value) of ∼57–89% across five housing categories. We also performed manual annotation on free-text notes of 120 patients (100 patients with housing issues based on the ICD-9 codes indicating a possible housing issue in their EHR structured data, 20 patients for each category of the housing categories, and a random sample of 20 additional patients who did not have any ICD-9 codes indicating a housing issue in their structured EHR). The manual annotation revealed high precision (positive predictive value) of the RegEx technique at the phrase, note, and patient-level (∼96, 95, and 94%, respectively). But the recall (sensitivity) was low at the phrase, note, and patient-level (∼30, 32, and 41%, respectively). The RegEx pattern matching approach that we applied in this study is considered a basic text mining technique with rigid flexibility and potentially high FN rates. For instance, any housing phrases not embedded in the RegEx patterns will be missed in the results. The high FN rates resulted in low recall (sensitivity) for the text mining technique and the RegEx algorithm failed to identify a high number of patients with actual housing issues. However, the high precision (positive predictive value) helped to know, with high certainty, that those identified as patients with housing issues indeed were suffering from those challenges.

Manually tagging EHR's free-text for SDOH variables is an exhausting task involving several annotators spending hundreds of hours to generate the "gold standard" text. Manually annotated gold standard text is required to both assess RegEx techniques as well as train, test and evaluate advanced NLP techniques. EHR data sources that also include survey-level SDOH information will be critical in future SDOH NLP research as survey data can be treated/assumed as the gold standard text, hence enabling researchers to train, test, and evaluate the accuracy of their NLP methods. This approach might result in lower false-negative instances and improve the recall (sensitivity) of the text mining/ NLP techniques. Alternatively, approximated SDOH factors associated with the residential location/address of patients can be assessed as a proxy to train and/or validate advanced NLP techniques (e.g., compare the NLP results with SDOH variables derived based on patient's residential address).

Our results were slightly different from other studies using rule-based systems to identify social needs in free-text provider notes. For instance, Conway et al. (55) tested the performance of Moonstone, a new, highly configurable rule-based clinical NLP system for extraction of information requiring inferencing from clinical notes derived from the Veterans Health Administration. Their system achieved a precision (positive predictive value) of 0.66 (lower than ∼94–96% at the phrase, note, and patient-level in our study) and a recall (sensitivity) of 0.87 (higher than ∼30-41% at phrase, note, and patient-level in our study) for phrases related to homelessness and marginally housed.

In another study, Dorr et al. (56) extracted the phenotypic profiles for four key psychosocial vital signs including housing insecurity or homelessness from EHR data. They used lexical associations expanded by expert input, then, for each psychosocial vital sign, and manually reviewed the retrieved charts. Their system achieved a precision (positive predictive value) of >0.90 in all psychosocial vital signs except for social isolation. Navathe et al. (8) utilized MTERMS, an NLP system validated for identifying clinical terms within medical record text to extract social factor information from physician notes. They customized and developed the MTERMS NLP system on a randomized 500 annotated physician note training set and tested the diagnostic characteristics. After development, they validated the system by studying the diagnostic characteristics of the system vs. a gold standard manual review of a new set of randomized 600 physician notes. They achieved a precision (positive predictive value) of 1.0 and a recall (sensitivity) of 0.66 for housing instability.

While beyond the scope of this study, future efforts should also incorporate more advanced text mining approaches such as statistical NLP techniques (e.g., embedding, word2vec, and deep learning such as the recursive neural network). Recent studies utilizing such advanced NLP techniques have shown promising results in identifying syndromes not encoded properly by EHR's structured data elements (44, 57–59), Other approaches such as creating text preparation tasks may help to improve the results of the text-mining/NLP techniques. These tasks may include detecting clinical templates and repeated copy/pastes of some information in the text. They may also include detecting various sections of clinical notes that may result in the detection of false positive or false negative phrases. For instance, omitting family history of SDOH challenges and keeping the mentions of patient's specific SDOH issues may result in lower false positive or false negative instances.

This study had several limitations. First, we only identified predefined housing phrases in the EHR's free text. Second, we did not use a statistical NLP approach to assess the likelihood of notes or patients having similar phrases addressing categories of housing issues. Hence, we could not calculate the TN rates for patients and notes with housing issues. Third, we only measured the stratified rates of comorbidity and utilization among patients having any phrases related to housing issues in their free-text notes.

Moreover, we did not evaluate the net effect of the housing issues on healthcare utilization using multivariate analysis. Future research should analyze the effect of housing issues on long-term healthcare utilization while adjusting for clinical variables. The study period might also limit the study results. As in the last few years, there have been a growing number of providers and practices that actively plan for assessing and documenting the SDOH challenges in the EHR. Therefore, there might be a higher number of TP and TN instances of housing issues in the free-text EHR, which would impact the performance of the text mining techniques.

Finally, we measured the availability of housing issues regardless of the underlying socioeconomic status of the patients. Future research should expand on the underlying population denominator to patients in high need of SDOH interventions (e.g., Medicaid patients) as well as comparing NLP results with geo-driven SDOH factors (e.g., comparing the neighborhood-level housing issues, measured based on patient's residential address, with individual-level social needs found in the EHR's free-text notes).

CONCLUSION

This study assessed the use of a pragmatic text mining methodology in identifying various SDOH housing factors in EHR's free text. The study results revealed a high precision (positive predictive value) for the assessed text mining approach but the recall (sensitivity) was low. The simplicity of this approach suggests its generalizability across the healthcare systems. The development of generalizable text mining methodologies with promising performance will enhance the value of EHRs to identify at-risk patients across different health systems, improve patient care and outcomes, and eventually mitigating socioeconomic disparities across individuals and communities.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: The dataset includes patients' information in electronic health records, which are confidential to patients and their providers. Requests to access these datasets should be directed to Elham Hatef, ehatef1@jhu.edu.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Institutional Review Board, Johns Hopkins Bloomberg School of Public Health. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR'S NOTE

This study attempts to evaluate a practical text mining approach (i.e., advanced pattern matching techniques) in identifying phrases referring to housing issues, an important SDOH domain affecting value-based healthcare providers, using EHR of a large multispecialty medical group in the New England region, United States.

AUTHOR CONTRIBUTIONS

EH supervised the development of the analysis plan, reviewed and interpreted the results, and led writing this paper. GS and MR performed the data analysis. AL, KE, CM, RB, and MS contributed to setting the overall scope and goal of the project as well as finalizing the manuscript. HK designed the overall scope and goals of the study and supervised the day-to-day operations of the project. All authors contributed significantly to the project and writing of the manuscript, reviewed the final paper, and provided comments as deemed necessary.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.697501/full#supplementary-material

REFERENCES

1. Office of National Coordinator for Health Information Technology (ONC-HIT). *Hospital Adoption of Electronic Health Record Technology to Meet Meaningful Use Objectives: 2008-2012.* (2013). Available online at: http://www.healthit.gov/sites/default/files/oncdatabrief10final.pdf (accessed March 30, 2021).
2. The Office of the National Coordinator for Health Information Technology (ONC). *Office-Based Physician Electronic Health Record Adoption. Health IT Quick-Stat #50.* (2016). Available online at: https://dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php (accessed March 30, 2021).
3. Blumenthal D, Tavenner M. The meaningful use regulation for electronic health records. *N Engl J Med.* (2010) 363:501–4. doi: 10.1056/NEJMp1006114
4. Jha AK. Meaningful use of electronic health records: the road ahead. *JAMA.* (2010) 304:1709–10. doi: 10.1001/jama.2010.1497
5. Jha AK, Burke MF, DesRoches C, Joshi MS, Kralovec PD, Campbell EG, et al. Progress toward meaningful use: hospitals' adoption of electronic health records. *Am J Manag Care.* (2011) 17:SP117–24.
6. The Office of the National Coordinator for Health Information Technology. *Meaningful Use Definition & Objectives*. Available online at: https://www.healthit.gov/providers-professionals/meaningful-use-definition-objectives (accessed March 30, 2021).
7. Kharrazi H, Anzaldi LJ, Hernandez L, Davison A, Boyd CM, Leff B, et al. The value of unstructured electronic health record data in geriatric syndrome case identification. *J Am Geriatr Soc.* (2018) 66:1499–507. doi: 10.1111/jgs.15411
8. Navathe AS, Zhong F, Lei VJ, Chang FY, Sordo M, Topaz M, et al. Hospital readmission and social risk factors identified from physician notes. *Health Serv Res.* (2018) 53:1110–36. doi: 10.1111/1475-6773.12670
9. Liwei W, Xiaoyang R, Ping Y, Hongfang L. Comparison of three information sources for smoking information in electronic health records. *Cancer Inform.* (2016) 2016:237–42. doi: 10.4137/CIN.S40604
10. Torres JM, Lawlor J, Colvin JD, Sills MR, Bettenhausen JL, Davidson A, et al. ICD social codes: an underutilized resource for tracking social needs. *Med Care.* (2017) 55:810–6. doi: 10.1097/MLR.0000000000000764
11. Oreskovic NM, Maniates J, Weilburg J, Choy G. Optimizing the use of electronic health records to identify high-risk psychosocial determinants of health. *JMIR Med Inform.* (2017) 5:e25. doi: 10.2196/medinform.8240
12. Hripcsak G, Forrest CB, Brennan PF, Stead WW. Informatics to support the IOM social and behavioral domains and measures. *J Am Med Inform Assoc.* (2015) 22:921–4. doi: 10.1093/jamia/ocv035
13. Cantor MN, Thorpe L. Integrating data on social determinants of health into electronic health records. *Health Affairs.* (2018) 37:585–90. doi: 10.1377/hlthaff.2017.1252
14. Arons A, DeSilvey S, Fichtenberg C, Gottlieb L. Documenting social determinants of health-related clinical activities using standardized medical vocabularies. *JAMIA Open.* (2019) 2:81–8. doi: 10.1093/jamiaopen/ooy051
15. Regenstrief Institute. *LOINC.* (2018). Available online at: https://loinc.org/sdh/ (accessed March 30, 202).
16. Kharrazi H, Hatef E, Lasser E, Woods B, Rouhizadeh M, Kim J, DeCamp L. *A Guide to Using Data From Johns Hopkins Epic Electronic Health Record for Behavioral, Social and Systems Science Research*. Baltimore: Johns Hopkins Medical Institute (2018).
17. Bae J, Ford EW, Kharrazi HH, Huerta TR. Electronic medical record reminders and smoking cessation activities in primary care. *Addict Behav.* (2018) 77:203–9. doi: 10.1016/j.addbeh.2017.10.009
18. Anzaldi LJ, Davison A, Boyd CM, Leff B, Kharrazi H. Comparing clinician descriptions of frailty and geriatric syndromes using electronic health records: a retrospective cohort study. *BMC Geriatr.* (2017) 17:248. doi: 10.1186/s12877-017-0645-7
19. Nichols LM, Taylor LA. Social determinants as public goods: a new approach to financing key investments in healthy communities. *Health Affairs.* (2018) 37:1223–30. doi: 10.1377/hlthaff.2018.0039
20. Ash A, Mick E, Ellis R, Kiefe C, Allison J, Clark M. Social determinants of health in managed care payment formulas. *JAMA Intern Med.* (2017) 177:1424–30. doi: 10.1001/jamainternmed.2017.3317
21. Centers for Medicare and Medicaid Services. *The Accountable Health Communities Health-Related Social Needs Screening Tool. AHC Screening Tool.* (2019). Available online at: https://innovation.cms.gov/Files/worksheets/ahcm-screeningtool.pdf (accessed March 30, 2021).
22. North Carolina Department of Health and Human Services. *Using Standardized Social Determinants of Health Screening Questions to Identify and Assist Patients With Unmet Health-Related Resource Needs in North Carolina. SDOH Screening Tool.* (2018). Available online at: https://files.nc.gov/ncdhhs/documents/SDOH-Screening-Tool_Paper_FINAL_20180405.pdf (accessed March 30, 2021).
23. LaForge K, Gold R, Cottrell E, Bunce AE, Proser M, Hollombe C, et al. How 6 organizations developed tools and processes for social determinants of health screening in primary care: an overview. *J Ambul Care Manage.* (2018) 41:2–14. doi: 10.1097/JAC.0000000000000221
24. Breslin E, Lambertino A, Heaphy D, Dreyfus T. *Medicaid and Social Determinants of Health: Adjusting Payment and Measuring Health Outcomes.* (2017). Available online at: https://www.healthmanagement.com/wp-content/uploads/SHVS_SocialDeterminants_HMA_July2017.pdf (accessed March 30, 2021).
25. Alley D, Asomugha C, Conway P, Sanghavi D. Accountable health communities—addressing social needs through medicare and medicaid. *New Eng J Med.* (2019) 347:8–11. doi: 10.1056/NEJMp1512532
26. U.S. Census Bureau. *American Community Survey.* (2019). Available online at: https://www.census.gov/programs-surveys/acs/ (accessed March 30, 2021).
27. U.S. Census Bureau. *American Housing Survey.* (2019). Available online at: https://www.census.gov/programs-surveys/ahs.html. (accessed March 30, 2021).
28. National Association of Community Health Centers. *The Protocol for Responding to and Assessing Patients' Assets, Risks, and Experiences (PRAPARE).* (2019). Available online at: http://www.nachc.org/research-and-data/prapare/ (accessed March 30, 2021).
29. *The Johns Hopkins ACG® System Version 12.0 User Documentation*. Available online at:https://www.hopkinsacg.org/document/acg-system-version-12-0-system-documentation-all-guides/ (accessed March 30, 2021).
30. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis.* (1987) 40:373–83. doi: 10.1016/0021-9681(87)90171-8
31. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. *Med Care.* (1998) 36:8–27. doi: 10.1097/00005650-199801000-00004
32. Moore BJ, White S, Washington R, Coenen N, Elixhauser A. Identifying increased risk of readmission and in-hospital mortality using hospital administrative data: the AHRQ elixhauser comorbidity index. *Med Care.* (2017) 55:698–705. doi: 10.1097/MLR.0000000000000735
33. van Walraven C, Austin PC, Jennings A, Quan H, Forster AJ. A modification of the elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med Care.* (2009) 47:626–33. doi: 10.1097/MLR.0b013e31819432e5
34. Kharrazi H, Chi W, Chang HY, Richards TM, Gallagher JM, Knudson SM, et al. Comparing population-based risk-stratification model performance using demographic, diagnosis and medication data extracted from outpatient

electronic health records versus administrative claims. *Med Care.* (2017) 55:789–96. doi: 10.1097/MLR.0000000000000754

35. Kharrazi H, Weiner JP. A practical comparison between the predictive power of population-based risk stratification models using data from electronic health records versus. *Med Care.* (2018) 56:202–3. doi: 10.1097/MLR.0000000000000849
36. Kan HJ, Kharrazi H, Leff B, Boyd C, Davison A, Chang HY, et al. Defining and assessing geriatric risk factors and associated health care utilization among older adults using claims and electronic health records. *Med Care.* (2018) 56:233–9. doi: 10.1097/MLR.0000000000000865
37. Chang HY, Richards TM, Shermock KM, Elder Dalpoas S, J Kan H, Alexander GC, et al. Evaluating the impact of prescription fill rates on risk stratification model performance. *Med Care.* (2017) 55:1052–60. doi: 10.1097/MLR.0000000000000825
38. Lemke KW, Gudzune KA, Kharrazi H, Weiner JP. Assessing markers from ambulatory laboratory tests for predicting high-risk patients. *Am J Manag Care.* (2018) 24:e190–5.
39. Kharrazi H, Chang HY, Heins SE, Weiner JP, Gudzune KA. Assessing the impact of body mass index information on the performance of risk adjustment models in predicting health care costs and utilization. *Med Care.* (2018) 56:1042–50. doi: 10.1097/MLR.0000000000001001
40. Kharrazi H, Gonzalez CP, Lowe KB, Huerta TR, Ford EW. Forecasting the maturation of electronic health record functions among US hospitals: retrospective analysis and predictive model. *J Med Internet Res.* (2018) 20:e10458. doi: 10.2196/10458
41. Adler-Milstein J, DesRoches CM, Kralovec P, Foster G, Worzala C, Charles D, et al. Electronic health record adoption in US hospitals: progress continues, but challenges persist. *Health Aff.* (2015) 34:2174–80. doi: 10.1377/hlthaff.2015.0992
42. Chan KS, Kharrazi H, Parikh MA, Ford EW. Assessing electronic health record implementation challenges using item response theory. *Am J Manag Care.* (2016) 22:e409–15.
43. Kharrazi H, Wang C, Scharfstein D. Prospective EHR-based clinical trials: the challenge of missing data. *J Gen Intern Med.* (2014) 29:976–8. doi: 10.1007/s11606-014-2883-0
44. Chen T, Dredze M, Weiner J, Hernandez L, Kimura J, Kharrazi H. Extraction of geriatric syndromes from electronic health record clinical notes: assessment of statistical natural language processing methods. *JMIR Med Inform.* (2019) 7:e13039. doi: 10.2196/13039
45. Dixon BE, Kharrazi H, Lehmann HP. Public health and epidemiology informatics: recent research and trends in the United States. *Yearb Med Inform.* (2015) 10:199–206. doi: 10.15265/IY-2015-012
46. Gamache R, Kharrazi H, Weiner JP. Public and population health informatics: the bridging of big data to benefit communities. *Yearb Med Inform.* (2018) 27:199–206. doi: 10.1055/s-0038-1667081
47. Kharrazi H, Weiner JP. IT-enabled community health interventions: challenges, opportunities, and future directions. *EGEMS.* (2014) 2:1117. doi: 10.13063/2327-9214.1117
48. Kharrazi H, Lasser EC, Yasnoff WA, Loonsk J, Advani A, Lehmann HP, et al. A proposed national research and development agenda for population health informatics: summary recommendations from a national expert workshop. *J Am Med Inform Assoc.* (2017) 24:2–12. doi: 10.1093/jamia/ocv210
49. Hatef E, Lasser EC, Kharrazi HH, Perman C, Montgomery R, Weiner JP. A population health measurement framework: evidence-based metrics for assessing community-level population health in the global budget context. *Popul Health Manag.* (2018) 21:261–70. doi: 10.1089/pop.2017.0112
50. Hatef E, Kharrazi H, VanBaak E, Falcone M, Ferris L, Mertz K, et al. A state-wide health it infrastructure for population health: building a community-wide electronic platform for maryland's all-payer global budget. *Online J Public Health Inform.* (2017) 9:e195. doi: 10.5210/ojphi.v9i3.8129
51. Kharrazi H, Horrocks D, Weiner JP. Use of HIEs for value-based care delivery: a case study of Maryland's HIE. In: Dixon B, editors. *Health Information Exchange: Navigating and Managing a Network of Health Information Systems.* Amsterdam: AP Elsevier (2016).
52. Lasser EC, Kim JM, Hatef E, Kharrazi H, Marsteller JA, DeCamp LR. Social and behavioral variables in the electronic health record: a path forward to increase data quality and utility. *Acad Med.* (2021) 96:1050–6. doi: 10.1097/ACM.0000000000004071
53. Hatef E, Ma X, Rouhizadeh M, Singh G, Weiner JP, Kharrazi H. Assessing the impact of social needs and social determinants of health on health care utilization: using patient- and community-level data. *Popul Health Manag.* (2021) 24:222–30. doi: 10.1089/pop.2020.0043
54. Tan M, Hatef E, Taghipour D, Vyas K, Kharrazi H, Gottlieb L, et al. Including social and behavioral determinants in predictive models: trends, challenges, and opportunities. *JMIR Med Inform.* (2020) 8:e18084. doi: 10.2196/18084
55. Conway M, Keyhani S, Christensen L, South BR, Vali M, Walter LC, et al. Moonstone: a novel natural language processing system for inferring social risk from clinical narratives. *J Biomed Semantics.* (2019) 10:6. doi: 10.1186/s13326-019-0198-0
56. Dorr D, Bejan CA, Pizzimenti C, Singh S, Storer M, Quinones A. Identifying patients with significant problems related to social determinants of health with natural language processing. *Stud Health Technol Inform.* (2019) 264:1456–7. doi: 10.3233/SHTI190482
57. Zolnoori M, Fung KW, Patrick TB, Fontelo P, Kharrazi H, Faiola A, et al. A systematic approach for developing a corpus of patient reported adverse drug events: a case study for SSRI and SNRI medications. *J Biomed Inform.* (2019) 90:103091. doi: 10.1016/j.jbi.2018.12.005
58. Zolnoori M, Fung KW, Fontelo P, Kharrazi H, Faiola A, Wu YSS, et al. Identifying the underlying factors associated with patients' attitudes toward antidepressants: qualitative and quantitative analysis of patient drug reviews. *JMIR Ment Health.* (2018) 5:e10726. doi: 10.2196/10726
59. Bettencourt-Silva JH, Mulligan N, Sbodio M, Segrave-Daly J, Williams R, Lopez V, et al. Discovering new social determinants of health concepts from unstructured data: framework and evaluation. *Stud Health Technol Inform.* (2020) 270:173–7. doi: 10.3233/SHTI200145

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

SYSTEMATIC REVIEW
published: 29 October 2021
doi: 10.3389/fpubh.2021.741334

Racial and Ethnic Disparities in Health Outcomes Among Long-Term Survivors of Childhood Cancer: A Scoping Review

Tegan J. Reeves[1], *Taylor J. Mathis*[1], *Hailey E. Bauer*[1], *Melissa M. Hudson*[1,2], *Leslie L. Robison*[1], *Zhaoming Wang*[1,3], *Justin N. Baker*[2] and *I-Chan Huang*[1]*

[1] *Department of Epidemiology and Cancer Control, St. Jude Children's Research Hospital, Memphis, TN, United States,* [2] *Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN, United States,* [3] *Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN, United States*

The five-year survival rate of childhood cancer has increased substantially over the past 50 yr; however, racial/ethnic disparities in health outcomes of survival have not been systematically reviewed. This scoping review summarized health disparities between racial/ethnic minorities (specifically non-Hispanic Black and Hispanic) and non-Hispanic White childhood cancer survivors, and elucidated factors that may explain disparities in health outcomes. We used the terms "race", "ethnicity", "childhood cancer", "pediatric cancer", and "survivor" to search the title and abstract for the articles published in PubMed and Scopus from inception to February 2021. After removing duplicates, 189 articles were screened, and 23 empirical articles were included in this review study. All study populations were from North America, and the mean distribution of race/ethnicity was 6.9% for non-Hispanic Black and 4.5% for Hispanic. Health outcomes were categorized as healthcare utilization, patient-reported outcomes, chronic health conditions, and survival status. We found robust evidence of racial/ethnic disparities over four domains of health outcomes. However, health disparities were explained by clinical factors (e.g., diagnosis, treatment), demographic (e.g., age, sex), individual-level socioeconomic status (SES; e.g., educational attainment, personal income, health insurance coverage), family-level SES (e.g., family income, parent educational attainment), neighborhood-level SES (e.g., geographic location), and lifestyle health risk (e.g., cardiovascular risk) in some but not all articles. We discuss the importance of collecting comprehensive social determinants of racial/ethnic disparities inclusive of individual-level, family-level, and neighborhood-level SES. We suggest integrating these variables into healthcare systems (e.g., electronic health records), and utilizing information technology and analytics to better understand the disparity gap for racial/ethnic minorities of childhood cancer survivors. Furthermore, we suggest national and local efforts to close the gap through improving health insurance access, education and transportation aid, racial-culture-specific social learning interventions, and diversity informed training.

Keywords: childhood cancer survivor, ethnicity, health disparities, health outcomes, race

OPEN ACCESS

Edited by:
Yi Guo,
University of Florida, United States

Reviewed by:
Xue Han,
Bristol Myers Squibb, United States
Huan Kuang,
University of Florida, United States

****Correspondence:***
I-Chan Huang
i-chan.huang@stjude.org

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: *15 July 2021*
Accepted: *30 September 2021*
Published: *29 October 2021*

Citation:
Reeves TJ, Mathis TJ, Bauer HE, Hudson MM, Robison LL, Wang Z, Baker JN and Huang I-C (2021) Racial and Ethnic Disparities in Health Outcomes Among Long-Term Survivors of Childhood Cancer: A Scoping Review. Front. Public Health 9:741334. doi: 10.3389/fpubh.2021.741334

INTRODUCTION

The shifting racial/ethnic makeup of the population indicates that in as early as 2045 the United States (US) will become a "minority majority population" country (1). By 2060, non-White individuals will make up more than 60% of the population (2). Racial/ethnic disparities in health are the race- and ethnicity-specific illnesses, injuries, or mortality (3) that disproportionately impact the marginalized groups. The growing diversity of the US population will be accompanied by growing health disparities.

Health equity is considered one of the four basic human rights (3), yet determinants of the inequity and effective implementation strategies to improve racial/ethnic disparity in health outcomes are still limited. Although life expectancy of the US general population has steadily increased since the 1950s in the US, non-Hispanic Black individuals have a 40% higher overall mortality rates (4) and non-Hispanic Black and Hispanic populations have a higher burden of chronic health conditions (e.g., cancer, heart disease, diabetes) (5) compared to non-Hispanic White individuals. In addition, minority populations often have lower healthcare utilization and access to quality care (6). In the context of oncology, health disparities in the US are significantly different in the rates of cancer screening, incidence, survival, treatment-related complications, and quality-of-life (7). Although the 5-yr relative cancer survival rates of childhood cancer have reached 94% among the child and 85% among adolescent survivors (8), there is evidence of lower patient-reported outcomes and survival rates (9–11) in minority vs. non-Hispanic White survivors. Furthermore, the mechanisms behind these disparities are understudied.

While scholars have yet to agree on specifics, it is clear that health disparities are influenced by multiple factors. Some argue that socioeconomic status (SES) is a stronger determinant of health outcomes than race *per se* (12). Others suggest that cultural (13) or population-level factors (14) contribute to health disparities in childhood cancer survivors. Neighborhoods with a higher proportion of Black or Hispanic residents are associated with higher poverty due to a lack in community investment and built environments (i.e., fast food restaurants, liquor stores, lack of green space) which decrease the opportunities for healthy eating and exercises (15). Disadvantaged neighborhood conditions (e.g., high crime rate, poor community support, collective efficacy or social capital) have shown elevated mortality through the mechanisms of practicing health behaviors (16).

The main objective of this study was to summarize the evidence of racial/ethnic disparities in health outcomes for survivors of childhood cancer based on a scoping review of previously published literature. We focused on race/ethnicity as the primary variable determining disparities in health outcomes, and viewed SES factors as confounding or mediating variables that explain the associations between race/ethnicity and health outcomes. This is because the vast majority of the studies selected are based on the cross-sectional design, and the true effect of SES factors on health outcomes cannot be determined (e.g., survivors having lower incomes may develop worse chronic health conditions, and worse health conditions may further lower survivors' incomes). Specifically, we aimed to elucidate the role that personal/family/community-level SES factors alongside other demographic and clinical factors might play to explain the associations between race/ethnicity and health outcomes. Based on these findings, we made recommendations toward improving health disparities for minority childhood cancer survivors, especially by identifying modifiable social determinants of health using information technology, integrating social determinant information into healthcare systems, and suggesting potential interventions for health outcomes improvement.

METHODS

In line with our aims, a scoping review was performed to aggregate evidence from empirical studies. Scoping reviews are particularly useful for complex/diverse issues (17) such as race/ethnicity, and generally precede systematic and meta-analyses (18).

Article Selection and Screening Process

We performed a literature search process according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) (19). Two independent investigators researched the title and abstract for articles published between inception and February 2021 in the PubMed and Scopus using the terms "race", "ethnicity", "childhood cancer", "pediatric cancer", and "survivor". In addition, the search was limited to articles published in the English language and available in full-text. The initial search yielded 26 articles from PubMed and 173 articles from Scopus. After removing duplicates, a combined total of 189 articles were prepared for screening.

Two independent investigators screened the articles for inclusion if these articles included the following criteria: health disparity (i.e. difference in outcome based on race/ethnicity), health outcomes/late effects, and any age range of the survivorship stage. We excluded articles if they met the following criteria: not reporting health outcomes/late effects (90 articles), race/ethnicity listed but only in descriptive statistics (64 articles), no full-text available (4 articles), and absence of IRB approval or otherwise proof of rigor (e.g., qualitative, opinion, review, briefs, and meta-analysis; 6 articles). In addition, two Swiss articles were removed from subsequent review because they did not include non-Hispanic Black or Hispanic survivors. We included two articles that use the term "non-White" (i.e., a combined concept for non-Hispanic Black and Hispanic) into our review.

Data Charting

We extracted data from each article according to the study design, race/ethnicity, health outcomes, and risk modulators of racial/ethnic disparity in health outcomes. We focused on marginalized/minoritized (20). US categories of non-Hispanic Black and Hispanic, and reported its association with health disparity.

We classified health outcomes of childhood cancer survivors by four distinct categories: (1) healthcare utilization, (2) patient-reported outcomes, (3) chronic health conditions, and (4) survival rates. For counting health outcomes of interest, if studies reported outcomes in more than one category, these studies were listed in different, separate outcome categories. Healthcare utilization outcomes included the concept of healthcare self-efficacy, initial and follow-up care visits, contact with general or cancer-specific healthcare providers, and use of hospital services. Patient-reported outcomes included the concept of health-related quality of life, symptom presence or

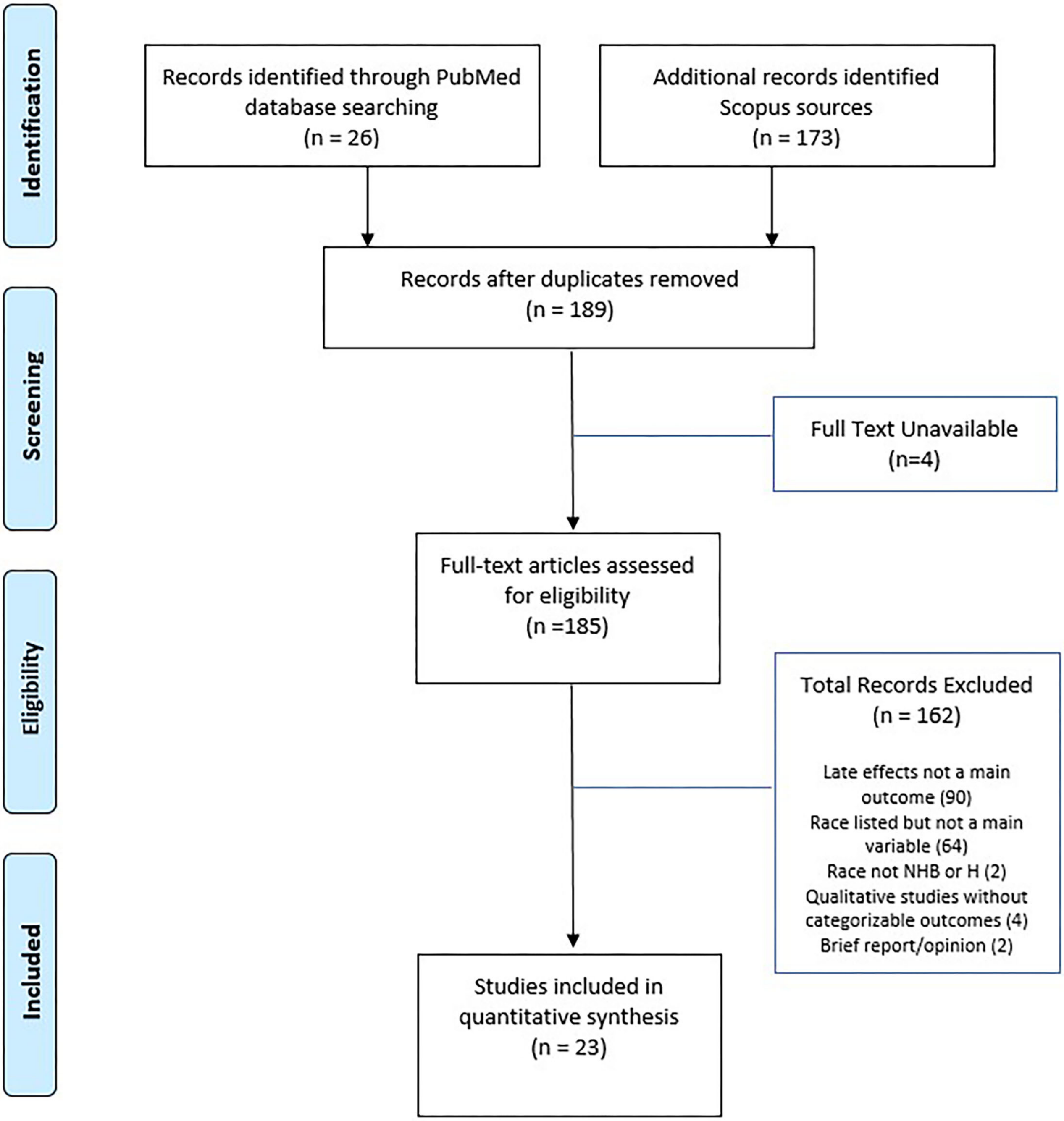

Identification
Screening
Eligibility
Included
Records identified through PubMed database searching (n = 26)
Additional records identified Scopus sources (n = 173)
Records after duplicates removed (n = 189)
Full Text Unavailable (n=4)
Full-text articles assessed for eligibility (n =185)
Total Records Excluded (n = 162)
Late effects not a main outcome (90)
Race listed but not a main variable (64)
Race not NHB or H (2)
Qualitative studies without categorizable outcomes (4)
Brief report/opinion (2)
Studies included in quantitative synthesis (n = 23)

FIGURE 1 | PRISMA diagram of study selection. Adapted from (21).

severity, adaptive functioning, and post-traumatic stress. Chronic health conditions represented individual health conditions (e.g., diabetes) or organ system-based condition groups (e.g., endocrine). Survival outcomes were categorized as all-cause or condition-specific survival rates. In addition, we reported the factors used to explain racial/ethnic disparities in health outcomes per the rationale of the articles or statistical modeling process (e.g., the mediating effects from the path analysis and

TABLE 1 | Characteristics of studies included in the scoping review.

References	Sample size	Age (years)	Population	Race/Ethnicity (%)	Health outcome category	Specific health outcomes
Armstrong et al. (22)	$N = 26443$	0–18	SEER	NHB (8.8); H (14.4)	Survival rates	All-cause mortality; non-recurrence/non-external mortality
Arpawong et al. (23)	$N = 94$	11–21	Treatment Center	Hispanic (English 27.6; ESL, 19.1)	Patient-reported outcomes	Post-traumatic growth, post-traumatic stress
Barrera et al. (24)	$N = 74$	8–16*	Canadian Children's	Black (5.4)	Patient-reported outcomes	Quality of life and emotional quality of life
Berkman et al. (25)	$N = 164316$	0–34*	SEER	Black (10.7)	Chronic health conditions and survival rates	Cardiovascular conditions, overall mortality
Castellino et al. (26)	$N = 8767$	≥18	CCSS	Hispanic (15.6)	Healthcare utilization, patient-reported outcomes	Screening, mental health
Choudhary et al. (27)	$N = 484$	2–36**	Sloan Kettering	Black (12.1); Hispanic (15.6)	Chronic health conditions	Vitamin-D deficiency i.e. chemiluminescent assay
Daly et al. (28)	$N = 866$	> 6	CHOA	48.5% Non-White	Healthcare utilization	Initial visit
Gance-Cleveland et al. (29)	$N = 321$	6–21*	Survivor Clinic	Black (2.4); Hispanic (29.9)	Chronic health conditions	Obesity
Kehm et al. (30)	$N = 31866$	0–19*	SEER	NHB (11.8); Hispanic (31.5)	Survival rates	Overall survival
Liu et al. (31)	$N = 13841$	8–58	CCSS	NHB (5); Hispanic (5.4)	Survival rates and chronic health conditions	Cardiovascular condition, overall mortality
Lu et al. (32)	$N = 10362$	18–38	CCSS	NHB (4.3); Hispanic (1.9)	Patient-reported outcomes	Pain
Meeske et al. (33)	$N = 86$	8–18*	CHLA	Hispanic (48%)	Patient-reported outcomes	Total and psychosocial function
Meeske et al. (34)	$N = 139$	14–25**+	CSP	Hispanic (US, 12.9; foreign born 43.8)	Patient-reported outcomes	Parent post-traumatic stress, depression
Milam et al. (35)	$N = 193$	14–25**	LA SEER	Hispanic (54.4)	Healthcare utilization	Follow-up care
Miller et al. (36)	$N = 193$	≥15	LA SEER	Hispanic (54.4)	Healthcare utilization	Healthcare self-efficacy
Miller et al. (37)	$N = 193$	≥15*	LA SEER	Hispanic (54.4)	Healthcare utilization	Information seeking
Oikonomou et al. (38)	$N = 88,418$	0–19*	SEER	Black (10.7); Other (8.1)	Chronic health conditions	Cardiovascular condition
Raghubar et al. (39)	$N = 114$	5–21*+	CHLA	Hispanic (29.8); Other (19.2)	Patient-reported outcomes	Adaptive functioning
Samaan et al. (40)	$N = 5956$	0–19**	SEER	Non-White (17.8)	Survival rates	Mortality to incidence ratio and relative survival trend
Santacroce et al. (41)	$N = 15$	37.9 +#	Clinic	Black (33.3)	Patient-reported outcomes	Uncertainty, anxiety, stress
Tobin et al. (42)	$N = 235$	14–25	CSP	Hispanic (56.2)	Patient-Reported outcome	Post-traumatic growth
Wasilewski-Masker et al. (43)	$N = 519$	12.1#	CHOA	Black (14)	Chronic health condition	Severity of symptoms (CTCEA)
Zebrack et al. (44)	$N = 6425$	≥18	CCSS	Non-White (6.4)	Patient-reported outcomes	Positive impact

*Less than 5 years since last treatment; **No information on years since last treatment; +Parent responses; #Mean age reported in years at survey/assessment; Reference Group: Non-Hispanic White.*

CCSS, Childhood Cancer Survivor Study; CHLA, Children's Hospital of Los Angeles; CHOA, Childhood Healthcare of Atlanta; CHOC, Children's Hospital of Orange County; CSP, Los Angeles Cancer Surveillance Program; SEER, Surveillance Epidemiology and End Results Program.

covariate-adjustment or interaction effects from the standard regression models).

RESULTS

Characteristics of the Selected Articles

Figure 1 presents the PRISMA flow diagram for the process of article selection. Of 185 full-text articles initially identified, 23 articles which met the inclusion/exclusion criteria were selected into full review and data extraction. **Table 1** summarizes the characteristics of the 23 selected articles, published between 2002 and 2020. All study populations included in the 23 articles were from North America. A majority (52%) of the selected articles were based on the US National Cancer Institute-funded Childhood Cancer Survivorship Study or the US Surveillance, Epidemiology, and End Results program (SEER). The size of the samples ranged from under 100 (5 articles) to over 10,000 (6 articles). The age range of survivors included in the 23 articles varied from adolescents (7 articles) and young adults (5 articles) to adults (5 articles) or all ages (6 articles). The distribution of race/ethnicity was calculated and presented as percentage with non-Hispanic White as the reference group (see **Table 1**). The average percentages were 6.9% for non-Hispanic Black and 4.5% for Hispanic, which were smaller compared to 13.4% for non-Hispanic Black and 18.5% for Hispanic in the general US population (45). Data abstracted from the selected studies were all cross-sectional in nature. A variety of statistical techniques were used to test the statistical difference and suggest the influential factors. The most used methods included multivariate modeling (13 articles) and ratio-based models [e.g., odds ratio (3 articles), proportional hazards (2 articles), or standardized ration (3 articles)]; other methods included analysis of covariance and flexible parameters model.

Disparities in Healthcare Utilization

Five articles reported racial/ethnic disparities in healthcare utilization (**Table 2**). The type of healthcare utilization disparities reported for non-Hispanic Black survivors included general medical contact (26) and an initial survivorship visit (28). The type of healthcare utilization disparities reported for Hispanic survivors included general medical contact (26), a cancer center visit (26), the use of follow-up care (35), health-care self-efficacy, defined as perceived control and confidence in managing healthcare (36), and seeking information from a hospital (37) or from family members (37). Across the articles, risk modulators included in analytic models for healthcare utilization disparities

TABLE 2 | Factors influencing disparities in healthcare utilization for childhood cancer survivors by race/ethnicity.

Source	Risk modulators	Results of unadjusted models	Results of adjusted models	Interpretation of findings
NON-HISPANIC BLACK				
(26)	SES (insurance, education, household income), age, diagnosis	**OR = 0.6; 95% CI: 0.5–0.9**	OR = 0.7; 95% CI: 0.5–1.0 (males)	Lower general medical contact attenuated by risk modulators.
(26)	SES (insurance, education, household income), age, diagnosis		**OR = 0.5; 95% CI: 0.3–0.7 (females)**	Lower general medical contact accounting for risk modulators.
(28)	Gender, age, treatment factors (year and age of diagnosis, diagnosis, therapy subsequent event), and logistic factors (insurance, distance from clinic)	**HR = 0.77, 95% CI: 0.64–0.94**	**HR= 0.64, 95% CI: 0.52–0.79**	Less likely to have initial survivorship visit.
HISPANIC				
(26)	SES (insurance, education, household income), age, diagnosis		**OR = O.6; 95% CI: 0.4–0.8**	Lower general medical contact accounting for risk modulators.

(26)	SES (insurance, education, household income), age, diagnosis		OR = 1.7, 95% CI: 1.2–2.3 (males)
			OR = 1.5, 95% CI: 1.1–2.0 (females)
				More likely to visit cancer center accounting for risk modulators.

(35)	Age, sex, social support, family influence, post traumatic growth, depressive symptoms, treatment, self-efficacy	OR = 0.55, 95% CI: 0.25–1.21, $p = 0.17$	**OR = 0.33, 95% CI: 0.11–0.96, $p = 0.03$**	Less likely to report previous use of follow-up care after accounting for risk modulators.
(36)	Age, sex, social support, family influence, post traumatic growth, depressive symptoms, treatment, self-efficacy	β = −0.38 (0.19), $p < 0.1$	**β = −0.42 (0.20), $p < 0.05$**	Lower health-care self-efficacy after accounting for risk modulators.
(37)	Age, sex, health insurance	**OR = 2.1, 95% CI: 1.17–3.79, $p < 0.05$**	**OR = 2.52, 95% CI: 1.19–5.30, $p < 0.05$**	Less likely to get information from hospital.
(37)	Age, sex, health insurance	**OR = 0.48, 95% CI: 0.24–0.98, $p < 0.05$**	OR = 0.50, 95% CI: 0.23–1.09, $p < 0.1$	Less likely to get information from family attenuated by risk modulators.

*Bold denotes statistical significance with $p < 0.05$; Reference group Non-Hispanic White or Caucasian; * Reference group listed as Non-Hispanic. SES, Socioeconomic Status; HR, Hazards Ratio; OR, Odd Ratio.*

TABLE 3 | Factors influencing disparities in patient-reported outcomes for childhood cancer survivors by race/ethnicity.

Source	Risk modulators	Results of unadjusted models	Results of adjusted models	Interpretation of findings
NON-HISPANIC BLACK				
(26)	SES (insurance, education, household income), age, diagnosis	OR = 0.8; 95% CI: 0.5–1.2	**OR = 0.6; 95% CI: 0.5–1.0 (females)**	Less adverse mental health after accounting for risk modulators.
(26)	SES (insurance, education, household income), age, diagnosis	**OR =1.7; 95% CI: 1.2–2.5**	OR = 1.2; 95% CI: 0.8–1.8 (females)	Higher functional impairment attenuated by risk modulators.
(32)	None	**HR = 1.91, 95% CI: 1.58–2.30, $p < 0.001$**		Higher reports of pain or abnormal sensation without accounting for risk modulators.
(32)	None	**HR =1.85, 95% CI: 1.54–2.22, $p < 0.001$**		Higher reports of migraines without accounting for risk modulators.
(32)	None	**HR = 1.68, 95% CI: 1.40–2.02, $p < 0.001$**		Higher reports of other frequent headaches without accounting for risk modulators.
(41)		**$p < 0.001$**		Higher parental uncertainty without accounting for risk modulators.
HISPANIC				
(32)	None	**HR =1.74, 95% CI: 1.27–2.39, $p = 0.001$**		Higher reports of pain or abnormal sensations without accounting for risk modulators.
(32)	None	**HR = 1.44, 95% CI: 1.06–1.96, $p = 0.02$**		Higher reports of other frequent headaches without accounting for risk modulators.
(23)	Demographics, disease/treatment factors, depressive symptoms, PTSS, optimism, QOL, SES	$p = 0.52$	**$p < 0.05$ (English primary language)**	Lower Post-traumatic Growth (PTG) accounting for risk modulators.
(33)	Diagnosis and fatigue	**$p = 0.02$**	**$p < 0.01$**	Lower psychosocial health after accounting for risk modulators.
(33)	Diagnosis and fatigue	**$p = 0.04$**	**$p < 0.01$**	Lower total reported quality of life after accounting for risk modulators.
(33)	Diagnosis and fatigue		**$p = 0.001$**	Lower school functioning accounting for risk modulators.
(33)	Diagnosis and fatigue		**$p = 0.01$**	Lower emotional functioning accounting for risk modulators.
(34)	Birthplace, education, income, stress, and treatment intensity	**$p < 0.0001$**	**$\beta = 14.20$ (3.95), $p = 0.0005$ (Foreign born)**	Higher parent post-traumatic stress.
(34)	Birthplace, education, income, stress, and treatment intensity	**$p = 0.002$**	**$\beta = 4.35$ (1.90), $p = 0.02$ (Foreign born); $\beta = 4.09$ (1.28), $p = 0.0002$ (US born)**	Higher rates of depression.
(39)	Family-level SES (parent education and family income)	**$p < 0.05$**	$p = 0.25$	Lower global adaptive functioning attenuated by risk modulators.
(39)	Family-level SES (parent education and family income)	**$p < 0.01$**	$p = 0.19$	Lower conceptual adaptive functioning attenuated by risk modulators.
(39)	Family-level SES (parent education and family income)	**$p < 0.01$**	$p = 0.48$	Lower social adaptive functioning attenuated by risk modulators.
(39)	Family-level SES (parent education and family income)	**$p < 0.05$**	$p = 0.15$	Lower practical adaptive functioning attenuated by risk modulators.
(42)	Age, sex, social support, family influence, PTG, depressive symptoms, treatment, self-efficacy		**OR= 0.25, 95% CI: 0.13-0.45**	Higher post-traumatic growth scores accounting for risk modulators.
NON-WHITE*				
(24)	Family income and caregiver education		**P=0.04**	Lower emotional quality of life accounting for risk modulators.

(Continued)

TABLE 3 | Continued

Source	Risk modulators	Results of unadjusted models	Results of adjusted models	Interpretation of findings
(33)	Diagnosis and fatigue	$p = 0.26$	$\mathbf{p = 0.04}$	Lower psychosocial functioning after accounting for risk modulators.
(33)	Diagnosis and fatigue	$p = 0.35$	$\mathbf{p = 0.04}$	Lower total reported quality of life after accounting for risk modulators.
(33)	Diagnosis and fatigue		$\mathbf{p = 0.01}$	Lower school functioning accounting for risk modulators.
(44)	Demographic and clinical variables		$\mathbf{p < 0.01}$	More positive impact of cancer in all five aspects of growth accounting for risk modulators.

*Bold denotes statistical significance with $p < 0.05$; Reference group Non-Hispanic White or Caucasian; * Listed as Other or both Non-Hispanic Black and Hispanic. SES, Socioeconomic Status; PTS, Post Traumatic Stress; QOL, Quality of Life; HR, Hazards Ratio; OR, Odd Ratio.*

included clinical factors (diagnosis, treatment), individual characteristics (age, sex, depressive symptoms, post-traumatic growth, self-efficacy), individual-/family-level SES (educational attainment, household income, health insurance coverage), and social/contextual factors (support, family influence).

In two articles, inclusion of individual characteristics and individual-/family-level SES in multivariable modeling attenuated the statistical significance for racial/ethnic disparities in healthcare utilization. Specifically, in an odds-ratio model, inclusion of individual-level and family-level SES and individual characteristics (cancer diagnosis and age at the time of study) removed statistical significance of the disparity in general medical contact among non-Hispanic Black male survivors (26). Similarly, inclusion of age, sex, and individual-level SES (i.e., health insurance) removed the statistical significance for receiving less cancer-related information from family members among Hispanic survivors (37). However, two articles found significant racial/ethnic disparity after adjusting for individual characteristics and social contextual factors in the multivariable analyses. Specifically, inclusion of cancer treatment, age, sex, social support, family influence, post-traumatic growth, depressive symptoms, and self-efficacy factors revealed significantly fewer previous receipt of follow-up care among Hispanic survivors compared to non-Hispanic White survivors (35). In addition, inclusion of cancer treatment, age, sex, social support, family influence, post-traumatic growth, and depressive symptoms factors revealed significantly lower healthcare self-efficacy for Hispanic survivors compared to non-Hispanic White survivors (36).

Disparities in Patient-Reported Outcomes

Nine articles have reported racial/ethnic disparities in patient-reported outcomes (**Table 3**). The types of patient-reported outcomes assessed for non-Hispanic Black survivors included quality-of-life (24), adverse mental health (26), functional impairment (26), pain or abnormal sensations (32), migraines (32), frequent headaches (32) and parental uncertainty about the child's health (41). The types of patient-reported outcomes assessed for Hispanic survivors included post-traumatic growth (23, 42), psychosocial health (33), quality-of-life (33), school functioning (33), emotional functioning (33), parental post-traumatic stress (34), depression (34), and conceptual, social and practical adaptive functioning (39), pain or abnormal sensations (32), and frequent headaches (32). In addition, the types of patient-report outcomes assessed for non-White survivors included psychosocial functioning (33), quality-of-life (33), school functioning (33), and positive impact of cancer (44). Across the articles, risk modulators included in multivariable modeling comprised clinical factors (diagnosis, treatment), individual characteristics (age, sex, depressive symptoms, post-traumatic stress, optimism, fatigue), individual-/family-level SES (educational attainment, health insurance, household income, parent educational attainment), and social/contextual factors (birthplace, language spoken at home).

In two articles, inclusion of individual characteristics and SES in multivariable modeling attenuated the significance for racial/ethnic disparities in patient-reported outcomes. Specifically, inclusion of cancer diagnosis, individual-level SES and age at study participation removed the significance for adverse mental health outcomes among non-Hispanic Black females (26). Similarly, inclusion of family-level SES removed the significance for poor global, conceptual, social, and practical adaptive functioning in Hispanic survivors (39). However, one article found that after adjusting for cancer diagnosis and fatigue symptoms, poorer psychosocial functioning and quality-of-life in minority (both non-Hispanic Black and Hispanic) survivors vs. non-Hispanic White survivors remained statistically significant (33).

Disparities in Chronic Health Conditions

Five articles reported racial/ethnic disparities in chronic health conditions (**Table 4**). The type of chronic health condition disparities assessed for non-Hispanic Black survivors included vitamin-D deficiency (27), subsequent neoplasms (31), cardiovascular disorders (31), cardiovascular risks (38), and serious/life-threatening health conditions (43). The type of chronic health condition disparities assessed for Hispanic survivors included vitamin-D deficiency (27), obesity (29), subsequent neoplasm (31), endocrine condition (31). Across the articles, risk modulators included in multivariable modeling included clinical factors (diagnosis, treatment), individual characteristics (age, sex, pubertal status), individual-level SES (educational attainment, income, health insurance), family-level SES (parent educational

TABLE 4 | Factors influencing disparities in chronic health conditions for childhood cancer survivors by race/ethnicity.

Source	Risk modulators	Results of unadjusted models	Results of adjusted models	Interpretation of findings
NON-HISPANIC BLACK				
(27)	Pubertal status	**OR = 3.11, 95% CI: 1.78–5.46**	**OR = 3.25, 95% CI: 1.83–5.78**	More likelihood of vitamin-D deficiency
(29)	Diagnosis and fatigue		OR = 2.06, 95% CI: 0.26 = 11.85, *p* = 0.436	Higher risk of obesity accounting for risk modulators.
(31)	Clinical/demographic variables		**RR = 0.6, 95% CI: 0.4–0.9, *p* = 0.009**	Higher rate of subsequent neoplasms accounting for risk modulators.
(31)	Clinical/demographic variables and treatment		**RR = 0.5, 95% CI: 0.3–0.8, *p* = 0.005**	Higher rate of subsequent neoplasms accounting for risk modulators.
(31)	Clinical/demographic variables, treatment, and SES (education, income, & insurance)		**RR = 0.6, 95% CI: 0.4–0.9, *p* = 0.01**	Higher rate of subsequent neoplasms accounting for risk modulator.
(31)	Clinical/demographic variables, treatment, SES (education, income, & insurance), and CVRF (obesity, diabetes, hypertension, and dyslipidemia)		**RR = 0.6, 95% CI: 0.4–0.9, *p* = 0.02**	Higher rate of subsequent neoplasms accounting for risk modulators.
(31)	Clinical/demographic variables		**RR = 1.9, 95% CI: 1.2–2.9, *p* = 0.005**	Higher grade cardiovascular conditions accounting for risk modulators.
(31)	Clinical/demographic variables and treatment		**RR = 1.8, 95% CI: 1.1–2.7, *p* = 0.01**	Higher grade cardiovascular conditions accounting for risk modulators.
(31)	Clinical/demographic variables, treatment, and SES (education, income, & insurance)		**RR = 1.6, 95% CI: 1.0–2.3, *p* = 0.04**	Higher grade cardiovascular conditions accounting for risk modulators.
(38)	More or less that years of diagnosis	HR = 0.98, 95% CI: 0.52–1.86, *p* = 0.95	**HR = 1.60, 95% CI: 1.05–2.43, *p* = 0.03**	Higher cardiovascular after five years accounting for risk modulators.
(43)	Treatment, diagnosis, age and gender	RR = 0.9, 95% CI: 0.7–1.2, *p* = 0.32	**RR = 1.5, 95% CI: 1.0–2.1, *p* = 0.03**	Higher severity (Grade 3-4) in health conditions after accounting for risk modulators.
HISPANIC				
(27)	Pubertal status	**OR = 2.08, 95% CI: 1.09–3.97**	**OR = 2.14, 95% CI: 1.11–4.13**	More likelihood of vitamin-D deficiency.
(29)	Diagnosis and fatigue		**OR= 2.29, 95% CI: 1.23–4.30**	Higher risk of obesity accounting for risk modulators.
(31)	Clinical/demographic variables, treatment, SES (education, income, & insurance), and CVRF (obesity, diabetes, hypertension, and dyslipidemia)		**RR= 1.6, 95% CI: 1.2–2.3, *p* = 0.005**	Higher rate of subsequent neoplasms accounting for risk modulators.
(31)	Clinical/demographic variables, treatment, and SES (education, income, & insurance)		**RR = 1.5, 95% CI: 1.1–1.2, *p* = 0.01**	Increased risk for endocrine conditions accounting for risk modulators.
(31)	Clinical/demographic variables, treatment, SES (education, income, & insurance), and CVRF (obesity, diabetes, hypertension, and dyslipidemia)		**RR = 1.6, 95% CI: 1.2–2.3, *p* = 0.005**	Increased risk for endocrine conditions accounting for risk modulators.

Bold denotes statistical significance with $p < 0.05$; Reference group Non-Hispanic White or Caucasian. CVRF, Cardiovascular Risk Factor; SES, Socioeconomic Status; HR, Hazards Ratio; OR, Odd Ratio; RR, Relative Ratio.

attainment, household income), and lifestyle health risk for chronic health conditions (BMI and cardiovascular risk factors including obesity, diabetes, hypertension and dyslipidemia).

In two articles, inclusion of clinical factors, individual characteristics, and SES in the modeling attenuated the significance for racial/ethnic disparities in chronic health condition. Specifically, one article found that inclusion of clinical factors, individual characteristics, and SES factors removed the significance of disparity in subsequent neoplasms for non-Hispanic Black survivors (31). Another article suggested that inclusion of clinical and demographic factors removed the significance of the disparity in serious/life-threatening health conditions for non-Hispanic Black survivors (43). However, one article found that disparities in subsequent neoplasms and cardiovascular disorders remained significant for non-Hispanic Black survivors and disparities in subsequent neoplasms and endocrine disorders remained significant for Hispanic survivors after adjusting for clinical, cardiovascular risk, and/or individual SES factors in the modeling (31).

Disparities in Survival Rates

Five articles reported racial/ethnic disparities in survival rates (**Table 5**). Type of survival outcomes assessed for non-Hispanic Black survivors included all-cause mortality (22), all-cause mortality including relative and standardized rates (31), subsequent malignancy mortality (22), risk of cardiovascular-specific death (25), and risk of any death (25). Types of survival metrics assessed for Hispanic survivors included all-cause standardized mortality rates (31). Type of survival metrics assessed for non-White survivors included hazard of death for survivors diagnosed with acute myeloid leukemia, astrocytoma, and non-astrocytoma CNS tumors (30) and mortality to incidence ratios (40). Across the articles, modulators included in multivariable modeling included clinical factors (time since cancer diagnosis, age at cancer diagnosis, cancer type), individual characteristics (age, sex), SES (educational attainment, income, health insurance), lifestyle health risk (cardiovascular risk factors such as obesity, diabetes, hypertension, dyslipidemia), neighborhood factors (census-track SES Index), and US national mortality rate (for the purpose of mortality standardization).

In two articles, inclusion of clinical and SES factors attenuated the significance for racial/ethnic disparities in survival outcomes. Specifically, one article found that inclusion of census-tract (i.e., neighborhood-level) SES removed the significance of death hazard for non-Hispanic Black survivors diagnosed with astrocytoma and non-astrocytoma CNS tumor, but not acute myeloid leukemia (30). Another article suggests that the adjustment individual and clinical factors removed the significance of cardiovascular-specific death for non-Hispanic Black survivors (25). However, another article found that disparities in all-cause relative mortality rates remained statistically significant for non-Hispanic Black and Hispanic survivors after adjusting for clinical factors, individual demographic and SES factors, and cardiovascular risk in the modeling (31). In addition, based on a path analysis focusing neighborhood socioeconomic determinants as the mediator, one article found significantly higher death hazard among non-White survivors of acute myeloid leukemia compared to non-Hispanic White survivors (30).

DISCUSSION

Compared to non-Hispanic White, non-Hispanic Black and Hispanic childhood cancer survivors suffer more from poorer health outcomes including healthcare utilization, patient-reported outcomes, chronic health conditions and survival rate. While there is an effect of race/ethnicity on health outcomes for childhood cancer survivors; there is not yet enough evidence to determine the true effect of SES across all outcomes given the cross-sectional design of previous studies. The current findings do show that embedded in race and ethnicity are a multitude of factors at the clinical (e.g., disease, treatment), individual (e.g., demographic, SES), and neighborhood (e.g., community SES) levels that may explain some of the disparities and poor health outcomes. However, the magnitude of racial/ethnic disparities changed in some but not all studies after adjusting for these risk modulators. As such, we see a complex interplay among these risk factors for health disparities. Future research is warranted to elucidate the complex associations between racial/ethnic and SES factors and health outcomes for childhood cancer survivors.

Disparity-Specific Risk Modulators

Potential risk modulators that explained the associations between race/ethnicity and health outcomes attempted in all articles were reviewed. Risk modulators commonly reported for healthcare utilization disparity included individual-level SES (26, 28). Social support and religious importance (35–37) also explained aspects of the racial/ethnic disparities. In patient-reported outcomes, risk modulators for racial/ethnic disparities included family-level SES (23, 24, 26, 39), family dynamics (34, 42), and treatment factors (33). Particularly for Hispanic survivors, family dynamics (e.g., language spoken at home) should be further investigated as they are potentially protective factors for poor patient-reported outcomes. In chronic health conditions, most articles found that racial/ethnic disparities remained statistically significant after risk modulators (e.g., clinical factors, individual demographic and SES factors, and cardiovascular risk factors) were included in the multivariable modeling (27, 29, 31, 38, 43). It is possible that underlying biological mechanisms (e.g., inherited genetic predisposition to disease risks and epigenetic modifications due to life experiences or environmental exposures) and disadvantaged neighborhood environments may elevate disparities in chronic health conditions beyond the influence of individual SES and clinical risk. In survival rate, individual-level and neighborhood-level SES (22, 30, 31), together with age at cancer diagnosis (25, 39) and years since diagnosis, played an important role for risk facilitation (26).

Bhatia et al. argues that the "burden of morbidity and mortality [between races] is comparable because mortality is mitigated by SES" (12, 46); however, our findings suggest that the effect of SES on disparities is less straightforward. In fact, individual-level, family-level, and neighborhood-level variables may have distinct impacts on health disparities. We found that adjustment for SES increased the magnitude of disparities in some patient-reported outcomes (i.e. adverse mental health for Black females, post-traumatic stress for Hispanic parents) rather than mitigated them (23, 26). Furthermore, various sources and datasets used to quantify SES risk modulators in the analysis across studies may complicate the interpretations of findings. While the majority of the articles in our review used individual-level SES or family-level SES), one article used

TABLE 5 | Factors influencing disparities in survival for childhood cancer survivors by race/ethnicity.

Source	Risk modulators	Results of unadjusted models	Results of adjusted models	Interpretation of findings
NON-HISPANIC BLACK				
(22)	US mortality rates, sex, year of diagnosis		**SMR = 6.67, 95% CI: 5.84–7.59**	Higher all-cause mortality risk accounting for risk modulators.
(22)	US mortality rates, sex, year of diagnosis		**SMR = 10.72, 95% CI: 7.18–15.40**	Higher mortality risk of subsequent malignancy accounting for risk modulators.
(25)	Age at diagnosis, time since diagnosis, cancer type	**HR = 1.75, 95% CI: 1.70–1.79**	**Age 0–14 HR = 1.26, 95% CI: 1.18–1.35 Age 15–35 HR= 1.88, 95% CI: 1.83–1.93**	Higher risk of any death.
(25)	Age at diagnosis, time since diagnosis, cancer type	**HR = 2.13, 95% CI: 1.85–2.46**	Age 0–14 HR = 1.08, 95% CI: 0.62–1.89 Age 15–34 HR = 1.33, 95% CI: 0.60–2.95	Higher cardiovascular disease death attenuated by risk modulators.
(31)	Clinical/demographic variables		**RR = 1.5, 95% CI: 1.1–2.0,** ***p*** **= 0.004**	Higher all-cause relative mortality rate accounting for risk modulators.
(31)	Clinical/demographic variables and treatment		**RR = 1.4, 95% CI: 1.1–1.9,** ***p*** **= 0.008**	Higher all cause relative mortality rate accounting for risk modulators.
(31)	Clinical/demographic variables, treatment, and SES (education, income, & insurance)		RR = 1.0, 95% CI: 0.8–1.4, $p = 0.88$	Higher all-cause relative mortality rate accounting for risk modulators.
(31)	Clinical/demographic variables, treatment, and SES (education, income, & insurance)		**SMR = 0.6, 95% CI: 0.4–0.8,** ***p*** **< 0.001**	Higher all-cause standardized mortality rate accounting for risk modulators.
(31)	Clinical/demographic variables, treatment, and SES (education, income, & insurance) and SVRF (obesity, diabetes, hypertension, and dyslipidemia)		**SMR = 0.6, 95% CI: 0.4–0.8,** ***p*** **< 0.001**	Higher all-cause standardized mortality rate accounting for risk modulators.
HISPANIC				
(31)	Clinical/demographic variables, treatment, and SES (education, income, & insurance) and SVRF (obesity, diabetes, hypertension, and dyslipidemia)		**SMR = 0.7, 95% CI: 0.6–1.0,** ***p*** **= 0.03**	Higher all-cause standardized mortality rate accounting for risk modulators.
NON-WHITE*				
(30)	Neighborhood-level SES index**	**Direct HR = 1.45, 95% CI: 1.15–1.84,** $p <$ **0.01**	**Indirect HR = 1.15, 95%CI: 1.03–1.29,** ***p*** **= 0.01**	Higher hazard of death for Acute Myeloid Leukemia survivors.
(30)	Neighborhood-level SES index**	**Direct HR = 1.80, 95% CI: 1.42–2.30,** $p <$ **0.0001**	Indirect HR = 1.08, 95% CI: 0.98–1.20, $p = 0.12$	Higher hazard of death for Astrocytoma survivors attenuated by risk modulators.
(30)	Neighborhood-level SES index**	**Direct HR = 1.41, 95% CI: 1.11–1.78,** ***p*** **< 0.01**	Indirect HR = 1.09, 95% CI; 0.97**–**1.22, $p = 0.14$	Higher hazard of death for non-astrocytoma CNS tumors attenuated by risk modulators.
(40)	None	**MIR = 27.4%,** ***p*** **= 0.001**		Higher mortality to incidence without accounting for risk modulators.

*Bold denotes statistical significance with p < 0.05; Reference group Non-Hispanic White or Caucasian; * Listed as other or Non-White; ** Tract SES Index, National Cancer Institute Census Tract-level socioeconomic status (SES) Index. CVRF, Cardiovascular Risk Factor; SES, Socioeconomic Status; HR, Hazards Ratio; OR, Odd Ratio. SMR, Standard Mortality Ration; RR, Relative Ratio; MIR, Mortality to Incidence Ratio.*

a validated composite SES index with seven specific indicators (proportion employed in working class occupations, proportion over 16 employed, education index, median household income, proportion below 200% poverty level, median rent and median house value) (30, 47) to capture the complex influences of different levels of SES. However, very few selected articles included neighborhood contextual factors in the analysis. In fact, neighborhood-level factors such as the built environment (e.g., green space) (48), accessibility to healthy food (49), and healthcare services (50) are increasingly considered key determinants of health outcomes for adult-onset cancer but not for pediatric cancer research. In addition, race/ethnicity-sensitive indices warrant consideration including crime-rate, incarceration, and residential segregation. The use of geospatial neighborhood metrics may provide useful information for understanding disparities in health outcomes thereby offering a more complete depiction of health disparities for childhood cancer survivors.

In addition to improving SES measurement for childhood cancer research, it is important to use a holistic and life-course approach to investigating risk of health disparities. Williams (4, 51) suggests that race is an antecedent for SES instead of a variable inside, and embedded in race and ethnicity are layered factors that may be inextricably linked. Geronimus et al., suggest the burden of physiological stress (i.e. allostatic load) of race, ethnicity, and low SES can accumulate over time (52), which in turn may link to health disparities in underserved minority breast cancer (53, 54) and general (54) populations. Cultural and familial factors can influence the impact of allostatic load (54, 55) which may explain the risk for poor patient-reported outcomes in minority survivors. Krieger (56) suggested a federal mandate to include and categorize individual-level data pertinent to racialized societal inequities and explicit justification of metrics used to categorize racial groups. Therefore, in addition to standard SES variables, the design and collection of standardized race/ethnic-specific risk modulators for childhood cancer survivors are needed.

Racial/Ethnic Disparity-Specific Interventions

Risk modulators that substantially impact health outcomes of individual childhood cancer survivors were SES, healthcare accessibility, and health insurance. Several studies suggested that neighborhood-level SES (30), individual-level SES (26, 31) and/or family-level SES (26, 39) plays a more significant role as compared to health insurance in explaining the effects of race/ethnicity on poor health outcomes in childhood cancer survivors. In fact, a population-based study found that improving health insurance coverage alone may disproportionately benefit non-Hispanic White with lower SES rather than racial/ethnic minorities (57, 58). A more inclusive, need-based financial assistance program for individual survivors should be considered for minority survivors to reduce the risk of health disparities. In addition, the first two to three years from cancer diagnosis (25, 38, 43) and primary caregiver education background and proximity/access to care (39) were associated with elevated risk of health disparities in minority childhood cancer survivors. Therefore, healthcare systems should assess the disparity status for minority childhood cancer survivors immediately following completion of therapy and provide social support or resources to address these issues (e.g., coordinating transportation aids for minority families) toward improving follow-up care and reducing disease burden.

Our findings highlight that individual-level factors, such as culture (28, 33, 34, 37) and sex (26, 28) may contribute to racial/ethnic disparities in health outcomes. Cultural beliefs (i.e. fatalism) and gender beliefs seem relevant to health disparities in non-Hispanic Black survivors (26), while family dynamics, such as foreign-born parents experiencing greater amounts of post-traumatic stress, may impact Hispanic survivors (34, 59). In addition, minority childhood cancer survivors who had better social skills (27) and post-traumatic growth (31) were associated with better health outcomes. Therefore, it is critical to provide culture-/race-/ethnicity-/gender-specific social and emotional learning (i.e. stress prevention) interventions and diversity informed training for healthcare navigators (i.e. social workers, hospital staff, researchers, etc.). Social and emotional learning interventions that acknowledge established race-/gender-related stigma are avenues to augment resilience and provide social support and belonging.

Racial/Ethnic Disparity in Era of Digital Health and Big Data

There is an opportunity to leverage health information technology to promote health equity for minority and underserved populations (60, 61). Emerging evidence has found that the use of eHealth and mHealth platforms can improve physiological and psychological well-being, health knowledge, and self-management skill in racial/ethnic minorities and underserved populations (62). Given the importance of visiting oncologists/primary physicians for follow-up care and maintaining healthy lifestyle among childhood cancer survivors, mHealth and eHealth technology represent the methods that may improve access to medical care (e.g., telemedicine consultation and remote lifestyle and psychological interventions), communication with healthcare providers (e.g., digital therapy and education, tailored supportive resources), and symptom monitoring and management (e.g., real-time symptom monitoring for identifying early signs of late effects) (62). However, the vast majority of current eHealth and mHealth applications are designed in the English language. Future efforts are warranted to ensure the provision of technology platforms that are multilingual and culturally and literately appropriate.

Improving medical informatics infrastructure within healthcare systems can facilitate the collection and assessment of social determinants of health data for cancer survivors on a regular-basis and integrate social determinant information into clinical decision-making process. Incorporating neighborhood-/community-level social determinant data into electronic health records (EHRs) will allow clinicians to provide tailored interventions that are clinically actionable based on the survivors' need and contextual influence. Given the big data stored in EHRs, the use of artificial intelligence analytics (e.g., machine learning and natural processing techniques) can help identify complex social determinants for individual minority survivors. Recent evidence suggests that implementation of machine learning approaches helps identify the patterns of social determinants for impaired health outcomes with superior performance compared to the use of traditional analytics (63).

Limitations

While this scoping review provides useful information for racial/ethnic disparities in health outcomes among childhood cancer survivors, the findings should be carefully interpreted. First, race/ethnicity data from all articles were self-reported. Self-reported race/ethnicity information is often arbitrary and poorly defined (16). Furthermore, based on the available data included in the articles, we only focused on two traditionally minoritized/marginalized groups and excluded other non-White minority groups from our review. For example, American Indian childhood cancer survivors with acute lymphoblastic leukemia have lower survival rates compared to other races/ethnicities (46). Second, characteristics and patients of the survivor populations included in our review were generally homogenous. As the majority of selected studies were derived from the US-based Childhood Cancer Survivorship Study or the Surveillance Epidemiology and End Results registry, health outcome data are likely to overlap in time, collection, and patients. As mentioned in the beginning of the Results section, the percentage of minority survivors in the selected study was far lower than the percentage of the US general population. Finally, this scoping review focused on non-Hispanic Black and Hispanic health disparities, which is an emerging topic supported by current research on minoritized populations (20). In fact, the studies selected into our review did not breakdown race/ethnicity into categories beyond the three minoritized categories reported. Some articles just reported White and Other. It is critical to evaluate health disparities across more categories and intersections of races and ethnicities in the future research. It is also important to use a community-based, culture-specific participatory research design to recruit and engage racial/ethnic minorities to in childhood cancer survivorship research for better understanding the gap while also elucidating clinical interventions (64).

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

JB and I-CH: concept and design. MH and LR: administrative support. TR and TM: collection and assembly of data. TR, TM, and I-CH: data analysis and interpretation. TR and I-CH: manuscript writing. All authors editing and final approval of manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

The research reported in this manuscript was supported by the National Cancer Institute under award number R01CA238368 (MPIs: I-CH, JB) and V-Foundation Robin Roberts Cancer Thrivership Fund under award number DT2020-014 (MPIs: ZW, I-CH). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

REFERENCES

1. US Census Bureau. Methodology, Assumptions, and Inputs for the 2017 National Population Projections. Washington (2018). Available online at: https://www2.census.gov/programs-surveys/popproj/technical-documentation/methodology/methodstatement17.pdf (accessed September 30, 2021).
2. US Census Bureau. American Community Survey demographic and housing estimates. Washington (2020). Available online at: https://www.census.gov/programs-surveys/acs/data.html (accessed September 30, 2021).
3. Ford CL, Airhihenbuwa CO. The public health critical race methodology: praxis for antiracism research. *Soc Sci Med.* (2010) 71:1390–8. doi: 10.1016/j.socscimed.2010.07.030
4. Williams DR. Race, socioeconomic status, and health. The added effects of racism and discrimination. *Ann N Y Acad Sci.* (1999) 896:173–88. doi: 10.1111/j.1749-6632.1999.tb08114.x
5. American Association for Cancer Research. Cancer Progress Report. Philidelphia (2020). Available online at: https://cancerprogressreport.aacr.org/progress/ (accessed September 30, 2021).
6. Heron M. Deaths: leading causes for 2017. *Natl Vital Stat Rep.* (2019) 68:1–77. Available online at: https://www.cdc.gov/nchs/data/nvsr/nvsr68/nvsr68_06-508.pdf
7. Viale PH. The American cancer society's facts & figures: 2020 edition. *J Adv Pract Oncol.* (2020) 11:135–6. doi: 10.6004/jadpro.2020.11.2.1
8. Huang IC, Ehrhardt MJ LI C, Mulrooney DA, Chamaitilly W, Srivastava D, et al. Longitudinal assessment of patient-reported cumulative symptom burden as an indicator of chronic health conditions in adult survivors of childhood cancer: a joint report of the St. Jude Lifetime Cohort (SJLIFE) and the Childhood Cancer Survivor Study (CCSS). *J Clin Oncol.* (2018) 36:10571. doi: 10.1200/JCO.2018.36.15_suppl.10571
9. Dixon SB Li N, Yasui Y, Bhatia S, Casillas JN, Gibson TM, et al. Racial and ethnic disparities in neurocognitive, emotional, and quality-of-life outcomes in survivors of childhood cancer: a report from the Childhood Cancer Survivor Study. *Cancer.* (2019) 125:3666–77. doi: 10.1002/cncr.32370
10. Robison LL, Hudson MM. Survivors of childhood and adolescent cancer: life-long risks and responsibilities. *Nat Rev Cancer.* (2014) 14:61–70. doi: 10.1038/nrc3634
11. Nolan VG, Krull KR, Gurney JG, Leisenring W, Robison LL, Ness KK. Predictors of future health-related quality of life in survivors of adolescent cancer. *Pediatr Blood Cancer.* (2014) 61:1891–4. doi: 10.1002/pbc.25037
12. Bhatia S, Gibson TM, Ness KK, Liu Q, Oeffinger KC, Krull KR, et al. Childhood cancer survivorship research in minority populations: a position paper from the Childhood Cancer Survivor Study. *Cancer.* (2016) 122:2426–39. doi: 10.1002/cncr.30072
13. Egede LE. Race, ethnicity, culture, and disparities in health care. *J Gen Intern Med.* (2006) 21:667–9. doi: 10.1111/j.1525-1497.2006.0512.x
14. Caplin DA, Smith KR, Ness KK, Hanson HA, Smith SM, Nathan PC, et al. Effect of population socioeconomic and health system factors on medical care of childhood cancer survivors: a report from the childhood cancer survivor study. *J Adolesc Young Adult Oncol.* (2017) 6:74–82. doi: 10.1089/jayao.2016.0016
15. Williams DR, Sternthal M. Understanding racial-ethnic disparities in health: sociological contributions. *J Health Soc Behav.* (2010) 51:S15–27. doi: 10.1177/0022146510383838
16. Gomez SL, Shariff-Marco S, DeRouen M, Keegan TH, Yen IH, Mujahid M, et al. The impact of neighborhood social and built environment factors across the cancer continuum: current research, methodological considerations, and future directions. *Cancer.* (2015) 121:2314–30. doi: 10.1002/cncr.29345
17. 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

18. Sucharew H, Macaluso M. Progress notes: methods for research evidence synthesis: the scoping review approach. *J Hosp Med.* (2019) 14:416–8. doi: 10.12788/jhm.3248
19. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ.* (2021) 372:n71. doi: 10.1136/bmj.n71
20. Brown LA, Strega S. *Research as Resistance, 2e: Revisiting Critical, Indigenous, and Anti-Oppressive Approaches.* Canadian Scholars' Press (2015).
21. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* (2009) 6:e1000097.
22. Armstrong GT, Pan Z, Ness KK, Srivastava D, Robison LL. Temporal trends in cause-specific late mortality among 5-year survivors of childhood cancer. *J Clin Oncol.* (2010) 28:1224–31. doi: 10.1200/JCO.2009.24.4608
23. Arpawong TE, Oland A, Milam JE, Ruccione K, Meeske KA. Post-traumatic growth among an ethnically diverse sample of adolescent and young adult cancer survivors. *Psychooncology.* (2013) 22:2235–44. doi: 10.1002/pon.3286
24. Barrera M, Atenafu EG, Schulte F, Bartels U, Sung L, Janzen L, et al. Determinants of quality of life outcomes for survivors of pediatric brain tumors. *Pediatr Blood Cancer.* (2017) 64:e26481. doi: 10.1002/pbc.26481
25. Berkman AM, Brewster AM, Jones LW Yu J, Lee JJ, Peng SA, et al. Racial differences in 20-year cardiovascular mortality risk among childhood and young adult cancer survivors. *J Adolesc Young Adult Oncol.* (2017) 6:414–21. doi: 10.1089/jayao.2017.0024
26. Castellino SM, Casillas J, Hudson MM, Mertens AC, Whitton J, Brooks SL, et al. Minority adult survivors of childhood cancer: a comparison of long-term outcomes, health care utilization, and health-related behaviors from the childhood cancer survivor study. *J Clin Oncol.* (2005) 23:6499–507. doi: 10.1200/JCO.2005.11.098
27. Choudhary A, Chou J, Heller G, Sklar C. Prevalence of vitamin D insufficiency in survivors of childhood cancer. *Pediatr Blood Cancer.* (2013) 60:1237–9. doi: 10.1002/pbc.24403
28. Daly A, Lewis RW, Vangile K, Masker KW, Effinger KE, Meacham LR, et al. Survivor clinic attendance among pediatric- and adolescent-aged survivors of childhood cancer. *J Cancer Surviv.* (2019) 13:56–65. doi: 10.1007/s11764-018-0727-3
29. Gance-Cleveland B, Linton A, Arbet J, Stiller D, Sylvain G. Predictors of overweight and obesity in childhood cancer survivors. *J Pediatr Oncol Nurs.* (2020) 37:154–62. doi: 10.1177/1043454219897102
30. Kehm RD, Spector LG, Poynter JN, Vock DM, Altekruse SF, Osypuk TL. Does socioeconomic status account for racial and ethnic disparities in childhood cancer survival? *Cancer.* (2018) 124:4090–7. doi: 10.1002/cncr.31560
31. Liu Q, Leisenring WM, Ness KK, Robison LL, Armstrong GT, Yasui Y, et al. Racial/ethnic differences in adverse outcomes among childhood cancer survivors: the childhood cancer survivor study. *J Clin Oncol.* (2016) 34:1634–43. doi: 10.1200/JCO.2015.66.3567
32. Lu Q, Krull KR, Leisenring W, Owen JE, Kawashima T, Tsao JCI, et al. Pain in long-term adult survivors of childhood cancers and their siblings: a report from the Childhood Cancer Survivor Study. *Pain.* (2011) 152:2616–24. doi: 10.1016/j.pain.2011.08.006
33. Meeske KA, Patel SK, Palmer SN, Nelson MB, Parow AM. Factors associated with health-related quality of life in pediatric cancer survivors. *Pediatr Blood Cancer.* (2007) 49:298–305. doi: 10.1002/pbc.20923
34. Meeske KA, Sherman-Bien S, Hamilton AS, Olson AR, Slaughter R, Kuperberg A, et al. Mental health disparities between Hispanic and non-Hispanic parents of childhood cancer survivors. *Pediatr Blood Cancer.* (2013) 60:1470–7. doi: 10.1002/pbc.24527
35. Milam JE, Meeske K, Slaughter RI, Sherman-Bien S, Ritt-Olson A, Kuperberg A, et al. Cancer-related follow-up care among Hispanic and non-Hispanic childhood cancer survivors: the Project Forward study. *Cancer.* (2015) 121:605–13. doi: 10.1002/cncr.29105
36. Miller KA, Wojcik KY, Ramirez CN, Ritt-Olson A, Freyer DR, Hamilton AS, et al. Supporting long-term follow-up of young adult survivors of childhood cancer: correlates of healthcare self-efficacy. *Pediatr Blood Cancer.* (2017) 64:358–63. doi: 10.1002/pbc.26209
37. Miller KA, Ramirez CN, Wojcik KY, Ritt-Olson A, Baezconde-Garbanati L, Thomas SM, et al. Prevalence and correlates of health information-seeking among Hispanic and non-Hispanic childhood cancer survivors. *Support Care Cancer.* (2018) 26:1305–13. doi: 10.1007/s00520-017-3956-5
38. Oikonomou EK, Athanasopoulou SG, Kampaktsis PN, Kokkinidis DG, Papanastasiou CA, Feher A, et al. Development and validation of a clinical score for cardiovascular risk stratification of long-term childhood cancer survivors. *Oncologist.* (2018) 23:965–73. doi: 10.1634/theoncologist.2017-0502
39. Raghubar KP, Orobio J, Ris MD, Heitzer AM, Roth A, Brown AL, et al. Adaptive functioning in pediatric brain tumor survivors: An examination of ethnicity and socioeconomic status. *Pediatr Blood Cancer.* (2019) 66:e27800. doi: 10.1002/pbc.27800
40. Samaan MC, Akhtar-Danesh N. The impact of age and race on longevity in pediatric astrocytic tumors: a population-based study. *Pediatr Blood Cancer.* (2015) 62:1567–71. doi: 10.1002/pbc.25522
41. Santacroce S. Uncertainty, anxiety, and symptoms of posttraumatic stress in parents of children recently diagnosed with cancer. *J Pediatr Oncol Nurs.* (2002) 19:104–11. doi: 10.1177/104345420201900305
42. Tobin J, Allem JP, Slaughter R, Unger JB, Hamilton AS, Milam JE. Posttraumatic growth among childhood cancer survivors: associations with ethnicity, acculturation, and religious service attendance. *J Psychosoc Oncol.* (2018) 36:175–88. doi: 10.1080/07347332.2017.1365799
43. Wasilewski-Masker K, Mertens AC, Patterson B, Meacham LR. Severity of health conditions identified in a pediatric cancer survivor program. *Pediatr Blood Cancer.* (2010) 54:976–82. doi: 10.1002/pbc.22431
44. Zebrack BJ, Stuber ML, Meeske KA, Phipps S, Krull KR, Liu Q, et al. Perceived positive impact of cancer among long-term survivors of childhood cancer: a report from the childhood cancer survivor study. *Psychooncology.* (2012) 21:630–9. doi: 10.1002/pon.1959
45. US Census Bureau. Quick Facts: Race and Hispanic Origin. Washington (2020). Available online at: https://www.census.gov/quickfacts/fact/table/US/PST045219 (accessed September 30, 2021).
46. Bhatia S. Disparities in cancer outcomes: lessons learned from children with cancer. *Pediatr Blood Cancer.* (2011) 56:994–1002. doi: 10.1002/pbc.23078
47. Yu M, Tatalovich Z, Gibson JT, Cronin KA. Using a composite index of socioeconomic status to investigate health disparities while protecting the confidentiality of cancer registry data. *Cancer Causes Control.* (2014) 25:81–92. doi: 10.1007/s10552-013-0310-1
48. Kish JK Yu M, Percy-Laurry A, Altekruse SF. Racial and ethnic disparities in cancer survival by neighborhood socioeconomic status in Surveillance, Epidemiology, and End Results (SEER) Registries. *J Natl Cancer Inst Monogr.* (2014) 2014:236–43. doi: 10.1093/jncimonographs/lgu020
49. Richardson LD, Norris M. Access to health and health care: how race and ethnicity matter. *Mt Sinai J Med.* (2010) 77:166–77. doi: 10.1002/msj.20174
50. Syed ST, Gerber BS, Sharp LK. Traveling towards disease: transportation barriers to health care access. *J Community Health.* (2013) 38:976–93. doi: 10.1007/s10900-013-9681-1
51. Williams DR. Race/ethnicity and socioeconomic status: measurement and methodological issues. *Int J Health Serv.* (1996) 26:483–505. doi: 10.2190/U9QT-7B7Y-HQ15-JT14
52. Geronimus AT, Hicken M, Keene D, Bound J. "Weathering" and age patterns of allostatic load scores among blacks and whites in the United States. *Am J Public Health.* (2006) 96:826–33. doi: 10.2105/AJPH.2004.060749
53. Linnenbringer E, Gehlert S, Geronimus AT. Black-white disparities in breast cancer subtype: the intersection of socially patterned stress and genetic expression. *AIMS Public Health.* (2017) 4:526–56. doi: 10.3934/publichealth.2017.5.526
54. Peek MK, Cutchin MP, Salinas JJ, Sheffield KM, Eschbach K, Stowe RP, et al. Allostatic load among non-Hispanic Whites, non-Hispanic Blacks, and people of Mexican origin: effects of ethnicity, nativity, and acculturation. *Am J Public Health.* (2010) 100:940–6. doi: 10.2105/AJPH.2007.129312
55. Maguire-Jack K, Lanier P, Lombardi B. Investigating racial differences in clusters of adverse childhood experiences. *Amn J Orthopsychiatry.* (2020) 90:106–14. doi: 10.1037/ort0000405
56. Krieger N. Structural racism, health inequities, and the two-edged sword of data: structural problems require structural solutions. *Front Public Health.* (2021) 9:301. doi: 10.3389/fpubh.2021.655447

57. Manuel JI. Racial/ethnic and gender disparities in health care use and access. *Health Serv Res.* (2018) 53:1407–29. doi: 10.1111/1475-6773.12705

58. Callison K, Nguyen BT. The effect of medicaid physician fee increases on health care access, utilization, and expenditures. *Health Serv Res.* (2018) 53:690–710. doi: 10.1111/1475-6773.12698

59. Tobin J, Miller KA, Baezconde-Garbanati L, Unger JB, Hamilton AS, Milam JE. Acculturation, mental health, and quality of life among hispanic childhood cancer survivors: a latent class analysis. *Ethn Dis.* (2018) 28:55–60. doi: 10.18865/ed.28.1.55

60. Kruse CS, Beane A. Health information technology continues to show positive effect on medical outcomes: systematic review. *J Med Internet Res.* (2018) 20:e41. doi: 10.2196/jmir.8793

61. Clauser SB, Wagner EH, Aiello Bowles EJ, Tuzzio L, Greene SM. Improving modern cancer care through information technology. *Am J Prev Med*. (2011) 40:S198–207. doi: 10.1016/j.amepre.2011.01.014

62. Armaou M, Araviaki E, Musikanski L. eHealth and mHealth interventions for ethnic minority and historically underserved populations in developed countries: an umbrella review. *Int J Community Well-Being.* (2020) 3:193–221. doi: 10.1007/s42413-019-00055-5

63. Seligman B, Tuljapurkar S, Rehkopf D. Machine learning approaches to the social determinants of health in the health and retirement study. *SSM Popul Health.* (2018) 4:95–9. doi: 10.1016/j.ssmph.2017.11.008

64. Elk R, Emanuel L, Hauser J, Bakitas M, Levkoff S. Developing and testing the feasibility of a culturally based tele-palliative care consult based on the cultural values and preferences of southern, rural african american and white community members: a program by and for the community. *Health Equity.* (2020) 4:52–83. doi: 10.1089/heq.2019.0120

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.

ORIGINAL RESEARCH
published: 05 November 2021
doi: 10.3389/fpubh.2021.742355

Disparities in Hepatocellular Carcinoma Survival by Insurance Status: A Population-Based Study in China

Jing Wu[1,2*], *Chengyu Liu*[1,2] *and Fengmei Wang*[3*]

[1] School of Pharmaceutical Science and Technology, Tianjin University, Tianjin, China, [2] Center for Social Science Survey and Data, Tianjin University, Tianjin, China, [3] The Department of Gastroenterology and Hepatology, Tianjin Third Central Hospital, Tianjin, China

Objective: Health disparities related to basic medical insurance in China have not been sufficiently examined, particularly among patients with hepatocellular carcinoma (HCC). This study aims to investigate the disparities in HCC survival by insurance status in Tianjin, China.

Methods: This retrospective analysis used data from the Tianjin Basic Medical Insurance claims database, which consists of enrollees covered by Urban Employee Basic Medical Insurance (UEBMI) and Urban and Rural Resident Basic Medical Insurance (URRBMI). Adult patients newly diagnosed with HCC between 2011 and 2016 were identified and followed until death from any cause, withdrawal from UEBMI or URRBMI, or the latest data in the dataset (censoring as of December 31st 2017), whichever occurred first. Patients' overall survival during the follow-up was assessed using Kaplan-Meier and extrapolated by six parametric models. The hazard ratio (HR) and 95% confidence intervals (CI) were calculated with the adjusted Cox proportional hazards model including age at diagnosis, sex, baseline comorbidities and complications, baseline healthcare resources utilization and medical costs, tumor metastasis at diagnosis, the initial treatment after diagnosis and antiviral therapy during the follow-up.

Results: Two thousand sixty eight patients covered by UEBMI ($N = 1{,}468$) and URRBMI ($N = 570$) were included (mean age: 60.6 vs. 60.9, $p = 0.667$; female: 31.8 vs. 27.7%, $p = 0.074$). The median survival time for patients within the UEBMI and URRBMI were 37.8 and 12.2 months, and the 1-, 3-, 5-, 10-year overall survival rates were 63.8, 50.2, 51.0, 33.4, and 44.4, 22.8, 31.5, 13.1%, respectively. Compared with UEBMI, patients covered by URRBMI had 72% (HR: 1.72; 95% CI: 1.47–2.00) higher risk of death after adjustments for measured confounders above. The survival difference was still statistically significant (HR: 1.49; 95% CI: 1.21–1.83) in sensitivity analysis based on propensity score matching.

Conclusions: This study reveals that HCC patients covered by URRBMI may have worse survival than patients covered by UEBMI. Further efforts are warranted to understand healthcare disparities for patients covered by different basic medical insurance in China.

Keywords: hepatocellular carcinoma, survival, health disparities, insurance, China

OPEN ACCESS

Edited by:
Yi Guo,
University of Florida, United States

Reviewed by:
Hui Shao,
University of Florida, United States
Wei Wang,
Merck, United States

****Correspondence:***
Jing Wu
jingwu@tju.edu.cn
Fengmei Wang
wangfengmeitj@126.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: *16 July 2021*
Accepted: *29 September 2021*
Published: *05 November 2021*

Citation:
Wu J, Liu C and Wang F (2021) Disparities in Hepatocellular Carcinoma Survival by Insurance Status: A Population-Based Study in China. Front. Public Health 9:742355. doi: 10.3389/fpubh.2021.742355

INTRODUCTION

Primary liver cancer is the sixth commonly diagnosed cancer and the third leading cause of cancer death worldwide, with about 905,667 new cases and 830,180 deaths in 2020 (1). China is the most afflicted country with almost half of global newly diagnosed patients and fatalities (410,038 new cases and 391,152 deaths in 2020) (2). Moreover, the prognosis for Chinese with primary liver cancer is inferior than other countries and regions, with a 5-year survival probability of only 14.1% (3). Hepatocellular carcinoma (HCC) accounts for ∼90% of all local primary liver cancer, followed by intrahepatic cholangiocarcinoma amongst other types (4). Effective HCC treatment options, depending on the tumor stage and the underlying liver function, include hepatectomy, liver transplantation, transarterial chemoembolization (TACE), ablation, radiotherapy, and systemic therapies. Previous studies have indicated that patients with cancer may alter treatment options to reduce the out-of-pocket expenses and ease their financial burden (5).

Health insurance positively affects cancer diagnostics and treatments as it decreases patients' financial burden (6, 7). A previous study reported that patients with no insurance were more likely to be diagnosed at an advanced stage for all cancers when compared to those with private insurance (7). Furthermore, many studies have claimed that insurance status might be an important prognostic factor because of its impact on access to health care (8–18). Two studies have reported that Medicare or commercial insurance, compared with Medicaid or no insurance, were associated with improved HCC survival in the United States (15, 16). This association was declared in several other cancers, such as breast, lung, colorectal, bladder, multiple myeloma, and follicular lymphoma (9–14). However, the relationship between insurance status and cancer survival has not been extensively studied in China.

As the largest developing country, China has launched basic medical insurance schemes in the 1990s. After more than ten years of development, near-universal health insurance coverage was achieved in 2011, which consisted of three schemes: Urban Employee Basic Medical Insurance (UEBMI) for enrollees; Urban Resident Basic Medical Insurance (URBMI) for children, students and other unemployed adult residents living in urban areas; and New Rural Cooperative Medical Scheme (NRCMS) for all residents living in rural areas (19, 20). As these three insurance schemes were initially designed for individuals with different affordability of healthcare services based on their financial situation, benefits packages were quite different. Compared with UEBMI, patients enrolled in URBMI or NRCMS were underinsured which meant that they had lower reimbursement rate and limited coverage (20). In 2016, the URBMI and NRCMS were merged to form the Urban and Rural Resident Basic Medical Insurance (URRBMI) to improve administrative efficiency (19). URRBMI and UEBMI covered 13.61 billion inhabitants accounting for 96.4% of the total Chinese population in 2020 (21). However, the differences in benefits packages still exist between the current two basic medical insurances in China (21).

So far, only two studies have reported the disparities in cancer survival related to basic medical insurance in China (22, 23). One study revealed that non-small cell lung cancer patients enrolled in insurance plans with higher reimbursement rate or broader coverage (UEBMI or Free Medical Care) had better survival rates than those with inadequate insurance (uninsured or NRCMS) (22). The other study suggested that underinsured patients (NRCMS) faced a higher risk of breast cancer-specific mortality (23). However, the relationship between basic medical insurance and HCC survival was not reported.

Tianjin, one of the four municipalities in China, is the largest coastal city located in the Northern part of mainland China, and ranks 7th among all 31 provinces/municipalities regarding Gross Domestic Product per capita (GDP). In addition, Tianjin is the first provincial-level region that have achieved the integration of URBMI and NRCMS schemes in China and has established a relatively comprehensive basic medical insurance system. By 2020, there were about 11.64 million enrollees (UEBMI: 6.18 million, URRBMI: 5.46 million) in the northern municipality (24). This study aims to investigate the disparities in HCC survival by insurance status in Tianjin, China.

MATERIALS AND METHODS

Data Source

This population-based study was conducted on data obtained from the Tianjin Basic Medical Insurance claims database (2008–2017), which consists of enrollees covered by UEBMI and URRBMI. The database consisted of inpatient, outpatient and pharmacy services claims. Enrollment history, patient demographics (age, sex, working status), dates of service, diagnoses, information on medical prescriptions and procedures, and related costs were recorded in this database. International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10) codes and medical records were used to identify the disease diagnoses. In addition, all-cause mortality information was included in a separate dataset, which could be linked by patients' unique identification number. This study was exempted from applying for ethical approval by the Safety and Ethics Committee of the School of Pharmaceutical Science and Technology, Tianjin University.

Study Population

Males and females, aged over 18 years, with a first discharge diagnosis or outpatient diagnosis of HCC (defined as ICD-10 code C22.0, supplemented by Chinese descriptions), between January 1st 2011 and December 31st 2016, were eligible for inclusion. According to the insurance at the time of diagnosis, patients were grouped into two categories, UEBMI and URRBMI. The date of the first recorded HCC diagnosis was defined as the index date, and the 12 months before the index date was defined as the baseline period. Patients who were not continuously enrolled in the UEBMI or URRBMI during the 12 months prior to the index date, and patients who had history of any malignant neoplasm during the baseline period, were excluded. The cohort was followed until death from any cause, withdrawal from UEBMI or URRBMI, or the latest data in the dataset (censoring as of December 31st 2017), whichever occurred first. All patients in this study were continuously enrolled in only one type of insurance (UEBMI or URRBMI) during the whole study period, including the baseline and follow-up periods.

Outcomes Measures

Overall survival, measured in months, was calculated from the index date to the date of death, December 31st 2017, or the last enrollment date, whichever occurred first. Survival for patients still alive at the end of their follow-up period were censored.

Covariates of Interest

The primary covariate of interest was the insurance status, which was described as UEBMI or URRBMI. Additional covariates of interest included the following: age at diagnosis (categorized as 18–44, 45–54, 55–64, 65–74, or >75 years), sex (male or female), baseline healthcare resources use and medical costs (any hospitalizations, average length of stay per hospitalization, any outpatient visits and total direct medical costs), baseline Charlson Comorbidity Index (CCI) score [computed using an algorithm provided by Quan et al. (25)], liver comorbidities and complications (including hepatitis, liver cirrhosis, fatty liver, alcoholic liver, liver failure, as well as portal hypertension, hepatorenal syndrome, ascites, esophageal variceal bleeding, hepatic encephalopathy, jaundice, and primary peritonitis), tumor status at diagnosis (metastasis or not) and initial treatment after diagnosis which may represented the severity of the disease to some extent and was broadly categorized as curative surgery (including hepatectomy and liver transplantation), non-curative surgery (including TACE and ablation), or no surgery. The ICD-10 codes used for the identification of liver comorbidities and complications were listed were reported in the **Supplementary Table S1**. In addition, antiviral therapy was considered since it could significantly improve the liver function of HCC patients (26, 27). Patients who had at least two prescriptions of antiviral medication during the follow-up period were defined as receiving antiviral therapy.

Statistical Analysis

Descriptive statistics were performed to estimate the patients' characteristics for the UEBMI cohort, the URRBMI cohort and all patients. The *t-test* and the chi-squared test were employed for continuous variable and categorical variables, respectively, to determine the significant differences in characteristics between the two cohorts.

Patients' overall survival during the follow-up period was estimated by the Kaplan-Meier method and compared with a log-rank test. The hazard ratio (HR) and 95% confidence intervals (CI) were calculated with adjusted Cox proportional hazards models. Age at diagnosis, sex, baseline healthcare resources utilization and medical costs were adjusted in the Model A. CCI score and baseline liver comorbidities and complications were additionally included in Model B on the basis of Model A. Model C was adjusted for tumor metastasis at diagnosis and initial treatment after diagnosis as the proxy of the severity of HCC on the basis of Model B. Model D was carried out with additional adjustment for antiviral therapy during the follow-up aiming at excluding the effect of non-anticancer therapy on HCC survival. The proportionality hazards assumption was tested by the Schoenfeld residual method. As the initial treatment may violate the proportionality assumption, the Model C and Model D were stratified by initial treatment (28). In addition, the Cox models were also adjusted for the calendar year; however, due to violating the proportionality assumption and the lack of statistically meaningful differences, it had not been included in the final model.

The lifetime survival beyond the follow-up period for the UEBMI cohort, the URRBMI cohort, and all patients were estimated by extrapolating the Kaplan-Meier survival curves. Six distributions for the parameters were considered, including Exponential, Weibull, Gompertz, Log-logistic, Log-normal and Generalized gamma. The lifetime in this study was defined as 100 years old based on the average life expectancy of Tianjin residents (81.79 years old in 2019) (29). The Log-normal model fitted better than other parameter distributions for all cohorts based on the assessment using Akaike's information criterion (AIC), Bayesian information criterion (BIC), and the visual inspection method (**Supplementary Table S2**, **Supplementary Figures S1–S3**). As the suboptimal distribution, Generalized gamma models were used for sensitivity analysis (**Supplementary Figure S4**).

In addition, to minimize potential bias, propensity scores were calculated by multivariate logistic regression including age, sex, CCI score, liver comorbidities and complications, tumor metastasis at diagnosis, as well as baseline healthcare resources utilization and medical costs (**Supplementary Table S3**). Two matched cohorts were identified using one-to-one nearest neighbor matching without replacement, with a caliper of 0.0008. Sensitivity analysis based on the cohorts after matching was performed to assess the robustness of the results.

The significant level was defined as two-sided alpha = 0.05. All statistical analyses were performed using Stata statistical software (version 13.0; StataCorp, College Station, Texas).

RESULTS

Baseline Characteristics

A total of 2,068 patients newly diagnosed with HCC were identified, of which UEBMI covered 1,468 and 570 were covered by URRBMI (**Figure 1**). The mean age of the total cohort was 60.7 years (UEBMI vs. URRBMI: 60.0 vs. 60.9, $p = 0.667$), with 30.7% females (UEBMI vs. URRBMI: 31.8 vs. 27.7%, $p = 0.074$). Compared with patients covered by UEBMI, those in the URRBMI cohort tended to use fewer healthcare resources (including shorter length of stay per hospitalization and fewer outpatient visits) with lower related medical costs during the baseline period and were with lower CCI scores, but were more likely to be diagnosed with severe liver diseases such as decompensated cirrhosis, liver failure and ascites (**Table 1**).

Short-Term Survival of HCC Patients

During the follow-up period (mean: 25.9 months, median: 16.7 months), 783 and 297 deaths were observed in the UEBMI and the URRBMI cohorts (52.3 vs. 52.1%, $p = 0.947$,

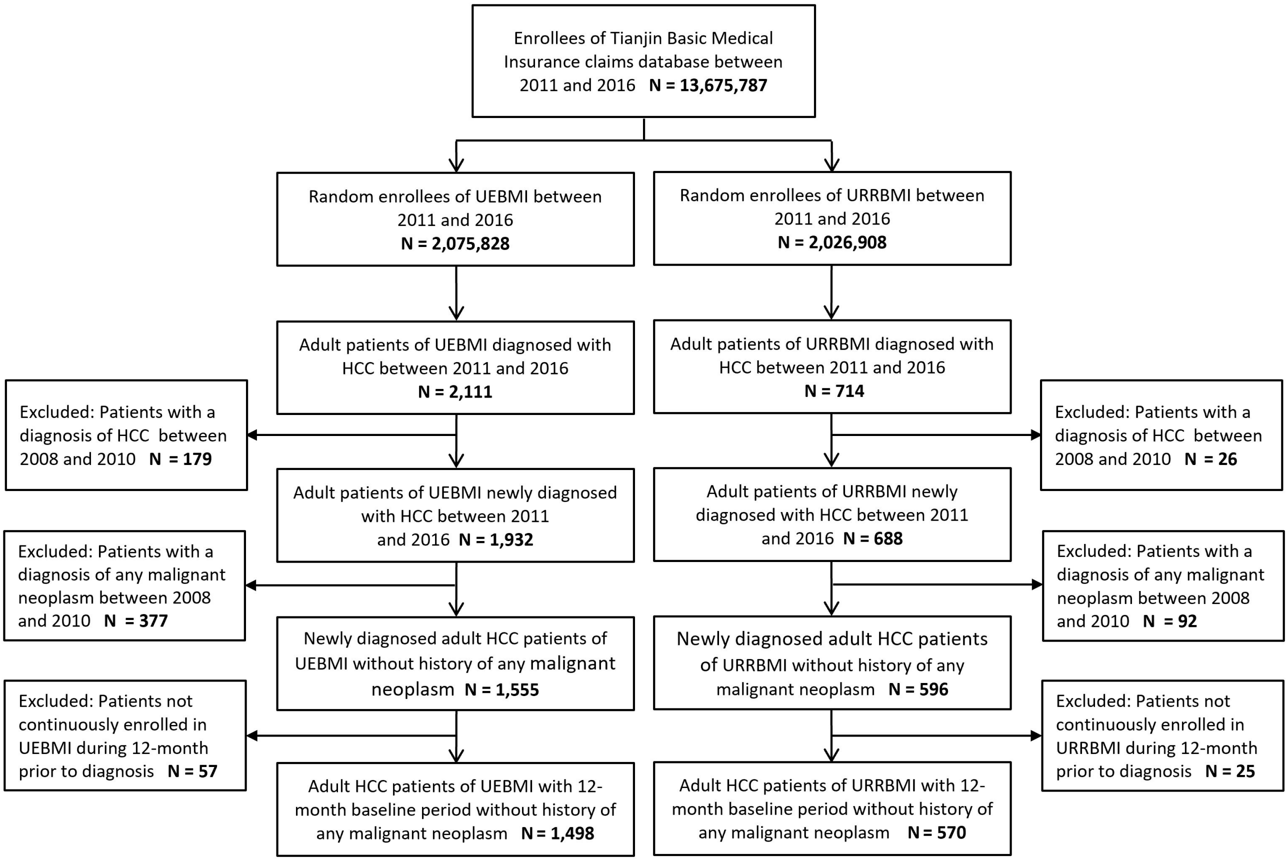

Enrollees of Tianjin Basic Medical Insurance claims database between 2011 and 2016 N = 13,675,787
Random enrollees of UEBMI between 2011 and 2016 N = 2,075,828
Random enrollees of URRBMI between 2011 and 2016 N = 2,026,908
Adult patients of UEBMI diagnosed with HCC between 2011 and 2016 N = 2,111
Adult patients of URRBMI diagnosed with HCC between 2011 and 2016 N = 714
Excluded: Patients with a diagnosis of HCC between 2008 and 2010 N = 179
Excluded: Patients with a diagnosis of HCC between 2008 and 2010 N = 26
Adult patients of UEBMI newly diagnosed with HCC between 2011 and 2016 N = 1,932
Adult patients of URRBMI newly diagnosed with HCC between 2011 and 2016 N = 688
Excluded: Patients with a diagnosis of any malignant neoplasm between 2008 and 2010 N = 377
Excluded: Patients with a diagnosis of any malignant neoplasm between 2008 and 2010 N = 92
Newly diagnosed adult HCC patients of UEBMI without history of any malignant neoplasm N = 1,555
Newly diagnosed adult HCC patients of URRBMI without history of any malignant neoplasm N = 596
Excluded: Patients not continuously enrolled in UEBMI during 12-month prior to diagnosis N = 57
Excluded: Patients not continuously enrolled in URRBMI during 12-month prior to diagnosis N = 25
Adult HCC patients of UEBMI with 12-month baseline period without history of any malignant neoplasm N = 1,498
Adult HCC patients of URRBMI with 12-month baseline period without history of any malignant neoplasm N = 570

FIGURE 1 | Sample selection flowchart. UEBMI, Urban Employee Basic Medical Insurance; URRBMI, Urban and Rural Resident Basic Medical Insurance; HCC, hepatocellular carcinoma.

Supplementary Table S4). 1-, 3-, 5-year overall survival rates among patients covered by UEBMI were 63.8, 51.0, and 44.4%, compared with 50.2, 33.4, and 22.8% among patients covered by URRBMI ($p < 0.001$; **Table 1**, **Figure 2**). There were also statistically significant differences in overall survival among patients in different subgroups (age of 18–44 vs. 45–54 vs. 55–64 vs. 65–74 vs. ≥75; male vs. female; CCI score ≤4 vs. CCI score >4; non-cirrhosis vs. compensated cirrhosis vs. decompensated cirrhosis) based on Kaplan-Meier methods and log-rank tests (**Supplementary Figure S5**). In addition, 1-, 3-, 5-year overall survival rates for all patients with HCC were shown in **Table 2**.

In the multiple Cox proportional hazards model that adjusted for measured confounders (**Table 3**), patients covered by URRBMI had a 72% higher risk of death than those covered by UEBMI (HR: 1.72; 95% CI: 1.47–2.00). Compared with adult patients younger than 45 years old, there was worse survival among patients who were at least 45 years old (45–54 years, HR: 1.79; 95% CI: 1.23–2.59; 55–64 years, HR: 2.45; 95% CI: 1.72–3.49; 65–74 years, HR: 3.43; 95% CI: 2.40–4.89; ≥75 years, HR: 5.25; 95% CI: 3.66–7.53). Moreover, male patients had a higher risk of death than females (HR: 1.74; 95% CI: 1.51–2.00). In addition, some factors also appeared to be associated with decreased or increased survival in the multiple Cox model, including compensated cirrhosis (HR: 1.34; 95% CI: 1.12–1.60), decompensated cirrhosis (HR: 1.85; 95% CI: 1.54–2.21), baseline outpatient visit (HR: 2.01; 95% CI: 1.72–2.34), tumor metastasis (HR: 2.58; 95% CI: 2.19–3.04), fatty liver disease (HR: 0.66; 95% CI: 0.48–0.91), antiviral therapy (HR: 0.52; 95% CI: 0.43–0.62). In addition, the results of the Model A, Model B and Model C with adjustment for fewer variables were shown in **Supplementary Table S5**.

Lifetime Survival of HCC Patients

Based on the total cohort's mean age (60.7 years old), overall survival curves were extrapolated to 40 years (i.e., 480 months) after the diagnosis of HCC to cover the lifetime horizon (**Figure 3**). The 10-year survival rates among the total cohort, the UEBMI cohort and the URRBMI cohort were 27.1, 31.5, and 13.1%, respectively. The results using Generalized gamma models did not vary significantly from those observed in the Log-normal models (**Supplementary Figure S6**).

Sensitivity Analysis

The survival difference was reduced but still statistically significant (URRBMI vs. UEBMI, HR: 1.49; 95% CI: 1.21–1.83) in the two cohorts after propensity score matching. Baseline characteristics, Kaplan-Meier survival curves and

TABLE 1 | Baseline characteristics for patients with HCC.

	Overall **(*N* = 2,068)**	**UEBMI** **(*N* = 1,498)**	**URRBMI** **(*N* = 570)**	***P***
Demographic characteristics				
Age [Mean (SD)]	60.7 (12.6)	60.6 (12.9)	60.9 (11.6)	0.667
Female [*N* (%)]	634 (30.7%)	476 (31.8%)	158 (27.7%)	0.074
Comorbidities and complications [*N* (%)]				
CCI score [Mean (SD)]	4.44 (2.16)	4.69 (2.23)	3.78 (1.81)	**<0.001**
Comorbidities related to the liver				
Hepatitis	916 (44.3%)	660 (44.1%)	256 (44.9%)	0.727
HBV	741 (35.8%)	525 (35.0%)	216 (37.9%)	0.227
HCV	82 (4.0%)	69 (4.6%)	13 (2.3%)	**0.015**
Cirrhosis of the liver	939 (45.4%)	663 (44.3%)	276 (48.4%)	0.089
Compensated cirrhosis	490 (23.7%)	365 (24.4%)	125 (21.9%)	0.244
Decompensated cirrhosis†	449 (21.7%)	298 (19.9%)	151 (26.5%)	**0.001**
Hepatic failure	266 (12.9%)	176 (11.7%)	90 (15.8%)	**0.014**
Fatty liver disease	92 (4.4%)	79 (5.3%)	13 (2.3%)	**0.003**
Alcoholic liver disease‡	52 (2.5%)	42 (2.8%)	10 (1.8%)	0.173
Ascites	366 (17.7%)	232 (15.5%)	134 (23.5%)	**<0.001**
Hepatic encephalopathy	202 (9.8%)	151 (10.1%)	51 (8.9%)	0.438
Jaundice	137 (6.6%)	104 (6.9%)	33 (5.8%)	0.346
Portal hypertension	82 (4.0%)	57 (3.8%)	25 (4.4%)	0.545
Esophageal variceal bleeding	54 (2.6%)	38 (2.5%)	16 (2.8%)	0.731
Primary peritonitis	52 (2.5%)	37 (2.5%)	15 (2.6%)	0.834
Hepatorenal syndrome	35 (1.7%)	31 (2.1%)	4 (0.7%)	**0.031**
All-cause resource utilization and costs				
Total cost [Mean(SD), CNY]	7,505 (16,870)	8,940 (17,501)	3,733 (14, 435)	**<0.001**
Any hospitalizations [*N* (%)]	428 (20.7%)	314 (21.0%)	114 (20.0%)	0.630
ALOS per hospitalization [Mean(SD)]	12.9 (10.5)	14.0 (11.4)	9.9 (6.6)	**<0.001**
Any outpatient visits [*N* (%)]	1,578 (76.3%)	1,424 (95.1%)	154 (27.0%)	**<0.001**

CCI score, Charlson Comorbidity Index score; HBV, hepatitis B virus; HCV, hepatitis C virus; CNY, Chinese yuan (year-2017 1 USD = 6.77 CNY); ALOS, Average length of stay.

†*Patients with liver cirrhosis who had the following symptoms were defined as decompensated liver cirrhosis: ascites; esophageal variceal bleeding; hepatorenal syndrome; portal hypertension; hepatic encephalopathy and jaundice; hepatic encephalopathy and primary peritonitis; jaundice and primary peritonitis.*

‡*Including alcoholic liver cirrhosis, alcoholic hepatitis, alcoholic fatty liver disease and alcoholic liver failure; hepatitis, liver cirrhosis, fatty liver disease and liver failure in this table only included non-alcoholic disease. Bold values means P < 0.05.*

multiple Cox proportional hazards models for the two cohorts after matching were reported in the supplementary (**Supplementary Tables S3, S6**, **Supplementary Figure S6**).

DISCUSSION

To the best of our knowledge, this is the first study to investigate the disparities in HCC survival by basic medical insurance in China as well as the first study to examine the discrepancies between UEBMI and URRBMI. In this population-based study, we found evidence of disparities in HCC survival by insurance status; the patients enrolled in URRBMI might have a higher risk of death than those enrolled in UEBMI whether during the follow-up period or over their lifetime.

Similar results were found in previous studies. In a study based on the data derived from Beijing Cancer Registry, underinsured (uninsured or NRCMS) patients with non-small cell lung cancer had shorter cancer-specific survival than well-insured (UEBMI or Free Medical Care) individuals (HR: 1.24; 95% CI: 1.03–1.49; $P = 0.021$) after adjusting for age, sex, cancer stage, smoking status, family history and residential area (22). Another study based on the Breast Cancer Information Management System in Sichuan West China Hospital has also suggested that patients covered by rural schemes (i.e., NRCMS) faced a higher risk of breast cancer-specific mortality (HR: 1.29; 95% CI: 1.00–1.65; $P = 0.046$) than those covered by urban schemes (URBMI, UEBMI, and/or commercial insurances) when adjusted for age, calendar year at diagnosis, ethnic group, education level, marital status, comorbidity, tumor characteristics (for example, histological type, hormone receptor status, tumor stage) and treatment (23). Compared with the previous studies, the HR of death for UEBMI and URRBMI cohorts in this study was larger (primary analysis: HR: 1.72; 95% CI: 1.47–2.00; sensitivity analysis HR: 1.49; 95% CI: 1.21–1.83). A possible reason might be that variables related to socioeconomic status (SES), such as educational level, income, and work status, were not sufficiently considered in this study. Enrollees in UEBMI

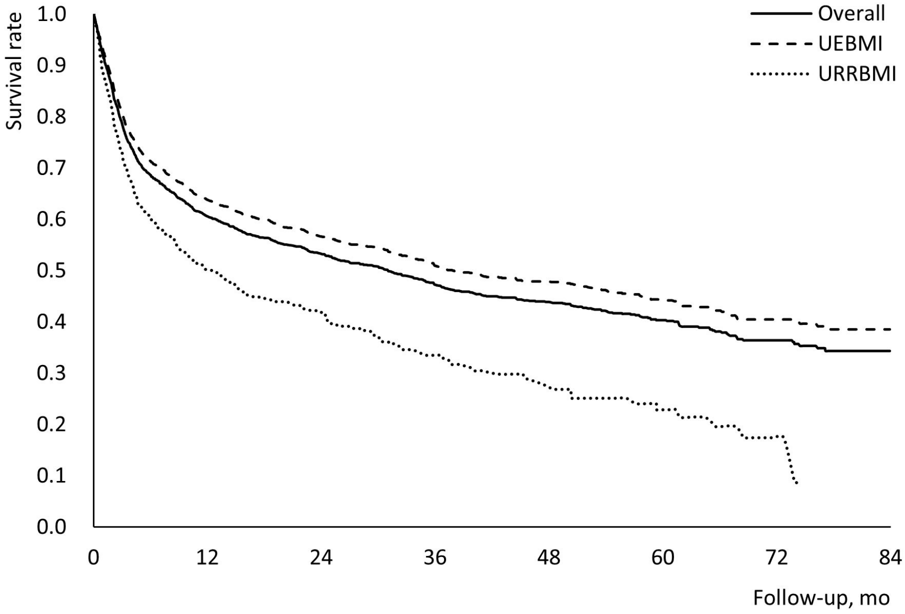


FIGURE 2 | Kaplan-Meier survival curves for patients with HCC during the follow-up period. UEBMI, Urban Employee Basic Medical Insurance; URRBMI, Urban and Rural Resident Basic Medical Insurance; HCC, hepatocellular carcinoma.

TABLE 2 | Patients' overall survival during the follow-up period.

	Overall	**UEBMI**	**URRBMI**
	(*N* = 2,068)	**(*N* = 1,498)**	**(*N* = 570)**
Overall survival, mo.			
Median	31.0	37.8	12.2
Mean[95%CI]	40.7* [39.0, 42.4]	43.8* [41.9, 45.7]	27.1* [39.0, 42.4]
Survival rate[95%CI]			
1-year	60.6 [58.4, 62.7]	63.8 [61.3, 66.2]	50.2 [45.5, 54.8]
3-year	47.3 [44.9, 49.6]	51.0 [48.3, 53.6]	33.4 [28.3, 38.6]
5-year	40.3 [37.8, 42.8]	44.4 [41.5, 47.2]	22.8 [17.1, 29.0]

**Largest observed analysis time was censored; mean was underestimated.*

TABLE 3 | Multivariate analysis for overall survival in patients with HCC.

	HR	***P***	**95%CI**
URRBMI (vs. UEBMI)	1.72	**<0.001**	1.47–2.00
Age (vs. 18–44)			
45–54	1.79	**0.002**	1.23–2.59
55–64	2.45	**<0.001**	1.72–3.49
65–74	3.43	**<0.001**	2.40–4.89
≥75	5.25	**<0.001**	3.66–7.53
Male (vs. female)	1.74	**<0.001**	1.51–2.00
CCI score	1.01	0.497	0.98–1.04
Compensated cirrhosis (vs. No)	1.34	**0.001**	1.12–1.60
Decompensated cirrhosis (vs. No)	1.85	**<0.001**	1.54–2.21
Hepatitis (vs. No)	0.92	0.314	0.78–1.08
Alcoholic liver disease (vs. No)	0.88	0.457	0.62–1.24
Fatty liver disease (vs. No)	0.66	**0.012**	0.48–0.91
Hepatic failure (vs. No)	0.99	0.890	0.81–1.19
Baseline total cost	1.00	0.561	1.00–1.00
Baseline ALOS	1.00	0.556	0.99–1.00
Any baseline outpatient visits (vs. No)	2.01	**<0.001**	1.72–2.34
Tumor metastasis at diagnosis (vs. No)	2.58	**<0.001**	2.19–3.04
Antiviral therapy during the follow-up (vs. No)	0.52	**<0.001**	0.43–0.62

The Cox model was stratified by initial treatment after diagnosis and was broadly categorized as curative surgery (including hepatectomy and liver transplantation), non-curative surgery (including transarterial chemoembolization [TACE] and ablation), or no surgery. CCI score, Charlson Comorbidity Index score; ALOS, Average length of stay. Bold values means $P < 0.05$.

always have a relatively higher SES and may pay close attention to health status, get more cancer screenings, and have full access to medical treatment. Additionally, some studies have also demonstrated that lower SES was associated with worse HCC-specific survival (30–32). Furthermore, previous studies examining the relationship between insurance and survival in other countries also have shown that patients with a good insurance status have better survival than those with poor insurance status, not only among HCC patients, but also among many other cancers (8–18).

Several mechanisms may contribute to the observed disparities in HCC survival between UEBMI and URRBMI. Firstly, patients with poor benefit packages are likely to have less access to healthcare (33). In this study, patients in the URRBMI cohort used fewer healthcare resources during the baseline period and had lower CCI scores. However, it did not mean that patients covered by URRBMI were in better health status, because they were found to be more likely to have some sorts of severe liver

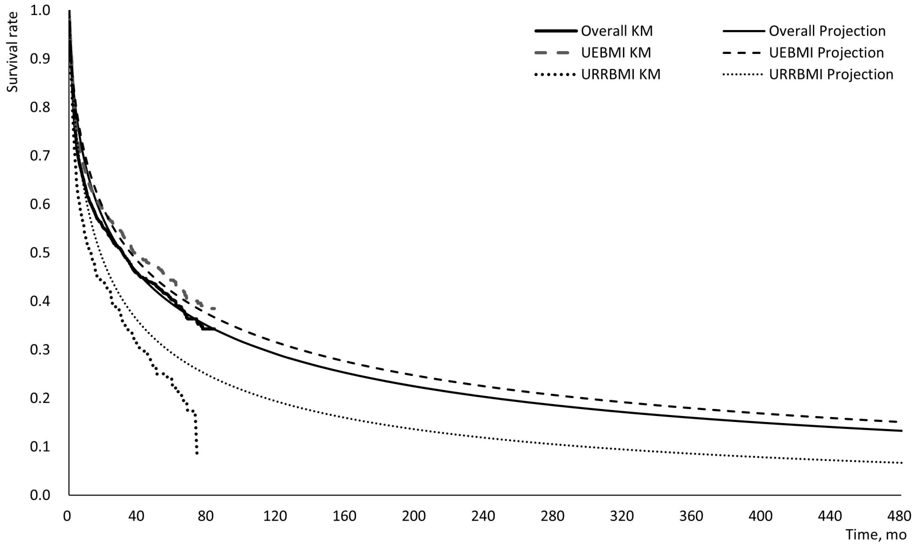


FIGURE 3 | Log-normal projection survival curves for patients with HCC during the lifetime. UEBMI, Urban Employee Basic Medical Insurance; URRBMI, Urban and Rural Resident Basic Medical Insurance; KM, Kaplan-Meier; HCC, hepatocellular carcinoma.

diseases including decompensated cirrhosis, liver failure and ascites during the baseline period. In addition, previous studies have reported that patients with inadequate insurance tended to receive cancer screening less frequently and were more likely to have an advanced stage of malignancy at diagnosis, which may be related to worse survival (22, 23, 34, 35). Even without the tumor stage variables in the database, this study showed that more patients in the URRBMI cohort had metastasized at diagnosis (**Supplementary Table S7**). In addition, when tumor metastasis at diagnosis and initial treatment after diagnosis were adjusted in the model, HR decreased from 2.14 (95% CI: 1.84–2.49, Model B) to 1.86 (95% CI: 1.59–2.16, Model C), which suggests that the survival disparity between UEBMI and URRBMI may exist before diagnosis (**Supplementary Table S5**). Secondly, the insurance status may also impact the treatment options, especially for uncovered therapies or with higher out-of-pocket expenses. Disparities in treatment by insurance status have been observed in the United States, with privately insured patients with HCC consistently being more likely to receive hepatectomy (35, 36). Some studies indicated that cancer patients with lower reimbursement rates were less likely to receive adjuvant chemotherapy and postoperative radiation therapy and were less likely to afford the high out-of-pocket expenses for an emerging therapy that significantly improved survival (e.g., targeted agents and immune agents) in China (23, 37). Herein, there were 11.7, 26.7, 61.6% and 11.6, 29.8, 58.6% patients with curative surgery, non-curative surgery, and no surgery for HCC patients in the UEBMI and the URRBMI cohorts, respectively ($P = 0.356$; **Supplementary Table S7**). There was no significant difference in receiving surgery between the two cohorts. Still, the preoperative and postoperative adjuvant therapy was not further analyzed due to insufficient power. Sorafenib was the only emerging drug approved for advanced HCC during the study period, but the basic medical insurance had not covered it by December 2017. Therefore, we could not examine whether more HCC patients enrolled in UEBMI had been treated with Sorafenib. In addition, the insurance status can be an indicator for health consciousness, health habits or socioeconomic status in this study, which might contribute to the survival (37, 38).

To understand the potential mechanisms contributing to the observed disparities in HCC survival, some additional analyses on the relationship between reimbursement rate (defined as the anti-cancer medical costs paid by basic medical insurance divided by the anti-cancer total costs in the insurance coverage) and HCC survival had been conducted. When the reimbursement rate was additionally adjusted in Model D, the HR of insurance type (URRBMI vs. UEBMI) decreased from 1.49 (95% CI: 1.21–1.83) to 1.42 (95% CI: 1.13–1.79) among matched cohorts (see **Supplementary Table S8**), which suggests that small part of the disparity in survival between UEBMI and URRBMI may be attributed to reimbursement rate. But further research is warranted to clarify the mechanisms by which health insurance affects survival.

Some factors also appeared to be associated with HCC survival in this study, consistent with previous studies. The risk of death increased with age, and patients who were 45 years old or older had a significantly higher risk of death than those younger than 45 years old. Males with HCC had worse survival than females, which was well-established in a previous study recruiting Americans (15, 39). Liver cirrhosis including compensated cirrhosis and decompensated cirrhosis were also related to the decreased survival, which was demonstrated among patients with HCC in Taiwan, China (40). Notably, some studies have indicated that hepatitis and liver cirrhosis are risk factors for HCC, and chronic hepatitis might lead to cirrhosis and then to HCC or other types of liver cancer. About 45% of patients had hepatitis (mainly HBV and HCV) or cirrhosis, respectively, before being diagnosed with HCC in this study. Therefore, regular screening and monitoring for patients with hepatitis or cirrhosis may contribute to the earlier diagnosis and better survival.

Antiviral therapy was also found to be associated with increased survival. To be mentioned, there were about 44.1 and 44.9% of patients in UEBMI and URRBMI cohorts with hepatitis during the baseline period, but the proportion of patients taking antiviral therapy were only 22.6 and 9.3% during the follow-up period (**Supplementary Table S7**). It is possible that some patients were cured during the baseline period. Still, HCC patients with hepatitis in the URRBMI cohort were less likely to receive antiviral treatment than those in the UEBMI cohort. Literature also reported that some antiviral regimens had better efficacy but were more expensive, and the benefits of these new antiviral regimens might not be accessible to all patients (41). Fatty liver disease was also associated with increased survival in the primary analysis (HR: 0.66; 95% CI: 0.48–0.91), but the association attenuated in the sensitivity analysis (HR: 0.54; 95% CI: 0.26–1.11). As fatty liver is a disease with no apparent clinical symptom, patients with URRBMI were more likely undiagnosed based on the discussion above. Therefore, the impact of fatty liver showed by the primary analysis might be biased. In addition, as there were no tumor stage variables in the database, we examined the tumor metastasis at diagnosis in the multiple Cox models, which was demonstrated to be associated with decreased HCC survival. Our findings highlight the importance of early screening and diagnosis for high-risk individuals.

Furthermore, this is also the first study to examine the survival of patients with HCC in mainland China. The median survival time was 31.0 months, which was similar to that of the Chinese patients in the U.S. (34.0 months) (42). The 1-, 3-, 5-, 10-year overall survival rates in this study were 60.6, 47.3, and 40.3%, respectively, which were slightly lower than in Taiwan, China (71.68, 57.14, and 47.82%) (40).

There are also some limitations to this study. Firstly, this study was conducted based on the Basic Medical Insurance claims database in Tianjin. The disparities in benefit packages by UEBMI and URRBMI may differ from those in other provinces. However, compared with UEBMI, patients enrolled in URRBMI continuously suffer from poorer benefit packages in almost all regions of China. Therefore, the results presented in this study, to a certain degree, could reflect the disparities in HCC survival by basic medical insurance in China. Secondly, this study did not examine the HCC-specific survival due to a lack of related information in the database. Nevertheless, studies that examined both cancer-specific survival and all-cause survival have reported similar results for the two outcome measures (15, 23). Thirdly, the database does not collect data on clinical characteristics (e.g., tumor stage), health behaviors (e.g., smoking, drinking), SES (e.g., education, income, work status) and private insurance. These factors likely differ between patients enrolled in UEBMI and URRBMI, especially SES and private insurance. If we were able to control for these factors, the HR of death for UEBMI and URRBMI cohorts in this study might decrease. Lastly, emerging therapies (i.e., Sorafenib) and some other prognostic factors (e.g., time to treatment, the preoperative and postoperative adjuvant therapy, complications related to the therapy and the treatment) related to treatment were not included, which might have an impact on HCC survival. Future studies using richer information on clinical characteristics, treatments, and SES are warranted to understand better the HCC survival disparities examined in this study.

CONCLUSION

This study reveals that HCC patients covered by URRBMI may have worse survival than patients covered by UEBMI. Further efforts are warranted to understand healthcare disparities for patients covered by different basic medical insurance in China.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study were available from the Tianjin Healthcare Security Administration. Due to the requirement from the data owner, these data could only be used for this study under the license, which could not be shared to others. Requests to access these datasets should be directed to Tianjin Healthcare Security Administration.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Safety and Ethics Committee of the School of Pharmaceutical Science and Technology in Tianjin University. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

JW and CL designed the study and supervised data collection. CL analyzed the data, interpreted the results, and drafted the manuscript. JW and FW critically reviewed the manuscript. All authors have read and agreed to the published version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.742355/full#supplementary-material

REFERENCES

1. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, et al. *International Agency for Research on Cancer. Global Cancer Observatory: Cancer Today*. Available online at: https://gco.iarc.fr/today (accessed March 15, 2021).
2. Liu Z, Jiang Y, Yuan H, Fang Q, Cai N, Suo C, et al. The trends in incidence of primary liver cancer caused by specific etiologies: results from the Global Burden of Disease Study 2016 and implications for liver cancer prevention. *J Hepatol*. (2019) 70:674–83. doi: 10.1016/j.jhep.2018.12.001
3. Bouzbid S, Hamdi-Chérif M, Zaidi Z, Meguenni K, Regagba D, Bayo S, et al. Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *Lancet*. (2018) 391:1023–75. doi: 10.1016/S0140-6736(17)33326-3
4. Zhou J, Sun HC, Wang Z, Cong WM, Wang JH, Zeng MS, et al. Guidelines for diagnosis and treatment of primary liver cancer in China (2017 Edition). *Liver Cancer*. (2018) 7:235–60. doi: 10.1159%2F000488035
5. Ellis L, Canchola AJ, Spiegel D, Ladabaum U, Haile R, Gomez SL. Trends in cancer survival by health insurance status in California from 1997 to 2014. *JAMA Oncol*. (2018) 4:317–23. doi: 10.1001/jamaoncol.2017.3846
6. Ward E, Halpern M, Schrag N, Cokkinides V, DeSantis C, Bandi P, et al. Association of insurance with cancer care utilization and outcomes. *CA-A Cancer J Clinicians*. (2008) 58:9–31. doi: 10.3322/CA.2007.0011
7. Ward EM, Fedewa SA, Cokkinides V, Virgo K. The association of insurance and stage at diagnosis among patients aged 55 to 74 years in the national cancer database. *Cancer J*. (2010) 16:614–21. doi: 10.1097/PPO.0b013e3181ff2aec
8. Hwang KT, Ju YW, Kim YA, Kim J, Oh S, Jung J, et al. Prognostic influence of Korean public medical insurance system on breast cancer patients. *Annals Surg Treat Res*. (2019) 96:58–69. doi: 10.4174/astr.2019.96.2.58
9. Jain V, Venigalla S, Sebro RA, Karakousis GC, Wilson RJ II, Weber KL, et al. Association of health insurance status with presentation, treatment and outcomes in soft tissue sarcoma. *Cancer Med*. (2019) 8:6295–304. doi: 10.1002/cam4.2441
10. Goldstein JS, Nastoupil LJ, Han X, Jemal A, Ward E, Flowers CR. Disparities in survival by insurance status in follicular lymphoma. *Blood*. (2018) 132:1159–66. doi: 10.1182/blood-2018-03-839035
11. Rice SR, Vyfhuis MAL, Scilla KA, Burrows WM, Bhooshan N, Suntharalingam M, et al. Insurance status is an independent predictor of overall survival in patients with stage III non-small-cell lung cancer treated with curative intent. *Clinical Lung Cancer*. (2020) 21:e130–e41. doi: 10.1016/j.cllc.2019.08.009
12. Niu X, Roche LM, Pawlish KS, Henry KA. Cancer survival disparities by health insurance status. *Cancer Med*. (2013) 2:403–11. doi: 10.1002/cam4.84
13. Perry AM, Brunner AM, Zou T, McGregor KL, Amrein PC, Hobbs GS, et al. Association between insurance status at diagnosis and overall survival in chronic myeloid leukemia: a population-based study. *Cancer*. (2017) 123:2561–9. doi: 10.1002/cncr.30639
14. Nazemi A, Ghodoussipour S, Pearce S, Bhanvadia S, Daneshmand S. Socioeconomic and insurance status are independent prognostic indicators of higher disease stage and worse prognosis in bladder cancer. *Urol Oncol*. (2019) 37:784–90. doi: 10.1016/j.urolonc.2019.04.021
15. Adler Jaffe S, Myers O, Meisner ALW, Wiggins CL, Hill DA, McDougall JA. Relationship between insurance type at diagnosis and hepatocellular carcinoma survival. *Cancer Epidemiol Biomark Prevent*. (2020) 29:300–7. doi: 10.1158/1055-9965.EPI-19-0902
16. Wang J, Ha J, Lopez A, Bhuket T, Liu B, Wong RJ. Medicaid and uninsured hepatocellular carcinoma patients have more advanced tumor stage and are less likely to receive treatment. *J Clin Gastroenterol*. (2018) 52:437–43. doi: 10.1097/MCG.0000000000000859
17. Jang JS, Shin DG, Cho HM, Kwon Y, Cho DH, Lee KB, et al. Differences in the survival of gastric cancer patients after gastrectomy according to the medical insurance status. *J Gastric Cancer*. (2013) 13:247–54. doi: 10.5230/jgc.2013.13.4.247
18. Naghavi AO, Echevarria MI, Grass GD, Strom TJ, Abuodeh YA, Ahmed KA, et al. Having Medicaid insurance negatively impacts outcomes in patients with head and neck malignancies. *Cancer*. (2016) 122:3529–37. doi: 10.1002/cncr.30212
19. Pan XF, Xu J, Meng Q. Integrating social health insurance systems in China. *Lancet*. (2016) 387:1274–5. doi: 10.1016/S0140-6736(16)30021-6
20. Fang H, Eggleston K, Hanson K, Wu M. Enhancing financial protection under China's social health insurance to achieve universal health coverage. *BMJ*. (2019) 365:l2378. doi: 10.1136/bmj.l2378
21. National Healthcare Security Administration. *The People's Republic of China Statistical Bulletin on the Development of Medical Security in 2020*. Available online at: http://www.nhsa.gov.cn/art/2021/6/8/art_7_5232.html (accessed July 15, 2020).
22. Wang Z, Yang L, Liu S, Li H, Zhang X, Wang N, et al. Effects of insurance status on long-term survival among non-small cell lung cancer (NSCLC) patients in Beijing, China: a population-based study. *Chinese J Cancer Res*. (2020) 32:596–604. doi: 10.21147/j.issn.1000-9604.2020.05.04
23. Xie Y, Valdimarsdóttir UA, Wang C, Zhong X, Gou Q, Zheng H, et al. Public health insurance and cancer-specific mortality risk among patients with breast cancer: a prospective cohort study in China. *Int J Cancer*. (2021) 148:28–37. doi: 10.1002/ijc.33183
24. Tianjin Municipal People's Government. *Tianjin Statistical Bulletin on National Economic and Social Development in 2020*. Available online at: http://stats.tj.gov.cn/tjsj_52032/tjgb/202103/t20210317_5386752.html (accessed March 17, 2021).
25. Quan H, Li B, Couris CM, Fushimi K, Graham P, Hider P, et al. Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries. *Am J Epidemiol*. (2011) 173:676–82. doi: 10.1093/aje/kwq433
26. Kamp WM, Sellers CM, Stein S, Lim JK, Kim HS. Direct-acting antivirals improve overall survival in interventional oncology patients with hepatitis C and hepatocellular carcinoma. *J Vascular Int Radiol*. (2020) 31:953–60. doi: 10.1016/j.jvir.2019.12.809
27. Singal AG, Rich NE, Mehta N, Branch A, Pillai A, Hoteit M, et al. Direct-acting antiviral therapy for HCV infection is associated with increased survival in patients with a history of hepatocellular carcinoma. *Gastroenterology*. (2019) 157:1253–63.e2. doi: 10.1053/j.gastro.2019.07.040
28. Kleinbaum DG, Klein M. *Survival Analysis: A Self-Learning Text (Statistics for Biology and Health)*. New York, NY: Springer (2011). doi: 10.1007/978-1-4419-6646-9
29. Tianjin Health Commission. *Tianjin Residents Health Status Report 2019*. Available online at: http://wsjk.tj.gov.cn/ZTZL1/ZTZL750/YQFKZL9424/FKDT1207/202009/t20200930_23945345.html (accessed March 21, 2021).
30. Sellers CM, Uhlig J, Ludwig JM, Taddei T, Stein SM, Lim JK, et al. The impact of socioeconomic status on outcomes in hepatocellular carcinoma: inferences from primary insurance. *Cancer Med*. (2019) 8:5948–58. doi: 10.1002/cam4.2251
31. Artinyan A, Mailey B, Sanchez-Luege N, Khalili J, Sun CL, Bhatia S, et al. Race, ethnicity, and socioeconomic status influence the survival of patients with hepatocellular carcinoma in the United States. *Cancer*. (2010) 116:1367–77. doi: 10.1002/cncr.24817
32. Abdel-Rahman O. Treatment choices and outcomes of non-metastatic hepatocellular carcinoma patients in relationship to neighborhood socioeconomic status: a population-based study. *Int J Clin Oncol*. (2020) 25:861–6. doi: 10.1007/s10147-020-01616-x
33. Meng Q, Fang H, Liu X, Yuan B, Xu J. Consolidating the social health insurance schemes in China: towards an equitable and efficient health system. *Lancet*. (2015) 386:1484–92. doi: 10.1016/S0140-6736(15)00342-6
34. Abdelsattar ZM, Hendren S, Wong SL. The impact of health insurance on cancer care in disadvantaged communities. *Cancer*. (2017) 123:1219–27. doi: 10.1002/cncr.30431
35. Zaydfudim V, Whiteside MA, Griffin MR, Feurer ID, Wright JK, Pinson CW. Health insurance status affects staging and influences treatment strategies in patients with hepatocellular carcinoma. *Annals Surg Oncol*. (2010) 17:3104–11. doi: 10.1245/s10434-010-1181-2
36. Hoehn RS, Hanseman DJ, Jernigan PL, Wima K, Ertel AE, Abbott DE, et al. Disparities in care for patients with curable hepatocellular carcinoma. *HPB*. (2015) 17:747–52. doi: 10.1111/hpb.12427
37. Li X, Zhou Q, Wang X, Su S, Zhang M, Jiang H, et al. The effect of low insurance reimbursement on quality of care for non-small cell lung cancer in China: a comprehensive study covering diagnosis, treatment, and outcomes. *BMC Cancer*. (2018) 18:683. doi: 10.1186/s12885-018-4608-y

38. Bittoni MA, Wexler R, Spees CK, Clinton SK, Taylor CA. Lack of private health insurance is associated with higher mortality from cancer and other chronic diseases, poor diet quality, and inflammatory biomarkers in the United States. *Prev Med*. (2015) 81:420–6. doi: 10.1016/j.ypmed.2015.09.016

39. Yang D, Hanna DL, Usher J, LoCoco J, Chaudhari P, Lenz HJ, et al. Impact of sex on the survival of patients with hepatocellular carcinoma: a Surveillance, Epidemiology, and End Results analysis. *Cancer*. (2014) 120:3707–16. doi: 10.1002/cncr.28912

40. Nguang SH, Wu CK, Liang CM, Tai WC, Yang SC, Ku MK, et al. Treatment and cost of hepatocellular carcinoma: a population-based cohort study in Taiwan. *Int J Environ Res Public Health*. (2018) 15:2655. doi: 10.3390/ijerph15122655

41. Younossi Z, Gordon SC, Ahmed A, Dieterich D, Saab S, Beckerman R. Treating medicaid patients with hepatitis C: clinical and economic impact. *Am J Managed Care*. (2017) 23:107–12.

42. Ren F, Zhang J, Gao Z, Zhu H, Chen X, Liu W, et al. Racial disparities in the survival time of patients with hepatocellular carcinoma and intrahepatic cholangiocarcinoma between Chinese patients and patients of other racial groups: a population-based study from 2004 to 2013. *Oncol Lett*. (2018) 16:7102–16. doi: 10.3892/ol.2018.9550

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

MINI REVIEW
published: 11 November 2021
doi: 10.3389/fpubh.2021.738253

Social Determinants of Health in Physiatry: Challenges and Opportunities for Clinical Decision Making and Improving Treatment Precision

Rosalynn R. Z. Conic [1†], Carolyn Geis [2] and Heather K. Vincent [2†]*

[1] *Department of Family Medicine and Public Health, University of California, San Diego, San Diego, CA, United States,*
[2] *Department of Physical Medicine and Rehabilitation, University of Florida, Gainesville, FL, United States*

OPEN ACCESS

Edited by:
Fei Wang,
Cornell University, United States

Reviewed by:
Yan Zhang,
Bristol Myers Squibb, United States
Yu Wang,
Oak Ridge Institute for Science and Education (ORISE), United States

****Correspondence:***
Heather K. Vincent
hkvincent@ufl.edu

† These authors have contributed equally to this work and share senior 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: *08 July 2021*
Accepted: *11 October 2021*
Published: *11 November 2021*

Citation:
Conic RRZ, Geis C and Vincent HK (2021) Social Determinants of Health in Physiatry: Challenges and Opportunities for Clinical Decision Making and Improving Treatment Precision. Front. Public Health 9:738253. doi: 10.3389/fpubh.2021.738253

Physiatry is a medical specialty focused on improving functional outcomes in patients with a variety of medical conditions that affect the brain, spinal cord, peripheral nerves, muscles, bones, joints, ligaments, and tendons. Social determinants of health (SDH) play a key role in determining therapeutic process and patient functional outcomes. Big data and precision medicine have been used in other fields and to some extent in physiatry to predict patient outcomes, however many challenges remain. The interplay between SDH and physiatry outcomes is highly variable depending on different phases of care, and more favorable patient profiles in acute care may be less favorable in the outpatient setting. Furthermore, SDH influence which treatments or interventional procedures are accessible to the patient and thus determine outcomes. This opinion paper describes utility of existing datasets in combination with novel data such as movement, gait patterning and patient perceived outcomes could be analyzed with artificial intelligence methods to determine the best treatment plan for individual patients in order to achieve maximal functional capacity.

Keywords: big data, physical function, outcomes, physiatry, physical medicine and rehabilitation, social determinants of health

INTRODUCTION

Physical medicine and rehabilitation, or physiatry, is a specialty that treats medical conditions affecting the brain, spinal cord, peripheral nerves, joints, muscle, bone, tendons and ligaments. The main treatment goal of physiatry is to maximize function and independent living. Physiatry care spans the entire continuum of health care from consultation in the acute care hospital to post-acute inpatient rehabilitation, home health, outpatient, and community re-integration. Patients move through these levels of care as they gain functional independence or have a need for ongoing care (**Figure 1**). Patients enter the healthcare system at different "starting points" in the care spectrum. At each phase of care and transition, physiatrists coordinate patient care and make critical decisions regarding rehabilitation needs based on medical status and functional progress. This decision-making is made more complex by the wide diversity of patient types, socioeconomic backgrounds, medical conditions, injury complexity, and patient-family perception of needs.

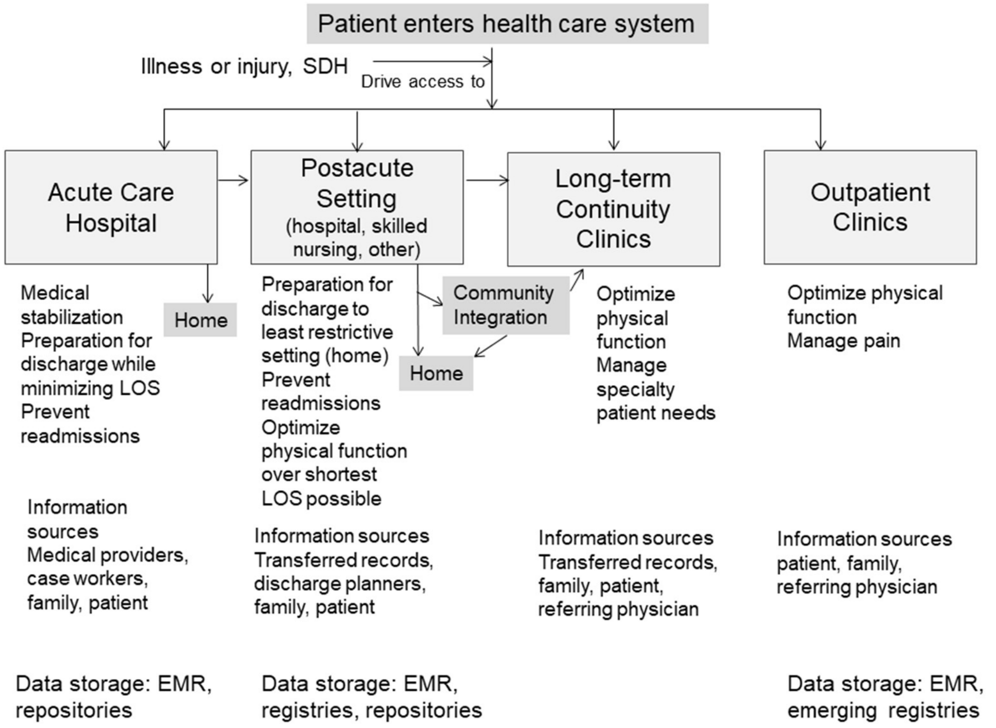

Patient enters health care system
Illness or injury, SDH
Drive access to
Acute Care Hospital
Postacute Setting (hospital, skilled nursing, other)
Long-term Continuity Clinics
Outpatient Clinics
Home
Community Integration
Home
Medical stabilization Preparation for discharge while minimizing LOS Prevent readmissions
Preparation for discharge to least restrictive setting (home) Prevent readmissions Optimize physical function over shortest LOS possible
Optimize physical function Manage specialty patient needs
Optimize physical function Manage pain
Information sources Medical providers, case workers, family, patient
Information sources Transferred records, discharge planners, family, patient
Information sources Transferred records, family, patient, referring physician
Information sources patient, family, referring physician
Data storage: EMR, repositories
Data storage: EMR, registries, repositories
Data storage: EMR, emerging registries

FIGURE 1 | Overview of possible patient pathways through the spectrum of physiatric care and variation of medical data sources and SDH captured and stored across rehabilitation settings.

Social determinants of health (SDH) are various social and economic factors, including education, healthcare access, and community support, which can impact health status and health outcomes (1). SDH guide resource allocation, discharge planning, access to outpatient rehabilitation services and other therapeutic interventions and progress assessment. However, despite their potential impact, SDH data are notoriously poorly collected and coded in the electronic health record (EHR) (2) which makes assessing their impact *post-hoc* challenging. Furthermore, specific SDH can contribute to disparities in outcomes among patient subgroups (3, 4). For example, SDH, including educational attainment, housing and living environment, and social support, influence rehabilitation outcomes with various post-acute conditions such as stroke (5–14), spinal cord injury (SCI) (15, 16), traumatic brain injury (TBI) (17–21), amputation (22–24), and chronic conditions such as osteoarthritis (25), chronic pain (26), and cardiopulmonary disease (27).

Due to wide availability of therapies and interventions, it is challenging for physiatrists to determine which patient subgroups will achieve the best outcomes. This challenge may be met through exploration of big data and artificial intelligence techniques. Presently, machine learning and artificial intelligence are not commonly used in this field to predict outcomes, but should be. We propose a critical reappraisal of data collection methods and development of a "biopsychosocial model" (28) that includes SDH and physical functional measures. In this opinion and perspective paper, we present: (1) SDH driving functional outcomes; (2) available big datasets relevant to physiatry and possible artificial intelligence application; and (3) new measurement and analysis methods that could improve care pathway mapping and functional outcomes in physiatry. The search terms "social determinants of health," "big data," "electronic health record," "physiatry," "rehabilitation," "physical medicine and rehabilitation" were used to identify relevant articles discussed herein. All relevant articles were reviewed and representative articles that included the main patient populations treated in physiatry are presented next.

SOCIAL DETERMINANTS OF HEALTH ON PHYSIATRY CARE PATHWAYS

SDH are vital to collaborative short and long-term goal setting with the patient and family, with establishing home safety parameters, setting expectations for rate and type of functional gains, and reintegration into social-vocational roles. In the outpatient setting, SDH affects symptom progression, mental health, social functioning and access to the amount or type of services obtained for a given diagnosis (29). Commonly measured SDH each care setting are summarized in **Supplementary Table 1**.

Acute Care Setting

In acute care, SDH are reviewed that could impact referral decisions and admissions into post-acute care. The decision to refer is described as "subjective" (30), yet referral of patients to the appropriate level of care ensures equitable access (31). Limiting or delaying access to services after severe injuries such as stroke or TBI can worsen functional disability and related outcomes^{27} and contributes to health disparities. For many conditions, early intensive rehabilitation can optimize functionality and re-engage patients back into life. SDH that affect referral to post-acute services include gender, race (29, 32), age, payor source (32), place of living (community alone, community with others, nursing home) (30), social support or living status, and geographic region (23). For medically-complex conditions, such as dysvascular amputation, inpatient rehabilitation referrals occur more often when the patient is married, has Medicaid and lives in a city; older, unmarried patients with history of nursing home residence are more often referred to skilled nursing facilities (SNF) (23). Patients with knee or hip arthroplasty may enter the rehabilitation pathway in post-acute care or outpatient settings depending on SDH, including age, gender and availability of caregiver at home. Younger patients and those with more family support are commonly referred to less intensive care settings (33). Among patients with hip fracture or joint replacement, SNF placement was more common in those with no insurance, Medicaid, and those who were Hispanic or black. SNFs are associated with less rigorous rehabilitation compared to an inpatient rehabilitation hospital (34, 35). Thus, a key transition at which functional outcomes is impacted is discharge to the next setting.

Post-acute Care Setting

The post-acute care setting shapes functional and clinical outcomes by rehabilitation prescription (type and volume of therapies). Inpatient rehabilitation hospitals are required to provide physician management at least 3 days per week, 24 h nursing care and at least 3 h of intensive rehabilitation therapy five times a week. Differences exist in the delivery of occupational, physical and speech-language therapy among post-acute settings for treatment of the same diagnosis (36). Gains in mobility and self-care are frequently better after inpatient rehabilitation compared to SNF (36, 37). Unfavorable outcomes in post-acute settings include long rehabilitation hospital stays, slow trajectory to achieve functional milestones (mobility, various activities), small functional gains, discharge to long-term care and acute care readmission. In general, worse outcomes occur with advanced age (15, 38–40), non-white race (19, 41), insurance type (42), less family support or living alone (23, 32, 40). Older patients are less able to engage in intensive rehabilitation therapies for SCI or hip fracture (3, 15). Some SDH, such as gender, have differential effects on rehabilitation outcomes. Specifically, female gender is associated with higher odds of discharge to home (43) and better supervision-level only status for more functional activities than men after stroke by discharge (10, 43, 44), but females demonstrate lower efficiency of functional improvement during rehabilitation than males after knee arthroplasty (45).

Readmission to acute care is differentially affected by SDH in different settings. For patients with knee arthroplasty receiving care in an inpatient rehabilitation hospital, advanced age and non-white race increased the odds for 90-day readmission (35). However, age, gender, race, marital status and living arrangement did not predict hospital readmissions for patients in a SNF, but medical conditions such as congestive heart failure did (46). Other evidence shows that patients with SCI are more likely to be readmitted multiple times if unemployed, female, have Medicaid (16, 47) or if rehabilitation was provided in a SNF (48). SDH in the context of the diagnoses and rehabilitation exposure will be important in future analytic methods for outcome prediction.

Reintegration Back Home

Successful community reengagement includes social, leisure, instrumental, vocational, school or volunteer participation. For some diagnoses like stroke, reengagement in community activities and self-care is best predicted by a supportive living situation (49, 50). In patients with TBI, community reintegration is complex, and strength of associations between SDH and outcomes vary widely. Scoping reviews found that white race, higher education, employment, level of disability and mood/affect contribute to reintegration (51). Conversely, poor housing is a risk factor for moderate-to-severe disability after hospital discharge for stroke (52). SDH are critical in the success of personal and societal engagement over the long-term.

Outpatient Setting

Common musculoskeletal conditions, such as arthritis and chronic back pain, disproportionately affect people who are non-white (black, Hispanic), older, have less than a high school level of education, low annual income, single, unemployed, and/or living in inner cities or rural areas (53–55). Job positions requiring more craft skills than managerial-professional skills are strongly related to back pain (56). Prospective evidence shows that pain symptom severity and disability are worse over time among non-white, less-educated individuals (26, 56, 57) and those with less social support (24). Neighborhood location and resources may influence effectiveness of long-term care for people in different geographical areas. For example, people with knee osteoarthritis who live in safe areas with better social cohesion and have resources for participation in physical activity have better mental health (25), which may improve health outcomes overall. In a mixed sample of individuals with stroke, cardiopulmonary disease and arthritis, social identification (social group membership in the community) fostered feelings of self-efficacy and confidence, which reduced disability (27). Our understanding of SDH effects on functional outcomes across all settings could be improved with the study of additional determinants related to rehabilitation access, quality and effectiveness. Additional determinants required to fully understand functional outcome trajectories are in **Supplementary Table 1**.

LEVERAGING BIG DATA AND EXPANDING MACHINE LEARNING IN PHYSIATRY

An exciting opportunity to improve prediction of functional improvement exists through the use of artificial intelligence. Based on existing evidence and state of the science, various machine learning algorithms already helped create predictive equations for standard functional measures after inpatient rehabilitation for stroke: Functional Independence Measure (FIM), 10-m walk test, 6-min walk test and Berg Balance Scale (58). Moreover, machine-learning modeling predicted 30-day hospital readmissions after discharge to post-acute care, using patient SDH and other characteristics (59).

Existing Datasets

Current datasets used in physiatry contain a mixture of institutional data obtained by EHR extraction. Specific registries and administrative datasets each have advantages and disadvantages, described **Supplementary Table 2**. Often, breadth, detail and consistency of data are sacrificed. Outcomes in PM&R are focused on functional outcomes rather than survival, and tracking and recording these data remains a major challenge to expanding datasets.

Many physiatry-specific datasets are focused on specific conditions, such as stroke or osteoarthritis, and contain limited SDH data (**Supplementary Table 3**). One of the more generalized datasets is the Uniform Data System for Medical Rehabilitation which has existed for almost 30 years and is used by approximately 70% of inpatient rehabilitation facilities in the U.S. and contains FIM data before, during and after completed rehabilitation (60). Similarly, the Model Systems for Burn, TBI and SCI have been in use over 20 years, and gather social, psychologic, functional data and patient outcomes (61). More recently, datasets are being developed which examine patient-reported outcomes for benchmarking Medicare payments. These include the American Academy of Physical Medicine and Rehabilitation registries (for low back pain, ischemic stroke), and the American Spine Registry created by the American Association of Neurologic Surgeons and American Academy of Orthopedic Surgeons (62, 63). SDH tend to be limited to age, gender, race/ethnicity, insurance type, housing situation and discharge location. This highlights the need to expand data collection to create better predictive models. Non-specific datasets (**Supplementary Table 3**) typically contain the "easy-to-collect" SDH like age, gender, race/ethnicity, insurance type, living situation (housing type, people in household), discharge location and readmissions. Functional status is often assessed by proxy for where the patient was discharged, and readmission to a hospital or another rehabilitation facility (64). Unfortunately, the physical/occupational therapy or rehabilitation type received, and functional performance are generally not present, as seen in the **Supplementary Tables 2, 3**.

Extraction of SDH, Rehabilitation Components and Key Words

Often, research does not present the rehabilitation elements or different proportions of time spent in specific activities like gait retraining, patient education or activities of daily living (65). The use of large datasets with detailed information about therapeutic activities and outcomes including SDH, functional assessment scores, and patient-reported outcome measures could improve treatment precision and optimize patient success. Natural language processing (NLP), language modeling and word embedding techniques could be used on provider notes to find items from patient interactions or audio files that are related to SDH and functionality (66). For example, NLP can be used to identify which patients are more likely to miss therapy, or functional recovery time could be predicted for resource allocation and treatment planning (67), as well as identify SDH impact on functional progress among physiatric patients.

Non-linear Modeling of Functional Change

Functional recovery in physiatry is rarely a linear process. Patients initiating care at lower functional levels receive more treatment, and more treatment is associated with longer recovery, likely because treatment was resourced according to need (68). This can be addressed by using non-linear modeling using supervised techniques such as non-linear regression, decision trees, non-linear support vector machines and unsupervised techniques like clustering and artificial neural networks (69). For example, non-linear modeling was applied to create a non-linear risk score for stroke which performed better than the Framingham Stroke Risk Score, and we postulate that this approach would also be successful in predicting functional outcomes following stroke or other diseases in the early or later recovery stages (70, 71). Furthermore, the effects of SDH on functional outcomes in physiatry is unlikely to be linear and their inclusion could have protective effects against health plan underpayment for treatment in high-risk vulnerable populations (72). In our view, non-linear modeling methods would help the field better establish which SDH impact which aspects of functionality during each stage of rehabilitation from acute to long-term. These techniques could immediately and positively change how treatment is applied to different patient diagnoses depending on the acuity of the condition. **Figure 2** provides a summary of these novel techniques.

PROPOSED NOVEL MEASUREMENT APPROACHES 194

Several challenges exist with interpretation of functional outcomes in physiatry. First, the level of functional impairment dictates the type and amount physiatry services provided and the long-term outcomes independent from SDH. Second, the health status (defined as comorbid health conditions, personal, social and environmental factors) preceding hospital admission or outpatient visit impacts rehabilitation outcomes. Physical function and mobility are embedded in many health measures, from post-acute care and surgical outcomes, to chronic frailty and disability; these are represented as a domain of human activity in the International Classification of Functioning, Disability, and Health (73). Yet, mobility and other functional activities remain under-studied, and commonly-used medical terminologies do not reflect functional status in the EHR. Health status impacts FIM scores, is linked to SDH and can be used for clinical decision-making or predicting functional outcomes. For example, gender-related differences exist in the health status factors that result in worse functional status after TBI for men (dementia, epilepsy, chronic cardiovascular pathology, mental health disorders) (74). Third, changes in SDH over time are

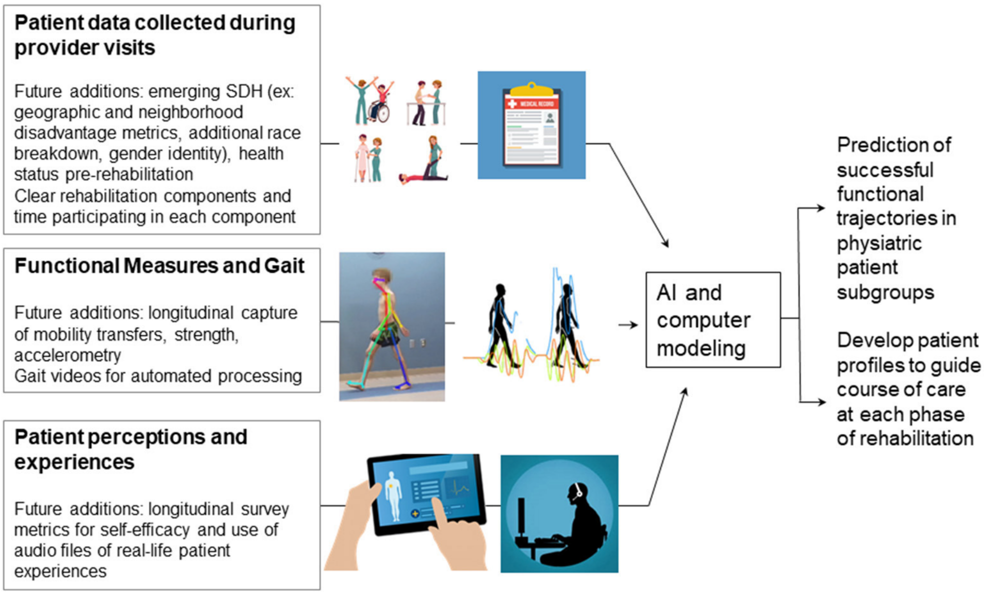

Patient data collected during provider visits
Future additions: emerging SDH (ex: geographic and neighborhood disadvantage metrics, additional race breakdown, gender identity), health status pre-rehabilitation
Clear rehabilitation components and time participating in each component
Functional Measures and Gait
Future additions: longitudinal capture of mobility transfers, strength, accelerometry
Gait videos for automated processing
Patient perceptions and experiences
Future additions: longitudinal survey metrics for self-efficacy and use of audio files of real-life patient experiences
AI and computer modeling
Prediction of successful functional trajectories in physiatric patient subgroups
Develop patient profiles to guide course of care at each phase of rehabilitation

FIGURE 2 | Proposed functional and performance metrics and processes to improve prediction of functional trajectories in patients with physiatric diagnoses. SDH, Social determinants of health; AI, Artificial intelligence.

rarely accounted for in physiatry research, a critical flaw that has been a barrier to understanding changes in function with different treatment approaches. Thus, linking SDH longitudinally to patient data and function is the next essential step in advancing treatment precision.

Approach 1. Establishing Functional Level and Health Status Prior to Disease

Physiatrists should have data regarding the patient's general health status and functional level prior to disease onset and be able to use these data to predict the extent of the patient's potential for recovery. These data would ideally be obtained prior to the disease, possibly at prior primary care physician visits or collected data from wearable technology such as FitBit®. Less optimal methods would be surveying the patient and/or their family and friends regarding their estimation of patient functional capacity.

Approach 2. Longitudinal Capture of SDH and Physical Function

The level of function at the start of rehabilitation coupled with health status and SDH, shape the trajectory and time-scale of recovery (58, 68). Supportive evidence includes widening disparities in FIM scores after stroke among white, black and Hispanic patients from rehabilitation to 12 months-post discharge; these different recovery patterns are strongly influenced by age (75). Also, there is population shifting among subgroups of patients undergoing physiatric treatment. Compared to years prior, individuals with non-vascular lower limb amputation today are more cognitively intact, but less physically functional and less able to afford prostheses—all of which can impact functional and clinical outcomes independent of other treatments provided (76). Longitudinal capture of SDH and physical function metrics will dramatically improve interpretation of treatment efficacy, disability fluctuation patterns, hospital readmissions, morbidity risk and mortality over time.

Approach 3. Capturing Movement and Gait Patterns in the EHR

Daily activity metrics that reflect community ambulation and physical activity patterns could be clinically useful to determine real-life functioning in the home and community (77). These metrics could include distance walked, daily step count and intensity of the steps taken; higher intensities are related to lower risk for major mobility disability (78) and predict independent living (79). Commercially-available triaxial accelerometers that produce raw acceleration output (Actigraph, Axivity, GeneActiv) can be used to determine average acceleration, intensity gradient or acceleration above which most active 30 min are captured. These raw data could be uploaded into the EHR on personal medical portals at specific follow-up intervals from the home or clinic.

Movement patterns produced during execution of functional tests provide insight on neuromotor strategies across a diverse range of patients. Gait metrics could be quickly extracted from 2D trajectories of body poses using single camera videos from the sagittal view (computer models available and freely shared) (80). Clinically-meaningful metrics could include gait speed, cadence, gait deviation index and knee flexion. Collection of gait metrics over time as part of routine care, coupled with SDH and clinical measures, would provide a complete picture of the patient experience and success with treatment. For example, lower gait speed was previously associated with age, literacy, and blue collar occupation (81).

Approach 4. Perceived Functional Outcomes and Self Efficacy

Inclusion of measures of perception of physical function and self-efficacy would inform how much functional limitation is modified by thoughts and feelings. Higher self-efficacy (82) directly relates to better community reintegration (83), functionality (24, 82), and independence in conditions such as amputation and osteoarthritis. Patient-perceived function and self-efficacy could be measured through traditional methods such as survey. We propose a new approach of capturing patient experiences through audio recording analysis. We envision a patient portal (accessible through phone or computer) in which patients could record changes in symptoms, pain and functional ability at specific time points after initiation of treatment or follow-up using standardized prompt questions from validated surveys or a diagnosis-specific question set. These audio files could be uploaded as part of the EHR. Additional free talking could supplement standardized responses and the language analyzed for key words that represent changes in well-being that may not otherwise be captured in EHR. These could include state of emotional well-being (such as, "feeling depressed," "sad"), SDH (including "lost my job," "retired," "moved to new area," "taking care of my husband," "got married") and physical function (examples could include "my knee pain is worse," can't drive anymore'). These methods could improve understanding of functional fluctuations over time in different patient subgroups.

MOVING FORWARD

As we move toward precision medicine, physiatry continues to face unique challenges such as insufficient datasets, difficulty with data access-sharing and lack of SDH and functional outcomes. Physiatry is uniquely positioned to: (1) implement new forms of data collection and integration such as movement and gait patterning, and (2) improve collection of SDH and patient-reported outcomes focusing on function. From a health system-wide perspective, we advocate for a consistent and standardized collection of SDH, health status and functional measures over time for diagnoses commonly treated in physiatry. Sources could include patient EHR, surveys, claims data, smart phone applications and wearable devices. Unique sources of data could include subcategories of race, "area deprivation scores" from the Neighborhood Atlas (84), and census tract data. Using artificial intelligence with the sources proposed here could help establish optimal treatment pathways for different patient subgroups, which in turn could improve preparation at each phase of rehabilitation care and treatment precision.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2021.738253/full#supplementary-material

REFERENCES

1. Centers for Disease Control. *Social Determinants of Health: Know What Affects Health* (2021).
2. Guo Y, Chen Z, Xu K, George TJ, Wu Y, Hogan W, et al. International classification of diseases, tenth revision, clinical modification social determinants of health codes are poorly used in electronic health records. *Medicine.* (2020) 99:e23818. doi: 10.1097/MD.0000000000023818
3. Siebens HC, Sharkey P, Aronow HU, Deutscher D, Roberts P, Munin MC, et al. Variation in rehabilitation treatment patterns for hip fracture treated with arthroplasty. *PMR.* (2016) 8:191–207. doi: 10.1016/j.pmrj.2015.07.005
4. Simões JL, Soares S, Sa-Couto P, Lopes C, Magina D, Melo E, et al. The influence of presurgical factors on the rehabilitation outcome of patients following hip arthroplasty. *Rehabil Nurs.* (2019) 44:189–202. doi: 10.1097/rnj.0000000000000126
5. Pournajaf S, Goffredo M, Agosti M, Massucci M, Ferro S, Franceschini M, et al. Community ambulation of stroke survivors at 6 months follow-up: an observational study on sociodemographic and sub-acute clinical indicators. *Eur J Phys Rehabil Med.* (2019) 55:433–41. doi: 10.23736/S1973-9087.18.05489-8
6. Sandel ME, Wang H, Terdiman J, Hoffman JM, Ciol MA, Sidney S, et al. Disparities in stroke rehabilitation: results of a study in an integrated health system in northern California. *PMR.* (2009) 1:29–40. doi: 10.1016/j.pmrj.2008.10.012
7. Tooth L, McKenna K, Goh K, Varghese P. Length of stay, discharge destination, and functional improvement: utility of the Australian national subacute and nonacute patient casemix classification. *Stroke.* (2005) 36:1519–25. doi: 10.1161/01.STR.0000169923.57038.a8
8. Meijer R, van Limbeek J, Kriek B, Ihnenfeldt D, Vermeulen M, de Haan R. Prognostic social factors in the subacute phase after a stroke for the discharge destination from the hospital stroke-unit. A systematic review of the literature. *Disabil Rehabil.* (2004) 26:191–7. doi: 10.1080/09638280310001636437
9. Dhand A, Lang CE, Luke DA, Kim A, Li K, McCafferty L, et al. Social network mapping and functional recovery within 6 months of ischemic stroke. *Neurorehabil Neural Repair.* (2019) 33:922–32. doi: 10.1177/1545968319872994
10. Hay CC, Graham JE, Pappadis MR, Sander AM, Hong I, Reistetter TA. The impact of one's sex and social living situation on rehabilitation outcomes after a stroke. *Am J Phys Med Rehabil.* (2020) 99:48–55. doi: 10.1097/PHM.0000000000001276
11. Kobylańska M, Kowalska J, Neustein J, Mazurek J, Wójcik B, Bełza M, et al. The role of biopsychosocial factors in the rehabilitation process of individuals with a stroke. *Work.* (2018) 61:523–35. doi: 10.3233/WOR-162823
12. Ouyang F, Wang Y, Huang W, Chen Y, Zhao Y, Dang G, et al. Association between socioeconomic status and post-stroke functional outcome in deprived rural southern China: a population-based study. *BMC Neurol.* (2018) 18:12. doi: 10.1186/s12883-018-1017-4
13. Delbari A, Keyghobadi F, Momtaz YA, Keyghobadi F, Akbari R, Kamranian H, et al. Sex differences in stroke: a socioeconomic perspective. *Clin Interv Aging.* (2016) 11:1207–12. doi: 10.2147/CIA.S113302
14. Everink IHJ, van Haastregt JCM, van Hoof SJM, Schols JMGA, Kempen GIJM. Factors influencing home discharge after inpatient

rehabilitation of older patients: a systematic review. *BMC Geriatr.* (2016) 16:5. doi: 10.1186/s12877-016-0187-4

15. Hsieh CH, DeJong G, Groah S, Ballard PH, Horn SD, Tian W. Comparing rehabilitation services and outcomes between older and younger people with spinal cord injury. *Arch Phys Med Rehabil.* (2013) 94(Suppl. 4):S175–86. doi: 10.1016/j.apmr.2012.10.038
16. DeJong G, Tian W, Hsieh CH, Junn C, Karam C, Ballard PH, et al. Rehospitalization in the first year of traumatic spinal cord injury after discharge from medical rehabilitation. *Arch Phys Med Rehabil.* (2013) 94(Suppl. 4):S87–97. doi: 10.1016/j.apmr.2012.10.037
17. Oyesanya TO, Moran TP, Espinoza TR, Wright DW. Regional variations in rehabilitation outcomes of adult patients with traumatic brain injury: a uniform data system for medical rehabilitation investigation. *Arch Phys Med Rehabil.* (2021) 102:68–75. doi: 10.1016/j.apmr.2020.07.011
18. Dahdah MN, Barnes S, Buros A, Dubiel R, Dunklin C, Callender L, et al. Variations in inpatient rehabilitation functional outcomes across centers in the traumatic brain injury model systems study and the influence of demographics and injury severity on patient outcomes. *Arch Phys Med Rehabil.* (2016) 97:1821–31. doi: 10.1016/j.apmr.2016.05.005
19. Graham JE, Radice-Neumann DM, Reistetter TA, Hammond FM, Dijkers M, Granger CV. Influence of sex and age on inpatient rehabilitation outcomes among older adults with traumatic brain injury. *Arch Phys Med Rehabil.* (2010) 91:43–50. doi: 10.1016/j.apmr.2009.09.017
20. Arango-Lasprilla JC, Rosenthal M, Deluca J, Cifu DX, Hanks R, Komaroff E. Functional outcomes from inpatient rehabilitation after traumatic brain injury: how do Hispanics fare? *Arch Phys Med Rehabil.* (2007) 88:11–8. doi: 10.1016/j.apmr.2006.10.029
21. Arango-Lasprilla JC, Ketchum JM, Cifu D, Hammond F, Castillo C, Nicholls E, et al. Predictors of extended rehabilitation length of stay after traumatic brain injury. *Arch Phys Med Rehabil.* (2010) 91:1495–504. doi: 10.1016/j.apmr.2010.07.010
22. Venkataraman K, Fong NP, Chan KM, Tan BY, Menon E, Ee CH, et al. Rehabilitation outcomes after inpatient rehabilitation for lower extremity amputations in patients with diabetes. *Arch Phys Med Rehabil.* (2016) 97:1473–80. doi: 10.1016/j.apmr.2016.04.009
23. Dillingham TR, Yacub JN, Pezzin LE. Determinants of postacute care discharge destination after dysvascular lower limb amputation. *PMR.* (2011) 3:336–44. doi: 10.1016/j.pmrj.2010.12.019
24. Miller MJ, Cook PF, Magnusson DM, Morris MA, Blatchford PJ, Schenkman ML, et al. Self-efficacy and social support are associated with disability for ambulatory prosthesis users after lower-limb amputation. *PMR.* (2021) 13:453–60. doi: 10.1002/pmrj.12464
25. Kowitt SD, Aiello AE, Callahan LF, Fisher EB, Gottfredson NC, Jordan JM, et al. How are neighborhood characteristics associated with mental and physical functioning among older adults with radiographic knee osteoarthritis? *Arthritis Care Res (Hoboken).* (2021) 73:308–17. doi: 10.1002/acr.24125
26. Fliesser M, De Witt Huberts J, Wippert PM. The choice that matters: the relative influence of socioeconomic status indicators on chronic back pain- a longitudinal study. *BMC Health Serv Res.* (2017) 17:800. doi: 10.1186/s12913-017-2735-9
27. Cameron JE, Voth J, Jaglal SB, Guilcher SJT, Hawker G, Salbach NM. “In this together”: Social identification predicts health outcomes (via self-efficacy) in a chronic disease self-management program. *Soc Sci Med.* (2018) 208:172–9. doi: 10.1016/j.socscimed.2018.03.007
28. Pincus T, Castrejon I. Low socioeconomic status and patient questionnaires in osteoarthritis: challenges to a “biomedical model” and value of a complementary “biopsychosocial model” *Clin Exp Rheumatol.* (2019) 120:18–23.
29. Odonkor CA, Esparza R, Flores LE, Verduzco-Gutierrez M, Escalon MX, Solinsky R, et al. Disparities in health care for black patients in physical medicine and rehabilitation in the United States: a narrative review. *PMR.* (2021) 13:180–203. doi: 10.1002/pmrj.12509
30. Longley V, Peters S, Swarbrick C, Bowen A. What factors affect clinical decision-making about access to stroke rehabilitation? A systematic review. *Clin Rehabil.* (2019) 33:304–16. doi: 10.1177/0269215518808000
31. Labberton AS, Barra M, Rønning OM, Thommessen B, Churilov L, Cadilhac DA, et al. Patient and service factors associated with referral and admission to inpatient rehabilitation after the acute phase of stroke in Australia and Norway. *BMC Health Serv Res.* (2019) 19:871. doi: 10.1186/s12913-019-4713-x
32. Ottenbacher KJ, Smith PM, Illig SB, Linn RT, Gonzales VA, Ostir GV, et al. Disparity in health services and outcomes for persons with hip fracture and lower extremity joint replacement. *Med Care.* (2003) 41:232–41. doi: 10.1097/01.MLR.0000044902.01597.54
33. Benz T, Angst F, Oesch P, Hilfiker R, Lehmann S, Mueller Mebes C, et al. Comparison of patients in three different rehabilitation settings after knee or hip arthroplasty: a natural observational, prospective study. *BMC Musculoskelet Disord.* (2015) 16:317. doi: 10.1186/s12891-015-0780-2
34. Freburger JK, Holmes GM, Ku LJE. Postacute rehabilitation care for hip fracture: who gets the most care? *J Am Geriatr Soc.* (2012) 60:1929–35. doi: 10.1111/j.1532-5415.2012.04149.x
35. Singh JA, Kallan MJ, Chen Y, Parks ML, Ibrahim SA. Association of race/ethnicity with hospital discharge disposition after elective total knee arthroplasty. *JAMA Netw Open.* (2019) 2:e1914259. doi: 10.1001/jamanetworkopen.2019.14259
36. Vincent HK, Vincent KR. Functional and economic outcomes of cardiopulmonary patients: a preliminary comparison of the inpatient rehabilitation and skilled nursing facility environments. *Am J Phys Med Rehabil.* (2008) 87:371–80. doi: 10.1097/PHM.0b013e31816dd251
37. Hong I, Goodwin JS, Reistetter TA, Kuo YF, Mallinson T, Karmarkar A, et al. Comparison of functional status improvements among patients with stroke receiving postacute care in inpatient rehabilitation vs skilled nursing facilities. *JAMA Netw Open.* (2019) 2:e1916646. doi: 10.1001/jamanetworkopen.2019.16646
38. Frankel RM, Stein T. Getting the most out of the clinical encounter: the four habits model. *J Med Pract Manage.* (2001) 16:184–91.
39. Cifu DX, Kreutzer JS, Marwitz JH, Rosenthal M, Englander J, High W. Functional outcomes of older adults with traumatic brain injury: a prospective, multicenter analysis. *Arch Phys Med Rehabil.* (1996) 77:883–8. doi: 10.1016/S0003-9993(96)90274-9
40. Jourdan C, Bayen E, Darnoux E, Ghout I, Azerad S, Ruet A, et al. Patterns of post-acute health care utilization after a severe traumatic brain injury: Results from the PariS-TBI cohort. *Brain Inj.* (2015) 29:701–8. doi: 10.3109/02699052.2015.1004646
41. Garcia JJ, Warren KL. Race/ethnicity matters: differences in poststroke inpatient rehabilitation outcomes. *Ethn Dis.* (2019) 29:599–608. doi: 10.18865/ed.29.4.599
42. Zhang X, Qiu H, Liu S, Li J, Zhou M. Prediction of prolonged length of stay for stroke patients on admission for inpatient rehabilitation based on the International Classification of Functioning, Disability, and Health (ICF) generic set: a study from 50 centers in China. *Med Sci Monit.* (2020) 26:e918811. doi: 10.12659/MSM.918811
43. Cations M, Lang C, Crotty M, Wesselingh S, Whitehead C, Inacio MC. Factors associated with success in transition care services among older people in Australia. *BMC Geriatr.* (2020) 20:496. doi: 10.1186/s12877-020-01914-z
44. Scrutinio D, Battista P, Guida P, Lanzillo B, Tortelli R. Sex differences in long-term mortality and functional outcome after rehabilitation in patients with severe stroke. *Front Neurol.* (2020) 11:84. doi: 10.3389/fneur.2020.00084
45. Vincent HK, Alfano AP, Lee L, Vincent KR. Sex and age effects on outcomes of total hip arthroplasty after inpatient rehabilitation. *Arch Phys Med Rehabil.* (2006) 87:461–7. doi: 10.1016/j.apmr.2006.01.002
46. Flanagan NM, Rizzo VM, James GD, Spegman A, Barnawi NA. Predicting risk factors for 30-day readmissions following discharge from post-acute care. *Prof Case Manag.* (2018) 23:139–46. doi: 10.1097/NCM.0000000000000261
47. Canori A, Kumar A, Hiremath SV. Factors associated with multiple hospital readmissions for individuals with spinal cord injury. *Commonhealth (Phila).* (2020) 1:57–61. doi: 10.15367/ch.v1i2.399
48. Cardenas DD, Hoffman JM, Kirshblum S, McKinley W. Etiology and incidence of rehospitalization after traumatic spinal cord

injury: a multicenter analysis. *Arch Phys Med Rehabil.* (2004) 85:1757–63. doi: 10.1016/j.apmr.2004.03.016

49. Erler KS, Sullivan V, Mckinnon S, Inzana R. Social support as a predictor of community participation after stroke. *Front Neurol.* (2019) 10:1013. doi: 10.3389/fneur.2019.01013
50. Elloker T, Rhoda AJ. The relationship between social support and participation in stroke: a systematic review. *Afr J Disabil.* (2018) 7:357. doi: 10.4102/ajod.v7i0.357
51. Kersey J, Terhorst L, Wu CY, Skidmore E. A scoping review of predictors of community integration following traumatic brain injury: a search for meaningful associations. *J Head Trauma Rehabil.* (2019) 34:E32–41. doi: 10.1097/HTR.0000000000000442
52. de Villiers L, Badri M, Ferreira M, Bryer A. Stroke outcomes in a socio-economically disadvantaged urban community. *S Afr Med J.* (2011) 101:345–8. doi: 10.7196/SAMJ.4588
53. Guillemin F, Carruthers E, Li LC. Determinants of MSK health and disability–social determinants of inequities in MSK health. *Best Pract Res Clin Rheumatol.* (2014) 28:411–33. doi: 10.1016/j.berh.2014.08.001
54. Vennu V, Abdulrahman TA, Alenazi AM, Bindawas SM. Associations between social determinants and the presence of chronic diseases: data from the osteoarthritis Initiative. *BMC Public Health.* (2020) 20:1323. doi: 10.1186/s12889-020-09451-5
55. Jonsdottir S, Ahmed H, Tómasson K, Carter B. Factors associated with chronic and acute back pain in wales, a cross-sectional study. *BMC Musculoskelet Disord.* (2019) 20:215. doi: 10.1186/s12891-019-2477-4
56. Fliesser M, De Witt Huberts J, Wippert PM. Education, job position, income or multidimensional indices? Associations between different socioeconomic status indicators and chronic low back pain in a German sample: a longitudinal field study. *BMJ Open.* (2018) 8:e020207. doi: 10.1136/bmjopen-2017-020207
57. Allen KD, Helmick CG, Schwartz TA, DeVellis RF, Renner JB, Jordan JM. Racial differences in self-reported pain and function among individuals with radiographic hip and knee osteoarthritis: the Johnston County Osteoarthritis Project. *Osteoarthritis Cartilage.* (2009) 17:1132–6. doi: 10.1016/j.joca.2009.03.003
58. Harari Y, O'Brien MK, Lieber RL, Jayaraman A. Inpatient stroke rehabilitation: prediction of clinical outcomes using a machine-learning approach. *J Neuroeng Rehabil.* (2020) 17:71. doi: 10.1186/s12984-020-00704-3
59. Howard EP, Morris JN, Schachter E, Schwarzkopf R, Shepard N, Buchanan ER. Machine-learning modeling to predict hospital readmission following discharge to post-acute care. *J Am Med Dir Assoc.* (2021) 22:1067–72.e29. doi: 10.1016/j.jamda.2020.12.017
60. Graham JE, Granger CV, Karmarkar AM, Deutsch A, Niewczyk P, Divita MA, et al. The uniform data system for medical rehabilitation: report of follow-up information on patients discharged from inpatient rehabilitation programs in 2002-2010. *Am J Phys Med Rehabil.* (2014) 93:231–44. doi: 10.1097/PHM.0b013e3182a92c58
61. Model Systems Knowledge Translation Center. *Directry of Model Systems.* (2021). Available online at: https://msktc.org/sci/model-system-centers (accessed July 7, 2021).
62. American Academy of Physical Medicine and Rehabilitation. *Patient Reported Outcomes Module.* (2021). Available online at: https://www.aapmr.org/quality-practice/registry/patient-reported-outcomes (accessed July 5, 2021).
63. American Spine Registry. *The National Quality Improvement Registry for Spine Care.* (2021). Available online at: https://www.americanspineregistry.org/ (accessed July 1, 2021).
64. Meddings J, Reichert H, Smith SN, Iwashyna TJ, Langa KM, Hofer TP, et al. The impact of disability and social determinants of health on condition-specific readmissions beyond Medicare risk adjustments: a cohort study. *J Gen Intern Med.* (2017) 32:71–80. doi: 10.1007/s11606-016-3869-x
65. Schumacher R, Müri RM, Walder B. Integrated health care management of moderate to severe TBI in older patients-a narrative review. *Curr Neurol Neurosci Rep.* (2017) 17:92. doi: 10.1007/s11910-017-0801-7
66. Reeves RM, Christensen L, Brown JR, Conway M, Levis M, Gobbel GT, et al. Adaptation of an NLP system to a new healthcare environment to identify social determinants of health. *J Biomed Inform.* (2021) 120:103851. doi: 10.1016/j.jbi.2021.103851
67. Ong CJ, Orfanoudaki A, Zhang R, Caprasse FPM, Hutch M, Ma L, et al. Machine learning and natural language processing methods to identify ischemic stroke, acuity and location from radiology reports. *PLoS ONE.* (2020) 15:e0234908. doi: 10.1371/journal.pone.0234908
68. Hart T, Kozlowski AJ, Whyte J, Poulsen I, Kristensen K, Nordenbo A, et al. Functional recovery after severe traumatic brain injury: an individual growth curve approach. *Arch Phys Med Rehabil.* (2014) 95:2103–10. doi: 10.1016/j.apmr.2014.07.001
69. Chatterjee P, Cymberknop LJ, Armentano R. Nonlinear systems in healthcare towards intelligent disease prediction. In: *Nonlinear Systems- Theoretical Aspects and Recent Applications.* Intech Open (2019). doi: 10.5772/intechopen.88163
70. Orfanoudaki A, Chesley E, Cadisch C, Stein B, Nouh A, Alberts MJ, et al. Machine learning provides evidence that stroke risk is not linear: the non-linear Framingham stroke risk score. *PLoS ONE.* (2020) 15:e0232414. doi: 10.1371/journal.pone.0232414
71. Kwakkel G, Kollen BJ. Predicting activities after stroke: what is clinically relevant? *Int J Stroke.* (2013) 8:25–32. doi: 10.1111/j.1747-4949.2012.00967.x
72. Irvin JA, Kondrich AA, Ko M, Rajpurkar P, Haghgoo B, Landon BE, et al. Incorporating machine learning and social determinants of health indicators into prospective risk adjustment for health plan payments. *BMC Public Health.* (2020) 20:608. doi: 10.1186/s12889-020-08735-0
73. Newman-Griffis D, Fosler-Lussier E. Automated coding of under-studied medical concept domains: linking physical activity reports to the International Classification of Functioning, Disability, and Health. *Front Digit Health.* (2021) 3:620828. doi: 10.3389/fdgth.2021.620828
74. Chan V, Sutton M, Mollayeva T, Escobar MD, Hurst M, Colantonio A. Data mining to understand how health status preceding traumatic brain injury affects functional outcome: a population-based sex-stratified study. *Arch Phys Med Rehabil.* (2020) 101:1523–31. doi: 10.1016/j.apmr.2020.05.017
75. Simmonds KP, Luo Z, Reeves M. Race/ethnic and stroke subtype differences in poststroke functional recovery after acute rehabilitation. *Arch Phys Med Rehabil.* (2021) 102:1473–81. doi: 10.1016/j.apmr.2021.01.090
76. Batten H, Kuys S, McPhail S, Varghese P, Mandrusiak A. Are people with lower limb amputation changing? A seven-year analysis of patient characteristics at admission to inpatient rehabilitation and at discharge. *Disabil Rehabil.* (2019) 41:3203–9. doi: 10.1080/09638288.2018.1492033
77. Gothe NP, Bourbeau K. Associations between physical activity intensities and physical function in stroke survivors. *Am J Phys Med Rehabil.* (2020) 99:733–8. doi: 10.1097/PHM.0000000000001410
78. Fanning J, Rejeski WJ, Chen SH, Nicklas BJ, Walkup MP, Axtell RS, et al. A case for promoting movement medicine: preventing disability in the LIFE Randomized Controlled Trial. *J Gerontol A Biol Sci Med Sci.* (2019) 74:1821–7. doi: 10.1093/gerona/glz050
79. Dunlop DD, Song J, Hootman JM, Nevitt MC, Semanik PA, Lee J, et al. One hour a week: moving to prevent disability in adults with lower extremity joint symptoms. *Am J Prev Med.* (2019) 56:664–72. doi: 10.1016/j.amepre.2018.12.017
80. Kidziński Ł, Yang B, Hicks JL, Rajagopal A, Delp SL, Schwartz MH. Deep neural networks enable quantitative movement analysis using single-camera videos. *Nat Commun.* (2020) 11:4054. doi: 10.1038/s41467-020-17807-z
81. Busch T de A, Duarte YA, Pires Nunes D, Lebrão ML, Satya Naslavsky M, dos Santos Rodrigues A, et al. Factors associated with lower gait speed among the elderly living in a developing country: a cross-sectional population-based study. *BMC Geriatr.* (2015) 15:35. doi: 10.1186/s12877-015-0031-2
82. Jackson T, Xu T, Jia X. Arthritis self-efficacy beliefs and functioning among osteoarthritis and rheumatoid arthritis patients: a meta-analytic review. *Rheumatology.* (2020) 59:948–58. doi: 10.1093/rheumatology/kez219
83. Gupta S, Jaiswal A, Norman K, DePaul V. Heterogeneity and its impact on rehabilitation outcomes and interventions for community reintegration in people with spinal cord injuries: an integrative review. *Top Spinal Cord Inj Rehabil.* (2019) 25:164–85. doi: 10.1310/sci2502-164

84. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible - the neighborhood Atlas. *N Engl J Med*. (2018) 378:2456–8. doi: 10.1056/NEJMp1802313

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: 10 March 2022
doi: 10.3389/fpubh.2022.778002

The Impact of Energy Consumption Revolution on Farmers' Happiness: An Empirical Analysis From China

*Zhiyao Xu and Rong Ge**

Institute of Natural Resources and Environmental Audits, School of Government Audit, Nanjing Audit University, Nanjing, China

This study divided the impact of energy consumption revolution on farmers' happiness into direct and indirect effects. We empirically tested these effects using the Chinese General Social Survey (CGSS) household data in 2015 and the mediation-moderation model. The results showed that: (1) The rural energy consumption revolution has increased the probability of farmers' happiness level by 22.7%. The direct effects with obvious marginal decrement accounted for the main part (over 90%) of the total effect, but the multi-dimensional mediating mechanism was not yet robust. (2) The revolution of rural energy consumption has slightly improved farmers' happiness through the mediating role of increased leisure activities, while the negative impact of increased use-cost on the happiness of low-income farmers was nearly significant. (3) Regional economy, household income, and energy type played negative roles when moderating the above process. To low-income households in the less-developed western region, the total effects were more evident in the aspect of electricity use. Hence, several policy recommendations have been further made, including inclusive energy and strategic synergies.

Keywords: happiness, energy consumption, farmer, China, mediation-moderation analysis

OPEN ACCESS

Edited by:
Fei Wang,
Cornell University, United States

Reviewed by:
Shi Qinghua,
Shanghai Jiaotong University, China
Rongbing Huang,
Zhejiang Gongshang University, China
Li Xu,
Jiang Xi Normal University, China

****Correspondence:***
Rong Ge
ge7218@163.com

Specialty section:
This article was submitted to Digital Public Health, a section of the journal Frontiers in Public Health

Received: *16 September 2021*
Accepted: *11 January 2022*
Published: *10 March 2022*

Citation:
Xu Z and Ge R (2022) The Impact of Energy Consumption Revolution on Farmers' Happiness: An Empirical Analysis From China. Front. Public Health 10:778002. doi: 10.3389/fpubh.2022.778002

INTRODUCTION

In response to the Fourth Industrial Revolution, the Chinese government has established the "Energy Consumption Revolution (ECR)" as the essential strategy for China's energy development and formulated concrete plans (1). Specifically, the "ECR" refers to the transition from the traditional energy consumption with intensive emission to the modern energy consumption with low emission (2). However, the Revolution has encountered numerous obstacles in the rural areas of China, the root cause is that Chinese rural population accounts for a high proportion (over 40%) of the total population while sparsely populated throughout China. Whether the Revolution can be successfully implemented in rural China, it depends on not only technological innovations, but also the demand for new energy sources from rural residents. It is crucial to continuously increase rural residents' satisfaction and perception of happiness in the Revolution, so as to motivate their active participation (3).

There is a body of scientific literature on rural energy consumption. First, studies have investigated the Energy Revolution and current status of rural energy consumption. Prior research used energy ladder model, energy stacking model and energy wave model to depict the general course of energy transition (4–6). Some researches described how rural energy consumption evolved from biomass energy (e.g., straw and fuelwood) to the new electrical energy sources [e.g., (4, 7)]. Although the consumption of new energy such as electricity and gas is rapidly growing in China, traditional biomass energy consumption still accounts for more than 60% of total energy consumption in rural China (8, 9). In the rural areas of Beijing, Tianjin and Hebei provinces, fuelwood and coal consumption even accounts for more than 70% of total energy consumption (10). Second, the studies investigated the issue of rural energy supply in China showed, although the commercialization of rural energy is developing fast, due to high costs of energy use, limited technological innovation and financial investment, there are still many challenges in sustainably supplying energy in rural China (11). The urgent problems which need to be solved for rural Energy Reform include the severe pollution from cheap coal energy, the high price and low utilization of clean gas and electricity energy, and the high operating cost of biogas (12). Third, after studying the energy demand in rural China, it is found that household income and the price of energy are the two key factors influencing the demand of rural household energy consumption (13–15). As income level rises, the need for energy upgrade increases for those households (16, 17). It has been proposed that the primary goal of China's energy consumption revolution is to achieve "coal-free" and "fuelwood-free" in the rural areas (2).

Prior research has also explored how energy upgrade impacts the socioeconomic and physical environment, as well as how the socioeconomic and physical environment impacts perceived happiness in rural China. Energy upgrade can significantly reduce the time women spend on housework, on the other hand, increase the time they can spend on leisure activities, increase leisure time and allow rural residents to socialize, entertain, and rest, leading to increased happiness (18). Furthermore, the reduction of traditional energy consumption, such as fuelwood and coal, has greatly decreased the emission of air pollutants such as CO_2, SO_2, and NO_x, which improves the air quality in rural China sharply (15, 19). Improvements for living environment can significantly increase farmers' happiness (20). Thereinto, air quality has been proved playing a significant role when impacting, especially on rural residents' happiness (21). However, there were few studies focus on the relationship between energy consumption and happiness in rural China. Considering the recent energy consumption revolution, this study aimed to quantify the impact of energy consumption on farmers' happiness.

MECHANISM ANALYSIS

Based on the analysis framework of mediation and moderation effects in social psychology, we constructed a conceptual model of the mechanism for energy consumption revolution improving farmers' happiness, as shown in **Figure 1**. The effects of energy consumption revolution on farmers' happiness include direct effect, mediation effect and moderation effect. Thereinto, the mediation effect consists of three dimensions of economy, society and environment, while the moderation effect contains three dimensionalities: region, economic status of respondents and energy type.

First, the direct effect refers to that the energy consumption revolution directly improves farmers' happiness. On the one hand, using new energy brings rural residents a more comfortable and convenient life, directly increasing people's sense of contentment and happiness; On the other hand, the installation and use of new energy devices bring farmers a strong 'demonstration effect' and pride among neighborhood (22). Although the impact of material consumption on people's happiness is non-linear (23, 24), the marginal effect brought by energy consumption revolution is often positive, especially for rural residents in regions with relatively backward economy (25).

Secondly, the mediation effect refers to that the rural energy consumption revolution indirectly improves farmers' happiness from three dimensions: economy, society and environment. Thereinto, (1) economic dimension refers to that energy consumption revolution enables farmers to devote more time to work and obtain higher income, thus increasing their sense of contentment and happiness (26, 27); meanwhile, the energy

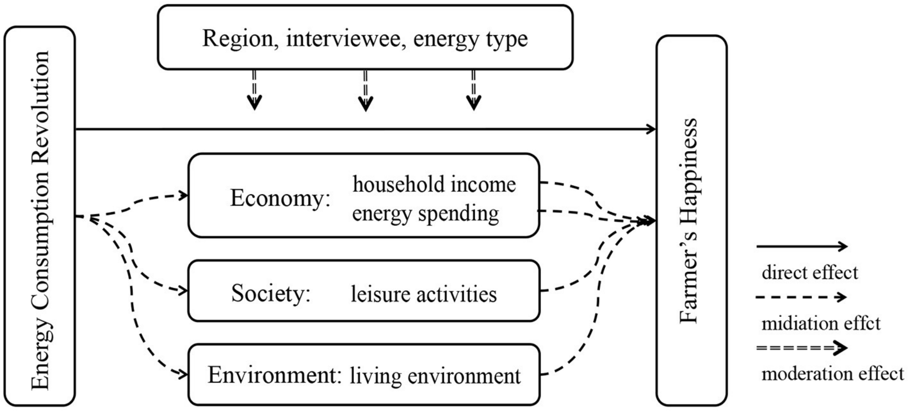


FIGURE 1 | The mechanism of energy consumption revolution effect on farmers' happiness.

consumption revolution may also have negative economic effects, such as resulting in higher energy use costs which will partly neutralize people's happiness. In this paper, household income and electricity cost were used to represent the mediators of these two aspects respectively. (2) The social dimension refers to the fact that the energy consumption revolution liberates farmers from complicated housework and enables them to enjoy more rest, leisure and entertainment time, thus gaining more sense of happiness (28). Here leisure time was used to proxy this mediator variable. (3) The environmental dimension means that the energy consumption revolution can improve the rural living environment and make farmers get more happiness (19). Here people's satisfaction with living environment is used as an intermediary.

Third, some key factors play moderating roles in the process of energy consumption revolution to improve farmers' happiness. (1) People's ability to accept new things differs from regional development levels, so there are certain differences in the mechanism of energy consumption revolution influencing happiness (20, 29). (2) People who are at different socioeconomic levels have significant structural differences in the sources of happiness, and they are various in the process of energy consumption revolution to improve happiness (30). (3) For different energy types, people show different acceptability, and the corresponding effects on people's happiness also vary.

Based on the above analysis, we put forward the following three hypotheses to be tested. (H1) Rural energy consumption revolution can directly improve farmers' happiness. (H2) Rural energy consumption revolution can indirectly improve farmers' happiness through economic, social and environmental dimensions. (H3) Region, people's socioeconomic level and energy type play regulatory roles in the process of consumption revolution to improve happiness.

MATERIALS AND METHODS

Data Source

We analyzed data from the Chinese General Social Survey (CGSS) which started in 2003. Every a few years, the CGSS randomly selects and surveys over 10,000 urban and rural households from all over China. Since 2015, the CGSS questionnaire has added an "Energy Module", which contains 115 questions related to energy use. In the 2015 CGSS, 10,967 households were interviewed in total. Therein, 3,653 households finished the Energy Module, while 1,472 of them were rural households. In this study, after removing households missing data on essential variables (e.g., happiness, electricity spending), we analyzed data from 1,320 rural households.

Model Variables

The mediation-moderation models were further applied based on the collected CGSS data. All the involved model variables were listed as follows:

Primary Outcome

The primary outcome variable was subjective happiness (*Happ*). The CGSS included the question "Do you think your life is happy?" with five response options: *very unhappy*, *relatively unhappy*, *not happy*, *relatively happy*, and *very happy*. We assigned values 1–5 to the responses, a higher value indicating the greater happiness.

Primary Predictor

The primary predictor was the response to the rural energy consumption revolution (*EneRef*). We defined *EneRef* as the consumption pattern shifting from traditional energy (e.g., fuelwood, straw, and coal) to modern clean energy (e.g., electricity, liquefied gas, natural gas, biogas, and solar energy) (2). If a household has completed the transition to modern energy in at least two of the three activities involving energy consumption (i.e., cooking, showering, and heating/cooling), the household is considered as having responded positively to the national call for energy consumption revolution ($EneRef_i = 1$). In contrast, if a household does not complete the transition as defined, then $EneRef_i = 0$ (9).

Mediators

The mediators included: (i) annual household income in natural logarithm (*Lginco*), (ii) monthly average electricity spending per person in natural logarithm (*Lgespp*), (iii) leisure activities (*Leis*), and (iv) satisfaction with living environment (*Envir*). Leisure activities were measured using the question "How often do you engage in leisure activities to rest or relax?", with response options being: *never*, *rarely*, *sometimes*, *often*, and *very often*. Satisfaction with living environment was measured using the question "Are you satisfied with the local government's performance on environmental protection?", with response options being: *very dissatisfied*, *dissatisfied*, *average*, *satisfied*, *and very satisfied*.

Moderators

The moderators included: (i) geographic region of the household (*Regid*): Eastern, Central, or Western China, (ii) perceived socioeconomic status (*Incox*), and (iii) Energy type (*Enetype*): fuelwood/coal, electricity, liquefied gas/natural gas, or new energy. Perceived socioeconomic status was measured using the question "How do you think about your socioeconomic status by comparing with your peers?", with response options being: *lower*, *almost the same*, *and higher* (20).

Covariates

Covariates included participant, household, and regional features. First, we considered nine participant features, consisting gender (*Sex*), age (*Age*), health status (*Heal*), political affiliation (*Poli*), years of education (*Edu*), marital status (*Marr*), employment status (*Work*), religion (*Reli*) and insurance (*Insu*). Therein, health status were categorized into *very unhealthy*, *relatively unhealthy*, *average*, *relatively healthy*, or *very healthy*; political affiliation were categorized into yes for members of the communist party or communist youth league, or no for non-members; marital status were categorized into married or unmarried/divorced/widowed; employment status were categorized into jobless, farmer, or non-farmer; religion were categorized into having any religion or having no

TABLE 1 | Descriptive statistics of the model variables.

Variable type	Variable name	Mean	Std	Min	Max
Outcome variable	Happiness (*Happ*)	3.8689	0.8262	1.0000	5.0000
Primary predictor	Rural energy consumption revolution (*EneRef*)	0.4970	0.5002	0.0000	1.0000
Mediators	Log of household economic income (*Lginco*)	3.2198	1.6160	0.0000	6.9996
	Log of average monthly electricity spending per person (*Lgespp*)	1.3935	0.4217	0.0000	3.3981
	Leisure activities (*Leis*)	3.2303	1.0124	0.0000	5.0000
	Habitat environment satisfaction (*Envir*)	3.4492	0.9604	0.0000	5.0000
Moderators	Region identification (*Regid*)	2.4447	0.96255	1.0000	4.0000
	Socioeconomic status level (*Incox*)	1.6523	0.5516	1.0000	3.0000
	Energy type (*Enetype*)	1.8121	0.4632	1.0000	3.0000
Covariates	Gender of interviewee (*Sex*)	0.4894	0.5001	0.0000	1.0000
	Age of interviewee (*Age*)	52.3909	15.2984	18.0000	91.0000
	Health status (*Heal*)	3.4545	1.1397	1.0000	5.0000
	Political appearance (*Poli*)	0.0818	0.2742	0.0000	1.0000
	Years of education (*Edu*)	6.5508	4.0692	0.0000	19.0000
	Marital status (*Marr*)	0.8265	0.3788	0.0000	1.0000
	Work status (*Work*)	0.8811	0.7287	0.0000	2.0000
	Religion (*Reli*)	0.1174	0.3220	0.0000	1.0000
	Number of children (*Child*)	2.1144	1.3426	0.0000	10.0000
	Son aged 18–35 (*Son*)	0.2492	0.4327	0.0000	1.0000
	Number of properties (*House*)	1.1212	0.4445	0.0000	5.0000
	Social insurance (*Insu*)	1.6788	0.6550	0.0000	4.0000

The statistics were based on the China General Social Survey (CGSS) in 2015.

religion; and insurance were measured as having each of the four insurance types: basic medical insurance, basic pension, commercial medical insurance, and commercial pension. Second, we considered three household features: quantity of children (*Child*), whether there was a son aged 18–35 (*Son*) (29), and number of houses (*House*). Third, we considered a regional characteristic (*Regn*), which represented the provinces in China.

The descriptive statistics for all the above variables can be found in **Table 1**.

Model Analysis

We conducted data analysis in three steps using STATA version 16. First, using subjective happiness as the outcome, we built benchmark regression models, take response to the rural energy consumption revolution and the covariates as the predictors (20, 29–31).

$$Happ_i = Ctem + \alpha_{xy} EneRef_i + \sum \gamma_m Cont_{mi} + \theta Regn_j + \varepsilon_i \quad (1)$$

In Equation (1), $Happ_i$ is the i_{th} household's subjective happiness, $EneRef_i$ is the binary variable indicating whether the farmer responded positively to the energy consumption revolution, $Cont_{mi}$ is a series of covariates including participant and household characteristics, $Regn_j$ is the fixed effect for the j_{th} province, $Ctem$ is a constant, and ε_i is the random error.

Second, to inspect potential mediation effects, we built mediation models by adding mediators into the above benchmark model as follows:

$$\begin{aligned} Medi_i &= Ctem + \alpha_{xz} EneRef_i + \sum \gamma_m Cont_{mi} \\ &\quad + \theta Regn_j + \varepsilon_i \qquad (2) \\ Happ_i &= Ctem + \alpha_{xy'} EneRef_i + \alpha_{zy} Medi_i + \sum \gamma_m Cont_{mi} \\ &\quad + \theta Regn_j + \varepsilon_i \qquad (3) \end{aligned}$$

In Equations (2) and (3), *Medi* is the potential mediators (i.e., *Lginco*, *Lgespp*, *Leis* or *Envir*). Since *Happ*, *Leis* and *Envir* are ordinal variables, we modeled these variables with ordered probability model. The other outcomes were modeled with linear regression based on the Ordinary Least Squares (OLS) method.

When using the ordered probability model in mediation analysis, the decomposition of total effect based on the traditional linear regression is not applicable. Therefore, we adopted the modified mediation effect test and decomposition method as described in (32, 33). In this method, the first step is to conduct a coefficient t-test of rural energy consumption revolution (*EneRef*) in the benchmark model (Equation 1). If α_{xy} is significant, we proceed to the second step; otherwise, the mediation effect is considered to be non-significant. The second step is to test models in Equation (2) and (3) separately. If the t-tests for both α_{xz} and α_{zy} are significant, we would skip third

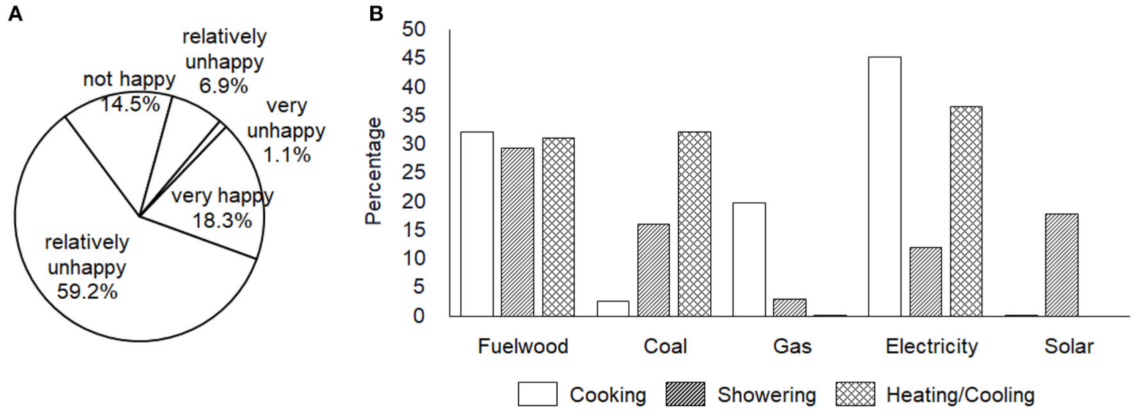


FIGURE 2 | The data distribution of happiness index **(A)** and energy consumption revolution **(B)** based on the China General Social Survey (CGSS) in 2015.

step and move forward to the fourth step; otherwise, we would proceed the third step to conduct an Iacobucci-z test:

$$z = \frac{z_{\alpha_{xz}} z_{\alpha_{zy}}}{\sqrt{z_{\alpha_{xz}}^2 + z_{\alpha_{zy}}^2 + 1}} \quad (4)$$

In Equation (4), z_{xz} and z_{zy} are the t-statistics of α_{xz} and α_{zy} in regression models 2 and 3. If the test results for Iacobucci-z is not significant, the mediation effect will be non-significant. If the Iacobucci-z test is significant, we proceed to the fourth step, in which we will determine whether the t-test of $\alpha_{xy'}$ is significant or not. If the t-test is non-significant, it will be a complete mediation effect. In contrast, if there is significant t-test, then it indicates a partial mediation effect and the Breen decomposition must be used to calculate the direct and indirect effects in the fifth step (33). In step 5, we define $\sigma_e = sqrt(3)\sigma_\varepsilon/\pi$, where σ_ε is the standard error of the random error ε_i in Equation (3), the direct effect is $\alpha_{xy'}/\sigma_e$, the mediation effect is $\alpha_{xz}\alpha_{zy}/\sigma_e$, and the total effect is $(\alpha_{xy'} + \Sigma\alpha_{xz}\alpha_{zy})/\sigma_e$.

Lastly, we built separate mediation models stratified by the moderators (i.e., *Regid, Incox,* and *Enetype*).

RESULTS

Respondents' Characteristics

In the 1320 rural households included in this analysis, there were respectively, 14 (1.1%), 91 (6.9%), 191 (14.5%), 782 (59.2%), and 242 (18.3%) households chose "*very unhappy*", "*relatively unhappy*", "*Just so-so*", "*relatively happy*" and "*very happy*" (**Figure 2A**). In cooking activity, 35% of the total households remained to use fuelwood and coal, while 45% and 20% chose electricity and gas, separately; for showering, 45% still used fuelwood and coal, while 18% and 12% used solar and electricity, separately; for heating/cooling, 63% still kept fuelwood and coal, while 37% used electricity (**Figure 2B**). Overall, a total of 656 (49.7%) households responded positively to the call for rural energy consumption revolution through various measures. However, only 23.5% of the 1,320 households completed the structural transition to modern energy in all of the three main energy consumption activities.

Benchmark Regression Models

We summarized the results from benchmark regression models in **Table 2**. Regression model 1 shows that, responding proactively to the energy consumption revolution may increase the probability of enhancing farmers' happiness level by 22.7%. The regression results for (2–1), (2–2), (2–3) and (2–4) of potential mediating variables *Lginco*, *Lgespp*, *Rest* and *Envir*, showed that the rural energy consumption revolution did not significantly increase household income, while it significantly increased per capita electricity expenditure and also the rest time of farmers. Likewise, it was inapparent in the impact on the satisfaction of rural living environment. Furthermore, regression models 1 and 3 show that older age, better health status, and possessing more house property were significantly associated with increasing of happiness, while having a son aged 18–35 was significantly associated with decreasing of happiness. It is not observed any significant association among happiness and other covariates including sex, political affiliation, years of education, marital status, employment status, religion, number of children in the household, and insurance.

Mediation Analysis

We summarized results from the mediation models in **Table 3**. As showed in the table, the tests of α_{xy}, α_{xz}, α_{zy} and $\alpha_{xy'}$ were t-test statics about coefficient of rural energy consumption revolution (*EneRef*) in the benchmark models 1, 2, and 3. The Iacobucci-z test was used to confirm the significance when only one of the α_{xz} and α_{zy} was significant. In the first step, α_{xy} test was significantly positive for all mediators. In the second step, the α_{xz} test was significantly positive for electricity expenditure (*Lgespp*) and leisure activities (*Leis*), but non-significant for household income (*Lginco*) and living environment satisfaction (*Envir*). In the third step, the α_{zy} test is significantly positive for leisure activities (*Leis*) and living environment satisfaction (*Envir*), but non-significant for household income (*Lginco*) and electricity

TABLE 2 | Benchmark regression results of the mediation model.

	Happ **(Model 1)**	*Lginco* **(Model 2–1)**	*Lgespp* **(Model 2–2)**	*Leis* **(Model 2–3)**	*Envir* **(Model 2–4)**	*Happ* **(Model 3)**
EneRef	0.227***	0.085	0.130***	0.196***	0.071	0.211***
	(0.070)	(0.086)	(0.025)	(0.065)	(0.067)	(0.073)
Lginco			0.004	−0.029	−0.019	−0.008
			(0.008)	(0.022)	(0.022)	(0.022)
Lgespp		0.043		−0.105	−0.076	−0.035
		(0.096)		(0.070)	(0.075)	(0.084)
Leis		−0.056	−0.016		0.002	0.088***
		(0.043)	(0.011)		(0.033)	(0.034)
Envir		−0.033	−0.012	0.003		0.089**
		(0.043)	(0.012)	(0.033)		(0.036)
Sex	−0.108	0.546***	−0.010	0.050	0.034	−0.112
	(0.068)	(0.085)	(0.025)	(0.064)	(0.063)	(0.069)
Age	0.009***	0.006	0.002*	−0.001	0.010***	0.009***
	(0.003)	(0.004)	(0.001)	(0.003)	(0.003)	(0.003)
Heal	0.261***	0.151***	−0.008	0.047	0.017	0.260***
	(0.034)	(0.039)	(0.011)	(0.030)	(0.032)	(0.034)
Poli	0.155	0.128	0.113**	0.154	0.009	0.156
	(0.112)	(0.140)	(0.049)	(0.106)	(0.104)	(0.117)
Edu	0.013	0.007	0.006*	0.030***	−0.000	0.012
	(0.010)	(0.013)	(0.003)	(0.010)	(0.010)	(0.011)
Marr	−0.004	0.492***	−0.066*	−0.119	0.098	−0.016
	(0.091)	(0.121)	(0.035)	(0.085)	(0.087)	(0.092)
Work	−0.078	0.787***	0.044**	−0.128***	0.030	−0.055
	(0.048)	(0.068)	(0.018)	(0.048)	(0.049)	(0.051)
Reli	0.142	0.060	0.040	0.183*	−0.092	0.160
	(0.100)	(0.129)	(0.040)	(0.108)	(0.104)	(0.102)
Child	0.046	−0.025	−0.008	0.057*	−0.014	0.044
	(0.032)	(0.039)	(0.011)	(0.031)	(0.034)	(0.032)
Son	−0.212***	−0.073	0.038	−0.051	0.085	−0.213***
	(0.076)	(0.092)	(0.027)	(0.069)	(0.073)	(0.077)
House	0.218**	0.243***	0.032	0.081	−0.044	0.227***
	(0.086)	(0.081)	(0.025)	(0.066)	(0.075)	(0.088)
Insu	0.019	0.116*	0.025	0.058	0.021	0.017
	(0.048)	(0.065)	(0.017)	(0.045)	(0.045)	(0.049)
Fixed effect	Y	Y	Y	Y	Y	Y
Sample size	1320	1320	1320	1320	1320	1320
R^2	0.071	0.261	0.100	0.042	0.021	0.091

****, **, and * indicate that the results are significant at 1, 5, and 10% levels, respectively.*

spending (*Lgespp*). In the fourth step, the Iacobucci-z test showed that the mediation effects of electricity spending (*Lgespp*) and living environment satisfaction (*Envir*) were not significant. In the fifth step, the $\alpha_{xy'}$ test revealed that only leisure activities (*Leis*) were a significant mediator of the relationship between energy consumption revolution and farmers' happiness. The other potential mediators were not significant.

We further applied the Breen decomposition to quantify the mediation effects. As seen in **Table 3**, in the relationship where rural energy consumption revolution significantly increased farmers' happiness the direct effect accounted for over 90% of the total effect and the mediation effect accounted for <10% of the total effect. All of the mediation effects were brought by increased leisure activities (*Leis*), while the mediation effect of economic [i.e., household income (*Lginco*) and electricity spending (*Lgespp*)] and environmental (*Envir*) factors was not significant.

Moderation Analysis

We summarized results from the moderation analysis in **Table 4**. As seen in the table, the effect of energy consumption revolution on farmers' happiness varied slightly across geographic regions. Overall, the direct effect of energy consumption revolution on increasing farmers' happiness decreased from the western to central-eastern regions, with the largest value being observed in the less-developed western region. Besides, farmers in the

TABLE 3 | Testing results and decomposition of the mediation effect of energy consumption revolution on improving farmers' happiness.

	Potential mediators	Test of α_{xy}	Test of α_{xz}	Test of α_{zy}	Iacob-z test	$\alpha_{xy'}$ test (direct effect)	Mediation effect	Total effect
Economic effect	*Lginco*	0.227*** (0.070)	0.085 (0.086)	−0.008 (0.022)	-		Non- significant	
	Lgespp	0.227*** (0.070)	0.130*** (0.025)	−0.035 (0.084)	−0.005 (−0.011)		Non-significant	
Social effect	*Leis*	0.227*** (0.070)	0.196*** (0.065)	0.088*** (0.034)	-	0.211*** (0.073)	0.017* (0.009)	0.227*** (0.070)
Environmental effect	*Envir*	0.227*** (0.070)	0.071 (0.067)	0.089** (0.036)	0.006 (0.006)		Non-significant	

*All regressions include all control variables and fixed effects; ***, **, and * indicate that the regression results are significant at 1, 5, and 10% levels, respectively.*

TABLE 4 | The moderation effect of energy consumption revolution on improving farmers' happiness.

	Group	Gross effect	Direct effect	Indirect effect	
Region (*Regid*)	Western regions	0.274** (0.118)	0.248** (0.121)	*Lginco*	Non-significant
				Lgespp	Non-significant
				Leis	Non-significant
				Envir	Non-significant
	Central-eastern regions	0.231*** (0.087)	0.209** (0.089)	*Lginco*	Non-significant
				Lgespp	Non-significant
				Leis	significant(+)
				Envir	Non-significant
Income level (*Incox*)	Low income	0.238** (0.116)	0.274** (0.118)	*Lginco*	Non-significant
				Lgespp	Near-significant(-)
				Leis	Non-significant
				Envir	Non-significant
	High income	0.177* (0.090)	0.142 (0.092)	*Lginco*	Non-significant
				Lgespp	Non-significant
				Leis	Significant(+)
				Envir	Non-significant
Energy type (*Enetype*)	Electricity	0.217** (0.085)	0.206** (0.087)	*Lginco*	Non-significant
				Lgespp	Non-significant
				Leis	Significant(+)
				Envir	Non-significant
	Gas (liquefied gas/natural gas)	0.267 (0.207)	0.221 (0.220)	*Lginco*	Non-significant
				Lgespp	Non-significant
				Leis	Non-significant
				Envir	Non-significant
	New energy	0.130 (0.213)	0.109 (0.228)	*Lginco*	Non-significant
				Lgespp	Non-significant
				Leis	non-significant
				Envir	Non-significant

*All regressions include all control variables and fixed effects; ***, ** and * indicate that the regression results are significant at 1, 5, and 10% levels, respectively.*

more-developed central-eastern region are more likely to prefer leisure activities when they have time. So, the leisure activities were significantly increased because of the energy consumption revolution in central-eastern China, which in turn significantly increased farmers' happiness.

Table 4 reported the moderation effect of household income (*Incox*) on the relationship between energy consumption revolution and farmers' happiness. We observed large differences in the effect of the energy consumption revolution on happiness across different household income levels. First, the direct effect of energy consumption revolution on happiness decreased from the low-income households to middle and high-income households. Second, farmers from middle- and high-income households cared more about leisure activities. The energy consumption revolution significantly increased the happiness of middle- and high-income farmers by increasing leisure activities, but the mediation effects of economic and environmental factors were not significant. Third, low-income farmers were more susceptible to the rising cost of clean energy. The energy consumption revolution near significantly reduced the happiness of low-income farmers because of the increasing electricity spending by energy consumption revolution.

The moderation effect of energy type (*Enetype*) was also reported in **Table 4**. Different types of energy had differential effects on farmers' happiness. First, the total effect of electricity, gas, and new energy on increasing farmers' happiness decreased in this order, with only electricity being significant. Second, electrical energy consumption significantly increased farmers' happiness by increasing leisure activities, while the effect of new energy sources such as liquefied gas, natural gas and especially solar energy was not significant.

Endogenous Treatment and Robustness Test

Endogenesis may come from a variety of complex factors, which can lead to systematic bias in estimates. As the model was based on cross-sectional data, propensity score matching (PSM) method was adopted to deal with the endogeneity problem (34). According to all control variables, 656 households were matched with 1:1 nearest neighbor, and 1312 matched regression samples were obtained. We found that there was a significant difference in density distribution between the control group and the treatment group before matching, and the density distribution of the control group was closer to that of the treatment group from all dimensions after matching, so the "systematic bias" between the control group and the treatment group could be better eliminated, and the propensity score matching achieved a good effect.

Further, we did a re-regression of the above benchmark and mediation model based on matched samples (**Table 5**). In terms of the magnitude, direction and significance of the key coefficients,the regression results before and after matching were consistent. Therefore, it is believed that there is no obvious bias in our models and the endogeneity problem will not have a systematic impact on the regression model.

Also, we did a serial of robustness tests. First, we adjusted the definition of Energy consumption Revolution from "If a household i has completed at least two of the three activities involving energy consumption, then $EneRef_i = 1$" to "If a household i has completed all the three activities involving energy consumption, then $EneRef_i = 1$". Second, we changed the regression method from Ordered Probit to Ordered Logit. Third, we did a Placebo test by manufacturing a treatment variable of Energy consumption Revolution with the random selection. We randomly selected 656 farmers from 1320 samples as the counterfactual treatment group of rural energy consumption revolution, constructed a pseudo-explanatory variable (*EneRef_fake*), and used it to estimate the above benchmark model. The above process was repeated 500 times and 1000 times respectively, and the density distribution of the regression coefficient (α_{xy}) of the primary explanatory variable (*EneRef_fake*) was obtained, as shown in **Figure 3**. This key variable, i.e., the *EneRef_fake*'s regression coefficient α_{xy}, was concentrated around 0, and the 1000 random results (right) were closer to 0 than the 500 random results (left). These results indicated that all these robustness tests were passed.

TABLE 5 | Regression results of benchmark model and mediation model after propensity score matching.

	Happ	*Lginco*	*Lgespp*	*Rest*	*Envir*	*Happ*
EneRef	0.233***	0.075	0.126***	0.187***	0.079	0.218***
	(0.071)	(0.087)	(0.025)	(0.065)	(0.068)	(0.072)
Lginco			0.004	−0.029	−0.018	−0.001
			(0.008)	(0.022)	(0.022)	(0.022)
Lgespp		0.043		−0.107	−0.071	−0.015
		(0.097)		(0.071)	(0.076)	(0.084)
Rest		−0.056	−0.016		0.008	0.087**
		(0.043)	(0.011)		(0.033)	(0.034)
Envir		−0.030	−0.012	0.009		0.101***
		(0.043)	(0.012)	(0.034)		(0.036)
Samples	1312	1312	1312	1312	1312	1312
R^2	0.073	0.261	0.101	0.044	0.023	0.091

*All regressions include all control variables and fixed effects; *** and ** indicate that the regression results are significant at 1 and 5% levels, respectively.*

DISCUSSION

Overall, our results show that the rural energy consumption revolution improved farmers' happiness. However, the mediation analyses show that the direct effect accounted for over 90% of the total effect; the revolution increased farmers' happiness is only through increasing leisure activities, but not through improving household income or living environment. To achieve ultimate success for energy consumption revolution to be successful, it must realize the full potential of the revolution by promoting its impact on the socioeconomic and environmental factors, so as to develop a multi-dimensional mechanism for increasing happiness.

While the rural energy consumption revolution has marginally increased farmers' happiness by increasing leisure activities, its negative impact on happiness for higher electricity expenditure is also significant especially to rural low-income households. In other words, the revolution is a double-edged sword, not only to increase leisure activities by liberating people from daily chores, but also to raise people's electricity expenditure, which leads to additional financial burden for rural low-income households. Therefore, when promoting the energy consumption revolution in rural areas, it is important to regulate electricity and gas prices, so as to guarantee that they

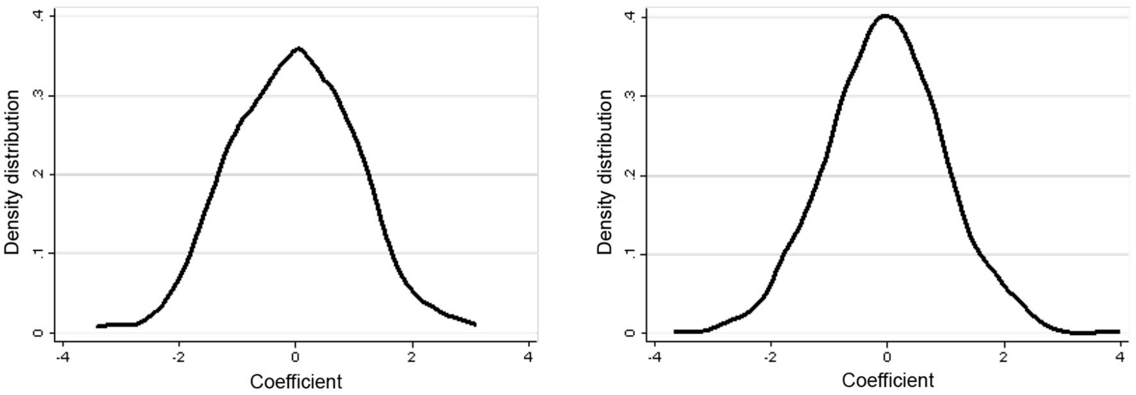


FIGURE 3 | Placebo test results using 500 and 1,000 random counterfactual explanatory variables. All regressions include all control variables and provincial fixed effects.

are affordable for farmers. In addition, the energy consumption revolution has yet to improve the rural living environment, which reflects a low energy utilization rate (i.e., incomplete energy upgrades) in rural China.

Our results also show that geographic region and household income level moderated the relationship between energy consumption revolution and farmers' happiness. It is found that rural residents of less-developed Western China or low-income households were more likely to respond to the energy consumption revolution. Besides, rural residents of Central-Eastern China, where are more-developed, including middle and high-income households, pay more attention to increasing leisure activities brought by the energy consumption revolution, while rural residents of Western China or low-income households were more sensitive to the spending of electricity use.

Results from our study are meaningful for policy implications. First, it is difficult to promote the rural energy consumption revolution only as a national policy. It is important to act synergistically, by implementing or integrating it with other relevant national policies, such as policy for 'Pollution Prevention and Abatement' (35) and 'Poverty Reduction' (36). It is observed that only 23.5% of the rural households have completed energy upgrade, it shows that the revolution did encounter obstacles in rural China. Lack of a multi-dimensional mechanism for improving farmers' happiness is the main reason for it. Therefore, it is necessary to synergistically implement the rural energy consumption revolution with the national targeted poverty alleviation policy as well as the national pollution prevention and control policy, forming a multi-pronged strategy that simultaneously targets the socioeconomic and environmental factors associated with farmers' happiness.

Second, from the perspectives of both equity and efficiency, to introduce an energy policy that favors the poor Western regions and low-income households are important for promoting the rural energy consumption revolution. In terms of equity, "Energy Poverty Reduction" (37) is an important part of the national target of poverty reduction policy. Our results suggest that promoting the energy consumption revolution in Western China and low-income households can increase happiness. As to efficiency, the poor Western China and low-income households should be paid more attention in the energy consumption revolution. The benefits of the new energy policy, including national investments in energy infrastructure, subsidies for terminal equipment (e.g., heater, refrigerator, and air conditioner), especially price regulations related energy use, should be introduced firstly to the poor rural areas, so that farmers' happiness can be increased. In return, it will further facilitate the rural energy consumption revolution.

DATA AVAILABILITY STATEMENT

Publicly available CGSS datasets were analyzed for this study, which can be found here http://cgss.ruc.edu.cn/. Further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

ZX conceived the model and analyzed results. RG collected the data. ZX and RG wrote the manuscript together. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the China National Philosophy and Social Science Foundation in 2021: Research on the dynamic mechanism and realization path of urbanization with county as an important carrier (21BJY017).

ACKNOWLEDGMENTS

We wish to thank the CGSS staff for their dedication to survey data collection and processing.

REFERENCES

1. Schwab K. The fourth industrial revolution: what it means, how to respond. *Economy, Culture & History Japan Spotlight Bimonthly.* (2016) 3–5.
2. Ni W, Jin Y, Ma L, Hu S. A Preliminary Discussion on Strategic Issues of Promoting the Revolution of Energy Consumption and Controlling the Total Energy Consumption in China. *Strategic Study of CAE*. (2015) 17:111–7.
3. Xi J. Secure a decisive victory in building a moderately prosperous society in all respects and Strive for the great victory of socialism with Chinese characteristics in the new era. *The Report of the 19th National Congress of the CPC.* People's Publishing House, Beijing (2017).
4. Leach G. The energy transition. *Energy Policy.* (1992) 8:116–23. doi: 10.1016/0301-4215(92)90105-B
5. Masera OR, Saatkamp BD, Kammen DM. From linear fuel switching to multiple cooking strategies: a critique and alternative to the energy ladder model. *World Dev*. (2000) 28:2083—103. doi: 10.1016/S0305-750X(00)00076-0
6. Zhang R, Wei TY, Sun J, Shi QY. Wave transition in household energy use. *Technol Forecast Soc Change*. (2016) 102:297–308. doi: 10.1016/j.techfore.2015.08.014
7. Barnes DF, Floor W. Biomass energy and the poor in the developing countries. *Int Aff.* (1999) 53:237–59.
8. Qiu H, Yan J, Li D, Han W. Residential energy consumption in rural China: current situation and determinants for future trend–an empirical study based on filed survey data of 4 provinces. *China Soft Sci.* (2015) 11:28–38. doi: 10.3969/j.issn.1002-9753.2015.11.004
9. Household Energy Consumption Research Group. *Chinese Household Energy Consumption Research Report (2015).* The annual report of national academy of development and strategy, Renmin University of China (2016).
10. Xu Y, Huang L. Beijing-Tianjin-Hebei rural energy consumption structure research. *J Hebei Univ Econ Bus*. (2018) 39:65–73. doi: 10.14178/j.cnki.issn1007-2101.2018.03.009
11. Xiong W, Fu Z, Wang P. Promotion of rural energy supply side reform under ecological environment construction. *Rural Econ.* (2017) 12:94–9.
12. Zhang J. Institutional analysis on sustainable supply of rural energy for livelihood. *Prod Res*. (2011) 2:25–7.
13. Jain G. Energy security issues at household level in India. *Energy Policy.* (2010) 38:2835–45. doi: 10.1016/j.enpol.2010.01.016
14. Zhang H, Mou J, Yin H. An empirical analysis of household energy consumption demand in rural areas of forest region-based on a double-extended linear expenditure system model. *Chin Rural Econ.* (2010) 7:64–74. doi: 10.3724/SP.J.1011.2010.01351

15. Shi Q, Peng X, Zhang R. A field survey on rural energy consumption in China - a case study of 2253 farming households in Jin-Qian-Zhejiang provinces. *Manage World*. (2014) 5:80–92.
16. Broadhead J, Bahdon J, Whiteman A. *Woodfuel Consumption Modelling and Results, Annex 2 in: Past Trends and Future Prospects for the Utilization of Wood for Energy*. Working Paper No. GFPOS/WP/05, Global Forest Products Outlook Study, FAO, Rome (2001)
17. Chen L, Heerink N, Berg M. Energy consumption in rural China: a household model for three villages in Jiangxi Province. *Ecol Econ*. (2006) 58:407–20. doi: 10.1016/j.ecolecon.2005.07.018
18. Zheng F, Liu J. The impact of household energy consumption structure on rural women's time allocation: a case study of Zhijin County, Guizhou Province. *J Agrotech Econ*. (2010) 10:72–81.
19. Zhang R, Wei TY, Glomsrød S, Shi QH. Bioenergy consumption in rural China: evidence from a survey in three provinces. *Energy Policy*. (2014) 75:136–45. doi: 10.1016/j.enpol.2014.08.036
20. Chen G, Li S. How can government make people happy? An empirical study of the impact of government quality on residents' happiness. *Manage World*. (2012) 8:55–67.
21. Yang J, Zhang Y. Pricing air pollution: an analysis based on happiness data. *J World Econ*. (2014) 37:162–88.
22. Zimmermann S. The Pursuit of subjective well-being through specific consumption choice. Available at SSRN 2484660.
23. Easterlin RA. Does economic growth improve the human lot? Some empirical evidence. *Nat Households Econ Growth*. (1974) 89:89–125. doi: 10.1016/B978-0-12-205050-3.50008-7
24. Easterlin RA, Morgan R, Switek M, Wang F. China's life satisfaction, 1990–2010. *Proc Natl Acad Sci USA*. (2012) 109:9775–80. doi: 10.1073/pnas.1205672109
25. Sun J, Wei W. A study on the relationship between economic welfare and happiness. *Consu Econ*. (2016) 32:80–8.
26. Ball R, Chernova K. Absolute income, relative income and happiness. *Social Indicator Research*. (2008) 88:497–529. doi: 10.1007/s11205-007-9217-0
27. An H, Ye J. The effect of housing price on happiness and its mechanism. *Guizhou Soc Sci*. (2018) 4:109–16.
28. Luorong Z. Leisure and happiness-a study of the changes of labor inputs of Tibetan farmers. *J Minzu Univ China*. (2016) 6:82–8.
29. Lu F, Liu G, Li H. The gender of children and parents' happiness. *Econ Res J*. (2017) 10:173–88.
30. Ma W, Wang X, Li H. The impact mechanism of income gap on happiness. *Econ Perspect*. (2018) 11:74–87.
31. Wen Z, Hou KT, Zhang L. A comparison of moderator and mediator and their applications. *Acta Psychol Sinica*. (2005) 37:268–74.
32. Buis ML. Direct and indirect effects in a logit model. *Stata J*. (2010) 10:11–29. doi: 10.1177/1536867X1001000104
33. Breen R, Karlson KB, Holm A. Total, direct, and indirect effects in logit and probit molels. *Sociol Methods Res*. (2013) 42:164–91. doi: 10.1177/0049124113494572
34. Leng C, Zhu Z. Research on the happiness effect of internet on rural residents. *S Chin J Econ*. (2018) 8:107–27. doi: 10.19592/j.cnki.scje.351019
35. World Bank Group. *Pollution Prevention and Abatement Handbook 1998: Toward Cleaner Production*. Washington, DC: The World Bank (1999).
36. Freeman HA, Ellis F, Allison E. Livelihoods and rural poverty reduction in Uganda. *World Dev*. (2003) 31:997–1013. doi: 10.1016/S0305-750X(03)00043-3
37. Cook CC, Duncan T, Jitsuchon S, Sharma A, Guobao W. *Assessing the Impact of Transport and Energy Infrastructure on Poverty Reduction*. Manila: Asian Development Bank (2005). Available online at: https://www.adb.org/sites/default/files/publication/27956/assessing-transport-energy.pdf

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 March 2022
doi: 10.3389/fpubh.2022.778463

Assessing the Documentation of Social Determinants of Health for Lung Cancer Patients in Clinical Narratives

*Zehao Yu, Xi Yang, Yi Guo, Jiang Bian and Yonghui Wu**

Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, United States

Social determinants of health (SDoH) are important factors associated with cancer risk and treatment outcomes. There is an increasing interest in exploring SDoH captured in electronic health records (EHRs) to assess cancer risk and outcomes; however, most SDoH are only captured in free-text clinical narratives such as physicians' notes that are not readily accessible. In this study, we applied a natural language processing (NLP) system to identify 15 categories of SDoH from a total of 10,855 lung cancer patients at the University of Florida Health. We aggregated the SDoH concepts into patient-level and assessed how each of the 15 categories of SDoH were documented in cancer patient's notes. To the best of our knowledge, this is one of the first studies to examine the documentation of SDoH in clinical narratives from a real-world lung cancer patient cohort. This study could guide future studies to better utilize SDoH information documented in clinical narratives.

Keywords: electronic health records, natural language processing, cancer, social determinants of health (SDOH), lung cancer

OPEN ACCESS

Edited by:
Angela Kaida,
Simon Fraser University, Canada

Reviewed by:
Liwei Wang,
Mayo Clinic, United States
James Cimino,
University of Alabama at Birmingham,
United States

****Correspondence:***
Yonghui Wu
Yonghui.wu@ufl.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: *16 September 2021*
Accepted: *28 February 2022*
Published: *28 March 2022*

Citation:
Yu Z, Yang X, Guo Y, Bian J and Wu Y (2022) Assessing the Documentation of Social Determinants of Health for Lung Cancer Patients in Clinical Narratives.
Front. Public Health 10:778463.
doi: 10.3389/fpubh.2022.778463

INTRODUCTION

As the second leading cause of death in the United States (US) (1), cancer has a long list of risk factors, ranging from biological traits to clinical characteristics to social determinants of health (SDoH) (2). In recent years, there is an increasing interest in examining how SDoH contribute to cancer risk and treatment outcomes (3). The Healthy People 2030 defined SDoH as "the conditions in the environments where people are born, live, learn, work, play, worship, and age that affect a wide range of health, functioning, and quality-of-life outcomes and risks" and categorized SDoH into 5 domains, including economic stability, education access and quality, healthcare access and quality, neighborhood and built environment, and social and community context (4). A recent study (5) reported that up to 75% of cancers occurrences are associated with SDoH rather than clinical factors. Other studies have shown that many SDoH contribute to individual cancer risk, influence the likelihood of survival, and affect cancer early prevention and health equity (2, 6, 7). A recent study (8) reported that SDoH factors such as poverty, lack of education, neighborhood disadvantage, and social isolation play important roles in breast cancer stage and survival. Many SDoH factors are also associated with the screening of cervical cancer, breast cancer, and lung cancer (9).

In the past decade, the rapid adoption of electronic health record (EHR) systems has made it possible to use the rich data elements (e.g., disease diagnoses, medications) captured in longitudinal patient's EHR data for cancer studies. However, it is challenging to extract SDoH from EHRs for assessing cancer outcomes as most SDoH were captured as free text in clinical notes rather than structured fields. In February 2018, the World Health Organization (WHO) defined structured codes to capture some of the SDoH. More specifically, the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM) Z codes (Z55–Z65) can be used to capture some SDoH. However, our previous study analyzed EHR data in a large clinical research network and showed that the use of SDoH Z codes is still quite low (10). Furthermore, it is unclear how well the SDoH information was documented in clinical notes for cancer patients. On the other hand, natural language processing (NLP) is the key technology to extract SDoH from clinical notes. NLP has been applied to extract various information such as diagnoses, lab tests, side effects from clinical narratives. We have explored many NLP models including state-of-the-art transformer-based NLP models in our previous studies (11–13). In a prior study (14), we have developed an NLP package to systematically extract SDoH from clinical notes using a subset of notes identified with a keyword matching pipeline.

In this study, we identified a cohort of lung cancer patients using ICD-9 and ICD-10 codes from the University of Florida Health (UF Health) system. We applied our NLP pipeline to systematically extract a total of 15 different categories of SDoH and examined the proportion of lung cancer patients who had various SDoH documented in clinical notes. This study is one of the earliest studies to examine how well the SDoH was documented in a real-world cohort of lung cancer patient, which will guide future studies exploring SDoH from clinical text for cancer studies.

METHODS

Study Population: Lung Cancer Patients

In this study, we obtained clinical notes from the UF Health integrated data repository (IDR), a secure clinical data warehouse (CDW) that aggregates data from UF Health's various clinical and administrative information systems including the Epic electronic medical record (EMR) system. We used ICD-9 codes (162*) and ICD-10 codes (C34*) to identify a cohort of lung cancer patients from the UF Health IDR between 2011 and 2020. Patients who had at least one of the ICD-9 or ICD-10 codes and aged at least 18 years were included in the cohort. For each patient, we collected all types of clinical notes associated with the patient, which were used as the resource to extract SDoH concepts.

An NLP Pipeline for Extracting SDoH

In our previous study (14), we created an SDoH corpus using 500 notes and developed an NLP pipeline that can extract 5 different categories of SDoH including gender, ethnicity, smoking, employment, and education from clinical narratives. In this study, we further extended the annotation with 10 new SDoH categories including race, alcohol use, drug use, marital status, occupation, language, physical activity, transportation, financial constraint, and social cohesion. Financial constraint indicates patients having a temporary or current financial problem but not in a poverty status (e.g., difficulty paying for the basics). Social cohesion indicates how well the patient connects to the society (e.g., attends religious service). Next, we re-trained the NLP model using this new corpus and developed an upgraded NLP pipeline that can extract a total number of 15 different categories of SDoH from clinical narratives. The transformer-based NLP model using the BERT architecture (15) was used in this study as it achieved the best performance in our previous study (14). BERT is a bidirectional transformer-based NLP model based on masked language modeling (MLM) and uses next-sentence prediction (NSP) to learn representations from text. This SDoH pipeline first identifies the SDoH concepts and then links them to various attributes including status, frequency, and negations. We reused the clinical transformer models developed in our previous study (11) implemented using the HuggingFace (16) package in PyTorch (17). We applied this NLP pipeline to all the clinical notes collected for our lung cancer patient cohort to extract SDoH concepts. Lastly, we aggregated the SDoH concepts at the patient level to assess the proportion of patients who had at least one SDoH concept documented in each of the 15 categories. When there were multiple SDoH instances extracted for one patients of the same category, we adopted majority voting strategy to keep the instance that most frequently documented in clinical notes.

RESULTS

We identified a total of 10,855 lung cancer patients in UF Health between 2011 and 2020 and collected a total of 1,798,409 clinical notes. **Table 1** shows a summary of statistics for the demographics of this lung cancer cohort. Most patients (>95%) in this lung cancer cohort are >50 years old; there are more female patients than male and the majority race is White (>72%).

Based on our previous annotation of 5 SDoH categories using 500 clinical notes (14), we further annotated additional 10 SDoH categories and extended the previous annotation from 1,876 SDoH concepts of 5 categories to a total of 5,015 concepts of 15 different SDoH categories. Following the standard NLP development procedure, we divided the annotation into a training set and a test set using a ratio of 4:1. We used the training set to optimize the parameter of a BERT model and used the test set to calculate evaluation scores. We reused the same experiment settings for batch size and learning rate identified from our previous study (14). Using the new extended corpus, the performance (micro average F1 score for all SDoH categories) of our SDoH NLP pipeline based on the BERT model improved from 0.8791 (precision: 0.8848 and recall: 0.8734) to 0.9216 (precision: 0.9298, recall: 0.9136).

We applied the BERT-based NLP pipeline and identified a total number of 5,408,148 SDoH concepts from 1,798,409 clinical notes of 10,855 lung cancer patients. We then aggregated the SDoH concepts at the patient level and calculated the distribution of SDoH concepts for each category. Majority voting was used

TABLE 1 | Summary of statistics for the lung cancer cohort.

Demographics	Sub-categories	Descriptive statistics (*N* = 10,855)	Percentage of the cohort (%)
Age	18–30	79	0.73
	30–40	112	1.03
	40–50	283	2.61
	50–60	1,433	13.20
	60–70	3,441	31.70
	70–80	3,568	32.87
	>80	1,939	17.86
Gender	Female	4,643	57.23
	Male	6,212	42.77
Race	White	7,834	72.17
	Africa American	1,517	13.98
	Asian	88	0.81
	American Indian or Alaska Native or Native Hawaiian or Other Pacific Islander	18	0.17
	Multi-Race	19	0.18
	Other*	1,379	12.70
Ethnics	Hispanic or Latino	210	1.93
	Not Hispanic or Latino	9,459	87.14
	Other*	1,186	10.93

TABLE 2 | Social determinants of health (SDoH) concepts identified from the lung cancer patient cohort.

SDoH category	Total number of concepts detected by NLP	Total number of patients has at least one SDoH	Percentage of patients has at least one SDoH for current category (%)
Gender	843,066	9,552	98.7
Alcohol use	223,214	9,195	95.0
Drug use	180,309	8,756	90.5
Marital status	167,457	8,655	89.5
Education	167,018	8,463	87.5
Occupation	142,306	8,345	86.3
Smoking	132,833	7,639	79.0
Race	144,980	7,376	76.2
Ethnicity	86,789	5,231	54.1
Language	83,539	5,173	53.5
Physical activity	55,842	3,092	32.0
Transportation	24,191	2,877	29.7
Financial constraint	113,220	2,766	28.6
Social cohesion	9,170	2,727	28.2
Employment status	843,066	2,110	21.8

when there were multiple SDoH instances identified for one SDoH category. **Table 2** shows the total number of SDoH concepts identified in each SDoH category and the percentage of patients with at least one SDoH concept in each category. Among the 15 SDoH categories, 3 categories (i.e., gender, alcohol use, and drug use) were extremely frequent-documented in the lung cancer patients, where over 90% of the patients in this cohort had at least one SDoH documented; 5 categories (i.e., marital status, education, occupation, smoking, race) were frequent-documented, where over 70% of the patients had at least one SDoH documented; 7 categories (i.e., ethnicity, language, physical activity, transportation, financial constraint, social cohesion, employment status) were not frequent-documented, where <60% of the patients had at least one SDoH documented.

DISCUSSION

Many SDoH are associated with cancer risk and cancer treatment outcomes. Yet, information related to SDoH is often unavailable in structured EHRs but is often documented in clinical notes as free text, making it challenging to examine SDoH in cancer research. In this study, we identified a cohort of lung cancer patients and applied our NLP system to extract SDoH concepts from 15 categories of SDoH. We examined the distribution of SDoH in each category and evaluated how frequent SDoH was documented for categories. This study will guide future cancer studies that aim to explore SDoH information from clinical notes.

DATA AVAILABILITY STATEMENT

The data analyzed in this study is subject to the following licenses/restrictions: we obtained clinical notes from the UF Health integrated data repository (IDR), a secure clinical data warehouse (CDW) that aggregates data from UF Health's various clinical and administrative information systems including the Epic electronic medical record (EMR) system. Requests to access these datasets should be directed to UF CSTI, info@ctsi.ufl.edu.

AUTHOR CONTRIBUTIONS

JB, YG, and YW were responsible for the overall design, development, and evaluation of this study. XY collected the data used in this study and involved in the results analysis. ZY conducted the experiments and data analysis. YW did the initial drafts and revisions of the manuscript. All authors reviewed the manuscript critically for scientific content and gave final approval of the manuscript for publication.

FUNDING

This study was partially supported by a Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-2018C3-14754), a Grant from National Institute on Aging 1R56AG 069880, a Grant from the National Cancer Institute, 1R01CA246418 R01, and the Cancer Informatics and eHealth core jointly supported by the UF Health Cancer Center and the UF Clinical and Translational Science Institute.

ACKNOWLEDGMENTS

The authors would like to thank the NVIDIA Corporation with the donation of the GPUs used for this research.

REFERENCES

1. GBD 2013 Mortality and Causes of Death Collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet Lond Engl*. (2015) 385:117–71. doi: 10.1016/S0140-6736(14)61682-2
2. Hiatt RA, Breen N. The social determinants of cancer: a challenge for transdisciplinary science. *Am J Prev Med*. (2008) 35:S141–150. doi: 10.1016/j.amepre.2008.05.006
3. Braveman P, Gottlieb L. The social determinants of health: it's time to consider the causes of the causes. *Public Health Rep*. (2014) 129 Suppl 2:19–31. doi: 10.1177/00333549141291S206
4. Social Determinants of Health. *Healthy People 2030 | health.gov*. Available online at: https://health.gov/healthypeople/objectives-and-data/social-determinants-health (accessed September 14, 2021).
5. Akushevich I, Kravchenko J, Akushevich L, Ukraintseva S, Arbeev K, Yashin A. *Cancer Risk and Behavioral Factors, Comorbidities, and Functional Status in the US Elderly Population*. ISRN Oncol (2011) 2011. Available online at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3197174/ (accessed February 3, 2019).
6. Matthews AK, Breen E, Kittiteerasack P. Social determinants of LGBT cancer health inequities. *Semin Oncol Nurs*. (2018) 34:12–20. doi: 10.1016/j.soncn.2017.11.001
7. Gerend MA, Pai M. Social determinants of Black-White disparities in breast cancer mortality: a review. *Cancer Epidemiol Biomark Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. (2008) 17:2913–23. doi: 10.1158/1055-9965.EPI-07-0633
8. Coughlin SS. Social determinants of breast cancer risk, stage, and survival. *Breast Cancer Res Treat*. (2019) 177:537–48. doi: 10.1007/s10549-019-05340-7
9. Lofters AK, Schuler A, Slater M, Baxter NN, Persaud N, Pinto AD, et al. Using self-reported data on the social determinants of health in primary care to identify cancer screening disparities: opportunities and challenges. *BMC Fam Pract*. (2017) 18:31. doi: 10.1186/s12875-017-0599-z
10. Guo Y, Chen Z, Xu K, George TJ, Wu Y, Hogan W, et al. International classification of diseases, tenth revision, clinical modification social determinants of health codes are poorly used in electronic health records. *Medicine*. (2020) 99:e23818. doi: 10.1097/MD.0000000000023818
11. Yang X, Bian J, Hogan WR, Wu Y. Clinical concept extraction using transformers. *J Am Med Inform Assoc JAMIA*. (2020) 27:1935–42. doi: 10.1093/jamia/ocaa189
12. Yang X, Bian J, Wu Y. Detecting medications and adverse drug events in clinical notes using recurrent neural networks. In: *International Workshop on Medication and Adverse Drug Event Detection* (2018). p. 1–6. Available online at: http://proceedings.mlr.press/v90/yang18a.html (accessed June 2, 2018).
13. Yang X, Bian J, Fang R, Bjarnadottir RI, Hogan WR, Wu Y. Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting. *J Am Med Inform Assoc JAMIA*. (2020) 27:65–72. doi: 10.1093/jamia/ocz144
14. Yu Z, Yang X, Dang C, Wu S, Adekkanattu P, Pathak J, et al. *A Study of Social and Behavioral Determinants of Health in Lung Cancer Patients Using Transformers-based Natural Language Processing Models*. arXiv210804949 Cs (2021). Available online at: http://arxiv.org/abs/2108.04949 (accessed September 15, 2021).
15. Devlin J, Chang MW, Lee K, Toutanova K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Available online at: https://arxiv.org/abs/1810.04805 (accessed October 30, 2018).
16. Wolf T, Debut L, Sanh V, Chaumond J, Delangue C, Moi A, et al. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv191003771 Cs. Available online at: http://arxiv.org/abs/1910.03771 (accessed March 5, 2021).
17. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. arXiv191201703 Cs. Available online at: http://arxiv.org/abs/1912.01703 (accessed March 5, 2021).

Author Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding institutions.

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