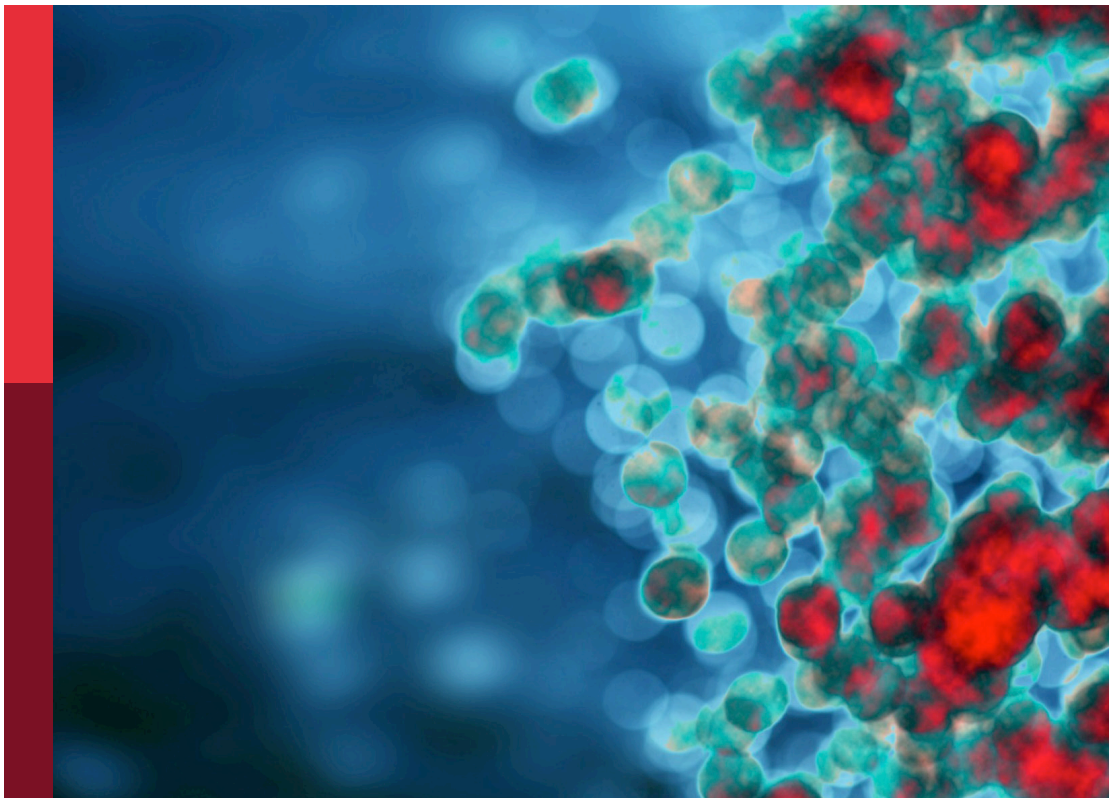
frontiers | Research Topics

Community series in towards precision medicine for immune-mediated disorders: advances in using big data and artificial intelligence to understand heterogeneity in inflammatory responses, volume II

Edited by
Alex Tsoi, Xu-jie Zhou and Yasmina Laouar

Published in
Frontiers in Immunology
Frontiers in Genetics



ISSN 1664-8714
ISBN 978-2-8325-5934-5
DOI 10.3389/978-2-8325-5934-5

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

Community series in towards precision medicine for immune-mediated disorders: advances in using big data and artificial intelligence to understand heterogeneity in inflammatory responses, volume II

Topic editors

Alex Tsoi — University of Michigan, United States

Xu-jie Zhou — Peking University, China

Yasmina Laouar — University of Michigan, United States

Citation

Tsoi, A., Zhou, X.-j., Laouar, Y., eds. (2025). *Community series in towards precision medicine for immune-mediated disorders: advances in using big data and artificial intelligence to understand heterogeneity in inflammatory responses, volume II*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5934-5

Table of contents

frontiers | Frontiers in Immunology

TYPE Editorial
PUBLISHED 13 January 2025
DOI 10.3389/fimmu.2025.1553004

OPEN ACCESS

EDITED AND REVIEWED BY
Betty Diamond,
Feinstein Institute for Medical Research,
United States

*CORRESPONDENCE
Xu-jie Zhou
zhouxujie@bjmu.edu.cn

†These authors have contributed equally to this work

RECEIVED 29 December 2024
ACCEPTED 02 January 2025
PUBLISHED 13 January 2025

CITATION
Zhou X-j, Laouar Y and Tsoi LC (2025) Editorial: Community series in towards precision medicine for immune-mediated disorders: advances in using big data and artificial intelligence to understand heterogeneity in inflammatory responses, volume II.
Front. Immunol. 16:1553004.
doi: 10.3389/fimmu.2025.1553004

Editorial: Community series in towards precision medicine for immune-mediated disorders: advances in using big data and artificial intelligence to understand heterogeneity in inflammatory responses, volume II

Xu-jie Zhou[1,2,3,4,5*†], Yasmina Laouar[6†] and Lam C. Tsoi[7,8,9,10†]

[1]Renal Division, Peking University First Hospital, Beijing, China, [2]Kidney Genetics Center, Peking University Institute of Nephrology, Beijing, China, [3]Peking University Institute of Nephrology, Key Laboratory of Renal Disease, National Health Commission, Beijing, China, [4]Key Laboratory of Chronic Kidney Disease Prevention and Treatment (Peking University), Ministry of Education, Beijing, China, [5]State Key Laboratory of Vascular Homeostasis and Remodeling, Peking University, Beijing, China, [6]Department of Microbiology and Immunology, Michigan Medicine, Ann Arbor, MI, United States, [7]Department of Dermatology, Michigan Medicine, Ann Arbor, MI, United States, [8]Department of Biostatistics, University of Michigan, Ann Arbor, MI, United States, [9]Gilbert S. Omenn Department of Computational Medicine and Bioinformatics, Michigan Medicine, Ann Arbor, MI, United States, [10]Mary H Weiser Food Allergy Center, Michigan Medicine, Ann Arbor, MI, United States

KEYWORDS

artificial intelligence, big data, heterogeneity, immune-mediated disorders, multi-omics, precision medicine

Editorial on the Research Topic

Community series in towards precision medicine for immune-mediated disorders: advances in using big data and artificial intelligence to understand heterogeneity in inflammatory responses, volume II

The advent of big data and artificial intelligence (AI) has ushered in a new era in understanding the complexities of immune-mediated disorders. This second volume of our Community Series continues to explore the frontier of precision medicine for these conditions, showcasing significant advancements in utilizing AI to comprehend inflammatory response heterogeneity and develop personalized treatment approaches (Figure 1).

AI has emerged as a powerful tool in deciphering immune system heterogeneity, offering unprecedented insights into the intricate networks of immune cells and their interactions (1, 2). Machine learning algorithms, particularly unsupervised learning techniques, have revolutionized our ability to analyze high-dimensional immunological data, identifying

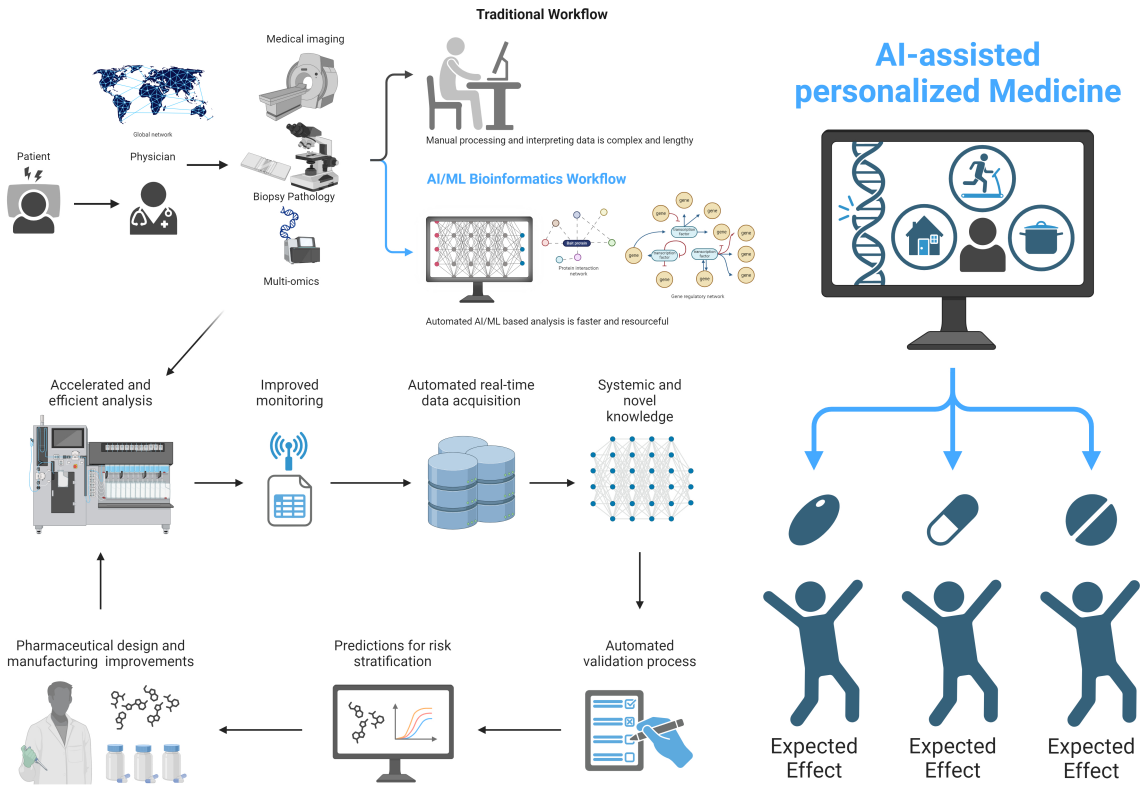


FIGURE 1
Significant advancements in utilizing AI to comprehend inflammatory response heterogeneity and develop personalized treatment approaches. This figure illustrates the transition from traditional workflow to AI-assisted personalized medicine in understanding and treating inflammatory responses. The traditional workflow is complex and time-consuming, involving manual processing and interpretation of data. AI-assisted methods can lead to more efficient and personalized approaches in medicine, particularly in the context of inflammatory responses. The expected effects of this transition include improved accuracy, faster processing times, and more tailored treatment strategies for individual patients.

distinct cell populations and characterizing their phenotypes based on marker expression. This approach enhances our understanding of immune cell subsets and their roles in diseases, paving the way for more targeted therapeutic interventions.

Machine learning for disease classification and prediction

Several studies in this Research Topic highlight the power of using machine learning (ML) to enhance disease classification and predict outcomes. Xu et al.'s work on predicting T-scores in IgA nephropathy using machine learning showcases the potential of AI in reducing the need for invasive kidney biopsies. This approach could significantly improve patient care by allowing for earlier diagnosis and more timely treatment initiation. The model's ability to predict pathological severity using routine clinical characteristics offers a valuable tool for clinicians, especially in cases where kidney biopsy is not feasible or advisable. Future research may focus on prospective validation studies and the integration of multi-omics data to enhance the model's predictive power.

Shi et al.'s application of machine learning in rheumatoid arthritis (RA) management represents a significant step towards personalized medicine. Their models for predicting treatment responses and disease progression could guide clinicians in selecting the most effective therapies for individual patients. This approach has the potential to optimize treatment outcomes, reduce adverse events, and improve overall patient care in RA management. Future development of user-friendly interfaces and decision support tools that seamlessly incorporate AI-derived insights into clinical practice may be of special importance.

Li et al.'s novel semi-supervised convolutional neural network (sscNOVA) for predicting functional regulatory variants in autoimmune diseases addresses the challenge of limited labeled data in genomics research. This innovative approach could accelerate the identification of disease-associated variants, potentially leading to improved risk prediction and the development of targeted therapies for autoimmune conditions.

Radiomics and imaging biomarkers

The application of radiomics in ankylosing spondylitis (AS) by Hu et al. represents a significant advance in non-invasive disease phenotyping. By extracting quantitative features from MRI scans, this approach offers a more objective method for assessing hip involvement in AS. The potential for early detection of hip

involvement could lead to more timely interventions, potentially slowing disease progression and improving patient outcomes. This study demonstrates the power of AI in extracting clinically relevant information from medical imaging data, potentially reducing the need for invasive diagnostic procedures.

Large-scale genetic and epidemiological analytics

The comprehensive analysis of skin disease comorbidities by Li et al. provides valuable insights into shared pathophysiology and risk factors. This large-scale epidemiological study leverages big data to uncover patterns of disease co-occurrence, potentially guiding the development of comprehensive patient care strategies that take into account the interconnectedness of different immune-mediated conditions.

Several studies in this Research Topic employ Mendelian randomization (MR) to investigate causal relationships between immune-mediated disorders and various comorbidities (Su et al., Xie et al., Hu et al., Li and Liu, Ye et al., Bai et al.). The discovery of common genetic factors between type 1 diabetes and other autoimmune disorders may pave the way for the creation of treatments that can target multiple conditions concurrently. The investigation of drug targets for Sjögren's syndrome through multi-omics MR and colocalization analyses offers a highly promising strategy for discovering new therapeutic targets[13]. The study on efgartigimod for refractory immune-mediated necrotizing myopathy highlights the potential of targeted therapies in managing challenging immune-mediated disorders (Yang et al.). This work demonstrates the importance of translating insights from basic immunology research into clinical practice, offering hope for patients with difficult-to-treat conditions. These studies provide robust evidence for shared genetic factors and potential causal pathways, which could inform drug repurposing efforts and the development of novel therapeutic strategies. The use of MR in these studies demonstrates the power of combining genetic data with advanced statistical techniques to uncover causal relationships that may not be apparent through traditional observational studies.

Clinical implications and translational value

The research presented in this Research Topic demonstrates the immense potential of big data and AI in advancing precision medicine for immune-mediated disorders. From improving diagnostic accuracy and treatment selection to uncovering novel drug targets and causal relationships, these studies offer a glimpse into the future of personalized healthcare.

The ML and radiomics approaches presented have the potential to significantly enhance diagnostic accuracy and enable earlier interventions. For example, the ability to predict T-scores in IgA nephropathy without invasive biopsies could lead to more timely treatment initiation and improved patient outcomes. Similarly, the radiomics-based phenotyping of hip involvement in AS could allow for earlier detection and management of this complication, potentially preventing long-term disability.

The ML models developed for RA management offer the promise of more personalized treatment strategies. By predicting individual patient responses to diverse therapies, these models have the potential to aid clinicians in selecting the most effective treatments, minimizing adverse events, and ultimately enhancing overall outcomes. This approach is consistent with the aims of precision medicine and could significantly elevate the quality of care for RA patients.

The large-scale epidemiological analysis of skin disease comorbidities provides a foundation for improved risk stratification. Clinicians could utilize this information to design tailored screening programs and preventive measures for patients with particular skin conditions, aiming to lower the occurrence of related comorbidities.

The MR studies in this Research Topic offer valuable insights into potential drug targets and repurposing opportunities. The discovery of common genetic factors between different autoimmune disorders may pave the way for the creation of treatments that can target multiple conditions concurrently, potentially revolutionizing the treatment landscape for immune-mediated diseases.

Challenges and future directions

Despite these advances, significant challenges remain in translating AI-driven approaches into clinical practice. The development of standardized protocols for data collection, processing, and integration across multiple modalities (e.g., imaging, genomics, clinical data) is crucial for the widespread acceptance of AI-driven approaches. Moreover, enhancing the interpretability of complex AI models is essential for fostering trust among clinicians and enabling their seamless integration into clinical decision-making frameworks.

To address these challenges, future research should focus on: *i)* Developing robust, externally validated AI models that incorporate diverse data types and account for population heterogeneity. *ii)* Creating interpretable AI algorithms that provide clinicians with clear rationales for their predictions and recommendations. *iii)* Designing user-friendly interfaces and clinical decision support tools that integrate AI-derived insights into existing clinical workflows. *iv)* Conducting large-scale, prospective clinical trials to demonstrate the real-world impact of AI-driven approaches on patient outcomes. *v)* Exploring the potential of federated learning and other privacy-preserving techniques to enable collaborative research while protecting patient data.

As we look forward to the third volume of this series, we anticipate studies that address these challenges and push the boundaries of AI applications in immunology. We further encourage submissions that integrate multi-omics data with clinical information to develop more comprehensive predictive models, explore the use of explainable AI techniques to enhance the interpretability of complex immunological models, and investigate the application of AI in real-time monitoring and prediction of immune responses, particularly in the context of immunotherapies. Additionally, we welcome research that

develops AI-driven approaches for personalized vaccine design and optimization, as well as studies that explore the use of AI in deciphering the complex interactions between the immune system and the microbiome. These areas of focus represent exciting frontiers in the application of AI to immunology and have the potential to significantly advance our understanding and treatment of immune-mediated disorders.

In conclusion, the studies presented in this Research Topic demonstrate the immense potential of AI in advancing our understanding of immune-mediated disorders and developing personalized treatment approaches. By harnessing the power of big data and AI, we are moving closer to realizing the promise of precision medicine in immunology, ultimately improving patient outcomes and quality of life. The advancement of precision medicine in immune-mediated disorders is evident, driven by big data and AI. Through sustained investment in these methodologies and overcoming forthcoming obstacles, we envision a future where customized, efficacious therapies for immune-mediated disorders are commonplace, not rare instances.

Author contributions

X-JZ: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. YL: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. LT: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. Support was provided by National Science Foundation of China (82370709); and the Joint Institute (JI) Collaboration Scholars Program at the University of Michigan Medical School. LCT was supported by the National Institutes of Health (NIH) grants R01AR080662 and UC2AR081033.

Acknowledgments

We acknowledge all the authors that have contributed to this Research Topic. The figure was created with BioRender.com.

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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

1. Zhou XJ, MacLeod AS, Tsoi LC. Editorial: advances in using big data and artificial intelligence to understand heterogeneity in inflammatory responses. *Front Immunol.* (2022) 13:948885. doi: 10.3389/fimmu.2022.948885

2. Zhou XJ, Zhong XH, Duan LX. Integration of artificial intelligence and multi-omics in kidney diseases. *Fundam Res*. (2023) 3:126–48. doi: 10.1016/j.fmre.2022.01.037

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 04 August 2023
DOI 10.3389/fimmu.2023.1224631

OPEN ACCESS

EDITED BY
Huji Xu,
Tsinghua University, China

REVIEWED BY
Youhua Xu,
Macau University of Science and Technology, Macao SAR, China
Jinxia Zhao,
Peking University Third Hospital, China

*CORRESPONDENCE
Xu-jie Zhou
zhouxujie@bjmu.edu.cn

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

RECEIVED 18 May 2023
ACCEPTED 21 July 2023
PUBLISHED 04 August 2023

CITATION
Xu L-L, Zhang D, Weng H-Y, Wang L-Z, Chen R-Y, Chen G, Shi S-F, Liu L-J, Zhong X-H, Hong S-D, Duan L-X, Lv J-C, Zhou X-J and Zhang H (2023) Machine learning in predicting *T*-score in the Oxford classification system of IgA nephropathy. *Front. Immunol.* 14:1224631. doi: 10.3389/fimmu.2023.1224631

Machine learning in predicting *T*-score in the Oxford classification system of IgA nephropathy

Lin-Lin Xu[1†], Di Zhang[2,3,4†], Hao-Yi Weng[2,3,4], Li-Zhong Wang[2,3,4], Ruo-Yan Chen[2,3,4], Gang Chen[2,3,4], Su-Fang Shi[1], Li-Jun Liu[1], Xu-Hui Zhong[5], Shen-Da Hong[6], Li-Xin Duan[7], Ji-Cheng Lv[1], Xu-Jie Zhou[1*] and Hong Zhang[1]

[1]Renal Division, Peking University First Hospital, Kidney Genetics Center, Peking University Institute of Nephrology, Key Laboratory of Renal Disease, Ministry of Health of China, Key Laboratory of Chronic Kidney Disease Prevention and Treatment, Peking University, Ministry of Education, Beijing, China, [2]Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China, [3]WeGene, Shenzhen Zaozhidao Technology, Shenzhen, China, [4]Shenzhen WeGene Clinical Laboratory, Shenzhen, China, [5]Department of Pediatrics, Peking University First Hospital, Beijing, China, [6]Institute of Medical Technology, Health Science Center of Peking University, Beijing, China, [7]The Sichuan Provincial Key Laboratory for Human Disease Gene Study, Research Unit for Blindness Prevention of Chinese Academy of Medical Sciences (2019RU026), Sichuan Academy of Medical Sciences and Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, China

Background: Immunoglobulin A nephropathy (IgAN) is one of the leading causes of end-stage kidney disease (ESKD). Many studies have shown the significance of pathological manifestations in predicting the outcome of patients with IgAN, especially *T*-score of Oxford classification. Evaluating prognosis may be hampered in patients without renal biopsy.

Methods: A baseline dataset of 690 patients with IgAN and an independent follow-up dataset of 1,168 patients were used as training and testing sets to develop the pathology *T*-score prediction (T_{pre}) model based on the stacking algorithm, respectively. The 5-year ESKD prediction models using clinical variables (base model), clinical variables and real pathological *T*-score (base model plus T_{bio}), and clinical variables and T_{pre} (base model plus T_{pre}) were developed separately in 1,168 patients with regular follow-up to evaluate whether T_{pre} could assist in predicting ESKD. In addition, an external validation set consisting of 355 patients was used to evaluate the performance of the 5-year ESKD prediction model using T_{pre}.

Results: The features selected by AUCRF for the T_{pre} model included age, systolic arterial pressure, diastolic arterial pressure, proteinuria, eGFR, serum IgA, and uric acid. The AUC of the T_{pre} was 0.82 (95% CI: 0.80–0.85) in an independent testing set. For the 5-year ESKD prediction model, the AUC of the base model was 0.86 (95% CI: 0.75–0.97). When the T_{bio} was added to the base model, there was an increase in AUC [from 0.86 (95% CI: 0.75–0.97) to 0.92 (95% CI: 0.85–0.98); P = 0.03]. There was no difference in AUC between the base model plus T_{pre} and the base model plus T_{bio} [0.90 (95% CI: 0.82–0.99) *vs.* 0.92 (95% CI: 0.85–0.98), P =

0.52]. The AUC of the 5-year ESKD prediction model using T_{pre} was 0.93 (95% CI: 0.87–0.99) in the external validation set.

Conclusion: A pathology *T*-score prediction (T_{pre}) model using routine clinical characteristics was constructed, which could predict the pathological severity and assist clinicians to predict the prognosis of IgAN patients lacking kidney pathology scores.

KEYWORDS

IgA nephropathy, machine learning, Oxford classification system, prediction model, end-stage kidney disease

1 Introduction

Immunoglobulin A (IgA) nephropathy (IgAN) is one of the most common forms of glomerulonephritis worldwide. The clinical manifestations are heterogeneous, ranging from asymptomatic proteinuria or microscopic hematuria to rapid deterioration in kidney function (1). It was reported that approximately 20%–30% of patients with IgAN would progress to kidney failure within 20 years (2). Therefore, early identification of high-risk patients with IgAN prone to ESKD is beneficial for early intervention in delaying disease progression. Great endeavors have been taken by many researchers to search for the risk factors for developing ESKD in patients with IgAN. Generally accepted risk factors affecting the progression of IgAN included decreased glomerular filtration rate (GFR), 24-h proteinuria >1 g/day, hypertension, and renal pathological manifestations (3–9). These risk factors have been used to build various scoring models for predicting the prognosis of IgAN based on traditional statistical methods (4, 10–14). However, these scoring models are constructed by the small sample sizes and different pathological scoring criteria, which may affect the accuracy and generalization of these scoring models. Moreover, the interactions between the characteristics and their effect on ESKD, the non-linear relationship among predictors, and the effects of therapeutic regimens make the interpretation of the data more complicated.

Machine learning, as a branch discipline of artificial intelligence, has obvious advantages in processing high-dimensional and sparse data. Machine learning algorithms can learn the relationship between input features and target outcomes as well as the relationship between features through a large amount of training data. Several studies have successfully constructed ESKD prediction models for patients with IgAN through machine learning algorithms (15–20). By comparing the performance of traditional statistical methods and different machine learning algorithms in predicting ESKD or halving of estimated glomerular filtration rate from baseline, Chen et al. showed that the XGBoost algorithm performed best (16). XGBoost, as a machine learning algorithm, assembles the weak prediction models to construct a prediction model (16, 21). Several studies have tried to construct event prediction models for a specific clinical outcome based on the XGBoost algorithm (22, 23). However, no matter whether it was a traditional prediction formula or a machine learning-based predictive model in IgAN, pathology scores showed consistently significant weighting among many parameters (15, 16, 19, 24). In 2009, the Oxford classification, an international consensus, was proposed to classify IgA nephropathy based on histopathological features to predict its prognosis and guide clinical treatment. The revised Oxford classification in 2017 divided IgAN into five categories, namely, "(1) mesangial hypercellularity (M); (2) endocapillary hypercellularity (E); (3) segmental glomerulosclerosis (S); (4) tubular atrophy/interstitial fibrosis (T); (5) cellular/fibrocellular crescents (C)" (25), which were shown to be the independent predictors in predicting renal outcome (24, 26). Since 2009, over 20 validation studies have tried to prove the predictive value of the MEST scores in some retrospective cohorts of patients with IgAN, which provided consistent evidence that the mesangial hypercellularity (M), segmental glomerulosclerosis (S), and tubular atrophy/interstitial fibrosis (T) each reliably provided prognostic value by univariate analysis (26), but T lesion was suggested to be the strongest predictor of renal survival. Hernan et al. summarized the results of these studies and found that M was of independent prognostic value in 5 out of 19, E in 4 out of 19, S in 7 out of 19, and T in 13 out of 19 (26). The *C*-score was adopted in the revised classification system in 2017, and three of the five prognostic studies on IgA nephropathy showed that *C*-score was associated with poor prognosis (26–28). In the constructed IgAN prognosis prediction models, it was observed that the T lesions showed greater weight in predicting prognosis compared with many other clinical and pathological parameters (14, 16). For example, in the prognosis prediction model constructed by Chen et al., there were three indexes that can be integrated to predict ESKD, namely, *T*, global sclerosis, and urine protein, among which the *T*-score ranked first in the weight of importance (16). However, the *T*-score is derived from the kidney biopsy, an invasive manipulation, sometimes refused by patients and cannot be repeated in clinical routine for detecting disease progression. Hence, it is of great significance to explore whether pathological T lesions can be predicted by the patient's clinical variables at the same time.

The purposes of our study are 1) to construct a pathology *T*-score (T_{pre}) prediction model based on the patient's clinical variables at the same time which may be able to predict whether

there is a pathological T lesion and 2) to evaluate whether the predicted T can be used to assist in predicting ESKD.

2 Methods

2.1 Study participants

This study had two independent datasets. Dataset 1, a baseline dataset without follow-up data, comprised 690 patients with IgAN. These patients received the kidney biopsy in our center but returned to local for follow-up. Dataset 2, a follow-up dataset (PKU-IgAN cohort), included 1,808 patients with IgAN who were registered and with long-term follow-up in the Peking University First Hospital IgAN database from 1997 to 2020 (29). All patients with IgAN were diagnosed based on the histologic and immunofluorescence study of the renal biopsy, and those with <8 glomeruli per biopsy section were excluded (29). After excluding 243 patients without blood lipid data, 28 patients presented at younger than 16 years of age, and 14 patients presented acute kidney failure, 1,523 patients in dataset 2 were finally enrolled in this study, consisting of 1,168 patients with Oxford MEST-C scores and 355 patients lacking Oxford MEST-C scores.

Finally, a total of 690 patients in dataset 1 and 1,168 patients with Oxford MEST-C scores in dataset 2 were enrolled in our study as the modeling group, and 355 patients without Oxford MEST-C scores in dataset 2 were enrolled in this study as the external validation group (Figure 1).

All clinical characteristics were collected at the time of the renal biopsy. The estimated glomerular filtration rate (eGFR) was calculated using the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) formula (30). Renal biopsies were categorized according to established criteria for the Oxford MEST-C scoring system (24, 26, 31). Mean arterial pressure (MAP, mm Hg) was defined as diastolic pressure plus a third of the pulse pressure. The end-stage kidney disease (ESKD) was defined as eGFR <15 ml/min/1.73 m^2, dialysis, or kidney transplantation. Our study was approved by the Ethics Committee of Peking University First Hospital (IRB number 2020Y197). Written informed consent was provided by all participants.

2.2 Pathology *T*-score prediction model

The pathology *T*-score prediction (T_{pre}) model, constructed by the stacking algorithm, was used to predict whether IgAN patients would have T lesions (yes or no). The stacking algorithm is an integrated machine learning algorithm that can summarize several

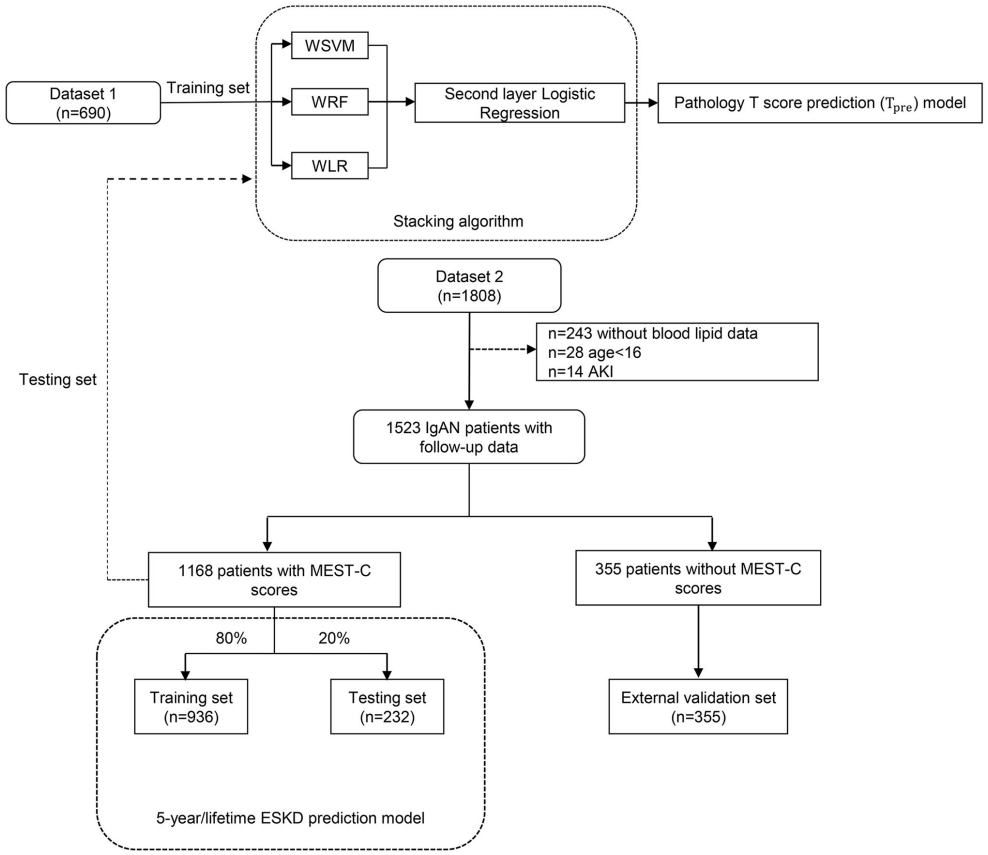


FIGURE 1
The flowchart of this study. WSVM, weighted support vector machine; WRF, weighted random forest; WLR, weighted logistic regression; AKI, acute kidney injury.

models and predict new observations. It utilizes the prediction of a collection of models as input for training a second-level model. This second-level model aims to find the best combination of the prediction of first-level models. Stacking can shield the capabilities of a range of well-performing models so that a better output prediction model can be achieved (32). In our study, we combined three machine learning algorithms, namely, support vector machine (SVM), random forest (RF), and logistic regression as first-level models, and then logistic regression as the second-level model to output the final probability of the binary *T*-score (with or without tubular atrophy/interstitial fibrosis, T_{pre}).

The input variables used in this model were chosen by AUCRF (33), a method using the random forest to find the optimal set for prediction. Variables entered into the AUCRF included age, sex, body mass index, systolic arterial pressure, diastolic arterial pressure, mean arterial pressure, hypertension, eGFR, proteinuria, microhematuria, history of gross hematuria, serum IgA, serum uric acid, serum triglycerides, total cholesterol, high-density lipoprotein, and low-density lipoprotein.

2.3 Five-year ESKD prediction model

Several studies have demonstrated the value of tubular atrophy/interstitial fibrosis (T) in predicting ESKD in patients with IgAN (16, 19, 24, 34, 35). To evaluate whether the predicted *T*-score could help predict ESKD and how effective it was, we constructed a 5-year ESKD prediction model based on the XGBoost algorithm. To illustrate the significance of tubular atrophy/interstitial fibrosis in predicting ESKD, we first constructed a 5-year ESKD prediction model with only clinical variables as input variables (base model). Then, the 5-year ESKD prediction model using clinical variables and the real pathological T lesions score (T_{bio}, T0 was assigned 0, T1 and T2 were assigned 1) was also developed (base model plus T_{bio}) to evaluate the additive value of atrophy/interstitial fibrosis (T) in predicting ESKD. Finally, to evaluate whether the value of T_{pre} in predicting ESKD of patients with IgAN was consistent with real pathological T lesions (T_{bio}) when the base model plus T_{bio} was trained in the training set, the T_{bio} of the testing set was replaced by the corresponding T_{pre} predicted by the pathology *T*-score prediction model and then the testing set was used to evaluate the model performance (the base model plus T_{pre}). For the base model plus T_{pre}, the purpose of training the model using real pathological *T*-score (T_{bio}) was for the model to learn the true value of T for predicting ESKD.

XGBoost is a kind of ensemble of the decision tree, whose advantages include higher-order interactions and complex non-linear relationships between the model features and the outcome (21). It has been shown to achieve impressive performance in predicting renal failure risk and provide explanations for variables by ranking their importance (16, 34). We also applied other machine learning algorithms to our data set for evaluating whether the predicted T could be used in ESKD prediction models based on different algorithms, including RF, penalized regression, artificial neural network (ANN), and SVM.

Characteristics selected by the Cox proportional hazards model were collected at the time of the renal biopsy at enrollment [age, sex, systolic arterial pressure, diastolic arterial pressure, proteinuria, eGFR, serum IgA, serum uric acid, serum triglycerides, total cholesterol, low-density lipoprotein, and history of previous use of renin–angiotensin system (RAS) inhibitors and immunosuppressants as well as pathological T lesions], whereas the binary outcome (ESKD within 5 years after diagnostic kidney biopsy, yes or no) represented the output data. For these variables, we imputed missing values to the means for continuous characteristics and the mode for categorical characteristics. Because of missing information on serum triglycerides, total cholesterol, and low-density lipoprotein in some cases, 243 patients without blood lipid data were excluded to avoid inaccuracy due to missing value filling (Figure 1).

To confirm that the T_{pre} can be used in the ESKD prediction model at multiple levels, we also constructed a lifetime ESKD prediction model based on XGBoost. The process and approach were the same as building the 5-year ESKD prediction model. The primary outcome was time-to-event ESKD. The survival time for the kidney without ESKD event was calculated from the kidney biopsy to the last follow-up.

The XGBoost was allowed to generate boosting trees at most 110 times, and the maximum depth of each tree was constrained to 5. To avoid overfitting, we further set the L2 regularization term on weights as 1 and stop training if the performance did not improve by more than 15 rounds. At last, the optimal prediction model parameters and architectures were selected by the five-fold cross-validation.

The patients of dataset 2 without Oxford MEST-C scores combined with the corresponding T_{pre} were used as an additional external validation set to evaluate the performance of the ESKD prediction model using T_{pre}.

2.4 Statistical analysis

The sociodemographic and clinical variables were calculated and expressed as the mean ± standard deviation for variables with approximately symmetrical distributions and as median (interquartile range 25th–75th percentile) for variables with skewed distribution. All categorical variables are expressed as frequencies and percentages. Univariate analyses based on the Cox proportional hazards model (36) were conducted to evaluate the association between the baseline clinical characteristics and ESKD event. Clinical characteristics associated with ESKD event in univariate analysis ($P < 0.05$) or if they were clinically relevant were used as input features of the 5-year ESKD prediction model.

For predicting 5-year ESKD status (yes or no) and *T*-score (0 or 1), the performance of the models was assessed by calculating the accuracy, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve (AUC). For predicting lifetime ESKD risk, we quantify the performance of the model by concordance statistic (*C*-statistic), which is a general concept of the area under the curve (AUC) for time-to-event survival data (37).

The *C*-statistic compares the rank of predicting probability and the rank of the survival time in the real world. The calibration ability of the models was assessed by the Hosmer–Lemeshow test and calibration scatter plot, in which *P*-value >0.05 indicated no very significant difference between the predicted probability predicted by the model and the true outcome frequencies during a certain time period. SPSS version 26.0 software and R 3.6.3 were used for the statistical analysis. All *P*-values were two-tailed, and $P < 0.05$ was considered statistically significant.

3 Results

3.1 Characteristics of the study participants

The clinical characteristics of 690 patients with IgAN in dataset 1 are shown in Table 1. The mean age of these patients was 32.38 ± 11.32 years at the time of renal biopsy. The male-to-female ratio was 1.2:1. The mean arterial pressure was 94.44 ± 14.02 mm Hg. The median value of eGFR was 84.66 (range, 63.32–107.50) ml/min per 1.73 m^2, and daily proteinuria was 1.38 (range, 0.66–2.89) g/day.

For the 1,168 follow-up patients with Oxford MEST-C scores in dataset 2, the mean age was 35.10 ± 11.73 years at the time of renal biopsy. The male-to-female ratio was 1:1. The mean arterial pressure was 93.59 ± 11.42 mm Hg. The eGFR was 85.91 (range, 60.94–107.23) ml/min per 1.73 m^2, and daily proteinuria was 1.27 (range, 0.66–2.45) g/day (Table 1). For the variables used to train the pathology *T*-score prediction (T_{pre}) model, there were no statistically significant differences in clinical parameters between dataset 1 and dataset 2 except for age (32.38 ± 11.32 *vs.* 35.10 ± 11.73, $P = 1.00 \times 10^{-6}$), serum IgA level (3.13 ± 1.21 *vs.* 3.29 ± 1.20, $P = 0.01$), and serum uric acid level (347.10 ± 114.95 *vs.* 367.63 ± 101.86, $P = 1.52 \times 10^{-4}$). Among these, 158 patients (13.53%) had reached the event of ESKD during the median 67.5-month follow-up. The unadjusted hazard ratios (HRs) between the different variables and ESKD are reported in Table 2. The risk of ESKD significantly increased for every 10.0 mm Hg increase in the MAP [HR: 1.34, 95% confidence interval (CI): 1.18–1.53, $P = 1.10 \times 10^{-5}$] and increased for every 1.0 g/day in the daily proteinuria (HR: 1.10, 95% CI: 1.05–1.15, $P = 1.60 \times 10^{-5}$). For each ml/min per 1.73 m^2 decrease in eGFR, the risk of ESKD increased by 4% (HR: 0.96, 95% CI: 0.96–0.97, $P = 1.24 \times 10^{-27}$). For each mg/dl increase in uric acid, the risk of ESKD increased by 38% (HR: 1.38, 95% CI: 1.29–1.49, $P = 1.47 \times 10^{-19}$). Moreover, there was the strongest association between the risk of ESKD and the presence of tubulointerstitial lesions (HR: 3.34, 95% CI: 2.73–4.07, $P = 1.72 \times 10^{-32}$).

TABLE 1 Baseline characteristics of patients with IgAN enrolled in this study to construct the pathology *T*-score prediction model at the time of kidney biopsy.

Characteristics	Training set (dataset 1)	Testing set (dataset 2 with MEST-C scores)	*P*-value
Patients (*n*)	690	1,168	
Age at biopsy, years	32.38 ± 11.32	35.10 ± 11.73	1.00×10^{-6}
Sex (male/female)	370/320	583/585	0.12
Systolic blood pressure, mm Hg	124.77 ± 18.28	123.67 ± 15.09	0.18
Diastolic blood pressure, mm Hg	79.28 ± 13.11	78.54 ± 11.00	0.22
Mean arterial pressure, mm Hg	94.44 ± 14.02	93.59 ± 11.42	0.17
eGFR, ml/min per 1.73 m^2	84.66 (63.32–107.50)	85.91 (60.94–107.23)	0.69
Proteinuria, g/day	1.38 (0.66–2.89)	1.27 (0.66–2.45)	0.10
Serum IgA level, g/l	3.13 ± 1.21	3.29 ± 1.20	0.01
Uric acid, μmol/l	347.10 ± 114.95	367.63 ± 101.86	1.52×10^{-4}
Triglycerides, mmol/l	1.61 (1.10–2.38)	1.62 (1.07–2.42)	0.64
Total cholesterol, mmol/l	4.70 (3.99–5.61)	4.77 (4.02–5.67)	0.23
Low-density lipoprotein, mmol/l	2.71 (2.12–3.33)	2.75 (2.23–3.38)	0.19
Renal biopsy, *n*/*n* (%)			
Mesangial (M) 1	560/690 (81.16%)	461/1,168 (39.47%)	3.37×10^{-68}
Endocapillary (E) 1	128/690 (18.55%)	400/1,168 (34.25%)	4.23×10^{-13}
Glomerular sclerosis (S) 1	225/690 (32.61%)	733/1,168 (62.76%)	3.33×10^{-36}
Tubulointerstitial damage (T1+T2)	182/690 (26.38%)	392/1,168 (33.56%)	1.00×10^{-3}

Data are expressed as mean ± SD, median (interquartile range), absolute, and percent frequency.
IgAN, immunoglobulin A nephropathy; eGFR, estimated glomerular filtration rate.

10.3389/fimmu.2023.1224631

TABLE 2 Risk estimated by Cox proportional hazard model for ESKD in patients of dataset 2 with Oxford MEST-C scores.

Risk factor	Non-ESKD (*n* = 1,010)	ESKD (*n* = 158)	*P*-value	HR (95% CI)
Age, years	35.21 ± 11.84	34.41 ± 11.02	0.55	1.00 (0.98–1.01)
Male (%)	482 (47.72%)	101 (63.92%)	1.96×10^{-4}	1.85 (1.34–2.57)
Systolic arterial pressure, mm Hg	123.00 ± 14.70	128.02 ± 16.77	3.00×10^{-6}	1.02 (1.01–1.03)
Diastolic arterial pressure, mm Hg	78.13 ± 10.66	81.18 ± 12.71	2.88×10^{-4}	1.03 (1.01–1.04)
Mean arterial pressure, mm Hg	93.09 ± 11.06	96.80 ± 13.10	1.10×10^{-5}	1.03 (1.02–1.04)
Proteinuria, g/day	1.17 (0.61–2.28)	1.99 (1.15–3.56)	1.60×10^{-5}	1.10 (1.05–1.15)
eGFR, ml/min per 1.73 m^2	89.13 (66.05–110.14)	53.69 (37.47–85.39)	1.24×10^{-27}	0.96 (0.96–0.97)
Serum IgA level, g/l	3.30 ± 1.22	3.17 ± 0.99	0.27	0.93 (0.81–1.06)
Uric acid, μmol/l	358.08 ± 97.12	429.22 ± 110.27	1.47×10^{-19}	1.01 (1.00–1.01)
Triglycerides, mmol/l	1.59 (1.06–2.37)	1.85 (1.13–2.69)	0.01	1.12 (1.03–1.23)
Total cholesterol, mmol/l	4.77 (4.03–5.66)	4.76 (3.99–5.84)	0.72	1.02 (0.93–1.11)
Low-density lipoprotein, mmol/l	2.75 (2.24–3.36)	2.75 (2.19–3.59)	0.77	1.02 (0.90–1.16)
Renal biopsy				
M0/M1	636/374 (62.97%/37.03%)	71/87 (44.94%/55.06%)	2.00×10^{-6}	2.14 (1.56–2.93)
E0/E1	668/342 (66.14%/33.86%)	100/58 (63.29%/36.71%)	0.36	1.16 (0.84–1.61)
S0/S1	402/608 (39.80%/60.20%)	33/125 (20.89%/79.11%)	2.10×10^{-5}	2.31 (1.57–3.39)
T0/T1+T2	723/287 (71.58%/28.42%)	53/105 (33.54%/66.46%)	1.72×10^{-32}	3.34 (2.73–4.07)
C0/C1+C2	421/589 (41.68%/58.32%)	54/104 (34.18%/65.82%)	1.00×10^{-3}	1.47 (1.17–1.86)
Therapy				
Renin–angiotensin system blocks	960 (95.05%)	151 (95.57%)	0.14	0.57 (0.26–1.21)
Corticosteroids/cytotoxic drugs	474 (46.93%)	107 (67.72%)	3.60×10^{-5}	2.02 (1.45–2.82)
Follow-up, months	67.50 (37.75–105.25)	67.50 (38.00–97.25)		

Data are expressed as mean ± SD, median (interquartile range), absolute, and percent frequency.
ESKD, end-stage kidney disease; CI, confidence interval; HR, hazard ratio; eGFR, estimated glomerular filtration rate.

3.2 Performance of the pathology *T*-score prediction model

Feature reductions were conducted using the AUCRF algorithm, which was used to select the optimal random forest model with the least number of predictive variables to predict the presence or absence of T lesions. Clinical variables with a probability of selection higher than 0.7 were selected in repeated cross-validation of the optimal random forest model (optimal AUC = 0.82). Finally, the features selected by AUCRF for the T prediction model included age, systolic arterial pressure, diastolic arterial pressure, proteinuria, eGFR, serum IgA, and uric acid (Figure 2). The 690 IgAN patients with Oxford MEST-C scores in dataset 1 as the training set were taken to develop a pathology *T*-score prediction model. The 1,168 IgAN patients with Oxford MEST-C scores in dataset 2 as the testing set were used only for reporting the performance of the model and were not used for development or fine-tuning.

If a predictive model has an AUC of higher than 0.75, it will be considered to have a good discriminating ability. The pathology T prediction model achieved a discrimination of 0.82 (95% CI: 0.80–0.85) [area under the receiver operating characteristic (ROC) curve (AUC)] in the testing set (Figure 3A). The ROC curve had 0.74 sensitivity and 0.77 specificity, which indicated that it had better clinical utility.

3.3 Performance of the 5-year ESKD prediction model

The unadjusted Cox regression analysis suggested that sex, systolic arterial pressure, diastolic arterial pressure, proteinuria, eGFR, uric acid, triglycerides, and tubular atrophy/interstitial fibrosis (T) were risk factors for developing ESKD (Table 2). A study supported elevated serum IgA as a causal factor in IgA nephropathy through Mendelian randomization (38). Some studies have suggested the association between the poor prognosis of renal disease and dyslipidemia. Higher triglycerides and cholesterol levels have been proven to be independent risk factors for the progression of kidney disease (39). Hence, clinical variables (age, sex, systolic arterial pressure, diastolic arterial pressure, proteinuria, eGFR, serum IgA, uric acid, triglycerides, total

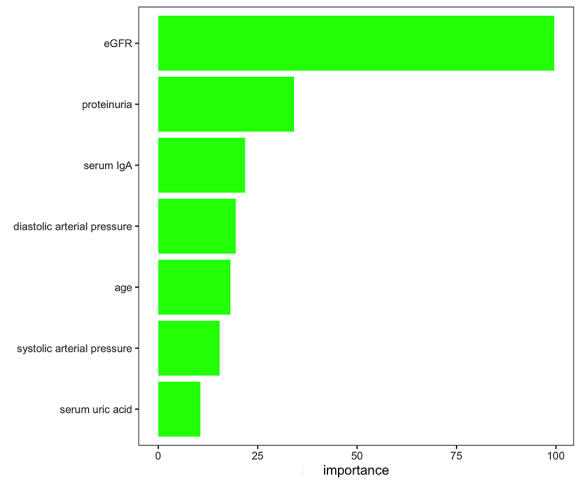


FIGURE 2
Variables selected by AUCRF for the pathology *T*-score prediction model. The importance scores of the clinical variables with a probability of selection higher than 0.7 in repeated cross-validation of the optimal random forest model to predict the presence or absence of T lesions.

cholesterol, low-density lipoprotein, history of previous use of RAS inhibitors and immunosuppressants) and the pathology T lesions (T_{bio}, T0 was assigned 0, T1 and T2 were assigned 1) were used as the input variables of the 5-year ESKD prediction model.

To make the predictive model achieve a good performance, the 1,168 follow-up IgAN patients with Oxford MEST-C scores in dataset 2 were randomly divided into training and testing sets at a ratio of 8:2. The training set included 936 patients and the testing set included 232 patients. The training set was used to perform five-fold cross-validation to select the optimal prediction model. The testing set was used to assess the performance.

The performance value of the 5-year ESKD prediction model using only the above clinical variables as input variables (base model) was 0.86 (95% CI: 0.75–0.97) in the test set (Figure 3B). To test whether the T_{bio} could improve the predictive performance of the 5-year ESKD prediction model, we added T_{bio} to the base model. An increase in AUC [from 0.86 (95% CI: 0.75–0.97) to 0.92 (95% CI: 0.85–0.98); P = 0.03] showed a better discriminating ability, which indicated that the T was important for judging the prognosis of patients with IgAN (Figure 3B). To test whether T_{pre} had a similar effect on judging the prognosis of IgAN patients, after training the 5-year ESKD prediction model with the training set, we replaced the T_{bio} in the testing set with the corresponding T_{pre} to see the discrimination effect. The AUC was 0.90 (95% CI: 0.82–0.99) in the testing set (Figure 3B). The performance of the base model plus T_{pre} did not differ from that of the base model plus T_{bio} [AUC for the base model plus T_{pre} 0.90 (95% CI: 0.82–0.99) *vs.* AUC for the base model plus T_{bio} 0.92 (95% CI: 0.85–0.98), P = 0.52, Table 3], which showed that the value of the T_{pre} in predicting the ESKD of patients was comparable to that of T_{bio}. The calibration of the three prediction models is shown in Figures 4A–C. The *P*-values for the Hosmer–Lemeshow test of the base model, the base model plus T_{bio}, and the base model plus T_{pre} were 0.42, 0.79, and 0.92, respectively, which indicated that these models had a good calibration. These results suggested the importance of T in predicting ESKD, and T_{pre} can be used to assist clinicians in assessing the prognosis of patients without pathology reports.

Table 4 shows the performance of the 5-year ESKD prediction model based on different machine learning algorithms in the testing set using T_{pre}. All models have good prediction performance, which

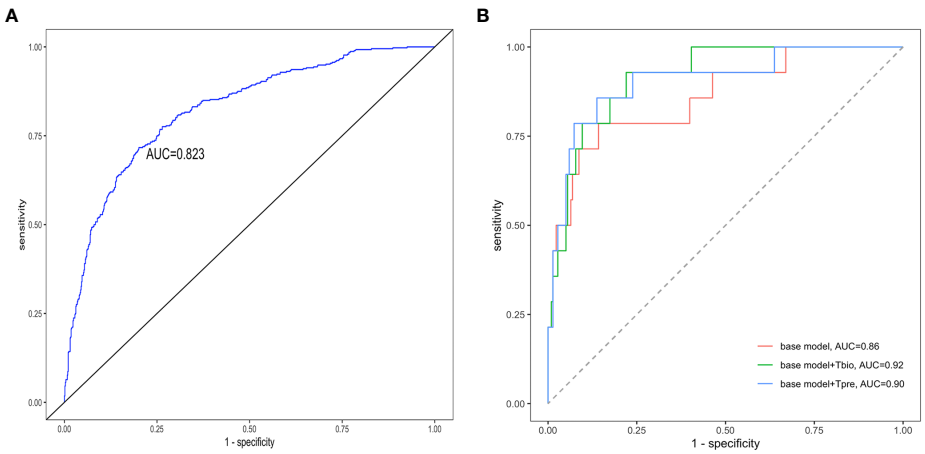


FIGURE 3
Receiver operating characteristic curves of the prediction models. The receiver operating characteristic curves for **(A)** the pathology *T*-score prediction (T_{pre}) model and **(B)** the 5-year ESKD prediction model. The base model was the 5-year ESKD prediction model based on the XGBoost algorithm with only clinical variables as input variables. The base model + T_{bio} was the 5-year ESKD prediction model based on XGBoost using clinical variables and the real pathological T lesions score (T_{bio}, T0 was assigned 0, and T1 and T2 were assigned 1). The base model + T_{pre} was when the base model plus T_{bio} was trained using clinical variables and T_{bio}, and the T_{bio} of the testing set was replaced by the corresponding T_{pre} predicted by the pathology *T*-score prediction model. The clinical variables used for the 5-year ESKD prediction model included age, sex, systolic arterial pressure, diastolic arterial pressure, proteinuria, eGFR, serum IgA, uric acid, triglycerides, total cholesterol, low-density lipoprotein, and history of previous use of renin–angiotensin system (RAS) inhibitors and immunosuppressants. AUC, area under the curve.

TABLE 3 Performance comparison for the prediction on 5-year ESKD status with different predictors in the testing subset.

Model	Accuracy	Sensitivity	Specificity	AUC
Clinical variables	0.85	0.79	0.86	0.86
Clinical variables plus T_{bio}	0.83	0.93	0.78	0.92
Clinical variables plus T_{pre}	0.86	0.86	0.86	0.90

The clinical variables include age, sex, systolic arterial pressure, diastolic arterial pressure, proteinuria, eGFR, serum IgA, uric acid, triglycerides, total cholesterol, low-density lipoprotein, and history of previous use of renin–angiotensin system (RAS) inhibitors and immunosuppressants.
T_{bio}, the real pathological T-score quantified as either 0 (absent) or 1 (T1 or T2); T_{pre}, the pathological T-score predicted by the baseline pathology T-score prediction (T_{pre}) model.

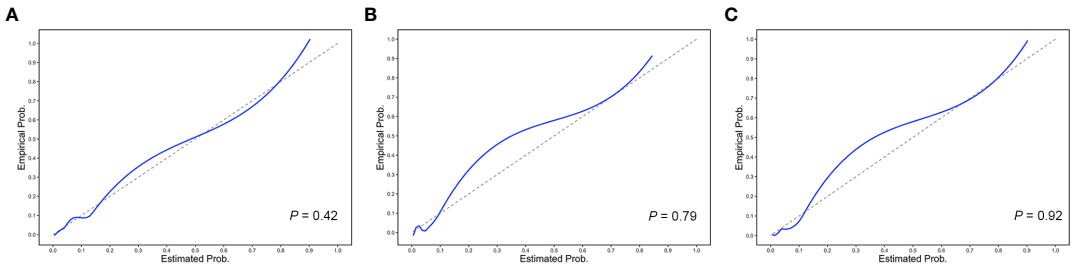


FIGURE 4
Calibration plots of the 5-year ESKD prediction models. The calibration plots for **(A)** the base model, **(B)** the base model plus T_{bio}, and **(C)** the base model plus T_{pre}. The *P*-values for the Hosmer–Lemeshow test of the base model, the base model plus T_{bio}, and the base model plus T_{pre} were 0.42, 0.79, and 0.92, respectively, which indicated that these models had a good calibration.

indicated that T_{pre} could be used in ESKD predictive models built on different algorithms.

For the lifetime ESKD prediction model based on XGBoost using only clinical variables (base model), the *C*-statistic was 0.82 (95% CI: 0.80–0.84) in the testing set. The discriminating ability of the base model plus T_{pre} was also comparable to the base model plus T_{bio} [*C*-statistic: 0.85 (95% CI: 0.83–0.86) *vs.* 0.85 (95% CI: 0.83–0.86), $P = 0.11$] in the testing set.

3.4 External validation of the ESKD prediction model using T_{pre}

The 355 patients without MEST-C scores in dataset 2 were included as the external validation population for evaluating the performance of the 5-year ESKD prediction model. Because patients did not have MEST-C scores, the T_{pre} predicted by the pathology *T*-score prediction model was used in the 5-year ESKD prediction model. The AUC of the 5-year ESKD prediction model using T_{pre} based on XGBoost was 0.93 (95% CI: 0.87–0.99). We listed the AUC of the applied other machine learning algorithms in Table 5.

In the lifetime ESKD prediction model using T_{pre}, the *C*-statistic was 0.92 (95% CI: 0.90–0.94). We have shown here that both models have a good performance in the external validation set, indicating the reliability of T_{pre} for assisting in evaluating the prognosis of IgAN.

4 Discussion

We developed a pathology *T*-score prediction (T_{pre}) model that can predict whether the patient with IgAN may have tubulointerstitial lesions at this time based on clinical variables when the patient did not undergo a renal biopsy or did not want to repeat the renal biopsy for progression assessment. We further constructed the 5-year/lifetime ESKD prediction model based on the XGBoost algorithm to confirm the importance of T in

TABLE 4 Performance of the 5-year ESKD prediction model using T_{pre} based on different machine learning algorithms in the testing set.

Model	Accuracy	Sensitivity	Specificity	AUC
XGBoost	0.86	0.86	0.86	0.90
Random forest	0.82	0.79	0.87	0.89
Penalized regression	0.80	0.93	0.80	0.88
Artificial neural network	0.78	0.86	0.77	0.86
Support vector machine	0.71	0.86	0.62	0.77

The model was trained using clinical variables and the T_{bio}, and the T_{bio} was replaced with the corresponding T_{pre} predicted by the pathology T-score prediction model in the test subset.

TABLE 5 Performance of the 5-year ESKD prediction model using T_{pre} based on different machine learning algorithms in the external validation set.

Model	Accuracy	Sensitivity	Specificity	AUC
XGBoost	0.82	1.00	0.81	0.93
Logistic regression	0.72	1.00	0.71	0.90
Artificial neural network	0.50	1.00	0.48	0.79
Support vector machine	0.87	0.67	0.87	0.74
Random forest	0.86	0.83	0.86	0.92

The characteristics used in the basic model include age, gender, SBP, DBP, eGFR, IgA, UTP, UA, TG, TCHO, LDL, history of corticosteroids/cytotoxic drugs, and renin–angiotensin system blockers.

predicting ESKD, and T_{pre} can replace the real pathological T lesions for assisting clinicians in evaluating the prognosis of IgAN patients without pathology reports. In addition, the ESKD prediction model built based on different machine learning algorithms had good discriminating ability by using clinical variables and T_{pre}, which indicated the reliability and universality of T_{pre} for assisting in evaluating the prognosis of IgAN.

For developing the pathology *T*-score (T_{pre}) prediction model, we first used the AUCRF algorithm to select the clinical variables that may be associated with the tubulointerstitial lesions. Feature selection before training the predictive model can prevent dimensional disaster, reduce training time, prevent overfitting, enhance model generalization ability, and enhance the understanding of features and feature values, which also determines the upper limit of the effect of a machine learning task. The AUCRF is based on the RF algorithm, which is used for feature reduction based on optimizing the area under the ROC curve (AUC) of the random forest (33). It was found that age, systolic arterial pressure, diastolic arterial pressure, proteinuria, eGFR, serum IgA, and uric acid may be the clinical characteristics associated with tubular atrophy/interstitial fibrosis. Mechanism studies are needed to explore the inherent causality of these correlations and predictive capability. There have been reports indicating the association between reduced initial eGFR, higher initial MAP, proteinuria, and tubular atrophy/interstitial fibrosis (31). Next, we used the stacking algorithm to construct the pathology *T*-score prediction (T_{pre}) model based on the clinical characteristics selected by the AUCRF. A single learner has over- or underfitting problems, and to obtain a learner with excellent generalization performance, we can train multiple individual learners to form a strong learner through a certain combination strategy. This method of integrating multiple individual learners is called ensemble learning. Stacking is one of the methods of ensemble learning. The advantage of integration is that different models can learn different features of the data, and the results after fusion tend to perform better (40). As our results showed, when we used an independent dataset as the testing set, the AUC of the pathological *T*-score prediction (T_{pre}) model reached 0.82, which indicates the good discriminating ability of this T_{pre} prediction model.

A host of studies have indicated that pathological T lesions play an important role in predicting prognosis (14, 35, 41). At the same time, most current ESKD prediction models based on different methods or algorithms all include pathology *T*-score (14, 16, 19). Nevertheless, a renal puncture is invasive, which may cause a series of complications and has a host of contraindications, such as severe hypertension, coagulation disorders, solitary kidney, and so on (42). Furthermore, the number of patients at high risk of renal puncture may increase in the near future because of the aging of the population and the increased use of anticoagulant medication (43). For the patients who lack the report of kidney biopsy or do not want to undergo repeat renal puncture for disease progression assessment and evaluation of the effect of drug therapy, the clinician could not assess the prognosis of these patients with IgAN by using the established ESKD prediction model. The pathology *T*-score prediction (T_{pre}) model we developed may solve this problem. We also constructed a 5-year/lifetime ESKD prediction model based on XGBoost to assess whether the value of T_{pre} in predicting ESKD of patients with IgAN was consistent with real pathological *T*-score. The performance of the base model plus T_{pre} was similar to the base model plus T_{bio}, which showed that the T_{pre} can replace the real pathological *T*-score for prognostic prediction.

As far as we know, this study is the first to construct a pathology *T*-score prediction model in IgA nephropathy. At the same time, it is also the first study to use a machine learning algorithm to identify clinical variables that may influence the development of tubular atrophy/interstitial fibrosis, which may be useful for assessing the prognosis and targeted medication guidance. However, there is a limitation in our study. The model has been developed and tested in a single-center cohort of patients with IgAN; therefore, multicenter prospective cohort and ethnic-based cohort studies are necessary, which will further confirm the reliability of the pathology *T*-score prediction model, expand the scope of application of the model, and provide possibilities for clinical application.

In conclusion, our pathology *T*-score prediction (T_{pre}) model is a reliable tool for predicting the presence or absence of pathological T lesions. At the same time, it can also be used to assist clinicians in predicting the prognosis of patients with IgAN. A prospective multicenter cohort study is necessary to explore the potential value and robustness of this T prediction tool in the management of IgA nephropathy.

Data availability statement

The data presented in the study are deposited in the GitHub repository (https://github.com/zhangd17-web/IGAN_MI).

Author contributions

Research idea and study design: HZ, X-JZ, LW, DZ, and LX. Data acquisition: LX, SS, X-JZ, and HZ. Data analysis/interpretation: LX, DZ, LW, HW, and X-JZ. Statistical analysis: LX and DZ. Supervision or mentorship: X-JZ, HZ, HW, LW, RC, GC, LL, SS, XZ, SH, LD, and JL. Each author contributed important intellectual content during manuscript drafting or revision and agrees to be personally accountable for the individual's own contributions and to ensure that questions pertaining to the accuracy or integrity of any portion of the work, even one in which the author was not directly involved, are appropriately investigated and resolved, with documentation in the literature if appropriate.

Funding

This work was supported by the National Science Foundation of China (82022010, 82131430172, 81970613, 82070733), Beijing Natural Science Foundation (Z190023), Academy of Medical Sciences—Newton Advanced Fellowship (NAFR13\1033), King's College London—Peking University Health Science Center Joint Institute for Medical Research (BMU2021KCL004), Fok Ying Tung Education Foundation (171030), Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (2019-I2M-5-046, 2020-JKCS-009), and National High Level Hospital Clinical Research Funding (Interdisciplinary Clinical Research Project of Peking University First Hospital, 2022CR41, 2022CR40). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank all the patients and researchers who participated in the study.

Conflict of interest

Authors DZ, HW, LW, RC and GC were employed by company WeGene.

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.

The reviewer JZ declared a shared parent affiliation with the authors LX, SS, LL, X-HZ, JL, X-JZ, and HZ to the handling editor at the time of review.

Publisher's note

References

1. Lai KN, Tang SC, Schena FP, Novak J, Tomino Y, Fogo AB, et al. IgA nephropathy. *Nat Rev Dis Primers* (2016) 2:16001. doi: 10.1038/nrdp.2016.1

2. Magistroni R, D'Agati VD, Appel GB, Kiryluk K. New developments in the genetics, pathogenesis, and therapy of IgA nephropathy. *Kidney Int* (2015) 88(5):974–89. doi: 10.1038/ki.2015.252

3. Le W, Liang S, Hu Y, Deng K, Bao H, Zeng C, et al. Long-term renal survival and related risk factors in patients with IgA nephropathy: results from a cohort of 1155 cases in a Chinese adult population. *Nephrol Dial Transplant* (2012) 27(4):1479–85. doi: 10.1093/ndt/gfr527

4. Goto M, Wakai K, Kawamura T, Ando M, Endoh M, Tomino Y. A scoring system to predict renal outcome in IgA nephropathy: a nationwide 10-year prospective cohort study. *Nephrol Dial Transplant* (2009) 24(10):3068–74. doi: 10.1093/ndt/gfp273

5. Beukhof JR, Kardaun O, Schaafsma W, Poortema K, Donker AJ, Hoedemaeker PJ, et al. Toward individual prognosis of IgA nephropathy. *Kidney Int* (1986) 29 (2):549–56. doi: 10.1038/ki.1986.33

6. Rekola S, Bergstrand A, Bucht H. Development of hypertension in IgA nephropathy as a marker of a poor prognosis. *Am J Nephrol* (1990) 10(4):290–5. doi: 10.1159/000168122

7. Radford MGJr., Donadio JVJr., Bergstralh EJ, Grande JP. Predicting renal outcome in IgA nephropathy. *J Am Soc Nephrol JASN* (1997) 8(2):199–207. doi: 10.1681/ASN.V82199

8. Chen D, Liu J, Duan S, Chen P, Tang L, Zhang L, et al. Clinicopathological features to predict progression of igA nephropathy with mild proteinuria. *Kidney Blood Pressure Res* (2018) 43(2):318–28. doi: 10.1159/000487901

9. Barbour SJ, Espino-Hernandez G, Reich HN, Coppo R, Roberts IS, Feehally J, et al. The MEST score provides earlier risk prediction in lgA nephropathy. *Kidney Int* (2016) 89(1):167–75. doi: 10.1038/ki.2015.322

10. Wakai K, Kawamura T, Endoh M, Kojima M, Tomino Y, Tamakoshi A, et al. A scoring system to predict renal outcome in IgA nephropathy: from a nationwide prospective study. *Nephrol Dial Transplant* (2006) 21(10):2800–8. doi: 10.1093/ndt/gfl342

11. Okonogi H, Utsunomiya Y, Miyazaki Y, Koike K, HIrano K, Tsuboi N, et al. A predictive clinical grading system for immunoglobulin A nephropathy by combining proteinuria and estimated glomerular filtration rate. *Nephron Clin Pract* (2011) 118(3): c292–300. doi: 10.1159/000322613

12. Xie J, Kiryluk K, Wang W, Wang Z, Guo S, Shen P, et al. Predicting progression of IgA nephropathy: new clinical progression risk score. *PloS One* (2012) 7(6):e38904. doi: 10.1371/journal.pone.0038904

13. Tanaka S, Ninomiya T, Katafuchi R, Masutani K, Tsuchimoto A, Noguchi H, et al. Development and validation of a prediction rule using the Oxford classification in IgA nephropathy. *Clin J Am Soc Nephrol CJASN* (2013) 8(12):2082–90. doi: 10.2215/CJN.03480413

14. Barbour SJ, Coppo R, Zhang H, Liu ZH, Suzuki Y, Matsuzaki K, et al. Evaluating a new international risk-prediction tool in igA nephropathy. *JAMA Internal Med* (2019) 179(7):942–52. doi: 10.1001/jamainternmed.2019.0600

15. Liu Y, Zhang Y, Liu D, Tan X, Tang X, Zhang F, et al. Prediction of ESRD in igA nephropathy patients from an asian cohort: A random forest model. *Kidney Blood Pressure Res* (2018) 43(6):1852–64. doi: 10.1159/000495818

16. Chen T, Li X, Li Y, Xia E, Qin Y, Liang S, et al. Prediction and risk stratification of kidney outcomes in igA nephropathy. *Am J Kidney Dis* (2019) 74(3):300–9. doi: 10.1053/j.ajkd.2019.02.016

17. Han X, Zheng X, Wang Y, Sun X, Xiao Y, Tang Y, et al. Random forest can accurately predict the development of end-stage renal disease in immunoglobulin a nephropathy patients. *Ann Trans Med* (2019) 7(11):234. doi: 10.21037/atm.2018.12.11

18. Konieczny A, Stojanowski J, Krajewska M, Kusztal M. Machine learning in prediction of igA nephropathy outcome: A comparative approach. *J Pers Med* (2021) 11 (4):312. doi: 10.3390/jpm11040312

19. Schena FP, Anelli VW, Trotta J, Di Noia T, Manno C, Tripepi G, et al. Development and testing of an artificial intelligence tool for predicting end-stage kidney disease in patients with immunoglobulin A nephropathy. *Kidney Int* (2021) 99 (5):1179–88. doi: 10.1016/j.kint.2020.07.046

20. Diciolla M, Binetti G, Di Noia T, Pesce F, Schena FP, Vågane AM, et al. Patient classification and outcome prediction in IgA nephropathy. *Comput Biol Med* (2015) 66:278–86. doi: 10.1016/j.compbiomed.2015.09.003

21. 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, California, USA: Association for Computing Machinery (2016). p. 785–94.

22. Khera R, Haimovich J, Hurley NC, McNamara R, Spertus JA, Desai N, et al. Use of machine learning models to predict death after acute myocardial infarction. *JAMA Cardiol* (2021) 6(6):633–41. doi: 10.1001/jamacardio.2021.0122

23. Pfaff ER, Girvin AT, Bennett TD, Bhatia A, Brooks IM, Deer RR, et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digital Health* (2022) 4(7):e532–e41. doi: 10.1016/S2589-7500(22)00048-6

24. Cattran DC, Coppo R, Cook HT, Feehally J, Roberts IS, Troyanov S, et al. The Oxford classification of IgA nephropathy: rationale, clinicopathological correlations, and classification. *Kidney Int* (2009) 76(5):534–45. doi: 10.1038/ki.2009.243

25. Rui Y, Yang Z, Zhai Z, Zhao C, Tang L. The predictive value of Oxford MEST-C classification to immunosuppressive therapy of IgA nephropathy. *Int Urol Nephrol* (2022) 54(4):959–67. doi: 10.1007/s11255-021-02974-9

26. Trimarchi H, Barratt J, Cattran DC, Cook HT, Coppo R, Haas M, et al. Oxford Classification of IgA nephropathy 2016: an update from the IgA Nephropathy Classification Working Group. *Kidney Int* (2017) 91(5):1014–21. doi: 10.1016/j.kint.2017.02.003

27. Schimpf JI, Klein T, Fitzner C, Eitner F, Porubsky S, Hilgers RD, et al. Renal outcomes of STOP-IgAN trial patients in relation to baseline histology (MEST-C scores). *BMC Nephrol* (2018) 19(1):328. doi: 10.1186/s12882-018-1128-6

28. Liu Y, Wei W, Yu C, Xing L, Wang M, Liu R, et al. Epidemiology and risk factors for progression in Chinese patients with IgA nephropathy. *Medicina clinica* (2021) 157 (6):267–73. doi: 10.1016/j.medcli.2020.05.064

29. Zhang Y, Guo L, Wang Z, Wang J, Er L, Barbour SJ, et al. External validation of international risk-prediction models of igA nephropathy in an asian-caucasian cohort. *Kidney Int Rep* (2020) 5(10):1753–63. doi: 10.1016/j.ekir.2020.07.036

30. Levey AS, Stevens LA, Schmid CH, Zhang YL, Castro AF3rd, Feldman HI, et al. A new equation to estimate glomerular filtration rate. *Ann Internal Med* (2009) 150 (9):604–12. doi: 10.7326/0003-4819-150-9-200905050-00006

31. Roberts IS, Cook HT, Troyanov S, Alpers CE, Amore A, Barratt J, et al. The Oxford classification of IgA nephropathy: pathology definitions, correlations, and reproducibility. *Kidney Int* (2009) 76(5):546–56. doi: 10.1038/ki.2009.168

32. Wolpert DH. Stacked generalization. *Neural Networks* (1992) 5(2):241–59. doi: 10.1016/S0893-6080(05)80023-1

33. Calle ML, Urrea V, Boulesteix AL, Malats N. AUC-RF: a new strategy for genomic profiling with random forest. *Hum heredity* (2011) 72(2):121–32. doi: 10.1159/000330778

34. Li Y, Chen T, Chen T, Li X, Zeng C, Liu Z, et al. An interpretable machine learning survival model for predicting long-term kidney outcomes in igA nephropathy. *AMIA Annu Symposium Proc AMIA Symposium* (2020) 2020:737–46.

35. Lv J, Shi S, Xu D, Zhang H, Troyanov S, Cattran DC, et al. Evaluation of the Oxford Classification of IgA nephropathy: a systematic review and meta-analysis. *Am J Kidney Dis* (2013) 62(5):891–9. doi: 10.1053/j.ajkd.2013.04.021

36. Prentice RL, Zhao S. Regression models and multivariate life tables. *J Am Stat Assoc* (2021) 116(535):1330–45. doi: 10.1080/01621459.2020.1713792

37. Park SH, Hahm MH, Bae BK, Chong GO, Jeong SY, Na S, et al. Magnetic resonance imaging features of tumor and lymph node to predict clinical outcome in node-positive cervical cancer: a retrospective analysis. *Radiat Oncol (London England)* (2020) 15(1):86. doi: 10.1186/s13014-020-01502-w

38. Liu L, Khan A, Sanchez-Rodriguez E, Zanoni F, Li Y, Steers N, et al. Genetic regulation of serum IgA levels and susceptibility to common immune, infectious, kidney, and cardio-metabolic traits. *Nat Commun* (2022) 13(1):6859. doi: 10.1038/s41467-022-34456-6

39. Trevisan R, Dodesini AR, Lepore G. Lipids and renal disease. *J Am Soc Nephrol JASN* (2006) 17(4 Suppl 2):S145–7. doi: 10.1681/ASN.2005121320

40. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol* (2018) 33(5):459–64. doi: 10.1007/s10654-018-0390-z

41. Myllymäki JM, Honkanen TT, Syrjänen JT, Helin HJ, Rantala IS, Pasternack AI, et al. Severity of tubulointerstitial inflammation and prognosis in immunoglobulin A nephropathy. *Kidney Int* (2007) 71(4):343–8. doi: 10.1038/sj.ki.5002046

42. Hergesell O, Felten H, Andrassy K, Kühn K, Ritz E. Safety of ultrasound-guided percutaneous renal biopsy-retrospective analysis of 1090 consecutive cases. *Nephrol Dial Transplant* (1998) 13(4):975–7. doi: 10.1093/ndt/13.4.975

43. Stiles KP, Yuan CM, Chung EM, Lyon RD, Lane JD, Abbott KC. Renal biopsy in high-risk patients with medical diseases of the kidney. *Am J Kidney Dis* (2000) 36 (2):419–33. doi: 10.1053/ajkd.2000.8998

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 08 January 2024
DOI 10.3389/fimmu.2023.1309549

OPEN ACCESS

EDITED BY
Paola Savoia,
Universitàdegli Studi del Piemonte Orientale, Italy

REVIEWED BY
Yan-na Wang,
Guangdong Provincial People's Hospital, China
Robert Swerlick,
Emory University, United States

*CORRESPONDENCE
Lam C. Tsoi
alextsoi@med.umich.edu

RECEIVED 08 October 2023
ACCEPTED 12 December 2023
PUBLISHED 08 January 2024

CITATION
Li Q, Patrick MT, Sreeskandarajan S, Kang J, Kahlenberg JM, Gudjonsson JE, He Z and Tsoi LC (2024) Large-scale epidemiological analysis of common skin diseases to identify shared and unique comorbidities and demographic factors.
Front. Immunol. 14:1309549.
doi: 10.3389/fimmu.2023.1309549

Large-scale epidemiological analysis of common skin diseases to identify shared and unique comorbidities and demographic factors

Qinmengge Li[1], Matthew T. Patrick[2], Sutharzan Sreeskandarajan[3], Jian Kang[1], J. Michelle Kahlenberg[2,4], Johann E. Gudjonsson[2], Zhi He[1] and Lam C. Tsoi[1,2,5]*

[1]Department of Biostatistics, University of Michigan, Ann Arbor, MI, United States, [2]Department of Dermatology, University of Michigan, Ann Arbor, MI, United States, [3]The Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, United States, [4]Rheumatology, Internal Medicine, University of Michigan, Ann Arbor, MI, United States, [5]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, United States

Introduction: The utilization of large-scale claims databases has greatly improved the management, accessibility, and integration of extensive medical data. However, its potential for systematically identifying comorbidities in the context of skin diseases remains unexplored.

Methods: This study aims to assess the capability of a comprehensive claims database in identifying comorbidities linked to 14 specific skin and skin-related conditions and examining temporal changes in their association patterns. This study employed a retrospective case-control cohort design utilizing 13 million skin/skin-related patients and 2 million randomly sampled controls from Optum's de-identified Clinformatics® Data Mart Database spanning the period from 2001 to 2018. A broad spectrum of comorbidities encompassing cancer, diabetes, respiratory, mental, immunity, gastrointestinal, and cardiovascular conditions were examined for each of the 14 skin and skin-related disorders in the study.

Results: Using the established type-2 diabetes (T2D) and psoriasis comorbidity as example, we demonstrated the association is significant (P-values<$1x10^{-15}$) and stable across years (OR=1.15-1.31). Analysis of the 2014-2018 data reveals that celiac disease, Crohn's disease, and ulcerative colitis exhibit the strongest associations with the 14 skin/skin-related conditions. Systemic lupus erythematosus (SLE), leprosy, and hidradenitis suppurativa show the strongest associations with 30 different comorbidities. Particularly notable associations include Crohn's disease with leprosy (odds ratio [OR] =6.60, 95% confidence interval [CI]: 3.09-14.08), primary biliary cirrhosis with SLE (OR=6.07, 95% CI: 4.93-7.46), and celiac disease with SLE (OR=6.06, 95% CI: 5.49-6.69). In addition, changes in associations were observed over time. For instance, the association between atopic dermatitis and lung cancer

demonstrates a marked decrease over the past decade, with the odds ratio decreasing from 1.75 (95% CI: 1.47-2.07) to 1.02 (95% CI: 0.97-1.07). The identification of skin-associated comorbidities contributes to individualized healthcare and improved clinical management, while also enhancing our understanding of shared pathophysiology. Moreover, tracking these associations over time aids in evaluating the progression of clinical diagnosis and treatment.

Discussion: The findings highlight the potential of utilizing comprehensive claims databases in advancing research and improving patient care in dermatology.

KEYWORDS

epidemiology, claims, skin disease, comorbidity, Optum

1 Introduction

Dermatological disorders are among the most common human diseases: more than a third of the global population suffers from some form of skin condition (1–5). While most skin disorders are not fatal, the burden on patients and society is severe; in fact, skin disorders are ranked the fourth leading cause of nonfatal disease burden globally (1). For instance, in a previous study, 60% of working patients noted significant work time lost, and 40% of non-working patients attributed their lack of work to psoriasis (6). In 1984, it was estimated that the cost for 2.3 million psoriasis outpatients in the US reached $1.5 billion per year (7), and a recent study reviewing the yearly cost for psoriasis nationwide increased the estimate to a range between $51.7 and $63.2 billion (8). Atopic dermatitis (AD) is another common skin condition that affects over 30 million patients in the US with a total annual cost of $4.2 billion in 2004 and $5.4 billion in 2016 (9). Although systemic lupus erythematosus (SLE), in which up to 70% patients exhibit skin manifestations, is relatively less common with a prevalence rate of around 10 per 10,000 in the US (10), the economic burden is significant, with a total annual cost estimated to be $13,735-$20,926 per patient (11). With these significant medical burden for the wide spectrum of dermatological disorders (12), the prevention and treatment of these conditions are critical issues for public health.

The associated comorbidities (i.e. co-occurrence of two different diseases (13)) for skin conditions contribute significantly to health and social burden. Numerous studies have found that skin disorders can be early manifestations of systemic diseases (13). Thus, it is important to assess patients' risk for having other conditions in addition to their primary skin disorder; furthermore, understanding skin-associated comorbidities can further the development of better healthcare management (14) by facilitating early diagnosis of associated systemic conditions (13). Comorbidity information can also advance the identification of shared pathophysiology and risk factors, which play an important role in preventive medicine. For instance, cardiovascular disease has been found to have a significant association with psoriasis and contributes largely to the 5-year shorter life expectancy of psoriatic patients (15). Although this connection has been well publicized, a survey conducted between 2009 to 2012 showed that many physicians were unaware of this association potentially increasing the risk of delayed diagnosis and inadequate treatment of the associated cardiovascular comorbidity (16, 17).

While small cohort studies have been conducted to identify associated demographic variables or co-occurring conditions for specific skin-diseases (4, 18, 19) and the availability of large-scale claims databases has advanced precision medicine and comorbidity identification (20), limited research has investigated the potential of using these resources to identify, in a systematic fashion, associated skin conditions and comorbidities. A prominent claims data system is Optum's de-identified Clinformatics® Data Mart Database (CDM) (21, 22), an organized medical claims database that supports large-scale retrospective cohort studies. By utilizing medical records dating from 2001 to 2018, we revealed specific/shared comorbidities for 14 different skin diseases. With the 18-year time span, the trajectory of disease-comorbidity associations was also studied (23).

Our work highlights that most of the potential skin/skin-related condition-comorbidity pairs are positively associated. We calculated the trend of the skin-comorbidity associations over time and illustrated that the association between type-2 diabetes (T2D) and psoriasis over time is significant, stable, and consistent with previously published studies, confirming the validity of using CDM data in the identification of skin/skin-related disease comorbidities. However, analysis of some disease conditions can be biased, for instance, the association between psoriatic arthritis (PsA) and rheumatoid arthritis (RA) can be inflated when using unrestricted CDM data. This observation manifests potential misdiagnosis for some disease pairs in claims data. The CDM data processing and analyses in skin disease comorbidity

identification can help inform the potentials and challenges in using large-scale claims data to study comorbidities and facilitate the development of individualized health care and optimization of clinical management.

2 Materials and methods

2.1 Data preparation

The data used in this study comes from CDM (21), a de-identified patient-level database provided by Optum, a national healthcare management company. The CDM database includes medical claims from various sources, including commercially insured patients, administrative services only patients, legacy medicare choice patients prior to 2006, and medicare advantage patients after 2006. It covers a span of 18 years, from 2001 to 2018, and includes over 63 million patients from all 50 U.S. states. However, the CDM cohort does not include patients insured by Medicaid, so the socioeconomic spectrum of the entire U.S. population is not fully represented in this dataset (22).

Our analysis focused on identifying comorbidities related to skin diseases. We began by selecting a total of 13,934,335 patients with at least one of the 14 skin conditions. These conditions were categorized into three groups: immune-mediated skin diseases (acne, rosacea, alopecia areata, vitiligo, psoriasis, atopic dermatitis, hidradenitis suppurativa, prurigo nodularis), non-immune-mediated skin diseases (aging, leprosy, pigmentation, melanoma), and skin-related disorders (systemic lupus erythematosus, psoriatic arthritis). For the control group, we randomly sampled 2 million unique patients from the entire CDM database, excluding those with any of the aforementioned 14 skin/skin-related diseases. We extracted and adjusted several demographic and socioeconomic variables for analysis, including age, sex, race, education level, income level, home ownership, and the number of adults and children in the household, to account for the higher socioeconomic sampling bias. To account for non-recorded comorbidities resulting from patients leaving the healthcare system, we also included the length of time patients stayed in the system as a covariate. In the subsequent analysis, we only included individuals with complete demographic and socioeconomic information leaving 7,553,273 patients and 726,230 controls. If a patient was diagnosed with two diseases within a 5-year time span, we considered those conditions to be co-occurring. This time range is based on empirical observations of the duration patients stay in the CDM system. We divided the full dataset into consecutive 5-year subsets (e.g., 2001-2005, 2002-2006, …, 2014-2018) and conducted separate analyses for each time interval. Figure 1 provides an overview of our study.

2.2 Statistical analysis

A descriptive analysis was performed to provide an overview of the dataset and the distribution of all covariates. Categorical variables such as sex, race, education level, home ownership, and income level were summarized as percentages for each category. Continuous/Integer variables such as age, the number of children, and the number of adults in the household were summarized as mean values with their corresponding standard deviations.

Logistic regression was employed to model the association between each skin disease and comorbidity pair while accounting for potential confounding covariates. Treating either skin/skin-related disorders or comorbidities as outcome variable can achieve this goal. Since the other aim of this work is to model the risk of skin/skin-related disorders, therefore, in the following analysis we treat skin/skin-related disorders as outcome variable and comorbidities and other demographics as predictors. Age was categorized into specific ranges (e.g., <10, 10-20, 20-30,…, 70-80,

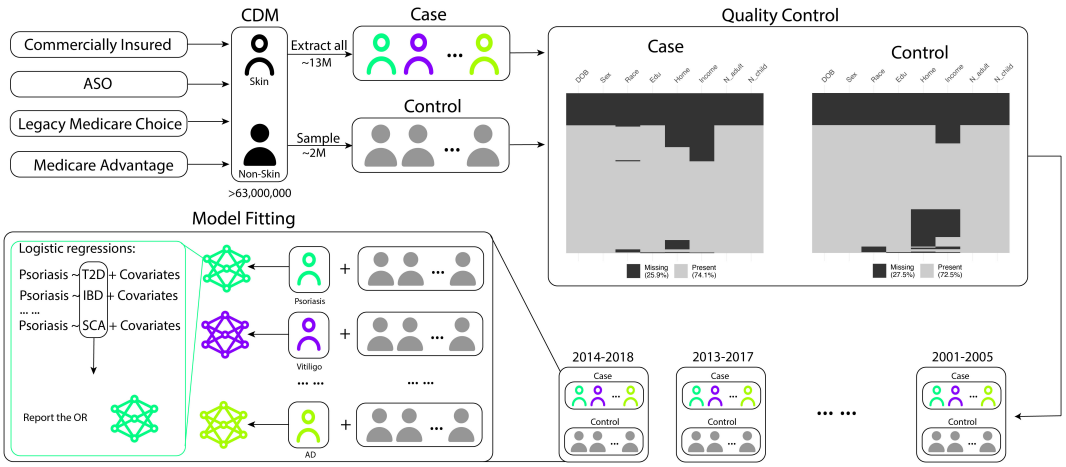


FIGURE 1
Data preprocessing and model fitting workflow. The flowchart illustrates the selection process for patients with skin-related conditions and the control group. Patients with the 14 skin-related conditions are initially extracted, and a separate control group of 2,000,000 patients is randomly sampled from the remaining cohort. Quality control steps are applied to remove patients with incomplete records. The subsequent statistical analyses involve comparing the extracted skin condition patients with the randomly sampled non-skin condition patients as the control group (psoriasis is used as an example in the above pipeline).

>80), allowing for non-linear patterns, with the reference category being age<10. As weight and height information was unavailable, the obesity diagnosis code was used as a surrogate to control for the impact of low or high BMI on disease associations. Male and European ancestry were chosen as the reference categories for sex and race, respectively. Education level was categorized as "below high school," "high school," "bachelor," and "above bachelor," with "below high school" as the reference category. Annual household income was categorized as "<$40k," "$40k-$49k," "$50k-$59k," "$60k-$74k," "$75k-$99k," and "$100k>," with "<$40k" as the reference category. The time lengths for each patient in the system were calculated as the number of years between the first and last recorded diagnosis. For patient i, the logistic regression model for the following comorbidity analysis is thus:

$$\begin{aligned} &logit\{\Pr(Skin_i|X_i)\} \\ &= \beta_0 + \beta_{comorbidity} \times X_{icomorbidity} + \beta_{obesity} \times X_i obesity + \boldsymbol{\beta}_{age} \\ &\times \boldsymbol{X}_{iage} + \beta_{sex} \times X_{isex} + \boldsymbol{\beta}_{race} \times \boldsymbol{X}_{irace} + \beta_{education} \times X_{ieducation} \\ &+ \boldsymbol{\beta}_{income} \times \boldsymbol{X}_{iincome} + \beta_{child} \times X_{ichild} + \beta_{adult} \times X_{iadult} + \beta_{time} \\ &\times X_{itime,} \end{aligned}$$

where $\beta_{comorbidity}$ is the parameter of interest indicating the association levels for a pair of skin/skin-related condition and comorbidity, which can be interpreted as the log odds ratio of developing the skin/skin-related disease between patients with or without the comorbidity.

3 Results

3.1 Summary statistics

The summary information for the cases and controls during the period of 2014-2018 is presented in Table 1 in addition to the US general population characteristics. When comparing the randomly controlled samples with the US general population, the CDM data represents older, higher income and education US population with less ethnic minorities. This further justifies controlling the socioeconomic factors in the logistic regression model for subsequent analysis. Consistent with previous studies (24–28), certain skin or skin-related disorders show a higher prevalence among women. For example, rosacea, alopecia areata, SLE, acne, and hidradenitis suppurativa (HS) have 67.6%, 73.7%, 86.3%, 67.6%, and 72.5% female patients, respectively, compared to 50.7% in the control group. We also found a higher proportion of European ancestry associated with the diagnosis of rosacea, aging (chronic exposure to sun or non-ionizing radiation), melanoma, and pigmentation (e.g. hyperpigmentation and freckles; detailed definition can be found in Supplementary Table 1), with percentages of 82.6%, 87.6%, 88.7%, and 81.8%, respectively, compared to the baseline composition of 72.2% Europeans in the control population. Conversely, the Hispanic and African American populations have lower proportions in most skin diseases compared to the control group, except for vitiligo (16.5%) and leprosy (14.5%) among Hispanics (control: 12.5%), and SLE (15.5%) and HS (18.6%) among African Americans (control: 10.5%). Patients of Asian heritage have a lower proportion of melanoma (0.9%) but a higher proportion of vitiligo (7.4%) and leprosy (8.9%) compared to the control group (4.8%). Furthermore, we observed that a higher education level is associated with a larger number of medical claims for skin disorders. Rosacea (30.5% above college), acne (35.4% above college), and pigmentation (30.1% above college) have the most significant elevation compared to the control group (18.8% above college). Similarly, a higher income level is linked to a stronger association with medical claims for skin conditions, with rosacea (53.4% income >$100k), acne (61.0% >$100k), and pigmentation (53.1% >$100k) showing the largest contrast compared to the control population (39.2% >$100k).

Figure 2 provides an overview of the demographic variables in our study. Figure 2A displays the prevalence of each skin disease and control categorized by gender. AD, pigmentation, and acne are the most prevalent skin conditions in the CDM data, and their prevalence remains consistent when comparing 2014-2018 records to those from 2001-2005 (Supplementary Figure 1A). The gender distributions for different skin conditions also remain consistent. Figure 2B presents the density of the time (in years) that patients stay in the CDM system, showing that approximately 60% of the patients stay within a 5-year time span. Figure 2C displays the age distribution of the control group and each skin disease group for the period between 2014-2018. This represents the ages of patients with skin-related disorders diagnosis in the system, and not necessarily represent the disease age of onset. Each disease exhibits a unique age distribution compared to the control group. For example, acne patients tend to be younger (29), while AD shows a bimodal pattern in age distribution, which is consistent with previous studies (30). We also observed that the median age for all skin conditions, except for acne, tends to be earlier in the 2001-2005 cohort (Supplementary Figure 1B) compared to the 2014-2018 cohort, whereas the age distribution for acne remains consistent over time.

3.2 Skin-comorbidity association trends across time

We first investigated the trend of associations between psoriasis and T2D (18, 31), a comorbidity pair that has been extensively studied before. Figure 3A provides a summary of adjusted Odds Ratios (ORs) with 95% confidence intervals (CIs) from the logistic regression model. We observed consistent and stable estimated ORs across different time periods, ranging between 1.15 and 1.31. To compare our findings with previous studies (18, 31) on the association between psoriasis and T2D, we included their OR estimates and corresponding 95% CIs. Due to smaller sample sizes, the 95% CIs of these earlier studies are wider compared to our analysis. Although their estimates show some variability, their point estimates for OR align closely with ours, and their 95% CIs encompass most of our estimates.

Furthermore, we explored the association trends of other disease pairs and highlighted notable findings in Figure 3. For instance, the association between AD and lung cancer (Figure 3B) has transitioned from a significant positive association in the period

TABLE 1 Descriptive analysis for CDM data.

		PN	Rosacea	AD	PsA	Psoriasis	Alopecia areata	Vitiligo	SLE	Acne	Aging	Melanoma	Pigmentation	Leprosy	HS	Control	Overall Skin	US population
N		146,796	351,026	1,458,417	48,241	272,913	108,462	31,914	67,718	801,150	480,415	73,928	1,297,949	235	36,364	470,414	5,148,043	323,100,000
Age		57.97 (19.22)	54.33 (18.44)	45 (26.15)	56.39 (14.44)	55.25 (18.02)	47.62 (19.04)	49.33 (21.62)	55.68 (16.05)	29.73 (16.87)	62.1 (15.75)	66.22 (14.56)	56.14 (18.64)	62.88 (19.62)	41.45 (16.77)	43.28 (22.83)	47.34 (23.24)	37.9 (median)
Gender	Female	55.78%	67.63%	57.14%	54.48%	52.83%	73.73%	52.90%	86.29%	67.59%	55.36%	44.02%	61.48%	58.72%	72.54%	50.68%	58.86%	51.01%
	Male	44.22%	32.37%	42.86%	45.52%	47.17%	26.27%	47.10%	13.71%	32.41%	44.64%	55.98%	38.52%	41.28%	27.46%	49.32%	41.13%	48.99%
Race	Asian	6.06%	2.11%	6.07%	2.72%	3.68%	6.70%	7.37%	3.25%	5.33%	1.27%	0.88%	2.68%	8.94%	3.13%	4.78%	4.60%	5.67%
	African American	9.05%	3.98%	8.49%	5.76%	7.02%	9.70%	9.34%	15.54%	6.72%	3.33%	3.88%	5.31%	8.51%	18.56%	10.50%	7.37%	13.31%
	Hispanic	8.32%	7.86%	10.38%	9.22%	9.25%	12.94%	16.45%	13.99%	10.13%	4.46%	3.58%	6.75%	14.47%	11.12%	12.52%	9.52%	17.79%
	European	73.34%	82.56%	71.73%	78.76%	76.54%	67.39%	63.13%	64.06%	73.81%	87.61%	88.65%	81.78%	65.53%	63.91%	72.20%	78.51%	61.27%
Education	Below High school	0.32%	0.20%	0.37%	0.36%	0.33%	0.40%	0.60%	0.54%	0.26%	0.09%	0.13%	0.15%	0.43%	0.38%	0.52%	0.29%	16.02%
	High School	22.29%	15.14%	20.17%	23.78%	23.05%	18.34%	20.46%	30.24%	13.59%	15.67%	18.32%	15.17%	31.91%	29.83%	26.40%	18.47%	27.57%
	Below Bachelor	54.53%	54.13%	53.91%	56.46%	54.51%	52.49%	51.87%	54.72%	50.72%	56.54%	57.33%	54.63%	51.06%	55.16%	54.26%	54.20%	45.77%
	Above Bachelor	22.86%	30.52%	25.55%	19.41%	22.12%	28.77%	27.06%	14.49%	35.43%	27.70%	24.22%	30.06%	16.60%	14.62%	18.81%	27.04	10.62%
Home Ownership	Own	90.42%	92.03%	88.59%	89.74%	89.07%	86.74%	88.60%	84.97%	86.86%	94.32%	94.05%	92.50%	91.91%	78.58%	85.06%	89.88%	63.7%
	Rent	9.58%	7.97%	11.41%	10.26%	10.93%	13.26%	11.40%	15.03%	13.14%	5.68%	5.95%	7.50%	8.09%	21.42%	14.94%	10.12%	36.3%
Household Income	<$40k	17.56%	10.30%	15.22%	16.83%	17.57%	14.86%	14.82%	25.94%	9.27%	11.20%	13.93%	10.68%	22.55%	25.46%	19.14%	13.31%	44.82% (<$49k)
	$40k-$49k	5.96%	4.42%	5.75%	6.06%	6.11%	5.83%	5.64%	7.70%	4.15%	4.68%	5.43%	4.40%	8.09%	8.26%	6.95%	5.19%	
	$50k-$59k	7.19%	5.74%	6.72%	7.24%	7.12%	6.59%	6.51%	8.15%	4.67%	6.15%	6.99%	5.68%	9.36%	8.33%	7.57%	6.19%	16.69% ($50k-$74k)
	$60k-$74k	10.93%	9.55%	10.15%	10.90%	10.73%	9.88%	9.95%	11.34%	7.47%	10.31%	11.21%	9.52%	8.09%	10.74%	10.89%	9.69%	
	$75k-$99k	16.77%	16.62%	15.87%	18.01%	16.83%	15.56%	15.87%	15.94%	13.45%	17.69%	18.56%	16.62%	20.00%	15.68%	16.22%	15.95%	12.08%
	>$100k	41.59%	53.38%	46.28%	40.97%	41.64%	47.28%	47.20%	30.93%	60.98%	49.97%	43.89%	53.08%	31.91%	31.53%	39.22%	49.67%	26.41%
Household member	#Adult	1.79 (1.19)	1.98 (1.26)	1.99 (1.24)	1.87 (1.21)	1.85 (1.22)	2.06 (1.3)	2.02 (1.28)	1.75 (1.18)	2.75 (1.49)	1.79 (1.19)	1.65 (1.1)	1.94 (1.26)	1.57 (1.04)	2.06 (1.36)	1.98 (1.25)	2.17 (1.37)	1.94 (0.00)
	#Children	0.26 (0.71)	0.34 (0.79)	0.62 (1.04)	0.24 (0.67)	0.29 (0.74)	0.45 (0.89)	0.48 (0.94)	0.21 (0.62)	0.64 (0.97)	0.21 (0.65)	0.13 (0.52)	0.33 (0.78)	0.21 (0.6)	0.36 (0.78)	0.59 (1.03)	0.48 (0.93)	0.59 (0.00)

PN, prurigo nodularis; AD, atopic dermatitis; PsA, psoriatic arthritis; SLE, systemic lupus erythematosus; HS, hidradenitis suppurativa. This data summarizes data between 2014-2018. The US data come from the US Census Bureau (https://www.census.gov/data/tables).

10.3389/fimmu.2023.1309549

10.3389/fimmu.2023.1309549

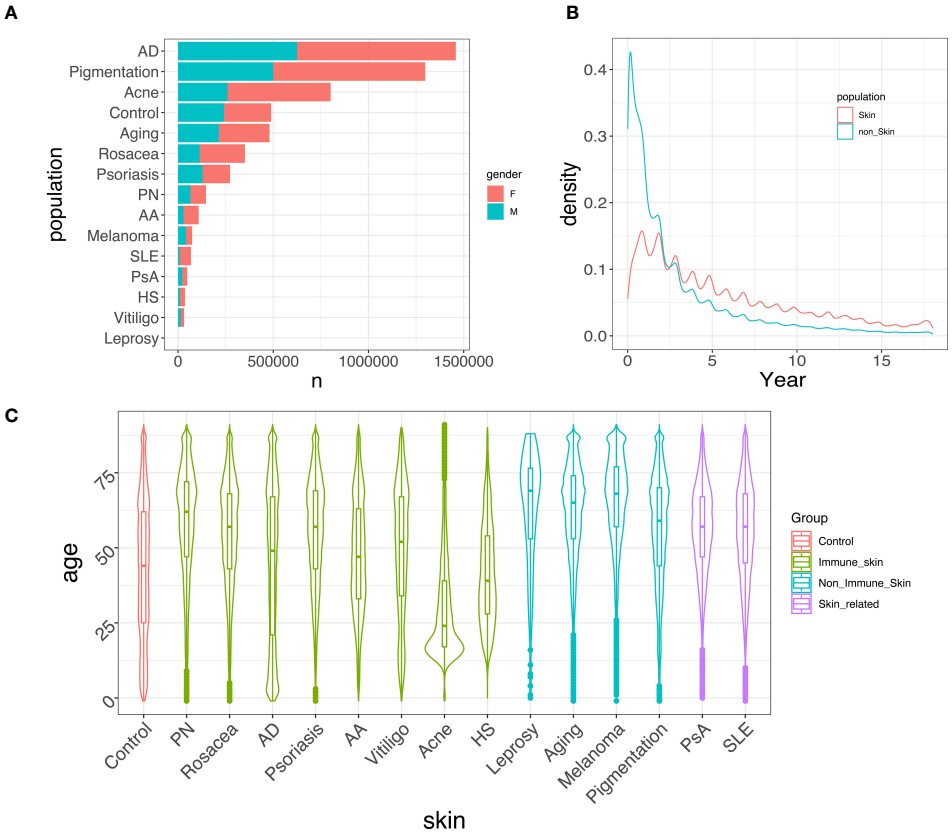


FIGURE 2
Data summary. **(A)** Gender-specific prevalence of each skin disease/control between 2014-2018. Females generally exhibit a higher prevalence than males in developing immune-mediated skin diseases. **(B)** Distribution of patients' time in the system, spanning from 2001 to 2018. Most patients stayed in the system for less than 5 years. **(C)** Age distribution of different skin/skin-related diseases/control between 2014-2018. Most skin diseases show a similar age distribution compared to the control group, while acne, AD, and HS tend to have a higher proportion of younger patients.

2001-2005 (OR: 1.62, 95% CI [1.34-1.97]) to a non-significant association in 2014-2018 (OR: 1.02, 95% CI [0.97-1.07]). While earlier studies from 2005 and 2012 reported positive associations between AD and lung cancer (32, 33), a more recent study in 2020 found that after adjusting for potential mediators such as smoking or smoking-related diseases, this association disappears (34). These findings suggest that improved treatment for AD in recent years or changes in modifying behaviors (such as smoking) may have played a role in reducing the risk of cancer for AD patients. In Figure 3C, we observed strong associations between PsA and RA across different years. Since many clinical measures of PsA are adopted from RA (35) and the specific diagnosis of RA and PsA require knowledge from rheumatologists (36), the strong associations may be attributed to miscoding. To explore this further, we conducted separate analyses for patients diagnosed exclusively in rheumatology clinics (red lines in Figure 3C), in addition to the analysis based on all clinics or providers (black lines in Figure 3C). The associations between PsA and RA from rheumatology clinics consistently exhibit weaker associations compared to the findings from the unrestricted data, while both analyses demonstrate a decreasing trend over time. Although this finding could indicate improving diagnosis accuracy for both rheumatology clinics and other clinics over time, special care is still needed when using medical claims to study disease comorbidities. Additionally, we also observed diminishing differences between the ORs estimated from rheumatology clinics and all clinics (i.e. unrestricted data). We regressed these ORs on both the first-order and second-order time covariates (Figure 3D), and found that the second-order term in the regression for all clinics is not significant ($p = 0.452$), indicating that the rate of ORs changing across years remains relatively constant. In contrast, the second-order term in the regression for rheumatology clinics is significant ($p < 1\times10^{-7}$), suggesting that the changing rate of ORs decreases across years.

3.3 Large-scale comorbidity identification

We conducted a large-scale association study to identify the comorbidities for the 14 skin/skin-related conditions using data from the period 2014-2018. We evaluated a total of 420 skin disease-comorbidity pairs by associating the concurrence of these conditions with 30 common human disorders, including respiratory, cancer, mental, immunological, gastrointestinal, cardiovascular, and diabetes conditions (Figure 4 with detailed association estimates, sample sizes and P-values in Supplementary Table 2). For the large-scale comorbidity analysis, we found that

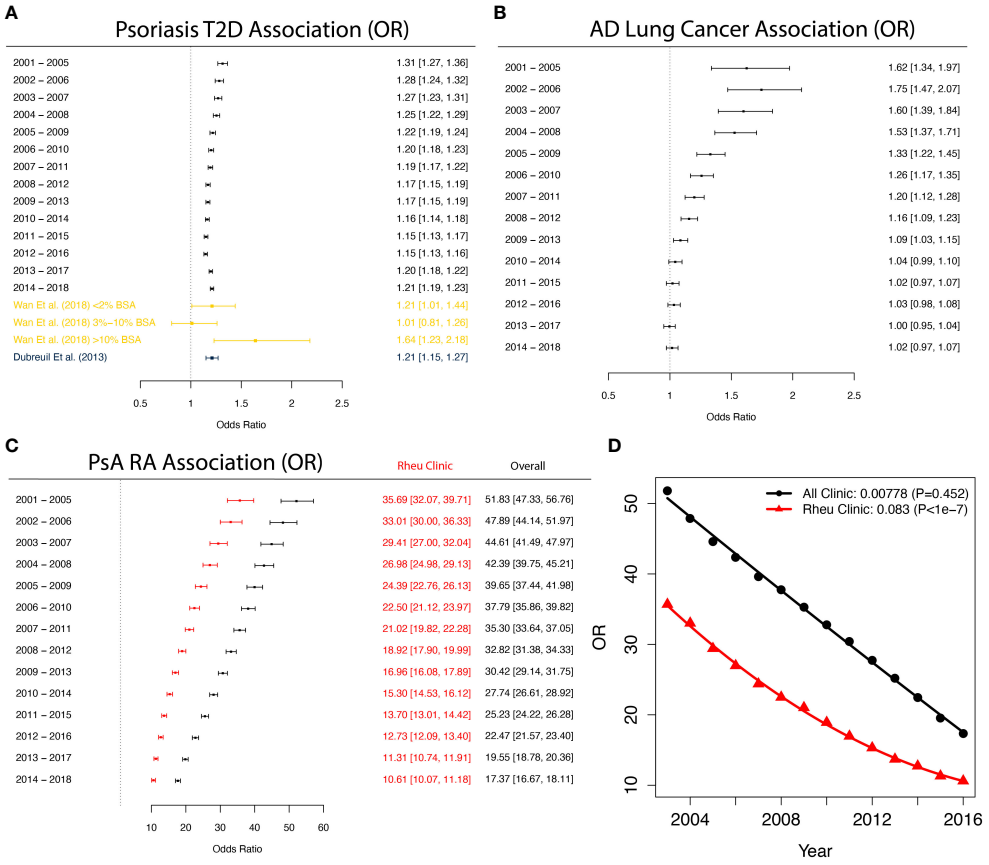


FIGURE 3
Forest plots of association across a year-to-year period. **(A)** Forest plot illustrating the odds ratio (OR) with confidence intervals (CIs) for the association between psoriasis and type 2 diabetes (T2D) in comparison to non-T2D patients. The OR and CI from this study are shown, along with the corresponding OR and CI from two previous studies for comparison. The findings indicate that the OR estimate from this study aligns with previous results, *but* featuring more precise CIs. **(B)** Forest plot showcasing the parameter estimate for the OR with CIs of developing atopic dermatitis (AD) in lung cancer patients compared to lung cancer-free patients. The results exhibit a declining trend in the association, which ultimately dissipates. **(C)** Forest plot displaying the parameter estimate for the OR with CIs between psoriatic arthritis (PsA) and rheumatoid arthritis (RA) based on all clinics and providers (black) and solely rheumatology clinics (red). The estimated associations derived from rheumatology clinics is *weaker* than that from all clinics, with both estimates showing a steady downward trend. This suggests the potential for more precise diagnoses in rheumatology clinics, as well as improved diagnosis accuracy over time in general. **(D)** Regression analysis of PsA vs RA odds ratios based on all clinics and rheumatology clinics, incorporating first-order and second-order time covariates. Estimates and P-values of the second-order time coefficients are shown in the legend. The significant second-order time coefficient from the rheumatology clinic estimate suggests a significant deceleration in the rate of change for ORs, while the rate of change for ORs from all clinics demonstrates a steady decline. For all figures, the control group consists of randomly sampled patients from the general CDM population.

most of the skin/skin-related condition-comorbidity associations are significant and positive, with the most prominent associated pairs being Crohn's disease and leprosy (OR=6.60, 95% CI: 3.09-14.08); primary biliary cirrhosis (PBC) and SLE (OR=6.07, 95% CI: 4.93-7.46); as well as celiac disease (CD) and SLE (OR=6.06, 95% CI: 5.49-6.69). These associations are consistent with previous literature: for instance, different studies have reported overlapping genetic signals between Crohn's disease and leprosy (37–39). For PBC and SLE, researchers have found the odds of developing PBC is 2.23 (CI: 1.26-3.96) times higher if patients have a family history of SLE (40). A 2016 study estimated the CD and SLE association to be 3.92 in OR (CI 2.55-6.03) (41). Our findings also reveal that patients diagnosed with melanoma have higher rates of being diagnosed with multiple cancers, including ovarian, lung, and prostate cancers. Additionally, we observed that diabetes has either no association or significant negative associations with acne, rosacea, aging, pigmentation, and melanoma. However, among all the skin conditions studied, leprosy patients exhibit the highest odds of co-diagnosis with type I diabetes (OR: 2.71, CI: 1.53-4.80). Our findings align with previous research demonstrating that the incidence of diabetes among leprosy patients is over seven times higher compared to control groups (14.2% vs. 2%) (42). Notably, when compared to the 2001-2005 cohort, the most notable associations remain consistent (Supplementary Figure 2), while less associations are observed for multiple cancers.

We presented the effect sizes (in log OR) of all comorbidities for each skin/skin-related condition in the 2014-2018 cohort in Supplementary Figure 3A. This highlights that patients with SLE, leprosy, and HS are more susceptible to other comorbid diagnoses. In Supplementary Figure 3B, we showed the effect sizes of skin/skin-related conditions within each comorbidity, revealing that celiac disease, Crohn's disease, and ulcerative colitis have the strongest

10.3389/fimmu.2023.1309549

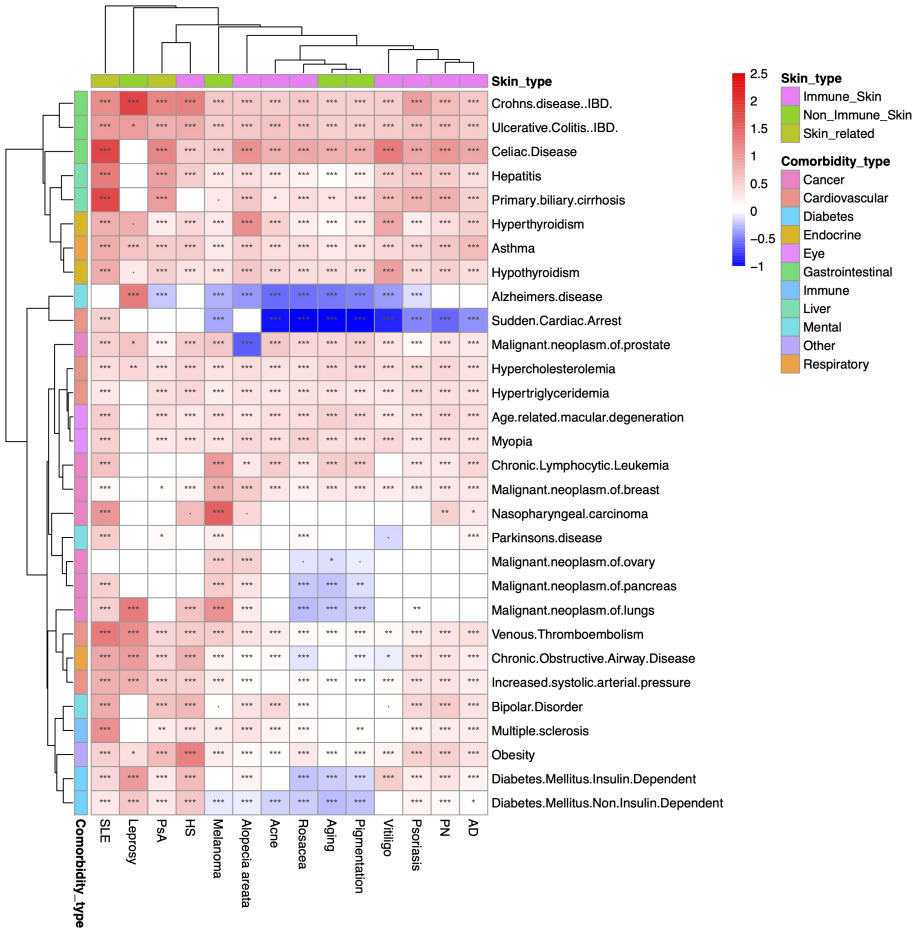


FIGURE 4
Heatmap of large-scale association results between 2014-2018. Heatmap representation of the associations between overall skin/skin-related conditions and potential comorbidities during the period of 2014-2018. The color intensity reflects the level of odds ratio (OR) association, while asterisks indicate the significance levels (***: $P<10^{-3}$; **: $10^{-3}\leq P<10^{-2}$; *: $10^{-2}\leq P<0.05$; ·: $0.05\leq P<0.01$). The findings suggest that the majority of associations between skin and skin-related conditions and comorbidities are both significant and positive. Particularly notable pairings include Crohn's disease with leprosy, primary biliary cirrhosis with systemic lupus erythematosus (SLE), and celiac disease with SLE. # The comorbidity analysis does not include rheumatological conditions due to the ambiguity of the phenotyping when using ICD codes and misdiagnosis.

average associations with the multiple different skin conditions studied in our analysis. We also provided the results for the 2001-2005 cohort in Supplementary Figure 4, which generally align with the findings from the 2014-2018 cohort. Additionally, we summarized the 2014-2018 prevalence of the most prevalent comorbidities within controls and patients with skin/skin-related diseases in Supplementary Table 3. These results further support that celiac disease is one of the most common comorbidities for patients suffering from skin/skin-related conditions.

4 Discussion

Identifying potential comorbidities, particularly those with modest associations, often requires a large sample size for adequate statistical power. Skin conditions, despite being prevalent, are known to have a high percentage of patients who do not seek medical advice, estimated at 73% (43). Consequently, studies in this domain may suffer from limited sample sizes and reduced power to detect weak associations (18, 31). However, leveraging the extensive sample size provided by the claims-based CDM database, we were able to uncover comorbidities even with mild associations. It is worth noting, however, that the CDM database does not include patients insured by Medicaid, which may impact the generalizability of the findings. To validate the CDM dataset, we evaluated the population summary statistics and confirmed their consistency with previous findings regarding overall prevalence, as well as age, ethnicity, and gender distributions. Additionally, we have showcased the well-established link between psoriasis and T2D as a proof-of-concept to further substantiate the validity of the CDM data. We also investigated other skin/skin-related diseases and comorbidities to determine association trends over time. We found that the PsA and RA association decreased dramatically across years. For a long time, PsA was considered to be a variant of RA (44, 45) due to limited knowledge and lack of more specific biomarkers (46). Since the proposition and clinical application of dactylitis as a hallmark and distinct feature of PsA, compared to RA in 1996 (47), and the

CASPAR criteria for PsA diagnosis in 2006 (48), our analysis suggests that potential mis-diagnosis is decreasing over time.

We also adopted a different approach to examine the comorbidity: for a particular skin condition (e.g. psoriasis) we randomly selected control patients from the remaining 13 cohorts consisting of patients with different skin conditions. The pipeline and results of this alternative analysis, depicted in Supplementary Figures 5 and 6, indicate a generally lower association between psoriasis and T2D compared to the original analysis. This suggests the existence of associations between T2D and other skin conditions within the dataset.

The comorbidity of skin diseases can arise from various mechanisms, and understanding these mechanisms can contribute to a deeper comprehension of disease pathogenesis and enhance diagnostic accuracy. The information on disease co-occurrence would enable researchers to explore shared pathogenesis between these related conditions, thereby advancing the understanding of both conditions. Additionally, comorbidities play a crucial role in dermatological diagnoses, aiding dermatologists in distinguishing different diseases more accurately. The presence of comorbidities can be influenced by treatments administered to patients. In other words, different therapeutic interventions, such as medications, surgeries, or other medical procedures, can have an impact on the occurrence or development of concurrent diseases in individuals with skin conditions. For instance, certain medications used to treat one condition may influence the immune system or physiological processes that could potentially lead to the onset or exacerbation of other diseases. Additionally, the side effects or interactions of medications can also contribute to the development of comorbidities. Moreover, confounding factors such as patients' lifestyle, quality of life, and living environment can also lead to disease co-occurrence (49). In this analysis, we accounted for potential confounders by adjusting for demographic and socioeconomic variables in the model. Lastly, misdiagnosis can contribute to the observed co-occurrence of two diseases. For example, PsA and RA are susceptible to misdiagnosis, as reported in previous studies (50). In our analysis, we observed a high association between these conditions; however, we also noticed a consistent temporal decrease in this association. This may be attributed to improved diagnostic criteria and a better understanding of disease mechanisms. Nevertheless, it is important to note that our association analysis does not completely eliminate the potential of misdiagnosis. We recommend that future systematic studies consider employing machine learning methods to correct phenotyping and address misdiagnosis as a preliminary step (51, 52).

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: The data analyzed in this study was obtained from Optum's de-identified Clinformatics® Data Mart Database, and cannot be directly shared to the public. Requests to access these datasets should be directed to Optum, connected@optum.com.

Author contributions

QL: Data curation, Formal Analysis, Investigation, Methodology, Software, Writing – original draft. MP: Data curation, Methodology, Supervision, Writing – review & editing. SS: Data curation, Writing – review & editing. JK: Supervision, Writing – review & editing. JMK: Writing – review & editing. JG: Writing – review & editing. KH: Methodology, Supervision, Writing – review & editing. LT: Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by the Michigan Mcubed 3.0 Classic Cube fund (JiK, KH, LT), National Psoriasis Foundation (LT, MP, and JG), and awards from the National Institutes of Health (K01AR072129 and R01AR080662 to LT; 1P30AR075043 to LT, MP, and JG; UC2 AR081033 to LT and JG). MP was also supported by the Dermatology Foundation.

Conflict of interest

JG has served as a consultant to AbbVie, Eli Lilly, Almirall, Celgene, BMS, Janssen, Prometheus, TimberPharma, Galderma, Novatis, MiRagen, AnaptysBio and has received research support from AbbVie, SunPharma, Eli Lilly, Kyowa Kirin, Almirall, Celgene, BMS, Janssen, Prometheus, and TimberPharma. LT has received research support from Janssen, Novartis, and Galderma.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2023.1309549/full#supplementary-material

References

1. Hay RJ, Johns NE, Williams HC, Bolliger IW, Dellavalle RP, Margolis DJ, et al. The global burden of skin disease in 2010: an analysis of the prevalence and impact of skin conditions. *J Invest Dermatol* (2014) 134(6):1527–34. doi: 10.1038/jid.2013.446

2. Johnson M-LT, Roberts J. Skin conditions and related need for medical care among persons 1-74 years, United States, 1971-1974. *Department Health Education Welfare Public Health Service Office* (1978).

3. Bickers DR, Lim HW, Margolis D, Weinstock MA, Goodman C, Faulkner E, et al. The burden of skin diseases: 2004: A joint project of the American Academy of Dermatology Association and the Society for Investigative Dermatology. *J Am Acad Dermatol* (2006) 55(3):490–500. doi: 10.1016/j.jaad.2006.05.048

4. Schofield J, Grindlay D, William H. *Skin conditions in the UK: a health needs assessment* (University of Nottingham: Centre of Evidence Based Dermatology UK) (2009) 2013.

5. Hay RJ, Fuller LC. The assessment of dermatological needs in resource-poor regions. *Int J Dermatol* (2011) 50(5):552–7. doi: 10.1111/j.1365-4632.2011.04953.x

6. Finlay AY, Coles E. The effect of severe psoriasis on the quality of life of 369 patients. *Br J Dermatol* (1995) 132(2):236–44. doi: 10.1111/j.1365-2133.1995.tb05019.x

7. Krueger G, Koo J, Lebwohl M, Menter A, Stern RS, Rolstad T. The impact of psoriasis on quality of life: results of a 1998 National Psoriasis Foundation patient-membership survey. *Arch Dermatol* (2001) 137(3):280–4.

8. Brezinski EA, Dhillon JS, Armstrong AW. Economic burden of psoriasis in the United States: a systematic review. *JAMA Dermatol* (2015) 151(6):651–8. doi: 10.1001/jamadermatol.2014.3593

9. Adamson AS. The economics burden of atopic dermatitis. *Manage Atopic Dermatitis* (2017) 1027:79–92. doi: 10.1007/978-3-319-64804-0_8

10. Khamashta M, Bruce I, Gordon C, Isenberg D, Ateka-Barrutia O, Gayed M, et al. The cost of care of systemic lupus erythematosus (SLE) in the UK: annual direct costs for adult SLE patients with active autoantibody-positive disease. *Lupus* (2014) 23 (3):273–83. doi: 10.1177/0961203313517407

11. Slawsky KA, Fernandes AW, Fusfeld L, Manzi S, Goss TF. A structured literature review of the direct costs of adult systemic lupus erythematosus in the US. *Arthritis Care Res* (2011) 63(9):1224–32. doi: 10.1002/acr.20502

12. Laughter MR, Maymone MB, Karimkhani C, Rundle C, Hu S, Wolfe S, et al. The burden of skin and subcutaneous diseases in the United States from 1990 to 2017. *JAMA Dermatol* (2020) 156(8):874–81. doi: 10.1001/jamadermatol.2020.1573

13. Wakkee M, Nijsten T. Comorbidities in dermatology. *Dermatologic clinics.* (2009) 27(2):137–47. doi: 10.1016/j.det.2008.11.013

14. Brunner PM, Silverberg JI, Guttman-Yassky E, Paller AS, Kabashima K, Amagai M, et al. Increasing comorbidities suggest that atopic dermatitis is a systemic disorder. *J Invest Dermatol* (2017) 137(1):18–25. doi: 10.1016/j.jid.2016.08.022

15. Balta I, Balta S, Demirkol S, Celik T, Ekiz O, Cakar M, et al. Aortic arterial stiffness is a moderate predictor of cardiovascular disease in patients with psoriasis vulgaris. *Angiology* (2014) 65(1):74–8. doi: 10.1177/0003319713485805

16. Parsi KK, Brezinski EA, Lin T-C, Li C-S, Armstrong AW. Are patients with psoriasis being screened for cardiovascular risk factors? A study of screening practices and awareness among primary care physicians and cardiologists. *J Am Acad Dermatol* (2012) 67(3):357–62. doi: 10.1016/j.jaad.2011.09.006

17. Kimball AB, Wu Y. Cardiovascular disease and classic cardiovascular risk factors in patients with psoriasis. *Int J Dermatol* (2009) 48(11):1147–56. doi: 10.1111/j.1365-4632.2009.04075.x

18. Wan MT, Shin DB, Hubbard RA, Noe MH, Mehta NN, Gelfand JM. Psoriasis and the risk of diabetes: a prospective population-based cohort study. *J Am Acad Dermatol* (2018) 78(2):315–22. doi: 10.1016/j.jaad.2017.10.050

19. Hemminki K, Liu X, Försti A, Sundquist J, Sundquist K, Ji J. Subsequent type 2 diabetes in patients with autoimmune disease. *Sci Rep* (2015) 5(1):1–8. doi: 10.1038/srep13871

20. Frey LJ, Bernstam EV, Denny JC. Precision medicine informatics. *J Am Med Inf Assoc* (2016) 23(4):668–70. doi: 10.1093/jamia/ocw053

21. Gunaseelan V, Kenney B, Lee JS-J, Hu HM. Databases for surgical health services research: Clinformatics Data Mart. *Surgery* (2019) 165(4):669–71. doi: 10.1016/j.surg.2018.02.002

22. O'Byrne ML, DeCost G, Katcoff H, Savla JJ, Chang J, Goldmuntz E, et al. Resource utilization in the first 2 years following operative correction for tetralogy of fallot: study using data from the optum's de-identified clinformatics data mart insurance claims database. *J Am Heart Assoc* (2020) 9(15):e016581. doi: 10.1161/JAHA.120.016581

23. Desai RJ, Solomon DH, Jin Y, Liu J, Kim SC. Temporal trends in use of biologic DMARDs for rheumatoid arthritis in the United States: a cohort study of publicly and privately insured patients. *J managed Care specialty pharmacy.* (2017) 23(8):809–14. doi: 10.18553/jmcp.2017.23.8.809

24. Kyriakis KP, Palamaras I, Terzoudi S, Emmanuelides S, Michailides C, Pagana G. Epidemiologic aspects of rosacea. *J Am Acad Dermatol* (2005) 53(5):918–9. doi: 10.1016/j.jaad.2005.05.018

25. Lundin M, Chawa S, Sachdev A, Bhanusali D, Seiffert-Sinha K, Sinha AA. Gender differences in alopecia areata. *J Drugs dermatology: JDD.* (2014) 13(4):409–13.

26. Wasef SZY. Gender differences in systemic lupus erythematosus. *Gender Med* (2004) 1(1):12–7. doi: 10.1016/S1550-8579(04)80006-8

27. Skroza N, Tolino E, Proietti I, Bernardini N, La Viola G, Nicolucci F, et al. Women and acne: any difference from males? A review of the literature. *Giornale italiano di dermatologia e venereologia: organo ufficiale Societa italiana di dermatologia e sifilografia.* (2016) 151(1):87–92.

28. Yee D, Collier EK, Atluri S, Jaros J, Shi VY, Hsiao JL. Gender differences in sexual health impairment in hidradenitis suppurativa: A systematic review. *Int J Women's Dermatol* (2021) 7(3):259–64. doi: 10.1016/j.ijwd.2020.10.010

29. Cunliffe W, Gould D. Prevalence of facial acne vulgaris in late adolescence and in adults. *Br Med J* (1979) 1(6171):1109–10. doi: 10.1136/bmj.1.6171.1109

30. Saeki H, Furue M, Furukawa F, Hide M, Ohtsuki M, Katayama I, et al. Guidelines for management of atopic dermatitis. *J Dermatol* (2009) 36(10):563–77. doi: 10.1111/j.1346-8138.2009.00706.x

31. Dubreuil M, Rho YH, Man A, Zhu Y, Zhang Y, Love TJ, et al. Diabetes incidence in psoriatic arthritis, psoriasis and rheumatoid arthritis: a UK population-based cohort study. *Rheumatology* (2014) 53(2):346–52. doi: 10.1093/rheumatology/ket343

32. Hagströmer L, Ye W, Nyrén O, Emtestam L. Incidence of cancer among patients with atopic dermatitis. *Arch Dermatol* (2005) 141(9):1123–7. doi: 10.1001/archderm.141.9.1123

33. Hwang CY, Chen YJ, Lin MW, Chen TJ, Chu SY, Chen CC, et al. Cancer risk in patients with allergic rhinitis, asthma and atopic dermatitis: a nationwide cohort study in Taiwan. *Int J cancer.* (2012) 130(5):1160–7. doi: 10.1002/ijc.26105

34. Mansfield KE, Schmidt SA, Darvalics B, Mulick A, Abuabara K, Wong AY, et al. Association between atopic eczema and cancer in England and Denmark. *JAMA Dermatol* (2020) 156(10):1086–97. doi: 10.1001/jamadermatol.2020.1948

35. Brent LH. Inflammatory arthritis: an overview for primary care physicians. *Postgraduate Med* (2009) 121(2):148–62. doi: 10.3810/pgm.2009.03.1987

36. Giacomelli R, Gorla R, Trotta F, Tirri R, Grassi W, Bazzichi L, et al. Quality of life and unmet needs in patients with inflammatory arthropathies: results from the multicentre, observational RAPSODIA study. *Rheumatology* (2015) 54(5):792–7. doi: 10.1093/rheumatology/keu398

37. Grant AV, Alter A, Huong NT, Orlova M, Van Thuc N, Ba NN, et al. Crohn's disease susceptibility genes are associated with leprosy in the Vietnamese population. *J Infect diseases.* (2012) 206(11):1763–7. doi: 10.1093/infdis/jis588

38. Schurr E, Gros P. A common genetic fingerprint in leprosy and Crohn's disease? *Mass Med Soc;* (2009) p:2666–8. doi: 10.1056/NEJMe0910690

39. Jung S, Park D, Lee H-S, Kim Y, Baek J, Hwang SW, et al. Identification of shared loci associated with both Crohn's disease and leprosy in East Asians. *Hum Mol Genet* (2022) 31(22):3934–44. doi: 10.1093/hmg/ddac101

40. Gershwin ME, Selmi C, Worman HJ, Gold EB, Watnik M, Utts J, et al. Risk factors and comorbidities in primary biliary cirrhosis: a controlled interview-based study of 1032 patients. *Hepatology* (2005) 42(5):1194–202. doi: 10.1002/hep.20907

41. Dahan S, Shor DB-A, Comaneshter D, Tekes-Manova D, Shovman O, Amital H, et al. All disease begins in the gut: celiac disease co-existence with SLE. *Autoimmun Rev* (2016) 15(8):848–53. doi: 10.1016/j.autrev.2016.06.003

42. Nigam P, Dayal S, Srivastava P, Joshi L, Goyal B, Gupta M. Diabetic status in leprosy. *Hansenologia Internationalis: hanseníase e outras doenças infecciosas.* (1979) 4 (1):7–14. doi: 10.47878/hi.1979.v4.35626

43. Basra MK, Shahrukh M. Burden of skin diseases. *Expert Rev Pharmacoeconomics Outcomes Res* (2009) 9(3):271–83. doi: 10.1586/erp.09.23

44. Wright V. Psoriatic arthritis. *Bull Rheum Dis* (1971) 21:627–32. doi: 10.1016/0049-0172(73)90035-8

45. Gladman DD, Ritchlin C. Clinical manifestations and diagnosis of psoriatic arthritis. *UpToDate* (2020). v18.

46. Avila R, Pugh DG, Slocumb CH, Winkelmann R. Psoriatic arthritis: a roentgenologic study. *Radiology* (1960) 75(5):691–702. doi: 10.1148/75.5.691

47. Roberton D, Cabral D, Malleson P, Petty R. Juvenile psoriatic arthritis: followup and evaluation of diagnostic criteria. *J Rheumatol* (1996) 23(1):166–70.

48. Taylor W, Gladman D, Helliwell P, Marchesoni A, Mease P, Mielants H. Classification criteria for psoriatic arthritis: development of new criteria from a large international study. *Arthritis Rheumatism: Off J Am Coll Rheumatol* (2006) 54(8):2665–73. doi: 10.1002/art.21972

49. Liu J-T, Yeh H-M, Liu S-Y, Chen K-T. Psoriatic arthritis: epidemiology, diagnosis, and treatment. *World J orthopedics.* (2014) 5(4):537. doi: 10.5312/wjo.v5.i4.537

50. Merola JF, Espinoza LR, Fleischmann R. Distinguishing rheumatoid arthritis from psoriatic arthritis. *RMD Open* (2018) 4(2):e000656. doi: 10.1136/rmdopen-2018-000656

51. Tang B, Wu Y, Jiang M, Denny JC, Xu H. Recognizing and encoding discorder concepts in clinical text using machine learning and vector space model. *CLEF (Working Notes).* (2013) 665.

52. Chen ML. *Machine Learning for the Classification of Dementia in the Presence of Mis-labelled Data*. (Imperial College London) (2013).

frontiers | Frontiers in Immunology

TYPE Methods
PUBLISHED 06 February 2024
DOI 10.3389/fimmu.2024.1323072

OPEN ACCESS

EDITED BY
Alex Tsoi,
University of Michigan, United States

REVIEWED BY
Poulami Dey,
University of Michigan, United States
Lin Zhang,
University of Michigan, United States

*CORRESPONDENCE
Fangyuan Shi
shify@nxu.edu.cn

RECEIVED 17 October 2023
ACCEPTED 15 January 2024
PUBLISHED 06 February 2024

CITATION
Li H, Yu Z, Du F, Song L, Gao Y and Shi F (2024) sscNOVA: a semi-supervised convolutional neural network for predicting functional regulatory variants in autoimmune diseases.
Front. Immunol. 15:1323072.
doi: 10.3389/fimmu.2024.1323072

sscNOVA: a semi-supervised convolutional neural network for predicting functional regulatory variants in autoimmune diseases

Haibo Li[1], Zhenhua Yu[1,2], Fang Du[1,2], Lijuan Song[1,2], Yang Gao[3] and Fangyuan Shi[1,2]*

[1]School of Information Engineering, Ningxia University, Yinchuan, China, [2]Collaborative Innovation Center for Ningxia Big Data and Artificial Intelligence Co-founded by Ningxia Municipality and Ministry of Education, Yinchuan, Ningxia University, Yinchuan, China, [3]School of Medical Technology, North Minzu University, Yinchuan, China

Genome-wide association studies (GWAS) have identified thousands of variants in the human genome with autoimmune diseases. However, identifying functional regulatory variants associated with autoimmune diseases remains challenging, largely because of insufficient experimental validation data. We adopt the concept of semi-supervised learning by combining labeled and unlabeled data to develop a deep learning-based algorithm framework, sscNOVA, to predict functional regulatory variants in autoimmune diseases and analyze the functional characteristics of these regulatory variants. Compared to traditional supervised learning methods, our approach leverages more variants' data to explore the relationship between functional regulatory variants and autoimmune diseases. Based on the experimentally curated testing dataset and evaluation metrics, we find that sscNOVA outperforms other state-of-the-art methods. Furthermore, we illustrate that sscNOVA can help to improve the prioritization of functional regulatory variants from lead single-nucleotide polymorphisms and the proxy variants in autoimmune GWAS data.

KEYWORDS

autoimmune disease, regulatory variant, semi-supervised, deep learning, genome wide association studies

Introduction

Autoimmune disease (AD) is a type of disease in which the immune system mistakenly attacks the body's own tissues and organs, resulting in symptoms such as myocarditis, skin rash, and joint pain, including asthma, type I diabetes, and systemic lupus erythematosus (1, 2). Family clustering of different autoimmune diseases suggests that genetic factors underlie common disease pathways (3), increasing the risk of certain autoimmune diseases by affecting the function of the immune system.

Recently, genome-wide association studies (GWAS) revealed that approximately 90% of disease-associated susceptibility variants are in noncoding regions (4). Now, we know that noncoding regions in the human genome harbor distinct regulatory elements, regulatory variants within these elements can potentially impact the regulation of gene expression (5), and hundreds of risk loci associated with autoimmune diseases have been identified (6)—for example, the G allele of the noncoding variant rs7216389 is associated with an increased risk of asthma (7). Although associations between variants and diseases can be identified (8), few regulatory variants were validated; it is still difficult to identify causal variants in autoimmune diseases (9).

Deep learning can now extract valuable information from complex genomic data, enabling the comprehension of regulatory variants linked to autoimmune diseases (10). Yousefian-Jazi et al. used a random forest model to identify regulatory variants associated with autoimmune diseases and studied their functionality, including the classification of putative causal variants for atopic dermatitis and inflammatory bowel disease (11). An integrated network-based approach called ARVIN was used to identify functional regulatory variants, and it was applied to seven autoimmune diseases (12). Lee et al. formulated the deltaSVM tool to predict several single-nucleotide polymorphisms (SNPs) associated with autoimmune diseases (13). Zhou et al. developed the ExPecto framework based on deep learning, enabling the prediction of mutation tissue-specific transcriptional effects, and experimentally validated predictions for four immune-related diseases (14). However, the data for functional regulatory variants in autoimmune diseases used by the previously mentioned tools is limited in quantity, either encompassing a smaller dataset or exclusively comprising variants from HGMD (15) and ClinVar (16). It is still difficult to systematically identify the function of regulatory variants in autoimmune diseases.

Given the lack of a "gold standard" dataset for functional regulatory variants, several unsupervised models were developed to identify functional regulatory variants, for example, MACIE (17), Eigen (18), and semi-supervised model GenoNet (19). Although unsupervised methods do not rely on labeled dataset, their capability may lag behind supervised methods when trained on a high-quality labeled dataset (17).

Here we develop sscNOVA, a semi-supervised convolutional neural network algorithm to identify functional regulatory variants from GWAS and eQTL dataset and explore the functional characteristics of regulatory variants in autoimmune diseases. We evaluate sscNOVA on the independent testing dataset and curated an experimentally validated testing dataset, and the results show that sscNOVA performs better than the state-of-the-art methods. sscNOVA could also identify the functional regulatory variants which are validated by the wet experiment and the candidate causal variants.

Results

Overview of sscNOVA

sscNOVA mainly includes the following modules: (1) acquiring and processing GWAS and ImmuNexUT data to construct the training data of sscNOVA, (2) 141 features related to 31 autoimmune diseases and 28 immune cell types are annotated by feature selection process, (3) training a supervised convolutional neural network (CNN) framework using GWAS and ImmuNexUT data and constructing a semi-supervised convolutional neural network framework (sscNOVA) with the GWAS data which do not have interactions with ImmuNexUT, and (4) evaluating the capability of the sscNOVA framework using GWAS and ImmuNexUT testing datasets as well as experimentally validated HGMD and ClinVar testing datasets (Figure 1).

Feature annotation, selection, and analysis

Variants in the GWAS catalog that have a significant association with autoimmune diseases are unevenly distributed across different autoimmune diseases, especially variants associated with asthma and systemic lupus erythematosus (Supplementary Figure 1). Merging variants from the GWAS catalog and eQTLs with autoimmune diseases, we find that most of the positive variants are more likely to enrich in T helper cells, monocytes, and dendritic cells across 28 immune cell types (Supplementary Figure 2), which is consistent with what has been reported (20). To annotate all variants, we adopt 21,907 features by the Sei framework (21). Feature selection methods are employed to reduce the feature number, while the annotation features are redundant. Ultimately, 141 features were selected with top feature importance which was calculated based on random forest, 150 features were selected by SelectKBest with mutual_info_classif method, and 40 sequence class features were provided by the Sei framework (details in "Methods" section). The T-SNE plot shows that the classification effect of 141 features is better than that of 150 features and 40 features (Figures 2A–C).

To compare the three feature selection methods, we train the CNN with a training dataset to test the model performance on the independent testing dataset (details in "Methods" section). According to the model performance on the independent testing dataset, when using the 141 features, the CNN model performs the best, achieving an area under curve (AUC) of 0.891 and an area under the precision–recall curve (AUPRC) of 0.893, which demonstrates that using 141 features is superior to using 150 features and 40 features (Figures 2D, E). These results indicate that the proposed method based on the CNN model has better performance for predicting regulatory variants in autoimmune diseases when using 141 features (Supplementary Figure 3).

Training and evaluation of sscNOVA

As the positive dataset in the CNN model only covers 10 autoimmune diseases, we adopt a semi-supervised learning approach to further improve the generalization ability of the model with the GWAS data which do not have interactions with the ImmuNexUT dataset (details in "Methods" section). As expected, sscNOVA shows an improvement in predictive

10.3389/fimmu.2024.1323072

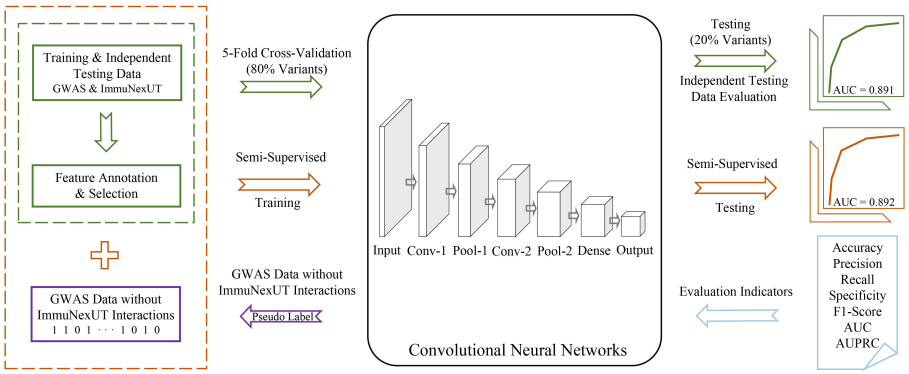


FIGURE 1
Overview of sscNOVA. sscNOVA takes VCF files as input and generates predicted probabilities for each variant as output. Among them, 80% of the intersection variants are designated as the training dataset (green solid box and arrow) for input into the convolutional neural network model (black solid box). The pre-training process employing a fivefold cross-validation training strategy, with 20% of the variants serving as an independent testing dataset for evaluating model performance (area under curve, AUC = 0.891, green curve). Based on the model's predicted probability values, an optimal threshold is identified, and pseudo-labels are assigned to these unlabeled genome-wide association studies data without ImmuNexUT intersection variants (purple solid box and arrow). Subsequently, the dataset with pseudo-labels is merged with the original training dataset (yellow dashed box), and the model undergoes another round of fivefold cross-validation training. In this cross-validation process, the model with the highest AUC is referred to as sscNOVA. Notably, sscNOVA achieves an AUC of 0.892 on the independent testing dataset (yellow curve). The performance of sscNOVA is evaluated using seven metrics (blue section).

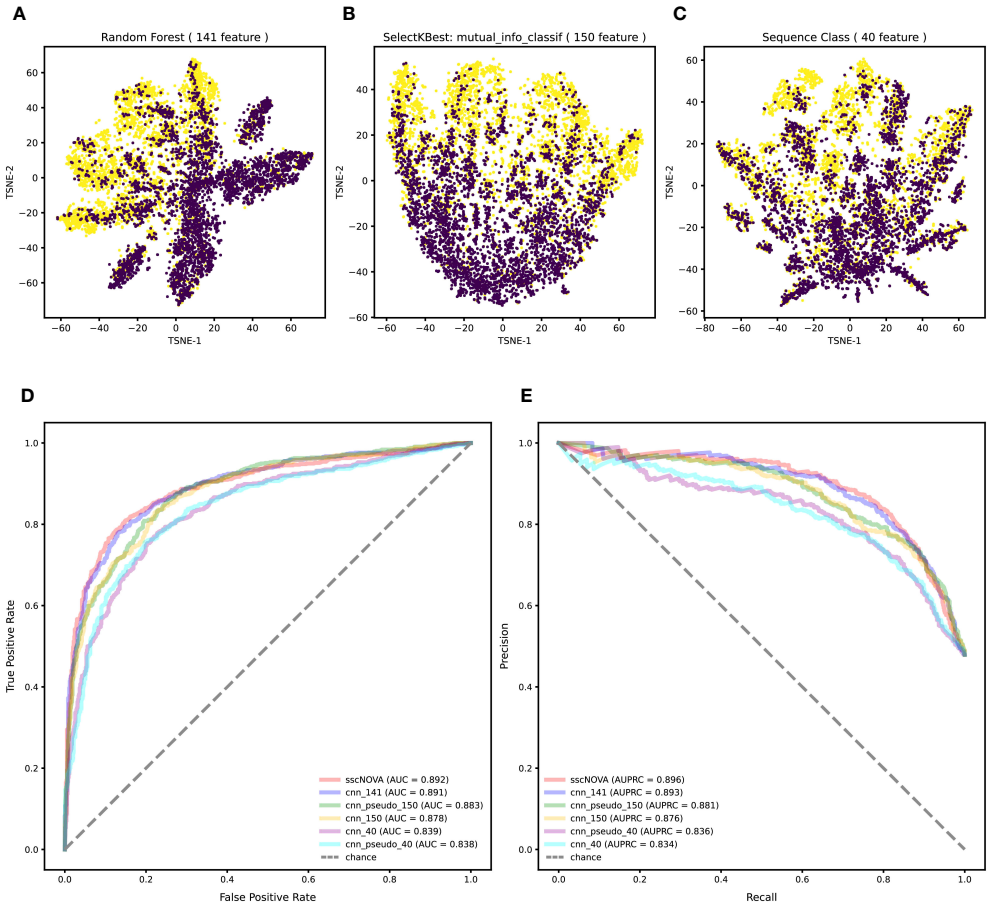


FIGURE 2
Feature selection and performance evaluation. **(A)** T-SNE plot of 141 features are chosen by the calculation of feature importance based on random forest. **(B)** T-SNE plot of 150 features selected by SelectKBest with mutual_info_classif method. **(C)** T-SNE plot of 40 features related to sequence classes which are provided by the Sei framework. **(D)** Comparison of the AUC between the 141, 150, and 40 features on the independent testing dataset with the convolutional neural network (CNN) model. **(E)** Comparison of the area under the precision–recall curve (AUPRC) between the 141, 150, and 40 features on the independent testing dataset with the CNN model.

10.3389/fimmu.2024.1323072

performance on the independent testing dataset; its AUC and AUPRC are 0.892 and 0.896, respectively (Figures 2D, E).

For the purpose of comparing the capability of CNN with other models, we construct three comparative models based on support vector machine (SVM), random forest, and transformer algorithms. Using the three types of features mentioned earlier, we apply the CNN model and these three models to perform fivefold cross-validation on the training dataset and evaluate their predictive performance on the independent testing dataset. According to the experimental results, we find that rf_141 achieves slightly higher AUC and AUPRC values, followed by the cnn_141 model (Figure 3A; Supplementary Figure 4). Afterward, we utilize the dataset containing pseudo-labeled data and train four models using identical methods. Though the AUC and AUPRC of sscNOVA on this dataset are slightly lower than rf_pseudo_141, sscNOVA still has the best recall (Figure 3B; Supplementary Figure 5). This suggests that sscNOVA is capable of accurately capturing features associated with positive variants, thereby reducing the risk of false negatives. This capability contributes to ensuring the effective identification of actual positive variants. The experimental results demonstrate that the pseudo-labeling method effectively alleviates the issue of limited labeled data and helps optimize the model's predictive performance.

Comparison on an experimentally curated testing dataset

To further validate the model performance, we use an experimentally curated testing dataset, in which positive variants include data from the HGMD and ClinVar databases (11), to evaluate four different models. Negative variants are obtained through three different methods: first, 190 negative variants are selected adjacent to positive variants (within ±1 kbp chromosomal positions); second, 118 negative variants are randomly selected from the human genome based on the chromosome numbers of positive variants; and third, 134 negative variants are selected adjacent to positive variants (within ±500 bp chromosomal positions). To compare the performance of the sscNOVA model on these three datasets, it is observed that the model performs best on the 190 negative variants selected adjacent to positive variants (Supplementary Table 1). Therefore, variants obtained through this method are chosen as the negative variants for the experimentally curated testing dataset. We observe that sscNOVA demonstrates excellent performance on both AUC and AUPRC metrics, ranking first (AUC = 0.658, AUPRC = 0.580) and showing significant improvement compared to the rf_141 model (Figure 3C; Supplementary Figures 6, 7). These results indicate that sscNOVA exhibits better generalization capabilities, allowing it to

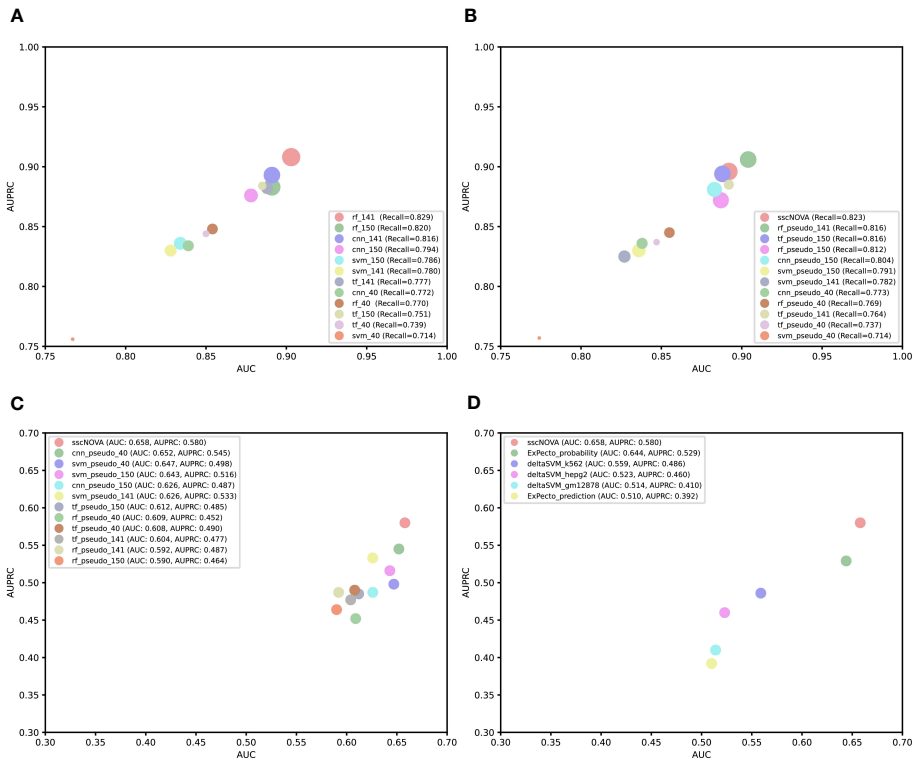


FIGURE 3
Comparison of performance among different models or tools. **(A)** Bubble plot of a different supervised model performance on the independent testing dataset. The x-axis is area under curve (AUC), the y-axis is area under the precision–recall curve (AUPRC), and the size of the bubble represents recall. **(B)** Bubble plot of a different semi-supervised model performance on the independent testing dataset. The x-axis is AUC, the y axis is AUPRC, and the size of the bubble represents recall. **(C)** Comparing convolutional neural network, support vector machine, random forest, and transformer algorithm models based on the experimentally curated testing dataset. The x-axis is AUC, and the y-axis is AUPRC. **(D)** Comparing sscNOVA, ExPecto, and deltaSVM tools based on the experimentally curated testing dataset. The calculation method involves weights for three types of cell lines for deltaSVM and employs two ExPecto score calculation methods. The x-axis is AUC, and the y-axis is AUPRC.

adapt better to new samples and data distributions. In addition, when training sscNOVA on the dataset containing pseudo-labeled data, the capability of sscNOVA on the experimentally curated testing dataset shows improvement in contrast to cnn_141 (Supplementary Figures 6, 7). Moreover, we compare sscNOVA with existing tools for predicting regulatory variants in autoimmune diseases. We evaluate the capability of sscNOVA, ExPecto, and deltaSVM on the experimentally curated testing dataset (details in "Methods" section). Based on the experimental results, the sscNOVA model achieves better performance than the state-of-the-art methods in identifying regulatory variants in autoimmune diseases (Figure 3D).

Prioritizing functional regulatory variants

The functional predictions of sscNOVA can be used to prioritize variants in GWAS. To illustrate the function of sscNOVA in this setting, we show two cases of variants with systemic lupus erythematosus and Crohn's disease risk. The 213-bp open chromatin regions containing the variant rs4385425 targeted by CRISPR-CAS9 showed increasing IKZF1 (Ikaros) expression in Jurkat cells (22). This variant is proxy to the sentinel rs11185603 (r^2 = 0.99) associated with systemic lupus erythematosus. sscNOVA predicts this variant as positive, with a score 0.944. As shown in the UCSC Genome Browser (23), rs4385425 falls into the intergenic region and peak region of H3K27ac (Figures 4A, B). Compared with allele A, allele C improves the binding affinity of two active enhancer makers, H3K27ac and H3K4me1 (24), in multiple lymphocyte cells.

An additional functional regulatory variant is rs212388, which was found to be associated with Crohn's disease. The authors show that the C allele of rs212388 has significantly lower levels of TAGAP mRNA in PBMCs. Moreover, data suggest that TAGAP deficiency was associated with infiltration and proinflammatory gene expression in $CD4^+$ T cells (25). As shown in the UCSC Genome Browser, rs212388 falls into the intro region of TAGAP (Figures 4C, D). The features of rs212388 show that this variant has significant changes in the open chromatin features of $CD4^+$ monocytes. The H3K27ac features in $CD4^+$ lymphocytes also show differences between alleles of rs212388.

Overall, we investigate that sscNOVA could be used to predict the functional regulatory variants in autoimmune GWAS but also prioritize the proxy variants that link with lead SNPs.

Methods

Data acquisition and process

Autoimmune disease-related data are downloaded from the GWAS catalog with GRCh38 human reference genome. A total of

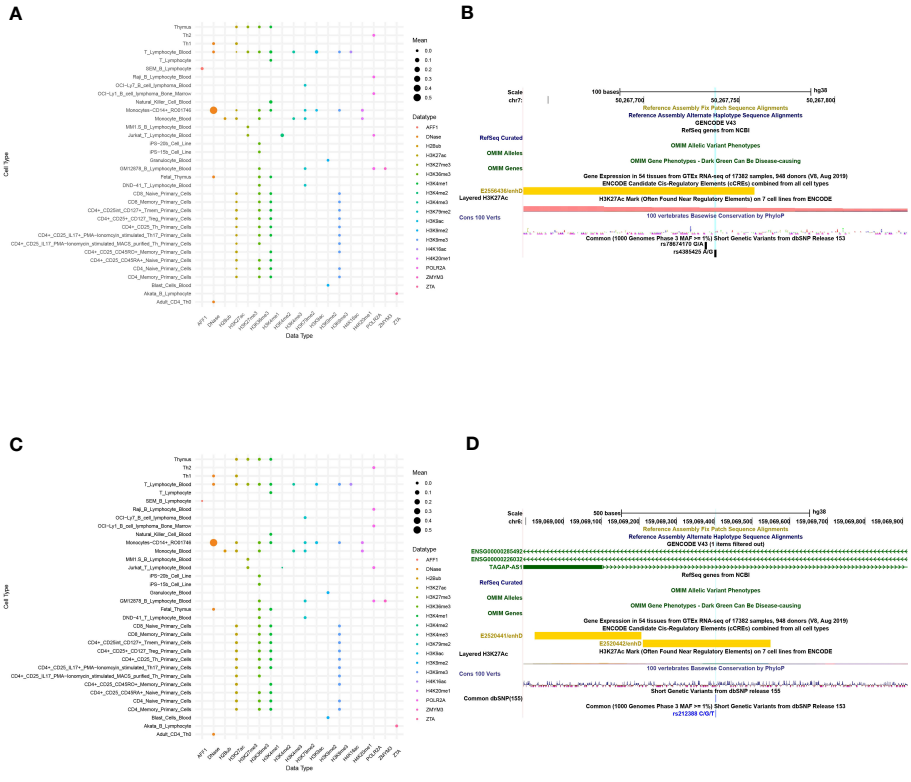


FIGURE 4
A total of 141 features of two variants, rs4385425 and rs212388, are produced in sscNOVA. **(A, C)** The 141 annotation features of variants rs4385425 and rs212388 with the same data type are merged with the average in each cell type to make the bubble plot. The x-axis is data type of annotations, and the y-axis is the cell type of annotations. **(B, D)** UCSC Genome Browser on Human with GRCh38 version is adopted to visualize the variants' genome features.

10,304 variants data are obtained, involving 31 autoimmune diseases such as asthma, rheumatoid arthritis, allergy, etc. The Immune Cell Gene Expression Atlas from the University of Tokyo (ImmuNexUT) data are downloaded from Ota M et al. (26) in the National Bioscience Database Centre (NBDC) website. This dataset includes two accession numbers, E-GEAD-398 and E-GEAD-420, which consist of expression quantitative trait loci (eQTLs) analysis data from 337 patients diagnosed with 10 different autoimmune diseases and 79 healthy volunteers, encompassing a total of 28 distinct immune cell subtypes. These datasets are used to identify associations between genetic variants and gene expressions. Among the datasets, E-GEAD-398 and E-GEAD-420 provide information on the correlation between gene expression levels and genotypes with 2,389,672 genetic variants records. E-GEAD-398 comprises variants with significant associations to autoimmune diseases, while E-GEAD-420 includes variants with non-significant associations to autoimmune diseases in addition to those found in E-GEAD-398. Take the intersection of variants associated with autoimmune diseases in E-GEAD-398 and GWAS as the positive variants of training dataset and independent testing dataset; for the corresponding negative variants, use the variants from E-GEAD-420.

Training dataset and independent testing dataset

The positive dataset was determined by taking the intersection of the processed GWAS catalog and ImmuNexUT numbered E-GEAD-398 variants to create 3,362 positive variants (Supplementary Figure 8). The negative dataset is created by selecting variants with a *P*-value greater than 0.1 and an allele frequency (AF) greater than 0.3 in ImmuNexUT data numbered E-GEAD-420, resulting in 3,670 negative variants (Supplementary Figure 8). After merging the positive dataset with the negative dataset, we randomly sampled the variants' data and split it into training and independent testing dataset in an 80% is to 20% ratio, as the 20% independent testing dataset does not participate in any model training process.

Experimentally curated testing dataset

We use the 140 positive variants utilized by Yousefian-Jazi et al. (https://github.com/jieunjung511/Autoimmune-research) (11). These variants come from HGMD and ClinVar, and a total of 118 positive variants conforming to the VCF format are obtained. Subsequently, we screen the variants within 1 kbp upstream and downstream of the chromosomal positions where the 118 positive variants are located, calculate the conservation values of these variants, and only retain the variants with a phastcons100way conservation value less than 0.5 and AF greater than 0.3. Therefore, the final experimentally curated testing dataset contains 118 positive variants and 190 negative variants (Supplementary Figure 8). In addition, we employ additional methods to obtain negative variants. One approach involves using a pseudo-random number generator on the GRCh37 genome to randomly select chromosomes and positions. This ensures that the chosen positions are not adjacent to known positive variants, resulting in the generation of 118 negative variants. The other method involves choosing 134 negative variants located within ±500 bp chromosomal positions adjacent to the positive variants.

Feature annotation and selection

After annotating the variants with 21,907 features from the Sei framework, feature selection is carried out to select the most informative and relevant features for the analysis, thus focusing on those that are more likely to be associated with the phenotype of interest or have potential functional significance (27).

Initially, 3,102 features related to immune cells are selected from the 21,907 features. Next, two methods, mutual_info_classif and f_classif of SelectKBest, are used to select 1,000, 800, 600, 400, and 200 features from the 3,102 immune-related cell features, respectively (Supplementary Figure 9). Mutual_info_classif method of SelectKBest shows better classification performance than f_classif (Supplementary Figure 10). Subsequently, we continue using mutual_info_classif to select 150, 100, and 50 features from the 3,102 immune-related cell features.

Additionally, we use the feature importance which was calculated based on random forest to select 141 features (Supplementary Figure 10). Three groups of features are compared by the performance trained with random forest model, which includes the 150 features selected by SelectKBest, 141 features selected by the top feature importance which was calculated based on random forest, and 40 features of sequence classes provided by the Sei framework. The T-distributed stochastic neighbor embedding (t-SNE) (28) plot shows that the classification performance is better with 141 features selected by using the random forest method (Supplementary Figure 10). Upon validation using the random forest model, the AUC and AUPRC based on the 141 features selected outperform those selected by other methods (Supplementary Figure 11). The mutual_info_classif method is superior to the f_classif method (details in Supplementary Table 2).

Method for constructing a pseudo-labeled dataset

We construct a pseudo-labeled dataset based on autoimmune disease-related GWAS data which do not have interactions with ImmuNexUT using a threshold and *t*-test method. First, we use the cnn_141 model to predict the probability of the GWAS data without ImmuNexUT interactions and subject them to a fivefold cross-validation. For each variant, five probability values are generated as predictions. First, the Student's *t*-test (29) is conducted to determine if the differences between these five probability values for each variant are statistically significant, with a *P*-value less than 0.05. If the *P*-value of this variant is less than 0.05, the variant is retained; otherwise, it is discarded. To find the optimal pseudo-label threshold for this variant, a parameter search is conducted. Then, using a threshold of 0.5 as a reference, we create five groups of thresholds with ±0, ± 0.1, ± 0.2, ± 0.3, and ±0.4 for all unlabeled variants. (Supplementary Figure 12). Next, we utilize the variants with pseudo-labeled data and the original training dataset to retrain the model and compare the models' performance. Through this approach, we identify the optimal threshold for applying pseudo-labels, which involves considering cnn_141

model-predicted probabilities greater than 0.9 as positive variants and those less than 0.1 as negative variants. In the end, we filter out 2,759 positive variants and 626 negative variants from 6,924 variants data, discarding 3,539 variants that did not satisfy the criteria.

Method for constructing a semi-supervised model

The approach to constructing sscNOVA involves using a trained model to predict variants from the GWAS data which do not have interactions with ImmuNexUT and then pseudo-labeling the unlabeled GWAS data using a threshold and *t*-test method. After that, we merge the dataset with pseudo-labeled data and the original training dataset and evaluate the model's capability using AUC on the independent testing dataset. The threshold corresponding to the highest AUC is selected as the final pseudo-labeling method. Using the same methods, we retrain the models with the augmented dataset.

Semi-supervised model architecture

Semi-supervised learning is a learning approach that combines supervised and unsupervised learning (30). In the presence of a small amount of labeled data, semi-supervised models infer the structure and features of unlabeled data to perform classification and prediction tasks, thereby enhancing model performance with limited labeled data (31). The semi-supervised sscNOVA model implementation consists of the following eight layers:

1. First convolutional layer: Let x be the input feature of length 141 and W be the convolutional kernel of size 5. The output y of the convolutional layer can be calculated as Equation 1:

$$y_i = GELU(\Sigma_{j=0}^{4} W_j \cdot x_{i+j} + b) \tag{1}$$

where i ranges from 0 to 136, and b is the bias term. The resulting output y will have a shape of (137, 32), the number 32 of which represents the quantity of distinct kernels applied to the input data.

2. First max-pooling layer: Given the (137, 32) output shape from the prior Conv1D layer, applying a max-pooling operation with a pool size of 2 reduces each feature map's length by half while keeping 32 feature maps. The output z of the max-pooling layer can be calculated by taking the maximum value within every consecutive two elements in each feature map as Equation 2:

$$z_{i,j} = max(y_{2i,j}, y_{2i+1,j}) \tag{2}$$

where i ranges from 0 to 67, and j ranges from 0 to 31. The resulting output z will have a shape of (68, 32).

3. Second convolutional layer: Let y be the previous output of shape (68, 32) and W' be the convolutional kernel of size 5 for the second convolutional layer, where the number of kernels is 64. The output z can be calculated as Equation 3:

$$z_{i,j} = GELU(\Sigma_{k=0}^{4} W_k^{'} \cdot y_{i+k,j} + b') \tag{3}$$

where i ranges from 0 to 63, j ranges from 0 to 63, k ranges from 0 to 4, and b' is the bias term. The resulting output z will have a shape of (64, 64).

4. Second max-pooling layer: The output w of the second max-pooling layer can be calculated similarly to the first pooling layer as Equation 4:

$$W_{i,j} = max(z_{2i,j}, z_{2i+1,j}) \tag{4}$$

where i ranges from 0 to 31, and j ranges from 0 to 63. The resulting output w will have a shape of (32, 64).

5. Flattening layer: The flattening operation reshapes the 2D array w into a 1D array v by concatenating its rows as Equation 5:

$$v_k = w_{i,j} \tag{5}$$

where $k = i\times64+j$, and k ranges from 0 to 2,047. The resulting output v will have a shape of (1, 2,048).

6. Fully connected (dense) layer: Let v be the input vector of size 2,048 and W'' be the weights of the dense layer. The output x of the dense layer can be calculated as Equation 6:

$$x_i = GELU(\Sigma_{j=0}^{2047} W_{j,i}^{''} \cdot v_j + b_i^{''}) \tag{6}$$

where i ranges from 0 to 15 and corresponds to the 16 specified units in the dense layer, j ranges from 0 to 2,047, and $b_i^{''}$ is the bias term. The resulting output x will have a shape of (16), which matches the number of units within the layer.

7. Dropout layer: The dropout layer performs an element-wise multiplication by a binary mask to apply dropout as Equation 7:

$$y_i = x_i \cdot m_i \tag{7}$$

where i ranges from 0 to 15, and m_i is a binary mask randomly set to 0 or 1 with a probability of 0.1.

8. Output dense layer: Let y be the output of the dropout layer and W''' be the weights of the output dense layer. The final output z can be calculated as Equation 8:

$$z = \sigma(\Sigma_{i=0}^{15} W_i^{'''} \cdot y_i + b''') \tag{8}$$

where σ is the sigmoid activation function, and b''' is the bias term.

The model's architecture is configured for training by utilizing the "binary_crossentropy" loss function (BCELoss). The loss function is as follows Equation 9:

$$BCELoss = -\frac{1}{N}\Sigma_{i=1}^{N}(y_i log(p_i) + (1 - y_i) log(1 - p_i)) \tag{9}$$

where N is the number of variants, y_i represents the actual label (0 or 1) of variant i, and p_i represents the predicted probability by the model that variant i belongs to the positive class. In this loss function, the term $y_i log(p_i)$ penalizes the model for inaccuracies when predicting positive variants, while $(1–y_i) log(1–p_i)$ penalizes inaccuracies in predicting negative variants. The objective of the model is to minimize this loss function to make its predictions closer to the actual labels.

In this neural network model, we opt to use Gaussian Error Linear Unit (GELU) (32) as the activation function, and it is applied in both the convolutional layers and the fully connected layers. Additionally, the "Adam" optimizer is adopted as the guiding algorithm responsible for the model's weight updates throughout the training process. Utilizing its default learning rate of 0.001, the Adam optimizer dynamically adjusts the learning rates for individual parameters (33). The training is conducted in 50 epochs.

sscNOVA functional significance score

For each variant i, $y_{prob}[i]$ is a probability value between 0 and 1, representing the model's prediction of the probability that it belongs to the positive class. Therefore, the scoring formula can be expressed as Equation 10:

$$f_{score}(i) = y_{prob}(i) = dense(flatten(pool(conv(i)))) \quad (10)$$

where i represents the i–th variant in the dataset; *conv*, *pool*, *flatten*, and *dense* represent one-dimensional convolution operation, maximum pooling operation, pooling result flattening, and full connection operation, respectively; and $f_{score}(i)$ represents the predicted probability of the i–th variant belonging to the positive class. The aim is to determine a threshold that achieves a balanced trade-off between these rates within the context of the specific dataset's characteristics, where values above the threshold are classified as positive and values below the threshold are classified as negative.

sscNOVA comparison with ExPecto and deltaSVM

When comparing with ExPecto, we try two methods to calculate the scores. The first method involves comparing the predicted variants labels from the ExPecto model with the true labels and then computing the evaluation metrics based on this comparison. Among them, ExPecto employs a minimum predictive effect threshold (>0.3), which is a threshold for log fold-change recommended by the official website (https://hb.flatironinstitute.org/expecto/about). The second method involves taking the absolute values of the ExPecto model's predicted probabilities and then normalizing and calculating the evaluation metrics based on the normalized probabilities and the true labels. To calculate the deltaSVM scores, the GM12878, K562, and HepG2 cell line models developed by deltaSVM are all tested.

Discussion

Identifying the functional impact of regulatory variants related to autoimmune diseases is a significant challenge in human genetics (34). Due to the scarcity of experimentally validated functional regulatory variants in autoimmune diseases, we adopt the idea of semi-supervised learning, combining labeled and unlabeled data, to develop a framework based on convolutional neural network algorithms to predict functional regulatory variants in autoimmune diseases. sscNOVA provides a feasible solution for the problem of limited gold standard data for regulatory variants in autoimmune diseases. By utilizing the information from unlabeled data, our algorithm helps the models gain more comprehensive information and further elevates the predictive performance. Moreover, the current model results represent the optimal model obtained after fine-tuning (Supplementary Table 3, 4).

Since sscNOVA is based on sequence prediction, it can predict various types of variants. To test whether sscNOVA can help find the rare variants or the variants have not been observed, we utilize the sscNOVA model to predict the validated rare or not previously observed variants in two studies in which the variants were validated by the MPRA assays (35, 36). The recall and AUC values in HeLa, LNCaP, and NPC cell lines indicate that sscNOVA has potential for identifying rare variants (Supplementary Figure 13). In contrast to traditional supervised learning methods, the idea of semi-supervised learning allows us to effectively utilize unlabeled samples in the presence of limited labeled samples, overcoming issues related to data sparsity and missing sample labels (37).

However, some challenges also exist—for instance, the insufficient number of experimentally validated functional regulatory variants may introduce label noise during model training (38), thus reducing prediction performance. It is expected that an increasing amount of experimentally validated variants data will become available, which can intensify prediction performance by leveraging high-confidence data. Due to the limited number of experimentally validated variants in autoimmune diseases, there is a decline in performance on the experimentally curated testing dataset. We localize the positional information of variants in both the independent testing dataset and the experimentally curated testing dataset. Additionally, we conduct a categorized analysis to assess the predictive capability of sscNOVA for each positional category. (Supplementary Figures 14A, B and Supplementary Table 5). We find that sscNOVA has better performance with variants falling into the intron and promoter regions, but variants in the intergenic regions might be missed out by sscNOVA. The annotations in intron and promoter regions are more abundant than those in intergenic regions, which may make it easier for the model to learn patterns of intron variants during the training phase (39, 40). Meanwhile, integrating more experimental validation and functional regulatory variants data will provide greater opportunities to improve predictive performance.

Furthermore, in the ever-evolving field of deep learning, there may be better feature annotation tools capable of capturing the interactions between regulatory regions more effectively. By combining appropriate feature selection methods and training strategies, it could improve the prediction of functional regulatory variants in autoimmune diseases and enhance the capability of model (41). In conclusion, a model based on semi-supervised deep learning can provide new insights and directions for the study of autoimmune diseases, facilitating further investigation into the pathogenesis of autoimmune diseases.

Data availability statement

The autoimmune diseases related GWAS data can be downloaded from https://www.ebi.ac.uk/gwas/docs/file-downloads (Version: All associations v1.0). The ImmuNexUT data can be downloaded from https://humandbs.biosciencedbc.jp/en/hum0214-v6. The source code and detail documentation of sscNOVA are available at https://github.com/NXU-Shilab/sscNOVA.

Author contributions

HL: Formal analysis, Methodology, Visualization, Writing – original draft. ZY: Formal analysis, Funding acquisition, Validation, Writing – review & editing. FD: Data curation, Funding acquisition, Writing – review & editing. LS: Data curation, Funding acquisition, Writing – review & editing. YG: Formal analysis, Writing – review & editing. FS: Conceptualization, Formal analysis, Funding acquisition, Supervision, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study was funded by the Key Research and Development Program of Ningxia (Special Talents) (grant number: 2022BSB03043 and 2022BSB03042), Natural Science Foundation of NingXia China (grant number: 2023A0896 and 2023AAC05006) and Research and Development Program of Ningxia (grant number: 2023BEG02009). This study was also funded by the Science and Technology Innovation Team of Ningxia (grant number: CXTD_2023_KJT_15).

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1323072/full#supplementary-material

References

1. Rioux JD, Abbas AK. Paths to understanding the genetic basis of autoimmune disease. *Nat Volume* (2005) 435:584–9. doi: 10.1038/nature03723

2. Rose NR. Prediction and prevention of autoimmune disease in the 21st Century: A review and preview. *Am J Epidemiol* (2016) 183:403–6. doi: 10.1093/aje/kwv292

3. Farh KKH, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* (2015) 518:337–43. doi: 10.1038/nature13835

4. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, et al. Pervasive sharing of genetic effects in autoimmune disease. *PloS Genet* (2011) 7(8):e1002254. doi: 10.1371/journal.pgen.1002254

5. Pang B, van Weerd JH, Hamoen FL, Snyder MP. Identification of non-coding silencer elements and their regulation of gene expression. *Nat Rev Mol Cell Biol* (2023) 24:383–95. doi: 10.1038/s41580-022-00549-9

6. Parkes M, Cortes A, Van Heel DA, Brown MA. Genetic insights into common pathways and complex relationships among immune-mediated diseases. *Nat Rev Genet* (2013) 14:661–73. doi: 10.1038/nrg3502

7. Ferreira MAR, Matheson MC, Tang CS, Granell R, Ang W, Hui J, et al. Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. *J Allergy Clin Immunol* (2014) 133:1564–71. doi: 10.1016/j.jaci.2013.10.030

8. Heyne HO, Karjalainen J, Karczewski KJ, Lemmelä SM, Zhou W, Havulinna AS, et al. Mono- and biallelic variant effects on disease at biobank scale. *Nature* (2023) 613:519–25. doi: 10.1038/s41586-022-05420-7

9. Perdigoto C. Genetic variation: Putting causal variants on the map. *Nat Rev Genet* (2018) 19:188–9. doi: 10.1038/nrg.2018.11

10. Jin S, Zeng X, Xia F, Huang W, Liu X. Application of deep learning methods in biological networks. *Brief Bioinform* (2021) 22:1902–17. doi: 10.1093/bib/bbaa043

11. Yousefian-Jazi A, Jung J, Choi JK, Choi J. Functional annotation of noncoding causal variants in autoimmune diseases. *Genomics* (2020) 112:1208–13. doi: 10.1016/j.ygeno.2019.07.006

12. Gao L, Uzun Y, Gao P, He B, Ma X, Wang J, et al. Identifying noncoding risk variants using disease-relevant gene regulatory networks. *Nat Commun* (2018) 9(1):702. doi: 10.1038/s41467-018-03133-y

13. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, et al. A method to predict the impact of regulatory variants from DNA sequence. *Nat Genet* (2015) 47:955–61. doi: 10.1038/ng.3331

14. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat Genet* (2018) 50:1171–9. doi: 10.1038/s41588-018-0160-6

15. Sharo AG, Zou Y, Adhikari AN, Brenner SE. ClinVar and HGMD genomic variant classification accuracy has improved over time, as measured by implied disease burden. *Genome Med* (2023) 15(1):51. doi: 10.1186/s13073-023-01199-y

16. Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: Improvements to accessing data. *Nucleic Acids Res* (2020) 48:835–44. doi: 10.1093/nar/gkz972

17. Li X, Yung G, Zhou H, Sun R, Li Z, Hou K, et al. A multi-dimensional integrative scoring framework for predicting functional variants in the human genome. *Am J Hum Genet* (2022) 109:446–56. doi: 10.1016/j.ajhg.2022.01.017

18. Ionita-Laza I, Mccallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* (2016) 48:214–20. doi: 10.1038/ng.3477

19. He Z, Liu L, Wang K, Ionita-Laza I. A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs. *Nat Commun* (2018) 9(1):5199. doi: 10.1038/s41467-018-07349-w

20. Ding J, Frantzeskos A, Orozco G. Functional genomics in autoimmune diseases. *Hum Mol Genet* (2020) 29:59–65. doi: 10.1093/hmg/ddaa097

21. Chen KM, Wong AK, Troyanskaya OG, Zhou J. A sequence-based global map of regulatory activity for deciphering human genetics. *Nat Genet* (2022) 54:940–9. doi: 10.1038/s41588-022-01102-2

22. Su C, Johnson ME, Torres A, Thomas RM, Manduchi E, Sharma P, et al. Mapping effector genes at lupus GWAS loci using promoter Capture-C in follicular helper T cells. *Nat Commun* (2020) 11(1):3294. doi: 10.1038/s41467-020-17089-5

23. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. *Genome Res* (2002) 12:996–1006. doi: 10.1101/gr.229102

24. Sungalee S, Liu Y, Lambuta RA, Katanayeva N, Donaldson Collier M, Tavernari D, et al. Histone acetylation dynamics modulates chromatin conformation and allele-specific interactions at oncogenic loci. *Nat Genet* (2021) 53:650–62. doi: 10.1038/s41588-021-00842-x

25. He R, Chen J, Zhao Z, Shi C, Du Y, Yi M, et al. T-cell activation Rho GTPase-activating protein maintains intestinal homeostasis by regulating intestinal T helper cells differentiation through the gut microbiota. *Front Microbiol* (2023) 13:1030947. doi: 10.3389/fmicb.2022.1030947

26. Ota M, Nagafuchi Y, Hatano H, Ishigaki K, Terao C, Takeshima Y, et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* (2021) 184:3006–3021.e17. doi: 10.1016/j.cell.2021.03.056

27. Marcos-Zambrano LJ, Karaduzovic-Hadziabdic K, Loncar Turukalo T, Przymus P, Trajkovik V, Aasmets O, et al. Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment. *Front Microbiol* (2021) 12:634511. doi: 10.3389/fmicb.2021.634511

28. Maaten LV, Hinton GE. Visualizing data using t-SNE. *J Mach Learn Res* (2008) 9:2579–605. Available at: https:////www.jmlr.org/papers/v9/vandermaaten08a.html.

29. Ho J, Tumkaya T, Aryal S, Choi H, Claridge-Chang A. Moving beyond P values: data analysis with estimation graphics. *Nat Methods* (2019) 16:565–6. doi: 10.1038/s41592-019-0470-3

30. Duarte JM, Berton L. A review of semi-supervised learning for text classification. *Artif Intell Rev* (2023) 56:9401–69. doi: 10.1007/s10462-023-10393-8

31. Xie Z, Chen J, Feng Y, He S. Semi-supervised multi-scale attention-aware graph convolution network for intelligent fault diagnosis of machine under extremely-limited labeled samples. *J Manuf Syst* (2022) 64:561–77. doi: 10.1016/j.jmsy.2022.08.007

32. Dubey SR, Singh SK, Chaudhuri BB. Activation functions in deep learning: A comprehensive survey and benchmark. *Neurocomputing* (2022) 503:92–108. doi: 10.1016/j.neucom.2022.06.111

33. Tang S, Zhu Y, Yuan S. An improved convolutional neural network with an adaptable learning rate towards multi-signal fault diagnosis of hydraulic piston pump. *Advanced Eng Inf* (2021) 50:101406. doi: 10.1016/j.aei.2021.101406

34. Caliskan M, Brown CD, Maranville JC. A catalog of GWAS fine-mapping efforts in autoimmune disease. *Am J Hum Genet* (2021) 108:549–63. doi: 10.1016/j.ajhg.2021.03.009

35. Kircher M, Xiong C, Martin B, Schubach M, Inoue F, Bell RJA, et al. Saturation mutagenesis of twenty disease-associated regulatory elements at single base-pair resolution. *Nat Commun* (2019) 10(1):3583. doi: 10.1038/s41467-019-11526-w

36. Weiss CV, Harshman L, Inoue F, Fraser HB, Petrov DA, Ahituv N, et al. The cis-regulatory effects of modern human-specific variants. *Elife* (2021) 10:e63713. doi: 10.7554/ELIFE.63713

37. Mallapragada PK, Jin R, Jain AK, Liu Y. SemiBoost: Boosting for semi-supervised learning. *IEEE Trans Pattern Anal Mach Intell* (2009) 31:2000–14. doi: 10.1109/TPAMI.2008.235

38. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, et al. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nat Commun* (2020) 11(1):5918. doi: 10.1038/s41467-020-19669-x

39. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* (2012) 489:57–74. doi: 10.1038/nature11247

40. Hong X, Scofield DG, Lynch M. Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol* (2006) 23:2392–404. doi: 10.1093/molbev/msl111

41. Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit Med* (2020) 3(1):30. doi: 10.1038/s41746-020-0229-3

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 10 April 2024
DOI 10.3389/fimmu.2024.1343480

OPEN ACCESS

EDITED BY
Xu-jie Zhou,
Peking University, China

REVIEWED BY
Chachrit Khunsriraksakul,
The Pennsylvania State University,
United States
Koshy Nithin Thomas,
Christian Medical College and Hospital, India

*CORRESPONDENCE
Jinfu Xu
jfxu@tongji.edu.cn

†These authors have contributed equally to this work

RECEIVED 23 November 2023
ACCEPTED 04 March 2024
PUBLISHED 10 April 2024

CITATION
Su Y, Zhang Y, Chai Y and Xu J (2024) Autoimmune diseases and their genetic link to bronchiectasis: insights from a genetic correlation and Mendelian randomization study.
Front. Immunol. 15:1343480.
doi: 10.3389/fimmu.2024.1343480

Autoimmune diseases and their genetic link to bronchiectasis: insights from a genetic correlation and Mendelian randomization study

Yue Su[1†], Youqian Zhang[2†], Yanhua Chai[1†] and Jinfu Xu[1*]

[1]Department of Respiratory and Critical Care Medicine, Shanghai Pulmonary Hospital, School of Medicine, Tongji University, Shanghai, China, [2]Health Science Center, Yangtze University, Jingzhou, Hubei, China

Background: Previous studies have demonstrated that autoimmune diseases are closely associated with bronchiectasis (BE). However, the causal effects between autoimmune diseases and BE remain elusive.

Methods: All summary-level data were obtained from large-scale Genome-Wide Association Studies (GWAS). The univariate Mendelian randomization (UVMR) was utilized to investigate the genetic causal correlation (r_g) of 12 autoimmune diseases and bronchiectasis, The Multivariable Mendelian Randomization (MVMR) method was used to explore the effects of the confounding factors. Further investigation was conducted to identify potential intermediate factors using mediation analysis. Finally, the linkage disequilibrium score regression (LDSC) method was used to identify genetic correlations among complex traits. A series of sensitivity analyses was performed to validate the robustness of the results.

Results: The LDSC analysis revealed significant genetic correlations between BE and Crohn's disease (CD) (r_g = 0.220, P = 0.037), rheumatoid arthritis (RA) (r_g = 0.210, P = 0.021), and ulcerative colitis (UC) (r_g = 0.247, P = 0.023). However, no genetic correlation was found with other autoimmune diseases (P > 0.05). The results of the primary IVW analysis suggested that for every SD increase in RA, there was a 10.3% increase in the incidence of BE (odds ratio [OR] = 1.103, 95% confidence interval [CI] 1.055-1.154, P = 1.75×10^{-5}, FDR = 5.25×10^{-5}). Furthermore, for every standard deviation (SD) increase in celiac disease (CeD), the incidence of BE reduced by 5.1% (OR = 0.949, 95% CI 0.902-0.999, P = 0.044, FDR = 0.044). We also observed suggestive evidence corresponding to a 3% increase in BE incidence with T1DM (OR = 1.033, 95% CI 1.001-1.066, P = 0.042, FDR = 0.063). Furthermore, MVMR analysis showed that RA was an independent risk factor for BE, whereas mediator MR analysis did not identify any mediating factors. The sensitivity analyses corroborated the robustness of these findings.

Conclusion: LDSC analysis revealed significant genetic correlations between several autoimmune diseases and BE, and further MVMR analysis showed that RA is an independent risk factor for BE.

KEYWORDS

autoimmune diseases, bronchiectasis, rheumatoid arthritis, Mendelian randomization, Crohn's disease

Introduction

Bronchiectasis (BE) is a chronic respiratory disease characterized by the clinical symptoms of cough, sputum production, and hemoptysis in the presence of abnormal, irreversible dilatation of the bronchi that can be diagnosed using high-resolution chest computed tomography (CT) (1, 2). There has been a marked increase in the overall prevalence of bronchiectasis worldwide. In China, the prevalence of bronchiectasis increased 2.31-fold between 2013 and 2017, from 75.48 to 174.45 per 100,000 (3). Moreover, the prevalence of BE in females is higher than in males and also increases with age (4, 5). Importantly, BE is a heterogeneous syndrome caused by several underlying factors, such as pulmonary infections, cystic fibrosis (CF), primary ciliary dyskinesia (PCD), immunodeficiency disorders, allergic bronchopulmonary aspergillosis (ABPA), and autoimmune diseases. Recently, the association between BE and autoimmune diseases has been well recognized, and available studies have suggested that the oral, lung, and gut microbiota may affect the autoimmunity and structural integrity of the airways that contribute to BE (6). Neel et al. suggested that BE is highly prevalent in anti-myeloperoxidase (MPO) antineutrophil cytoplasmic autoantibody (ANCA)-associated vasculitis, and anti-MPO patients with BE have a higher risk of peripheral neuropathy (7). A systematic review and meta-analysis by Martin et al. demonstrated that BE may be a common extra-articular manifestation of rheumatoid arthritis (RA) (8), and anti-cyclic citrullinated peptide (CCP) antibodies (ACPAs) are associated with more severe RA-BE. However, the causal effects between BE and autoimmune diseases remain unclear.

Mendelian Randomization (MR) represents a methodological approach employing genetic variants as instrumental variables (IVs) sourced from genome-wide association studies (GWAS) to evaluate the causal relationship between a risk factor (exposure) and a resultant outcome (9). Contrary to traditional observational analyses, MR offers a more accurate estimation of the causal effect by considerably reducing the impact of confounders (10). The linkage disequilibrium score (LDSC) regression serves as a tool for estimating trait heritability, reflecting the percentage of trait variance ascribed to genetic determinants. Furthermore, LDSC assesses the genetic correlation between various traits using GWAS-derived summary statistics (11, 12). The objective of this research was to explore the plausible causal linkage between BE and autoimmune disorders.

Materials and methods

Study design

The foundational data for this investigation was retrieved from publicly available summary-level datasets from GWAS. Univariate Mendelian Randomization (UVMR), Multivariable Mendelian Randomization (MVMR), genetic correlation, and colocalization analyses were used to elucidate the causal interplay between autoimmune disorders and outcome phenotypes.

The selection of Instrumental Variables (IVs) for exposure was grounded in a tripartite criterion: i) the nominated genetic determinant, earmarked as the instrumental variable, must display a robust affiliation with the exposure; ii) the genetic determinant must not be intertwined with any potential confounders; and iii) the influence of genetic determinants on the outcome is channeled exclusively through its interaction with the exposure, thus eliminating the prospect of secondary routes (13). The architectural blueprint of the MR is illustrated in Figure 1 and Table 1, along with Supplementary File 1, which provides a comprehensive exposition of the summary statistics data repositories.

It is imperative to note that all encompassed GWAS investigation procured endorsements from the relevant academic oversight committees. Given that our study was based on a secondary analysis of publicly disclosed datasets, further ethical vetting was not required.

In order to preserve the integrity of our Mendelian Randomization approximations, the chosen Single Nucleotide Polymorphisms (SNPs) were obligated to align with the ensuing benchmarks:

Genetic instrument selection

(1) Each of the SNPs selected as IVs established a notable resonance with stipulated exposure at a genome-wide significance threshold ($p< 5\times10^{-8}$).

(2) Rigorous scrutiny ensured that the SNPs did not have associations with possible confounders nor shared interdependence, thereby mitigating biases originating from linkage disequilibrium ($r^2 < 0.001$, clumping distance = 10,000 kb).

Genetic instrument validation

(3) We used F-statistics (where $F = beta^2/se^2$, with beta symbolizing the SNP-exposure nexus and variance denoted by se) to assess the potency of the instrumental variables (14). An elevated F-statistic indicates pronounced instrumental vigor. Consequently, it was essential that all integrated SNPs exhibit an F-statistic transcending 10.

(4) We used the MR-Steiger filtration method to enhance the reliability of our conclusions, thereby ruling out variables that are more related to the outcomes than exposures (15).

(5) In the event of an SNP's absence from the outcome database, we used the SNiPa digital repository (accessible at http://snipa.helmholtz-muenchen.de/snipa3/) to locate a particular SNP. This platform used genotype data from a European cohort obtained from the 1000 Genomes Project Phase 3. Therefore, a surrogate SNP, reflecting linkage disequilibrium ($r2 > 0.8$) with the primary SNP was identified.

(6) The SNP's footprint on exposure juxtaposed with its impact on the outcome must mirror the identical allele. An SNP found to be discordant in this regard was invariably excised.

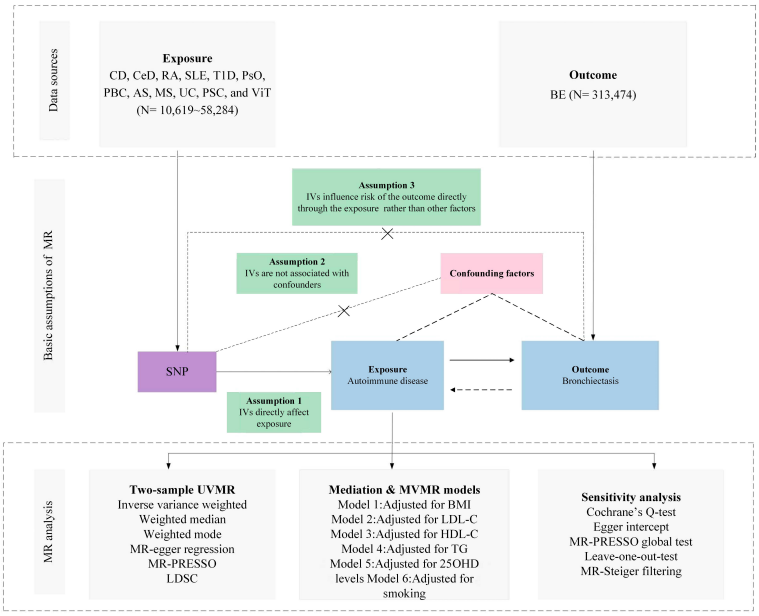


FIGURE 1
Overview of research design and analysis strategy. Overview of the research design. The MR framework is based on three fundamental MR assumptions, with MVMR analyses adjusting for six mediating factors for positive results. MR, Mendelian Randomization; MVMR: Multivariate Mendelian Randomization; UVMR, Univariate Mendelian Randomization; BMI, Body Mass Index; SNP, Single Nucleotide Polymorphism; MR- PRESSO, MR Pleiotropy Residual Sum and Outlier; LDL-C, Low Density Lipoprotein Cholesterol; HDL-C, High Density Lipoprotein Cholesterol; TG, Triglyceride; 25OHD, 25-hydroxyvitamin D; CD, Crohn's disease; CeD, Celiac disease; MS, Multiple sclerosis; RA, Rheumatoid arthritis; SLE, Systemic lupus erythematosus; UC, Ulcerative colitis; TID, Type 1 diabetes; PsO, Psoriasis; PSC, Primary sclerosing cholangitis; PBC, Primary biliary cirrhosis; AS, Ankylosing spondylitis; VIT, Vitiligo; BE, Bronchiectasis; LDSC, linkage disequilibrium score regression.

TABLE 1 Detailed information of data sources.

Explore or Outcome	Ref	Consortium	Ancestry	Participants
Phenotypes				
CD	28067908	de Lange KM et al	European	12,194 cases and 28,072 controls
CeD	22057235	Trynka et al	European	12,041 cases and 12,228 controls
MS	24076602	IMSGC	European	14,498 cases and 24,091 controls
RA	33310728	Ha E et al	European	14,361 cases and 43,923 controls
SLE	26502338	Bentham J et al	European	5,201 cases and 9,066 controls
UC	28067908	de Lange KM et al	European	12,366 cases and 33,609 controls
T1D	32005708	Forgetta V et al	European	9,266 cases and 15,574 controls
PsO	23143594	Tsoi LC et al	European	10,588 cases and 22,806 controls
PSC	27992413	IPSCSG	European	2,871 cases and 12,019 controls
PBC	34033851	Cordell HJ et al	European	8,021 cases and 16,489 controls
AS	23749187	Cortes A et al	European	9,069 cases and 1,550 controls
ViT	27723757	Jin Y et al	European	2,853 cases and 37,405 controls
BE	36653562	FinnGen Consortium	European	2,188 cases and 311,286 controls
Adjustment of the model				
LDL-C	24097068	GLGC	96% European	173,082 individuals
HDL-C	24097068	GLGC	96% European	187,167 individuals

(Continued)

TABLE 1 Continued

Explore or Outcome	Ref	Consortium	Ancestry	Participants
Adjustment of the model				
TG	24097068	GLGC	96% European	177,861 individuals
25OHD levels	32059762	Manousaki D et al.	European	441,291 individuals
Smoking	30643251	GSCAN	European	1,200,000 individuals
BMI	30239722	GIANT	European	694,649 individuals

BMI, body mass index; GWAS and Sequencing Consortium of Alcohol and Nicotine use; GIANT: Genetic Investigation of Anthropometric Traits; CD, Crohn's disease; CeD, Celiac disease; MS, Multiple sclerosis; RA, Rheumatoid arthritis; SLE, Systemic lupus erythematosus; UC, Ulcerative colitis; T1D, Type 1 diabetes; PsO, Psoriasis; PSC, Primary sclerosing cholangitis; PBC, Primary biliary cirrhosis; AS, Ankylosing spondylitis; ViT, Vitiligo; BE, Bronchiectasis; LDL-C, Low Density Lipoprotein Cholesterol; HDL-C, High Density Lipoprotein Cholesterol; TG, Triglyceride; 25OHD, 25-hydroxyvitamin D; GLGC, Global Lipids Genetics Consortium; IMSGC, International Multiple Sclerosis Genetics Consortium; IPSCSG, International PSC Study Group; Ref, reference (PUBMED ID).

Source of exposure and outcome phenotypes

For autoimmune diseases, all from large abstract-level GWAS studies, ulcerative colitis (UC) and Crohn's disease (CD) from de Lange KM et al. (16), celiac disease (CeD) from Trynka et al. (17), multiple sclerosis (MS) from International Multiple Sclerosis Genetics Consortium (IMSGC) (18), RA from Ha E et al. (19), systemic lupus erythematosus (SLE) from Bentham J et al. (20), type 1 diabetes (T1D) from Forgetta V et al. (21), psoriasis (PsO) from Tsoi LC et al. (22), primary sclerosing cholangitis (PSC) from International PSC Study Group (IPSCSG) (23), primary biliary cirrhosis (PBC) from Cordell HJ et al. (24), ankylosing spondylitis (AS) from Cortes A et al. (25), vitiligo (ViT) from Jin Y et al. (26), and for the outcome phenotype BE from FinnGen (R9) Consortium (27).

Data sources for possible mediators

We further obtained genetic associations for Body Mass Index (BMI) from the Genetic Investigation of Anthropometric Traits (GIANT) consortium (28), smoking from GWAS and Sequencing Consortium of Alcohol and Nicotine use (GSCAN) (29), triglycerides (TG), Low Density Lipoprotein Cholesterol (LDL-C) and High Density Lipoprotein Cholesterol (HDL-C) from Global Lipids Genetics Consortium (GLGC) (30), 25-hydroxyvitamin D (25OHD) levels from Manousaki D et al. (31).

Statistical analyses

Primary MR analysis

For the UVMR study, the Wald ratio test was used for exposure with only one instrument, and the multiplicative random-effects inverse-variance-weight (IVW) method was implemented for the causative assessment of multiple IVs (comprising two or more). This approach was further enhanced by incorporating both the MR-Egger and weight median techniques. The weightage in IVW is directly related to each SNP's Wald ratio estimate and inversely correlated with the variance estimate of each SNP's Wald ratio (32). When all genetic markers are judged valid, IVW provides estimates that are both consistent and efficient. Conversely, the weight median method stands out when over half of the genetic markers are deemed questionable, and the MR-Egger approach is adopted when all genetic markers are refutable (33). Stringent adjustment for multiple comparisons was performed using the False Discovery Rate (FDR). Following this adjustment, a P-value < 0.05 was considered indicative of a significant causal relationship. However, instances where the raw P-value was below 0.05, but the FDR-adjusted P-value exceeded this threshold were regarded as tentative.

Given the potential confounding effects of factors, such as BMI, smoking habits, lipid profiles (LDL-C, HDL-C, and TG), and 25OHD levels on the progression from exposure to outcome, subsequent MVMR analyses were performed. This study aimed to accurately quantify the direct causative effects of exposure on the results. When juxtaposed with the UVMR paradigm, the primary supposition of MVMR focus on genetic variability associated with one or more exposures, whereas the succeeding assumptions harmonize with the UVMR framework (34). A refined investigation was undertaken to ascertain the magnitude of mediation by certain factors. The initial step was to obtain the MR effect projections for exposure in relation to the outcome phenotypes using the IVW approach. Thereafter, multivariate MR analysis was performed to ascertain the impact of nine mediating factors on the outcome while concurrently considering exposure attributes. The indirect influence of the exposure was determined by multiplying the resulting estimates for each outcome. Finally, the division of the mediation effect by the overarching effect provided insight into the relative contribution of the mediators to the overall outcome.

Genetic correlation analysis

The LDSC regression, specifically tailored for GWAS summary data, serves as a robust approach for dissecting genetic correlations across complex diseases and traits. Notably, LDSC efficiently differentiates genuine polygenic signals from potential confounders such as cryptic relatedness and population stratification (35). A consequential genetic correlation, both

statistically and quantitatively robust, signifies that an overarching phenotypic correlation is not merely attributable to environmental confounders (35). The LDSC tool, accessible at (https://github.com/bulik/ldsc), was used to scrutinize the genetic intersections between exposure and an array of outcome phenotypes.

Sensitivity analysis

Within the framework of UVMR analysis, several tests were conducted to validate its rigor and authenticity. The heterogeneity of the selected genetic variants was assessed using Cochran's Q test, wherein a *P*-value of < 0.05 indicated pronounced discrepancies among the scrutinized SNPs (36). Employing the MR-Egger regression (37), this investigation discerned the potential for directional pleiotropy within the MR context. MR-Egger's intercept, with a *P*-value < 0.05, signified the presence of consequential directional pleiotropy despite the inherent limitations of this methodology (38). The MR Pleiotropy Residual Sum and Outlier (MR-PRESSO) approach was used to identify probable outliers and delve into horizontal pleiotropy, which was inferred when the global p-value was less than 0.05 (39). By excluding such outliers, the data correction was refined. An ensuing leave-one-out analysis elucidated the impact of singular SNPs on collective outcomes (40).

R^2 was calculated using the formula 2×MAF×(1-MAF) ×$beta^2$, where MAF denotes the minor allele frequency for each designated SNP. The cumulative values provided a coefficient essential for power computation (41). The determination of statistical potency was anchored on the mRnd platform (42) and is accessible at https://shiny.cnsgenomics.com/mRnd/.

Results

Genetic instrument selection and genetic correlation between phenotypes

The SNPs of each autoimmune disease were screened according to the genetic instrument selection process described above. Power calculations for bidirectional univariable MR analyses between autoimmune diseases including CD, CeD, MS, RA, SLE, UC, T1D, PsO, PSC, PBC, AS, ViT and BE, were performed. The study reported F- statistics exceeding 60 for all instrumental variants, signifying a robust reduction in bias from weak instruments. The SNPs selected as IVs ranged from 15 to 83, accounting for an explained variance of 2.59% to 1535.64% (Supplementary Table 1).

LDSC genetic correlation analyses were conducted to estimate the genetic correlation between different autoimmune diseases and BE. LDSC analysis revealed significant genetic correlations between BE and CD (r_g = 0.220, *P* = 0.037), RA (r_g = 0.210, *P* = 0.021), and UC (r_g = 0.247, *P* = 0.023) (Supplementary Table 2). However, no genetic correlation was found with other autoimmune diseases (*P* > 0.05). The SNP-based liability-scale heritability (h^2) ranged from 0.1% to 232.99%. Additionally, the genetic correlation between each autoimmune disease and BE was analyzed (Figure 2; Supplementary Table 3).

Association of genetically predicted autoimmune diseases with BE

A scatter plot illustrates the causal relationship between each autoimmune disease and BE (Supplementary Figure 1). After adjusting for multiple comparisons, the primary IVW analysis provided strong evidence for two causal relationships (Figure 3). Specifically, for each standard deviation (SD) increase in genetically predicted RA, there was a 10.3% increase in the incidence of BE (odds ration [OR] = 1.103, 95% CI 1.055-1.154, $P = 1.75\times10^{-5}$, FDR = 5.25×10^{-5}). Furthermore, for every SD increase in CeD, the incidence of BE was reduced by 5.1% (OR = 0.949, 95% CI 0.902-0.999, *P* = 0.044, FDR = 0.044). We also observed suggestive evidence corresponding to a 3% increase in BE incidence with T1DM (OR = 1.033, 95% CI 1.001-1.066, *P* = 0.042, FDR = 0.063). Additionally, we had 96%, 100%, and 92% statistical power to detect the associations of CeD, RA, and T1D with BE, with OR values of 1.103, 0.949, and 1.033, respectively (Supplementary Table 1). No other causal relationship evidence was found (P > 0.05, FDR > 0.05) (Table 2). Furthermore, MVMR analysis showed that RA was an independent risk factor for BE, whereas mediator MR analysis did not identify any mediating factors (Figure 4).

To avoid excessive bias effects, Cochran's Q test was performed to analyze the sensitivity of the MR results, and no evidence of heterogeneity was observed (*P>0.05*). Moreover, no horizontal pleiotropy was identified using the MR-Egger intercept test (*P>0.05*) or the MR-PRESSO global test (*P>0.05*). These analyses confirmed the robustness of the findings (Table 3). Leave-one-out analysis did not reveal any horizontal pleiotropy and further confirmed that the causal relationship was not influenced by any individual SNP (Supplementary File 1).

Discussion

In this study, we performed a comprehensive MR analysis to investigate the relationship between autoimmune diseases and BE. The results of LDSC analysis revealed significant genetic correlations between BE and CD, RA, and UC. However, beyond the aforementioned genetic correlations, no other genetic correlations were observed. Moreover, our objective in utilizing the MR analysis was to mitigate bias and confounding factors and identify causal associations. Interestingly, we found suggestive evidence of an association between T1D and BE. The MVMR analysis substantiated RA as an independent risk factor for BE, whereas the mediation MR analysis did not reveal any mediating model. While observational studies have inherent limitations, such

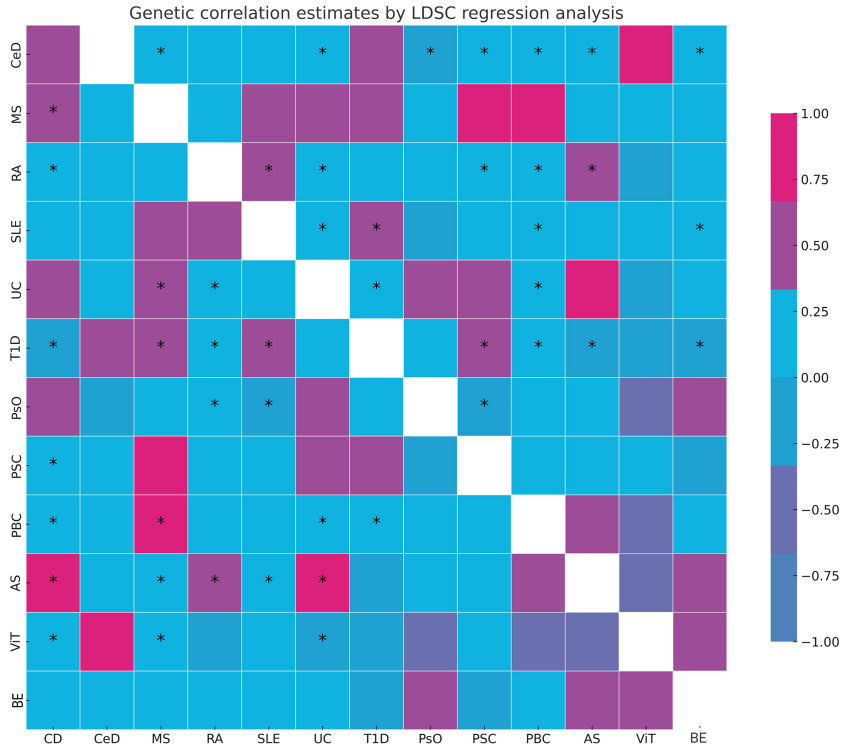


FIGURE 2
Summary of genetic correlation results. *: represents the presence of genetic correlation, $P<0.05$. LDSC, linkage disequilibrium score; CD, Crohn's disease; CeD, Celiac disease; MS, Multiple sclerosis; RA, Rheumatoid arthritis; SLE, Systemic lupus erythematosus; UC, Ulcerative colitis; TID, Type 1 diabetes; PsO, Psoriasis; PSC, Primary sclerosing cholangitis; PBC, Primary biliary cirrhosis; AS, Ankylosing spondylitis; ViT, Vitiligo; BE, Bronchiectasis.

as potential confounders and ambiguous causality, our MR approach aimed to mitigate these biases, providing clarity to these associations.

BE is characterized by damaged and dilated bronchi and is one of the most common pulmonary manifestations in patients with RA (43). Persistent pulmonary inflammation can inflict irreversible damage to the bronchi, culminating in BE (44). This notion is further supported by Lake et al., who suggested that pulmonary nodules, pleurisy, and air trapping in patients with RA might elevate the risk of anomalous pulmonary dilation (45). Additionally, Jin et al. found that the systemic inflammatory milieu in patients with RA might increase their susceptibility to other inflammatory disorders (46). Such inflammation can impair the bronchial walls, leading to BE. Moreover, Quirke et al. demonstrated that BE is a potent model for the initiation of autoimmunity in RA via bacterial infection of the lungs (47). CeD pathophysiologically correlates with autoimmune damage to the small intestine (48). This autoimmune response can potentially affect the lungs, wherein

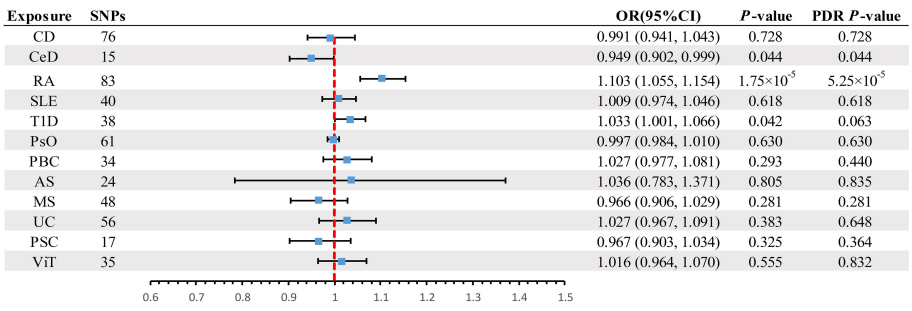


FIGURE 3
Summary of IVW results for the main UVMR analysis methods. IVW, Inverse variance weight; UVMR, Univariate Mendelian Randomization; SNP, Single Nucleotide Polymorphism; FDR, False Discovery Rate; OR, odds ratio; CI, confidence interval; CD, Crohn's disease; CeD, Celiac disease; MS, Multiple sclerosis; RA, Rheumatoid arthritis; SLE, Systemic lupus erythematosus; UC, Ulcerative colitis; T1D, Type 1 diabetes; PsO, Psoriasis; PSC, Primary sclerosing cholangitis; PBC, Primary biliary cirrhosis; AS, Ankylosing spondylitis; ViT, Vitiligo.

TABLE 2 Summary of UVMR analysis results.

Methods	SNPs	Crohn's disease (CD)						SNPs	Celiac disease (CeD)					
		OR	or_lci95	or_uci95	beta	P-value	FDR		OR	or_lci95	or_uci95	beta	P-value	FDR
Inverse variance weight	76	0.991	0.941	1.043	-0.009	0.728	0.728	15	0.949	0.902	0.999	-0.052	0.044	0.044
MR Egger	76	0.939	0.819	1.078	-0.063	0.375	0.5625	15	0.917	0.852	0.987	-0.087	0.039	0.044
Weight median	76	0.956	0.886	1.031	-0.045	0.238	0.5625	15	0.933	0.882	0.987	-0.069	0.015	0.044
Methods	SNPs	Rheumatoid arthritis (RA)						SNPs	Systemic lupus erythematosus (SLE)					
		OR	or_lci95	or_uci95	beta	P-value	FDR		OR	or_lci95	or_uci95	beta	P-value	FDR
Inverse variance weight	83	1.103	1.055	1.154	0.098	1.75E-05	5.25E-05	40	1.009	0.974	1.046	0.009	0.618	0.618
MR Egger	83	1.098	1.026	1.175	0.094	0.008	1.20E-02	40	0.954	0.884	1.031	-0.047	0.242	0.5055
Weight median	83	1.087	1.012	1.167	0.083	0.022	2.20E-02	40	1.027	0.973	1.084	0.027	0.337	0.5055
Methods	SNPs	Type 1 diabetes (T1D)						SNPs	Psoriasis (PsO)					
		OR	or_lci95	or_uci95	beta	P-value	FDR		OR	or_lci95	or_uci95	beta	P-value	FDR
Inverse variance weight	38	1.033	1.001	1.066	0.033	0.042	0.063	61	0.997	0.984	1.010	-0.003	0.630	0.63
MR Egger	38	1.071	1.021	1.122	0.068	0.007	0.021	61	0.991	0.975	1.008	-0.009	0.312	0.468
Weight median	38	1.035	0.990	1.082	0.034	0.131	0.131	61	0.989	0.973	1.005	-0.011	0.182	0.468
Methods	SNPs	Primary biliary cirrhosis (PBC)						SNPs	Ankylosing spondylitis (AS)					
		OR	or_lci95	or_uci95	beta	P-value	FDR		OR	or_lci95	or_uci95	beta	P-value	FDR
Inverse variance weight	34	1.027	0.977	1.081	0.027	0.293	0.4395	24	1.036	0.783	1.371	0.035	0.805	0.835
MR Egger	34	1.051	0.915	1.208	0.050	0.484	0.484	24	1.259	0.783	2.026	0.231	0.352	0.835
Weight median	34	1.080	1.004	1.163	0.077	0.040	0.12	24	1.041	0.715	1.515	0.040	0.835	0.835
Methods	SNPs	Multiple sclerosis (MS)						SNPs	Ulcerative colitis (UC)					
		OR	or_lci95	or_uci95	beta	P-value	FDR		OR	or_lci95	or_uci95	beta	P-value	FDR
Inverse variance weight	48	0.966	0.906	1.029	-0.035	0.281	0.281	56	1.027	0.967	1.091	0.027	0.383	0.648
MR Egger	48	0.886	0.780	1.007	-0.121	0.070	0.105	56	1.043	0.872	1.246	0.042	0.648	0.648
Weight median	48	0.912	0.829	1.004	-0.092	0.059	0.105	56	1.030	0.942	1.126	0.029	0.516	0.648
Methods	SNPs	Primary sclerosing cholangitis (PSC)						SNPs	Vitiligo (ViT)					
		OR	or_lci95	or_uci95	beta	P-value	FDR		OR	or_lci95	or_uci95	beta	P-value	FDR
Inverse variance weight	17	0.967	0.903	1.034	-0.034	0.325	0.364	35	1.016	0.964	1.070	0.016	0.555	0.8325

(Continued)

10.3389/fimmu.2024.1343480

10.3389/fimmu.2024.1343480

TABLE 2 Continued

Methods	SNPs	Primary sclerosing cholangitis (PSC)						Vitiligo (VT)						
		OR	or_lci95	or_uci95	beta	P-value	FDR	SNPs	OR	or_lci95	or_uci95	beta	P-value	FDR
MR Egger	17	0.883	0.792	0.984	-0.124	0.040	0.12	35	1.140	0.949	1.369	0.131	0.170	0.51
Weight median	17	0.963	0.887	1.045	-0.038	0.364	0.364	35	0.995	0.926	1.070	-0.005	0.903	0.903

damage to the intestine may precipitate the migration of inflammatory cells to the lungs, causing bronchitis. Dellaripa et al. also drew attention to dysregulated immune responses, suggesting that lungs are potential targets for autoimmune diseases (49). The primary hallmark of T1D is hyperglycemia, which stems from an immune attack on pancreatic β-cells. Barrett et al. suggested that microvascular damage correlated with T1DM might compromise the airway blood supply, contributing to BE (50). Lewis et al. have found that cystic fibrosis-associated diabetes (CFRD) often leads to poorer clinical outcomes in patients with CF including increased in pulmonary exacerbations, poorer lung function, and early mortality (51).

Emerging research has probed possible shared genetic pathways between autoimmune diseases and BE. Juge et al. have identified shared genetic susceptibilities between RA and respiratory ailments (52). Moreover, both CeD and T1D have been linked to gut microbiota dysbiosis (53, 54). An MR study by Huang et al. delineated a causal relationship between the gut microbiome and pulmonary diseases (55), hinting at the potential influence of the gut microbiota on pulmonary health and the predisposition to BE. Finally, as discussed by Litman et al., certain medications for autoimmune diseases may inadvertently exacerbate or induce pulmonary conditions (56).

The differences in the results between the MR and LDSC may be attributed to their distinct methodologies. MR relies on the use of genetic variants as instruments to infer causality, which assumes that these genetic variants affect the outcome solely through their impact on the exposure of interest and are not influenced by unmeasured confounding factors. Differently, LDSC focuses on quantifying genetic similarities between phenotypes and diseases. A significant genetic correlation detected by LDSC indicated shared genetic variations across multiple loci between the phenotypes. However, it is important to note that LDSC does not necessarily imply a causal relationship. In light of our findings, it is evident that there may be a causal relationship between BE and RA, and direct genetic correlations were detected using LDSC.

Our study has several strengths. First, our MR approach holistically analyzed the causative relationships between autoimmune diseases and BE. Second, the unique identification of SNPs as IVs in the European population minimized potential population stratification biases. Third, we employed rigorous methods with an F-statistic exceeding 10, reducing the biases from weak instruments. Fourth, we evaluated the confounding influence of the MVMR. Fifth, we relied on myriad sensitivity analyses based on statistical models and 'leave-one-out' techniques to enhance the reliability of the results. However, this study has several limitations. First, because of the lack of IVs achieving genome-wide significance for the outcomes, reverse causation inference was unfeasible. Second, summary-level GWAS data precluded subgroup analyses of autoimmune diseases and BE. Third, the sequencing and analysis methods for each autoimmune disease and BE may differ, contributing to the distinct results. Lastly, due to the summary-level GWAS data, the demographic data of the studies are absent, and further subgroup analysis of confounding factors, such as age and gender on autoimmune diseases and BE remains unknown.

10.3389/fimmu.2024.1343480

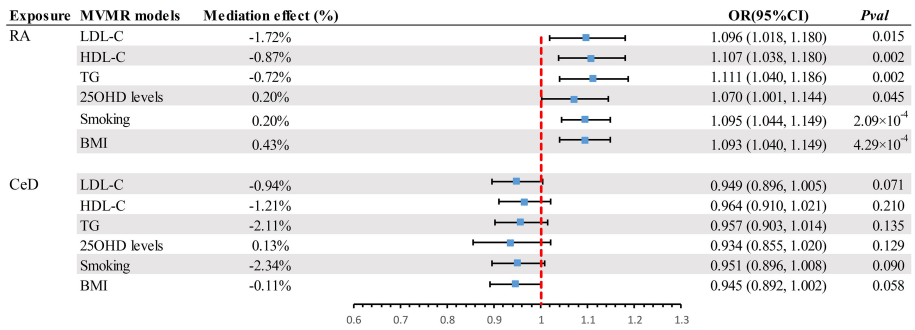


FIGURE 4
Summary of MVMR results. Estimating the impact of all positive exposure factors on the outcome through potential mediators. MVMR, Multivariate Mendelian Randomization; BMI, body mass index; OR, odds ratio; CI, confidence interval; LDL-C, Low Density Lipoprotein Cholesterol; HDL-C, High Density Lipoprotein Cholesterol; TG, Triglyceride; 250HD, 25-hydroxyvitamin D; CeD, Celiac disease; RA, Rheumatoid arthritis.

TABLE 3 Summary of sensitivity results.

Exposure	Outcome	MR-Egger intercept			MR-PRESSO global test			Cochrane's Q			Steiger_test		
		Intercept	SE	*Pval*	RSS_{obs}	*P*-value	*Outlier*	*Q*	*Q_df*	*Q_pval*	Direction	*Pval*	Filtered SNPs
CD	BE	0.009	0.011	0.414	84.328	0.300	*NA*	81.699	75	0.279	TRUE	0	NA
CeD		0.020	0.016	0.238	20.231	0.257	*NA*	17.773	14	0.217	TRUE	0	NA
MS		0.016	0.011	0.136	45.697	0.630	*NA*	42.816	47	0.646	TRUE	0	NA
RA		0.001	0.006	0.858	84.445	0.484	*NA*	82.900	82	0.451	TRUE	0	NA
SLE		0.022	0.014	0.118	37.818	0.551	*NA*	35.118	39	0.648	TRUE	0	NA
UC		-0.001	0.014	0.966	48.172	0.657	*NA*	46.698	55	0.780	TRUE	0	NA
T1D		-0.020	0.010	0.054	31.404	0.811	*NA*	29.506	37	0.805	TRUE	0	NA
PsO		0.012	0.012	0.309	80.749	0.061	*rs73695700*	78.790	60	0.052	TRUE	0	NA
PSC		0.040	0.020	0.064	27.022	0.113	*NA*	22.246	16	0.135	TRUE	0	NA
PBC		-0.006	0.018	0.729	42.556	0.190	*NA*	38.110	33	0.248	TRUE	0	NA
AS		-0.013	0.013	0.330	28.893	0.331	*NA*	26.232	23	0.290	TRUE	0	NA
ViT		-0.031	0.024	0.207	41.969	0.251	*rs28688825*	39.443	34	0.240	TRUE	0	NA

All results are after removing outliers and re-running the MR analysis. CD, Crohn's disease; CeD, Celiac disease; MS, Multiple sclerosis; RA, Rheumatoid arthritis; SLE, Systemic lupus erythematosus; UC, Ulcerative colitis; T1D, Type 1 diabetes; PsO, Psoriasis; PSC, Primary sclerosing cholangitis; PBC, Primary biliary cirrhosis; AS, Ankylosing spondylitis; ViT, Vitiligo; BE, Bronchiectasis; SNP, Single Nucleotide Polymorphisms.

Conclusion

LDSC analysis suggested significant genetic correlations between several autoimmune diseases and BE, and further MVMR analysis showed that RA was an independent risk factor for BE. These results provide genetic evidence for further mechanistic and clinical studies aimed at understanding the association between BE and autoimmune diseases.

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.

Author contributions

YS: Conceptualization, Data curation, Formal analysis, Methodology, Writing – original draft. YZ: Conceptualization, Data curation, Formal analysis, Methodology, Software, Writing – original draft. YC: Writing – original draft. JX: Conceptualization, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was sponsored by the National Natural Science Foundation of China (82330070, 81925001 to JX), the Innovation Program of Shanghai Municipal Education Commission (202101070007-E00097 to JX); the Program of Shanghai Municipal Science and Technology Commission (21DZ2201800 to JX). Shanghai Pujiang Program (22PJD065 to YS).

Acknowledgments

We thank all GWAS participants and investigators for making the summary statistics data publicly available.

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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1343480/full#supplementary-material

References

1. Elborn JS, Blasi F, Haworth CS, Ballmann M, Tiddens HAWM, Murris-Espin M, et al. Bronchiectasis and inhaled tobramycin: A literature review. *Respir Med*. (2022) 192:106728. doi: 10.1016/j.rmed.2021.106728

2. O'Donnell AE. Bronchiectasis — A clinical review. *New Engl J Med*. (2022) 387:533–45. doi: 10.1056/NEJMra2202819

3. Feng J, Sun L, Sun X, Xu L, Liu L, Liu G, et al. Increasing prevalence and burden of bronchiectasis in urban Chinese adults, 2013–2017: a nationwide population-based cohort study. *Respir Res*. (2022) 23:111. doi: 10.1186/s12931-022-02023-8

4. Quint JK, Millett ERC, Joshi M, Navaratnam V, Thomas SL, Hurst JR, et al. Changes in the incidence, prevalence and mortality of bronchiectasis in the UK from 2004 to 2013: a population-based cohort study. *Eur Respir J*. (2016) 47:186–93. doi: 10.1183/13993003.01033-2015

5. Henkle E, Chan B, Curtis JR, Aksamit TR, Daley CL, Winthrop KL. Characteristics and health-care utilization history of patients with bronchiectasis in US medicare enrollees with prescription drug plans, 2006 to 2014. *Chest*. (2018) 154:1311–20. doi: 10.1016/j.chest.2018.07.014

6. Leung JM, Olivier KN. Bronchiectasis and connective tissue diseases. *Curr Pulmonol Rep*. (2016) 5:169–76. doi: 10.1007/s13665-016-0154-8

7. Néel A, Espitia-Thibault A, Arrigoni P-P, Volteau C, Rimbert M, Masseau A, et al. Bronchiectasis is highly prevalent in anti-MPO ANCA-associated vasculitis and is associated with a distinct disease presentation. *Semin Arthritis Rheum*. (2018) 48:70–6. doi: 10.1016/j.semarthrit.2017.12.002

8. Martin LW, Prisco LC, Huang W, McDermott G, Shadick NA, Doyle TJ, et al. Prevalence and risk factors of bronchiectasis in rheumatoid arthritis: A systematic review and meta-analysis. *Semin Arthritis Rheum*. (2021) 51:1067–80. doi: 10.1016/j.semarthrit.2021.08.005

9. Tin A, Kottgen A. Mendelian randomization analysis as a tool to gain insights into causes of diseases: A primer. *J Am Soc Nephrol*. (2021) 32:2400–7. doi: 10.1681/ASN.2020121760

10. Davies NM, Holmes MV. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ*. (2018) 362:k601. doi: 10.1136/bmj.k601

11. Bulik-Sullivan BK, Neale BM. LD score regression distinguishes confounding from polygenicity in GWAS. *Nat Genet*. (2015) 47:291–5. doi: 10.1038/ng.3211

12. Tashman KC, Cui R, O'Connor LJ, Neale BM, Finucane HK. Significance testing for small annotations in stratified LD-Score regression. *medRxiv*. (2021). doi: 10.1101/2021.03.13.21249938

13. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Stat Med*. (2008) 27:1133–63. doi: 10.1002/sim.3034

14. Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int J Epidemiol*. (2016) 45:1961–74. doi: 10.1093/ije/dyw220

15. Hemani G, Tilling K, Davey Smith G. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PloS Genet*. (2017) 13:e1007081. doi: 10.1371/journal.pgen.1007081

16. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet*. (2017) 49:256–61. doi: 10.1038/ng.3760

17. Trynka G, Hunt KA, Bockett NA, Romanos J, Mistry V, Szperl A, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet*. (2011) 43:1193–201. doi: 10.1038/ng.998

18. International Multiple Sclerosis Genetics Consortium (IMSGC), Beecham AH, Patsopoulos NA, Xifara DK, Davis MF, Kemppinen A, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet*. (2013) 45:1353–60. doi: 10.1038/ng.2770

19. Ha E, Bae S-C, Kim K. Large-scale meta-analysis across East Asian and European populations updated genetic architecture and variant-driven biology of rheumatoid arthritis, identifying 11 novel susceptibility loci. *Ann Rheum Dis*. (2021) 80:558–65. doi: 10.1136/annrheumdis-2020-219065

20. Bentham J, Morris DL, Graham DSC, Pinder CL, Tombleson P, Behrens TW, et al. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat Genet*. (2015) 47:1457–64. doi: 10.1038/ng.3434

21. Forgetta V, Manousaki D, Istomine R, Ross S, Tessier MC, Marchand L, et al. Rare genetic variants of large effect influence risk of type 1 diabetes. *Diabetes*. (2020) 69:784–95. doi: 10.2337/db19-0831

22. Tsoi LC, Spain SL, Knight J, Ellinghaus E, Stuart PE, Capon F, et al. Identification of 15 new psoriasis susceptibility loci highlights the role of innate immunity. *Nat Genet*. (2012) 44:1341–8. doi: 10.1038/ng.2467

23. Ji S-G, Juran BD, Mucha S, Folseraas T, Jostins L, Melum E, et al. Genome-wide association study of primary sclerosing cholangitis identifies new risk loci and quantifies the genetic relationship with inflammatory bowel disease. *Nat Genet*. (2017) 49:269–73. doi: 10.1038/ng.3745

24. Cordell HJ, Fryett JJ, Ueno K, Darlay R, Aiba Y, Hitomi Y, et al. An international genome-wide meta-analysis of primary biliary cholangitis: Novel risk loci and candidate drugs. *J Hepatol*. (2021) 75:572–81. doi: 10.1016/j.jhep.2021.04.055

25. International Genetics of Ankylosing Spondylitis Consortium (IGAS), Cortes A, Hadler J, Pointon JP, Robinson PC, Karaderi T, et al. Identification of multiple risk variants for ankylosing spondylitis through high-density genotyping of immune-related loci. *Nat Genet*. (2013) 45:730–8. doi: 10.1038/ng.2667

26. Jin Y, Andersen G, Yorgov D, Ferrara TM, Ben S, Brownson KM, et al. Genome-wide association studies of autoimmune vitiligo identify 23 new risk loci and highlight key pathways and regulatory variants. *Nat Genet*. (2016) 48:1418–24. doi: 10.1038/ng.3680

27. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*. (2023) 613:508–18. doi: 10.1038/s41586-022-05473-8

28. Pulit SL, Stoneman C, Morris AP, Wood AR, Glastonbury CA, Tyrrell J, et al. Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of European ancestry. *Hum Mol Genet*. (2019) 28:166–74. doi: 10.1093/hmg/ddy327

29. Liu M, Jiang Y, Wedow R, Li Y, Brazel DM, Chen F, et al. Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nat Genet*. (2019) 51:237–44. doi: 10.1038/s41588-018-0307-5

30. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. (2013) 45:1274–83. doi: 10.1038/ng.2797

31. Manousaki D, Mitchell R, Dudding T, Haworth S, Harroud A, Forgetta V, et al. Genome-wide association study for vitamin D levels reveals 69 independent loci. *Am J Hum Genet*. (2020) 106:327–37. doi: 10.1016/j.ajhg.2020.01.017

32. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. *Elife*. (2018) 7:e34408. doi: 10.7554/eLife.34408

33. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol*. (2016) 40:304–14. doi: 10.1002/gepi.21965

34. Burgess S, Thompson SG. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am J Epidemiol*. (2015) 181:251–60. doi: 10.1093/aje/kwu283

35. Tobin MD, Minelli C, Burton PR, Thompson JR. Commentary: development of Mendelian randomization: from hypothesis test to "Mendelian deconfounding". *Int J Epidemiol*. (2004) 33:26–9. doi: 10.1093/ije/dyh016

36. Kulinskaya E, Dollinger MB, Bjørkestøl K. On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta-analysis of risk difference. *Res Synth Methods*. (2020) 11:920. doi: 10.1002/jrsm.1446

37. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol*. (2017) 32:377–89. doi: 10.1007/s10654-017-0255-x

38. Wu F, Huang Y, Hu J, Shao Z. Mendelian randomization study of inflammatory bowel disease and bone mineral density. *BMC Med*. (2020) 18:312. doi: 10.1186/s12916-020-01778-5

39. Verbanck M, Chen C-Y, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet*. (2018) 50:693–8. doi: 10.1038/s41588-018-0099-7

40. Cheng H, Garrick DJ, Fernando RL. Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction. *J Anim Sci Biotechnol*. (2017) 8:38. doi: 10.1186/s40104-017-0164-6

41. Guan W, Steffen BT, Lemaitre RN, Wu JHY, Tanaka T, Manichaikul A, et al. Genome-wide association study of plasma N6 polyunsaturated fatty acids within the CHARGE consortium. *Circ Cardiovasc Genet*. (2014) 7:321–31. doi: 10.1161/CIRCGENETICS.113.000208

42. Brion M-JA, Shakhbazov K, Visscher PM. Calculating statistical power in Mendelian randomization studies. *Int J Epidemiol*. (2013) 42:1497–501. doi: 10.1093/ije/dyt179

43. Duarte AC, Porter J, Leandro MJ. Bronchiectasis in rheumatoid arthritis. A clinical appraisial. *Joint Bone Spine*. (2020) 87:419–24. doi: 10.1016/j.jbspin.2019.12.006

44. Wilczynska MM, Condliffe AM, McKeon DJ. Coexistence of bronchiectasis and rheumatoid arthritis: revisited. *Respir Care*. (2013) 58:694–701. doi: 10.4187/respcare.01857

45. Overview of pleuropulmonary diseases associated with rheumatoid arthritis - UpToDate. Available at: https://www.uptodate.com/contents/overview-of-pleuropulmonary-diseases-associated-with-rheumatoid-arthritis (Accessed 31 Aug 2023).

46. Jin Z, Hao D, Song Y, Zhuang L, Wang Q, Yu X. Systemic inflammatory response index as an independent risk factor for ischemic stroke in patients with rheumatoid arthritis: a retrospective study based on propensity score matching. *Clin Rheumatol*. (2021) 40:3919–27. doi: 10.1007/s10067-021-05762-z

47. Quirke A-M, Perry E, Cartwright A, Kelly C, De Soyza A, Eggleton P, et al. Bronchiectasis is a model for chronic bacterial infection inducing autoimmunity in rheumatoid arthritis. *Arthritis Rheumatol*. (2015) 67:2335–42. doi: 10.1002/art.39226

48. Catassi C, Verdu EF, Bai JC, Lionetti E. Coeliac disease. *Lancet*. (2022) 399:2413–26. doi: 10.1016/S0140-6736(22)00794-2

49. MD PFD. Autoimmune lung disease: Early recognition and treatment helps, in: *Harvard Health*. (2020). Available at: https://www.health.harvard.edu/blog/autoimmune-lung-disease-early-recognition-and-treatment-helps-2020062420339 (Accessed 31 Aug 2023).

50. Barrett EJ, Liu Z, Khamaisi M, King GL, Klein R, Klein BEK, et al. Diabetic microvascular disease: an endocrine society scientific statement. *J Clin Endocrinol Metab*. (2017) 102:4343–410. doi: 10.1210/jc.2017-01922

51. Lewis C, Blackman SM, Nelson A, Oberdorfer E, Wells D, Dunitz J, et al. Diabetes-related mortality in adults with cystic fibrosis. Role of genotype and sex. *Am J Respir Crit Care Med*. (2015) 191:194–200. doi: 10.1164/rccm.201403-0576OC

52. Juge P-A, Borie R, Kannengiesser C, Gazal S, Revy P, Wemeau-Stervinou L, et al. Shared genetic predisposition in rheumatoid arthritis-interstitial lung disease and familial pulmonary fibrosis. *Eur Respir J*. (2017) 49:1602314. doi: 10.1183/13993003.02314-2016

53. Thomas DA, Rosenthal GA, Gold DV, Dickey K. Growth inhibition of a rat colon tumor by L-canavanine. *Cancer Res*. (1986) 46:2898–903.

54. Zhou H, Sun L, Zhang S, Zhao X, Gang X, Wang G. Evaluating the causal role of gut microbiota in type 1 diabetes and its possible pathogenic mechanisms. *Front Endocrinol (Lausanne)*. (2020) 11:125. doi: 10.3389/fendo.2020.00125

55. Huang S, Li J, Zhu Z, Liu X, Shen T, Wang Y, et al. Gut microbiota and respiratory infections: insights from Mendelian randomization. *Microorganisms*. (2023) 11:2108. doi: 10.3390/microorganisms11082108

56. Mleczko M, Gerkowicz A, Krasowska D. Chronic inflammation as the underlying mechanism of the development of lung diseases in psoriasis: A systematic review. *Int J Mol Sci*. (2022) 23:1767. doi: 10.3390/ijms23031767

frontiers | Frontiers in Immunology

TYPE Review
PUBLISHED 10 June 2024
DOI 10.3389/fimmu.2024.1409555

OPEN ACCESS

EDITED BY
Xu-jie Zhou,
Peking University, China

REVIEWED BY
Hiufung Yip,
Hong Kong Baptist University, Hong Kong SAR, China
Miha Lavric,
University of Maribor, Slovenia

*CORRESPONDENCE
Dongyi He
dongyihe@medmail.com.cn
Fubo Wang
wangfubo@gxmu.edu.cn

RECEIVED 30 March 2024
ACCEPTED 24 May 2024
PUBLISHED 10 June 2024

CITATION
Shi Y, Zhou M, Chang C, Jiang P, Wei K, Zhao J, Shan Y, Zheng Y, Zhao F, Lv X, Guo S, Wang F and He D (2024) Advancing precision rheumatology: applications of machine learning for rheumatoid arthritis management.
Front. Immunol. 15:1409555.
doi: 10.3389/fimmu.2024.1409555

Advancing precision rheumatology: applications of machine learning for rheumatoid arthritis management

Yiming Shi[1,2,3], Mi Zhou[1,3], Cen Chang[1,3], Ping Jiang[1,2,3], Kai Wei[1,2,3], Jianan Zhao[1,2,3], Yu Shan[1,2,3], Yixin Zheng[1,2,3], Fuyu Zhao[1,2,3], Xinliang Lv[4], Shicheng Guo[1], Fubo Wang[5,6]* and Dongyi He[1,2,3]*

[1]Department of Rheumatology, Shanghai Guanghua Hospital of Integrative Medicine, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [2]Guanghua Clinical Medical College, Shanghai University of Traditional Chinese Medicine, Shanghai, China, [3]Institute of Arthritis Research in Integrative Medicine, Shanghai Academy of Traditional Chinese Medicine, Shanghai, China, [4]Traditional Chinese Medicine Hospital of Inner Mongolia Autonomous Region, Hohhot, Inner Mongolia Autonomous Region, China, [5]Guangxi Key Laboratory for Genomic and Personalized Medicine, Guangxi Collaborative Innovation Center for Genomic and Personalized Medicine, Guangxi Medical University, Nanning, Guangxi, China, [6]Department of Urology, Affiliated Tumor Hospital of Guangxi Medical University, Guangxi Medical University, Nanning, Guangxi, China

Rheumatoid arthritis (RA) is an autoimmune disease causing progressive joint damage. Early diagnosis and treatment is critical, but remains challenging due to RA complexity and heterogeneity. Machine learning (ML) techniques may enhance RA management by identifying patterns within multidimensional biomedical data to improve classification, diagnosis, and treatment predictions. In this review, we summarize the applications of ML for RA management. Emerging studies or applications have developed diagnostic and predictive models for RA that utilize a variety of data modalities, including electronic health records, imaging, and multi-omics data. High-performance supervised learning models have demonstrated an Area Under the Curve (AUC) exceeding 0.85, which is used for identifying RA patients and predicting treatment responses. Unsupervised learning has revealed potential RA subtypes. Ongoing research is integrating multimodal data with deep learning to further improve performance. However, key challenges remain regarding model overfitting, generalizability, validation in clinical settings, and interpretability. Small sample sizes and lack of diverse population testing risks overestimating model performance. Prospective studies evaluating real-world clinical utility are lacking. Enhancing model interpretability is critical for clinician acceptance. In summary, while ML shows promise for transforming RA management through earlier diagnosis and optimized treatment, larger scale multisite data, prospective clinical validation of interpretable models, and testing across diverse populations is still needed. As these gaps are addressed, ML may pave the way towards precision medicine in RA.

KEYWORDS

ML, rheumatoid arthritis, precision medicine, diagnosis, treatment

1 Introduction

Rheumatoid arthritis (RA) is a prevalent autoimmune disorder characterized by inflammation and discomfort in numerous small joints, potentially leading to joint deformity and impaired functionality. Furthermore, it ranks among the primary contributors to chronic disability (1). Furthermore, RA not only impacts the joints but also has implications for other bodily systems, including the cardiovascular and respiratory systems, leading to an elevated susceptibility to conditions such as myocardial infarction, stroke, and pulmonary fibrosis (2, 3). Chronic illnesses and persistent pain can result in psychological distress for patients, manifesting as symptoms of depression and anxiety (4). Hence, it is imperative to promptly identify individuals with a high susceptibility to RA in order to facilitate early diagnosis and anticipate the potential severity of disease progression. Furthermore, the timely administration of efficacious medications is essential in impeding the advancement of the disease.

The phrase "machine learning (ML)" surged in popularity in the late 1990s in the field of artificial intelligence (5). In the past decade, ML has made significant advancements as a result of the increased availability of data and improvements in algorithms, enabling the identification of complex patterns and correlations within datasets (6). The biomedical field has experienced a significant increase in data volume, ranging from molecular details to comprehensive information on the human body system, due to advancements in high-throughput sequencing technologies, electronic health records, and medical imaging (7). Healthcare providers and researchers are currently facing a growing number of clinical challenges, leading them to explore ways to enhance decision-making effectiveness, refine personalized treatment strategies, and optimize resource allocation methods. ML is uniquely positioned to extract valuable patterns and insights from large datasets, potentially automating and enhancing the efficiency of healthcare decision-making and services. The incremental incorporation of biomedicine with various disciplines, including computational science, mathematics, and statistics, has spurred interdisciplinary partnerships, leading to accelerated progress in the application of ML in the field of biomedicine (8). In the clinical practice of RA, Rheumatoid Factor (RF) and Anti-Citrullinated Protein Antibody (ACPA) serve as crucial diagnostic biomarkers for RA, playing key roles in its diagnosis. However, approximately 20-25% of RA patients are seronegative, posing challenges to early diagnosis and potentially leading to delayed diagnosis and treatment (9). With the advent and development of biologics, significant progress has been made in the treatment of RA. Nevertheless, many RA patients exhibit poor responses to drug treatments, failing to achieve sustained remission (10), and currently, it is not possible to predict which treatment drugs will have the best therapeutic effect on individual patients. The accumulation of biomedical big data may provide new insights into better understanding the heterogeneity of RA (11). With the increase in data volume and complexity, traditional statistical analysis methods have become insufficient, especially when dealing with nonlinear relationships and complex interactions between variables (12). These unmet needs pose challenges to the precision medicine of RA. Using ML techniques for data processing and pattern recognition to build predictive models for RA can assist clinicians in making more accurate data-driven decisions (13). Therefore, understanding the prevalent ML algorithms in RA, their effectiveness, and potential applications is crucial. Our study is dedicated to evaluating recent literature on applications of ML in RA classification and outcome prediction, with the goal of offering a dependable benchmark for reference and guiding future research endeavors. By enhancing the utilization of sophisticated modeling in RA and advocating for precision medicine in the field, our work aims to propel advancements in RA treatment and management.

2 ML algorithms to enhance precision rheumatology

ML, a crucial component of artificial intelligence, is divided into two main categories: supervised and unsupervised learning. Supervised learning employs labeled training datasets to identify patterns and relationships. Upon training, the model can predict or classify new data inputs, yielding corresponding results. This method utilizes a range of algorithms, such as logistic regression, random forests, gradient boosting, and decision trees. Each algorithm contributes uniquely to the robustness and accuracy of predictive outcomes, making supervised learning integral to advancements in data-driven research methodologies (14). Supervised learning is divided into two principal methodologies: classification and regression (15). Classification methodologies segregate patients according to distinct characteristics (16). By employing datasets comprising genetic information, gene expression profiles, and clinical indicators from patients with RA, algorithms can be trained to identify RA patients within populations, as well as to ascertain which patients exhibit optimal responses to specific treatments. Regression models, on the other hand, are designed to predict continuous outcomes (17), such as disease activity scores and response rates to treatments in RA patients, thus facilitating personalized monitoring and management to optimize treatment efficacy. In contrast, unsupervised learning explores inherent patterns and relationships in datasets without predetermined labels (18). Clustering algorithms, an exemplary application of unsupervised learning, automatically group data into multiple clusters to maximize intra-cluster similarity and minimize inter-cluster similarity, aiding significantly in RA research by identifying potential patient subgroups who may exhibit favorable responses to specific treatments or distinct disease progression patterns. Deep learning, employing Artificial Neural Network (ANN) technologies, enhances the analysis and prediction of complex data through sophisticated non-linear mapping relationships (19). Particularly, Convolutional Neural Networks (CNNs) in deep learning architectures are adept in processing image data (20), enabling automatic feature learning from multiple convolutional layers which assist physicians in identifying early signs of arthritis or disease progression in X-ray or Magnetic Resonance Imaging (MRI)

images of RA patients. In summary, supervised and unsupervised learning each serve specific roles, while deep learning technologies enhance the capability of these methods to process complex data, thereby effectively advancing the field of precision rheumatology.

In the preprocessing phase, data cleaning and organization are paramount, involving the removal of duplicates and correction of anomalies (21). Furthermore, feature engineering plays a critical role in identifying predictors (x) that significantly influence the target variable (y) through strategic selection and transformation of data, a crucial task in supervised learning. Accurate feature selection not only enhances the precision of the model but also its interpretability. When constructing predictive models, addressing the challenge of managing a large volume of available features is commonplace. While the use of advanced and efficient algorithms is vital, ineffective predictive information derived from these features, or the presence of numerous irrelevant variables, can impair model performance. Implementing key feature selection strategies is crucial, including statistical filtering, wrapper methods, and advanced embedded techniques (22–24). For instance, Random Forest assesses feature importance by calculating their contribution to model accuracy (25), whereas Logistic Regression identifies key influencing factors by analyzing the magnitude and direction of coefficients (26). Through rigorous feature selection, the dimensionality and complexity of the dataset are effectively reduced, thereby enhancing the interpretability and practical application of the predictive model in clinical decision-making (22). For example, identifying RA patients with specific genetic mutations through feature selection has indicated that these individuals respond more positively to methotrexate, a principal drug for RA treatment. This insight assists physicians in devising targeted treatment plans, thereby improving therapeutic outcomes.

ML algorithms are increasingly recognized as powerful analytical tools in the field of RA research. As depicted in Figure 1, they provide assistance across multiple domains, including diagnosis, disease progression forecasting, prediction of treatment responses, and identification of potential complications. These computational tools are guiding the field towards a more refined and individualized approach, allowing clinicians and researchers to explore the complexities of RA with greater accuracy.

3 ML models in precision diagnosis and therapeutics for RA

A variety of predictive models have been built using ML algorithms in RA research. Presented in Table 1 is the appraisal of performance when these ML models serve as classifiers across a multitude of data types from various sources. The functionalities of these classifiers include identification of individuals at risk for RA, diagnosis and differentiation of subtypes, discrimination of disease activity levels, forecasting of treatment outcomes as effective or ineffective, and predicting the presence or absence of comorbidities.

3.1 Stratification of RA risk cohorts

Identifying individuals at risk for RA is crucial for early intervention, which has been shown to yield substantially better outcomes when applied during the preclinical stages rather than after the overt development of clinically significant arthritis (70). Specifically, by identifying individuals at high risk and conducting

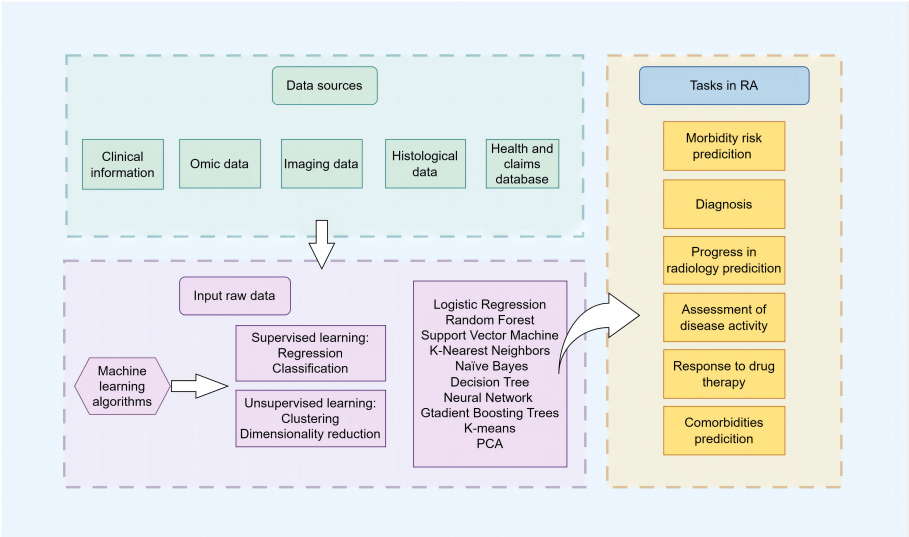


FIGURE 1
Schematic overview of clinical prediction in RA using ML The schematic illustrates the comprehensive workflow and applications of ML algorithms in the management of RA. It encapsulates the stepwise process from data collection, including electronic health records, imaging, and multi-omics data, through data preprocessing and feature engineering, to model training and validation phases. The central part of the diagram highlights the primary domains of ML application in RA: risk prediction, diagnosis and subtype classification, prediction of disease activity and progression, treatment response, and comorbidity identification for RA. It emphasizes the iterative optimization of models and the synergy between clinical and computational insights aimed at advancing early diagnosis, personalized treatments, and patient outcomes in RA management.

TABLE 1 Application of ML in RA.

Task	Sample Size	Features	ML algorithms	Performance	Ref
Risk Prediction	Training set: RA patients: n = 599 Controls: n = 1673 Test set 1:RA: n = 125 Controls: n = 349 Test set 2:RA: n = 127 Controls: n = 355 Test set 3:RA: n = 127 Controls: n = 355	9 SNPs	LR, SVM, Naïve Bayes, RF, XGBoost	AUC > 0.9	(27)
	RA or no arthritis: n =17,366 Training set: n = 8683 Validation set: n = 4342 Test set: n = 4341	Age, gender, race, high BMI, gout, diabetic, smoked, sleep, blood pressure, patient health questionnaire, income to poverty ratio	Bayes	validation set: AUC = 0.826 test set: AUC = 0.805	(28)
	Training cohort: RA: n=47 non-RA: n=64 Test cohort: UA: n = 62	the Leiden prediction rule, 12-gene risk metric	SVM	AUC = 0.84	(29)
	UA: n = 72, RA: n = 8, HD: n = 13	cpg sites, clinical parameters	LR, SVM, RF	AUC: 0.875-1	(30)
Diagnosis	hand radiograph images: Training set: RA: n = 256 OA: n = 262 Normal: n = 231, Others: n = 242; Validation set: RA: n = 56 OA: n = 57 Normal: n = 51 Others: n = 53; Test set: RA: n = 56 OA: n = 58 Normal: n = 51 Others: n = 53	–	CNNs	Classification of RA and normal: AUC = 0.97 Classification of RA and OA and normal: Acc = 0.806 Classification of RA and OA and normaland others: Acc = 0.844	(31)
	1337 RA ultrasound images of 208 patients	–	DL	Classification of synovial proliferation or not: Group1/Group2/ Group3: AUC = 0.863/0.861/ 0.886 Classification of healthy and diseased: Group1/Group2/ Group3: AUC=0.848/ 0.864/0.916	(32)
	Training set: HC: n = 100 RA: n = 100 Validation set: HC: n = 18 RA: n = 20	hand images, Age, gripforce	BayesNet, NaïveBayes, Logistic, k-NN, RF,etc.	Classification of RA and HC Acc = 0.947 Sen = 0.95 Spe = 0.944 AUC = 0.971	(33)
	Training set: GSE93272, GSE45291, GSE74143, GSE65010, GSE15573, GSE61635, GSE65391,	15 key genes		AUC > 0.85	(34)

(Continued)

TABLE 1 Continued

Task	Sample Size	Features	ML algorithms	Performance	Ref
	GSE138458, GSE143272, GSE113469, GSE50772 Test set: GSE55457,		LASSO, SVM, RF, XGBoost, BPNN, CNN		
	GSE93272, GSE17755	MAPK3, ACTB, ACTG1, VAV2, PTPN6, ACTN1	LASSO	Training set: AUC= 0.801 Validation set: AUC= 0.979	(35)
	Uninflamed: n = 10 Resolving arthritis: n = 9 Early RA: n = 17 Established RA: n = 12	cytokine, chemokine	GMLVQ	RA vs. non-inflamed group: AUC = 0.996 Early RA vs. resolved arthritis group: AUC = 0.764	(36)
	Training set: GSE12021, GSE55235, GSE55457, GSE55584 Validation set: Dataset1: GSE89408 Dataset2: GSE77298, GSE153015	m6A methylation regulators	RF, Rpart, LASSO, XGBoost, LR	Classification of RA and HC AUC = 0.85 (IGF2BP3) AUC = 0.85 (YTHDC2)	(37)
	Serum of 225 RA patients and 100 HC Discovery set: n = 243 Validation set: n = 82	26 metabolites and lipids	LR, RF, SVM	Classification of RA and HC: AUC = 0.91 Sen = 0.897 Spe = 0.906	(38)
	Test cohort: RA: n=36 OA: n=18 HC: n=18 Validation cohort: RA: n=24 OA: n=12 HC: n=12	3 groups of differentially expressed proteins	RF	Classification of RA: AUC = 0.9949 Classification of ACPA-positive RA patients: AUC = 0.9913 Classification of ACPA-negative RA patients: AUC = 1.0	(39)
	IBD: n = 14, MS: n = 7, RA: n = 5, JIA: n = 3, SLE: n = 3, T1D: n = 2, BS: n = 2, AS: n = 2, APS: n = 1、PSC: n = 1, MG: n = 1, ReA: n = 1	gut microbiome	RF, SVM , XGBoost, Ridge Regression	Classification of RA and IBD: AUC > 0.86 Classification of RA and MS: AUC > 0.96	(40)
	Discovery cohort: 167 RA and 91 controls Validation cohort: 12 SLE、32 RA and 32 controls	miR-22-3p, miR-24-3p, miR-96-5p, miR-134-5p, miR-140-3p, miR-627-5p	LASSO, RF, LR	Classification of RA and non-RA: AUC = 0.71 Classification of ACPA-positive RA and others: AUC = 0.73 Classification of ACPA-negative RA and others: AUC = 0.73	(41)
	H&E-stained images of TKR explant synovium (OA: n = 147, RA: n = 60) Training set: n = 166 Test set: n = 41	14 pathologist-scored features、computer vision-quantified cell density	RF	Classification of RA and OA AUC = 0.91	(42)
	129 synovial tissue samples RA: n = 123 OA: n = 6	histologic scoring	SVM	Classification of the high inflammatory subtype and others: AUC = 0.88 Classification of the low inflammatory subtype and others: AUC = 0.71 Classification of the mixed subtype and others: AUC = 0.59	(43)

(Continued)

TABLE 1 Continued

Task	Sample Size		Features	ML algorithms	Performance	Ref
Disease activity/ imaging progression	Hanyang Bae RA Cohort: No progression: n = 118 Severe progression: n = 120 NARAC Cohort: No progression: n = 68 Severe progression: n = 86		genetic and clinical factors	SVM	Classification of radiologic progression and no progression AUC = 0.7872	(44)
	ultrasound images from RA patients Training set: n = 1678 Testing set: n = 322		–	CNN	Distinguishing class 0 from the other classes: AUC = 0.96 Distinguishing class1 from class 2 and 3 classes: AUC = 0.94 Distinguishing class 2 from class 3 classes: AUC = 0.93	(45)
	135 visits from 41 patients		dose percentage change, the DAS-28 ESR score, ESR, disease duration, CRP, and the duration of remission at study entry	LR, KNN, NB, RF, Stacking-Meta Classifier	Classification of flare yes and. flare no AUC: 0.72 - 0.81	(46)
	stable RA patients: n = 130 training set: n = 104 test set: n= 26		baseline serum proteomics	LASSO, XGBoost	Classification of flare and remission AUC = 0.8	(47)
	2 electronic health record platforms UH Cohort: n = 578 (Training Cohort : Test Cohort: n= 116) SNH Cohort: n= 242 (Training Cohort: n = 125, Test: n = 117)		medications, patient demographics, laboratories, and prior measures of disease activity.	DL	Classification of controlled and uncontrolled UH training model test in UH Test Cohort: AUC = 0.91 UH training model test in SNH test Cohort: AUC = 0.74	(48)
	300 RA patients		laboratory data, medicare claims and medications	LR	Classification of high/ moderate and low disease activity/ remission AUC = 0.76	(49)
	Optum dataset:n = 68,608 Externally validatiation: IBM CCAE: n = 75,579 IBM MDCD: n = 7,537 IBM MDCR: n = 36,090		health service utilization, demographics, prescription claims for immunosuppressants, steroids, DMARDs, pain medications, and other comorbid conditions.	regularized LASSO, LR, RF, GBM	90-day TAR: AUC (IBM CCAE) = 0.77, AUC (IBM MDCR) = 0.75, AUC (IBM MDCD) = 0.77, 730-day TAR: AUC = 0.71	(50)
Terapeutic response	MTX	All patients with new onset RA Training cohort: n = 26 Validation cohort: n = 21	metagenomic, clinical-pharmacogenetic variables	RF	AUC = 0.84	(51)
		Training dataset: ESPOIR: n = 493 EAC: n = 239 External validation dataset: Treach: n = 138	DAS28, creatininemia, leucocytes, lymphocytes, AST, ALT, swollen joints count and corticosteroids co-treatment.	LR, RF, LightGBM, CatBoost	Training dataset: AUC = 0.73 External validation set: AUC = 0.72	(52)
		349 RA patients: Training set: n = 279 Test set: n = 70	95 haplotypes and 5 non-genetic factors	NN, SVM, LR, EN, RF, Boosted Trees	AUC: 0.776 - 0.828 Sen: 0.656 - 0.813 Spe: 0.684 - 0.868	(53)

(Continued)

TABLE 1 Continued

Task	Sample Size		Features	ML algorithms	Performance	Ref
		82 RA patients: good responders: n = 42 poor responders/nonresponders: n = 43	gene expression	L2-regularized LR, RF, network-based approach	predictive utility between 4 weeks and pretreatmen: acc = 0.61, AUC = 0.78 predictive utility at the 4-week time point: acc = 0.68, AUC = 0.78.	(54)
	TNFi	Discovery cohort: n = 74(52 responders and 22 non responders) Validation cohort: n = 25(14 responders and 11 non responders)	clinical and molecular parameters	LR	AUC = 0.91	(55)
		Training dataset: n = 1892 Testing dataset: n = 680	demographic, clinical, and genetic markers	linear models, CART, SVM, GPR	Training dataset: AUC = 0.66 Testing dataset: AUC = 0.615	(56)
		Synovial tissue samples: RA: n = 256, OA: n = 41 NC: n = 36; Genes: n = 11,769	pathway and DEG	NB, DT, KNN, SVM	For infliximab response: Pathway-driven model AUC = 0.87, AUPR = 0.78; DEG-driven mode AUC = 0.92, AUPR = 0.86	(57)
		179 RA patients: Training set: n = 141 Validation set: n = 38	9 clinical parameters	NN	Response to infliximab AUC = 0.75	(58)
		responders: n = 23 non-responders: n = 16	clinical data, flow cytometry measurements, protein measurements and transcriptomics data	Linear, non-linear, kernel-based	response to TNFi AUC = 0.81	(59)
		Training set: n = 161 Validation set: n = 118	DAS28, lymphocytes, ALT, neutrophils, Age, weight and ever smoked	LR, RF, XGBoost, CatBoost	Response to Etanercept: Training set: AUC = 0.74 Validation set: AUC = 0.70 Response to monoclonal anti-TNF antibodies: Training set: AUC = 0.74 Validation set: AUC = 0.71	(60)
	Other drugs	R4RA synovial biopsies: n = 164	gene expression, clinical data and histological data	elastic net regression, GBM	For rituximab response AUC = 0.744 For tocilizumab response AUC = 0.681 For refractory state: AUC = 0.686	(61)
		1204 patients treated with bDMARDs	age, rheumatoid factor, ESR, disease duration, CRP	Lasso, Ridge, SVM , RF, XGBoost	Acc: 0.528 - 0.729 AUC: 0.511 - 0.694	(62)
		Training set: n = 625 Independent test set: n = 322	PtGA	RF, XGBoost, ANN, SVM	Acc = 0.726 AUC = 0.638 F1 score = 0.841	(63)
		Training set: 51 MR and 85 NR	DAS-28	CART	Training set: AUC = 0.89 Sen = 0.88	(64)

(Continued)

TABLE 1 Continued

Task	Sample Size	Features	ML algorithms	Performance	Ref
	External validation cohort: 35 MR and 47 NR			Spe = 0.94 Validation cohort: AUC = 0.82	
Comorbidities	487 patients diagnosed with RA and osteoporosis Training set: n = 340 Testing set: n = 147	baseline demographic, clinical test indicators	RF, ANN, SVM, XGBoost, DT	Training set: AUC = 0.878 Testing set: AUC = 0.872	(65)
	2374 RA patients	clinical features, medication, laboratory results	LR, RF, XGBoost, LightGBM	AUC = 0.75 Acc =0.68 F1 score = 0.7	(66)
	2 atherosclerosis and 2 RA datasets	NFIL3, EED, GRK2, MAP3K11, RMI1, TPST1	LASSO, RF	AUC: 0.723 to 1	(67)
	Training cohort: RA+CHD: n = 294 RA: n = 718 Validation cohort: RA+CHD: n = 70 RA: n = 204	age, hypertension, anti-CCP antibody positivity, rheumatoid factor positivity, a high ESR, high CRP levels, and dyslipidemia of LDL-c, TC, triglycerides and HDL-c	GBDT, KNN, LR, RF, XGBoost, SVM	AUC = 0.77 Sen = 0.639 Spe = 0.772	(68)
	RA-ILD: n = 75 RA-non-ILD: n = 78	age, KL-6, D-dimer, CA19-9	LASSO, RF, PLS	AUC = 0.928 Sen = 0.83 Spe = 0.81	(69)

Acc, accuracy; ADA, adaptive boosting; ALT, alanine aminotransferase; AST, aspartate aminotransferase; APS, antiphospholipid syndrome; AS, ankylosing spondylitis; AUPR, area under the precision-recall; BMI, body mass index; BS, behcet's syndrome; b/tsDMARDs, biologic or targeted synthetic disease modifying antirheumatic drugs; CART, classification and regression tree; CA19-9,carbohydrate antigen 19-9; CCP, cyclic citrullinated peptide; CHD, coronary heart disease; CRP, c-reactive protein; DAS 28, disease activity score-28; DEG, differentially expressed gene; DL, deep learning; DT, decision tree; EN, elastic nets; ESR, erythrocyte sedimentation rate; GBDT, gradient boosting decision tree; GBM, gradient-boosted machine; GPR, gaussian process regression; HC, healthy control; HDL, high density lipoprotein; IBD, inflammatory bowel disease; ILD, interstitial lung disease; JIA, juvenile idiopathic arthritis; KL-6, Krebs von den Lungen-6; KNN, k-nearest-neighbors; LASSO, least absolute shrinkage and selection operator; LDL, low density lipoprotein; LR, logistic regression; MG, myasthenia gravis; MR, multi-refractory; MS, multiple sclerosis; MTX, methotrexate; Non-ILD, rheumatoid arthritis-without interstitial lung disease; NB, naïve bayes; NN, neural networks; NR, non-refractory; OA, osteoarthritis; OP, osteoporosis; PLS, partial least square; PRS, polygenic risk score; PSC, primary sclerosing cholangitis; PtGA, patient global assessment of disease activity; R, responders; RA, rheumatoid arthritis; ReA, reactive arthritis; RF, random forest; SEN, sensitivity; SLE, systemic lupus erythematosus; SNH, safety-net hospital cohort; SNP, single nucleotide polymorphism; SPE, specificity; SVM, support vector machine; TAR, time at risk; TC, total cholesterol; T1D, type 1 diabetes; TNFi, tumor necrosis factor inhibitor; TKR, total knee replacement; UH, university hospital cohort; XGBoost, eXtreme Gradient Boosting.

regular medical examinations and monitoring RA-related biomarkers, such as inflammation levels and autoantibodies, early detection of the disease can utilize the 'window of opportunity' for therapeutic intervention. Early interventions can help prevent severe radiographic damage and disability, thus significantly improving patient prognosis (71). The exact etiology of RA remains not fully understood; however, it is known that genetic and environmental factors, as well as their interactions, influence the onset and progression of RA (72). ML, as an effective data analysis tool, is capable of processing and interpreting large volumes of diverse data, ranging from genetic factors to lifestyle choices. ML can uncover potential risk patterns within complex genetic and environmental datasets, assisting clinicians in making more accurate disease predictions and risk assessments.

Predictive modeling harnessing ML techniques to pinpoint individuals at an elevated risk for RA can be principally segregated into two domains: forecasting the incident risk in asymptomatic persons and assessing the progression likelihood in symptomatic patients with undifferentiated arthritis towards RA. The detection of RA susceptibility in the broad population leans on the analysis of genetic variants alongside common clinical risk indicators such as family history, age, and gender. A study found nine single nucleotide polymorphisms (SNPs) linked to RA, by combining these variations into a risk score and using ML algorithms, researchers were able to accurately distinguish RA patients from those without the condition, exhibiting five-fold cross-validated AUCs surpassing the 0.9 threshold (27). 11 risk factors for RA were identified from National Health and Nutrition Examination Survey (NHANES) data and used to create a Bayesian logistic regression model, which was refined using a Genetic Algorithm. The model showed high predictive accuracy with an AUC of 0.826 on the validation set (28). These findings highlight the potential of machine learning strategies in predicting risk populations for RA. Genetic risk scores derived from SNPs can help identify an individual's potential genetic risks, thereby providing a crucial foundation for personalized medicine (73). However, translating these studies into clinical decision support tools faces obstacles, primarily ensuring the equal applicability of Polygenic risk score (PRS) across populations (74). In reality, PRS exhibits limited transferability among populations, and its clinical

utility in RA remains undetermined, necessitating substantial investment in extensive data collection across diverse ethnic groups and methodological research to enhance genetic prediction in admixed individuals (75). Another critical issue is the interpretability of genetic findings in participants, requiring clinicians to possess the capacity to comprehend and interpret data (76). Furthermore, privacy and security of the involved genetic data must be adequately ensured. Federated learning, as a distributed machine learning technique, aims to achieve collaborative modeling while ensuring data privacy, security, and legal compliance (77). Participants can train their local models using their proprietary data, and through iterative training, each participant contributes to the construction of a global model without sharing their data externally (78). This approach fosters collaboration among multiple medical institutions, facilitating the sharing of model learning outcomes (79).

The likelihood of individuals with undifferentiated arthritis (UA), who exhibit joint symptoms without fulfilling the full diagnostic criteria, subsequently progressing to RA poses a clinical conundrum. Accurate prediction of this progression can facilitate early diagnosis and intervention for those at risk, while concurrently preventing overtreatment and diminishing both the health repercussions and superfluous healthcare expenditures for those unlikely to develop RA (80). Models are increasingly geared towards the evaluation of dynamic variables, reflecting shifts correlated with disease activity, such as gene expression profiles, epigenetic modifications, and a spectrum of detailed symptomatic and clinical markers.

A notable investigation sought to unearth clinically pertinent predictive biomarkers from peripheral blood CD4 T cells in UA patients, employing a support vector machine (SVM) classification model. This approach demonstrated that an integration of the pre-established Leiden predictive rule with a 12-gene risk indicator notably enhanced the prognostic capability from the original (AUC=0.74) to a significantly improved accuracy for seronegative UA patients (AUC=0.84) (29). A comparative analysis of three distinct ML algorithms revealed that a SVM model, which integrated DNA methylation profiles from 40 CpG sites with clinical parameters including disease activity score (DAS) and RF, effectively distinguished individuals with UA who were predisposed to developing RA within one year, achieving an AUC range of 0.85 to 1 (30).

Contemporary studies report promising predictive performance in identifying at-risk individuals within the general population and in forecasting RA development in patients with UA, and that the features having the greatest impact on predictive outcomes were identified and selected as much as possible during model training in order to simplify the model and potentially improve performance and generalizability. More important than performance, however, is the potential for practical clinical application, and future studies will need to examine the generalizability of the model by testing it in populations of multiple ethnicities and regions, and tracking the progression of individuals to RA in larger prospective cohorts to observe the accuracy of the model.

3.2 Diagnosis and subtype classification of RA

The diagnostic framework for RA, especially in the context of seronegative RA, is intricate and often obstructed by the absence of potent biomarkers, impeding early detection and management (47). Investigations are thus aimed at the identification of new biomarkers to bridge this gap.

Non-invasive imaging techniques are pivotal in elucidating inflammatory activity and its effects on joint morphology, especially when serological markers are indistinct or inconclusive. These tools are indispensable for both diagnostic purposes and for monitoring treatment efficacy (81). Furthermore, the application of ML algorithms in the analysis of imaging data presents a sophisticated approach to patient classification (82). Üreten K et al. presented a model of a Visual Geometry Group-16 (VGG-16) neural network for hand radiographs augmented by transfer learning to distinguish RA patients from non-RA patients, which achieved an AUC of 0.97 (31). Ultrasound imaging of the metacarpophalangeal joints in RA patients has been categorized for classification purposes, employing a DenseNet-based deep learning model in several regions of interest, significant efficacy was demonstrated in distinguishing between synovial proliferation and healthy and diseased synovium, as evidenced by AUCs exceeding 0.8 (32). Additionally, research has been conducted utilizing hand RGB images and gripforce as features to develop a random forest model with an AUC of 0.97 for distinguishing between individuals with RA and control subjects, thereby offering a supplementary diagnostic tool for RA (33). Image-based predictive models have shown notable performance in research settings, accurately differentiating RA patients from others in various cohorts, thereby contributing to the precision and efficiency of RA diagnosis. These models facilitate the early detection of abnormal changes within the joints, enabling timely intervention and ultimately delaying the progression of RA. However, their clinical application still faces significant challenges. A primary obstacle is the interpretability of the models. Owing to the 'black box' nature of deep learning models, the decision-making processes are opaque and difficult to comprehend, which may affect both physician and patient trust and understanding of model predictions (83). To address this limitation, some well-known methods can be utilized: The Class Activation Mapping (CAM) technique helps in understanding the regions of interest within images as attended by the model (84); Shapley Additive exPlanations (SHAP) elucidate the global impact of each feature on the model (85); and Local Interpretable Model-agnostic Explanations (LIME) explicate the local prediction process for individual samples (86). Collectively, these methods provide interpretability tools that enhance comprehension of the model's decision-making process and improve its interpretability. Future studies are also suggested to involve multi-center collaborations to enhance image collection with the intent to further refine and generalize these diagnostic models.

In RA, both individual analyses and integrative omics studies have accumulated a vast amount of data, providing insights into the mechanisms of RA from multiple perspectives. Genomics identifies genetic variations associated with RA, revealing potential genetic mechanisms influencing gene expression (87). Epigenetic modifications, including DNA methylation, histone modifications, chromatin remodeling, and non-coding RNA, play crucial roles in maintaining normal gene expression patterns. Epigenomics studies these modifications to reveal gene expression and regulatory mechanisms in RA, offering insights into the diverse molecular processes involved (88). Transcriptomics, by analyzing the variations in gene expression under different conditions, provides a detailed elucidation of which genes are upregulated or downregulated in RA. This process not only involves the regulation at the genetic level but also directly affects the production and function of the corresponding proteins (89). Proteomics provides a comprehensive analysis of protein composition, expression levels, and modification states, elucidating the interactions and connections among proteins that may play key roles in RA inflammation and immune response processes (90). Metabolomics provides insights into the shifts in metabolic states and pathways during the progression of RA. These changes are potentially influenced by alterations in gene and protein activities. Furthermore, metabolites themselves can play a modulatory role, affecting gene transcription and protein expression, thereby forming a complex interplay that influences disease dynamics (91). Host genomic variations significantly influence the composition of the gut microbiota, which can synthesize, regulate, or degrade endogenous small molecules or macromolecules, resulting in metabolic changes. Utilizing metagenomics and related techniques reveals the role of gut microbiota in the development of RA by influencing metabolic pathways and modulating the host immune system (92). Omic studies are characterized by the generation of vast, high-dimensional datasets. ML algorithms are critically employed for visualization and processing such information—finding patterns, crafting predictive models, and examining large-scale, multi-omic data to identify biomarkers and pathways implicated in disease progression (93, 94). Existing research has integrated multimodal data and employed various machine learning algorithms to develop high-performance diagnostic models for RA. Key genes highly correlated with RA phenotypes have been identified through the application of weighted gene co-expression network analysis (WGCNA) and differential gene expression (DEG) analysis on RA blood sample microarray datasets. These genes have been deployed as features to assess the performance of six ML models, with five demonstrating commendable efficacy (AUC > 0.85) (34). Through the sourcing of RA patient peripheral blood sample microarray datasets from the GEO database, a platelet-related signature risk score model was formulated, comprised of six genes, using the Least Absolute Shrinkage and Selection Operator (LASSO) algorithm. The model exhibited AUCs of 0.801 and 0.979 across the training and validation sets, respectively (35). Employing the Generalized Matrix Learning Vector Quantization (GMLVQ) method, mRNA expression profiles of cytokines and chemokines from synovial biopsies were analyzed, leading to the identification of two gene sets. These sets were instrumental in generating a model capable of differentiating between various arthritis types, with AUC scores reaching 0.996 and 0.764 for distinguishing diagnosed RA from non-inflammatory cases and early-stage RA from self-remitting arthritis, respectively (36). By focusing on the expression of 19 N6-methyladenosine (m6A) methylation regulators, diagnostic models have been established to separate RA from non-RA conditions. A subset of these regulators, particularly IGF2BP3 and YTHDC2, demonstrated accuracies and AUCs exceeding 0.8 across most ML models, indicating the potential diagnostic importance of m6A methylation profiles (37). A multi-variable classification model, incorporating 26 metabolites and lipids, was devised utilizing three ML algorithms. The logistic regression model, in particular, stood out for its ability to differentiate seropositive and seronegative RA from normal controls within an independent validation cohort, securing an AUC of 0.91, thus showcasing that a holistic metabolomic and lipidomic approach grounded in Liquid Chromatography-Mass Spectrometry (LC-MS) can effectively segregate RA cases (38). Serum antigens were analyzed in patient cohorts with RA, osteoarthritis (OA), and healthy controls. Subsequently, distinct biomarker sets were identified for the differentiation of RA, ACPA-positive RA, and ACPA-negative RA using feature selection through the Random Forest algorithm. The model demonstrated exceptional performance with AUC values of 0.9949, 0.9913, and 1.0, respectively, establishing a proteomics-based diagnostic model for RA (39). Furthermore, leveraging metagenomic data to predict the microbiomic characteristics of the gut in autoimmune diseases has been demonstrated to discriminate between various types of autoimmune disorders (40).

Histopathology, as a fundamental pillar in confirming disease diagnosis, stands as the definitive standard for the verification of numerous ailments (95). Overlap of symptoms in certain pathologies may obscure the principal etiology responsible for articular manifestations; in such instances, tissue biopsy, particularly of synovial tissue, proves invaluable. Following Total Knee Arthroplasty (TKA), synovial samples from 147 OA and 60 RA individuals were subjected to hematoxylin and eosin (H&E) staining. Utilization of a Random Forest Algorithm, integrating pathologist-derived scores with computer vision-generated cellular density measures, led to the construction of an optimal discriminative model for OA and RA, achieving a model AUC of 0.91 (42). This serves as a potent discriminative tool for RA assessment. Orange et al. utilized consensus clustering of gene expression data from synovial tissues of patients with RA to identify three distinct synovial subtypes: high-inflammatory, low-inflammatory, and mixed. They subsequently employed a support vector ML algorithm to distinguish between these subtypes based on histological features, achieving area under the curve values of 0.88, 0.71, and 0.59, respectively (43).

Despite the high performance of ML-derived predictive models for RA diagnosis, concerns on potential model overfitting due to limited sample sizes, which may exaggerate effect sizes, cannot be overlooked. Additionally, independent evaluation of the research methodology, data processing, and outcomes by an external party ensures the accuracy and reliability of the research findings.

Validation of these models in diverse datasets, supplemented by molecular biology experimentation, is imperative for evaluating true diagnostic merit. Predictive models relying on histopathological data encounter additional challenges, including the necessity for manual feature annotation by pathologists and the invasiveness of the procedure, compounded by technical and sample handling issues. External validation is a critical quality control measure, ensuring that model utility and accuracy in diagnosing RA reflect true clinical relevance and potential for widespread application. The diagnosis of RA extends beyond segregating RA from healthy subjects or OA patients. Future investigations must address the diagnostic capacity of predictive model-derived markers in distinguishing seronegative RA from other inflammatory arthritides, such as psoriatic arthritis, reactive arthritis, or spondyloarthritis. Concomitantly, safeguarding against confounding variables and maintaining diversity within patient cohorts are essential to render the model universally applicable.

3.3 Prediction of disease activity and imaging progression in RA

Radiographic deterioration in RA is characterized by the degree of articular damage and the presence of distinct lesions such as joint space narrowing, bone erosion, and osteoporosis, as revealed through diagnostic imaging modalities including X-rays, magnetic resonance imaging, or computed tomography scans (96). The quantification and prognostication of structural joint impairment traditionally hinge on clinical expertise, underscoring the necessity for an automated, bias-free evaluation method. A study utilizing SVM modeling on cohorts comprising 374 Korean and 399 North American patients with incipient RA identified SNPs correlated with radiographic progression. An integrated model encompassing SNPs with clinical parameters exhibited optimal performance, yielding a mean ten-fold cross-validation AUC of 0.78, providing a more satisfactory distinction between severe and non-severe progression (44).

Radiological damage bears a significant association with disease activity in RA, with heightened activity posing an increased risk for osseous impairment. CNNs trained on ultrasound imagery of RA joints, have facilitated the automatic grading of disease activity, achieving an overall classification accuracy of 83.9% (45). Vodencarevic et al. used data from 135 consultations with 41 RA patients to predict flare incidents during biologic disease-modifying antirheumatic drugs (DMARDs) tapering in remission. They combined multiple ML models to achieve an AUC of 0.81 (46). Furthermore, baseline serum proteomics from 130 stable RA patients in clinical remission was analyzed for biomarkers predictive of future disease flares, employing LASSO and eXtreme Gradient Boosting (XGBoost) algorithms to construct predictive models. The XGBoost model exhibited superior performance in differentiating between relapsed and non-relapsed patients with an AUC of 0.80 (47).

The expansive volume of patient intelligence and clinical information harbored in electronic medical records (EMR) and electronic health records (EHR) constitutes a substantial body of data ripe for investigation (97, 98). Nonetheless, hindrances such as imbalances in data record quantities across patients, omissions of pivotal information, and the variability in patient conditions and therapeutic outcomes over time contribute to the complex temporal nature of the data (48). Conventional ML techniques encounter constraints concerning data pre-processing, time-series analysis capacity, and the simplification of intricate relational processing (99). Deep learning integrated with structured EHR data, have been deployed to prognosticate disease activity during subsequent outpatient rheumatology consultations, wherein the model trained on the UH cohort manifested an AUC of 0.91 for internal validation and 0.74 for external cohort testing (48). Feldman et al. endeavored to enhance the precision of RA disease activity evaluation by integrating electronic medical records and claims data, achieving an AUC of 0.76 in discriminating high/moderate from low disease activity/remission (49). Chandran et al. employed the use of biologic agents or tofacitinib as a surrogate for distinguishing disease severity indicators, with the model accurately predicting both current and future disease activity validated across various databases with AUCs exceeding 0.7 (50).

The aforementioned results substantiate the viability of employing routinely documented clinical and laboratory data to assess and forecast disease activity in RA. With the progressive advancements in information technology, an extensive array of data has become accessible, prompting researchers to explore ML methodologies for the extraction of RA patient records from electronic health record data, thereby enabling the study of substantial populations at minimal expense. Algorithms trained via ML are progressively leveraged with EMR for clinical investigations. These algorithms function by detecting specifiable patterns in the data associated with RA, yet systematic disparities in EMR data quality present hurdles for model generalizability. Despite these challenges, high-caliber investigations are somewhat limited and the dependability and transferability of pertinent ML methods remain largely undetermined, rendering periodic evaluation of algorithm performance imperative. The current research trend involves the utilization of thousands of digitally annotated images obtained from large-scale observational studies, clinical trials, and electronic medical records, along with clinical data, to automatically classify and quantify the extent of joint damage and activity scores in RA using ML algorithms (100–102).

3.4 Prediction of RA treatment response

In the realm of RA therapeutics, a plethora of options including nonsteroidal anti-inflammatory drugs (NSAIDs), glucocorticoids, conventional synthetic DMARDs, biologic DMARDs, and oral small molecules have been made available (103). The selection of appropriate treatments continues to challenge clinicians owing to the vast range of alternatives and the prevalent trial-and-error approach in therapeutic prescription, exacerbated by a lack of comprehensive knowledge regarding drug efficacy and safety across distinct patient demographics (53).

Methotrexate (MTX) stands as the quintessential first-line therapy in RA treatment strategies (104). Investigation into

whether disparities in the gut microbiome across individuals could serve as predictive markers for MTX efficacy in newly onset RA was conducted by Artacho et al. Fecal samples from 26 new-onset RA patients, procured prior to MTX treatment, were analyzed using 16S ribosomal RNA (16S rRNA) and shotgun sequencing. Subsequent construction of a predictive model via random forests revealed that a response to MTX treatment at 4 months could be anticipated, with an AUC of 0.84, based on colony characterization (51). Additional research involving ML algorithms applied to clinical and biological data from 493 and 239 patients across two cohorts, aimed to predict MTX treatment response at 9 months. Notably, the Light Gradient Boosting Machine (LightGBM) model acquired AUCs of 0.73 and 0.72 in training and external validation sets, respectively (52). Lim et al. analyzed exome sequencing data from 349 RA patients and predicted treatment response to MTX using six ML algorithms. They identified 95 genetic factors and 5 non-genetic factors that influenced response. The predictions had strong performance with AUCs between 0.776 and 0.828 in the test set (53). Plant et al. utilized whole blood samples from RA patients initiating MTX treatment, both before and 4 weeks after commencement, conducting gene expression profiling to foretell treatment response at 6 months. Application of an L2 regularized logistic regression yielded an AUC of 0.78 (54). The development of these predictive models has contributed significantly towards identifying patients who are more likely to respond favorably to, or may not derive benefit from, MTX treatment.

Anti-tumor necrosis factor (anti-TNF) agents have been established as pivotal second-line therapeutic agents following methotrexate. A prospective multicenter study recruited 104 RA patients and 29 healthy donors to discover predictive biomarkers for anti-TNF treatment using ML. A hybrid model combining clinical and molecular variables achieved a high AUC value of 0.91 (55). The DREAM RA Responder Challenge introduced a novel approach to predicting anti-TNF treatment response by proposing an optimal model that incorporates Gaussian Process Regression (GPR) and integrates demographic, clinical, and genetic markers. This model accurately predicts the Disease Activity Score in patients 24 months post-baseline assessment and categorizes treatment response according to the EULAR response criteria, effectively identifying non-responders to anti-TNF therapy with an AUC of 0.6 in cross-validation data (56). Kim et al. utilized 11 datasets containing 256 synovial tissue samples, integrating RA-associated pathway activation scores and four ML types, and found that the SVM model performed the best, with an AUC of 0.87 using the pathway-driven model and an AUC of 0.9 using the DEG-driven model (57).

Recent research has emphasized the potential benefits of integrating diverse datasets for the purpose of treatment decision-making. ML algorithms have demonstrated efficacy in enhancing the precision of response prediction for TNF inhibitors and MTX. Furthermore, ML methodologies are being increasingly utilized in forecasting treatment responses to a range of other biologic therapies (61–64). Clinical data may be limited by trial design, including inclusion and exclusion criteria.Using deep learning technology for cluster analysis on RA patients has revealed the connection between patient characteristics and treatment response (105). Advancements in spatial omics technologies enable a comprehensive and spatially intact analysis of synovial tissue in RA patients. This approach allows for precise localization of cells, exploration of cellular interactions, assessment of cell type distributions, and identification of disease-associated molecular markers (106). Integrating traditional multi-omics with spatial data, spatial multi-omics elucidates the complexity and dynamics of biological processes across various levels, including their interactions and influences on each other. This approach deepens our understanding of the pathological mechanisms of RA and enhances our knowledge of its spatial heterogeneity (107). The biopsy-driven RA randomized clinical trial (R4RA), which utilizes spatial omics to create synovial biopsy gene maps, provides a paradigm for predicting drug treatment responses and refining therapeutic strategies. This is crucial for achieving personalized medicine and optimizing treatment outcomes. Despite some progress, spatial omics in RA research is still in its early stages. Numerous challenges remain, such as high costs, high demands on sample handling, patient acceptance, ethical issues, and the need for advanced computational tools for data integration (108). Overcoming these challenges will be crucial for developing accurate, interpretable, and clinically applicable predictive models. In summary while opportunities exist for refining the accuracy of these predictions, progress is evident in this area of study. In the future, using a larger, more comprehensive datase, appropriate algorithms, and methods in parameter optimization, improving model features and validating against independent cohorts may further improve the discriminative power of predictive models.

3.5 Prediction of comorbidities related to RA

ML is also gaining attention in the prediction of comorbidities associated with RA. Focus within extant research has primarily been oriented towards the identification of risk factors for osteoporosis (65, 66), assessment of cardiovascular risk (67, 68), and the prediction of interstitial lung disease development (69) in individuals with RA. Current models pertaining to comorbidities are limited in both quantity and accuracy, with constraints stemming from various sources, notably the scarcity of comprehensive comorbidity data within RA patient cohort datasets. Furthermore, there is significant variability in data quality across different cohorts. To overcome these obstacles, future research should prioritize the accumulation of larger, more robust datasets and improve integration among diverse data sources.Simultaneously, there is a necessity for the advancement of algorithms with broader applicability, thereby enabling the utilization of ML in the prediction of complications associated with RA.

4 Conclusion and outlook

Integrating data from diverse sources allows ML models to yield more comprehensive and precise predictions for the diagnosis and treatment outcomes of RA. However, more focus and effort are needed to create predictive models for comorbidities related to RA. Recent research has demonstrated the potential of multimodal learning to

improve clinical prediction accuracy. The optimal performing model under specific conditions often necessitates an extensive comparative analysis. Beyond frequently used metrics such as AUC, accuracy, sensitivity, specificity, and F1 score, the employment of cross-validation, the statistical tests applied, the model's computational cost, the data requirements, and accessibility, the adoption of multimodal learning approaches aims to refine clinical predictions. Efforts should be made to improve the clinical operability of models, utilize external datasets from diverse origins for validation, assess the model's generalizability, monitor its long-term performance, and evaluate its strengths and weaknesses through multidimensional approaches rather than relying on a single performance metric. Although ML models have demonstrated impressive predictive prowess in research settings, it is imperative to establish their practicality and effectiveness in real-world clinical scenarios. To cultivate trust and acceptance among medical practitioners, it is essential to enhance the interpretability of these models. This can be achieved by prioritizing simplicity in experimental design or by employing tools that enhance model interpretability. Finally, but importantly, the privacy and ethical implications of big biological data should be emphasized and protected.

Author contributions

YMS: Data curation, Visualization, Writing – original draft. MZ: Data curation, Formal analysis, Writing – review & editing. CC: Data curation, Formal analysis, Writing – review & editing. PJ: Data curation, Formal analysis, Writing – review & editing. KW: Data curation, Formal analysis, Writing – review & editing. JZ: Data curation, Formal analysis, Writing – review & editing. YS: Data curation, Formal analysis, Writing – review & editing. YZ: Data curation, Formal analysis, Writing – review & editing. FZ: Data curation, Formal analysis, Writing – review & editing. XL: Data curation, Formal analysis, Writing – review & editing. SG: Conceptualization, Writing – review & editing. FW: Supervision, Writing – review & editing. DH: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was funded by the National Natural Science Funds of China (82074234, 82004166 and 82071756), Shanghai Chinese Medicine Development Office, National Administration of Traditional Chinese Medicine, Regional Chinese Medicine (Specialist) Diagnosis and Treatment Center Construction Project-Rheumatology, State Administration of Traditional Chinese Medicine, Shanghai Municipal Health Commission, East China Region-based Chinese and Western Medicine Joint Disease Specialist Alliance, and Shanghai He Dongyi Famous Chinese Medicine Studio Construction Project (SHGZS-202220).

Acknowledgments

Figure 1 was created by Figdraw (www.figdraw.com).

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

References

1. Cross M, Smith E, Hoy D, Carmona L, Wolfe F, Vos T, et al. The global burden of rheumatoid arthritis: estimates from the global burden of disease 2010 study. *Ann Rheum Dis*. (2014) 73:1316–22. doi: 10.1136/annrheumdis-2013-204627

2. Johnson TM, Sayles HR, Baker JF, George MD, Roul P, Zheng C, et al. Investigating changes in disease activity as a mediator of cardiovascular risk reduction with methotrexate use in rheumatoid arthritis. *Ann Rheum Dis*. (2021) 80:1385–92. doi: 10.1136/annrheumdis-2021-220125

3. Redente EF, Aguilar MA, Black BP, Edelman BL, Bahadur AN, Humphries SM, et al. Nintedanib reduces pulmonary fibrosis in a model of rheumatoid arthritis-associated interstitial lung disease. *Am J Physiol Lung Cell Mol Physiol*. (2018) 314: L998–L1009. doi: 10.1152/ajplung.00304.2017

4. Ng KJ, Huang KY, Tung CH, Hsu BB, Wu CH, Koo M, et al. Modified rheumatoid arthritis impact of disease (RAID) score, a potential tool for depression and anxiety screening for rheumatoid arthritis. *Joint Bone Spine*. (2019) 86:805–7. doi: 10.1016/j.jbspin.2019.04.007

5. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet*. (2015) 16:321–32. doi: 10.1038/nrg3920

6. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. *Science*. (2015) 349:255–60. doi: 10.1126/science.aaa8415

7. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst*. (2014) 2:3. doi: 10.1186/2047-2501-2-3

8. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. *Nature*. (2018) 559:547–55. doi: 10.1038/s41586-018-0337-2

9. Coffey CM, Crowson CS, Myasoedova E, Matteson EL, Davis JM 3rd. Evidence of diagnostic and treatment delay in seronegative rheumatoid arthritis: missing the window of opportunity. *Mayo Clin Proc*. (2019) 94:2241–8. doi: 10.1016/j.mayocp.2019.05.023

10. Conigliaro P, Triggianese P, De Martino E, Fonti GL, Chimenti MS, Sunzini F, et al. Challenges in the treatment of rheumatoid arthritis. *Autoimmun Rev*. (2019) 18:706–13. doi: 10.1016/j.autrev.2019.05.007

11. Zhao J, Guo S, Schrodi SJ, He D. Molecular and cellular heterogeneity in rheumatoid arthritis: mechanisms and clinical implications. *Front Immunol*. (2021) 12:790122. doi: 10.3389/fimmu.2021.790122

12. Lo-Ciganic WH, Huang JL, Zhang HH, Weiss JC, Wu Y, Kwoh CK, et al. Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions. *JAMA Netw Open*. (2019) 2:e190968. doi: 10.1001/jamanetworkopen.2019.0968

13. Warnat-Herresthal S, Schultze H, Shastry KL, Manamohan S, Mukherjee S, Garg V, et al. Swarm Learning for decentralized and confidential clinical machine learning. *Nature*. (2021) 594:265–70. doi: 10.1038/s41586-021-03583-3

14. Goodswen SJ, Barratt JLN, Kennedy PJ, Kaufer A, Calarco L, Ellis JT. Machine learning and applications in microbiology. *FEMS Microbiol Rev*. (2021) 45:fuab015. doi: 10.1093/femsre/fuab015

15. Jiang T, Gradus JL, Rosellini AJ. Supervised machine learning: A brief primer. *Behav Ther*. (2020) 51:675–87. doi: 10.1016/j.beth.2020.05.002

16. Gitto S, Cuocolo R, Annovazzi A, Anelli V, Acquasanta M, Cincotta A, et al. CT radiomics-based machine learning classification of atypical cartilaginous tumours and appendicular chondrosarcomas. *EBioMedicine*. (2021) 68:103407. doi: 10.1016/j.ebiom.2021.103407

17. Kulin M, Fortuna C, De Poorter E, Deschrijver D, Moerman I. Data-driven design of intelligent wireless networks: an overview and tutorial. *Sensors (Basel)*. (2016) 16:790. doi: 10.3390/s16060790

18. Williamson DJ, Burn GL, Simoncelli S, Griffié J, Peters R, Davis DM, et al. Machine learning for cluster analysis of localization microscopy data. *Nat Commun*. (2020) 11:1493. doi: 10.1038/s41467-020-15293-x

19. Gao T, Lu W. Machine learning toward advanced energy storage devices and systems. *iScience*. (2020) 24:101936. doi: 10.1016/j.isci.2020.101936

20. Bajić F, Orel O, Habijan M. A multi-purpose shallow convolutional neural network for chart images. *Sensors (Basel)*. (2022) 22:7695. doi: 10.3390/s22207695

21. Schwendicke F, Samek W, Krois J. Artificial intelligence in dentistry: chances and challenges. *J Dent Res*. (2020) 99:769–74. doi: 10.1177/0022034520915714

22. Peng H, Fan Y. Feature selection by optimizing a lower bound of conditional mutual information. *Inf Sci (N Y)*. (2017) 418-419:652–67. doi: 10.1016/j.ins.2017.08.036

23. Yang L, Jiang H, Ding X, Liao Z, Wei M, Li J, et al. Modulation of sleep architecture by whole-body static magnetic exposure: A study based on EEG-based automatic sleep staging. *Int J Environ Res Public Health*. (2022) 19:741. doi: 10.3390/ijerph19020741

24. Tasci E, Jagasia S, Zhuge Y, Sproull M, Cooley Zgela T, Mackey M, et al. RadWise: A rank-based hybrid feature weighting and selection method for proteomic categorization of chemoirradiation in patients with glioblastoma. *Cancers (Basel)*. (2023) 15:2672. doi: 10.3390/cancers15102672

25. Liang Y, Zhang ZQ, Liu NN, Wu YN, Gu CL, Wang YL. MAGCNSE: predicting lncRNA-disease associations using multi-view attention graph convolutional network and stacking ensemble model. *BMC Bioinf*. (2022) 23:189. doi: 10.1186/s12859-022-04715-w

26. Chen Y, Luo M, Cheng Y, Huang Y, He Q. A nomogram to predict prolonged stay of obesity patients with sepsis in ICU: Relevancy for predictive, personalized, preventive, and participatory healthcare strategies. *Front Public Health*. (2022) 10:944790. doi: 10.3389/fpubh.2022.944790

27. Lim AJW, Tyniana CT, Lim LJ, Tan JWL, Koh ET, TTSH Rheumatoid Arthritis Study Group, et al. Robust SNP-based prediction of rheumatoid arthritis through machine-learning-optimized polygenic risk score. *J Transl Med*. (2023) 21:92. doi: 10.1186/s12967-023-03939-5

28. Lufkin L, Budišić M, Mondal S, Sur S. A bayesian model to analyze the association of rheumatoid arthritis with risk factors and their interactions. *Front Public Health*. (2021) 9:693830. doi: 10.3389/fpubh.2021.693830

29. Pratt AG, Swan DC, Richardson S, Wilson G, Hilkens CM, Young DA, et al. A CD4 T cell gene signature for early rheumatoid arthritis implicates interleukin 6-mediated STAT3 signalling, particularly in anti-citrullinated peptide antibody-negative disease. *Ann Rheum Dis*. (2012) 71:1374–81. doi: 10.1136/annrheumdis-2011-200968

30. de la Calle-Fabregat C, Niemantsverdriet E, Cañete JD, Li T, van der Helm-van Mil AHM, Rodríguez-Ubreva J, et al. Prediction of the progression of undifferentiated arthritis to rheumatoid arthritis using DNA methylation profiling. *Arthritis Rheumatol*. (2021) 73:2229–39. doi: 10.1002/art.41885

31. Üreten K, Maraş HH. Automated classification of rheumatoid arthritis, osteoarthritis, and normal hand radiographs with deep learning methods. *J Digit Imaging*. (2022) 35:193–9. doi: 10.1007/s10278-021-00564-w

32. Wu M, Wu H, Wu L, Cui C, Shi S, Xu J, et al. A deep learning classification of metacarpophalangeal joints synovial proliferation in rheumatoid arthritis by ultrasound images. *J Clin Ultrasound*. (2022) 50:296–301. doi: 10.1002/jcu.23143

33. Alarcón-Paredes A, Guzmán-Guzmán IP, Hernández-Rosales DE, Navarro-Zarza JE, Cantillo-Negrete J, Cuevas-Valencia RE, et al. Computer-aided diagnosis based on hand thermal, RGB images, and grip force using artificial intelligence as screening tool for rheumatoid arthritis in women. *Med Biol Eng Comput*. (2021) 59:287–300. doi: 10.1007/s11517-020-02294-7

34. Xiao J, Wang R, Cai X, Ye Z. Coupling of co-expression network analysis and machine learning validation unearthed potential key genes involved in rheumatoid arthritis. *Front Genet*. (2021) 12:604714. doi: 10.3389/fgene.2021.604714

35. Liu Y, Jiang H, Kang T, Shi X, Liu X, Li C, et al. Platelets-related signature based diagnostic model in rheumatoid arthritis using WGCNA and machine learning. *Front Immunol*. (2023) 14:1204652. doi: 10.3389/fimmu.2023.1204652

36. Yeo L, Adlard N, Biehl M, Juarez M, Smallie T, Snow M, et al. Expression of chemokines CXCL4 and CXCL7 by synovial macrophages defines an early stage of rheumatoid arthritis. *Ann Rheum Dis*. (2016) 75:763–71. doi: 10.1136/annrheumdis-2014-206921

37. Geng Q, Cao X, Fan D, Gu X, Zhang Q, Zhang M, et al. Diagnostic gene signatures and aberrant pathway activation based on m6A methylation regulators in rheumatoid arthritis. *Front Immunol*. (2022) 13:1041284. doi: 10.3389/fimmu.2022.1041284

38. Luan H, Gu W, Li H, Wang Z, Lu L, Ke M, et al. Serum metabolomic and lipidomic profiling identifies diagnostic biomarkers for seropositive and seronegative rheumatoid arthritis patients. *J Transl Med*. (2021) 19:500. doi: 10.1186/s12967-021-03169-7

39. Han P, Hou C, Zheng X, Cao L, Shi X, Zhang X, et al. Serum antigenome profiling reveals diagnostic models for rheumatoid arthritis. *Front Immunol*. (2022) 13:884462. doi: 10.3389/fimmu.2022.884462

40. Volkova A, Ruggles KV. Predictive metagenomic analysis of autoimmune disease identifies robust autoimmunity and disease specific microbial signatures. *Front Microbiol*. (2021) 12:621310. doi: 10.3389/fmicb.2021.621310

41. Ormseth MJ, Solus JF, Sheng Q, Ye F, Wu Q, Guo Y, et al. Development and validation of a microRNA panel to differentiate between patients with rheumatoid arthritis or systemic lupus erythematosus and controls. *J Rheumatol*. (2020) 47:188–96. doi: 10.3899/jrheum.181029

42. Mehta B, Goodman S, DiCarlo E, Jannat-Khah D, Gibbons JAB, Otero M, et al. Machine learning identification of thresholds to discriminate osteoarthritis and rheumatoid arthritis synovial inflammation. *Arthritis Res Ther*. (2023) 25:31. doi: 10.1186/s13075-023-03008-8

43. Orange DE, Agius P, DiCarlo EF, Robine N, Geiger H, Szymonifka J, et al. Identification of three rheumatoid arthritis disease subtypes by machine learning integration of synovial histologic features and RNA sequencing data. *Arthritis Rheumatol*. (2018) 70:690–701. doi: 10.1002/art.40428

44. Joo YB, Kim Y, Park Y, Kim K, Ryu JA, Lee S, et al. Biological function integrated prediction of severe radiographic progression in rheumatoid arthritis: a nested case control study. *Arthritis Res Ther*. (2017) 19:244. doi: 10.1186/s13075-017-1414-x

45. Christensen ABH, Just SA, Andersen JKH, Savarimuthu TR. Applying cascaded convolutional neural network design further enhances automatic scoring of arthritis disease activity on ultrasound images from rheumatoid arthritis patients. *Ann Rheum Dis*. (2020) 79:1189–93. doi: 10.1136/annrheumdis-2019-216636

46. Vodencarevic A, Tascilar K, Hartmann F, Reiser M, Hueber AJ, Haschka J, et al. Advanced machine learning for predicting individual risk of flares in rheumatoid arthritis patients tapering biologic drugs. *Arthritis Res Ther*. (2021) 23:67. doi: 10.1186/s13075-021-02439-5

47. O'Neil LJ, Hu P, Liu Q, Islam MM, Spicer V, Rech J, et al. Proteomic approaches to defining remission and the risk of relapse in rheumatoid arthritis. *Front Immunol*. (2021) 12:729681. doi: 10.3389/fimmu.2021.729681

48. Norgeot B, Glicksberg BS, Trupin L, Lituiev D, Gianfrancesco M, Oskotsky B, et al. Assessment of a deep learning model based on electronic health record data to forecast clinical outcomes in patients with rheumatoid arthritis. *JAMA Netw Open*. (2019) 2:e190606. doi: 10.1001/jamanetworkopen.2019.0606

49. Feldman CH, Yoshida K, Xu C, Frits ML, Shadick NA, Weinblatt ME, et al. Supplementing claims data with electronic medical records to improve estimation and classification of rheumatoid arthritis disease activity: A machine learning approach. *ACR Open Rheumatol*. (2019) 1:552–9. doi: 10.1002/acr2.11068

50. Chandran U, Reps J, Stang PE, Ryan PB. Inferring disease severity in rheumatoid arthritis using predictive modeling in administrative claims databases. *PloS One*. (2019) 14:e0226255. doi: 10.1371/journal.pone.0226255

51. Artacho A, Isaac S, Nayak R, Flor-Duro A, Alexander M, Koo I, et al. The pretreatment gut microbiome is associated with lack of response to methotrexate in new-onset rheumatoid arthritis. *Arthritis Rheumatol*. (2021) 73:931–42. doi: 10.1002/art.41622

52. Duquesne J, Bouget V, Cournède PH, Fautrel B, Guillemin F, de Jong PHP, et al. Machine learning identifies a profile of inadequate responder to methotrexate in rheumatoid arthritis. *Rheumatol (Oxford)*. (2023) 62:2402–9. doi: 10.1093/rheumatology/keac645

53. Lim AJW, Lim LJ, Ooi BNS, Koh ET, Tan JWL, TTSH RA Study Group, et al. Functional coding haplotypes and machine-learning feature elimination identifies predictors of Methotrexate Response in Rheumatoid Arthritis patients. *EBioMedicine*. (2022) 75:103800. doi: 10.1016/j.ebiom.2021.103800

54. Plant D, Maciejewski M, Smith S, Nair N, Maximising Therapeutic Utility in Rheumatoid Arthritis Consortium, the RAMS Study Group, Hyrich K, et al. Profiling of gene expression biomarkers as a classifier of methotrexate nonresponse in patients with rheumatoid arthritis. *Arthritis Rheumatol*. (2019) 71:678–84. doi: 10.1002/art.40810

55. Luque-Tévar M, Perez-Sanchez C, Patiño-Trives AM, Barbarroja N, Arias de la Rosa I, Abalos-Aguilera MC, et al. Integrative clinical, molecular, and computational analysis identify novel biomarkers and differential profiles of anti-TNF response in rheumatoid arthritis. *Front Immunol*. (2021) 12:631662. doi: 10.3389/fimmu.2021.631662

56. Guan Y, Zhang H, Quang D, Wang Z, Parker SCJ, Pappas DA, et al. Machine learning to predict anti-tumor necrosis factor drug responses of rheumatoid arthritis

patients by integrating clinical and genetic markers. *Arthritis Rheumatol.* (2019) 71:1987–96. doi: 10.1002/art.41056

57. Kim KJ, Kim M, Adamopoulos IE, Tagkopoulos I. Compendium of synovial signatures identifies pathologic characteristics for predicting treatment response in rheumatoid arthritis patients. *Clin Immunol.* (2019) 202:1–10. doi: 10.1016/j.clim.2019.03.002

58. Miyoshi F, Honne K, Minota S, Okada M, Ogawa N, Mimura T. A novel method predicting clinical response using only background clinical data in RA patients before treatment with infliximab. *Mod Rheumatol.* (2016) 26:813–6. doi: 10.3109/14397595.2016.1168536

59. Yoosuf N, Maciejewski M, Ziemek D, Jelinsky SA, Folkersen L, Müller M, et al. Early prediction of clinical response to anti-TNF treatment using multi-omics and machine learning in rheumatoid arthritis. *Rheumatol (Oxford).* (2022) 61:1680–9. doi: 10.1093/rheumatology/keab521

60. Bouget V, Duquesne J, Hassler S, Cournède PH, Fautrel B, Guillemin F, et al. Machine learning predicts response to TNF inhibitors in rheumatoid arthritis: results on the ESPOIR and ABIRISK cohorts. *RMD Open.* (2022) 8:e002442. doi: 10.1136/rmdopen-2022-002442

61. Rivellese F, Surace AEA, Goldmann K, Sciacca E, Çubuk C, Giorli G, et al. Rituximab versus tocilizumab in rheumatoid arthritis: synovial biopsy-based biomarker analysis of the phase 4 R4RA randomized trial. *Nat Med.* (2022) 28:1256–68. doi: 10.1038/s41591-022-01789-0

62. Koo BS, Eun S, Shin K, Yoon H, Hong C, Kim DH, et al. Machine learning model for identifying important clinical features for predicting remission in patients with rheumatoid arthritis treated with biologics. *Arthritis Res Ther.* (2021) 23:178. doi: 10.1186/s13075-021-02567-y

63. Lee S, Kang S, Eun Y, Won HH, Kim H, Lee J, et al. Machine learning-based prediction model for responses of bDMARDs in patients with rheumatoid arthritis and ankylosing spondylitis. *Arthritis Res Ther.* (2021) 23:254. doi: 10.1186/s13075-021-02635-3

64. Novella-Navarro M, Benavent D, Ruiz-Esquide V, Tornero C, Díaz-Almirón M, Chacur CA, et al. Predictive model to identify multiple failure to biological therapy in patients with rheumatoid arthritis. *Ther Adv Musculoskelet Dis.* (2022) 14:1759720X221124028. doi: 10.1177/1759720X221124028

65. Chen R, Huang Q, Chen L. evelopment and validation of machine learning models for prediction of fracture risk in patients with elderly-onset rheumatoid arthritis. *Int J Gen Med.* (2022) 15:7817–29. doi: 10.2147/IJGM.S380197

66. Lee C, Joo G, Shin S, Im H, Moon KW. Prediction of osteoporosis in patients with rheumatoid arthritis using machine learning. *Sci Rep.* (2023) 13:21800. doi: 10.1038/s41598-023-48842-7

67. Liu F, Huang Y, Liu F, Wang H. Identification of immune-related genes in diagnosing atherosclerosis with rheumatoid arthritis through bioinformatics analysis and machine learning. *Front Immunol.* (2023) 14:1126647. doi: 10.3389/fimmu.2023.1126647

68. Wei T, Yang B, Liu H, Xin F, Fu L. Development and validation of a nomogram to predict coronary heart disease in patients with rheumatoid arthritis in northern China. *Aging (Albany NY).* (2020) 12:3190–204. doi: 10.18632/aging.v12i4

69. Qin Y, Wang Y, Meng F, Feng M, Zhao X, Gao C, et al. Identification of biomarkers by machine learning classifiers to assist diagnose rheumatoid arthritis-associated interstitial lung disease. *Arthritis Res Ther.* (2022) 24:115. doi: 10.1186/s13075-022-02800-2

70. Karlson EW, van Schaardenburg D, van der Helm-van Mil AH. Strategies to predict rheumatoid arthritis development in at-risk populations. *Rheumatol (Oxford).* (2016) 55:6–15. doi: 10.1093/rheumatology/keu287

71. Burgers LE, Raza K, van der Helm-van Mil AH. Window of opportunity in rheumatoid arthritis - definitions and supporting evidence: from old to new perspectives. *RMD Open.* (2019) 5:e000870. doi: 10.1136/rmdopen-2018-000870

72. Hazlewood GS, Barnabe C, Tomlinson G, Marshall D, Devoe DJ, Bombardier C. Methotrexate monotherapy and methotrexate combination therapy with traditional and biologic disease modifying anti-rheumatic drugs for rheumatoid arthritis: A network meta-analysis. *Cochrane Database Syst Rev.* (2016) 2016:CD010227. doi: 10.1002/14651858.CD010227.pub2

73. Nahon P, Bamba-Funck J, Layese R, Trépo E, Zucman-Rossi J, Cagnot C, et al. Integrating genetic variants into clinical models for hepatocellular carcinoma risk stratification in cirrhosis. *J Hepatol.* (2023) 78:584–95. doi: 10.1016/j.jhep.2022.11.003

74. Martin AR, Kanai M, Kamatani Y, Okada Y, Neale BM, Daly MJ. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet.* (2019) 51:584–91. doi: 10.1038/s41588-019-0379-x

75. Ruan Y, Lin YF, Feng YA, Chen CY, Lam M, Guo Z, et al. Improving polygenic prediction in ancestrally diverse populations. *Nat Genet.* (2022) 54:573–80. doi: 10.1038/s41588-022-01054-7

76. Hao L, Kraft P, Berriz GF, Hynes ED, Koch C, Kumar PKV, et al. Development of a clinical polygenic risk score assay and reporting workflow. *Nat Med.* (2022) 28:1006–13. doi: 10.1038/s41591-022-01767-6

77. Li H, Cai Z, Wang J, Tang J, Ding W, Lin CT, et al. FedTP: federated learning by transformer personalization. *IEEE Trans Neural Netw Learn Syst.* (2023). doi: 10.1109/TNNLS.2023.3269062

78. Gu X, Sabrina F, Fan Z, Sohail S. A review of privacy enhancement methods for federated learning in healthcare systems. *Int J Environ Res Public Health.* (2023) 20:6539. doi: 10.3390/ijerph20156539

79. Haggenmüller S, Schmitt M, Krieghoff-Henning E, Hekler A, Maron RC, Wies C, et al. Federated learning for decentralized artificial intelligence in melanoma diagnostics. *JAMA Dermatol.* (2024) 160:303–11. doi: 10.1001/jamadermatol.2023.5550

80. van den Berg R, Ohrndorf S, Kortekaas MC, van der Helm-van Mil AHM. What is the value of musculoskeletal ultrasound in patients presenting with arthralgia to predict inflammatory arthritis development? A systematic literature review. *Arthritis Res Ther.* (2018) 20:228. doi: 10.1186/s13075-018-1715-8

81. Jo J, Tian C, Xu G, Sarazin J, Schiopu E, Gandikota G, et al. Photoacoustic tomography for human musculoskeletal imaging and inflammatory arthritis detection. *Photoacoustics.* (2018) 12:82–9. doi: 10.1016/j.pacs.2018.07.004

82. Madani A, Arnaout R, Mofrad M, Arnaout R. Fast and accurate view classification of echocardiograms using deep learning. *NPJ Digit Med.* (2018) 1:6. doi: 10.1038/s41746-017-0013-1

83. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. *NPJ Digit Med.* (2019) 2:43. doi: 10.1038/s41746-019-0122-0

84. Lei Y, Tian Y, Shan H, Zhang J, Wang G, Kalra MK. Shape and margin-aware lung nodule classification in low-dose CT images *via* soft activation mapping. *Med Image Anal.* (2020) 60:101628. doi: 10.1016/j.media.2019.101628

85. Rynazal R, Fujisawa K, Shiroma H, Salim F, Mizutani S, Shiba S, et al. Leveraging explainable AI for gut microbiome-based colorectal cancer classification. *Genome Biol.* (2023) 24:21. doi: 10.1186/s13059-023-02858-4

86. Lee WY, Lee Y, Lee S, Kim YW, Kim JH. A machine learning approach for recommending herbal formulae with enhanced interpretability and applicability. *Biomolecules.* (2022) 12:1604. doi: 10.3390/biom12111604

87. Lee YG, Choi SC, Kang Y, Kim KM, Kang CS, Kim C. Constructing a reference genome in a single lab: the possibility to use oxford nanopore technology. *Plants (Basel).* (2019) 8:270. doi: 10.3390/plants8080270

88. Sun Y, Chen BR, Deshpande A. Epigenetic regulators in the development, maintenance, and therapeutic targeting of acute myeloid leukemia. *Front Oncol.* (2018) 8:41. doi: 10.3389/fonc.2018.00041

89. Rodríguez-Molina JB, West S, Passmore LA. Knowing when to stop: Transcription termination on protein-coding genes by eukaryotic RNAPII. *Mol Cell.* (2023) 83:404–15. doi: 10.1016/j.molcel.2022.12.021

90. Graves PR, Haystead TA. Molecular biologist's guide to proteomics. *Microbiol Mol Biol Rev.* (2002) ;66:39–63. doi: 10.1128/MMBR.66.1.39-63.2002

91. Guo H, Guo H, Zhang L, Tang Z, Yu X, Wu J, et al. Metabolome and transcriptome association analysis reveals dynamic regulation of purine metabolism and flavonoid synthesis in transdifferentiation during somatic embryogenesis in cotton. *Int J Mol Sci.* (2019) 20:2070. doi: 10.3390/ijms20092070

92. Smeekens SP, Huttenhower C, Riza A, van de Veerdonk FL, Zeeuwen PL, Schalkwijk J, et al. Skin microbiome imbalance in patients with STAT1/STAT3 defects impairs innate host defense responses. *J Innate Immun.* (2014) 6:253–62. doi: 10.1159/000351912

93. Tarazona S, Balzano-Nogueira L, Gómez-Cabrero D, Schmidt A, Imhof A, Hankemeier T, et al. Harmonization of quality metrics and power calculation in multi-omic studies. *Nat Commun.* (2020) 11:3092. doi: 10.1038/s41467-020-16937-8

94. Yi D, Bayer T, Badenhorst CPS, Wu S, Doerr M, Höhne M, et al. Recent trends in biocatalysis. *Chem Soc Rev.* (2021) 50:8003–49. doi: 10.1039/D0CS01575J

95. Brown MV, McDunn JE, Gunst PR, Smith EM, Milburn MV, Troyer DA, et al. Gunst PR Cancer detection and biopsy classification using concurrent histopathological and metabolomic analysis of core biopsies. *Genome Med.* (2012) 4:33. doi: 10.1186/gm332

96. Yang S, Hollister AM, Orchard EA, Chaudhery SI, Ostanin DV, Lokitz SJ, et al. Quantification of bone changes in a collagen-induced arthritis mouse model by reconstructed three dimensional micro-CT. *Biol Proced Online.* (2013) 15:8. doi: 10.1186/1480-9222-15-8

97. Liao KP, Kurreeman F, Li G, Duclos G, Murphy S, Guzman R, et al. Associations of autoantibodies, autoimmune risk alleles, and clinical diagnoses from the electronic medical records in rheumatoid arthritis cases and non-rheumatoid arthritis controls. *Arthritis Rheumatol.* (2013) 65:571–81. doi: 10.1002/art.37801

98. Kurreeman F, Liao K, Chibnik L, Hickey B, Stahl E, Gainer V, et al. Genetic basis of autoantibody positive and negative rheumatoid arthritis risk in a multi-ethnic cohort derived from electronic health records. *Am J Hum Genet.* (2011) 88:57–69. doi: 10.1016/j.ajhg.2010.12.007

99. Li H, Guan Y. Multilevel modeling of joint damage in rheumatoid arthritis. *Adv Intell Syst.* (2022) 4:2200184. doi: 10.1002/aisy.202200184

100. Sun D, Nguyen TM, Allaway RJ, Wang J, Chung V, Yu TV, et al. RA2-DREAM challenge community. A crowdsourcing approach to develop machine learning models to quantify radiographic joint damage in rheumatoid arthritis. *JAMA Netw Open.* (2022) 5:e2227423. doi: 10.1001/jamanetworkopen.2022.27423

101. Fiorentino MC, Cipolletta E, Filippucci E, Grassi W, Frontoni E, Moccia S. A deep-learning framework for metacarpal-head cartilage-thickness estimation in ultrasound rheumatological images. *Comput Biol Med.* (2022) 141:105117. doi: 10.1016/j.compbiomed.2021.105117

102. Andersen JKH, Pedersen JS, Laursen MS, Holtz K, Grauslund J, Savarimuthu TR, et al. Neural networks for automatic scoring of arthritis disease activity on ultrasound images. *RMD Open*. (2019) 5:e000891. doi: 10.1136/rmdopen-2018-000891

103. Singh JA, Hossain A, Mudano AS, Tanjong Ghogomu E, Suarez-Almazor ME, Buchbinder R, et al. Biologics or tofacitinib for people with rheumatoid arthritis naive to methotrexate: a systematic review and network meta-analysis. *Cochrane Database Syst Rev*. (2017) 5:CD012657. doi: 10.1002/14651858

104. Bluett J, Riba-Garcia I, Verstappen SMM, Wendling T, Ogungbenro K, Unwin RD, et al. Development and validation of a methotrexate adherence assay. *Ann Rheum Dis*. (2019) 78:1192–7. doi: 10.1136/annrheumdis-2019-215446

105. Kalweit M, Burden AM, Boedecker J, Hügle T, Burkard T. Patient groups in Rheumatoid arthritis identified by deep learning respond differently to biologic or targeted synthetic DMARDs. *PloS Comput Biol*. (2023) 19:e1011073. doi: 10.1371/journal.pcbi.1011073

106. Jain S, Eadon MT. Spatial transcriptomics in health and disease. *Nat Rev Nephrol*. (2024). doi: 10.1038/s41581-024-00841-1

107. Wu H, Dixon EE, Xuanyuan Q, Guo J, Yoshimura Y, Debashish C, et al. High resolution spatial profiling of kidney injury and repair using RNA hybridization-based in *situ* sequencing. *Nat Commun*. (2024) 15:1396. doi: 10.1038/s41467-024-45752-8

108. Kiessling P, Kuppe C. Spatial multi-omics: novel tools to study the complexity of cardiovascular diseases. *Genome Med*. (2024) 16:14. doi: 10.1186/s13073-024-01282-y

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 12 June 2024
DOI 10.3389/fimmu.2024.1419363

OPEN ACCESS

EDITED BY
Xu-jie Zhou,
Peking University, China

REVIEWED BY
Jin Li,
Hainan Medical University, China
Konstantine Halkidis,
University of Kansas Medical Center,
United States

*CORRESPONDENCE
Shengao Qin
shengaoqin123@163.com
Guowu Ma
mgw640242000@aliyun.com
Fan Zhang
fanzhang2023@tongji.edu.cn

RECEIVED 18 April 2024
ACCEPTED 31 May 2024
PUBLISHED 12 June 2024

CITATION
Bai Y, Wang J, Feng X, Xie L, Qin S, Ma G and Zhang F (2024) Identification of drug targets for Sjögren's syndrome: multi-omics Mendelian randomization and colocalization analyses.
Front. Immunol. 15:1419363.
doi: 10.3389/fimmu.2024.1419363

Identification of drug targets for Sjögren's syndrome: multi-omics Mendelian randomization and colocalization analyses

Yingjie Bai[1,2], Jiayi Wang[1,2], Xuefeng Feng[1,2], Le Xie[3], Shengao Qin[4,5]*, Guowu Ma[1,2,6]* and Fan Zhang[7]*

[1]School of Stomatology, Dalian Medical University, Dalian, China, [2]Academician Laboratory of Immune and Oral Development & Regeneration, Dalian Medical University, Dalian, China, [3]Shanghai Engineering Research Center of Tooth Restoration and Regeneration & Tongji Research Institute of Stomatology & Department of Oral Implantology, Stomatological Hospital and Dental School, Tongji University, Shanghai, China, [4]Salivary Gland Disease Center and Beijing Key Laboratory of Tooth Regeneration and Function Reconstruction, Beijing Laboratory of Oral Health and Beijing Stomatological Hospital, Capital Medical University, Beijing, China, [5]Beijing Laboratory of Oral Health, Capital Medical University, Beijing, China, [6]Department of Stomatology, Stomatological Hospital Affiliated School of Stomatology of Dalian Medical University, Dalian, China, [7]Department of Stomatology, Shanghai East Hospital, School of Medicine, Tongji University, Shanghai, China

Background: Targeted therapy for Sjögren's syndrome (SS) has become an important focus for clinicians. Multi-omics-wide Mendelian randomization (MR) analyses have provided new ideas for identifying potential drug targets.

Methods: We conducted summary-data-based Mendelian randomization (SMR) analysis to evaluate therapeutic targets associated with SS by integrating DNA methylation, gene expression and protein quantitative trait loci (mQTL, eQTL, and pQTL, respectively). Genetic associations with SS were derived from the FinnGen study (discovery) and the GWAS catalog (replication). Colocalization analyses were employed to determine whether two potentially relevant phenotypes share the same genetic factors in a given region. Moreover, to delve deeper into potential regulation among DNA methylation, gene expression, and protein abundance, we conducted MR analysis to explore the causal relationship between candidate gene methylation and expression, as well as between gene expression and protein abundance. Drug prediction and molecular docking were further employed to validate the pharmacological activity of the candidate drug targets.

Results: Upon integrating the multi-omics data, we identified three genes associated with SS risk: TNFAIP3, BTN3A1, and PLAU. The methylation of cg22068371 in BTN3A1 was positively associated with protein levels, consistent with the negative effect of cg22068371 methylation on the risk of SS. Additionally, positive correlations were observed between the gene methylation of PLAU (cg04939496) and expression, as well as between expression and protein levels. This consistency elucidates the promotional effects of PLAU on SS risk at the DNA methylation, gene expression, and protein levels. At the protein level, genetically predicted TNFAIP3 (OR 2.47, 95% CI 1.56–3.92) was positively associated with SS risk, while BTN3A1 (OR 2.96E-03, 95% CI 2.63E-04–3.33E-02) was negatively

associated with SS risk. Molecular docking showed stable binding for candidate drugs and target proteins.

Conclusion: Our study reveals promising therapeutic targets for the treatment of SS, providing valuable insights into targeted therapy for SS. However, further validation through future experiments is warranted.

KEYWORDS

Sjögren's syndrome, Mendelian randomization, drug target, methylation, gene expression, protein, proteomics, genetics

1 Introduction

Sjögren's syndrome (SS) is a refractory autoimmune disease pathologically characterized by progressive destruction of exocrine glands, involving several systemic organs such as the oral cavity, eyes, kidneys, liver, lungs, joints, and nerves (1). SS is associated with a significantly higher incidence of non-Hodgkin's lymphoma compared to other autoimmune disease, making it one of the diseases closely associated with malignancy (2, 3). The efficacy of drugs such as lubricants, glucocorticoids, and immunosuppressants, which are commonly used in the clinical treatment of SS, is not always effective and there is a certain degree of adverse reactions, such as local allergies, gastrointestinal damage, and skin lesions (4). Therefore, exploring drug targets for the treatment of SS is of far-reaching clinical significance and can provide theoretical support for the development of new drugs for the treatment of SS.

Finding drug targets through genetic means can not only greatly improve the efficiency of drug development but also save a lot of human and material resources (5, 6). In addition, proteins, as key regulators of molecular pathways, have widely emerged as a major source of drug targets (7, 8). It has been demonstrated that disease-related protein drug targets supported by genetic associations have a higher likelihood of gaining market approval (5). Therefore, constructing drug targets based on genetic information is a more effective approach to developing drugs.

Mendelian randomization (MR) analyses, which utilize genetic variation as an instrumental variable to enhance inferences about causal relationships between exposures and outcomes, have been widely employed in drug target development and drug repurposing. In contrast to observational studies, MR circumvents the influence of environmental and self-adoption factors because genetic variants are randomly allocated at the time of conception. With advancements in high-throughput genomic and proteomic technologies in plasma and cerebrospinal fluid, MR-based strategies have facilitated the identification of potential therapeutic targets for numerous diseases such as inflammatory bowel disease, multiple sclerosis, and colorectal cancer (9–11). In this study, we systematically identified molecular signatures of genes associated with SS risk by integrating DNA methylation, gene expression, and protein abundance data, providing comprehensive directions for future research and potential therapeutic targets.

Abbreviations: SS, Sjögren's syndrome; QTL, Quantitative trait loci; MR, Mendelian randomization; SMR, Summary-data-based Mendelian randomization; GWAS, Genome-wide association studies; HEIDI, Heterogeneity in the dependent instrument; SNPs, Single nucleotide polymorphisms; FDR, False discovery rate, PPH4: Posterior probability of H4; OR, Odds ratio; CI, Confidence interval; TNFAIP3, Tumor necrosis factor a induced protein 3; NF-KB, Nuclear factor kB; SLE, Systemic lupus erythematosus; BTN3A1, Butyrophilin 3A1.

2 Materials and methods

2.1 Data sources for DNA methylation, gene expression and protein quantitative trait loci

The schematic illustration of the identification of drug targets for SS and the study design is illustrated in Figures 1 and 2. Methylated quantitative trait loci (mQTL) data were obtained from SNP-CpG associations in the blood of individuals of European ancestry from 1980 by McRae et al. (12). The blood expression quantitative trait loci (eQTL) dataset was extracted from the eQTLGen consortium (https://eqtlgen.org/), comprising 31,684 individuals, 16,987 genes, and 31,684 cis eQTLs derived from blood samples, primarily from healthy European individuals (13). The protein quantitative trait loci (pQTL) dataset was derived from a large-scale pQTL study of 35,559 Icelanders, with summary statistics extracted for genetic associations at the level of 4907 circulating proteins (14).

2.2 SS data sources

Genome-wide association studies (GWAS) data for the SS discovery cohort were obtained from FinnGen Release 10 (https://www.finngen.fi/en). The study was conducted on individuals of

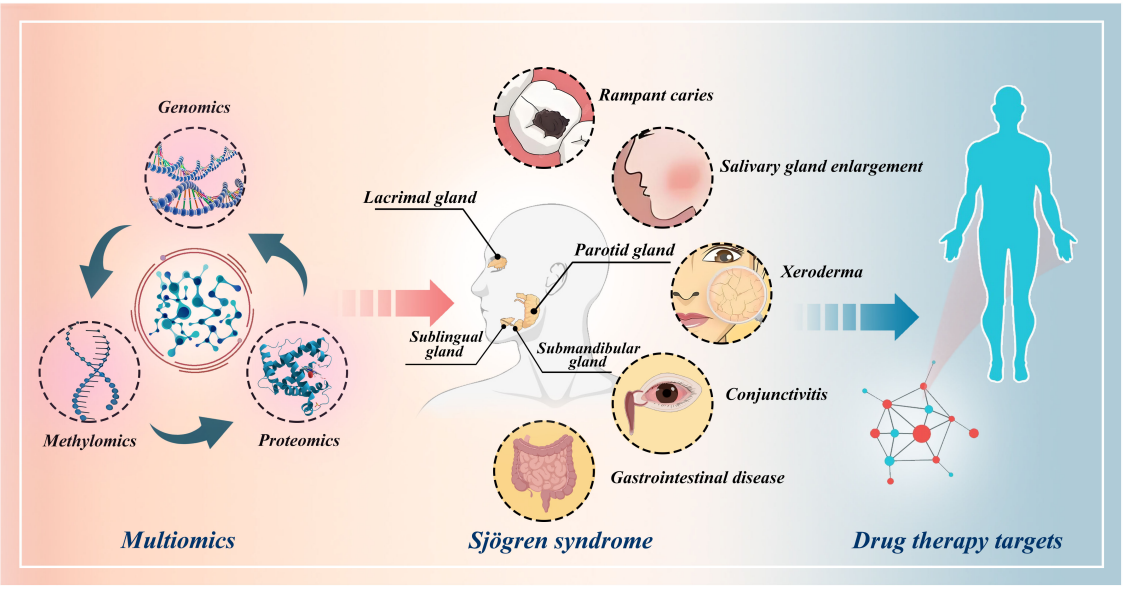


FIGURE 1
Schematic illustration of the identification of drug targets for Sjögren's syndrome through multi-omics Mendelian randomization study.

European ancestry and comprised a total of 2,735 SS cases and 399,355 control cases. SS patients were identified based on ICD-10 code M35.0, ICD-9 code 7102, or ICD-8 code 73490 (primarily relying on ICD-10 codes). The validation cohort was sourced from the GWAS Catalog GCST90018920 and included 1,599 SS cases and 658,316 control cases (https://www.ebi.ac.uk/gwas/).

2.3 Summary-data-based MR analysis

Summary-data-based Mendelian Randomization (SMR) analysis is a statistical method based on the principles of Mendelian randomization that uses genetic variation (single nucleotide polymorphisms, SNP) as an instrumental variable to assess the causal relationship between an exposure and an outcome, and is mainly applied for causal inference between genes and complex diseases or traits, especially when direct randomized controlled trials are not feasible. Compared to MR analysis, SMR analysis relies on pooled results from genome-wide association studies (GWAS) rather than individual-level data, an approach that is more favorable in terms of privacy protection and data sharing. SMR analysis can be combined at the multi-omics level to help researchers explore potential causal relationships between specific drug targets and diseases. In this study, we used SNPs as instrumental variables, mQTL, eQTL, pQTL as exposures, and SS as outcomes. The SMR analysis was conducted using SMR 1.3.1 software (https://yanglab.westlake.edu.cn/software/smr/) (15).

We screened for the top associated cis-QTL by defining a chromosome window centered around the target gene (± 1000 kb) and passing a *P*-value threshold of 5.0×10^{-8}. The Heterogeneity in Dependent Instrument (HEIDI) test was primarily employed to assess whether a gene SNP-mediated phenotype resulted from a linkage disequilibrium reaction, with the criterion of *P*-HEIDI > 0.01. If the *P*-value of the HEIDI test was less than 0.01, it indicated a heterogeneous association, suggesting possible pleiotropy. A false discovery rate (FDR) of $\alpha = 0.05$, based on the Benjamini-Hochberg method, was applied for multiple testing. Associations with FDR-corrected *P*-values < 0.05 and P-HEIDI > 0.01 were analyzed for colocalization.

2.4 Colocalization analysis

Colocalization analysis can be utilized to genetically co-localize two potentially related phenotypes, determining whether they share common genetic causal variants within a given region. We conducted colocalization analyses to assess whether SS and the identified mQTLs, eQTLs, or pQTLs are influenced by linkage disequilibrium. Five exclusivity hypotheses were examined in the colocalization analyses: 1) No association with any of the traits (H0); 2) Association with trait 1 only (H1); 3) Association with trait 2 only (H2); 4) Causal variants for the two traits are different (H3); 5) Causal variants for the two traits (H4) are the same. For pQTL-GWAS colocalization, eQTL-GWAS, and mQTL-GWAS, the colocalization region windows were set at ±1000 kb, ± 1000 kb, and ±500 kb, respectively. A posterior probability of H4 (PPH4) greater than 0.70 was considered strong evidence for colocalization.

2.5 Integrating results at the multi-omics level of evidence

To achieve a comprehensive understanding of the association of gene-related regulation with SS across different levels, we integrated results from three distinct gene regulatory layers. Considering that proteins represent the final expression products of genes and are

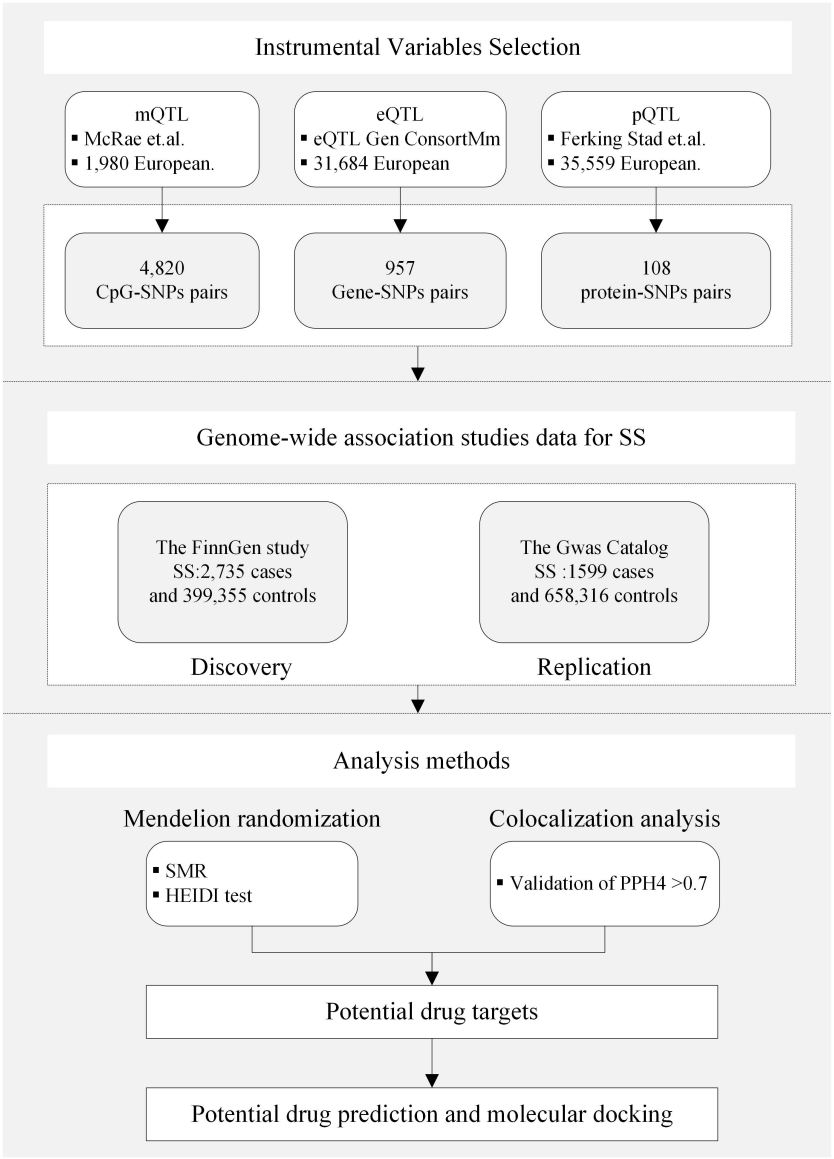


FIGURE 2
Study design. QTL, quantitative trait loci; SS, Sjögren's syndrome; SNP, single nucleotide polymorphisms; SMR, summary-based Mendelian randomization; HEIDI, heterogeneity in the dependent instrument; PPH4, posterior probability of H4.

prime targets for drug therapy, genes associated with SS at the protein level were prioritized as high-quality candidates. Based on this principle, the final candidate genes were categorized into two tiers: 1) Tier 1 genes: These genes were defined as having associations with SS at protein abundance level (FDR-corrected P-value < 0.05), PPH4 of colocalization > 0.7, and associations with SS at gene methylation or expression level (original P-value < 0.05); 2) Tier 2 genes: These genes were defined as having associations with SS at protein abundance level (FDR-corrected P-value < 0.05), and associations with SS at both gene methylation and expression levels (FDR-corrected P-value < 0.05), PPH4 of colocalization > 0.7. Moreover, to delve deeper into potential regulation among methylation, expression, and protein abundance, we conducted MR analysis and colocalization analysis to explore the causal relationship between related DNA methylation and expression, as well as between gene expression and protein abundance.

2.6 Candidate drug prediction and molecular docking

Predicting drug candidates through drug targets is a critical step in drug discovery and development. We searched each of the key genes in the DrugBank database to obtain information about the drugs associated with these genes (https://go.drugbank.com/) (16). DrugBank is a comprehensive drug database that contains information about the pharmacological properties, targets, and other information about drugs. DrugBank is often used in

conjunction with other databases and tools to explore multi-targeted mechanisms of action of a drug and its potential therapeutic effects.

To further understand the interaction between drug candidates and targets, molecular docking technique was used in this study. The drug structure data and target protein structure data were obtained from the PubChem Compound Database (https://pubchem.ncbi.nlm.nih.gov/), and the Protein Data Bank (http://www.rcsb.org/), respectively (17). We employed semi-flexible docking to form stable complexes. Protein pretreatment (removal of water molecules and excess ligands, addition of hydrogen atoms) was accomplished using PyMOL 2.4. AutoDock Tools 1.5.6 was used to generate PDBQT files for docking simulations. Molecular docking analysis was performed using AutoDock Vina 1.2.2 (http://autodock.scripps.edu/) (18). Binding energies less than -5 kcal/mol were defined to indicate effective ligand-receptor binding, while binding energies less than -7 kcal/mol indicated strong binding activity.

3 Results

3.1 DNA methylation and SS

A total of 4820 CpG sites were identified as associated with SS risk ($P < 0.05$) (Supplementary Table S1). After correction for multiple testing and colocalization analysis, we identified a total of 154 CpG sites associated with SS ($P_{(FDR)} < 0.05$, PPH4 > 0.70) (Table 1, Supplementary Table S1). For instance, cg18909389 (OR 0.35, 95% CI 0.31–0.41) and cg12257344 (OR 0.33, 95% CI 0.28–0.38), located in CLIC1, as well as cg00355613 (OR 0.27, 95% CI 0.22–0.33), cg15745284 (OR 0.28, 95% CI 0.23–0.34), cg21289669 (OR 0.23, 95% CI 0.18–0.29), and cg07518714 (OR 0.27, 95% CI 0.22–0.34), located in TNXB, were negatively associated with SS risk. Additionally, cg05571472 (OR 6.13, 95% CI 4.33–8.69), located in C6orf48, was positively associated with SS risk. In the validation cohort, many CpG sites such as C6orf25 (cg06132876), PLAU (cg04939496), and TNXB (cg07237769) were replicated (Supplementary Table S2).

TABLE 1 Associations of DNA methylation with Sjögren's syndrome (SS).

Gene	Probe ID	OR (95% CI)	*P* value	PPH4
CLIC1	cg18909389	0.35 (0.31–0.41)	7.75E-46	0.98
TRIM31	cg11100081	0.59 (0.55–0.64)	6.48E-45	<0.01
CLIC1	cg12257344	0.33 (0.28–0.38)	1.67E-44	0.98
TNXB	cg00355613	0.27 (0.22–0.33)	3.72E-36	0.98
HLA-DMB	cg13524037	2.47 (2.14–2.86)	1.69E-34	<0.01
HLA-DPB1	cg14373797	0.8 (0.77–0.83)	2.22E-34	<0.01
C6orf27	cg05239811	0.25 (0.2–0.31)	8.98E-34	0.06
TNXB	cg15745284	0.28 (0.23–0.34)	3.94E-33	0.93
TNXB	cg21289669	0.23 (0.18–0.29)	4.47E-32	0.97
TNXB	cg07518714	0.27 (0.22–0.34)	8.58E-32	0.97
HLA-DPA1	cg05751055	0.51 (0.45–0.57)	1.25E-29	<0.01
TNXB	cg21642103	0.19 (0.14–0.26)	3.37E-28	0.98
TNXB	cg15014577	0.18 (0.14–0.25)	2.29E-27	0.97
COL11A2	cg22122760	0.43 (0.37–0.51)	1.16E-26	<0.01
HLA-DRA	cg08882389	0.18 (0.13–0.25)	1.30E-26	0.12
TNXB	cg11493661	0.17 (0.12–0.24)	1.61E-25	0.98
C6orf48	cg05571472	6.13 (4.33–8.69)	2.08E-24	0.96
CLIC1	cg18402034	0.14 (0.09–0.2)	3.65E-24	0.92
XXbac-BPG308K3.6	cg06608359	0.56 (0.5–0.63)	4.55E-23	1.00
GPSM3	cg21386484	0.31 (0.24–0.39)	8.05E-23	0.78

OR, odds ratio; CI, confidence interval; PPH4, posterior probability of H4.

3.2 Gene expression and SS

A total of 957 genes were identified as associated with SS risk ($P < 0.05$) (Supplementary Table S3). After correcting for multiple testing ($P_{(FDR)} < 0.05$) and conducting colocalization analysis (PPH4 > 0.7), genetically predicted CA8 (OR 0.58, 95% CI 0.43–0.77), BACH2 (OR 0.51, 95% CI 0.36–0.72), RP4–555D20.2 (OR 0.59, 95% CI 0.44- 0.78), RP11–148O21.4 (OR 0.78, 95% CI 0.70–0.87), BLK (OR 0.73, 95% CI 0.64–0.84), KIAA1683 (OR 0.83, 95% CI 0.75–0.91), RP11–148O21.2 (OR 0.45, 95% CI 0.32–0.65), TNXA (OR 0.32, 95% CI 0.27–0.38), VSIG10 (OR 0.75, 95% CI 0.65–0.86), and WSB2 (OR 0.72, 95% CI 0.62–0.84) were negatively correlated with SS risk. Conversely, genetically predicted PLAU (OR 1.77, 95% CI 1.40–2.24), FAM167A (OR 1.20, 95% CI 1.11–1.30), MIF4GD (OR 1.41, 95% CI 1.18–1.69), and SYNGR1 (OR 1.21, 95% CI 1.10–1.33) were positively associated with SS risk (Figure 3). The associations of FAM167A, BLK, RP11–148O21.2, RP11–148O21.4, RP11–148O21.6, SYNGR1, MIF4GD, and CA8 were replicated in the validation cohort (Supplementary Table S4).

10.3389/fimmu.2024.1419363

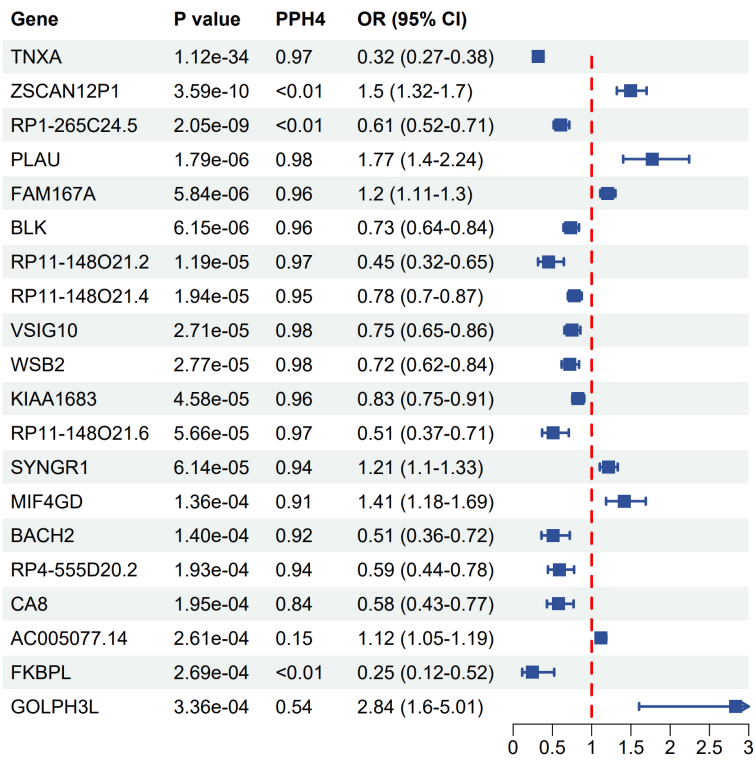


FIGURE 3
Forest plot of associations between gene expression with SS. OR, odds ratio; CI, confidence interval; PPH4, posterior probability of H4.

3.3 Protein and SS

A total of 108 proteins were associated with SS risk at the $P < 0.05$ level (Supplementary Table S5). After adjusting for multiple tests, 8 proteins were associated with the risk of Sjögren at the $P_{(FDR)} < 0.05$ level. HSPA1B (OR 2.41E-03, 95% CI 3.42E-04–1.70E-02), LY6G6D (OR 2.88E-03, 95% CI 2.73E-04–3.03E-02), BTN3A1 (OR 2.96E-03, 95% CI 2.63E-04 -3.33E-02), SFTA2 (OR 0.08, 95% CI 0.02–0.26), HSPA1L (OR 0.31, 95% CI 0.17–0.56), and VARS1 (OR 0.27, 95% CI 0.14–0.53) were observed to be negatively correlated with SS risk. Conversely, PLAU (OR 1.61, 95% CI 1.32–1.95) and TNFAIP3 (OR 2.47, 95% CI 1.56–3.92) were positively associated with SS risk (Figure 4). The results of the colocalization analysis found high supportive colocalization evidence for BTN3A1 (PPH4 = 0.86) and TNFAIP3 (PPH4 = 0.90). BTN3A1 (OR 0.01, 95% CI 6.31E-04–0.09, $P_{(FDR)}$ = 0.036) was replicated in the validation cohort (Supplementary Table S6).

3.4 Integrating evidence from multi-omics levels

After integrating evidence at the multi-omics level, we identified 2 tier 1 genes, TNFAIP3 and BTN3A1, and the tier 2 gene PLAU

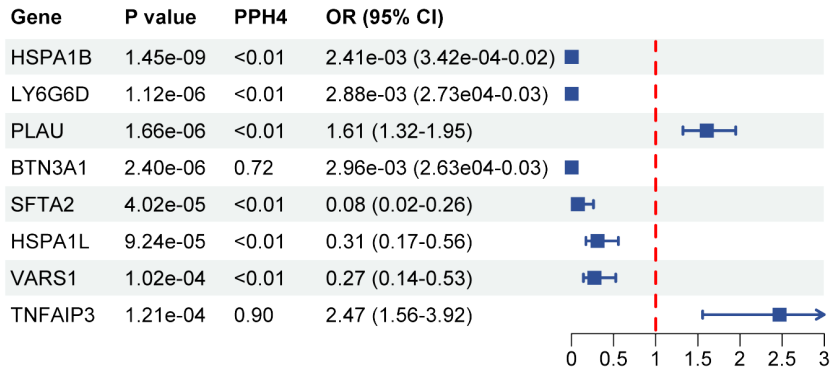


FIGURE 4
Forest plot of associations between protein with SS. OR, odds ratio; CI, confidence interval; PPH4, posterior probability of H4.

(Table 2, Figure 5). In the validation cohort, BTN3A1 was replicated at the level of circulating proteins ($P_{(FDR)}$ = 0.036) (Supplementary Table S6). In exploring the association between gene methylation, expression, and protein abundance, we found that the methylation of cg22068371 in BTN3A1 was positively associated with protein levels, which is consistent with the negative effect of cg22068371 methylation on the risk of SS (Supplementary Table S7). Positive correlations were also observed between the gene methylation of PLAU (cg04939496) and gene expression, as well as between gene expression and protein levels, which were corroborated with the positive effect on SS risk. Strong colocalization supportive evidence was observed between the methylation of BTN3A1 (cg22068371) and protein abundance, and between the gene methylation of PLAU (cg04939496) and expression.

3.5 Molecular docking

We identified drug candidates related to the target proteins through DrugBank, and the corresponding IDs of drug and protein structure data can be viewed in Table 3. The molecular docking of these drugs and proteins encoded by these corresponding target genes was performed using AutoDock Vina 1.2.2. The coordinate of the docking box for protein BTN3A1 was x: y: z= 17.074: -36.189: -7.092. The coordinate of the docking box for protein PLAU was x: y: z= 17.074: -0.176: 18.957. The coordinate of the docking box for protein TNFAIP3 was x: y: z= 20.145: 15.764: 21.938. The drug candidates were attached to their protein targets through hydrogen bonding and strong electrostatic interactions (Figure 6). PLAU-Amiloride (-7.4 kcal/mol) and TNFAIP3-Sulfasalazine (-7.3 kcal/mol) had the lowest binding energies and were considered to be the most potential binding mode between ligand and protein.

4 Discussion

Genes are specific sequences on DNA molecules. They encode proteins or RNAs that regulate gene expression, which can serve as new targets for drug development, i.e., drugs can bind specifically to these molecules, thereby modulating their function or expression. To our knowledge, this study represents the first attempt to utilize MR to identify potential drug targets for SS. We integrated results from multi-omics level evidence, reinforcing the causal relationship between genes and SS risk. Additionally, we combined SMR and colocalization analyses to pinpoint common drivers between potential therapeutic targets and SS risk, while excluding potential confounders. Our study pinpointed TNFAIP3, BTN3A1, and PLAU as potential drug targets for SS. Notably, BTN3A1 was also found to be associated with SS in the validation cohort using a similar analytical approach, underscoring the reliability of the potential drug targets identified in this study.

TNFAIP3 was identified as positively associated with SS risk with high colocalization support. Tumor necrosis factor alpha-induced protein 3 (TNFAIP3) is a crucial nuclear factor κB (NF-κB) regulatory protein that modulates NF-κB expression and apoptosis through multiple pathways (19). Associations between TNFAIP3 and various autoimmune diseases, including SS, rheumatoid arthritis, systemic lupus erythematosus (SLE), and systemic sclerosis, have been documented (16–18). TNFAIP3 has also been identified as one of the susceptibility loci for SS by GWAS (20). Activation of the NF-κB pathway in activated B cells is a key step in the pathogenesis of primary SS (21). The TNFAIP3 gene encodes the A20 protein, essential for the development and functional expression of dendritic cells, B and T cells, and macrophages. The A20 protein serves as a critical negative regulator of NF-κB, and reduced negative regulatory activity of A20 may permit excessive immunoreactivity, leading to increased auto-reactivity (22, 23). Notably, our study found that the top single nucleotide polymorphism (SNP) associated with SS located in TNFAIP3 was rs5029939, which is similar to previous findings that this SNP has been associated with various autoimmune diseases, including SLE, systemic sclerosis, and other autoimmune disorders (24–26). Therefore, we hypothesize that rs5029939 may also be a genetic risk factor for SS susceptibility, although further experimental validation is warranted.

Butyrophilin 3A1 (BTN3A1) is a type I transmembrane protein belonging to the immunoglobulin (Ig) superfamily, with immunomodulatory and antigen-presenting functions. It has been implicated in autoimmune diseases, diabetes mellitus, multiple sclerosis, and cancer (27). Several SNPs, including rs1796520, rs3857550, rs3208733, rs6912853, and rs10456045, of BTN3A1 have been associated with SLE patients (28, 29). Our MR analysis provides

TABLE 2 Tier of genetically predicted methylation, expression, and protein of candidate gene with SS.

Gene	Tier	mQTL				eQTL			pQTL		
		Probe	OR (95% CI)	*P* value	$P_{(FDR)}$ value	OR (95% CI)	*P* value	$P_{(FDR)}$ value	OR (95% CI)	*P* value	$P_{(FDR)}$ value
BTN3A1	Tier 1	Cg 22068371	0.47 (0.25–0.88)	0.018	0.570	1.12 (0.96–1.30)	0.163	0.840	2.96E-03 (2.63E04–0.03)	2.40E-06	3.78E-04
TNFAIP3	Tier 1	–				4.35 (1.45–13)	0.009	0.439	2.47 (1.56–3.92)	1.21E-04	0.014
PLAU	Tier 2	Cg 04939496	1.35 (1.18–1.54)	6.73E-06	9.11E-04	1.77 (1.4–2.24)	1.79E-06	4.11E-04	1.61 (1.32–1.95)	1.66E-06	3.19E-04

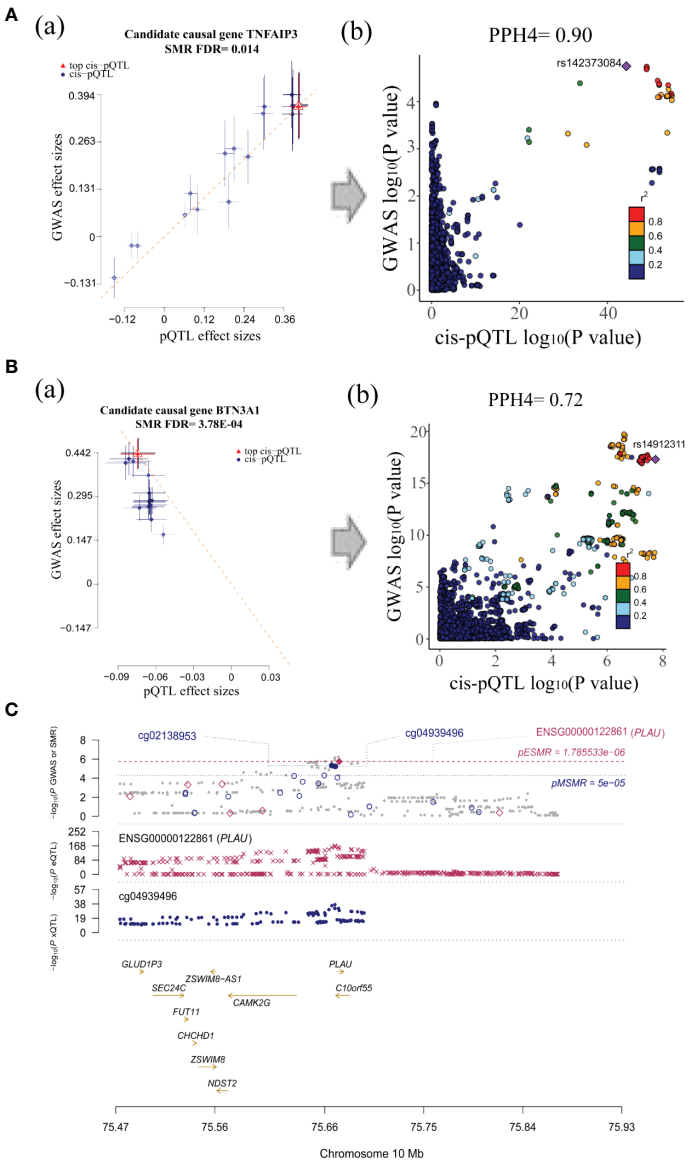


FIGURE 5
Visualization for associations between candidate causal genes and SS. **(A)** The SMR (a) and colocalization analysis (b) between TNFAIP3 protein and SS GWAS. **(B)** The SMR (a) and colocalization analysis (b) between BTN3A1 protein and SS GWAS (all SMR FDR < 0.05; HEIDI test $P > 0.01$; PPH4 of colocalization > 0.7, the r^2 value indicates the linkage disequilibrium (LD) between the variants and the top SNPs.). **(C)** Associations between PLAU methylation, expression and SS GWAS.

TABLE 3 Docking results of potential targets with drugs.

Target	PDB ID	Drug	PubChem ID	Binding energy (kcal/mol)
TNFAIP3	2VFJ	Acetylcysteine	12035	-4.2
TNFAIP3	2VFJ	Aminosalicylic acid	4649	-5.0
TNFAIP3	2VFJ	Mesalamine	4075	-5.0
TNFAIP3	2VFJ	Sulfasalazine	5339	-7.3
BTN3A1	4F80	Valproic acid	3121	-1.7
PLAU	1C5W	Amiloride	16231	-7.4

evidence that the top SNP rs149123117, located in BTN3A1, is a protective factor against SS, possibly linked to the up-regulation of cg22068371 methylation leading to increased BTN3A1 protein levels.

Plasminogen activator urokinase (PLAU) is a protease involved in fibrinolysis, ECM remodeling, and growth factor activation (30). While most reports on PLAU have been associated with cancers such as breast, colorectal, and esophageal cancers, there is limited evidence of its association with SS. However, in our study, PLAU was found to be associated with an increased risk of SS in terms of gene expression and methylation level. Positive correlations were observed between the gene methylation of PLAU (cg04939496) and expression, as well as between expression and protein levels, supporting the promotional effects of PLAU on SS risk across different regulatory levels.

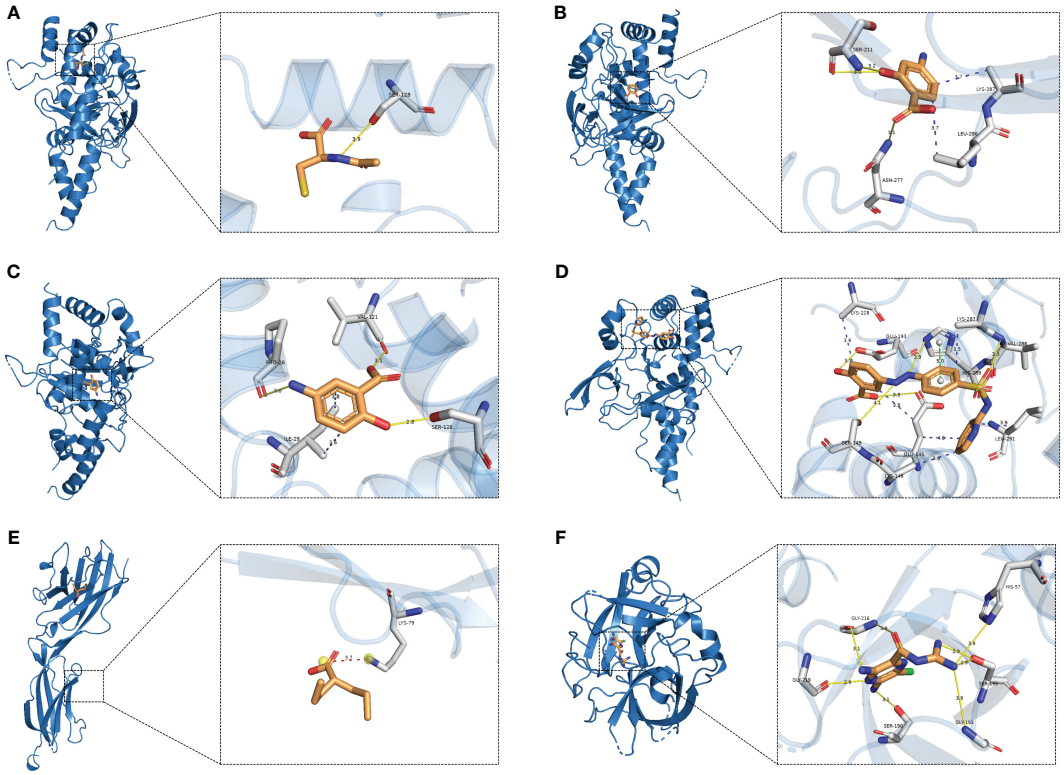


FIGURE 6
Molecular docking. **(A)** TNFAIP3-Acetylcysteine; **(B)** TNFAIP3-Aminosalicylic; **(C)** TNFAIP3-Mesalamine; **(D)** TNFAIP3-Sulfasalazine; **(E)** BTN3A1-Valproic acid; **(F)** PLAU-Amiloride.

Our study has some limitations: Firstly, it focused on the relationship between cis-mQTL, -eQTL, -pQTL, and SS, potentially overlooking other regulatory and environmental factors contributing to disease complexity. Although colocalization analysis was used to mitigate bias from linkage disequilibrium, horizontal pleiotropy may still persist. Additionally, the study predominantly involved individuals of European origin, necessitating further research and validation in individuals of other ethnicities for broader applicability. Furthermore, the eQTL dataset derived from blood may not fully capture tissue-specific regulatory mechanisms, warranting further tissue-specific validation. Though molecular docking predicted the interactions of potential drugs and targets, its feasibility may need to be validated by additional *in vitro* and *in vivo* experiments.

5 Conclusions

In conclusion, our study identifies TNFAIP3, BTN3A1, and PLAU as potential targets for SS by integrating the potential causal relationship of DNA methylation, gene expression, and protein abundance with SS. These findings provide important insights for targeted therapy of SS, although further experimental validation is required.

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 authors.

Author contributions

YB: Conceptualization, Data curation, Investigation, Methodology, Software, Writing – original draft. JW: Data curation, Methodology, Software, Writing – original draft. XF: Data curation, Methodology, Software, Writing – original draft. LX: Investigation, Writing – original draft. SQ: Supervision, Writing – review & editing, Conceptualization. GM: Funding acquisition, Supervision, Writing – review & editing. FZ: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This

research was funded by the National Natural Science Foundation of China (grant number 62171077).

Acknowledgments

We thank all the participants in this study.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1419363/full#supplementary-material

References

1. Stefanski AL, Tomiak C, Pleyer U, Dietrich T, Burmester GR, Dörner T. The diagnosis and treatment of Sjögren's syndrome. *Dtsch Arztebl Int*. (2017) 114(20):354–61. doi: 10.3238/arztebl.2017.0354

2. Stergiou IE, Poulaki A, Voulgarelis M. Pathogenetic mechanisms implicated in Sjögren's syndrome lymphomagenesis: a review of the literature. *J Clin Med*. (2020) 9 (12):3794. doi: 10.3390/jcm9123794

3. Nocturne G, Pontarini E, Bombardieri M, Mariette X. Lymphomas complicating primary Sjögren's syndrome: from autoimmunity to lymphoma. *Rheumatol (Oxford)*. (2021) 60(8):3513–21. doi: 10.1093/rheumatology/kez052

4. Brito-Zerón P, Retamozo S, Kostov B, Baldini C, Bootsma H, De Vita S, et al. Efficacy and safety of topical and systemic medications: a systematic literature review informing the EULAR recommendations for the management of Sjögren's syndrome. *RMD Open*. (2019) 5(2):e001064. doi: 10.1136/rmdopen-2019-001064

5. Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, et al. The support of human genetic evidence for approved drug indications. *Nat Genet*. (2015) 47(8):856–60. doi: 10.1038/ng.3314

6. Hingorani AD, Kuan V, Finan C, Kruger FA, Gaulton A, Chopade S, et al. Improving the odds of drug development success through human genomics: modelling study. *Sci Rep*. (2019) 9(1):18911. doi: 10.1038/s41598-019-54849-w

7. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. *Nat Rev Drug Discov*. (2017) 16 (1):19–34. doi: 10.1038/nrd.2016.230

8. Zheng J, Haberland V, Baird D, Walker V, Haycock PC, Hurle MR, et al. Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat Genet*. (2020) 52(10):1122–31. doi: 10.1038/s41588-020-0682-6

9. Chen J, Xu F, Ruan X, Sun J, Zhang Y, Zhang H, et al. Therapeutic targets for inflammatory bowel disease: proteome-wide Mendelian randomization and colocalization analyses. *EBioMedicine*. (2023) 89:104494. doi: 10.1016/j.ebiom.2023.104494

10. Lin J, Zhou J, Xu Y. Potential drug targets for multiple sclerosis identified through Mendelian randomization analysis. *Brain*. (2023) 146(8):3364–72. doi: 10.1093/brain/awad070

11. Sun J, Zhao J, Jiang F, Wang L, Xiao Q, Han F, et al. Identification of novel protein biomarkers and drug targets for colorectal cancer by integrating human plasma proteome with genome. *Genome Med*. (2023) 15(1):75. doi: 10.1186/s13073-023-01229-9

12. McRae AF, Marioni RE, Shah S, Yang J, Powell JE, Harris SE, et al. Identification of 55,000 replicated DNA methylation QTL. *Sci Rep*. (2018) 8(1):17605. doi: 10.1038/s41598-018-35871-w

13. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet*. (2021) 53(9):1300–10. doi: 10.1038/s41588-021-00913-z

14. Ferkingstad E, Sulem P, Atlason BA, Sveinbjornsson G, Magnusson MI, Styrmisdottir EL, et al. Large-scale integration of the plasma proteome with genetics and disease. *Nat Genet*. (2021) 53(12):1712–21. doi: 10.1038/s41588-021-00978-w

15. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet*. (2016) 48(5):481–7. doi: 10.1038/ng.3538

16. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res*. (2018) 46 (D1):D1074–82. doi: 10.1093/nar/gkx1037

17. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res*. (2021) 49(D1):D1388–95. doi: 10.1093/nar/gkaa971

18. Morris GM, Huey R, Olson AJ. Using autodock for ligand-receptor docking. *Curr Protoc Bioinf*. (2008) 24(1):8–14. doi: 10.1002/0471250953.bi0814s24

19. Momtazi G, Lambrecht BN, Naranjo JR, Schock BC. Regulators of A20 (TNFAIP3): new drug-able targets in inflammation. *Am J Physiol Lung Cell Mol Physiol*. (2019) 316(3):L456–l469. doi: 10.1152/ajplung.00335.2018

20. Khatri B, Tessneer KL, Rasmussen A, Aghakhanian F, Reksten TR, Adler A, et al. Genome-wide association study identifies Sjögren's risk loci with functional implications in immune and glandular cells. *Nat Commun*. (2022) 13(1):4287. doi: 10.1038/s41467-022-30773-y

21. Musone SL, Taylor KE, Nititham J, Chu C, Poon A, Liao W, et al. Sequencing of TNFAIP3 and association of variants with multiple autoimmune diseases. *Genes Immun*. (2011) 12(3):176–82. doi: 10.1038/gene.2010.64

22. Zhang M, Peng LL, Wang Y, Wang JS, Liu J, Liu MM, et al. Roles of A20 in autoimmune diseases. *Immunol Res*. (2016) 64(2):337–44. doi: 10.1007/s12026-015-8677-6

23. Johnsen SJ, Gudlaugsson E, Skaland I, Janssen EAM, Jonsson MV, Helgeland L, et al. Low protein A20 in minor salivary glands is associated with lymphoma in primary Sjögren's syndrome. *Scand J Immunol*. (2016) 83(3):181–7. doi: 10.1111/sji.12405

24. Gaballah H, Abd-Elkhalek R, Hussein O, Abd El-Wahab . Association of TNFAIP3 gene polymorphism (rs5029939) with susceptibility and clinical phenotype of systemic lupus erythematosus. *Arch Rheumatol*. (2021) 36(4):570–6. doi: 10.46497/ArchRheumatol.2022.8769

25. Dieudé P, Guedj M, Wipff J, Ruiz B, Riemekasten G, Matucci-Cerinic M, et al. Association of the TNFAIP3 rs5029939 variant with systemic sclerosis in the European Caucasian population. *Ann Rheum Dis*. (2010) 69(11):1958–64. doi: 10.1136/ard.2009.127928

26. Zhang MY, Yang XK, Pan HF, Ye DQ. Associations between TNFAIP3 gene polymorphisms and systemic lupus erythematosus risk: an updated meta-analysis. *Hla*. (2016) 88(5):245–52. doi: 10.1111/tan.12908

27. Vantourout P, Laing A, Woodward MJ, Zlatareva I, Apolonia L, Jones AW, et al. Heteromeric interactions regulate butyrophilin (BTN) and BTN-like molecules governing γδ T cell biology. *Proc Natl Acad Sci USA*. (2018) 115(5):1039–44. doi: 10.1073/pnas.1701237115

28. Tang YY, Xu WD, Fu L, Liu XY, Huang AF. Synergistic effects of BTN3A1, SHP2, CD274, and STAT3 gene polymorphisms on the risk of systemic lupus erythematosus: a multifactorial dimensional reduction analysis. *Clin Rheumatol*. (2024) 43(1):489–99. doi: 10.1007/s10067-023-06765-8

29. Xu WD, Yang C, Li R, Tang YY, Wang DC, Huang AF, et al. Association of BTN3A1 gene polymorphisms with systemic lupus erythematosus in a Chinese Han population. *Int J Rheum Dis*. (2024) 27(3):e15112. doi: 10.1111/1756-185X.15112

30. Li Z, Chen C, Wang J, Wei M, Liu G, Qin Y, et al. Overexpressed PLAU and its potential prognostic value in head and neck squamous cell carcinoma. *PeerJ*. (2021) 9: e10746. doi: 10.7717/peerj.10746

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 09 July 2024
DOI 10.3389/fimmu.2024.1377915

OPEN ACCESS

EDITED BY
Alex Tsoi,
University of Michigan, United States

REVIEWED BY
Shi Xue Dai,
Guangdong Provincial People's Hospital, China
Shuai Wang,
University of Texas Southwestern Medical Center, United States
Vincent Salvatore Gallicchio,
Clemson University, United States

*CORRESPONDENCE
Tie-mei Liu
ltm@jlu.edu.cn

RECEIVED 28 January 2024
ACCEPTED 25 June 2024
PUBLISHED 09 July 2024

CITATION
Li H-y and Liu T-m (2024) Platelet indices and inflammatory bowel disease: a Mendelian randomization study.
Front. Immunol. 15:1377915.
doi: 10.3389/fimmu.2024.1377915

Platelet indices and inflammatory bowel disease: a Mendelian randomization study

Hong-yang Li and Tie-mei Liu*

Department of Blood Transfusion, China-Japan Union Hospital of Jilin University, Changchun, China

Background: Platelets play a significant role in the innate and adaptive processes of immunity and inflammation. Inflammatory bowel disease (IBD) is an autoimmune disease that is widely understood to be caused by a combination of genetic predisposition, aberrant immune responses, etc.

Methods: To examine the relationships between genetically determined platelet indices and IBD, we conducted a Mendelian randomization (MR) study. Data associated with platelet count (PLT), mean platelet volume (MPV), platelet distribution width (PDW), plateletcrit (PCT) were used from the UK Biobank. The outcome data, including IBD, Crohn's disease (CD), ulcerative colitis (UC), were from the FinnGen database. The inverse variance-weighted (IVW), MR-Egger, weighted median methods were used for MR analyses.

Results: The MR estimations from the IVW approach show a significant connection between PLT and IBD. Similarly, PCT and IBD have a relationship following the IVW and MR-Egger approaches. While PLT and PCT have strong relationships with CD, according to the findings of all three approaches respectively. Nevertheless, PDW was the only relevant indicator of UC. The only significant result was IVW's.

Conclusion: Our findings suggest that the fluctuation of platelet indicators is of great significance in the development of IBD. PLT and PCT have a close association with IBD and CD, respectively; PDW only has a connection with UC. Platelets play an important role in the progression of IBD (UC, CD).

KEYWORDS

Mendelian randomization, platelet indices, inflammatory bowel disease, Crohn's disease, ulcerative colitis

Abbreviations: CD, Crohn's disease; CI, Confidence intervals; EIM, Extra intestinal manifestation; IBD, Inflammatory bowel disease; IVs, Instrumental variables; IVW, Inverse variance-weighted; GWAS, Genome-wide association studies; MKs, Megakaryocytes; MPV, Mean platelet volume; MR, Mendelian randomization; MR-PRESSO, MR pleiotropy residual sum and outlier; OR, Odds ratios; PCT, Platelet crit; PDW, Platelet distribution width; PLT, Platelet count; SNPs, Single-nucleotide polymorphisms; UC, Ulcerative colitis; WM, Weighted Median.

Introduction

Platelets are blood cells in plasma that are well recognized for their critical role in sustaining blood hemostasis (1). Megakaryocytes (MKs) create billions of them every day. MKs perceive and respond to inflammatory stress, and they engage in host immunological responses, according to emerging data (2). Platelet count (PLT), mean platelet volume (MPV), platelet width of distribution (PDW), and plateletcrit (PCT) are major platelet indicators in clinical practice that may be utilized to indicate platelet biochemical and functional changes (3). Platelets also play important roles in innate and adaptive immunity and inflammation, and they are the first blood cells to respond to wound-healing and tissue-repair mechanisms (1). Small platelets manage to maintain vascular integrity when faced with challenges of infection, sterile inflammation, and even malignancy, where they aid in hemostasis and serve as early responders to microbial threats (4). Because of their quick recruitment dynamics, these tiny, anucleate cell fragments are the first cells to form not just at sites of damage but also at sites of inflammation (5). Intravital imaging indicated that platelets are recruited and behave as individual cells rather than clots in the inflamed microvasculature, indicating that the hemostatic mechanism is unique to classical thrombosis and hemostasis. Unlike the well-defined processes of hemostasis following vascular trauma, inflammation-associated hemorrhage, also known as inflammatory bleeding, is a simplified summary of a phenomenon that occurs in a variety of disease settings, including sterile inflammation, microbial infection, and malignant tumors (6–8). Predilection sites include mucosal membranes, with epistaxis, gum bleeding, gastrointestinal bleeds, and hematuria being the most common bleeding episodes in thrombocytopenia patients. Platelet-mediated hemostasis without clot formation is critical to maintaining vascular integrity under these conditions (9, 10).

The autoimmune illness known as inflammatory bowel disease (IBD) is a chronic, relapsing condition that has caused significant health problems and is becoming more commonplace worldwide (11, 12). It is well accepted that genetic predisposition, environmental variables, and abnormal immune responses combine to cause IBD (13). The two main IBD subtypes, ulcerative colitis (UC) and Crohn's disease (CD), can differ significantly in terms of their molecular, immunological, morphological, and clinical features (14). Rectal bleeding, diarrhea, stomach discomfort, fever, anemia, and weight loss are some of the symptoms of UC (15, 16). CD may impact any region of the digestive tract in addition to causing diarrhea and abdominal pain (17). Up to 29.3 percent of IBD patients have at least one extra intestinal manifestation (EIM), which can have an effect on many systems, according to a Swiss cohort study (18). As per the European Crohn's and Colitis Organization, at least one EIM is experienced by up to 50% of people with IBD (19). Because of its great prevalence, IBD not only drastically lowers patients' quality of life but also places a major financial and medical burden on society (20), additionally accompanied by a number of issues or EIM (21). The most common areas of the body affected by the various types of EIMs are the musculoskeletal system, mucocutaneous system, ocular system, hepatobiliary tract, and oral cavity. There's a chance that other systems, including the pancreatic, pulmonary, cardiovascular, and urogenital systems are also at play (22, 23). Hematological EIMs haven't been thoroughly acknowledged or verified yet. Although the exact pathogenesis of EIMs is still unknown, it often involves dysregulated immunological responses, environmental factors, genetic vulnerability, and microbiota dysbiosis (19). Therefore, in order to obtain better prevention and control, it is essential to investigate the pathophysiology and risk factors of IBD. Determining causative relationships and possible risk factors for IBD represents an emerging public health concern.

A recent research by Vallet et al., which was published in the Journal of Clinical Investigation (24), demonstrates how the locations of megakaryocytes and the quality of platelet production alter with illness. Considering the vital role platelets play in coagulation, wound healing, tissue damage repair, immunological response, and inflammatory infections. Thus, assessments of platelet indices that reflect platelet bioactivity may be extremely important for tracking the onset, course, management, and prognosis of IBD.

In conclusion, it has not been established that platelet indices and IBD (UC and CD) are causally related. However, conventional observational study designs are limited in their ability to establish causality regarding the function of platelets in the development of IBD because of significant methodological constraints like reverse causation and residual confounding. A different strategy is the Mendelian randomization (MR) design, which makes use of genetic variations as instrumental variables (IVs) for an exposure in order to establish the causal relationship between the exposure and the outcome (25–28). By employing genetic variation as an indicator of causation, MR can remove the confounding bias seen in observational research. As alleles follow the principle of random assignment, different genotypes result in different intermediate phenotypes. If this phenotype represents an individual's exposure characteristic, the association effect between genotype and disease can describe the impact of exposure factors on illness. This effect is unaffected by confounding factors and reverse causal associations, as in traditional epidemiological studies (25, 29). The MR study concept is founded on Mendel's rule and functions similarly to a randomized controlled trial (RCT) but without the high expense (30).

In the current investigation, we employed a two-sample MR analysis to ascertain the association between platelet indices (PLT, MPV, PDW, and PCT) and IBD (UC and CD). It suggests that an IV-induced modifiable exposure caused the result. Therefore, we think the single-nucleotide polymorphisms (SNPs) used as research instruments had a modifying impact on the platelet indices, proving a positive causal relationship between the SNPs and the probability of developing IBD. However, interventions aimed at targeting the exposure are unlikely to be effective if there is a non-causal link between the exposure and the outcome.

Materials and methods

Study design

In order to investigate the associations between platelet indices (PLT, MPV, PDW, and PCT) and IBD (UC and CD), we used a

two-sample MR design. Strengthening the Reporting of Observational Studies in Epidemiology Using Mendelian Randomization and the Fundamentals of MR were adhered to in the design of our study (31). Additionally, these selections underwent an MR analysis and satisfied three fundamental presumptions (Figure 1): Three things are relevant about the instrumental variables: (1) they are directly correlated with the exposure; (2) they are unaffected by confounders; and (3) genetic variations only influence outcomes through exposure (32). The purpose of the univariable MR study was to explore the relationship between platelet indices (PLT, MPV, PDW, and PCT) and IBD (UC and CD). The research design employed is shown in Figure 2.

Data source

The genetic tools for the four platelet indices (PLT, MPV, PDW, and PCT) were chosen from a genome-wide association study (GWAS) that involved 408,112 participants in the UK Biobank (33). Every participant was descended from Europeans. Data from the FinnGen collaboration, which became publicly available in May 2021, was utilized to determine the outcomes. Which enrolled 218,792 European participants (cases/controls for IBD: 5,673/213,119; CD: 940/217,852); and 218,507 participants (cases/controls for UC: 2,701/215,806) (34). Since 2017, FinnGen has been a large-scale national effort that aims to improve human medicine by gathering genetic data and health record information from Finnish health registries and Biobanks, respectively. The detailed information on all traits involved was summarized in Table 1. Since all of the data are GWAS summary statistics that are available to the public, no further ethical approval or informed permission was needed.

Selection of instrumental variables

IVs were chosen as independent SNPs at genomewide significance ($P<5\times10^{-8}$) for every exposure taken into account in univariable MR analysis. Pairwise linkage disequilibrium, or independent SNPs, were found using criteria of ($r^2<0.001$, clumping window=10,000 kb). To find and eliminate outlier instruments, MR pleiotropy residual sum and outlier (MR-PRESSO) analyses were carried out. The cumulative strength of the chosen SNPs was assessed using the F-statistic ($F = beta^2/se^2$), where beta denoted the exposure's effect value and se denoted the exposure's standard error. This helped to prevent weak instrument bias. F>10 is required to access the whole SNPs collection (35). The F-statistics used in the univariable MR analyses are provided in Supplementary Table 1.

Statistical analysis

Reverse causation can lead to an incorrect inference that the exposure and the outcome are causally connected if variations in the outcomes that exhibit greater relationships with outcomes than with exposures are employed in the MR analyses (36). Consequently, we must exclude the SNPs that have an outcome of $P<5\times10^{-8}$. And then, prior to analysis, we first harmonized exposure and outcome data to make alignments on effect alleles to the forward strand, if it is specified or could be inferred based on the allele frequency. Ambiguous SNPs with non-concordant alleles and palindromic SNPs that may create uncertainty regarding the identification of the effect allele in the exposure and outcome GWASs were excluded for further MR analyses (37, 38). After identifying the IV sets using the aforementioned selection criteria,

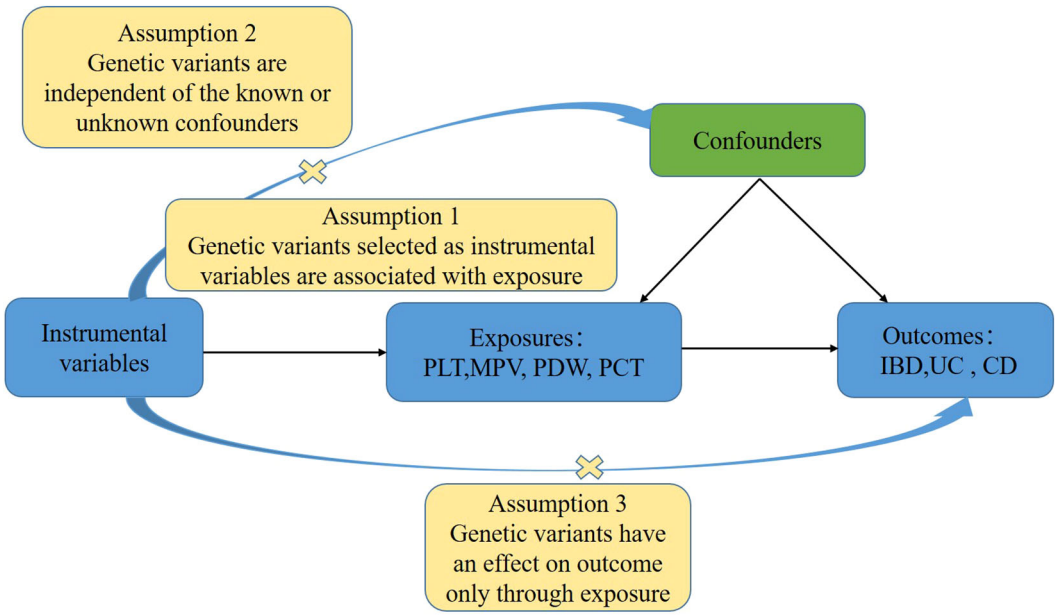


FIGURE 1
The basic principles of the MR study show the three principal assumptions.

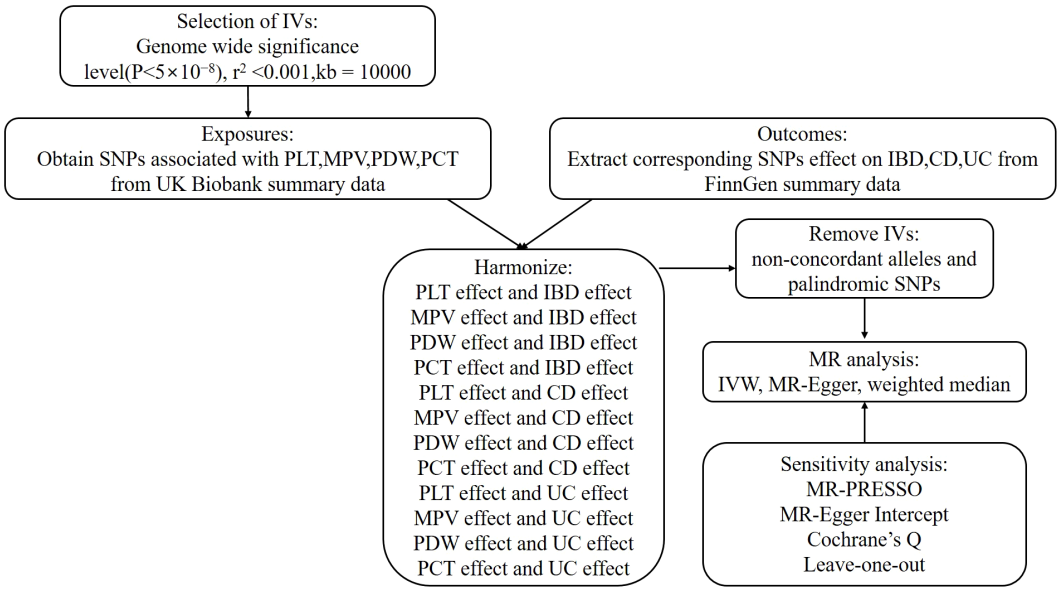


FIGURE 2
An overview of the study design.

we estimated the total effects using MR analysis. We performed a significance analysis using the IVW approach. Assuming that all SNPs are legitimate instrumental factors, this technique yields the maximum power estimate. When all IVs are genuine and horizontal pleiotropy is balanced, this method yields unbiased estimates of causal links even in the presence of variability across SNPs (39). The MR-Egger regression was used in secondary analyses to account for pleiotropy and assess the findings' robustness. Although its power is limited, the MR-Egger method can identify and rectify directional pleiotropy. Even in the event that the second and third assumptions are false, it accounts for the directed pleiotropic effects of genetic instruments (40). The MR-Egger test produces a consistent causal estimate and a valid test of the null causal hypothesis, even in the case when all genetic variations are invalid (40). Nevertheless, MR-Egger shows poor statistical accuracy and is vulnerable to outlying genetic variations (41). The weighted median approach is the third method. It is substantially and continuously more accurate than the MR Egger approach and more resilient to violations of causal effects (42). It is predicated on the supposition that more than half of the IVs are believable. Furthermore, outliers and high-leverage genetic variants won't have an impact on it (42). Otherwise, the IVW outcomes took precedence. The OR and accompanying 95% CI on the outcome risk of corresponding unit changes in exposure were used to represent the MR results. To evaluate the relative risk brought on by the existence of the illness of interest, the OR and 95% CI were shown. P<0.05 was used to indicate statistical significance in the univariable MR analysis for the findings of sensitivity analyses on the causal effects of exposures and outcomes. To depict the MR data, scatterplots, forest plots, and funnel plots were created in the interim.

We also assessed horizontal pleiotropy for significant estimates using the intercept tests of MR-Egger regression and MR-PRESSO. MR-Egger regression yielded an intercept, and intercept values that differ from zero indicate pleiotropy (here assessed using a p-value <0.05), which was suggestive of an overall directional pleiotropy (43). Using the global and SNP-specific observed residual sum of

TABLE 1 Detail of the data for the cohort population.

Trait	Gwas ID	Data source	Sample size	Case/ control	Number of SNPs	Population	Year
IBD	finn-b-K11_IBD	FinnGen	218792	5,673/213,119	16,380,466	European	2021
CD	finn-b-K11_KELACROHN	FinnGen	218792	940/217,852	16,380,466	European	2021
UC	finn-b-K11_UC_STRICT	FinnGen	218507	2,701/215,806	16,380,466	European	2021
PLT	ebi-a-GCST90002402	UK Biobank	408,112	/	40,299,783	European	2020
MPV	ebi-a-GCST90002395	UK Biobank	408,112	/	40,299,375	European	2020
PDW	ebi-a-GCST90002401	UK Biobank	408,112	/	40,300,122	European	2020
PCT	ebi-a-GCST90002400	UK Biobank	408,112	/	40,299,196	European	2020

squares, the MR-PRESSO method screened for general horizontal pleiotropy (global test) and outliers (outlier test), with a significant threshold of 0.05 (44). Additionally, after eliminating outliers, it provided causal estimates and contrasted the raw values with the distortion. Additionally, 10,000 distribution points were allocated. By gradually eliminating each IV, leave-one-out analysis was performed in order to identify bias caused by a heterogeneous variation. In order to identify heterogeneity ($p<0.05$ shows heterogeneity), we also calculated the Cochrane's Q value, which allowed us to identify the existence of pleiotropy (45). Each SNP's heterogeneity in terms of causative effects was assessed using Cochran's Q value (46). For the second and third assumptions to be satisfied, horizontal pleiotropy must be assessed (38). R statistical program (version 4.3.1, R Foundation for Statistical Computing, Vienna, Austria, 2023; https://www.R-project.org) was used for all statistical analyses, together with the Two-Sample MR and MR-PRESSO Packages (38).

Results

Selection of instrumental variables

Altogether, 477 index SNPs were shown to be possible genetic IVs for IBD, 482 SNPs for CD, and 479 SNPs for UC when PLT was taken into account as an exposure factor. In the presence of MPV as an exposure factor, 453 index SNPs were shown to be putative genetic IVs for IBD, 455 SNPs for CD, and 454 SNPs for UC, in that order. PDW as an exposure factor led to the identification of 379 index SNPs as putative genetic IVs for IBD, 378 SNPs for CD, and 375 SNPs for UC, in that order. In the case of PCT as an exposure factor, possible genetic IVs for IBD, CD, and UC were found to be 452 index SNPs, 454 SNPs, and 453 SNPs, respectively. Not only have all of these SNPs been harmonized and palindromic SNPs with intermediate allele frequencies removed, but they have also undergone the MR-PRESSO test, which was run in order to identify and eliminate outlier IVs. Once the outlier IVs were eliminated, MR estimations were reexamined. Thus, the SNPs listed above were taken into account for the MR analysis. Furthermore, each SNP's F-value was greater than 10, which suggests that there is a minimal possibility of weak instrumental variable bias.

Mendelian randomization analysis

Overall, there was inconsistency in the results from the three approaches used to establish a causal relationship between platelet indicators (PLT, MPV, PDW, and PCT) and IBD (UC and CD). According to the IVW method's MR estimations, there is a significant correlation between PLT and IBD (OR:1.11, 95% CI:1.02 to 1.21, P:0.013). However, IBD was not associated with the findings of the MR-Egger or weighted median techniques (OR:1.14, 1.11,95%CI:0.99 to 1.32,0.98 to 1.27, P:0.079,0.095), respectively. Likewise, there is a close link between PCT and IBD. IVW produced the following results: OR:1.10, 95%CI:1.01 to 1.20, P:0.034. OR:1.19, 95%CI:1.02 to 1.39, P:0.023 was the MR-Egger. However, there was no significant difference using the weighted median approach (OR:1.10, 95%CI:0.95 to 1.28, P:0.2). PLT and PCT were related to CD, whereas PDW was connected to UC, according to further study of the two subtypes. IVW (OR:1.35, 95% CI:1.15 to 1.59, P:0.0003), MR-Egger (OR:1.43, 95%CI:1.07 to 1.90, P:0.015), and weighted median (OR:1.41, 95%CI:1.06 to 1.86, P:0.017) were the values obtained from PLT to CD. PLT and CD have strong relationships, according to the findings of all three approaches. A comparison between PCT and CD revealed similarities in the IVW (OR:1.27, 95%CI:1.06 to 1.52, P:0.011), MR-Egger (OR:1.89, 95%CI:1.38 to 2.59, $P:9.3\times10^{-5}$), and weighted median (OR:1.36, 95%CI:1.01 to 1.85, P:0.046). PCT was closely associated to CD, according to the findings of all three methodologies. However, the only elevated factor with regard to UC was PDW. And only IVW's finding (OR:1.14, 95%CI:1.01 to 1.29, P:0.032) was remarkable. We found no relationship between other platelet indices and IBD, CD, and UC; the detailed results and scatterplots are listed in Figures 3, 4. And the forest plots and funnel plots are shown in Supplementary Figures 1, 2.

Sensitivity analysis

While some of the Cochran Q test findings showed heterogeneity, the major outcome of the random effects IVW analysis allowed for some heterogeneity. All except one of the p-values for the MR-Egger intercept were greater than 0.05. The results and details are provided in Supplementary Table 2. Furthermore, our results' robustness was further validated by the fact that leave-one-out analysis failed to find any outlier IVs (Supplementary Figure 3). Additionally, following the global MR-PRESSO testing, we had to exclude a few SNPs. However, there were all significant SNPs after removing the outliers. The MR-PRESSO distortion test results showed the causal effect of genetically predicted platelet indices on IBD (CD, UC) after correction by removing outliers. On the other hand, genetically predicted platelet indices were shown to raise the risk of IBD (CD, UC) in both corrected and uncorrected data (Table 2).

Discussion

This is the first MR research that we are aware of that examines the relationship between platelet indices and IBD (UC and CD). The purpose of the current study was to investigate the relationship between IVs of the four platelet indices and IBD (CD, UC). We discovered in the univariable MR that a rise in IBD and CD was correlated with the amounts of PLT and PCT predicted by the provided genetics, while PDW was linked to UC. But there was no significant correlation between other platelet indicators and IBD (CD, UC). According to these results, PLT and PCT are the essential characteristics that generate favorable correlations between IBD and CD. PDW may only relevant to UC.

A two-sample MR analysis of the relationship between platelet indices and IBD was conducted for this investigation. There was

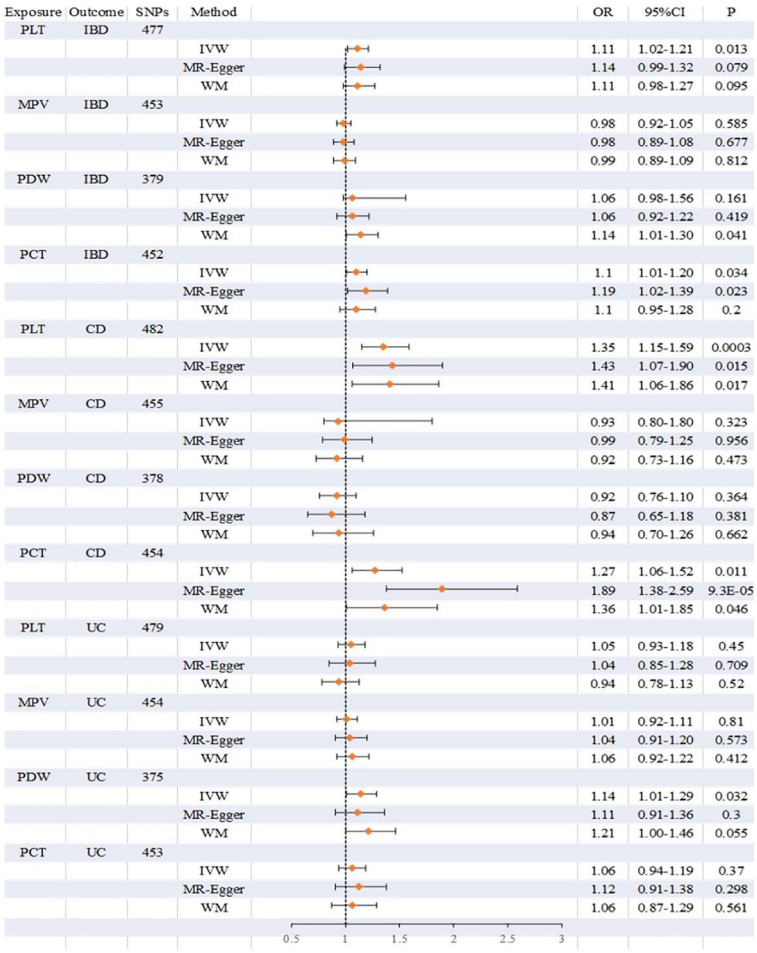


FIGURE 3
Detailed results on the association between platelet indices (PLT, MPV, PDW, PCT) and IBD, CD and UC.

shown to be a strong relationship between platelet indices and IBD. In order to better understand the association between platelet indices and IBD and to develop therapies for the disease, a greater study of the correlations between various platelet indices and IBD utilizing bigger and more diverse data sources is necessary. However, although they are categorized as IBD, CD, and UC, they are not the same in terms of pathophysiology, symptoms, complications, natural courses, and sequelae. In addition to severely impairing a patient's quality of life, CD and UC both increase mortality and financial burden (12, 14, 47). Although the exact cause of IBD (CD, UC) is still unclear, genetic vulnerability, environmental factors, and the gut microbiome may all be significant (48). Further evidence of these two distinct situations was found in our research.

As is well known, PLT counts the number of platelets per unit volume of blood, PCT represents the proportion of blood volume occupied by platelets, and MPV indicates the average size of platelets. As a result, PCT is connected with the products of MPV and PLT, and may be thought of as a sort of analog of the total platelet volume. PDW, in comparison to PLT, PCT and MPV, is another significant metric. Thus, elevated indices may suggest that platelets play a part in understanding the IBD process (CD, UC). In our study, we have found there is a relevance between PLT, PDW, and PCT with IBD (CD, UC), so the platelet indices reflect this phenomenon and may be useful indicators for assessing the course of IBD (CD, UC). In clinical practice, it is important to highlight the independent and prominent roles that PLT, PCT, and PDW play among the four platelet indices.

Excessive clotting or unusual bleeding are the outcomes of elevated platelet levels (49). Because of the close involvement of their membrane receptors at different stages of the blood-coagulation cascade (50), a sequence of biochemical reactions that take place in the body in response to injury or damage to blood vessels, platelets play a critical role as the defenders of the integrity of the blood vasculature. The exterior membrane of platelets is extremely active and functional, expressing different integrin, glycoproteins, and antigens (1). These membrane constituents play a crucial role in coordinating the intricate interplay between platelets and sub endothelial structures that are exposed due to blood vessel wall damage. Additionally, proteins that make up fibrin clots and plasma coagulation factors and activators interact with biomolecules produced on platelet membranes. Membrane glycoproteins identify blood clotting factors and play a key role in platelet adherence and activation. Platelet membranes strongly

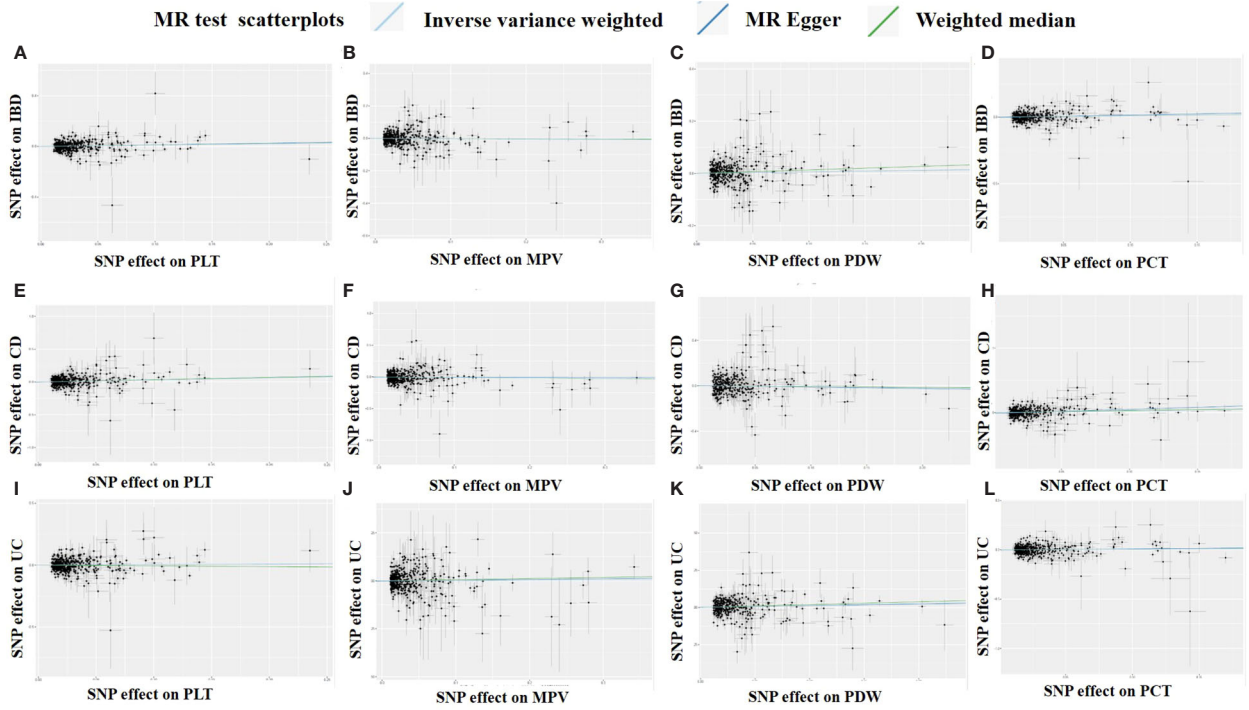


FIGURE 4
Scatter plots of the MR analysis. **(A)** PLT on IBD. **(B)** MPV on IBD. **(C)** PDW on IBD. **(D)** PCT on IBD. **(E)** PLT on CD. **(F)** MPV on CD. **(G)** PDW on CD. **(H)** PCT on CD. **(I)** PLT on UC. **(J)** MPV on UC. **(K)** PDW on UC. **(L)** PCT on UC.

express GPIIb/IIIa, GPIb-IX-V, GPVI, and P2Y12, all of which are essential in the hemostatic process that comes before the wound-healing phase (51). The immunological response of the body is improved by platelets. It has been demonstrated that platelet-derived CD40L may stimulate monocyte differentiation into dendritic cells (DC), DC maturation, and co-stimulatory molecule upregulation (52). This role of platelet-derived CD40L may be particularly important for autoimmune illnesses like systemic lupus erythematosus, where platelets stimulate B-cell secretion of antibodies via inducing DC differentiation and type-I interferon release (53). But IBD is an autoimmune disease that recurs frequently, causing intestinal bleeding, inflammatory responses,

TABLE 2 The MR-PRESSO test's results.

Exposure	Outcome	Raw			Outlier corrected			Global P	Number of outliers	Distortion P
		OR	95%CI	P	OR	95%CI	P			
PLT	IBD	1.10	1.01–1.20	0.031	1.11	1.02–1.21	0.009	<1e-04	5	0.787
MPV		0.99	0.92–1.05	0.640	0.98	0.92–1.05	0.575	<1e-04	2	0.947
PDW		1.06	0.98–1.56	0.216	-	-	-	<1e-04	NA	NA
PCT		1.10	1.00–1.20	0.116	1.10	1.01–1.20	0.058	<1e-04	1	0.822
PLT	CD	1.35	1.15–1.59	0.0003	-	-	-	0.502	NA	NA
MPV		0.93	0.80–1.80	0.324	-	-	-	7e-04	NA	NA
PDW		0.92	0.76–1.11	0.371	0.92	0.76–1.10	0.455	0.006	1	0.810
PCT		1.27	1.06–1.52	0.011	1.27	1.06–1.52	0.008	0.031	1	0.945
PLT	UC	1.05	0.93–1.18	0.432	1.05	0.93–1.18	0.436	<1e-04	3	0.978
MPV		1.01	0.92–1.11	0.829	1.01	0.92–1.11	0.699	<1e-04	1	0.908
PDW		1.14	1.01–1.29	0.032	1.14	1.01–1.29	0.016	<1e-04	4	0.910
PCT		1.06	0.94–1.19	0.374	1.06	0.94–1.19	0.430	<1e-04	1	0.909

and EIMs such as cardiovascular problems. Furthermore, the precise aspects of its pathophysiology are yet unknown, but they appear to be linked to immune response problems and genetic predisposition. So combining the function of platelets and the MR results we obtained, platelet-related indices are indeed closely related to IBD and predict its occurrence and development.

We discovered the link between platelet indices and IBD (UC and CD), as previously mentioned. However, three presumptions relevance, independence, and exclusion-restriction are necessary for IVs to be valid in MR. The second and third assumptions, however, are dependent on every potential confounding factor of the exposure-outcome connection, both measurable and unmeasured, and only the first can be completely empirically evaluated. To provide a consistent estimate of the causative effect, all genetic variations included in the research as IVs must meet the MR assumptions for the IVW method (42). Both the weighted median and the MR-Egger methods were used to verify this. Even in cases where all genetic effects are null due to violations of the third assumption mentioned above, the MR-Egger approach reliably predicts the genuine causal impact under a lesser assumption (54). However, if all genetic variants have a comparable degree of connection with the exposure, then MR-Egger regression estimates become less accurate. On the other hand, if no single genetic variation accounts for more than 50% of the weight, the weighted median approach will yield a consistent estimate only if at least 50% of the weight originates from legitimate genetic variants. When it comes to faulty genetic variations, the weighted median method permits a more widespread violation of the MR assumptions than the MR-Egger method does (42). Therefore, we think that the remaining results suggest a causal relationship between platelet indices and IBD, even if an MR-Egger technique observation yielded a non-significant estimate.

Although we have identified a relationship between platelet indices and IBD through the MR study. There were a few more restrictions on this study. Firstly, it is probable that the putative gender-specific effects on the relationship were overlooked since we did not separate platelet indices and IBD (UC and CD) by gender. The UK Biobank sample was used for the GWAS of characteristics linked to platelet indices, while FinnGen provided data on IBD (CD, UC). As a result, bias and sample overlap are possible in relation to this fact (55). Furthermore, even though steps have been taken to identify and eliminate outlier SNPs, we cannot totally rule out the possibility that heterogeneity will have an impact on the results. Moreover, our work has demonstrated a causal association between platelet indices and IBD (UC and CD); nevertheless, additional research is necessary as the specific underlying processes are still unclear. Then, even with an MR research design, confounding cannot be totally minimized because the risk factors for IBD (CD, UC) comprise not just genetic variables but also other factors, such as environmental ones. Finally, the study only contained four platelet indices; more hematological indicators associated with platelets may exist, meaning that the relative importance of PLT, PCT, and PDW may need to be adjusted when considering other features.

Conclusions

Evidence supporting PLT, PCT, and PDW as distinct and predominant features explaining the relationship to IBD (CD, UC) may be found in the current MR investigation. Comprehending the function of platelets and their associated characteristics is beneficial for both public and clinical health. To strengthen the case for antiplatelet medication as the main preventive measure in IBD patients, stratified randomized controlled trials are also required. Our MR investigation showed that PLT and PCT had a connection to IBD and CD meanwhile that PDW had a relation to UC. To a certain extent, platelets and their associated characteristics influence the development of IBD (UC, CD). A possible preventative method for IBD might involve focusing on these characteristics. Further research is required to determine the precise mechanism and validate the therapeutic benefits of this kind of preventive therapy.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements. Written informed consent was not obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article because online public data does not require informed consent.

Author contributions

HL: Data curation, Formal analysis, Investigation, Software, Writing – original draft, Writing – review & editing. TL: Conceptualization, Funding acquisition, Writing – review & editing.

Funding

The authors declare financial support was received for the research, and/or publication of this article. This work was supported by the grants from National Key Research and Development Project of China (No. 2023YFC2413100).

Acknowledgments

The authors acknowledge and thank the investigators of the original GWAS studies for sharing the summary data used in this study.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1377915/full#supplementary-material

References

1. Burnouf T, Chou ML, Lundy DJ, Chuang EY, Tseng CL, Goubran H. Expanding applications of allogeneic platelets, platelet lysates, and platelet extracellular vesicles in cell therapy, regenerative medicine, and targeted drug delivery. *J BioMed Sci*. (2023) 30:79. doi: 10.1186/s12929-023-00972-w

2. Khatib-Massalha E, Méndez-Ferrer S. Megakaryocyte diversity in ontogeny, functions and cell-cell interactions. *Front Oncol*. (2022) 12:840044. doi: 10.3389/fonc.2022.840044

3. Pogorzelska K, Krętowska A, Krawczuk-Rybak M, Sawicka-Żukowska M. Characteristics of platelet indices and their prognostic significance in selected medical condition - a systematic review. *Adv Med Sci*. (2020) 65:310–5. doi: 10.1016/j.advms.2020.05.002

4. Kaiser R, Escaig R, Nicolai L. Hemostasis without clot formation: how platelets guard the vasculature in inflammation, infection, and Malignancy. *Blood*. (2023) 142:1413–25. doi: 10.1182/blood.2023020535

5. Nicolai L, Massberg S. Platelets as key players in inflammation and infection. *Curr Opin Hematol*. (2020) 27:34–40. doi: 10.1097/MOH.0000000000000551

6. Ho-Tin-Noé B, Boulaftali Y, Camerer E. Platelets and vascular integrity: how platelets prevent bleeding in inflammation. *Blood*. (2018) 131:277–88. doi: 10.1182/blood-2017-06-742676

7. Ho-Tin-Noé B, Demers M, Wagner DD. How platelets safeguard vascular integrity. *J Thromb Haemost*. (2011) 9 Suppl 1:56–65. doi: 10.1111/j.1538-7836.2011.04317.x

8. Ho-Tin-Noé B, Goerge T, Wagner DD. Platelets: guardians of tumor vasculature. *Cancer Res*. (2009) 69:5623–6. doi: 10.1158/0008-5472.CAN-09-1370

9. Nachman RL, Rafii S. Platelets, petechiae, and preservation of the vascular wall. *N Engl J Med*. (2008) 359:1261–70. doi: 10.1056/NEJMra0800887

10. Wéra O, Lecut C, Servais L, Hego A, Delierneux C, Jiang Z, et al. P2X1 ion channel deficiency causes massive bleeding in inflamed intestine and increases thrombosis. *J Thromb Haemost*. (2020) 18:44–56. doi: 10.1111/jth.14620

11. Ng SC, Shi HY, Hamidi N, Underwood FE, Tang W, Benchimol EI, et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *Lancet*. (2017) 390:2769–78. doi: 10.1016/S0140-6736(17)32448-0

12. Kaplan GG. The global burden of IBD: from 2015 to 2025. *Nat Rev Gastroenterol Hepatol*. (2015) 12:720–7. doi: 10.1038/nrgastro.2015.150

13. Ramos GP, Papadakis KA. Mechanisms of disease: inflammatory bowel diseases. *Mayo Clin Proc*. (2019) 94:155–65. doi: 10.1016/j.mayocp.2018.09.013

14. Roda G, Chien Ng S, Kotze PG, Argollo M, Panaccione R, Spinelli A, et al. Crohn's disease. *Nat Rev Dis Primers*. (2020) 6:22. doi: 10.1038/s41572-020-0156-2

15. Bu F, Ding Y, Chen T, Wang Q, Wang R, Zhou JY, et al. Total flavone of Abelmoschus Manihot improves colitis by promoting the growth of Akkermansia in mice. *Sci Rep*. (2021) 11:20787. doi: 10.1038/s41598-021-00070-7

16. Kim HY, Cheon JH, Lee SH, Min JY, Back SY, Song JG, et al. Ternary nanocomposite carriers based on organic clay-lipid vesicles as an effective colon-targeted drug delivery system: preparation and in *vitro*/*in vivo* characterization. *J Nanobiotechnology*. (2020) 18:17. doi: 10.1186/s12951-020-0579-7

17. Levison SE, Fisher P, Hankinson J, Zeef L, Eyre S, Ollier WE, et al. Genetic analysis of the Trichuris muris-induced model of colitis reveals QTL overlap and a novel gene cluster for establishing colonic inflammation. *BMC Genomics*. (2013) 14:127. doi: 10.1186/1471-2164-14-127

18. Vavricka SR, Rogler G, Gantenbein C, Spoerri M, Prinz Vavricka M, Navarini AA, et al. Chronological order of appearance of extraintestinal manifestations relative to the time of IBD diagnosis in the swiss inflammatory bowel disease cohort. *Inflammation Bowel Dis*. (2015) 21:1794–800. doi: 10.1097/MIB.0000000000000429

19. Hedin CRH, Vavricka SR, Stagg AJ, Schoepfer A, Raine T, Puig L, et al. The pathogenesis of extraintestinal manifestations: implications for IBD research, diagnosis, and therapy. *J Crohns Colitis*. (2019) 13:541–54. doi: 10.1093/ecco-jcc/jjy191

20. Alatab S, Sepanlou SG, Ikuta K, Vahedi H, Bisignano C, Safiri S, et al. The global, regional, and national burden of inflammatory bowel disease in 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet Gastroenterol Hepatol*. (2020) 5:17–30. doi: 10.1016/s2468-1253(19)30333-4

21. Marotto D, Atzeni F, Ardizzone S, Monteleone G, Giorgi V, Sarzi-Puttini P. Extra-intestinal manifestations of inflammatory bowel diseases. *Pharmacol Res*. (2020) 161:105206. doi: 10.1016/j.phrs.2020.105206

22. Harbord M, Annese V, Vavricka SR, Allez M, Barreiro-de Acosta M, Boberg KM, et al. The first European evidence-based consensus on extra-intestinal manifestations in inflammatory bowel disease. *J Crohns Colitis*. (2016) 10:239–54. doi: 10.1093/ecco-jcc/jjv213

23. Rogler G, Singh A, Kavanaugh A, Rubin DT. Extraintestinal manifestations of inflammatory bowel disease: current concepts, treatment, and implications for disease management. *Gastroenterology*. (2021) 161:1118–32. doi: 10.1053/j.gastro.2021.07.042

24. Valet C, Magnen M, Qiu L, Cleary SJ, Wang KM, Ranucci S, et al. Sepsis promotes splenic production of a protective platelet pool with high CD40 ligand expression. *J Clin Invest*. (2022) 132:e153920. doi: 10.1172/JCI153920

25. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int J Epidemiol*. (2003) 32:1–22. doi: 10.1093/ije/dyg070

26. Sekula P, Del Greco MF, Pattaro C, Köttgen A. Mendelian randomization as an approach to assess causality using observational data. *J Am Soc Nephrol*. (2016) 27:3253–65. doi: 10.1681/ASN.2016010098

27. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum Mol Genet*. (2014) 23:R89–98. doi: 10.1093/hmg/ddu328

28. Lawlor DA, Harbord RM, Sterne JA, Timpson N, Davey Smith G. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Stat Med*. (2008) 27:1133–63. doi: 10.1002/sim.3034

29. Burgess S, Butterworth A, Thompson SG. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet Epidemiol*. (2013) 37:658–65. doi: 10.1002/gepi.21758

30. Zhu X. Mendelian randomization and pleiotropy analysis. *Quant Biol*. (2021) 9:122–32. doi: 10.1007/s40484-020-0216-3

31. Skrivankova VW, Richmond RC, Woolf BAR, Davies NM, Swanson SA, VanderWeele TJ, et al. Strengthening the reporting of observational studies in epidemiology using mendelian randomisation (STROBE-MR): explanation and elaboration. *BMJ*. (2021) 375:n2233. doi: 10.1136/bmj.n2233

32. Davies NM, Holmes MV, Davey Smith G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ*. (2018) 362:k601. doi: 10.1136/bmj.k601

33. Vuckovic D, Bao EL, Akbari P, Lareau CA, Mousas A, Jiang T, et al. The polygenic and monogenic basis of blood traits and diseases. *Cell*. (2020) 182:1214–1231.e1211. doi: 10.1016/j.cell.2020.08.008

34. Kurki MI, Karjalainen J, Palta P, Sipilä TP, Kristiansson K, Donner KM, et al. FinnGen provides genetic insights from a well-phenotyped isolated population. *Nature*. (2023) 613:508–18. doi: 10.1038/s41586-022-05473-8

35. Burgess S, Thompson SG. Avoiding bias from weak instruments in Mendelian randomization studies. *Int J Epidemiol*. (2011) 40:755–64. doi: 10.1093/ije/dyr036

36. Luo J, Xu Z, Noordam R, van Heemst D, Li-Gao R. Depression and inflammatory bowel disease: A bidirectional two-sample mendelian randomization study. *J Crohns Colitis*. (2022) 16:633–42. doi: 10.1093/ecco-jcc/jjab191

37. Hartwig FP, Davies NM, Hemani G, Davey Smith G. Two-sample Mendelian randomization: avoiding the downsides of a powerful, widely applicable but potentially fallible technique. *Int J Epidemiol*. (2016) 45:1717–26. doi: 10.1093/ije/dyx028

38. Hemani G, Zheng J, Elsworth B, Wade KH, Haberland V, Baird D, et al. The MR-Base platform supports systematic causal inference across the human phenome. *ELife*. (2018) 7:e34408. doi: 10.7554/eLife.34408

39. Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan N, Thompson J. A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat Med*. (2017) 36:1783–802. doi: 10.1002/sim.7221

40. Bowden J, Davey Smith G, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*. (2015) 44:512–25. doi: 10.1093/ije/dyv080

41. Bowden J, Del Greco MF, Minelli C, Davey Smith G, Sheehan NA, Thompson JR. Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *Int J Epidemiol*. (2016) 45:1961–74. doi: 10.1093/ije/dyw220

42. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol*. (2016) 40:304–14. doi: 10.1002/gepi.21965

43. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol*. (2017) 32:377–89. doi: 10.1007/s10654-017-0255-x

44. Verbanck M, Chen CY, Neale B, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet*. (2018) 50:693–8. doi: 10.1038/s41588-018-0099-7

45. Greco MF, Minelli C, Sheehan NA, Thompson JR. Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Stat Med*. (2015) 34:2926–40. doi: 10.1002/sim.6522

46. Bowden J, Del Greco MF, Minelli C, Zhao Q, Lawlor DA, Sheehan NA, et al. Improving the accuracy of two-sample summary-data Mendelian randomization: moving beyond the NOME assumption. *Int J Epidemiol*. (2019) 48:728–42. doi: 10.1093/ije/dyy258

47. Kobayashi T, Siegmund B, Le Berre C, Wei SC, Ferrante M, Shen B, et al. Ulcerative colitis. *Nat Rev Dis Primers*. (2020) 6:74. doi: 10.1038/s41572-020-0205-x

48. Ananthakrishnan AN, Bernstein CN, Iliopoulos D, Macpherson A, Neurath MF, Ali RAR, et al. Environmental triggers in IBD: a review of progress and evidence. *Nat Rev Gastroenterol Hepatol*. (2018) 15:39–49. doi: 10.1038/nrgastro.2017.136

49. Gregg D, Goldschmidt-Clermont PJ. Cardiology patient page. Platelets and cardiovascular disease. *Circulation*. (2003) 108:e88–90. doi: 10.1161/01.CIR.0000086897.15588.4B

50. Schenone M, Furie BC, Furie B. The blood coagulation cascade. *Curr Opin Hematol*. (2004) 11:272–7. doi: 10.1097/01.moh.0000130308.37353.d4

51. Gremmel T, Frelinger AL3rd, Michelson AD. Platelet physiology. *Semin Thromb Hemost*. (2016) 42:191–204. doi: 10.1055/s-00000077

52. Kaneider NC, Kaser A, Tilg H, Ricevuti G, Wiedermann CJ. CD40 ligand-dependent maturation of human monocyte-derived dendritic cells by activated platelets. *Int J Immunopathol Pharmacol*. (2003) 16:225–31. doi: 10.1177/039463200301600307

53. Duffau P, Seneschal J, Nicco C, Richez C, Lazaro E, Douchet I, et al. Platelet CD154 potentiates interferon-alpha secretion by plasmacytoid dendritic cells in systemic lupus erythematosus. *Sci Transl Med*. (2010) 2:47ra63. doi: 10.1126/scitranslmed.3001001

54. Liu Z, Ye T, Sun B, Schooling M, Tchetgen ET. Mendelian randomization mixed-scale treatment effect robust identification and estimation for causal inference. *Biometrics*. (2023) 79:2208–19. doi: 10.1111/biom.13735

55. Burgess S, Davies NM, Thompson SG. Bias due to participant overlap in two-sample Mendelian randomization. *Genet Epidemiol*. (2016) 40:597–608. doi: 10.1002/gepi.21998

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 29 August 2024
DOI 10.3389/fimmu.2024.1413560

OPEN ACCESS

EDITED BY
Xu-jie Zhou,
Peking University, China

REVIEWED BY
Xiaofei Hu,
Army Medical University, China
Ping Zhu,
Air Force Medical University, China

*CORRESPONDENCE
Xin Lou
louxin@301hospital.com.cn
Feng Huang
fhuang@301hospital.com.cn

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

RECEIVED 07 April 2024
ACCEPTED 12 August 2024
PUBLISHED 29 August 2024

CITATION
Hu Z, Wang Y, Ji X, Xu B, Li Y, Zhang J, Liu X, Li K, Zhang J, Zhu J, Lou X and Huang F (2024) Radiomics-based machine learning model to phenotype hip involvement in ankylosing spondylitis: a pilot study. *Front. Immunol.* 15:1413560. doi: 10.3389/fimmu.2024.1413560

Radiomics-based machine learning model to phenotype hip involvement in ankylosing spondylitis: a pilot study

Zhengyuan Hu[1†], Yan Wang[2†], Xiaojian Ji[1†], Bo Xu[3], Yan Li[1], Jie Zhang[1], Xingkang Liu[1], Kunpeng Li[1], Jianglin Zhang[1], Jian Zhu[1], Xin Lou[2*] and Feng Huang[1*]

[1]Department of Rheumatology and Immunology, The First Medical Center, Chinese PLA General Hospital, Beijing, China, [2]Department of Radiology, The First Medical Center, Chinese PLA General Hospital, Beijing, China, [3]Basic Research Center for Medical Science, Academy of Medical Science, Zhengzhou University, Zhengzhou, Henan, China

Objectives: Hip involvement is an important reason of disability in patients with ankylosing spondylitis (AS). Unveiling the potential phenotype of hip involvement in AS remains an unmet need to understand its biological mechanisms and improve clinical decision-making. Radiomics, a promising quantitative image analysis method that had been successfully used to describe the phenotype of a wide variety of diseases, while it was less reported in AS. The objective of this study was to investigate the feasibility of radiomics-based approach to profile hip involvement in AS.

Methods: A total of 167 patients with AS was included. Radiomic features were extracted from pelvis MRI after image preprocessing and feature engineering. Then, we performed unsupervised machine learning method to derive radiomics-based phenotypes. The validation and interpretation of derived phenotypes were conducted from the perspectives of clinical backgrounds and MRI characteristics. The association between derived phenotypes and radiographic outcomes was evaluated by multivariable analysis.

Results: 1321 robust radiomic features were extracted and four biologically distinct phenotypes were derived. According to patient clinical backgrounds, phenotype I (38, 22.8%) and II (34, 20.4%) were labelled as high-risk while phenotype III (24, 14.4%) and IV (71, 42.5%) were at low risk for hip involvement. Consistently, the high-risk phenotypes were associated with higher prevalence of MRI-detected lesion than the low-risk. Moreover, phenotype I had significant acute inflammation signs than phenotype II, while phenotype IV was enthesitis-predominant. Importantly, the derived phenotypes were highly predictive of radiographic outcomes of patients, as the high-risk phenotypes were 3 times more likely to have radiological hip lesion than the low-risk [27 (58.7%) vs 16 (28.6%); adjusted odds ratio (OR) 2.95 (95% CI 1.10, 7.92)].

Conclusion: We confirmed for the first time, the clinical actionability of profiling hip involvement in AS by radiomics method. Four distinct phenotypes of hip involvement in AS were identified and importantly, the high-risk phenotypes could predict structural damage of hip involvement in AS.

KEYWORDS

radiomics, spondylitis, ankylosing, hip involvement, machine learning, magnetic resonance imaging

Introduction

Ankylosing spondylitis (AS) is a chronic inflammatory disease that primarily involves the spine, sacroiliac joints and peripheral joints, which could potentially lead to significant morbidity and disability (1). Hip involvement is a prevalent manifestation and an important cause of disability in AS. It is also associated with spine damage, function impairment, increased disease burden and poor prognosis in AS (2, 3). Magnetic resonance image (MRI) can detect early hip lesion in AS and plays an important role in the diagnosis of hip involvement in AS (4). However, MRI-detected hip lesions like joint effusion, subchondral bone marrow edema (BME) were not AS-specific, they could also appear in a wide spectrum of clinical entities such as osteoarthritis, stress injury, femoral head avascular necrosis, joint infection and inflammatory disorders (5, 6). Moreover, it is prone to overestimate the prevalence of hip involvement in AS if we only rely on the present of abnormal MRI lesions (7) and the gold-standard MRI definition of hip involvement in AS is still lacking. Therefore, a new method that accurately predicts hip involvement in AS is urgently needed.

Radiomics has gained increasing attention over the last decade as a promising quantitative image analysis method that had been successfully used in patient phenotyping and prediction of treatment response in a wide variety of diseases (8, 9). Generally, radiomic features were firstly extracted from regions of interest (ROIs) in routine images like CT or MRI. Then, the radiomic features containing crucial information about disease were progressed by artificial intelligent techniques like machine learning (ML) or deep learning methods. Radiomics was initiated in oncology studies and extended to musculoskeletal diseases in the last few years (10). Moreover, ML-based deciphering of complex diseases, such as sepsis, heart failure, ARDS and COVID-19 (11–14), had successfully identified biologically distinct phenotypes and facilitated the understanding of their biological mechanisms. Therefore, we hypothesized that radiomics is a promising method in profiling of hip involvement in AS. We did this pilot study to evaluate the clinical actionability of using radiomics data to phenotype AS patients with symptomatic hip involvement and predict structural damage of hip joint in AS.

Materials and methods

We retrospectively investigated AS patients with hip joint pain and who underwent pelvis MRI exams since January 2019 to September 2022, at the First Medical Center of the Chinese People's Liberation Army (PLA) General Hospital, a tertiary referral center in Beijing. All enrolled patients met the following criteria: they were diagnosed with AS according to the 1984 modified New York criteria (15) and whose MRI imaging fulfilled the quality criteria for reading. Patients with other comorbidities that potentially result in hip joint pain were excluded. Socio-demographic data, type of previous anti-inflammatory medication (non-steroidal anti-inflammatory drugs (NSAIDs) and tumor necrosis factor inhibitors (TNFi)) and clinical assessments were obtained from medical records. Clinical assessments included age at onset, disease duration, peripheral arthritis history, serum inflammatory markers level (C-reactive protein (CRP) and erythrocyte sedimentation rate (ESR)) and HLA-B27 status. Furthermore, X-rays of anterior–posterior pelvis were collected and the severity of structure damage of hip joint was assessed by the Bath ankylosing spondylitis radiology hip index (BASRI-hip) (16). Research ethics approval was granted by the Ethical Committee of the Chinese PLA General Hospital (S2023-375-01) and informed consent was waived due to the retrospective nature of the study. Our works were conducted in accordance with the Declaration of Helsinki.

MRI image acquisition and preprocessing

As the real-world background, patients underwent MRI exams in 8 MRI scanners at our hospital. The parameters of different scanners were detailed in Supplementary Table S1. To correct the heterogeneity of radiomic features caused by different scanners, we used a practical realignment approach, the comBat compensation method (17). This method realigns image-derived data in a single space in which the batch effect is discarded. This method enables pooling data from different scanners and centers without a substantial loss of statistical power caused by intra- and inter-center variability

(18, 19). Image preprocessing was conducted as a fixed bin size of 25 for image discretization was used to filter noise from images and all images were resampled at the same voxel size ($1 \times 1 \times 1$ mm^3) to standardize the voxel spacing. A detailed workflow of the steps involved in our study was summarized in Figure 1.

Image evaluation and region segmentation

Conventional MRI characteristics of hip joint were reported by two musculoskeletal radiologists (reader 1 and reader 2). The severity of structure damage of hip joint was also assessed by reader1, according to the BASRI-hip. The presence of joint effusion, BME and enthesitis was considered as active inflammatory changes, whereas sclerosis, subchondral erosion, joint space narrowing and fat lesion were termed as structural damage of hip involvement (7). We defined active inflammatory changes and chronic structural damage with reference to previously reported method (7). Additionally, we used a qualitative method to define these lesions: the presence of a defined lesion in any slice of hip MRI was considered positive for that lesion. A senior radiologist would also be brought into making the final conclusion if there was disagreement between the two observers. Then, a fellowship-trained operator (reader 3) delineated the entire hip joint, composed of the femur, acetabulum, and joint space, as regions of interest (ROI). The reader delineated the ROIs with reference to the range of proximal hip femur, acetabulum and hip joint capsule in slices on an open-source software, 3D Slicer (Version 5.0.3). The ROIs were drawn manually slice by slice in the axial axis, by using edge-based tool and then fine-tuned by the smoothing tool in 3D Slicer (Figure 2).

Radiomic features extraction and selection

Radiomic features were extracted in the open-source radiomics platform, Pyradiomics (version 3.0.1), in Python (version 3.7). Radiomic features were defined according to the Image Biomarkers Standardization Initiative (IBSI) (20) and fell into the following categories: first-order (n=18), shape (n=8) and texture (n=75) features. Moreover, 14 image filters were applied and high-order features (n=1210) were extracted after decompositions of the original images by the filters. A list of all radiomic features and detailed explanation were provided in Supplementary Table S2.

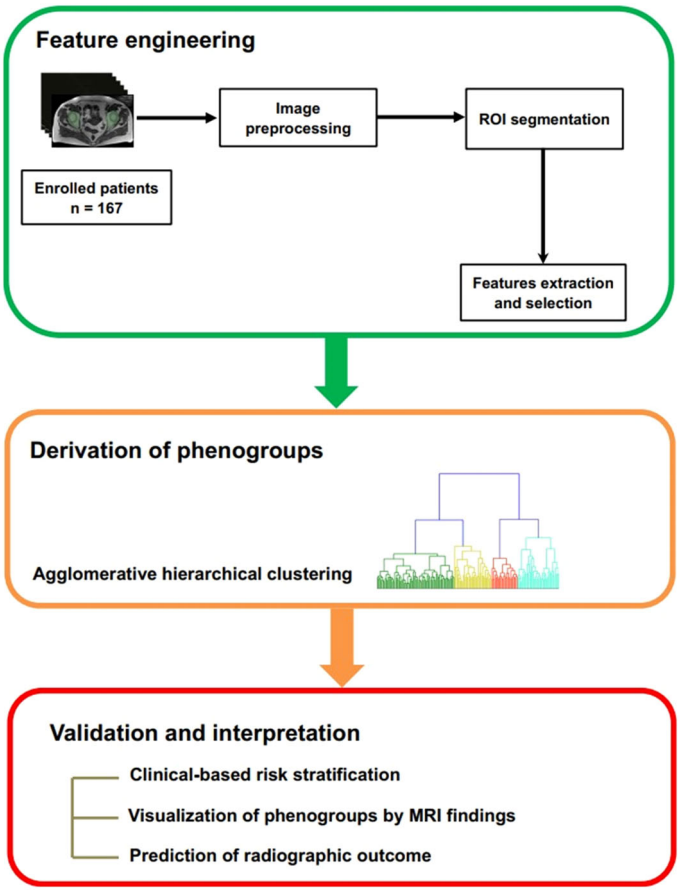


FIGURE 1
Workflow for the development and validation of the radiomics-based machine learning model. ROI: region of interest.

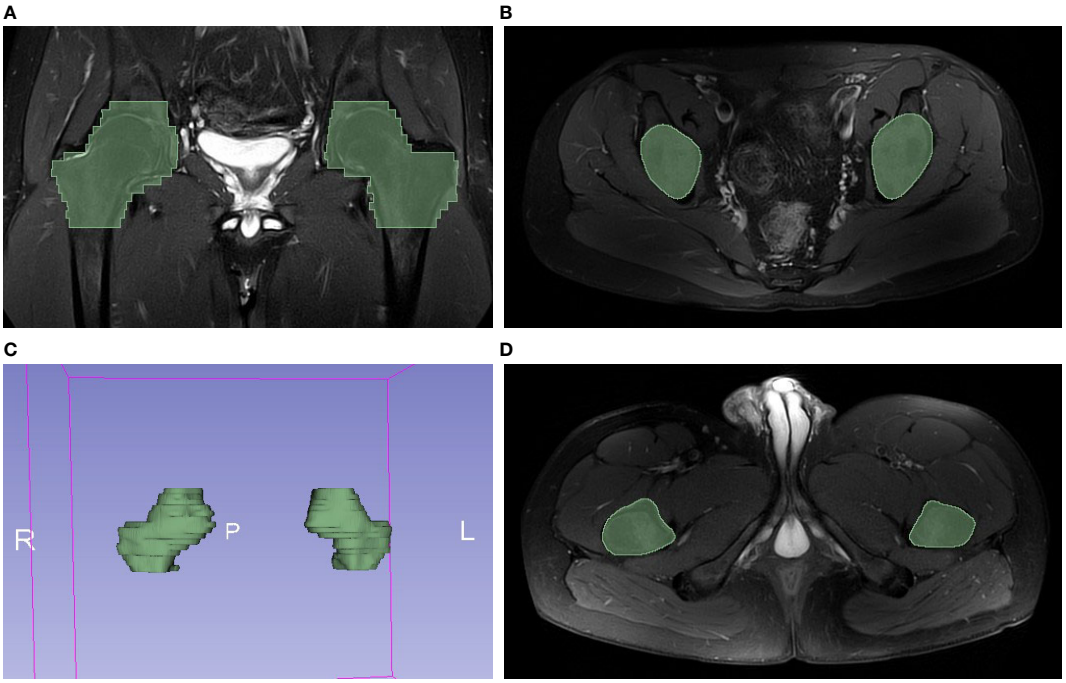


FIGURE 2
Example of hip MRI slices showed the range of handcrafted segmentation. **(A)** Regions of interest (ROI) of bilateral hips were labeled with green color in coronal plane. **(B)** The first slide containing ROI in axial plane. **(C)** The reconstructed 3D volume of ROI. **(D)** The last slide containing ROI in axial plane.

Redundancy was checked and radiomic features with invariance were removed. Additionally, to assess the reliability of manual segmentation process, another observer (reader 1) delineated 15 randomly selected patients, after training session and consensus meeting with reader 3. Then, inter-observer (reader 1 and 3) and intra-observer (reader 3 twice) intraclass correlation (ICC) were calculated to evaluate the reliability of extracted radiomic features. Only features with good reproducibility that both inter-observer and intra-observer ICC $\geq$ 0.75 were considered in further analyses. All selected features were normalized by Z-score standardization before the next step.

Phenotype derivation, validation and interpretation

Once radiomic features were selected and prepared, unsupervised agglomerative hierarchical clustering with Euclidean distance calculation and Ward linkage criterion was applied to identify radiomics-based patient clusters. Dendrogram that visualizes the clustering procedure and distances between the clusters at different layers was prepared to help determine the optimal number of clusters (phenotypes).

The validation of derived phenotypes was conducted in three ways. First, we characterized the derived phenotypes by clinical backgrounds. In detail, we evaluated inter-groups differences of clinical factors associated with hip involvement, such as juvenile-onset, disease duration, cigarette smoking, TNFi treatment and serum inflammation markers. Second, we interpreted phenotyping results by profiling the heterogeneity of MRI-detected hip lesions between phenotypes. Third, we assessed the radiographic outcomes of hip involvement by the BASRI-hip criteria, to evaluate the performance of radiomics-based phenotyping to predict hip joint structural damage.

Validation of radiomic-derived phenotypes

To evaluate the robustness and reliability of the phenotypes obtained from unsupervised agglomerative hierarchical clustering, we performed a consensus clustering algorithm using the 'ConsensusClusterPlus' package (version 1.62.0). This method involves conducting multiple iterations of clustering on resampled data and then measuring the consistency of the resulting clusters across these iterations (21).

The performance of consensus clustering was assessed using the consensus matrix, cumulative distribution function (CDF) curve, relative alterations in the area under the CDF curve (Delta Area Plot), and cluster-consensus plot, in order to help determine the optimal number of phenotypes and evaluate whether the derived phenotypes are reasonable.

Statistics

Descriptive statistical analysis was performed using SPSS Statistics (version 22; IBM Corp.). Missing data were addressed using multiple imputation by 5 iterations, assuming they were missing at random. Implementation of other work is based on Python (version 3.7) and R programming language (version 4.2.1). The ICC coefficient was calculated by the two-way mixed effect models and consistency method, by using R package 'psych' package (version 2.2.9). Unsupervised agglomerative hierarchical clustering and the formation of dendrogram were based on Python package 'scikit-learn' (version 0.22.1). Chord diagrams were created using R package 'circlize' (version 0.4.15). We used binary logistic regression to estimate odds ratios (ORs) and 95% CIs of having radiological hip involvement across the derived-phenotypes. For all analyses, two-sided *P* values <0.05 were considered significant.

Results

Patients and MRI imaging findings

A total of 167 patients were admitted into our study.146 patients were males (87.4%), the median age (interquartile range (IQR)) was 31.0 (26.0–37.0) years. They had established AS with median disease duration (IQR) of 6 (2.0–10.0) years and their median age (IQR) at disease onset was 23.0 (20.2–28.0). HLA-B27 positive rate was 88.6% and 18 (10.8%) individuals were identified

TABLE 1 Characteristics and MRI findings of patients among different phenogroups.

	Total (n= 167)	Phenogroup I (n= 38)	Phenogroup II (n= 34)	Phenogroup III (n= 24)	Phenogroup IV (n= 71)	*P* value
Clinical characteristics						
Age, yrs	31.0 (26.0–37.0)	29.0 (22.0, 33.0)	32.0 (26.0, 37.3)	30.0 (25.3, 35.8)	34.0 (28.0, 37.0)	0.125
Male	146 (87.4%)	30 (78.9%)	28 (82.4%)	24 (100.0%)	64 (90.1%)	**0.046**
JAS	18 (10.8%)	8 (21.1%)	6 (17.6%)	1 (4.2%)	3 (4.2%)	**0.015**
Age at onset, yrs	23.0 (20.2, 28.0)	21.0 (18.5, 24.0)	25.0 (20.8, 28.3)	23.0 (20.2, 28.5)	25.0 (22.0, 30.0)	0.125
Disease duration, yrs	6.0 (2.0, 10.0)	7.0 (3.0, 12.0)	5.0 (2.0, 13.3)	5.0 (3.0, 8.5)	6.0 (2.0, 10.0)	0.840
HLA-B27 (+)	148 (88.6%)	35 (92.1%)	32 (94.1%)	21 (87.5%)	60 (84.5%)	0.483
Peripheral arthritis history	70 (41.9%)	12 (31.6%)	11 (32.4%)	12 (50.0%)	35 (49.3%)	0.165
Enthesitis history	71 (42.5%)	18 (47.4%)	11 (32.4%)	10 (41.7%)	32 (45.1%)	0.579
Smoking status						0.712
None	127 (76.0%)	30 (78.9%)	26 (76.5%)	16 (66.7%)	55 (77.5%)	
Ever smokers	40 (24.0%)	8 (21.1%)	8 (23.5%)	8 (33.3%)	16 (22.5%)	
Alcohol consumption						0.143
None	145 (86.8%)	35 (92.1%)	32 (94.1%)	18 (75.5%)	60 (84.5%)	
With drinking habit	22 (13.2%)	3 (7.9%)	2 (5.9%)	6 (25.0%)	11 (15.5%)	
ESR, mm/h	7.0 (2.0, 18.0)	17.0 (7.0, 49.5)	8.5 (2.0, 19.3)	4.0 (2.0, 11.5)	6.0 (2.0, 13.0)	< **0.001**
CRP, mg/L	3.4 (1.0, 10.9)	6.5 (2.3, 29.5)	5.6 (1.0, 13.7)	4.1 (1.0, 9.6)	3.0 (0.5, 8.3)	**0.021**
NSAIDs	161 (96.4%)	36 (94.7%)	32 (94.1%)	22 (91.7%)	71 (100.0%)	0.148
TNFi	88 (52.7%)	19 (50.0%)	17 (50.0%)	17 (70.8%)	35 (49.3%)	0.303
TNFi duration, month	4.0 (0.0, 24.0)	30.0 (13.0, 48.0)	20.0 (11.5, 38.0)	20.0 (6.0, 27.0)	21.0 (11.0, 36.0)	0.905
MRI findings						
Joint effusion	147 (88.0%)	36 (94.7%)	29 (85.3%)	20 (83.3%)	67 (94.4%)	0.174
BME	75 (44.9%)	22 (57.9%)	13 (38.2%)	12 (50.0%)	28 (39.4%)	0.230
Enthesitis-t	61 (36.5%)	16 (42.1%)	12 (35.3%)	6 (25.0%)	27 (38.0%)	0.582
Enthesitis-i	10 (6.0%)	6 (15.8%)	0	0	4 (5.6%)	**0.023**
Enthesitis-p	34 (20.4%)	12 (31.6%)	1 (2.9%)	4 (16.7%)	17 (23.9%)	**0.009**

(Continued)

TABLE 1 Continued

	Total (n= 167)	Phenogroup I (n= 38)	Phenogroup II (n= 34)	Phenogroup III (n= 24)	Phenogroup IV (n= 71)	*P* value
MRI findings						
Sclerosis	9 (5.4%)	4 (10.5%)	3 (8.8%)	0	2 (2.8%)	0.169
Erosion	62 (37.1%)	23 (60.5%)	13 (38.2%)	5 (20.8%)	21 (29.6%)	**0.004**
Fat lesion	59 (35.3%)	14 (36.8%)	13 (38.2%)	5 (20.8%)	27 (38.0%)	0.472
Narrowing	38 (22.8%)	17 (44.7%)	7 (20.6%)	2 (8.3%)	12 (16.9%)	**0.002**
Radiological outcomes, (missing = 65)						
BASRI-hip	1.0 (1.0, 3.0)	2.0 (1.0, 4.0)	2.0 (1.0, 3.0)	1.0 (0, 2.0)	1.0 (1.0, 2.0)	**0.027**
Radiological-defined hip involvement	45/102 (44.1%)	16/26 (61.5%)	11/20 (55.0%)	3/13 (23.1%)	13/43 (30.2%)	**0.019**

Data are n (%) for categorical variables and median (interquartile range) for continuous variables, respectively. JAS, juvenile-onset ankylosing spondylitis; ESR, erythrocyte sedimentation rate; CRP, C-reactive protein; NSAIDs, non-steroidal anti-inflammatory drugs; TNFi, tumor necrosis factor inhibitor; BME, bone marrow edema; Enthesitis-t, enthesitis at greater femoral trochanter; Enthesitis-i, enthesitis at ischial tuberosity; Enthesitis-p, enthesitis at pubic symphysis; BASRI-hip, Bath ankylosing spondylitis radiology hip index. Bold text highlighted significant differences.

as juvenile-onset AS (JAS). Among the 167 patients, 70 (41.9%) or 71 (42.5%) patients had history of peripheral arthritis or enthesitis, respectively. Besides, 40 (24.0%) patients were ever-smokers and 22 (13.2%) patients had drinking habit.

Joint effusion was the most frequent MRI finding (147, 88.0%), followed by BME (75, 44.9%), erosion (62, 37.1%), fat lesion (59, 35.3%), joint space narrowing (38, 22.8%) and sclerosis (9, 5.4%). Enthesitis was also a prevalent MRI finding and three subtypes were

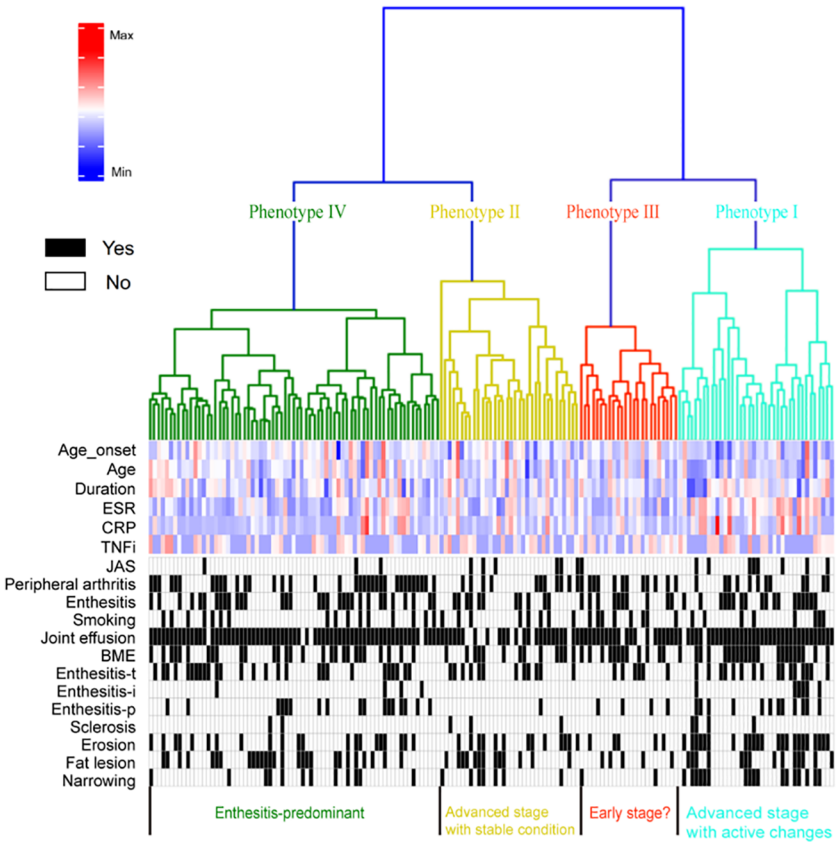


FIGURE 3
Dendrogram shows the process of unsupervised hierarchical clustering. Heatmap shows results of the cluster analysis of patient clinical profiles and MRI-detected lesions. ESR, erythrocyte sedimentation rate; CRP, C-reactive protein; TNFi, tumor necrosis factor inhibitor; JAS, juvenile-onset ankylosing spondylitis; Peri_history, Peripheral arthritis history; E_history, Enthesitis history; BME, bone marrow edema; Enthesitis-t, enthesitis at greater femoral trochanter; Enthesitis-i, enthesitis at ischial tuberosity; Enthesitis-p, enthesitis at pubic symphysis.

identified based on anatomic location: ischial tuberosity (enthesitis-i, 10 (6.0%)), greater femoral trochanter (enthesitis-t, 61 (36.5%)) and pubic symphysis (enthesitis-p, 34 (20.4%)). Detailed patient characteristics and MRI findings were shown in Table 1.

Radiomic features and phenotypes derivation

1422 radiomic features were extracted based on T2WI MRI images. After removing redundant and instable features, 1321 robust radiomic features were identified and used for model construction. The agglomerative hierarchical clustering model identified four phenotypes of patients (Figure 3). Characteristics including demographics, clinical variables, serum inflammation markers and previous treatments across the four phenotypes were presented in Table 1.

Phenotype I consisted of 38 (22.8%) patients. Compared to the others, it included more younger (median age 29.0 years, IQR (22.0, 33.0)) and JAS (8, 21.1%) patients. Besides, patients in phenotype I had longer AS duration (7.0 (3.0, 12.0)) and significantly elevated serum inflammatory markers (17.0 (7.0, 49.5) and 6.5 (2.3, 29.5) for ESR and CRP, respectively). Phenotype II consisted of 34 (20.4%) patients. As similar to phenotypes I, phenotypes II included patients with high rate of juvenile-onset (6, 17.6%) and elevated serum inflammatory markers (8.5 (2.0, 19.3) and 5.6 (1.0, 13.7) for ESR and CRP, respectively). The TNFi use rate in phenotypes II was similar to that in phenotype I (50.0% vs 50.0%, P=0.593) but phenotypes II had shorter duration of TNFi use than phenotypes I (20.0 (11.5, 38.0) vs 30.0 (13.0, 48.0), P=0.043).

Phenotype III consisted of 24 (14.4%) patients and phenotype IV included 71 (42.5%) patients. They shared similar characteristics that patients were neither apt to be JAS (4.2% and 4.2% for phenotype III and IV, respectively) nor had elevated serum inflammatory markers (ESR 4.0 (2.0, 11.5) and 6.0 (2.0, 13.0), CRP 4.1 (1.0, 9.6) and 3.0 (0.5, 8.3) for phenotype III and IV, respectively). As for TNFi treatment, the duration of TNFi use in phenotype III (20.0 (6.0, 27.0)) and IV (21.0 (11.0, 36.0)) were comparable to phenotype II (20.0 (11.5, 38.0), despite more frequent TNFi use in phenotype III (50.0%, 70.8% and 49.3% for phenotype II, III and IV, respectively, P= 0.905).

Therefore, according to their exposure on known clinical factors associated with hip involvement, phenotype I and II could be

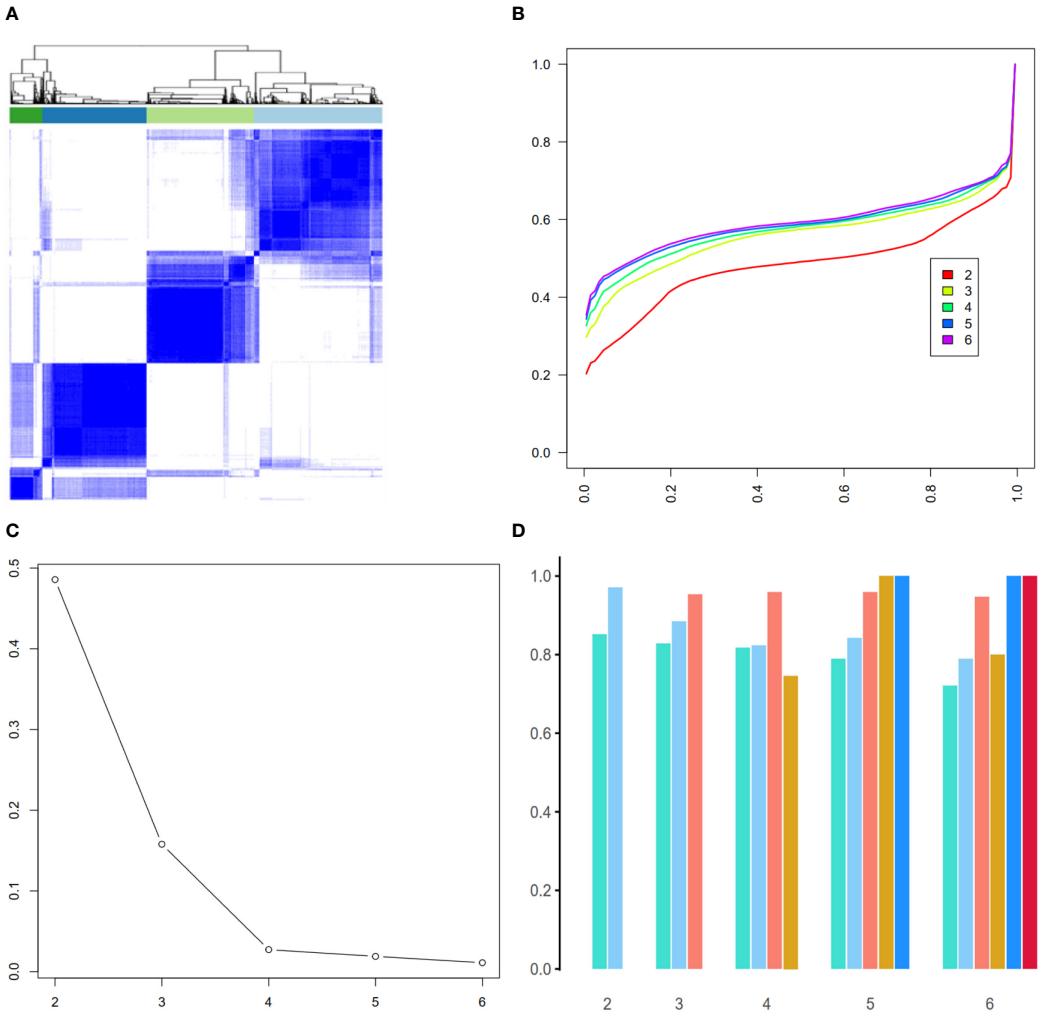


FIGURE 4
Validation of radiomic-derived phenotypes by consensus clustering. **(A)**: Consensus matrix when k = 4. **(B)** Consensus CDF curves when k=2 to 6. **(C)** Relative alterations in CDF Delta area plot. **(D)** Cluster-consensus value of each phenotype when k=2 to 6.

labelled as high-risk while phenotype III and IV were at low-risk for hip involvement in AS.

Validation of radiomic-derived phenotypes by consensus clustering

To assess the robustness of the derived 4-phenotype structure of radiomics data, we performed consensus clustering to validate the radiomics-based phenotypes. Based on the consensus matrix (Figure 4A), CDF curve (Figure 4B), Delta area plot (Figure 4C), k = 4 was identified as the optimal value for phenotyping the AS patients. Additionally, as expected, these four phenotypes had high cluster-consensus values (Figure 4D), indicating strong stability among the radiomic-derived phenotypes.

Interpretation of four phenotypes by MRI findings

Both phenotype I and II manifested high prevalence of structural lesion. More specifically, the high-risk phenotypes were associated with significantly higher prevalence of erosive lesion [36 (50.0%) vs 26 (27.4%), odds ratio (OR) 2.65 (95% CI 1.39, 5.06)] and joint space narrowing [24 (33.3%) vs 14 (14.7%), OR 2.89 (95% CI 1.37, 6.12)] than the low-risk, whereas they did not differ for sclerosis and fat lesion. In contrast, phenotype II had lower prevalence of active lesions than phenotype I (joint effusion (85.3% vs 94.7%, P=0.243), BME (38.2% vs 57.9%, P=0.096), enthesitis-t (35.3% vs 42.1%, P=0.554), enthesitis-i (0 vs 15.8%, P=0.026) and enthesitis-p (2.9% vs 31.6%, P=0.002)), which reflected that phenotype II had severe structural damage but less active inflammatory lesions on MRI.

As for acute inflammatory signs, the high-risk phenotypes had comparable prevalence of joint effusion [65 (90.3%) vs 87 (91.6%), OR 0.46 (95% CI 0.18, 1.19)], BME [35 (48.6%) vs 40 (42.1%), OR 1.30 (95% CI 0.70, 2.41)] and enthesitis-t [28 (38.9%) vs 33 (34.7%), OR 1.20 (95% CI 0.63, 2.26)] than the low-risk phenotypes. Nevertheless, phenotype I and IV had significantly higher prevalence of enthesitis-i (15.8% and 5.6%, respectively, P=0.023) and enthesitis-p (31.6% and 23.9%, respectively, P=0.009) compared to phenotype II and phenotype III (enthesitis-i: 0 for both, enthesitis-p: 2.9% and 16.7%, respectively). MRI findings across the 4 phenotypes were presented in Table 1 and inter-group differences were visualized in Figures 3, 5.

Prediction of radiographic outcomes by phenotypes

102 patients received pelvis X-ray exams at a 2-year interval after taking MRI exams. Patients in phenotype I and II had significantly higher BASRI-hip scores than phenotype III and IV (median (IQR) of scores were 2.0 (1.0, 4.0), 2.0 (1.0, 3.0), 1.0 (0, 2.0) and 1.0 (1.0,2.0), respectively, P=0.027). Likewise, after adjusting for confounding factors including JAS, age, duration, smoking status and ESR, the high-risk phenotypes (phenotype I and II) were 3 times more likely to have radiological-defined hip involvement

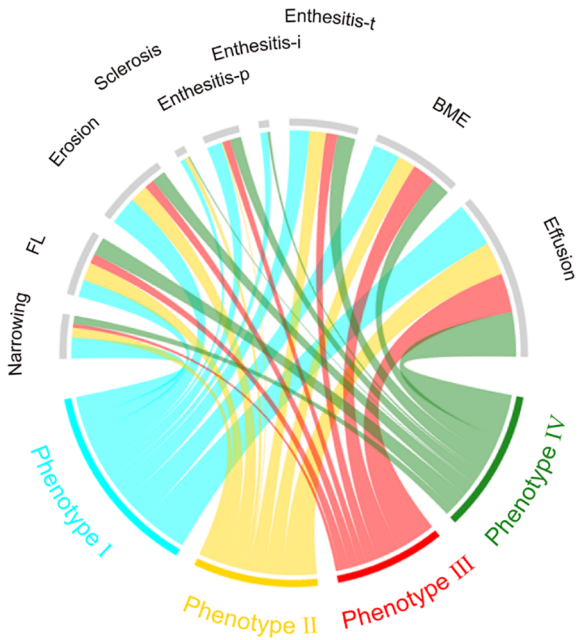


FIGURE 5
Chord diagrams showing differences in MRI findings among phenotypes. BME, bone marrow edema; Enthesitis-t, enthesitis at greater femoral trochanter; Enthesitis-i, enthesitis at ischial tuberosity; Enthesitis-p, enthesitis at pubic symphysis; FL, Fat lesion.

(BASRI-hip ≥ 2) than the low-risk [27 (58.7%) vs 16 (28.6%), adjusted OR 2.95 (95% CI 1.10, 7.92)].

Therefore, according to clinical behaviors, MRI characteristics and radiographic outcomes, patients in phenotype I and II could be labeled as "advanced-stage hip involvement". Patients in phenotype I concomitantly exhibited significant acute inflammation signs and demanded anti-inflammatory therapy, especially TNFi treatment. Phenotype III and IV were assumed as "early-stage hip involvement", and phenotype IV was enthesitis-predominant, whereas patients in phenotype III were not yet identified based on the current variables.

Discussion

Hip involvement is prevalent in AS and constitutes an important reason of disability in AS (2, 3). There remains unmet need that a method can make early and accurate identification of hip involvement in AS, as early detection means the opportunity to get timely treatments. Radiomics has gained increasing attention in the last few years, as a promising quantitative image analyzing method used for differential diagnosis, prognosis analysis and identification of responders to therapy (22, 23). In this pilot study, four distinct phenotypes of AS-related hip involvement were identified by the integration of MRI radiomics data and unsupervised ML approach. This study is, to the best of our knowledge, the first to apply radiomics-based approach to profile hip involvement in AS. Our study validated the clinical actionability of using radiomics approach to detect hip involvement in AS, which offers opportunities for the foundation of a novel method, the MRI radiomics, to diagnose hip involvement in AS.

A 4-phenotype structure of radiomics data were derived and it was validated from the perspectives of clinical backgrounds, MRI signs and radiographic outcomes. Firstly, phenotype I and II were labelled as high-risk clinical pattern, in that they included more patients exposed to risk factors associated with hip involvement than the other two phenotypes (low-risk clinical pattern). Then, we used conventional MRI findings to validate the phenotyping structure and interpreted the radiomics-based phenotypes, since the 'black-box' nature of artificial intelligence-based approaches often provides results that are difficult to understand (24). Practitioners are more familiar with the clinical implications of MRI findings rather than radiomic features. Importantly, the significantly increased prevalence of MRI-detected structural damage on high-risk than low-risk phenotypes vigorously supported such clinical patterns. Additionally, patients in phenotype I had notable acute inflammation signs besides the presence of structural damage while phenotype IV was assumed as "enthesitis-predominant", given the prominent enthesitis findings on MRI. The profiling of phenotype III was challenging since it had limited cases number (only 24 patients). Patients in phenotype III were young and less likely exposed to risk factors associated with hip involvement, we carefully inferred that their nonspecific MRI findings may derive from other origins of hip joint pain, such as stress injury, acute bone marrow edema syndrome or femoroacetabular impingement (25, 26), besides the possibility that they represent a stage, probably the early stage, in the progression of AS-related hip involvement.

The radiographic outcomes of hip involvement strongly supported the current phenotyping results. After adjusting for confounding factors, patients with high-risk phenotypes were associated with 3.0-fold higher odds of having radiological hip involvement than the low-risk (ORa 2.95 (95% CI 1.10, 7.92)). This finding suggested that radiomics-derived phenotyping could predict the radiographic outcome of hip involvement in AS, which makes the radiomics method a promising tool in the early identification of hip involvement in AS. Additionally, consensus clustering analysis significantly enhances the credibility and robustness of our findings. These results endorse that the derived phenotypes are not only statistically sound but also clinically interpretable and meaningful.

Among the reported MRI findings associated with hip involvement in AS, we don't know which were of predictive power for worse outcome or which could discriminate it from other reasons of hip pain. Our study provided some indirective evidence for this question. Joint effusion is an indirective MRI finding of hip synovitis and BME is linked to bone marrow capillary wall damage and leakage (5). Joint effusion and BME were quite common MR findings in AS patients with hip joint pain (7) but they had a low-level variance among the 4 phenotypes. Erosion, sclerosis and joint space narrowing were structural lesion findings in MRI, their roles were quite limited since the target was early diagnosis of hip involvement. Focal fat infiltration likely reflects post-inflammatory tissue metaplasia: since the inflammation recedes, fat metaplasia develops in its place (27, 28). The prevalence of fatty lesion was comparable in phenotype I, II and IV (36.8%, 38.2% and 38.0%, respectively), despite it subtle decreased in phenotype III (20.8%). We also found that enthesitis was a prevalent MRI finding in each phenotype and it comprised one distinct phenotype of patients. Further studies are needed to dissect the pathophysiologic significance of fat lesion and enthesitis in hip joints and their value in sorting out AS-related hip involvement from other origins of hip joint pain. It is noteworthy that we evaluated the described MRI signs in a crude mode that whether they existed or not and the emergence of sophisticated methods such as morphological feature analysis, quantitative scoring and radiomic feature analysis, had shed light on exploring of AS-specific MRI findings (10, 29, 30).

Our study has several limitations that should be acknowledged. Firstly, there existed sampling bias due to various factors, including relatively young population and a geographical area where AS population had limited biologics use (31), which may render a relative high prevalence of hip involvement. Additionally, we enrolled patients with AS (radiographic axial SpA) rather than non-radiographic axial SpA, which was assumed as the pre-stage of axial SpA (1). Further researches are needed to investigate whether our observations persist across racial, ethnic and the whole SpA groups. Secondly, we did not set out a specific prediction model or scoring system for the prediction of hip involvement in AS, which we believe requires further developed tools as well as external validation. Rather, we aimed to ascertain the potential of MRI radiomics approach to profile hip involvement in AS. We believed

that the novelty predominantly lies in the described methodology, and perhaps less so in the detected four phenotypes, despite that they were comprehensively validated. Finally, patients in phenotype III were not yet identified and the underlying cellular or molecular level heterogeneity across the four phenotypes were not studied.

In conclusion, our results serve as a proof-of-concept that unsupervised ML methods could turn complex radiomics data into interpretable and clinically meaningful classification of hip involvement in AS. Our findings illuminate a promising approach to identify hip involvement in AS and its added value in clinical decision making should be evaluated in prospective studies.

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.

Ethics statement

The studies involving humans were approved by the Ethical Committee of the Chinese PLA General Hospital. The studies were conducted in accordance with the local legislation and institutional requirements. The ethics committee/institutional review board waived the requirement of written informed consent for participation from the participants or the participants' legal guardians/next of kin because the retrospective nature of the study.

Author contributions

ZH: Writing – original draft. YW: Writing – original draft. XJ: Writing – review & editing. BX: Writing – review & editing. YL: Writing – review & editing. JZha: Writing – review & editing. XKL: Writing – review & editing. KL: Writing – review & editing. JLZ: Writing – review & editing. JZhu: Writing – review & editing. XL: Writing – review & editing. FH: Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was supported by National Key R&D Program of China (2021ZD0140409 to KL), National Natural Science Foundation of China (82151309 and 82327803 to XL), and Youth Independent Innovation Science Fund Project of Chinese PLA General Hospital (22QNFC139 to XJ).

Acknowledgments

We would like to show our gratitude to all the participants of this pilot study. We also thank the hospital staff members for the convenience they provided in collecting the information that used in this study.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1413560/full#supplementary-material

References

1. Sieper J, Poddubnyy D. Axial spondyloarthritis. *Lancet*. (2017) 390:73–84. doi: 10.1016/S0140-6736(16)31591-4

2. Vander Cruyssen B, Munoz-Gomariz E, Font P, Mulero J, de Vlam K, Boonen A, et al. Hip involvement in ankylosing spondylitis: epidemiology and risk factors associated with hip replacement surgery. *Rheumatology*. (2010) 49:73–81. doi: 10.1093/rheumatology/kep174

3. Vander Cruyssen B, Vastesaeger N, Collantesestévez E. Hip disease in ankylosing pondylitis. *Curr Opin Rheumatol*. (2013) 25:448–54. doi: 10.1097/BOR.0b013e3283620e04

4. Zheng Y, Zhang K, Han Q, Hao Y, Liu Y, Yin H, et al. Application and preliminary validation of the hip inflammation MRI scoring system (HIMRISS) in spondyloarthritis. *Int J Rheum Dis*. (2019) 22:228–33. doi: 10.1111/1756-185X.13451

5. Vassalou EE, Spanakis K, Tsifountoudis IP, Karantanas AH. MR imaging of the hip: an update on bone marrow edema. *Semin Musculoskelet Radiol*. (2019) 23:276–88. doi: 10.1055/s-0039-1677872

6. Patel S. Primary bone marrow oedema syndromes. *Rheumatol (Oxford)*. (2014) 53:785–92. doi: 10.1093/rheumatology/ket324

7. Huang ZG, Zhang XZ, Hong W, Wang GC, Zhou HQ, Lu X, et al. The application of MR imaging in the detection of hip involvement in patients with ankylosing spondylitis. *Eur J Radiol*. (2013) 82:1487–93. doi: 10.1016/j.ejrad.2013.03.020

8. Chen Q, Zhang L, Liu S, You J, Chen L, Jin Z, et al. Radiomics in precision medicine for gastric cancer: opportunities and challenges. *Eur Radiol*. (2022) 32:5852–68. doi: 10.1007/s00330-022-08704-8

9. Shin J, Seo N, Baek SE, Son NH, Lim JS, Kim NK, et al. MRI radiomics model predicts pathologic complete response of rectal cancer following chemoradiotherapy. *Radiology*. (2022) 303:351–58. doi: 10.1148/radiol.211986

10. Fritz B, Yi PH, Kijowski R, Fritz J. Radiomics and deep learning for disease detection in musculoskeletal radiology: an overview of novel MRI- and CT-based approaches. *Invest Radiol*. (2023) 58:3–13. doi: 10.1097/RLI.0000000000000907

11. Seymour CW, Kennedy JN, Wang S, Chang CH, Elliott CF, Xu Z, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA*. (2019) 321:2003–17. doi: 10.1001/jama.2019.5791

12. Cikes M, Sanchez-Martinez S, Claggett B, Duchateau N, Piella G, Butakoff C, et al. Machine learning-based phenotypeing in heart failure to identify responders to cardiac resynchronization therapy. *Eur J Heart Fail*. (2019) 21:74–85. doi: 10.1002/ejhf.1333

13. Maddali MV, Churpek M, Pham T, Rezoagli E, Zhuo H, Zhao W, et al. Validation and utility of ARDS subphenotypes identified by machine-learning models using clinical data: an observational, multicohort, retrospective analysis. *Lancet Respir Med*. (2022) 10:367–77. doi: 10.1016/S2213-2600(21)00461-6

14. Su C, Zhang Y, Flory JH, Weiner MG, Kaushal R, Schenck EJ, et al. Clinical subphenotypes in COVID-19: derivation, validation, prediction, temporal patterns, and interaction with social determinants of health. *NPJ Digit Med*. (2021) 4:110. doi: 10.1038/s41746-021-00481-w

15. van der Linden S, Valkenburg HA, Cats A. Evaluation of diagnostic criteria for ankylosing spondylitis. *A proposal modification New York criteria. Arthritis Rheum*. (1984) 27:361–68. doi: 10.1002/art.1780270401

16. MacKay K, Brophy S, Mack C, Doran M, Calin A. The development and validation of a radiographic grading system for the hip in ankylosing spondylitis: The Bath Ankylosing Spondylitis Radiology Hip Index. *J Rheumatol*. (2000) 27:2866–72.

17. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. (2007) 8:118–27. doi: 10.1093/biostatistics/kxj037

18. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. *J Nucl Med*. (2018) 59:1321–28. doi: 10.2967/jnumed.117.199935

19. Orlhac F, Lecler A, Savatovski J, Goya-Outi J, Nioche C, Charbonneau F, et al. How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. *Eur Radiol*. (2021) 31:2272–80. doi: 10.1007/s00330-020-07284-9

20. Zwanenburg A, Vallieres M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. (2020) 295:328–38. doi: 10.1148/radiol.2020191145

21. Lai Y, Lin P, Lin F, Chen M, Lin C, Lin X, et al. Identification of immune microenvironment subtypes and signature genes for Alzheimer's disease diagnosis and risk prediction based on explainable machine learning. *Front Immunol*. (2022) 13:1046410. doi: 10.3389/fimmu.2022.1046410

22. Chen J, Meng T, Xu J, Ooi JD, Eggenhuizen PJ, Liu W, et al. Development of a radiomics nomogram to predict the treatment resistance of Chinese MPO-AAV patients with lung involvement: a two-center study. *Front Immunol*. (2023) 14:1084299. doi: 10.3389/fimmu.2023.1084299

23. Ye L, Miao S, Xiao Q, Liu Y, Tang H, Li B, et al. A predictive clinical-radiomics nomogram for diagnosing of axial spondyloarthritis using MRI and clinical risk factors. *Rheumatol (Oxford)*. (2022) 61:1440–47. doi: 10.1093/rheumatology/keab542

24. Castelvecchi D. Can we open the black box of AI? *Nature*. (2016) 538:20–3.

25. Hodnett PA, Shelly MJ, MacMahon PJ, Kavanagh EC, Eustace SJ. MR imaging of overuse injuries of the hip. *Magn Reson Imaging Clin N Am*. (2009) 17:667–79. doi: 10.1016/j.mric.2009.06.005

26. Riley GM, McWalter EJ, Stevens KJ, Safran MR, Lattanzi R, Gold GE. MRI of the hip for the evaluation of femoroacetabular impingement; past, present, and future. *J Magn Reson Imaging*. (2015) 41:558–72. doi: 10.1002/jmri.24725

27. Renson T, de Hooge M, De Craemer AS, Deroo L, Lukasik Z, Carron P, et al. Progressive increase in sacroiliac joint and spinal lesions detected on magnetic resonance imaging in healthy individuals in relation to age. *Arthritis Rheumatol*. (2022) 74:1506–14. doi: 10.1002/art.42145

28. Koo BS, Song Y, Shin JH, Lee S, Kim TH. Evaluation of disease chronicity by bone marrow fat fraction using sacroiliac joint magnetic resonance imaging in patients with spondyloarthritis: A retrospective study. *Int J Rheum Dis*. (2019) 22:734–41. doi: 10.1111/1756-185X.13485

29. Mori V, Sawicki LM, Sewerin P, Eichner M, Schaarschmidt BM, Oezel L, et al. Differences of radiocarpal cartilage alterations in arthritis and osteoarthritis using morphological and biochemical magnetic resonance imaging without gadolinium-based contrast agent administration. *Eur Radiol*. (2019) 29:2581–88. doi: 10.1007/s00330-018-5880-6

30. Han Q, Lu Y, Han J, Luo A, Huang L, Ding J, et al. Automatic quantification and grading of hip bone marrow oedema in ankylosing spondylitis based on deep learning. *Mod Rheumatol*. (2022) 32:968–73. doi: 10.1093/mr/roab073

31. Nikiphorou E, van der Heijde D, Norton S, Landewé RB, Molto A, Dougados M, et al. Inequity in biological DMARD prescription for spondyloarthritis across the globe: results from the ASAS-COMOSPA study. *Ann Rheum Dis*. (2018) 77:405–11. doi: 10.1136/annrheumdis-2017-212457

frontiers | Frontiers in Immunology

TYPE Methods
PUBLISHED 02 September 2024
DOI 10.3389/fimmu.2024.1381002

OPEN ACCESS

EDITED BY
Xu-jie Zhou,
Peking University, China

REVIEWED BY
Chenglin Sun,
Jilin University, China
Xiaoyong Yu,
Shaanxi Provincial Hospital of Traditional Chinese Medicine, China

*CORRESPONDENCE
Yunyan Ye
lhlyeyunyan@nbu.edu.cn

RECEIVED 02 February 2024
ACCEPTED 18 July 2024
PUBLISHED 02 September 2024

CITATION
Ye Y, Dai L, Gu H, Yang L, Xu Z and Li Z (2024) The causal relationship between immune cells and diabetic retinopathy: a Mendelian randomization study. *Front. Immunol.* 15:1381002. doi: 10.3389/fimmu.2024.1381002

The causal relationship between immune cells and diabetic retinopathy: a Mendelian randomization study

Yunyan Ye[1]*, Lei Dai[2], Hong Gu[1], Lan Yang[1], Zhangxing Xu[1] and Zhiguo Li[1]

[1]Department of Ophthalmology, Li Huili Hospital Affiliated with Ningbo University, Ningbo, China, [2]Department of Hepato-Pancreato-Biliary Surgery, Li Huili Hospital Affiliated with Ningbo University, Ningbo, China

Purpose: This article explored the causal relationship between immune cells and diabetic retinopathy (DR) using single nucleotide polymorphisms (SNPs) as an instrumental variable and Mendelian randomization (MR).

Methods: Statistical data were collected from a publicly available genome-wide association study (GWAS), and SNPs that were significantly associated with immune cells were used as instrumental variables (IVs). Inverse variance weighted (IVW) and MR–Egger regression were used for MR analysis. A sensitivity analysis was used to test the heterogeneity, horizontal pleiotropy, and stability of the results.

Results: We investigated the causal relationship between 731 immune cells and DR risk. All the GWAS data were obtained from European populations and from men and women. The IVW analysis revealed that HLA DR on CD14+ CD16- monocytes, HLA DR on CD14+ monocytes, HLA DR on CD33-HLA DR+, HLA DR on CD33+ HLA DR+ CD14- on CD33+ HLA DR+ CD14dim, and HLA DR on myeloid dendritic cells may increase the risk of DR ($P<0.05$). HLA DR to CD14-CD16- cells, the monocytic myeloid-derived suppressor cell absolute count, the SSC-A count of CD4+ T cells, and terminally differentiated CD4+ T cells may be protective factors against DR ($P<0.05$). The sensitivity analysis indicated no heterogeneity or pleiotropy among the selected SNPs. Furthermore, gene annotation of the SNPs revealed significant associations with 10 genes related to the risk of developing PDR and potential connections with 12 other genes related to PDR.

Abbreviations: DR, diabetic retinopathy; SNPs, single nucleotide polymorphisms; MR, Mendelian randomization; GWAS, genome-wide association study; IVSs, instrumental variables; IVW, Inverse variance weighted; PDR, proliferative diabetic retinopathy; LOO. leave one-out; MDSCs, myeloid-derived suppressor cells; T1D, type 1 diabetes mellitus; DN, diabetic nephropathy; DRGen, Diabetic Retinopathy Genomics.

Conclusion: Monocytes and T cells may serve as new biomarkers or therapeutic targets, leading to the development of new treatment options for managing DR.

KEYWORDS

Mendelian randomization, immune cells, diabetic retinopathy, proliferative diabetic retinopathy, causal effect

1 Background

Diabetic retinopathy (DR) is one of the most common microvascular complications of diabetes and affects 30% to 50% of diabetic patients. DR can progress to proliferative diabetic retinopathy (PDR) when the severity of ischemia increases, leading to neovascularization, fibroplasia, and retinal detachment, which are the leading causes of blindness and visual impairment in diabetic individuals. Diabetes is expected to affect 415 million people worldwide by 2024, more than one-third of whom suffer from DR, making it a serious global health issue (1, 2). Current DR treatment mostly focuses on regulating blood sugar, blood pressure, and lipid levels to slow down the disease and lower the risk of DR; however, there is still a high number of diabetes patients who develop PDR (3, 4). Early detection and diagnosis of DR, as well as systematic therapy, can prevent persistent vision loss; however, diagnosis and treatment of DR are often delayed due to a lack of resources for early DR screening (5). As a result, identifying more precise and sensitive biomarkers is critical for facilitating early detection of DR and understanding its pathophysiology (6).

Immune cells play a crucial role in the onset and progression of DR. In DR, there is frequent and persistent white blood cell adhesion to the vascular wall, which may result in capillary occlusion and retinal ischemia (7). They also play an important role in the pathogenesis of late PDR and can contribute to neovascularization, vitreous hemorrhage, and traction retinal detachment (8). A recent prospective study demonstrated that the number of circulating neutrophils increases while the number of T cells decreases during the initial stages and progression of DR (9). However, previous research on the pathophysiology of DR mostly relies on association analysis of observational cohorts, which cannot achieve causal association inference. Furthermore, the causal relationships between various immune cells and DR have not been investigated; therefore, there is limited existing evidence regarding immune cell types related to DR and their causal associations.

Mendelian randomization (MR) is a popular causal inference method in which the genetic variation associated with exposure is employed as an instrumental variable (IV) for assessing the causal effect of exposure on outcomes. It remains unaffected by common complicating variables such as acquired environment, life behavior, and habits, allowing it to minimize the reverse causal effect while maintaining maximum validity (10). Compared to traditional randomized controlled trials and observational research, MR can significantly reduce expenses and shorten study periods. It is widely employed in studies investigating the causal association of complex disorders, and the genome-wide association study (GWAS) dataset is expanding rapidly. These findings also provide a solid foundation for further MR research. With the advent of big data, the growth of epidemiological methodologies, and the demand for precision medicine, the application of MR for etiology mining will emerge as a new area of future research (11).

At present, no studies have been conducted to properly investigate the causal relationship between immune cells and DR using MR. Further investigation and study are required for diabetes. In this study, MR analysis was performed to investigate the causal relationship between immune cells and DR.

2 Methods

2.1 Study design

We used MR analysis to evaluate the causal relationship between 731 immune cells and DR. In this study, immune cells were used as exposure factors and represented by X, whereas single nucleotide polymorphisms (SNPs) that were strongly linked with X were used as instrumental variables (IVs). The outcome variable was diabetic retinopathy. Figure 1 depicts a schematic view of the study design, as well as the three essential MR assumptions (12).

2.2 Data source

The analysis was conducted using published summary statistics from the International Working Unit (IEU) Open GWAS project (https://gwas.mrcieu.ac.uk/), and it included 731 immune cells, two DR datasets (Finn-b-DM_RETINOPATHY and finn-b-H7_RETINOPATHYDIAB), and two PDR datasets (finn-b-DM_RETINA_PROLIF and finn-b-H7_RETINOPATHYDIAB_PROLIF). Validation was performed using the datasets finn-b-H7_RETINOPATHYDIAB and finn-b-H7_RETINOPATHYDIAB_PROLIF. The study was conducted on European individuals, including both men and women, and the summary data are provided in Table 1. The current analysis did not require

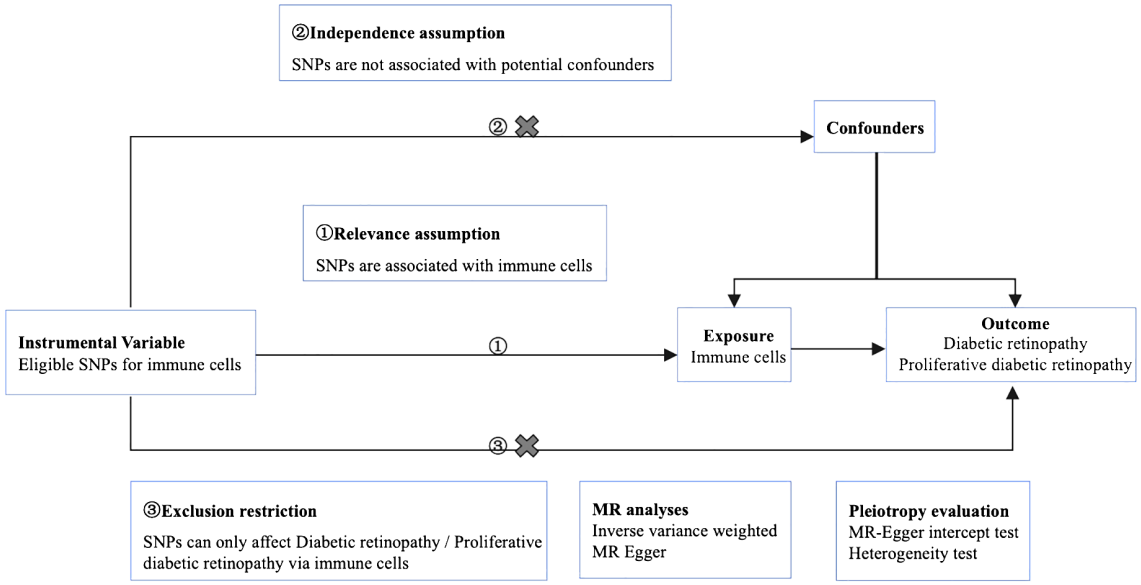


FIGURE 1
Mendelian randomization diagram of the association of immune cells with diabetic retinopathy. ① The genotype must be associated with the exposure to be studied. ② Genotypes must be independent of confounding factors. ③ Genotype is only associated with outcome by influencing the exposure factors to be studied.

ethics approval because all of the included GWASs received ethical review board approval and informed consent, as indicated in their individual original manuscripts.

2.3 Selection and validation of SNPs

The selected SNPs were related to immune cells at a genome-wide significance threshold of $p < 1\times10^{-5}$. Second, pairwise linkage disequilibrium was used to assess the independence of the selected SNPs. When $r^2 > 0.001$ (clumping window of 10,000 kb) was reached, the SNP that correlated with more SNPs or had a higher *P*-value was removed (Figure 1, ①). Phenoscanner was used to minimize the impact of improper SNPs (Figures 1, ②③). The *F*-statistic was subsequently used to validate the strength of each SNP. When the *F*-statistic exceeded 10, SNPs were deemed powerful enough to minimize the effects of potential bias. Furthermore, the SNPs listed above were retrieved from the GWAS summary data of DR and PDR, with a minimum $r^2 > 0.8$. The information from the datasets listed above was summarized (12).

2.4 Mendelian randomization analysis

The causal association study was conducted using inverse variance weighting (IVW) and MR–Egger regression. The discrepancy in intercept terms, as indicated by the intercept of the MR-Egger analysis, revealed horizontal pleiotropy in the study. Cochrane's Q value and accompanying *P*-values were used to assess heterogeneity among the selected IVs, with $P > 0.05$ indicating no heterogeneity. In addition, a leave-one-out (LOO) analysis was performed to observe whether a particular SNP had a disproportionate effect on the overall estimations. Forest plots were used to visualize the MR analysis results, while scatter plots and funnel plots were utilized to assess the stability of the MR data (13, 14).

TABLE 1 Detailed information of datasets.

Data source	Phenotype	Sample size	Cases	Population	Adjustment
IEU Open GWAS project	Immune cells	-	-	European	-
finn-b-DM_RETINOPATHY	Diabetic retinopathy (DM_RETINOPATHY)	-	14584	European	Males and Females
finn-b-H7_RETINOPATHYDIAB	Diabetic retinopathy (H7_RETINOPATHYDIAB)	-	3646	European	Males and Females
finn-b-DM_RETINA_PROLIF	Proliferative diabetic retinopathy (DM_RETINA_PROLIF)	-	8681	European	Males and Females
finn-b-H7_RETINOPATHYDIAB_PROLIF	Proliferative diabetic retinopathy (H7_RETINOPATHYDIAB_PROLIF)	-	1382	European	Males and Females

The statistical power was calculated using an online tool at: http://cnsgenomics.com/shiny/mRnd/ (15). We used the following formula to calculate R^2: (2×EAF×(1−EAF)×beta2)/[(2×EAF×(1−EAF)×beta2)+(2×EAF×(1−EAF)×N×SE(beta)2) (16).

2.5 SNP annotation

The SNPs were annotated using online tools (https://biit.cs.ut.ee/gprofiler/snpense). g: SNPense maps a collection of human SNP rs-codes to gene names, along with chromosome positions and expected variant effects. Mapping was allowed only for variations that coincided with at least one protein coding Ensembl gene. All underlying data were extracted from Ensembl variation data.

2.6 Statistical methods

All the statistical analyses were conducted using R 4.1.0 software and R packages. IVW and MR−Egger analyses were performed using the TwoSample MR package (α= 0.05), meta-analysis using the meta package, and a statistically significant difference was indicated by P < 0.05. If the null hypothesis was rejected, random effects IVW was utilized rather than fixed effects IVW (17). Additionally, the Forest Plats package was used to generate forest plots.

3 Results

3.1 Selected SNPs

A total of 6,196 SNPs in the DR and 6,186 in the PDR MR analyses were used, respectively (Supplementary Table 2). We obtained the degree of phenotype overlap from the FinnGen database. Among diabetic retinopathy phenotypes, there is a 57.47% sample overlap between the DM_RETINOPATHY cohort and the H7_retinydiab cohort. In terms of the proliferative diabetic retinopathy phenotype, there is a 25.82% sample overlap between the DM_RETINA_PROLIF cohort and the H7_retinyDIAB_prolif cohort (Supplementary Table Overlap).

3.2 MR analysis results

MR analysis was performed to explore the causal effects of immune cells on DR, and the IVW method was used as the primary analysis. According to MR analysis using the finn-b-DM_RETINOPATHY dataset, IVW analysis revealed that 30 immune cells were significantly associated with DR. In total, 18 immune cells were found to increase the risk of DR; for example, HLA DR was found in CD33+ HLA DR+ CD14- (OR=1.229, 95% CI=1.178-1.283, *P*<0.001), and HLA DR was found in CD33+ HLA DR+ CD14dim (OR=1.323, 95% CI=1.239-1.413, *P*<0.001). Furthermore, 12 immune cells, such as HLA-DR on CD14-CD16 + cells (odds ratio (OR)=0.798, 95% CI=0.748-0.852, *P*<0.001) and on CD4+ T cells (OR=0.477, 95% CI=0.403-0.565, *P*<0.001), may decrease the risk of DR (Figure 2).

The IVW analysis based on the FinN-B-H7_RETINOPATHYDIAB dataset revealed that 36 immune cells were significantly associated with DR. Among these, HLA-DR among CD33+ HLA DR+ CD14- (OR=1.716, 95% CI=1.531-1.924, *P*<0.001) and HLA-DR among CD33+ HLA DR+ CD14dim (OR=2.240, 95% CI=1.968-2.550, *P*<0.001) were identified as two of the 15 immune cells that may increase the risk of DR. Additionally, HLA-DR among CD14- CD16- (OR=0.686, 95% CI=0.605-0.778, *P*<0.001) and SSC-A among CD4+ T cells (OR=0.196, 95% CI=0.136-0.282, *P*<0.001) were identified as two of the 21 immune cells that may decrease the risk of DR (Figure 3).

Merged MR analysis results from the FINN-B-H7_RETINYDIAB and FINN-b-DM_RETINOPATHY datasets revealed 10 immune cells. HLA-DR on myeloid dendritic cells, HLA-DR on CD14+ CD16- monocytes, HLA-DR on CD33+ HLA-DR+ CD14-, HLA-DR on CD14+ monocytes, HLA-DR on CD33- HLA-DR+, and HLA-DR on CD33+ HLA-DR+ CD14dim are six immune cells that may be risk factors for DR. Additionally, the following four immune cells may serve as protective factors for DR: monocytic myeloid-derived suppressor cell absolute count, terminally differentiated CD4+ T cell, HLA-DR on CD14- CD16-, and SSC-A on CD4+ T cells (Table 2A).

After merging the MR analysis results from the finn-b-DM_RETINA_PROLIF and finn-b-H7_RETINOPATHYDIAB_PROLIF datasets, 10 immune cell types were obtained. These include HLA DR on dendritic cells, HLA DR on myeloid dendritic cells, HLA DR on CD14+ CD16- monocytes, HLA DR on CD33+ HLA DR+ CD14- cells, CD4 on CD39+ activated CD4 regulatory T cells, HLA DR on CD14+ monocytes, HLA DR on CD33+ HLA DR+ CD14dim, and HLA DR on CD33- HLA DR+. These eight immune cell types may be risk factors for DR. The other two immune cell types may act as protective factors for PDR: HLA DR on CD14- CD16- and SSC-A on CD4+ T cells (Table 2B).

3.3 Sensitivity analysis

According to the merging of the two DR datasets, Cochran's Q *P*-value revealed no heterogeneity among SNPs in DR and immune cell HLA DR on CD14-CD16- or HLA DR on CD33+ HLA DR+ CD14- or DR (*P* > 0.05, Table 3A). Furthermore, the MR−Egger intercept ruled out the possibility of horizontal pleiotropy for these associations. The LOO sensitivity analysis revealed that no individual SNP disproportionately affected the overall estimates (Figure 4). Additionally, scatter plots and funnel plots also indicated the stability of the results (Figure 4).

After merging the two PDR datasets, we detected no heterogeneity in the Cochran's Q *P*- value among the SNPs of PDR and immune CD4+ T cells among the CD39+ activated CD4+ regulatory T cells or HLA DR among the CD33+ HLA DR+ CD14- cells (*P* > 0.05, Table 3B). Furthermore, the MR−Egger intercept ruled out the possibility of horizontal pleiotropy for these associations. The LOO sensitivity analysis revealed that no individual SNP disproportionately affected the overall estimates

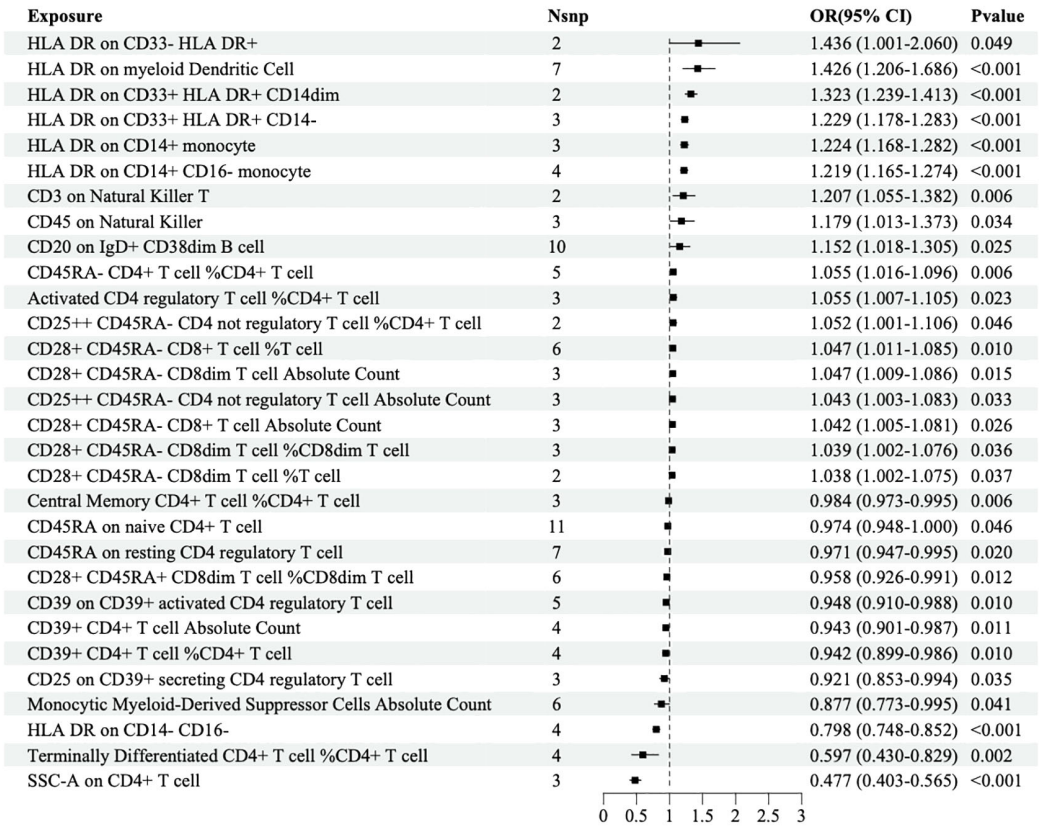


FIGURE 2
Forest map of MR causal effect between immune cells and DR (finn-b-DM_RETINOPATHY).

(Figure 5). In addition, the scatter plots and funnel plots also indicated the stability of the results (Figure 5).

In this study, we obtained seven immune cell features that were causally related to DR and PDR from two datasets. In the power calculations, the power of all the above immune cell features was >0.99, indicating that this study has sufficient statistical power (Table 3; Supplementary Power-1; Supplementary Power-2).

3.4 Meta-analysis

The MR data for immune cells from two DR patient datasets were merged through meta-analysis. If Cochran's Q P value was <0.05, the random effects model was adopted. We identified six immune cells that have a risk effect on DR, namely, HLA DR on CD14+ CD16- monocytes, HLA DR on CD14+ monocytes, HLA DR on CD33- HLA DR+, HLA DR on CD33+ HLA DR+ CD14-, HLA DR on CD33+ HLA DR+ CD14dim, and HLA DR on myeloid dendritic cells. Additionally, we found four immune cells that have a protective effect against DR including HLA-DR on CD14- CD16-, monocytic myeloid-derived suppressor cell absolute count, and SSC-A on CD4+ T cells and terminally differentiated CD4+ T cells (Supplementary DR-meta).

The MR data for immune cells from two PDR patient datasets were merged through meta-analysis. We identified a risk effect of eight immune cells on PDR, including CD4+ on CD39+ activated CD4 regulatory T cells, HLA DR on CD14+ CD16- monocytes, HLA DR on CD14+ monocytes, HLA DR on CD33- HLA DR+, HLA DR on CD33+ HLA DR+ CD14-, HLA DR on CD33+ HLA DR+ CD14dim, HLA DR on dendritic cells, and HLA DR on myeloid dendritic cells. Additionally, we discovered the protective effects of two immune cell types on DR, namely, HLA DR on CD14- CD16- T cells and SSC-A on CD4+ T cells (Supplementary PDR-meta).

The risk factors associated with the two phenotypes identified from the four datasets included six immune cell types: HLA-DR on CD14+ CD16 monocytes, HLA-DR on CD14+ monocytes, HLA-DR on CD33-HLA-DR+, HLA-DR on CD33+ HLA-DR+ CD14-, HLA-DR on CD33+ HLA-DR+ CD14dim, and HLA-DR on myeloid dendritic cells. Among the two phenotypes identified from the four datasets, HLA-DR to CD14-CD16- and SSC-A to CD4+ T cells were protective factors. Figure 6 shows the HLA-DR on CD33+ cells, HLA-DR+CD14- cells, and HLA-DR on CD14-CD16- cells.

3.5 SNP annotation

Immune cell SNPs strongly associated with DR were annotated, and 10 genes potentially connected with PDR were identified. HLA-DPA1, CD33, HLA-DOB, and NEK7 may serve as protective factors

10.3389/fimmu.2024.1381002

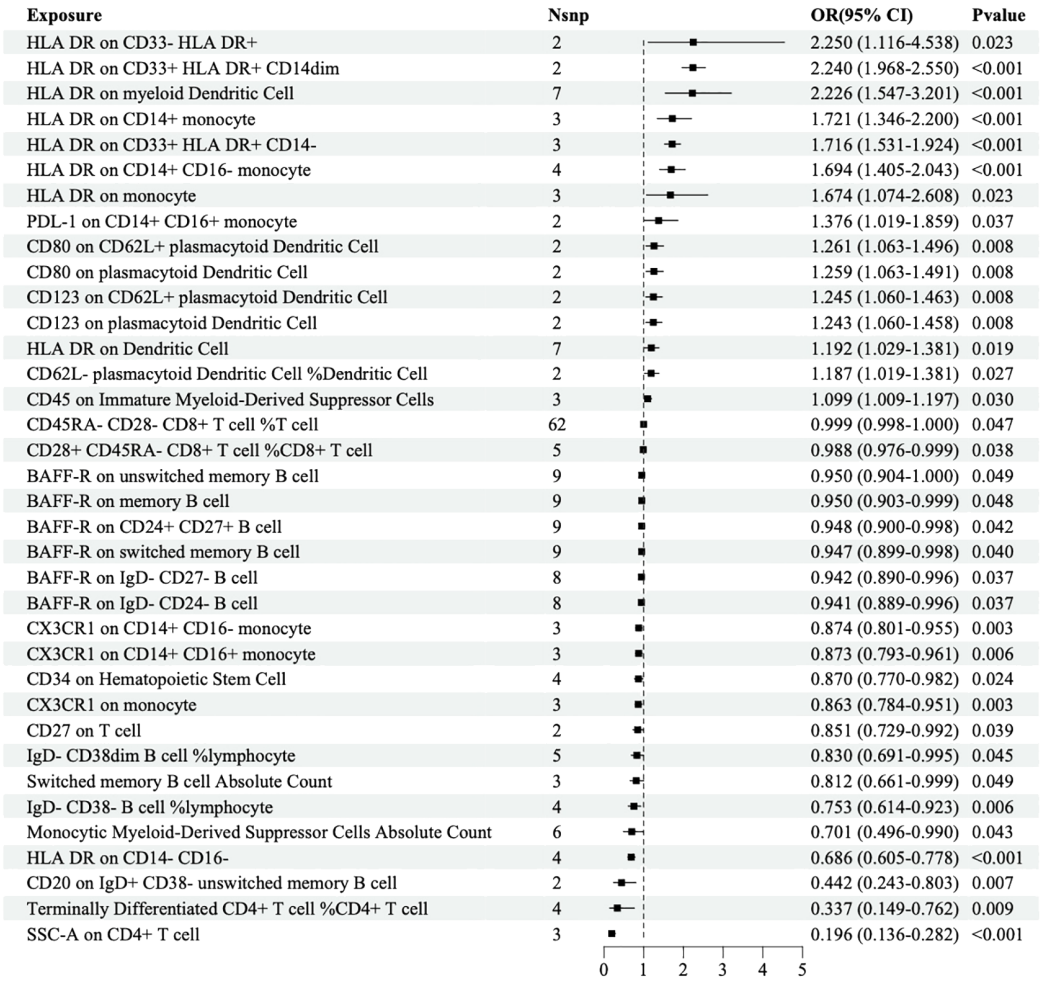


FIGURE 3
Forest map of MR causal effect between immune cells and DR (finn-b-H7_RETINOPATHYDIAB).

for DR, while TSBP1-AS1, LYZ, ENSG00000233183, MICB, GABBR1, and FCGR3A may act as risk factors for DR (Table 4A).

Immune cell SNPs strongly associated with PDR were annotated, revealing 12 genes potentially related to PDR. CD4, RPL3P2, LYZ, TSBP1-AS1, ENSG00000233183, MICB, TVP23A, GABBR1, FCGR3A, and CIITA may be risk factors for PDR, however, HLA-DPA1 and TSBP1-AS1 might provide protection against it (Table 4B).

4 Discussion

A previous study revealed that immune system disorders and inflammation play important roles in the pathogenesis of DR. Further research into the specific role of immune mechanisms in DR, as well as the identification of more specific and sensitive biomarkers, will provide a new foundation and strategies for the early clinical diagnosis and treatment of DR (18). This MR study adds to the evidence supporting a causal connection between immune cells and DR. In DR, immune cells HLA DR on CD33+ HLA DR+ CD14- and HLA DR on CD33+ HLA DR+ CD14dim may be risk factors, while immune cells HLA DR on CD14-CD16- may be protective. In PDR, CD4+ T cells on CD39+ active CD4 regulatory T cells, HLA DR cells on CD14+ monocytes, and HLA DR cells on CD33+ HLA DR+ CD14dim may be risk factors. Immune cells, specifically those harboring SSC-A on CD4+ T cells, may provide protection. The findings of this study indicate that immune cells HLA DR on CD33+ HLA DR+ CD14- and HLA DR+ CD14- have a causative influence in both the datasets of DR and PDR, which may increase the risk of developing DR. Both types express molecules known as human leukocyte antigen (HLA) -driven receptors that are important markers for antigen-presenting cells (APCs). However, CD33+ cells are primarily present in monocytes or macrophages where they play a role in innate immunity and inflammation by exhibiting phagocytic activity along with cytokine production at sites experiencing

10.3389/fimmu.2024.1381002

TABLE 2A MR results of causal links.

Data source	Classification	Trait type	Panel	Nsnp	Methods	OR (95%CI)	P-value	FDR	Power
finn-b-DM_RETINOPATHY	HLA DR on myeloid Dendritic Cell	MFI	cDC	7	Inverse variance weighted	1.426 (1.206-1.686)	3.32E-05	2.45E-03	1.000
	HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	4	Inverse variance weighted	1.219 (1.165-1.274)	4.41E-18	1.14E-15	1.000
	HLA DR on CD14- CD16-	MFI	Monocyte	4	Inverse variance weighted	0.798 (0.748-0.852)	1.19E-11	1.02E-09	1.000
	HLA DR on CD14 + monocyte	MFI	Monocyte	3	Inverse variance weighted	1.224 (1.168-1.282)	2.38E-17	3.07E-15	1.000
	SSC-A on CD4+ T cell	Morphological parameter	TBNK	3	Inverse variance weighted	0.477 (0.403-0.565)	1.04E-17	1.78E-15	1.000
	HLA DR on CD33+ HLA DR+ CD14-	MFI	Myeloid cell	3	Inverse variance weighted	1.229 (1.178-1.283)	1.91E-21	9.86E-19	1.000
	HLA DR on CD33+ HLA DR+ CD14dim	MFI	Myeloid cell	2	Inverse variance weighted	1.323 (1.239-1.413)	8.11E-17	8.37E-15	1.000
finn-b-H7_RETINOPATHYDIAB	HLA DR on myeloid Dendritic Cell	MFI	cDC	7	Inverse variance weighted	2.226 (1.547-3.201)	1.60E-05	1.18E-03	1.000
	HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	4	Inverse variance weighted	1.694 (1.405-2.043)	3.50E-08	3.61E-06	1.000
	HLA DR on CD14- CD16-	MFI	Monocyte	4	Inverse variance weighted	0.686 (0.605-0.778)	4.49E-09	5.79E-07	1.000
	HLA DR on CD14 + monocyte	MFI	Monocyte	3	Inverse variance weighted	1.721 (1.346-2.200)	1.47E-05	1.18E-03	1.000
	SSC-A on CD4+ T cell	Morphological parameter	TBNK	3	Inverse variance weighted	0.196 (0.136-0.282)	1.63E-18	2.81E-16	1.000
	HLA DR on CD33+ HLA DR+ CD14-	MFI	Myeloid cell	3	Inverse variance weighted	1.716 (1.531-1.924)	1.89E-20	4.86E-18	1.000
	HLA DR on CD33+ HLA DR+ CD14dim	MFI	Myeloid cell	2	Inverse variance weighted	2.240 (1.968-2.550)	2.99E-34	1.54E-31	1.000

TABLE 2B MR results of causal links.

Data source	Classification	Trait type	Panel	Nsnp	Methods	OR (95% CI)	P-value	FDR	Power
finn-b-DM_RETINA_PROLIF	HLA DR on myeloid Dendritic Cell	MFI	cDC	7	Inverse variance weighted	1.690 (1.327-2.153)	2.15E-05	1.59E-03	1.000

(Continued)

10.3389/fimmu.2024.1381002

TABLE 2B Continued

Data source	Classification	Trait type	Panel	Nsnp	Methods	OR (95% CI)	P-value	FDR	Power
	HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	4	Inverse variance weighted	1.369 (1.245-1.505)	8.20E-11	1.06E-08	1.000
	HLA DR on CD14 + monocyte	MFI	Monocyte	3	Inverse variance weighted	1.380 (1.218-1.565)	4.59E-07	3.94E-05	1.000
	CD4 on CD39+ activated CD4 regulatory T cell	MFI	Treg	3	Inverse variance weighted	1.224 (1.107-1.354)	8.53E-05	5.50E-03	0.994
	SSC-A on CD4+ T cell	Morphological parameter	TBNK	3	Inverse variance weighted	0.348 (0.270-0.449)	3.78E-16	6.50E-14	1.000
	HLA DR on CD33+ HLA DR+ CD14-	MFI	Myeloid cell	3	Inverse variance weighted	1.383 (1.305-1.466)	9.65E-28	3.62E-25	1.000
	HLA DR on CD33+ HLA DR+ CD14dim	MFI	Myeloid cell	2	Inverse variance weighted	1.599 (1.469-1.740)	1.40E-27	3.62E-25	1.000
finn-b-H7_RETINOPATHYDIAB_PROLIF	HLA DR on myeloid Dendritic Cell	MFI	cDC	7	Inverse variance weighted	2.603 (1.645-4.121)	4.46E-05	3.83E-03	1.000
	HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	4	Inverse variance weighted	1.900 (1.536-2.350)	3.27E-09	4.22E-07	1.000
	HLA DR on CD14 + monocyte	MFI	Monocyte	3	Inverse variance weighted	1.943 (1.474-2.560)	2.40E-06	2.48E-04	1.000
	CD4 on CD39+ activated CD4 regulatory T cell	MFI	Treg	3	Inverse variance weighted	1.532 (1.202-1.954)	5.73E-04	4.23E-02	0.992
	SSC-A on CD4+ T cell	Morphological parameter	TBNK	3	Inverse variance weighted	0.142 (0.100-0.202)	8.73E-28	4.50E-25	1.000
	HLA DR on CD33+ HLA DR+ CD14-	MFI	Myeloid cell	3	Inverse variance weighted	1.879 (1.525-2.315)	3.15E-09	4.22E-07	1.000
	HLA DR on CD33+ HLA DR+ CD14dim	MFI	Myeloid cell	2	Inverse variance weighted	2.654 (2.162-3.257)	1.06E-20	2.73E-18	1.000

inflammation. However, lymphocytes more widely express another type called CD14-, which has stronger associations with adaptive immune responses potentially contributing to diseases or immune response regulation.

Numerous studies have demonstrated that regulatory T cells and monocytes play a significant role in the pathogenesis of DR. The activation of immunoinflammatory cells and proinflammatory substances in the retinal tissue of DR patients contributes to the occurrence and progression of DR (19–22). Leukocyte adhesion stasis; neutrophil increase; abnormal expression of T cells, B lymphocytes, mononuclear/macrophages, and other immune cells; elevated concentrations of inflammatory and proangiogenic factors; and increased levels of anti-pericytes and anti-endothelial cell antibodies were found in the serum, vitreous, and retinal tissues of DR animal models and patients (23–25). YUAN et al. used a gene expression microarray for immunoinfiltration analysis. They found that in DR samples, there was significant overexpression ($P<0.05$) of seven types of immune cells: original B cells, plasma cells, memory CD4+ T cells, regulatory T cells (Tregs), MO macrophages, M1 macrophages, and neutrophils ($P<0.05$). The activated memory

10.3389/fimmu.2024.1381002

TABLE 3A Evaluation of heterogeneity and pleiotropy.

Data source	Classification	Trait type	Panel	Nsnp	Heterogeneity			Horizontal pleiotropy		
					I^2 (%)	Cochran's Q	P-value	Egger intercept	SE	P-value
finn-b-DM_RETINOPATHY	HLA DR on myeloid Dendritic Cell	MFI	cDC	7	97	178.303	<0.001	-0.137	0.075	0.127
	HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	4	0	1.611	0.657	-0.018	0.050	0.751
	HLA DR on CD14- CD16-	MFI	Monocyte	4	1	3.016	0.389	-0.035	0.046	0.521
	HLA DR on CD14 + monocyte	MFI	Monocyte	3	0	1.634	0.442	-0.052	0.066	0.574
	SSC-A on CD4+ T cell	Morphological parameter	TBNK	3	69	6.544	0.038	0.401	0.720	0.677
	HLA DR on CD33+ HLA DR+ CD14-	MFI	Myeloid cell	3	0	1.089	0.580	0.018	0.031	0.662
	HLA DR on CD33+ HLA DR+ CD14dim	MFI	Myeloid cell	2	0	0.164	0.685	-	-	-
finn-b-H7_RETINOPATHYDIAB	HLA DR on myeloid Dendritic Cell	MFI	cDC	7	97	218.238	<0.001	-0.288	0.166	0.144
	HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	4	78	13.641	0.003	-0.297	0.145	0.177
	HLA DR on CD14- CD16-	MFI	Monocyte	4	0	2.821	0.420	-0.098	0.082	0.355
	HLA DR on CD14 + monocyte	MFI	Monocyte	3	86	14.335	<0.001	-0.476	0.128	0.168
	SSC-A on CD4+ T cell	Morphological parameter	TBNK	3	74	7.610	0.022	1.531	0.874	0.330
	HLA DR on CD33+ HLA DR+ CD14-	MFI	Myeloid cell	3	46	3.732	0.155	-0.109	0.061	0.325
	HLA DR on CD33+ HLA DR+ CD14dim	MFI	Myeloid cell	2	0	0.042	0.837	-	-	-

TABLE 3B Evaluation of heterogeneity and pleiotropy.

Data source	Classification	Trait type	Panel	Nsnp	Heterogeneity			Horizontal pleiotropy		
					I^2 (%)	Cochran's Q	P-value	Egger intercept	SE	P-value
finn-b-DM_RETINA_PROLIF	HLA DR on myeloid Dendritic Cell	MFI	cDC	7	97	228.471	<0.001	-0.198	0.109	0.128
	HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	4	63	8.183	0.042	-0.143	0.080	0.217
	HLA DR on CD14 + monocyte	MFI	Monocyte	3	77	8.758	0.013	-0.241	0.084	0.213
	CD4 on CD39+ activated CD4 regulatory T cell	MFI	Treg	3	0	1.522	0.467	-0.223	0.289	0.583
	SSC-A on CD4+ T cell	Morphological parameter	TBNK	3	77	8.869	0.012	1.134	0.480	0.255
	HLA DR on CD33+ HLA DR+ CD14-	MFI	Myeloid cell	3	12	2.284	0.319	-0.032	0.051	0.647

(Continued)

TABLE 3B Continued

Data source	Classification	Trait type	Panel	Nsnp	Heterogeneity			Horizontal pleiotropy		
					I^2 (%)	Cochran's Q	P-value	Egger intercept	SE	P-value
	HLA DR on CD33+ HLA DR + CD14dim	MFI	Myeloid cell	2	0	0.359	0.549	-	-	-
finn-b-H7_RETINOPATHYDIAB_PROLIF	HLA DR on myeloid Dendritic Cell	MFI	cDC	7	96	136.103	<0.001	-0.371	0.207	0.133
	HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	4	57	7.015	0.071	-0.350	0.154	0.152
	HLA DR on CD14 + monocyte	MFI	Monocyte	3	72	7.234	0.027	-0.523	0.202	0.235
	CD4 on CD39+ activated CD4 regulatory T cell	MFI	Treg	3	0	0.174	0.917	0.265	0.696	0.769
	SSC-A on CD4+ T cell	Morphological parameter	TBNK	3	29	2.798	0.247	1.644	1.009	0.350
	HLA DR on CD33+ HLA DR+ CD14-	MFI	Myeloid cell	3	60	4.980	0.083	-0.204	0.096	0.281
	HLA DR on CD33+ HLA DR + CD14dim	MFI	Myeloid cell	2	0	0.510	0.475	-	-	-

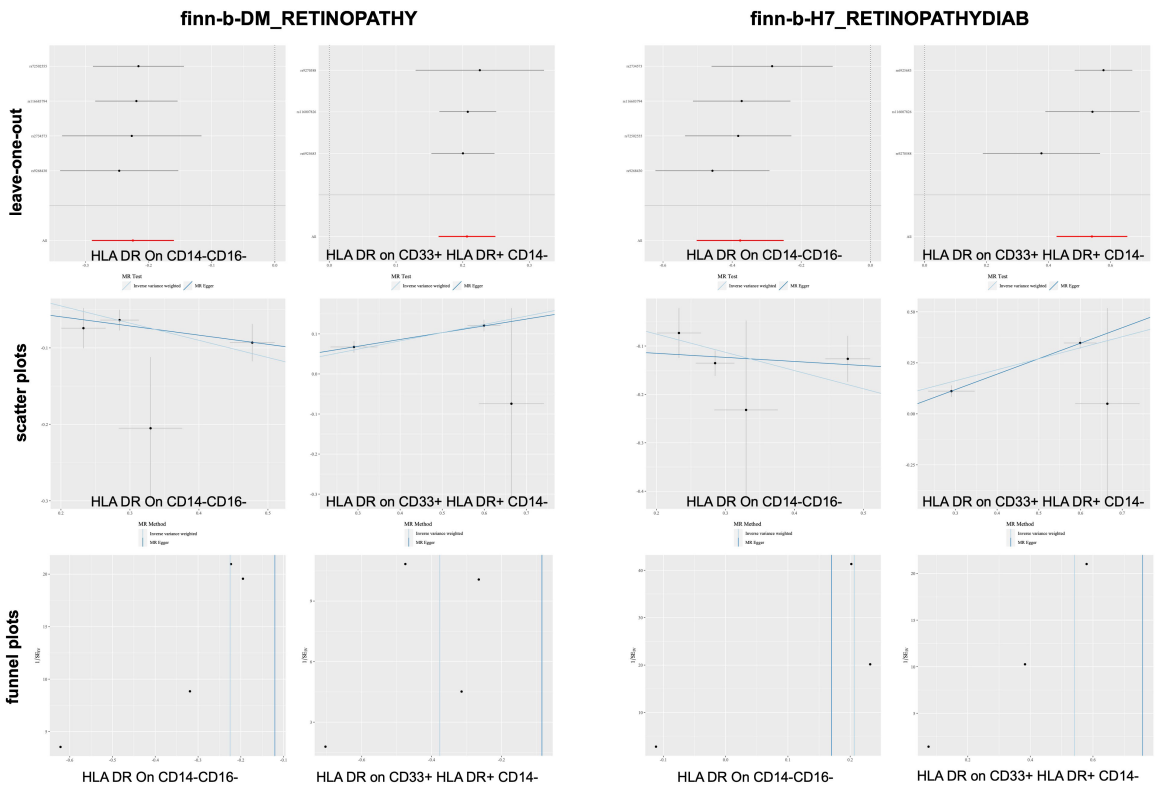


FIGURE 4
Sensitivity analysis for DR datasets.

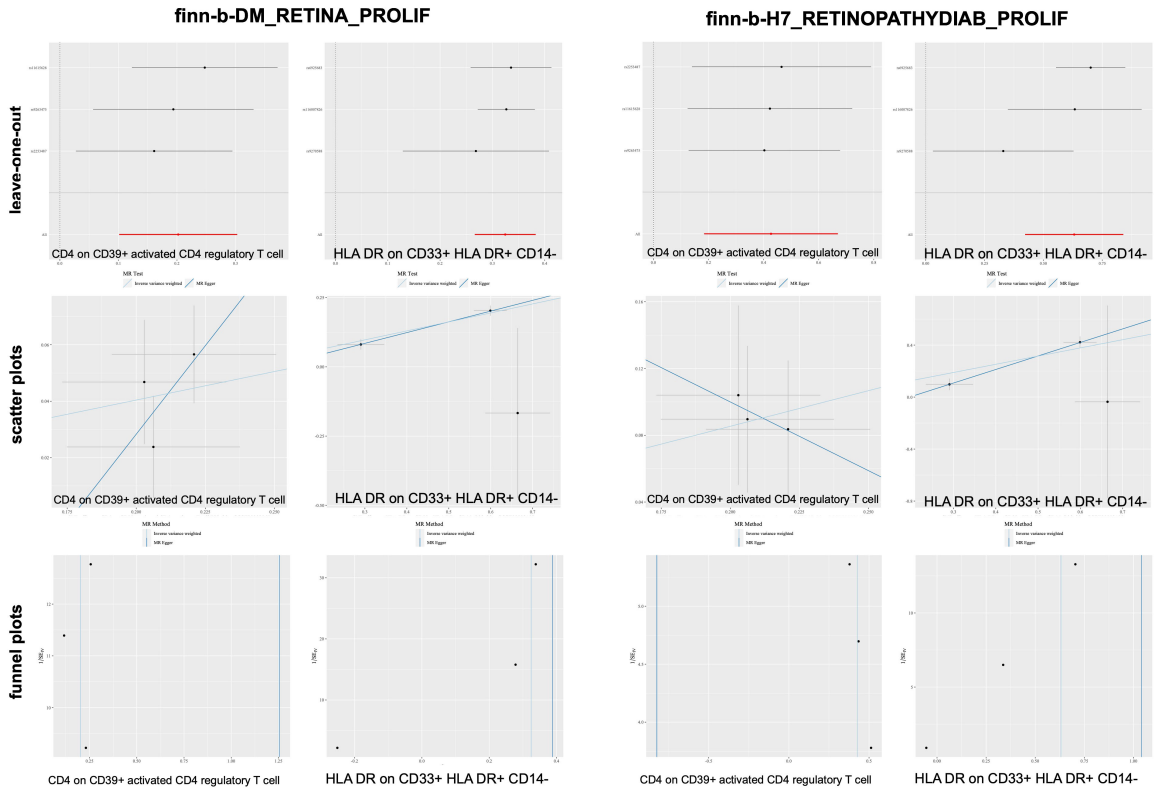


FIGURE 5
Sensitivity analysis for PDR datasets.

CD4+ T-cell module had the highest correlation and differential expression ($P< 0.001$). Activated NK cells showed low expression among immune cells ($P<0.05$) (26). Our study also revealed an increased risk for DR associated with HLA DR on CD14+ CD16- monocytes, HLA DR on CD33+ HLA DR+ CD14- cells, HLA DR on CD14+ monocytes, and HLA DR on CD33+ HLA DR+ CD14dim (all belonging to the monocyte population). Additionally, PDR patients showed a significant increase in CD4+ T cells on CD39+ activated CD4+ regulatory T cells. Two immune-associated target genes in DR, DLGAP5 and AURKB, were found to be enriched in pathways relevant to memory CD4+ cells. These findings suggest that DR is closely related to the activation of regulatory T cells and monocytes.

Our research revealed a significant increase in HLA-DR on CD33 +HLA-DR+CD14- cells in both DR and PDR, indicating a strong correlation between microangiopathy in DR and the activation of myeloid-derived suppressor cells (MDSCs). MDSCs are diverse cell types that can effectively suppress T cell responses. Under normal conditions, these cells develop into dendritic cells, macrophages, and granulocytes. However, in pathological conditions such as infection, inflammation, or cancer, the differentiation of these cells stops resulting in their accumulation (27–29). Initially classified as HLA-DR-CD33+ or CD14-CD11b+ cells, both of which are populations of cells with T cell inhibitory activity (30, 31), human MDSCs can be further subdivided into granulocytic CD14– and monocytic CD14+ MDSCs (32, 33). One study found that patients with type 1 diabetes mellitus (T1D) have significantly greater numbers of MDSCs in their peripheral blood with M-MDSCs (CD14+ CD33+ HLA-DR–) being the most prevalent subset of MDSCs. Compared to diabetic patients without kidney disease, diabetic patients with kidney disease had a substantial increase in the number of total MDSCs and a rise in the percentage of CD14- cells (34). An imbalance of immune active cells is directly linked to the development of DR, as evidenced by the aberrant activation and expression of immune cells in the ocular tissue of DR patients and the association with DR. There are many similarities between diabetic nephropathy (DN) and DR. DN and DR are both microvascular complications resulting from diabetes, which are complex illnesses with diverse manifestations (35). If immune cell activation is effectively inhibited, delays in the onset of DR disease can be expected.

Our study indicates that CD33 is implicated in DR. Additionally, these findings reveal a set of genetic variants associated with proangiogenic and inflammatory pathways that may contribute to the pathogenesis of DR. Further investigation into these variants is necessary and may lead to the development of novel biomarkers and new therapeutic targets for DR. Previous research has shown that

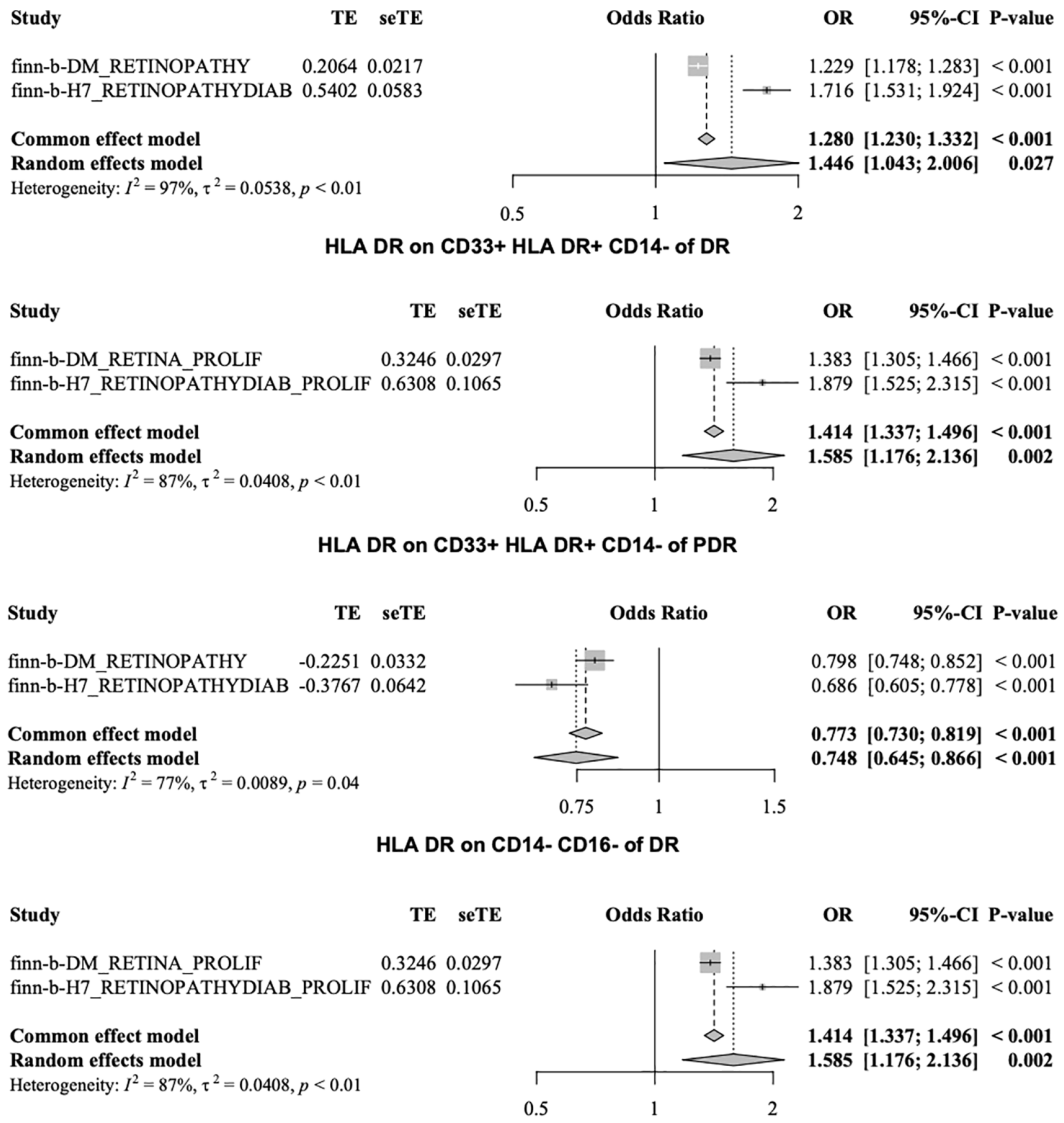


FIGURE 6
Results of meta-analysis.

genes associated with angiogenesis and inflammatory pathways play a crucial role in the onset of DR (36–39). The Diabetic Retinopathy Genomics (DRGen) study revealed the involvement of Kruppel Like Factor 17 (KLF17), Zinc Finger Protein 395 (ZNF395), Myeloid cell surface antigen (CD33), Pleckstrin Homology Domain-Containing Family G Member 5 (PLEKHG5), NK2 Homeobox 3 (*NKX2.3*), and Collagen Type XVIII Alpha 1 Chain (COL18A1) in the progression of DR. These genes have been shown to be involved in angiogenesis and inflammatory pathways (40).

In MR studies, genetic variations that are substantially associated with an exposure are used as IVs to investigate the potential causal relationship between an exposure and a specific outcome of interest. Since genetic variants are randomly assigned at conception, MR estimates are not influenced by confounding factors, reverse causality, or measurement error (41). Inference typically relies on SNPs identified as IVs in GWASs. The current study was conducted in a rather conservative manner and supported by a comprehensive sensitivity analysis due to the strong assumptions underlying MR research (42). To ensure the robustness of the results, several measures were taken. Firstly, to minimize any bias resulting from demographic variability, only European populations were included in the analysis. Secondly, considering that both disease risk factors and immune cells are complex polygenic phenotypes that can be influenced by various genetic and environmental factors simultaneously (pleiotropy), we assessed potential pleiotropic effects through LOO and examined

10.3389/fimmu.2024.1381002

TABLE 4A SNP annotation.

Classification	Trait type	Panel	SNP	Chr	Start	End	Strand	Gene_ids	Gene_names
HLA DR on CD14- CD16-	MFI	Monocyte	rs116683794	6	33066337	33066337	+	ENSG00000231389	HLA-DPA1
HLA DR on CD14- CD16-	MFI	Monocyte	rs2734573		-1	-1			
HLA DR on CD14- CD16-	MFI	Monocyte	rs72502555		-1	-1			
HLA DR on CD14- CD16-	MFI	Monocyte	rs9268430	6	32377652	32377652	+	ENSG00000225914	TSBP1-AS1
HLA DR on CD14 + CD16- monocyte	MFI	Monocyte	rs150649461		-1	-1			
HLA DR on CD14 + CD16- monocyte	MFI	Monocyte	rs1800973	12	69350234	69350234	+	ENSG00000090382	LYZ
HLA DR on CD14 + CD16- monocyte	MFI	Monocyte	rs80032720		-1	-1			
HLA DR on CD14 + CD16- monocyte	MFI	Monocyte	rs9270585		-1	-1			
HLA DR on CD14 + monocyte	MFI	Monocyte	rs1800973	12	69350234	69350234	+	ENSG00000090382	LYZ
HLA DR on CD14 + monocyte	MFI	Monocyte	rs80032720		-1	-1			
HLA DR on CD14 + monocyte	MFI	Monocyte	rs9270585		-1	-1			
HLA DR on CD33 + HLA DR + CD14-	MFI	Myeloid cell	rs116007826		-1	-1			
HLA DR on CD33 + HLA DR + CD14-	MFI	Myeloid cell	rs6925683	6	33926515	33926515	+	ENSG00000233183	ENSG00000233183
HLA DR on CD33 + HLA DR + CD14-	MFI	Myeloid cell	rs9270588		-1	-1			
HLA DR on CD33 + HLA DR + CD14dim	MFI	Myeloid cell	rs142186496	6	31505930	31505930	+	ENSG00000204516	MICB
HLA DR on CD33 + HLA DR + CD14dim	MFI	Myeloid cell	rs9270588		-1	-1			
HLA DR on myeloid dendritic cell	MFI	cDC	rs116007826		-1	-1			
HLA DR on myeloid dendritic cell	MFI	cDC	rs2858885		-1	-1			
HLA DR on myeloid dendritic cell	MFI	cDC	rs29221	6	29621347	29621347	+	ENSG00000204681	GABBR1
HLA DR on myeloid dendritic cell	MFI	cDC	rs35525122		-1	-1			

(Continued)

TABLE 4A Continued

Classification	Trait type	Panel	SNP	Chr	Start	End	Strand	Gene_ids	Gene_names
HLA DR on myeloid dendritic cell	MFI	cDC	rs55971447	1	1.62E+08	1.62E+08	+	ENSG00000203747, ENSG00000273112, ENSG00000289768	FCGR3A, ENSG00000273112, ENSG00000289768
HLA DR on myeloid dendritic cell	MFI	cDC	rs6925683	6	33926515	33926515	+	ENSG00000233183	ENSG00000233183
HLA DR on myeloid dendritic cell	MFI	cDC	rs9267650		-1	-1			
SSC-A on CD4+ T cell	Morphological parameter	TBNK	rs113243185		-1	-1			
SSC-A on CD4+ T cell	Morphological parameter	TBNK	rs148031710		-1	-1			
SSC-A on CD4+ T cell	Morphological parameter	TBNK	rs9271536		-1	-1			

"+" strand: sense strand, or coding strand.
"-" strand: antisense strand or template strand.

TABLE 4B SNP annotation.

Classification	Trait type	Panel	SNP	Chr	Start	End	Strand	Gene_ids	Gene_names
CD4 on CD39+ activated CD4 regulatory T cell	MFI	Treg	rs11615628	12	6794465	6794465	+	ENSG00000010610	CD4
CD4 on CD39+ activated CD4 regulatory T cell	MFI	Treg	rs2253487	6	31281350	31281350	+	ENSG00000227939	RPL3P2
CD4 on CD39+ activated CD4 regulatory T cell	MFI	Treg	rs9263475		-1	-1			
HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	rs150649461		-1	-1			
HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	rs1800973	12	69350234	69350234	+	ENSG00000090382	LYZ
HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	rs80032720		-1	-1			
HLA DR on CD14+ CD16- monocyte	MFI	Monocyte	rs9270585		-1	-1			
HLA DR on CD14 + monocyte	MFI	Monocyte	rs1800973	12	69350234	69350234	+	ENSG00000090382	LYZ
HLA DR on CD14 + monocyte	MFI	Monocyte	rs80032720		-1	-1			
HLA DR on CD14 + monocyte	MFI	Monocyte	rs9270585		-1	-1			
HLA DR on CD33+ HLA DR+ CD14-	MFI	Myeloid cell	rs116007826		-1	-1			
HLA DR on CD33+ HLA DR+ CD14-	MFI	Myeloid cell	rs6925683	6	33926515	33926515	+	ENSG00000233183	ENSG00000233183

(Continued)

10.3389/fimmu.2024.1381002

TABLE 4B Continued

Classification	Trait type	Panel	SNP	Chr	Start	End	Strand	Gene_ids	Gene_names
HLA DR on CD33+ HLA DR+ CD14-	MFI	Myeloid cell	rs9270588		-1	-1			
HLA DR on CD33+ HLA DR+ CD14dim	MFI	Myeloid cell	rs142186496	6	31505930	31505930	+	ENSG00000204516	MICB
HLA DR on CD33+ HLA DR+ CD14dim	MFI	Myeloid cell	rs9270588		-1	-1			
HLA DR on myeloid dendritic cell	MFI	cDC	rs116007826		-1	-1			
HLA DR on myeloid dendritic cell	MFI	cDC	rs2858885		-1	-1			
HLA DR on myeloid dendritic cell	MFI	cDC	rs29221	6	29621347	29621347	+	ENSG00000204681	GABBR1
HLA DR on myeloid dendritic cell	MFI	cDC	rs35525122		-1	-1			
HLA DR on myeloid dendritic cell	MFI	cDC	rs55971447	1	1.62E+08	1.62E+08	+	ENSG00000203747, ENSG00000273112, ENSG00000289768	FCGR3A, ENSG00000273112, ENSG00000289768
HLA DR on myeloid dendritic cell	MFI	cDC	rs6925683	6	33926515	33926515	+	ENSG00000233183	ENSG00000233183
HLA DR on myeloid dendritic cell	MFI	cDC	rs9267650		-1	-1			
SSC-A on CD4+ T cell	Morphological parameter	TBNK	rs113243185		-1	-1			
SSC-A on CD4+ T cell	Morphological parameter	TBNK	rs148031710		-1	-1			
SSC-A on CD4+ T cell	Morphological parameter	TBNK	rs9271536		-1	-1			

"+" strand: sense strand, or coding strand.

the intercept of MR-Egger regression. These approaches consistently yielded results suggesting reliable causal estimations. Data from European populations were utilized in this study to select a representative sample. By using MR methodology, it is possible to minimize the impact of reverse causation and confounding variables on estimation accuracy while producing trustworthy causal effect estimates based on observational research findings. Furthermore, GWAS data with large sample sizes were employed for these studies which significantly enhanced test efficiency compared to small-sample models relying on individual data points.

This study has certain limitations. First, there will inevitably be batch differences across the various datasets analyzed in this study due to its use of a public database. There are issues with the cohesiveness of integrating multiple databases in this study, and further efforts are needed to improve the accuracy of causal inference. Second, the research was limited to individuals with European ancestry, making it challenging to generalize the findings to other demographic groups. Third, residual and unmeasured confounders may still exist as the study was unable to determine whether demographic stratification and other potential confounders had an impact on its findings.

5 Conclusion

This study's findings emphasized the complex network of connections between the immune system and DR, as it demonstrated causal relationships between various immune cells and DR through MR analysis. HLA-DR on CD14+ CD16 monocytes, HLA-DR on CD14+ monocytes, HLA-DR on CD33-HLA-DR+, HLA-DR on CD33+ HLA-DR+ CD14-, HLA-DR on CD33+ HLA-DR+ CD14dim, and HLA-DR on myeloid dendritic cells may increase the risk of DR. Additionally, HLA-DR to CD14-CD16- and SSC-A to CD4+ T cells may be protective factors against DR. These findings could open new avenues for investigating the biological causes of DR and pave the way for research into earlier intervention and treatment.

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.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

YY: Writing – original draft, Methodology. LD: Writing – original draft, Formal analysis. HG: Writing – original draft, Validation. LY: Writing – original draft, Visualization. ZX: Writing – review & editing. ZGL: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

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.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1381002/full#supplementary-material

References

1. Cheung N, Chee ML, Klein R, Klein BE, Shea S, Cotch MF, et al. Incidence and progression of diabetic retinopathy in a multi-ethnic US cohort: the Multi-Ethnic Study of Atherosclerosis. *Br J Ophthalmol*. (2022) 106:1264–8. doi: 10.1136/bjophthalmol-2021-318992

2. Sabanayagam C, Sultana R, Banu R, Rim T, Tham YC, Mohan S, et al. Association between body mass index and diabetic retinopathy in Asians: the Asian Eye Epidemiology Consortium (AEEC) study. *Br J Ophthalmol*. (2022) 106:980–6. doi: 10.1136/bjophthalmol-2020-318208

3. Nishi K, Nishitsuka K, Yamamoto T, Yamashita H. Factors correlated with visual outcomes at two and four years after vitreous surgery for proliferative diabetic retinopathy. *PloS One*. (2021) 16:e0244281. doi: 10.1371/journal.pone.0244281

4. Wang S, Pan X, Jia B, Chen S. Exploring the correlation between the systemic immune inflammation index (SII), systemic inflammatory response index (SIRI), and type 2 diabetic retinopathy. *Diabet Metab Syndrome Obes*. (2023) 16:3827–36. doi: 10.2147/DMSO.S437580

5. Malhotra NA, Muste J, Hom GL, Conti TF, Greenlee TE, Singh RP. Race and socioeconomic status in anti-VEGF treatment of diabetic macular edema. *Ophthalmic Surgery Lasers Imaging Retina*. (2021) 52:578–85. doi: 10.3928/23258160-20211018-01

6. Żuchnik M, Rybkowska A, Szczuraszek P, Szczuraszek H, Bętkowska P, Radulski J, et al. Olko, Type 2 diabetes-factors of occurrence and its complications. *Qual Sport*. (2023) 10:32–40. doi: 10.12775/QS.2023.10.01.003

7. Yang M, Wang X, Han Y, Li C, Wei L, Yang J, et al. Targeting the NLRP3 inflammasome in diabetic nephropathy. *Curr Medicinal Chem*. (2021) 28:8810–24. doi: 10.2174/0929867328666210705153109

8. Wei H, Gu Q. SOX4 promotes high-glucose-induced inflammation and angiogenesis of retinal endothelial cells by activating NF-κB signaling pathway. *Open Life Sci*. (2022) 17:393–400. doi: 10.1515/biol-2022-0045

9. Alchujyan N, Hovhannisyan M, Movsesyan N, Melkonyan A, Shaboyan V, Aghajanova Y, et al. Sexual dimorphism in alternative metabolic pathways of L-arginine in circulating leukocytes in young people with type 1 diabetes mellitus. *Endocrine Res*. (2021) 46:149–59. doi: 10.1080/07435800.2021.1920608

10. Yoshikawa M, Asaba K, Nakayama T. Causal effect of atrial fibrillation/flutter on chronic kidney disease: A bidirectional two-sample Mendelian randomization study. *PloS One*. (2021) 16:e0261020. doi: 10.1371/journal.pone.0261020

11. Lee K, Lim C-Y. Mendelian randomization analysis in observational epidemiology. *J Lipid Atheroscl*. (2019) 8:67–77. doi: 10.12997/jla.2019.8.2.67

12. Consortium GP. A global reference for human genetic variation. *Nature*. (2015) 526:68. doi: 10.1038/nature15393

13. Burgess S, Small DS, Thompson SG. A review of instrumental variable estimators for Mendelian randomization. *Stat Methods Med Res*. (2017) 26:2333–55. doi: 10.1177/0962280215597579

14. Wang C, Zhu D, Zhang D, Zuo X, Yao L, Liu T, et al. Causal role of immune cells in schizophrenia: Mendelian randomization (MR) study. *BMC Psychiatry*. (2023) 23:590. doi: 10.1186/s12888-023-05081-4

15. Brion M-JA, Shakhbazov K, Visscher PM. Calculating statistical power in Mendelian randomization studies. *Int J Epidemiol*. (2013) 42:1497–501. doi: 10.1093/ije/dyt179

16. Papadimitriou N, Dimou N, Tsilidis KK, Banbury B, Martin RM, Lewis SJ, et al. Physical activity and risks of breast and colorectal cancer: a Mendelian randomisation analysis. *Nat Commun*. (2020) 11:597. doi: 10.1038/s41467-020-14389-8

17. Burgess S, Thompson SG. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur J Epidemiol*. (2017) 32:377–89. doi: 10.1007/s10654-017-0255-x

18. Lechner J, O'Leary OE, Stitt AW. The pathology associated with diabetic retinopathy. *Vision Res*. (2017) 139:7–14. doi: 10.1016/j.visres.2017.04.003

19. Rangasamy S, McGuire PG, Franco Nitta C, Monickaraj F, Oruganti SR, Das A. Chemokine mediated monocyte trafficking into the retina: role of inflammation in alteration of the blood-retinal barrier in diabetic retinopathy. *PloS One*. (2014) 9:e108508. doi: 10.1371/journal.pone.0108508

20. Urbančič M, Štunf Š., MilutinovićŽivin A, Petrovič D. GlobočnikPetrovič, Epiretinal membrane inflammatory cell density might reflect the activity of proliferative diabetic retinopathy. *Invest Ophthalmol Visual Sci*. (2014) 55:8576–82. doi: 10.1167/iovs.13-13634

21. Takeuchi M, Sato T, Tanaka A, Muraoka T, Taguchi M, Sakurai Y, et al. Elevated levels of cytokines associated with Th2 and Th17 cells in vitreous fluid of proliferative diabetic retinopathy patients. *PloS One*. (2015) 10:e0137358. doi: 10.1371/journal.pone.0137358

22. Chernykh V, Varvarinsky E, Smirnov E, Chernykh D, Trunov AN. Proliferative and inflammatory factors in the vitreous of patients with proliferative diabetic retinopathy. *Indian J Ophthalmol*. (2015) 63:33. doi: 10.4103/0301-4738.151464

23. Sasongko M, Wong T, Jenkins A, Nguyen T, Shaw J, Wang J. Circulating markers of inflammation and endothelial function, and their relationship to diabetic retinopathy. *Diabetic Med*. (2015) 32:686–91. doi: 10.1111/dme.12640

24. Wang R-t, Zhang J-r, Li Y, Liu T, Yu K-j. Neutrophil–lymphocyte ratio is associated with arterial stiffness in diabetic retinopathy in type 2 diabetes. *J Diabetes its Complicat*. (2015) 29:245–9. doi: 10.1016/j.jdiacomp.2014.11.006

25. He J, Wang H, Liu Y, Li W, Kim D, Huang H. Blockade of vascular endothelial growth factor receptor 1 prevents inflammation and vascular leakage in diabetic retinopathy. *J Ophthalmol*. (2015) 2015. doi: 10.1155/2015/605946

26. L.-H. YUAN L.-J. ZHANG, X. LIU, Y.-Y. QI. Identification of key immune related genes in diabetes retinopathy based on weighted gene co-expression network. *Int Eye Sci*. (2023) 16:1343–51. doi: 10.3980/j.issn.1672-5123.2023.8.20

27. Filipazzi P, Huber V, Rivoltini L. Phenotype, function and clinical implications of myeloid-derived suppressor cells in cancer patients. *Cancer Immunology Immunother*. (2012) 61:255–63. doi: 10.1007/s00262-011-1161-9

28. Atretkhany K-SN, Drutskaya M. Myeloid-derived suppressor cells and proinflammatory cytokines as targets for cancer therapy. *Biochem (Moscow)*. (2016) 81:1274–83. doi: 10.1134/S0006297916110055

29. Youn JI, Gabrilovich DI. The biology of myeloid-derived suppressor cells: the blessing and the curse of morphological and functional heterogeneity. *Eur J Immunol*. (2010) 40:2969–75. doi: 10.1002/eji.201040895

30. Gabrilovich DI, Nagaraj S. Myeloid-derived suppressor cells as regulators of the immune system. *Nat Rev Immunol*. (2009) 9:162–74. doi: 10.1038/nri2506

31. Talmadge JE, Gabrilovich DI. History of myeloid-derived suppressor cells. *Nat Rev Cancer*. (2013) 13:739–52. doi: 10.1038/nrc3581

32. Greten TF, Manns MP, Korangy F. Myeloid derived suppressor cells in human diseases. *Int Immunopharmacol*. (2011) 11:802–7. doi: 10.1016/j.intimp.2011.01.003

33. Ning G, She L, Lu L, Liu Y, Zeng Y, Yan Y, et al. Analysis of monocytic and granulocytic myeloid-derived suppressor cells subsets in patients with hepatitis C virus infection and their clinical significance. *BioMed Res Int*. (2015) 2015. doi: 10.1155/2015/385378

34. Hassan M, Raslan HM, Eldin HG, Mahmoud E, Alm-elhuda Abd Elwajed H. CD33+ HLA-DR–myeloid-derived suppressor cells are increased in frequency in the peripheral blood of type1 diabetes patients with predominance of CD14+ Subset. *Open Access Macedonian J Med Sci*. (2018) 6:303. doi: 10.3889/oamjms.2018.080

35. Krolewski AS, Skupien J, Rossing P, Warram JH. Fast renal decline to end-stage renal disease: an unrecognized feature of nephropathy in diabetes. *Kidney Int*. (2017) 91:1300–11. doi: 10.1016/j.kint.2016.10.046

36. Horikawa N, Abiko K, Matsumura N, Hamanishi J, Baba T, Yamaguchi K, et al. Expression of vascular endothelial growth factor in ovarian cancer inhibits tumor immunity through the accumulation of myeloid-derived suppressor cells. *Clin Cancer Res*. (2017) 23:587–99. doi: 10.1158/1078-0432.CCR-16-0387

37. Iwanicki MP, Brugge JS. Transcriptional regulation of metastatic [Id] entity by KLF17. *Genome Biol*. (2009) 10:1–3. doi: 10.1186/gb-2009-10-11-244

38. Maystadt I, Rezsöhazy R, Barkats M, Duque S, Vannuffel P, Remacle S, et al. The nuclear factor κB–activator gene PLEKHG5 is mutated in a form of autosomal recessive lower motor neuron disease with childhood onset. *Am J Hum Genet*. (2007) 81:67–76. doi: 10.1086/518900

39. Murat A, Migliavacca E, Hussain SF, Heimberger AB, Desbaillets I, Hamou M-F, et al. Modulation of angiogenic and inflammatory response in glioblastoma by hypoxia. *PloS One*. (2009) 4:e5947. doi: 10.1371/journal.pone.0005947

40. Cabrera AP, Mankad RN, Marek L, Das R, Rangasamy S, Monickaraj F, et al. Genotypes and phenotypes: a search for influential genes in diabetic retinopathy. *Int J Mol Sci*. (2020) 21:2712. doi: 10.3390/ijms21082712

41. Yavorska OO, Burgess S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int J Epidemiol*. (2017) 46:1734–9. doi: 10.1093/ije/dyx034

42. Larsson SC. Mendelian randomization as a tool for causal inference in human nutrition and metabolism. *Curr Opin lipidol*. (2021) 32:1–8. doi: 10.1097/MOL.0000000000000721

frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 16 September 2024
DOI 10.3389/fgene.2024.1335839

OPEN ACCESS

EDITED BY
Alex Tsoi,
University of Michigan, United States

REVIEWED BY
Yuxiang Fei,
China Pharmaceutical University, China
Jing-guo Chen,
The Second Affiliated Hospital of Xi'an Jiaotong University, China

*CORRESPONDENCE
Naxin Liu,
liunaxin68@wzhospital.cn
Shaoliang Han,
slhan88@126.com
Sen Li,
lzz1840@wmu.edu.cn

RECEIVED 15 January 2024
ACCEPTED 22 August 2024
PUBLISHED 16 September 2024

CITATION
Xie W, Jiang H, Chen Y, Yu Z, Song Y, Zhang H, Li S, Han S and Liu N (2024) Relationship between type 1 diabetes and autoimmune diseases in european populations: A two-sample Mendelian randomization study. *Front. Genet.* 15:1335839.
doi: 10.3389/fgene.2024.1335839

Relationship between type 1 diabetes and autoimmune diseases in european populations: A two-sample Mendelian randomization study

Weidong Xie[1], Haojie Jiang[1], Yao Chen[2], Zhaojie Yu[3], Yaoyu Song[3], Huanhao Zhang[4], Sen Li[5]*, Shaoliang Han[1]* and Naxin Liu[1]*

[1]Department of Gastrointestinal Surgery, The First Affiliated Hospital of Wenzhou Medical University, Wenzhou, China, [2]Department of Medical Oncology, Sir Run Run Shaw Hospital, School of Medicine, Graduate School, Zhejiang University, Hangzhou, China, [3]The First School of Medicine, School of Information and Engineering, Wenzhou Medical University, Wenzhou, China, [4]School of Public Health and Management, Wenzhou Medical University, Wenzhou, China, [5]School of Basic Medicine, Wenzhou Medical University, Wenzhou, China

Background: Previous studies have suggested an association between Type 1 diabetes (T1D) and autoimmune diseases (AIDs), but the causal relationship remains unclear. Therefore, this study utilizes publicly available Genome-Wide Association Studies (GWAS) databases and employs a two-sample Mendelian Randomization (MR) approach to explore the causal relationships between T1D and systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), and inflammatory bowel disease (IBD).

Methods: Summary GWAS data for T1D, SLE, RA, and IBD were downloaded from open GWAS databases and the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC). We employed a series of methods to select instrumental variables closely related to T1D. To enhance the reliability of our conclusions, we applied multiple robust analytical methods, with the inverse variance weighted (IVW) method as the primary approach. Validation and meta-analysis were conducted using the FinnGen consortium. Additionally, we assessed heterogeneity, pleiotropy, and sensitivity to ensure the robustness of our conclusions.

Results: A potential causal association was found between T1D and SLE (OR = 1.37, 95% CI = 1.26 – 1.49, $P < 0.001$), which was further confirmed by meta-analysis. Similarly, a potential causal association was found between T1D and RA (OR = 1.32, 95% CI = 1.17 – 1.50, $P < 0.001$), and this was also confirmed by meta-analysis. Although the association between T1D and IBD showed $P < 0.05$, the leave-one-out test did not pass, and further meta-analysis indicated no significant statistical association between them.

Conclusion: Our study reveals the relationships between T1D and three clinically common autoimmune diseases (SLE, RA, and IBD). This research supplements previous studies and provides a reference for future clinical work.

KEYWORDS

systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), inflammatory bowel disease (IBD), type 1 diabetes(T1D), Mendelian randomization(MR)

Introduction

Autoimmune diseases (AIDs) are a group of complex chronic diseases of unknown etiology characterized by defects in immune tolerance. Common autoimmune diseases include systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), and inflammatory bowel disease (IBD) (Gao et al., 2021). In the United States, autoimmune diseases are one of the leading causes of death among young and middle-aged women (Cooper and Stroehla, 2003). Additionally, because these conditions are often lifelong, they impose a significant burden on both society and individuals (Roberts and Erdei, 2020; Rose, 2016).

Type 1 diabetes (T1D) is an autoimmune disease characterized by insulin deficiency and resultant hyperglycemia (DiMeglio et al., 2018). It commonly occurs in individuals aged 10–14 years (DiMeglio et al., 2018; Maahs et al., 2010). The current understanding is that its pathogenesis may be related to a T-cell-mediated autoimmune process targeting pancreatic β-cells, with its incidence increasing globally (Vehik and Dabelea, 2011). The relationship between T1D and autoimmune diseases has long been both intriguing and perplexing. Clinically, it has been observed that patients with T1D often have other autoimmune diseases, such as dermatological and rheumatic conditions (Popoviciu et al., 2023). Research indicates that T1D and other autoimmune diseases may share certain pathways or genes (Szymczak et al., 2021). However, the causal relationship between T1D and other autoimmune diseases remains unclear.

Observational studies may struggle to correctly determine causality or may produce spurious associations due to the presence of some unavoidable biases (Boyko, 2013). Therefore, in this study, we use Mendelian Randomization (MR) to further investigate the causal relationship between T1D and three clinically common autoimmune diseases (SLE RA and IBD). Mendelian Randomization uses genetic variation as an instrumental variable for the exposure, thereby determining the causal relationship between the exposure and the outcome (Davey Smith and Hemani, 2014; Yarmolinsky et al., 2018). This method can avoid reverse causation and potential confounding biases, making the results more convincing (Zoccali et al., 2006).

Materials and method

Study Design

Mendelian Randomization (MR) studies typically use single nucleotide polymorphisms (SNPs) as instrumental variables (IVs). Conducting an MR analysis requires meeting the following three assumptions (Figure 1): (1) the IVs are strongly associated with the exposure; (2) the IVs are not associated with potential confounders; (3) the IVs influence the outcome only through the exposure. The data used in this study are publicly available and free, thus no further ethical review or patient consent is required.

Data sources

To ensure the robustness of the results and the generalizability of the conclusions, we selected databases from two different sources for each outcome. Details of the data are shown in Table 1.

SNPs related to T1D were obtained from a large Genome-Wide Association Studies (GWAS) study, which included 9,266 cases and 15,574 controls (Forgetta et al., 2020).

SNPs related to SLE were obtained from a large GWAS study that included 5,201 cases and 9,066 controls (Bentham et al., 2015). Moreover, SLE data from the Finnish database (FinnGen) included 538 cases and 213,145 controls.

SNPs related to RA were obtained from a large GWAS study that included 14,361 cases and 43,923 controls (Okada et al., 2014). Moreover, RA data from the Finnish database (FinnGen) included 6,236 cases and 147,221 controls.

SNPs related to IBD were obtained from a study by the International Inflammatory Bowel Disease Genetics Consortium (IIBDGC), which is the largest genetic database for IBD globally. This study included 31,665 cases and 33,977 controls after quality control (QC) (Liu et al., 2015). In addition, IBD data from the Finnish database (FinnGen) included 5,673 cases and 213,119 controls.

When multiple GWAS databases were available, we prioritized those with larger sample sizes, more SNPs, and greater citation frequency by researchers.

Meta-analysis

To validate the robustness of the results, we further verified the outcomes within the FinnGen consortium. Subsequently, we conducted a meta-analysis to further ascertain the relationship between T1D and the different autoimmune diseases. In the meta-analysis, a random effects model was used if heterogeneity ($p < 0.05$) was present; if no heterogeneity was detected ($p > 0.05$), a fixed effects model was employed.

Selection of genetic instruments

To ensure adherence to the assumptions of Mendelian Randomization, we selected instrumental variables based on the following criteria (Gagliano Taliun and Evans, 2021): we used a threshold of $p < 5 \times 10^{-8}$ as the primary filter to ensure that the SNPs were strongly associated with the characteristics of T1D. Moreover, we excluded SNPs in linkage disequilibrium (LD) ($R^2 < 0.001$, clumped at 10,000 kb). We also calculated the F-statistic to test for bias due to weak instruments, using the formula: $F = \beta^2/se^2$ (Wang et al., 2022; Zhao et al., 2023; Li and Martin, 2002). An F-statistic greater than 10 was required to minimize bias from weak instruments (Burgess et al., 2011).

Statistical analysis

In this study, MR analysis was conducted using the TwoSampleMR package (version 0.5.6) and R software (version 4.2.1) (Yavorska and Burgess, 2017). Meta-analysis was performed using Review Manager (version 5.4). The primary analysis method was the Inverse Variance Weighted (IVW) approach, which combines the Wald ratio estimates of each SNP to produce a pooled estimate (Pierce and Burgess, 2013). Supplementary analyses included: (1). Weighted Median (Bowden et al., 2016). This method can provide consistent estimates of causal effects

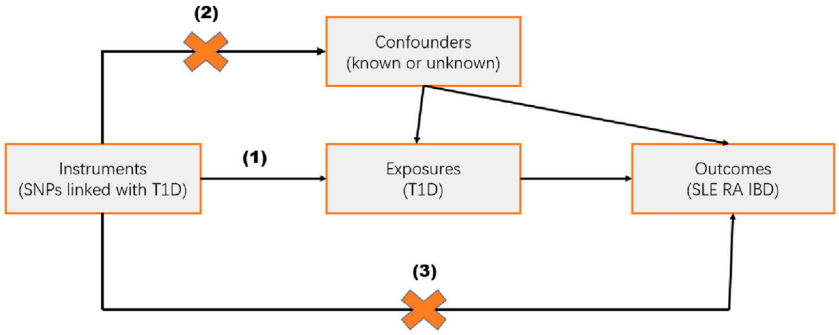


FIGURE 1
Study Design. Mendelian randomization studies are based on three assumptions: (1) the instrumental variable (IV) is strongly associated with the exposure; (2) the IV is independent of known or unknown confounders; (3) the IV influences the outcome only through the exposure.

TABLE 1 Data sources.

Phenotype	Data source	Sample size (cases/controls)
Exposure		
T1D	Vincenzo Forgetta et al	9266/15574
outcome		
SLE	James Bentham et al	5201/9066
SLE	FinnGen	538/213145
RA	Yukinori Okada et al	14361/43923
RA	FinnGen	6236/147221
IBD	IIBDGC	31665/33977
IBD	FinnGen	5673/213119

even if up to 50% of the instruments are invalid; (2). MR Egger (Bowden et al., 2015). This method offers consistent estimates of pleiotropy even if all instruments are invalid; 3. MR-PRESSO (Verbanck et al., 2018). This method identifies outliers with horizontal pleiotropy and is most effective when less than 50% of the instruments exhibit horizontal pleiotropy. Cochran's Q test was used to detect heterogeneity (Greco et al., 2015). The intercept test from MR Egger regression was employed to evaluate horizontal pleiotropy (Bowden et al., 2015).

Results

Selection of instrumental variables

We selected IVs based on the criteria outlined above. Ultimately, we identified 44 SNPs to be used as IVs for T1D. Moreover, all F-statistics were greater than 10, indicating the absence of weak instrument bias (Supplementary Table S1).

Relationship between T1D and SLE

In this study, we found that T1D exhibited a positive association with SLE (OR = 1.37, 95% CI = 1.26–1.49, $p < 0.001$). This result remained robust even after removing outliers using the MR-PRESSO method (OR = 1.11, 95% CI = 1.02–1.20, $p = 0.018$) (Figures 2, 4; Table 2). Within the FinnGen consortium, T1D continued to show a positive association with SLE (OR = 1.18, 95% CI = 1.10–1.27, $p < 0.001$) (Figures 3, 4; Table 2). Meta-analysis further confirmed the relationship between the two (OR = 1.27, 95% CI = 1.10–1.46, $p = 0.001$) (Figure 5).

Relationship between T1D and RA

In our study, we found a positive association between T1D and RA (OR = 1.32, 95% CI = 1.17–1.50, $p < 0.001$). This positive association persisted even after removing outliers using the MR-PRESSO method (OR = 1.26, 95% CI = 1.21–1.31, $p < 0.001$) (Figures 2, 4; Table 2). This conclusion was also validated using data from the FinnGen consortium (OR = 1.17, 95% CI = 1.07–1.27, $p < 0.001$) (Figures 3, 4; Table 2). Meta-analysis further confirmed the relationship between T1D and RA (OR = 1.23, 95% CI = 1.09–1.39, $p = 0.001$) (Figure 5).

Relationship between T1D and IBD

In this study, our analysis indicated a negative association between T1D and IBD (OR = 0.88, 95% CI = 0.83–0.94, $p <$

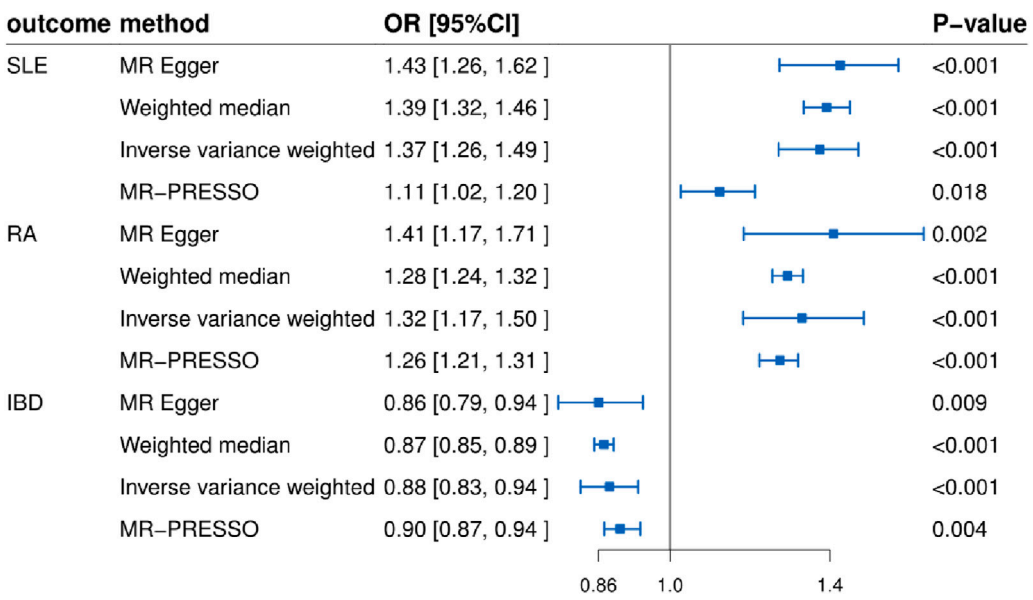


FIGURE 2
Forest plots showing the relationship between T1D and SLE, RA, and IBD (non-FinnGen databases).

0.001) (Figures 2, 4; Table 2). However, it is noteworthy that the leave-one-out analysis suggested that this relationship might be disproportionately influenced by a single SNP, indicating potential bias in the results (Supplementary Figure S1). In the FinnGen consortium, a similar trend was observed between T1D and IBD (OR = 0.96, 95% CI = 0.93–1.00, p = 0.043) (Figures 3, 4; Table 2), but the leave-one-out analysis still indicated potential bias (Supplementary Figure S1). Meta-analysis showed no significant statistical difference (OR = 0.92, 95% CI = 0.85–1.01, p = 0.07) (Figure 5).

MR sensitivity analysis results

The MR Egger regression intercept indicated limited evidence of horizontal pleiotropy (Table 3). For T1D and SLE, as well as T1D and RA, the leave-one-out analysis demonstrated that the causal associations were not unduly influenced by any single SNP. However, the leave-one-out analysis suggested that the causal association between T1D and IBD might be disproportionately affected by a single SNP. Heterogeneity tests for each group are presented in Table 3. The forest plots and volcano plots provide a more visual representation of the heterogeneity (Supplementary Figures S2, S3).

Discussion

In this study, we used MR to evaluate the causal relationships between T1D and several clinically common autoimmune diseases. Our research indicated that genetic susceptibility to T1D was associated with an increased risk of both SLE and RA, but not with IBD.

T1D is a complex chronic disease that is often found to co-occur with other autoimmune diseases in clinical settings (Zeglaoui et al., 2010; Çetın et al., 2013). A study from Sweden involving 3,093 participants demonstrated a significant association between T1D and RA (OR = 4.9, 95% CI = 1.8, 13.1), which is consistent with our findings (Liao et al., 2009). Although previous views suggested that T1D is not an independent risk factor for RA (Popoviciu et al., 2023), our analysis indicated a possible causal relationship, and Zhernakova et al. have also identified shared genetic risk loci between T1D and RA (Zhernakova et al., 2007). Similarly, a study based on the HealthFacts database showed that patients with T1D are more likely to develop SLE, another rheumatic disease (1325/158865) (Bao et al., 2019), compared to an incidence rate of approximately 23.2 per 100,000 in the general North American population (Popoviciu et al., 2023). Additionally, both RA and SLE are more commonly co-morbid in female T1D patients than in males (Bao et al., 2019; Bao et al., 2018). Therefore, clinicians should be vigilant in preventing rheumatic diseases in T1D patients, especially in females, to reduce potential risks and economic burdens on patients.

Although SLE and RA are distinct diseases, they both fall under the category of rheumatic diseases. Previous researches have shown that RA and SLE share familial aggregation (Cardenas-Roldan et al., 2013), genetic (Cui et al., 2013; Orozco et al., 2011; Marquez et al., 2017), molecular mechanisms (Higgs et al., 2011), and targeted therapies (Petitdemange et al., 2020), which might partially explain why both are associated with T1D. Studies have indicated that the interleukin two receptor subunit alpha (IL2RA) gene is closely related to the onset of T1D (Pahkuri et al., 2023), and IL2RA is also implicated in the pathogenesis of SLE and RA (Gorji et al., 2019; van Steenbergen et al., 2015). Our study also identified that mutations in the IL2RA gene (rs12722495) might contribute to the associations observed between these conditions.

TABLE 2 MR analysis results for T1D with SLE, RA, and IBD.

Outcome	Data source	Methods	OR	95%CI	P-Value
SLE	James Bentham et al	MR-Egger	1.43	1.26–1.62	P<0.001
		Weighted median	1.39	1.32–1.46	P<0.001
		IVW	1.37	1.26–1.49	P<0.001
		MR-PRESSO	1.11	1.02–1.20	P = 0.018
SLE	FinnGen	MR-Egger	1.18	1.06–1.32	P = 0.005
		Weighted median	1.19	1.08–1.31	P<0.001
		IVW	1.18	1.10–1.27	P<0.001
		MR-PRESSO	1.18	1.10–1.26	P<0.001
RA	Yukinori Okada et al	MR-Egger	1.41	1.17–1.71	P = 0.002
		Weighted median	1.28	1.24–1.32	P<0.001
		IVW	1.32	1.17–1.50	P<0.001
		MR-PRESSO	1.26	1.21–1.31	P<0.001
RA	FinnGen	MR-Egger	1.24	1.09–1.40	P = 0.002
		Weighted median	1.11	1.06–1.17	P<0.001
		IVW	1.17	1.07–1.27	P<0.001
		MR-PRESSO	1.07	1.02–1.12	P = 0.006
IBD	IIBDGC	MR-Egger	0.86	0.79–0.94	P = 0.009
		Weighted median	0.87	0.85–0.89	P<0.001
		IVW	0.88	0.83–0.94	P<0.001
		MR-PRESSO	0.90	0.87–0.94	P = 0.004
IBD	FinnGen	MR-Egger	0.95	0.90–1.00	P = 0.081
		Weighted median	0.96	0.92–1.01	P = 0.113
		IVW	0.96	0.93–1.00	P = 0.043
		MR-PRESSO	0.98	0.94–1.02	P = 0.234

The causal relationship between T1D and IBD has long been debated. A study from Denmark indicated a significant association between T1D and IBD (Halling et al., 2017). However, other studies have found no significant association between the two (Cohen et al., 2008; Lu et al., 2020), which aligns with our findings. Although our results confirm some previous clinical studies, several important points deserve attention: Firstly, T1D commonly occurs in individuals aged 10–14 years (DiMeglio et al., 2018; Maahs et al., 2010), whereas IBD tends to develop in young and middle-aged adults (He et al., 2022). This study targeted an adult population. For the pediatric population, a study involving 1,200 cases found an association between IBD and diabetes (Kappelman et al., 2011). Additionally, research from Austria and Germany observed a higher incidence of IBD in children with T1D compared to their age-matched peers (Jasser-Nitsche et al., 2021). Therefore, the relationship between T1D and early-onset IBD in children warrants further investigation. Secondly, although our study did not find a statistically significant relationship between T1D and IBD, the P-value was close to 0.05, suggesting a potential negative association trend. Previous studies have shown that protein tyrosine phosphatase non-receptor type 22 (PTPN22) plays an opposing role in Crohn's disease compared to T1D (Barrett et al., 2008). Research indicates that PTPN22 knockdown activates inflammatory signaling pathways, leading to Crohn's disease (Spalinger et al., 2013). Conversely, PTPN22 knockdown does not increase the risk of T1D and may even confer protective effects (Zheng and Kissler, 2013). Similarly, risk alleles for T1D, such as Interleukin 27 (IL-27), Interleukin 10 (IL-10), and interleukin-18 receptor 1 (IL18RA), have been found to prevent Crohn's disease. Major histocompatibility complex (MHC) alleles strongly associated with T1D risk also appear to prevent both Crohn's disease and ulcerative colitis (Wang et al., 2010). In contrast, PTPN22 is implicated in promoting the development of RA and SLE (Liao et al., 2009; Deng and Tsao, 2010). This intriguing phenomenon may relate to the "direction" of genetic variants: if a variant is associated with multiple

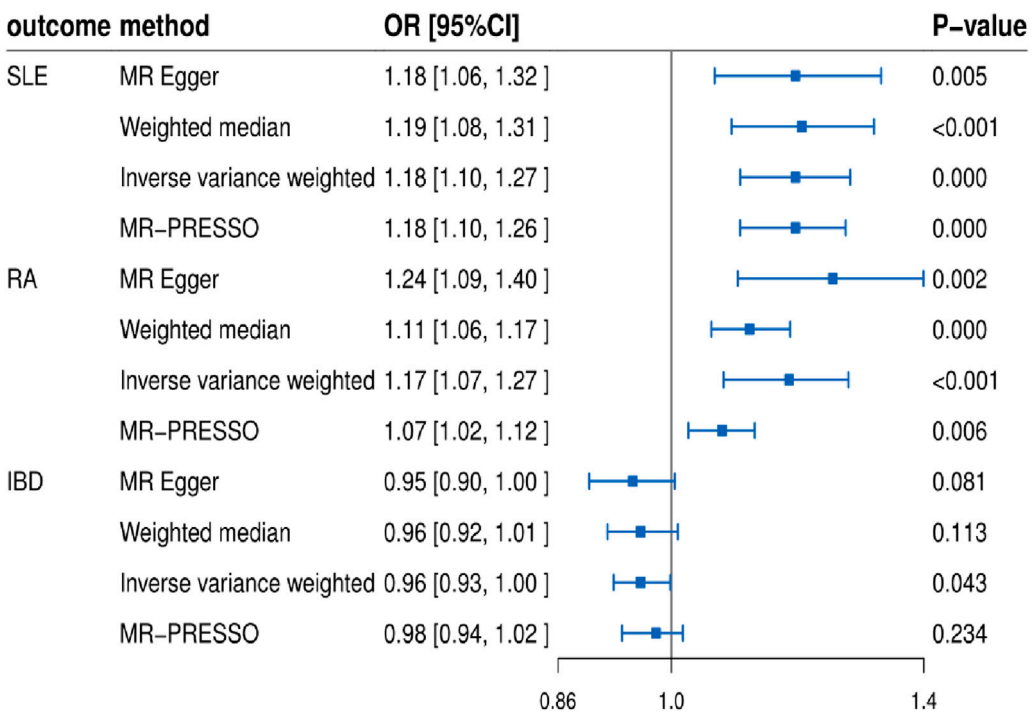


FIGURE 3
Forest plots showing the relationship between T1D and SLE, RA, IBD (FinnGen database).

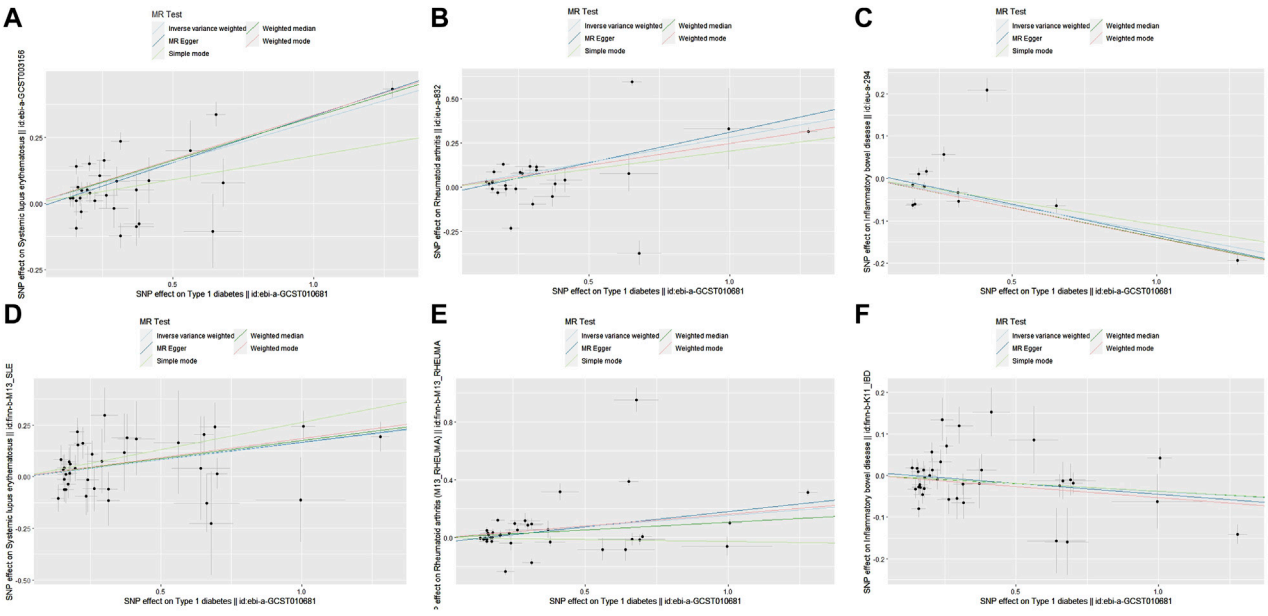


FIGURE 4
Scatter plots showing the relationship between T1D and SLE, RA, IBD. **(A)** T1D and SLE (non-FinnGen databases); **(B)** T1D and RA (non-FinnGen databases) **(C)** T1D and IBD (non-FinnGen databases); **(D)** T1D and SLE (FinnGen database) **(E)** T1D and RA (FinnGen database); **(F)** T1D and IBD (FinnGen database).

autoimmune diseases but in opposite directions, it is more likely to be involved in pathways related to immune function, exhibiting contrasting characteristics (Wang et al., 2010).

Compared with traditional research methods, our study has several advantages. Firstly, we used Mendelian Randomization to evaluate the relationship between T1D and other autoimmune

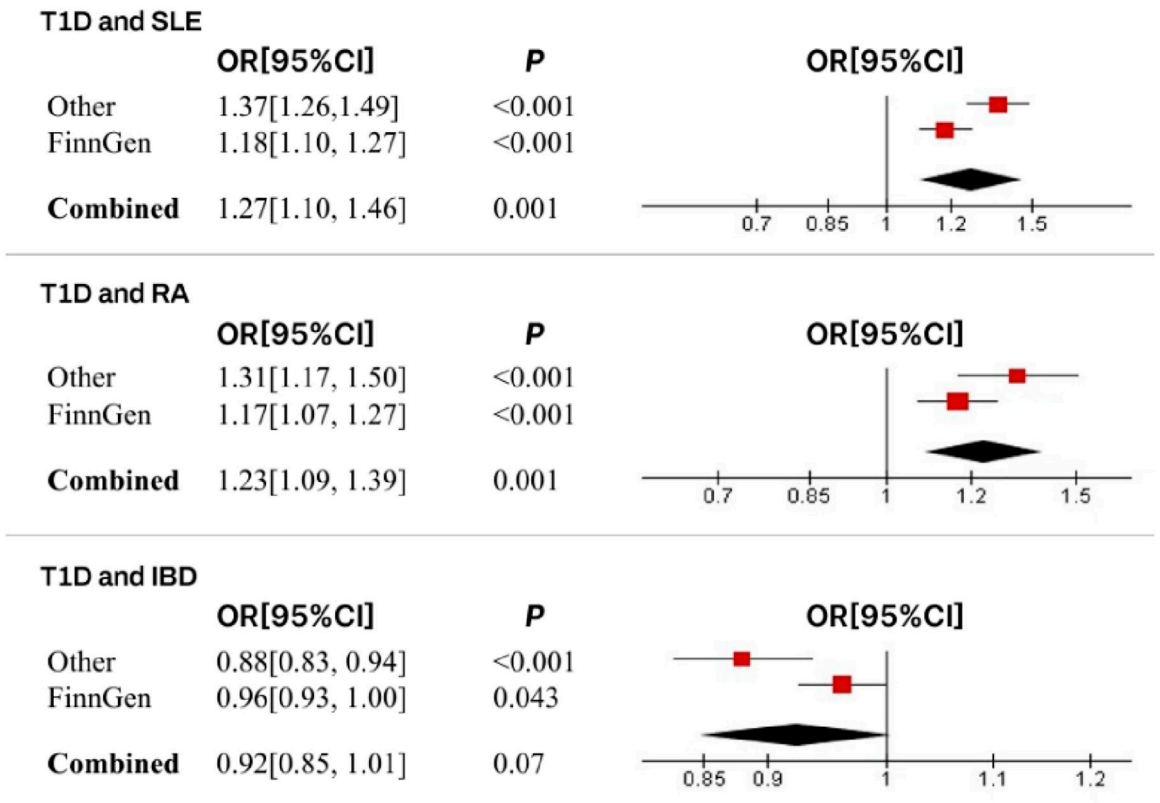


FIGURE 5
Meta-analysis results and forest plots for the relationship between T1D and SLE, RA, and IBD.

TABLE 3 Heterogeneity and Horizontal Pleiotropy Test Results

Outcome	Pleiotropy		Heterogeneity		Data source
	Intercept	P	Q	P	
SLE	−0.022	P = 0.37	139	P < 0.001	James Bentham et al
SLE	−0.002	P = 0.94	47	P = 0.08	FinnGen
RA	−0.032	P = 0.39	720	P < 0.001	Yukinori Okada et al
RA	−0.033	P = 0.21	543	P < 0.001	FinnGen
IBD	0.012	P = 0.58	161	P < 0.001	IIBDGC
IBD	0.007	P = 0.54	116	P < 0.001	FinnGen

diseases. This method significantly reduces the impact of confounding bias and reverse causation. Secondly, our instrumental variables were derived from large-scale GWAS, providing reliable and robust SNP data, which helps avoid bias from weak instruments. Additionally, we conducted meta-analyses to further assess the reliability of our results.

However, our study has some limitations. Firstly, although we used various methods to analyze horizontal pleiotropy, we cannot entirely rule out the presence of potential horizontal pleiotropy. Secondly, Mendelian Randomization itself may face new issues, such as the "winner's curse" (Yarmolinsky et al., 2018). Lastly, our study was limited by its focus on a specific ethnic group, so the conclusions might not be generalizable to other populations.

Conclusion

In summary, our study further substantiated the causal relationships between T1D and both RA and SLE, while no association was found between T1D and IBD. These findings suggested that in managing patients with T1D, attention should be given to preventing RA and SLE to reduce potential complications and economic burdens for patients.

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 authors.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participant's legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

WX: Supervision, Writing–original draft, Writing–review and editing. HJ: Conceptualization, Investigation, Methodology, Project administration, Software, Writing–review and editing. YC: Conceptualization, Project administration, Validation, Writing–review and editing. ZY: Validation, Writing–review and editing. YS: Validation, Writing–review and editing. HZ: Validation, Writing–review and editing. SL: Writing–review and editing, Funding acquisition. SH: Writing–review and editing, Project administration. NL: Writing–review and editing, Funding acquisition, Project administration.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This project was sponsored by the Zhejiang Medical and Health Project (Key Talents) (2014RCA014), Wenzhou Municipal Science and Bureau (Y20210184), Zhejiang Provincial Health Department Medical Support Discipline–Nutrition (11-ZC24), Natural Science Foundation of Zhejiang Province (LQ20H020002), General Research Project of Zhejiang Provincial Department of Education (Y201942047) and Wenzhou Science and Technology Program (Y20180060).

Acknowledgments

We are grateful to all organizations that provide publicly available data.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2024.1335839/full#supplementary-material

References

Bao, Y. K., Salam, M., Parks, D., McGill, J. B., and Hughes, J. (2018). High prevalence of systemic rheumatic diseases in women with type 1 diabetes. *J. Diabetes Complicat.* 32 (8), 737–739. doi:10.1016/j.jdiacomp.2018.06.001

Bao, Y. K., Weide, L. G., Ganesan, V. C., Jakhar, I., McGill, J. B., Sahil, S., et al. (2019). High prevalence of comorbid autoimmune diseases in adults with type 1 diabetes from the HealthFacts database. *J. Diabetes* 11 (4), 273–279. doi:10.1111/1753-0407.12856

Barrett, J. C., Hansoul, S., Nicolae, D. L., Cho, J. H., Duerr, R. H., Rioux, J. D., et al. (2008). Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* 40 (8), 955–962. doi:10.1038/ng.175

Bentham, J., Morris, D. L., Graham, D. S. C., Pinder, C. L., Tombleson, P., Behrens, T. W., et al. (2015). Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* 47 (12), 1457–1464. doi:10.1038/ng.3434

Bowden, J., Davey Smith, G., and Burgess, S. (2015). Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* 44 (2), 512–525. doi:10.1093/ije/dyv080

Bowden, J., Davey Smith, G., Haycock, P. C., and Burgess, S. (2016). Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet. Epidemiol.* 40 (4), 304–314. doi:10.1002/gepi.21965

Boyko, E. J. (2013). Observational research-opportunities and limitations. *J. Diabetes Complicat.* 27 (6), 642–648. doi:10.1016/j.jdiacomp.2013.07.007

Burgess, S., Thompson, S. G., and CRP CHD Genetics Collaboration (2011). Avoiding bias from weak instruments in Mendelian randomization studies. *Int. J. Epidemiol.* 40 (3), 755–764. doi:10.1093/ije/dyr036

Cardenas-Roldan, J., Rojas-Villarraga, A., and Anaya, J. M. (2013). How do autoimmune diseases cluster in families? A systematic review and meta-analysis. *BMC Med.* 11, 73. doi:10.1186/1741-7015-11-73

Çetın, D., Ünübol, M., Güney, E., Karaoğlu, A. Ö., Meteoğlu, İ., and Bozkurt, G. (2013). Coexistence of type 1 diabetes mellitus and Crohn's disesae. *Turk J. Gastroenterol.* 24 (5), 451–452. doi:10.4318/tjg.2013.0513

Cohen, R., Robinson, D., Paramore, C., Fraeman, K., Renahan, K., and Bala, M. (2008). Autoimmune disease concomitance among inflammatory bowel disease patients in the United States, 2001-2002. *Inflamm. Bowel Dis.* 14 (6), 738–743. doi:10.1002/ibd.20406

Cooper, G. S., and Stroehla, B. C. (2003). The epidemiology of autoimmune diseases. *Autoimmun. Rev.* 2 (3), 119–125. doi:10.1016/s1568-9972(03)00006-5

Cui, Y., Sheng, Y., and Zhang, X. (2013). Genetic susceptibility to SLE: recent progress from GWAS. *J. Autoimmun.* 41, 25–33. doi:10.1016/j.jaut.2013.01.008

Davey Smith, G., and Hemani, G. (2014). Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* 23 (R1), R89–R98. doi:10.1093/hmg/ddu328

Deng, Y., and Tsao, B. P. (2010). Genetic susceptibility to systemic lupus erythematosus in the genomic era. *Nat. Rev. Rheumatol.* 6 (12), 683–692. doi:10.1038/nrrheum.2010.176

DiMeglio, L. A., Evans-Molina, C., and Oram, R. A. (2018). Type 1 diabetes. *Lancet* 391 (10138), 2449–2462. doi:10.1016/S0140-6736(18)31320-5

Forgetta, V., Manousaki, D., Istomine, R., Ross, S., Tessier, M. C., Marchand, L., et al. (2020). Rare genetic variants of large effect influence risk of type 1 diabetes. *Diabetes* 69 (4), 784–795. doi:10.2337/db19-0831

Gagliano Taliun, S. A., and Evans, D. M. (2021). Ten simple rules for conducting a mendelian randomization study. *PLoS Comput. Biol.* 17 (8), e1009238. doi:10.1371/journal.pcbi.1009238

Gao, Z. W., Wang, X., Zhang, H. Z., Lin, F., Liu, C., and Dong, K. (2021). The roles of adenosine deaminase in autoimmune diseases. *Autoimmun. Rev.* 20 (1), 102709. doi:10.1016/j.autrev.2020.102709

Gorji, A. E., Roudbari, Z., Alizadeh, A., and Sadeghi, B. (2019). Investigation of systemic lupus erythematosus (SLE) with integrating transcriptomics and genome wide association information. *Gene* 706, 181–187. doi:10.1016/j.gene.2019.05.004

Greco, M. F., Minelli, C., Sheehan, N. A., and Thompson, J. R. (2015). Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome. *Stat. Med.* 34 (21), 2926–2940. doi:10.1002/sim.6522

Halling, M. L., Kjeldsen, J., Knudsen, T., Nielsen, J., and Hansen, L. K. (2017). Patients with inflammatory bowel disease have increased risk of autoimmune and inflammatory diseases. *World J. Gastroenterol.* 23 (33), 6137–6146. doi:10.3748/wjg.v23.i33.6137

He, B. J., Liu, Z. K., Shen, P., Sun, Y. X., Chen, B., Zhan, S. Y., et al. (2022). Epidemiological study on the incidence of inflammatory bowel disease in Yinzhou District, Ningbo City from 2011 to 2020. *Beijing Da Xue Xue Bao Yi Xue Ban.* 54 (3), 511–519. doi:10.19723/j.issn.1671-167X.2022.03.017

Higgs, B. W., Liu, Z., White, B., Zhu, W., White, W. I., Morehouse, C., et al. (2011). Patients with systemic lupus erythematosus, myositis, rheumatoid arthritis and scleroderma share activation of a common type I interferon pathway. *Ann. Rheum. Dis.* 70 (11), 2029–2036. doi:10.1136/ard.2011.150326

Jasser-Nitsche, H., Bechtold-Dalla Pozza, S., Binder, E., Bollow, E., Heidtmann, B., Lee-Barkley, Y. H., et al. (2021). Comorbidity of inflammatory bowel disease in children and adolescents with type 1 diabetes. *Acta Paediatr.* 110 (4), 1353–1358. doi:10.1111/apa.15643

Kappelman, M. D., Galanko, J. A., Porter, C. Q., and Sandler, R. S. (2011). Association of paediatric inflammatory bowel disease with other immune-mediated diseases. *Arch. Dis. Child.* 96 (11), 1042–1046. doi:10.1136/archdischild-2011-300633

Li, B., and Martin, E. B. (2002). An approximation to the F distribution using the chi-square distribution. *Comput. statistics and data analysis* 40 (1), 21–26. doi:10.1016/s0167-9473(01)00097-4

Liao, K. P., Gunnarsson, M., Källberg, H., Ding, B., Plenge, R. M., Padyukov, L., et al. (2009). Specific association of type 1 diabetes mellitus with anti-cyclic citrullinated peptide-positive rheumatoid arthritis. *Arthritis Rheum.* 60 (3), 653–660. doi:10.1002/art.24362

Liu, J. Z., van Sommeren, S., Huang, H., Ng, S. C., Alberts, R., Takahashi, A., et al. (2015). Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* 47 (9), 979–986. doi:10.1038/ng.3359

Lu, S., Gong, J., Tan, Y., and Liu, D. (2020). Epidemiologic association between inflammatory bowel diseases and type 1 diabetes mellitus: a meta-analysis. *J. Gastrointestin Liver Dis.* 29 (3), 407–413. doi:10.15403/jgld-798

Maahs, D. M., West, N. A., Lawrence, J. M., and Mayer-Davis, E. J. (2010). Epidemiology of type 1 diabetes. *Endocrinol. Metab. Clin. North Am.* 39 (3), 481–497. doi:10.1016/j.ecl.2010.05.011

Marquez, A., Vidal-Bralo, L., Rodríguez-Rodríguez, L., González-Gay, M. A., Balsa, A., González-Álvaro, I., et al. (2017). A combined large-scale meta-analysis identifies COG6 as a novel shared risk locus for rheumatoid arthritis and systemic lupus erythematosus. *Ann. Rheum. Dis.* 76 (1), 286–294. doi:10.1136/annrheumdis-2016-209436

Okada, Y., Wu, D., Trynka, G., Raj, T., Terao, C., Ikari, K., et al. (2014). Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* 506 (7488), 376–381. doi:10.1038/nature12873

Orozco, G., Eyre, S., Hinks, A., Bowes, J., Morgan, A. W., Wilson, A. G., et al. (2011). Study of the common genetic background for rheumatoid arthritis and systemic lupus erythematosus. *Ann. Rheum. Dis.* 70 (3), 463–468. doi:10.1136/ard.2010.137174

Pahkuri, S., Ekman, I., Vandamme, C., Näntö-Salonen, K., Toppari, J., Veijola, R., et al. (2023). DNA methylation differences within INS, PTPN22 and IL2RA promoters in lymphocyte subsets in children with type 1 diabetes and controls. *Autoimmunity* 56 (1), 2259118. doi:10.1080/08916934.2023.2259118

Petitdemange, A., Blaess, J., Sibilia, J., Felten, R., and Arnaud, L. (2020). Shared development of targeted therapies among autoimmune and inflammatory diseases: a systematic repurposing analysis. *Ther. Adv. Musculoskelet. Dis.* 12, 1759720X20969261. doi:10.1177/1759720X20969261

Pierce, B. L., and Burgess, S. (2013). Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *Am. J. Epidemiol.* 178 (7), 1177–1184. doi:10.1093/aje/kwt084

Popoviciu, M. S., Kaka, N., Sethi, Y., Patel, N., Chopra, H., and Cavalu, S. (2023). Type 1 diabetes mellitus and autoimmune diseases: a critical review of the association and the application of personalized medicine. *J. Pers. Med.* 13 (3), 422. doi:10.3390/jpm13030422

Roberts, M. H., and Erdei, E. (2020). Comparative United States autoimmune disease rates for 2010-2016 by sex, geographic region, and race. *Autoimmun. Rev.* 19 (1), 102423. doi:10.1016/j.autrev.2019.102423

Rose, N. R. (2016). Prediction and prevention of autoimmune disease in the 21st century: a review and preview. *Am. J. Epidemiol.* 183 (5), 403–406. doi:10.1093/aje/kwv292

Spalinger, M. R., Lang, S., Weber, A., Frei, P., Fried, M., Rogler, G., et al. (2013). Loss of protein tyrosine phosphatase nonreceptor type 22 regulates interferon-γ-induced signaling in human monocytes. *Gastroenterology* 144 (5), 978–988. doi:10.1053/j.gastro.2013.01.048

Szymczak, F., Colli, M. L., Mamula, M. J., Evans-Molina, C., and Eizirik, D. L. (2021). Gene expression signatures of target tissues in type 1 diabetes, lupus erythematosus, multiple sclerosis, and rheumatoid arthritis. *Sci. Adv.* 7 (2), eabd7600. doi:10.1126/sciadv.abd7600

van Steenbergen, H. W., van Nies, J. A. B., Ruyssen-Witrand, A., Huizinga, T. W. J., Cantagrel, A., Berenbaum, F., et al. (2015). IL2RA is associated with persistence of rheumatoid arthritis. *Arthritis Res. Ther.* 17 (1), 244. doi:10.1186/s13075-015-0739-6

Vehik, K., and Dabelea, D. (2011). The changing epidemiology of type 1 diabetes: why is it going through the roof? *Diabetes Metab. Res. Rev.* 27 (1), 3–13. doi:10.1002/dmrr.1141

Verbanck, M., Chen, C. Y., and Neale, B. (2018). Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* 50 (5), 693–698. doi:10.1038/s41588-018-0099-7

Wang, B., Liu, D., Song, M., Wang, W., Guo, B., and Wang, Y. (2022). Immunoglobulin G N-glycan, inflammation and type 2 diabetes in East Asian and European populations: a Mendelian randomization study. *Mol. Med.* 28 (1), 114. doi:10.1186/s10020-022-00543-z

Wang, K., Baldassano, R., Zhang, H., Qu, H. Q., Imielinski, M., Kugathasan, S., et al. (2010). Comparative genetic analysis of inflammatory bowel disease and type 1 diabetes implicates multiple loci with opposite effects. *Hum. Mol. Genet.* 19 (10), 2059–2067. doi:10.1093/hmg/ddq078

Yarmolinsky, J., Wade, K. H., Richmond, R. C., Langdon, R. J., Bull, C. J., Tilling, K. M., et al. (2018). Causal inference in cancer epidemiology: what is the role of mendelian randomization? *Cancer Epidemiol. Biomarkers Prev.* 27 (9), 995–1010. doi:10.1158/1055-9965.EPI-17-1177

Yavorska, O. O., and Burgess, S. (2017). MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* 46 (6), 1734–1739. doi:10.1093/ije/dyx034

Zeglaoui, H., Landolsi, H., Mankai, A., Ghedira, I., and Bouajina, E. (2010). Type 1 diabetes mellitus, celiac disease, systemic lupus erythematosus and systemic scleroderma in a 15-year-old girl. *Rheumatol. Int.* 30 (6), 793–795. doi:10.1007/s00296-009-0988-2

Zhao, S. S., Yiu, Z. Z. N., Barton, A., and Bowes, J. (2023). Association of lipid-lowering drugs with risk of psoriasis: a mendelian randomization study. *JAMA Dermatol* 159 (3), 275–280. doi:10.1001/jamadermatol.2022.6051

Zheng, P., and Kissler, S. (2013). PTPN22 silencing in the NOD model indicates the type 1 diabetes-associated allele is not a loss-of-function variant. *Diabetes* 62 (3), 896–904. doi:10.2337/db12-0929

Zhernakova, A., Alizadeh, B. Z., Bevova, M., van Leeuwen, M. A., Coenen, M. J. H., Franke, B., et al. (2007). Novel association in chromosome 4q27 region with rheumatoid arthritis and confirmation of type 1 diabetes point to a general risk locus for autoimmune diseases. *Am. J. Hum. Genet.* 81 (6), 1284–1288. doi:10.1086/522037

Zoccali, C., Testa, A., Spoto, B., Tripepi, G., and Mallamaci, F. (2006). Mendelian randomization: a new approach to studying epidemiology in ESRD. *Am. J. Kidney Dis.* 47 (2), 332–341. doi:10.1053/j.ajkd.2005.10.027

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 22 October 2024
DOI 10.3389/fimmu.2024.1447182

OPEN ACCESS

EDITED BY
Alex Tsoi,
University of Michigan, United States

REVIEWED BY
Yue-Bei Luo,
Central South University, China
Shigeaki Suzuki,
Keio University, Japan

*CORRESPONDENCE
YaWen Zhao
18813187041@163.com

†PRESENT ADDRESSES
MengTing Yang,
Department of Neurology, Peking University First Hospital, Beijing, China
JingChu Yuan,
Department of Neurology, Peking University First Hospital, Beijing, China
YiKang Wang,
Department of Neurology, Peking University First Hospital, Beijing, China
YaWen Zhao,
Department of Neurology, Peking University First Hospital, Beijing, China

‡These authors have contributed equally to this work

RECEIVED 11 June 2024
ACCEPTED 01 October 2024
PUBLISHED 22 October 2024

CITATION
Yang M, Yuan J, Wang Y, Hao H, Zhang W, Wang Z, Yuan Y and Zhao Y (2024) Treatment of refractory immune-mediated necrotizing myopathy with efgartigimod.
Front. Immunol. 15:1447182.
doi: 10.3389/fimmu.2024.1447182

Treatment of refractory immune-mediated necrotizing myopathy with efgartigimod

MengTing Yang[1†‡], JingChu Yuan[1†‡], YiKang Wang[1†], HongJun Hao[1,2], Wei Zhang[1,2], ZhaoXia Wang[1,2], Yun Yuan[1,2] and YaWen Zhao[1*†]

[1]Department of Neurology, Peking University First Hospital, Beijing, China, [2]Beijing Key Laboratory of Neurovascular Disease Discovery, Beijing, China

Objective: We aimed to explore the efficacy and safety of efgartigimod in patients with refractory immune-mediated necrotizing myopathy (IMNM).

Methods: This open-label pilot observational study included seven patients with refractory IMNM, all of whom received intravenous efgartigimod treatment. The clinical response was assessed after 4 weeks of efgartigimod treatment according to the 2016 American College of Rheumatology–European League Against Rheumatism response criteria for adult idiopathic inflammatory myopathy. Serum levels of immunoglobulin as well as anti–signal recognition particle (SRP) and anti–3-hydroxy-3-methylglutaryl-CoA reductase (HMGCR) antibodies were measured using enzyme-linked immunosorbent assays and commercial line immunoblot assays. Safety assessments included evaluations of adverse events and severe adverse events.

Results: The seven patients with refractory IMNM included five cases with anti-HMGCR antibodies and two cases within anti-SRP antibodies. Four of the seven patients achieved clinical responses. The total improvement score for the responders at 4 weeks were 32.5, 40.0, 47.5, and 70.0, and those at 8 weeks were 27.5, 47.5, 57.5, and 70.0. In comparison to the responsive patients, the non-responsive patients had longer durations [8 (-) versus 2 (1–5) years, P = 0.03], and more chronic myopathic features by muscle biopsy (67% versus 0%, P = 0.046). Serum immunoglobulin G levels (11.2 ± 2.5 versus 5.7 ± 2.5, P = 0.007) and anti-HMGCR/SRP antibody levels (97.2 ± 6.9 versus 41.8 ± 16.8, P = 0.002) were decreased after treatment compared with baseline levels. Adverse events were reported in one of the seven patients, who showed mild headache.

Conclusions: Despite its small size, our study demonstrated that promoting the degradation of endogenous immunoglobulin G may be effective for patients with IMNM. Efgartigimod may be a promising option for cases of refractory IMNM to shorten duration and minimize chronic myopathic features.

KEYWORDS

immune-mediated necrotizing myopathy, efgartigimod, refractory, anti-signal recognition particle, anti-3-hydroxy-3-methylglutaryl-CoA reductase

1 Introduction

Immune-mediated necrotizing myopathy (IMNM) is a major subgroup of idiopathic inflammatory myopathy characterized by severe proximal weakness and high creatine kinase (CK) levels (1–3). Based on the type of myositis-specific antibodies (MSAs) involved, IMNM can be further classified as anti–signal recognition particle (SRP) myopathy, anti–3-hydroxy-3-methylglutaryl-CoA reductase (HMGCR) myopathy, or seronegative IMNM (1, 4, 5). The relevant autoantibodies bind to target autoantigens in the muscle fibers, potentially leading to the formation of the membrane attack complex and muscle necrosis (1, 6–8).

Compared with other idiopathic inflammatory myopathy subtypes, IMNM has been considered a form of refractory myositis (9, 10), as 27% (11) to 50% (10) of patients with IMNM continue to experience severe muscle weakness even after intensive treatment. Because anti-SRP myopathy and anti-HMGCR myopathy are caused by MSAs, new biotherapies targeting B lymphocytes, such as rituximab (9, 10), ofatumumab (12), and belimumab (13), have been used to treat refractory IMNM, with positive responses in some patients. Therapeutic plasma exchange has also induced positive clinical and laboratory responses in patients with refractory IMNM (14). Those studies indicated that IMNM may benefit from rapid deletion of circulating immunoglobulin (Ig) G to remove pathogenic antibodies and improve patient symptoms.

The neonatal Fc receptor (FcRn) plays a crucial role in extending the lifespan of IgG antibodies by protecting them from lysosomal degradation and recycling them back into circulation (15, 16). Targeting this receptor could present a novel therapeutic approach for IgG-mediated diseases, as inhibiting the FcRn leads to decreased overall IgG and pathological autoantibody levels (15, 16). The development and severity of IMNM are closely linked to the presence and levels of MSAs (17, 18). A recent study showed that efgartigimod can reduce circulating IgG levels, potentially preventing further muscle necrosis and promoting muscle fiber regeneration in a mouse model of IMNM (8). These findings support the investigation of the therapeutic efficacy of efgartigimod in patients with IMNM. In this study, we evaluated the therapeutic effects of IgG reduction via efgartigimod treatment in patients with refractory IMNM.

2 Materials and methods

2.1 Patient registry

This was an observational cohort study that included seven patients who were diagnosed with IMNM according to clinical, serological, and pathological criteria (1) at the Department of Neurology at Peking University First Hospital from January to May 2024. Serum IIM antibodies, including those against Nucleosome Remodeling Deacetylase Complex Subunit Mi-2 Alpha (Mi-2α), Nucleosome Remodeling Deacetylase Complex Subunit Mi-2 Beta (Mi-2β), Transcription Intermediary Factor 1 Gamma (TIF1-γ), Melanoma Differentiation-Associated Gene 5 (MDA5), Nuclear Matrix Protein 2 (NXP2), SUMO-Activating Enzyme Subunit 1 (SAE1), Histidyl-tRNA Synthetase (Jo-1), Threonyl-tRNA Synthetase (PL-7), Alanyl-tRNA Synthetase (PL-12), Glycyl-tRNA Synthetase (EJ), Isoleucyl-tRNA Synthetase (OJ), SRP, HMGCR, Ku Autoantigen (Ku), Polymyositis-Scleroderma Autoantigen 100 kDa (PM-Scl100), Polymyositis-Scleroderma Autoantigen 75 kDa (PM-Scl75), and SSA/Ro52 Autoantigen (Ro52), were detected using Euroline Myositis Profile immunoblot assays (Euroimmun, Lubeck, Germany) according to the manufacturer's instructions. The band intensity was reported relative to grayscale intensity as measured on a CanonScan LIDE 100 Scanner (Canon, Tokyo, Japan) using Line Scan scanning software (Euroimmun, Lubeck, Germany). The intensity of anti-SRP or anti-HMGCR antibodies in the study patients was strongly positive, with values exceeding 50. Anti-nuclear antibody was tested by an immunofluorescence assay using Hep-2010 cell line at a dilution of 1:100. Refractory criteria were defined as disease worsening or relapse after treatment with high-dose glucocorticoids and at least one immunosuppressant at a known effective dose for at least 3 months (1, 11, 19). The following exclusion criteria were applied: 1) treated with intravenous Ig or plasma exchange within the past month, and rituximab or eculizumab within the past 6 months; 2) had hepatitis virus B or C infection, other severe infection, or malignancy; 3) had low IgG serum levels (<6 g/L); 4) were pregnant, lactating, or planning to become pregnant; 5) had a history of infection requiring hospitalization within the 8 weeks prior to screening; 6) previously documented lack of clinical response to plasmapheresis; 7) vaccinated within 4 weeks before screening; or 8) had a history of malignancy. A written informed consent was obtained from all patients.

2.2 Data collection

Before efgartigimod treatment, we collected baseline data on the patients' demographics, clinical manifestations, laboratory tests, electromyography results, and medication history. Serum biomarker data—including total IgG, IgA, and IgM levels, and the intensity of anti-SRP and anti-HMGCR antibodies—were also collected at baseline. Thigh muscle magnetic resonance imaging was performed on all patients before treatment. Fatty replacement of muscle was graded on T1-weighted imaging (T1WI) sequences using the scale proposed by Mercuri et al. (20), and muscle edema was graded on the basis of T2 Short Tau Inversion Recovery (T2-STIR) sequences using a four-point scale (21). Muscle biopsy was performed for all patients before treatment. Muscle specimens were assessed histologically and with immunohistochemical staining for major histocompatibility complex (MHC) class I, membrane attack complex, CD3, CD4, CD8, CD20, and CD68. To exclude various muscular dystrophies, immunohistochemical staining was performed with autoantibodies against dystrophin, α- to δ-sarcoglycans, α- and β-dystroglycans, and dysferlin.

2.3 Outcome assessment and response criteria

Patients were followed from the initiation of efgartigimod and through the whole treatment period of combined therapy with low-

to-moderate–dose oral prednisone or tacrolimus. Three patients received low-dose prednisone (prednisone at ≤10 mg/day or equivalent) (22). Two patients received moderate-dose prednisone (prednisone at 10–30 mg/day or equivalent) (22). Four patients received tacrolimus. The concomitant oral medication regimens were unchanged during the treatment period. Efgartigimod (10 mg/kg) was administered as four infusions per cycle (one infusion per week). Clinical response was assessed using the total improvement score (TIS) according to the 2016 American College of Rheumatology–European League Against Rheumatism clinical response criteria for myositis after 4 and 8 weeks of treatment (23). The TIS (0–100) was determined by summing the scores according to the core set measures (CSMs) listed by the International Myositis Assessment and Clinical Studies Group (IMACS) to provide a quantitative measure of improvement for each patient (23). The CSMs included the Manual Muscle Testing–8 scale (MMT-8), Childhood Myositis Assessment Scale (CMAS), Physician Global Activity visual analog scale (VAS), Patient Global Activity VAS, Health Assessment Questionnaire (HAQ), Myositis Disease Activity Assessment Tool (MDAAT) Extramuscular Disease Activity VAS, and CK level. The TIS thresholds in adult patients for minimal, moderate, and major improvement were ≥20, ≥40, and ≥60 points, respectively; those in pediatric patients for minimal, moderate, and major improvement were ≥30, ≥45, and ≥70 points. The serum Ig level and MSA intensity were assessed at baseline and 4 weeks after the final infusion. Safety assessments included evaluations of adverse events (AEs), severe AEs, clinical laboratory tests, and vital signs, as well as physical examinations. The probucol of time schedule is shown in Figure 1.

2.4 Statistical analysis

Statistical analysis was performed using SPSS 26.0. Categorical variables are reported as numbers or percentages. The mean or median with standard deviation or interquartile range (IQR), respectively, was used to represent the central values of the data, depending on the normality of the distribution of the curve. We used Fisher's exact test for comparisons of categorical variables. To compare the parameters before and after efgartigimod treatment, we use paired t-tests for comparisons of means and Wilcoxon rank sum tests for analyses of data with a non-normal distribution. Where $P < 0.05$, a difference was considered significant.

3 Results

3.1 Baseline characteristics of patients

All patients were women, with a median age at disease onset of 21 years (10–32 years). Five patients were anti-HMGCR–positive and two patients were anti-SRP–positive (Table 1). The median duration of the disease was 6 years (2–8 years). All patients presented with a history of proximal muscle weakness. The median peak CK level at initial presentation was 7,234.0 IU/L (3,006.0–10,010.0). Other clinical features included myalgia in two patients, skin rashes in two patients, and muscle atrophy in two patients. Skin rashes were reported only in patients with anti-HMGCR myopathy. One patient presented with rashes on the anterior chest, which resolved spontaneously before treatment. Another patient had patchy alopecia with erythema. No patients presented with dyspnea, dysphagia, interstitial lung disease, cardiac insufficiency, Raynaud's phenomenon, arthritis, or concomitant cancer/rheumatic disease. Anti-Ku autoantibodies were found in one patient, anti-Ro52 autoantibodies were found in one patient, and antinuclear antibodies were found in two patients. Electromyography revealed irritable myopathy changes in all patients. Muscle edema was observed in six of the seven patients by thigh muscle magnetic resonance imaging, with an average total muscle edema score of 8.1. Fatty infiltration of muscle was present in all patients, with an average total fatty infiltration score of 16.1. Muscle biopsies from all patients showed scattered necrotic and regenerating muscle fibers. Muscle biopsies from two patients, each exhibiting a dystrophic-like progression, muscle atrophy, and severe fatty replacement in MRI,

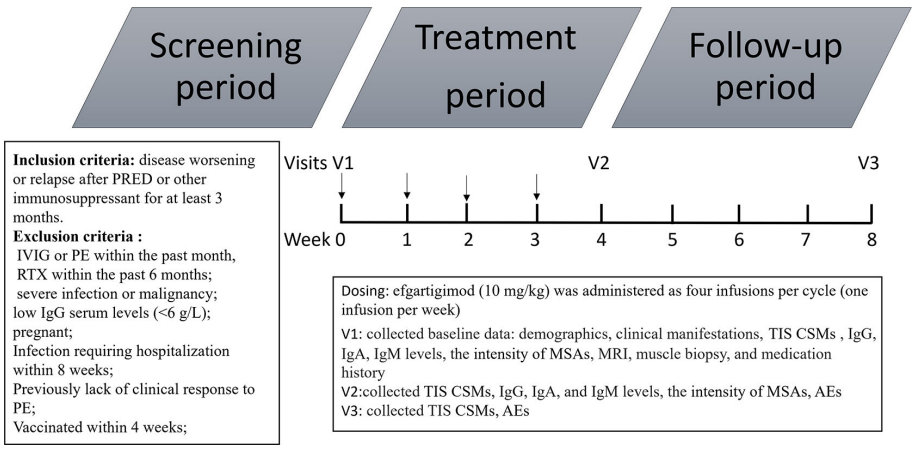


FIGURE 1
Protocol of time schedule. PRED, prednisone; IVIG, intravenous immunoglobulin; PE, plasma exchange; TIS, total improvement score; CSM, core set measures; MRI, magnetic resonance imaging; MSAs, myositis-specific antibodies AE, adverse event.

TABLE 1 Baseline demographics and clinical characteristics of patients with IMNM.

	Patient 1	Patient 2	Patient 3	Patient 4	Patient 5	Patient 6	Patient 7
Sex	Female	Female	Female	Female	Female	Female	Female
MSAs	HMGCR	HMGCR	SRP	HMGCR	HMGCR	SRP	HMGCR
Age at onset, years	13	21	32	5	25	56	10
Duration, years	7	6	8	23	2	1	2
Muscle weakness	+	+	+	+	+	+	+
Myalgia	–	–	–	–	+	+	–
Muscle atrophy	–	+	–	+	–	–	–
Skin rashes	+	–	–	+	–	–	–
Peak CK level, IU/L	2034	7,000	10,121	3,006	13,537	7,234	10,010
EMG	+	+	+	+	+	+	+
Muscle edema in MRI	0	16	5	8	11	9	8
Fatty infiltration in MRI	3	31	34	35	3	4	3
Muscle biopsy	Necrosis pattern	Chronic pattern	Necrosis pattern	Chronic pattern	Necrosis Pattern	Necrosis Pattern	Necrosis Pattern
Previous medication	PRED, MTX, TAC, and IVIG	PRED, MTX, TAC, CTX, IVIG, RTX, and OFA	MTX, Aza, CTX, and IVIG	PRED, MTX, Aza, TAC, RTX, and OFA	PRED, TAC, IVIG, and RTX	PRED and MTX	PRED, TAC, IVIG, and RTX

IMNM, immune-mediated necrotizing myopathy; HMGCR, 3-hydroxy-3-methylglutaryl-CoA reductase; I SRP, signal recognition particle; CK, creatine kinase; EMG, electromyogram; MRI, magnetic resonance imaging; PRED, prednisone; MTX, methotrexate; CTX, cyclophosphamide; TAC, tacrolimus; Aza, azathioprine; OFB, ofatumumab; RTX, rituximab; IVIG, intravenous immunoglobulin.

also revealed chronic myopathic features with endomysial fibrosis and greater variations in fiber size (Figure 2). All patients were initially treated with high-dose prednisone and received various additional immunotherapies for 5 years (2–8 years), including methotrexate in five, tacrolimus in five, azathioprine in two, cyclophosphamide in two, intravenous Ig in five, rituximab in four, and ofatumumab in two.

3.2 Clinical response to treatment

Efgartigimod demonstrated early disease control in four of the seven (57%) patients within 4 weeks of treatment. Four patients (one with anti-HMGCR and three with anti-SRP antibodies) attained minimal to major improvement in 4 weeks, which persisted 8 weeks after efgartigimod treatment. The TIS for the responders at 4 weeks were 32.5, 40.0, 47.5, and 70.0, and those at 8 weeks were 27.5, 47.5, 57.5, and 70.0 (Figure 3). Physician Global Activity [3.0 (IQR, 1.0–5.0) versus 3.0 (IQR, 0.0–5.0), P = 0.046] at 4 weeks after treatment was significantly better than that in baseline. There were statistically significant improvements at 8 weeks after treatment compared with baseline in the following CSMs (Figure 1, Supplementary Table S1): Physician Global Activity [3.0 (IQR, 1.0–5.0) versus 3.0 (IQR, 0.0–5.0), $P = 0.046$] and CK levels [478.0 (184.0–608.0) versus 296.0 (123.0–502.0) IU/L, $P = 0.04$]. Other CSMs—such as MMT-8, CMAS, Patient Global Activity VAS, HAQ, and Extramuscular Disease Activity—showed no significant improvement 4 or 8 weeks after treatment (Figure 3, Supplementary Table S1). In comparison to the responsive patients, the non-responsive patients had longer durations [8 (-) versus 2 (1–5) years, $P = 0.03$] and more chronic myopathic features by muscle biopsy (67% versus 0.0%, $P = 0.046$) (Figure 2, Supplementary Table S2). Subgroup analysis indicated that the beneficial effects of efgartigimod were evident regardless of autoantibody status and dosage of steroids and/or additional non-steroidal immunosuppressive drugs (Supplementary Table S3).

3.3 Serum Igs and anti-desmoglein antibody levels

Serum IgG levels significantly decreased after treatment compared with baseline levels (11.2 ± 2.5 versus 5.7 ± 2.5, $P = 0.007$; Figure 4A), with no differences observed between responders and non-responders. Serum IgG levels decreased from baseline for anti-HMGCR myopathy (mean, 38%) and anti-SRP myopathy (mean, 53%) at the end of the induction phase. There were no clinically relevant changes from the baseline levels of IgA and IgM (Figures 4B, C). MSA intensity significantly decreased post-treatment compared with baseline (97.2 ± 6.9 versus 41.8 ± 16.8, $P = 0.002$; Figure 4D), with no distinction between responders and non-responders. The

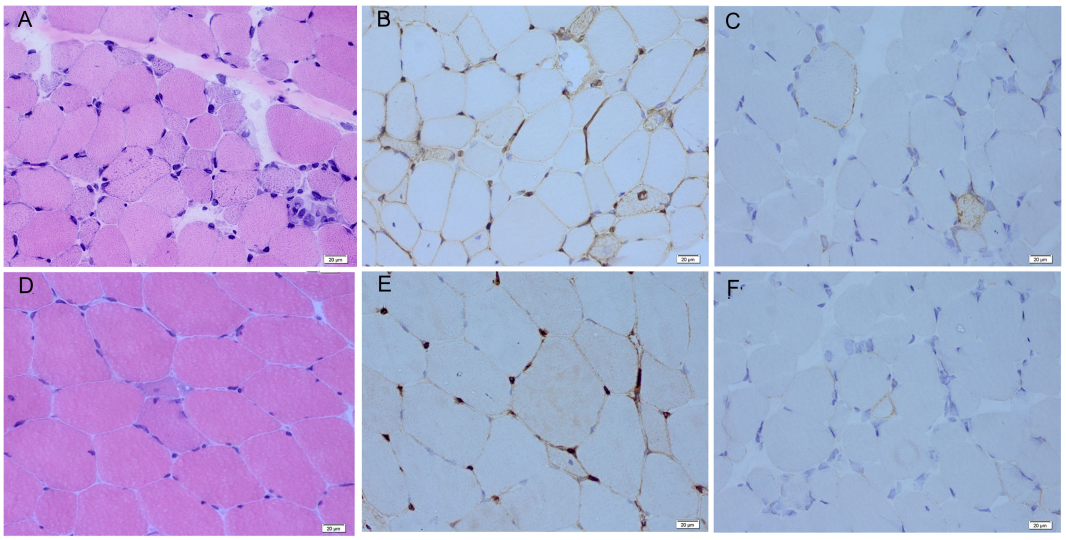


FIGURE 2
Myopathological features of responsive and non-responsive patients. **(A)** Scattered myofiber necrosis and regeneration with endomysial fibrosis are observed by hematoxylin and eosin staining in the non-responsive patients. **(B)** Diffuse sarcolemmal MHC-I deposition is seen by MHC-I immunohistochemical staining in the non-responsive patients. **(C)** Non-necrotic myofibers with little sarcolemmal MAC deposition are observed by MAC immunohistochemical staining in the non-responsive patients. **(D)** Scattered myofiber necrosis and regeneration are seen by hematoxylin and eosin staining in the responsive patients. **(E)** Diffuse sarcolemmal MHC-I deposition is observed by MHC-I immunohistochemical staining in the responsive patients. **(F)** Non-necrotic myofibers with little sarcolemmal MAC deposition are observed on MAC immunohistochemical staining in the responsive patients. MHC-I, major histocompatibility complex–I; MAC, membrane attack complex.

intensity of HMGCR antibodies decreased by a mean of 56% from baseline and 60% for SRP antibodies at the end of the induction phase. Subgroup analysis indicated that the changes in serum IgG levels and antibody levels were evident regardless of concomitant medications (Supplementary Table S3).

3.4 Safety data

During the 8-week study period, AEs were reported in one of the seven patients on efgartigimod, who experienced mild headache. A slightly abnormal differential leukocyte count was detected in one

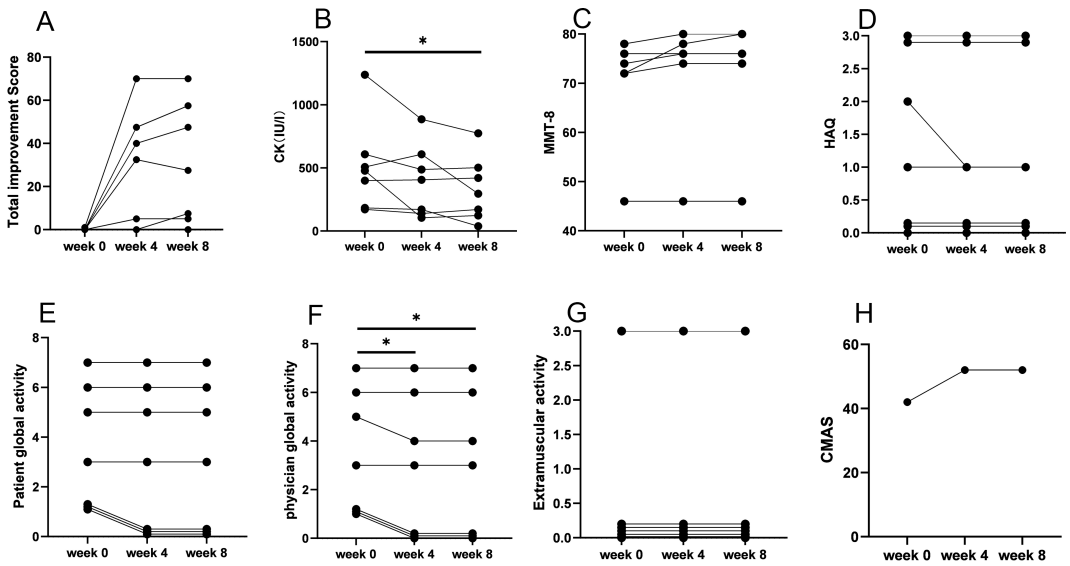


FIGURE 3
IMACS CSM in patients with IMNM at baseline and after treatment with efgartigimod. **(A)** Total improvement score. **(B)** CK level. **(C)** MMT-8 score. **(D)** HAQ score. **(E)** Patient Global Activity. **(F)** Physician Global Activity. **(G)** Extramuscular activity. **(H)** CMAS score. IMACS, International Myositis Assessment and Clinical Studies Group; CSMs, core set measures; IMNM, immune-mediated necrotizing myopathy; CK, creatine kinase; MMT-8, Manual Muscle Testing–8; HAQ, Health Assessment Questionnaire; CMAS, Childhood Myositis Assessment Scale; *$P < 0.05$.

10.3389/fimmu.2024.1447182

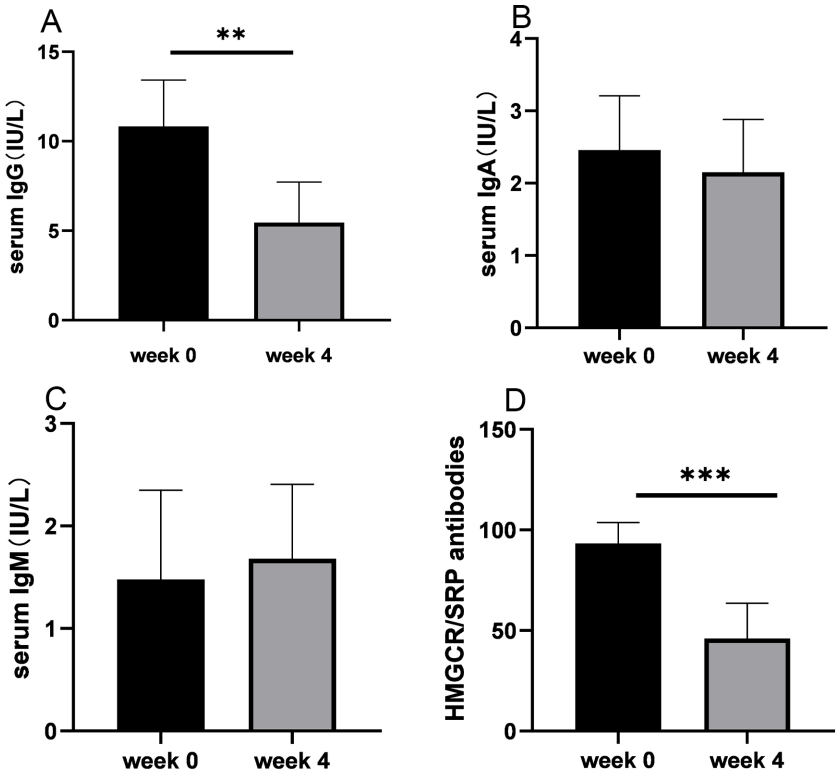


FIGURE 4
Serum levels of immunoglobulins and autoantibodies before and after treatment with efgartigimod. **(A)** Serum IgG levels were significantly decreased after treatment compared with baseline levels. **(B)** Serum IgA levels showed no significant difference after treatment. **(C)** Serum IgM levels showed no significant difference after treatment. **(D)** Staining intensity of anti-HMGCR/SRP antibodies was significantly reduced after treatment compared with baseline staining. Ig, immunoglobulin; HMGCR, 3-hydroxy-3-methylglutaryl-CoA reductase; SRP, signal recognition particle. $^{**}P < 0.01$ and $^{***}P < 0.005$.

case. None of the patients required efgartigimod dose reduction or withdrawal owing to AEs. There were no severe AEs (Table 2).

4 Discussion

We present a single-center, retrospective case series using efgartigimod for refractory IMNM. The seven patients with refractory IMNM included five cases with anti-HMGCR antibodies and two cases within anti-SRP antibodies. All patients presented with proximal muscle weakness and high CK levels. Extramuscular symptoms included skin rashes in two cases with anti-HMGCR antibodies. The Dermatomyositis (DM)-like rashes have been reported in anti-HMGCR myopathy with a frequency ranging from 38% (24) to 43% (25) in previous studies. Despite multiple immunosuppressants, all patients had precipitous declines in strength and quality of life, which led to a trial of efgartigimod treatment.

Although the beneficial effect of FcRn antagonism in refractory IMNM may be attributable to a combination of mechanisms, strategies to deplete pathogenic antibodies have been shown to have a profound impact on patients' responses to therapy. In our study, four patients experienced rapid symptomatic improvement within 4 weeks of efgartigimod treatment. The initial clinical improvement after efgartigimod was dramatic, similar to results in the context of myasthenia gravis (16). Moreover, these improvements persisted even 4 weeks after discontinuation of efgartigimod, indicating that its therapeutic effects are durable. Physician global activity and serum CK levels were significantly

TABLE 2 Summary of AEs in all patients.

AEs	Number (n = 7)
Any AEs	1/7
Any severe AEs	0/7
Any AEs leading to discontinuation of study drug	0/7
Infusion-related reaction event	0/7
Most common adverse events	1/7
Headache	1/7
Nasopharyngitis	0/7
Nausea	0/7
Diarrhea	0/7
Upper respiratory tract infection	0/7
Urinary tract infection	0/7

AEs, adverse events.

improved after treatment, whereas extramuscular symptom (skin rashes) showed no improvement. The concomitant oral medication regimens, which included prednisone and tacrolimus, were low and unchanged during treatment, as was reported in previous studies (26, 27). We evaluated whether the benefits of efgartigimod were consistent across key patient clinical characteristics. Disease control was similar in patients regardless of autoantibody status and/or concomitant medications, suggesting efgartigimod contributed to clinical efficacy.

Patients with a poor outcome in our study had longer durations and more chronic myopathic features by muscle biopsy. In the setting of chronic muscle damage, immune dysregulation and abnormal fibro-adipogenic progenitor differentiation can occur, leading to differentiation into fat cells or fibroblasts, progressive tissue fibrosis, and loss of normal tissue architecture, ultimately causing irreversible damage to the muscle (28). Therefore, although there is no consensus protocol for efgartigimod in IMNM, we suggest that an initial trial of efgartigimod for early-stage disease should be considered. Clinicopathological changes should be considered during patient selection. Muscular dystrophy–like pathology should be an exclusion criterion in further studies, as those pathological changes are currently untreatable.

The pharmacokinetic parameters in this study (10 mg/kg) were in line with data from other studies (15, 27). Efgartigimod rapidly decreased circulating IgG levels from baseline in patients, including autoantibodies, which has also been reported in myasthenia gravis and primary immune thrombocytopenia (16, 29). During the efgartigimod induction phase, early reductions of approximately 50% from baseline in total serum IgG and anti-SRP/HMGCR antibodies were observed after 4 weeks of treatment. Julien et al. also reported that administration of efgartigimod could decrease IgG levels and anti-HMGCR antibodies to prevent further necrosis and allow muscle fiber regeneration in a humanized mouse model of IMNM (8). It is noteworthy that both total IgG and pathogenic antibodies levels were reduced in non-responsive patients, suggesting that these patients may have disease with a non-IgG–mediated mechanism. We found the serum levels of IgA and IgM are not affected by efgartigimod, which has also been reported in myasthenia gravis and healthy volunteers (30, 31). These data reflect the mechanism of efgartigimod action of selective IgG reduction, which leads to incomplete IgG reduction without altering other Ig levels (31, 32).

The primary outcome of the study was safety, and efgartigimod was well tolerated, with few AEs. Mild headache is a well-known side effect of efgartigimod treatment and was reported in 16% of patients with primary immune thrombocytopenia (29) and in 29% of patients with myasthenia gravis (16). Most AEs resolve spontaneously or rapidly upon treatment without the need to discontinue efgartigimod (29). Transient decreases in blood leukocyte levels were observed and were also found in 7 of the 20 healthy volunteers (31). Several studies presented upper respiratory tract infections and urinary tract infections (30); however, a higher rate of infection was not observed in our patients. The efgartigimod did not inhibit production of protective IgG and the risk of infections is unaltered during efgartigimod treatment (31).

Our study preliminarily explored the efficacy and safety of efgartigimod in patients with refractory IMNM. However, all participants underwent only a single-treatment cycle, which raises uncertainty regarding the sustainability of efgartigimod therapy in this patient population. To enhance therapeutic outcomes, it may be advantageous to adopt a sequential treatment approach with efgartigimod aimed at achieving sustained reductions in IgG levels. The ADVANCE study showed the effectiveness and well toleration of efgartigimod using a treatment regimen of either once per week or biweekly for adults with primary immune thrombocytopenia (29). The median interval between treatment cycles in the ADAPT (16) and ADAPT+ (33) studies, which was determined by clinical evaluation of each participant with myasthenia gravis, was approximately 5.8 to 7.3 weeks. Thus, regular monitoring of IgG levels, clinical symptoms, and AEs is essential to identify the optimal timing for subsequent doses during efgartigimod treatment for IMNM. Additionally, previous research has indicated that combination therapy with telitacicept and the faster-acting efgartigimod may represent an effective and safe therapeutic approach for refractory myasthenia gravis (34). Given the close association of IMNM with antibody-mediated pathogenesis, B-cell–targeting treatments to suppress antibody production could also be complementary to efgartigimod (1).

There are several limitations to this study. First, the majority of study participants had previously received various third-line treatments with poor outcomes; therefore, the results may not be generalizable to treatment-naïve patients with IMNM. Second, another mitigating factor is the time from diagnosis to initial treatment with efgartigimod, as well as the duration of acute decline in strength, both of which may mark more extensive muscle damage that may not be reversible by reducing pathologic antibody levels. Some participants included may have been too far advanced in the course of the disease to respond to efgartigimod. Third, small sample size and short observation period limit the ability to evaluate sustained efficacy and rare AEs and may not adequately represent the broader patient population. Finally, we used commercial line immunoblot assay to observe the relative levels of HMGCR and SRP antibodies due to technical factor. We suggest the importance of establishing available titer assays such as quantitative Enzyme linked immunosorbent assay (ELISA) for HMGCR and SRP antibodies, which may be better to track the efficacy of efgartigimod. Future studies are necessary to evaluate the effectiveness of efgartigimod for IMNM more systematically, which may entail establishing a registry of IMNM patient cases and large, prospective studies to assess clinical outcomes using a standardized approach with defined biomarkers and validated clinical endpoints.

In conclusion, to our knowledge, this is the first study evaluating the efficacy and safety of an FcRn inhibitor for the treatment of refractory IMNM. Our findings suggest that efgartigimod may be an encouraging option for refractory IMNM cases. Although a prospective clinical trial remains to be performed, our study demonstrated that promoting the degradation of endogenous IgG may be effective for patients with IMNM, which may pave the way for the efficient design of future trials in idiopathic inflammatory myopathy.

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.

Ethics statement

The studies involving humans were approved by the Institutional Review Board of Peking University First Hospital [No.2019(181)]. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

MY: Data curation, Software, Writing – original draft, Writing – review & editing. JY: Writing – original draft, Writing – review & editing. YW: Data curation, Writing – original draft. HH: Conceptualization, Data curation, Writing – review & editing. WZ: Formal analysis, Writing – review & editing. ZW: Conceptualization, Investigation, Writing – review & editing. YY: Conceptualization, Investigation, Writing – review & editing. YZ: Conceptualization, Formal analysis, Funding acquisition, Investigation, Resources, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by the National High Level Hospital Clinical Research Funding (Youth Clinical Research Project of Peking University First Hospital); Funder: Ya Wen Zhao; Grant number: 2024YC27.

Acknowledgments

We thank Amanda Holland, PhD, from Liwen Bianji (Edanz) (www.liwenbianji.cn) for editing the English text of a draft of this manuscript.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1447182/full#supplementary-material

References

1. Allenbach Y, Mammen AL, Benveniste O, Stenzel W. 224th ENMC International Workshop:: Clinico-sero-pathological classification of immune-mediated necrotizing myopathies Zandvoort, The Netherlands, 14-16 October 2016. *Neuromuscular disorders: NMD*. (2018) 28:87–99. doi: 10.1016/j.nmd.2017.09.016

2. Merlonghi G, Antonini G, Garibaldi M. Immune-mediated necrotizing myopathy (IMNM): A myopathological challenge. *Autoimmun Rev*. (2022) 21:102993. doi: 10.1016/j.autrev.2021.102993

3. Wang L, Liu L, Hao H, Gao F, Liu X, Wang Z, et al. Myopathy with anti-signal recognition particle antibodies: clinical and histopathological features in Chinese patients. *Neuromuscular disorders: NMD*. (2014) 24:335–41. doi: 10.1016/j.nmd.2014.01.002

4. Ohnuki Y, Suzuki S, Uruha A, Oyama M, Suzuki S, Kulski JK, et al. Association of immune-mediated necrotizing myopathy with HLA polymorphisms. *Hla*. (2023) 101:449–57. doi: 10.1111/tan.v101.5

5. Ma X, Xu L, Ji S, Li Y, Bu B. The clinicopathological distinction between seropositive and seronegative immune-mediated necrotizing myopathy in China. *Front neurol*. (2021) 12:670784. doi: 10.3389/fneur.2021.670784

6. Bergua C, Chiavelli H, Allenbach Y, Arouche-Delaperche L, Arnoult C, Bourdenet G, et al. *In vivo* pathogenicity of IgG from patients with anti-SRP or anti-HMGCR autoantibodies in immune-mediated necrotising myopathy. *Ann rheumatic diseases*. (2019) 78:131–9. doi: 10.1136/annrheumdis-2018-213518

7. Julien S, Vadysirisack D, Sayegh C, Ragunathan S, Tang Y, Briand E, et al. Prevention of anti-HMGCR immune-mediated necrotising myopathy by C5 complement inhibition in a humanised mouse model. *Biomedicines*. (2022) 10:2036. doi: 10.3390/biomedicines10082036

8. Julien S, van der Woning B, De Ceuninck L, Briand E, Jaworski T, Roussel G, et al. Efgartigimod restores muscle function in a humanized mouse model of immune-mediated necrotizing myopathy. *Rheumatol (Oxford England)*. (2023) 62:4006–11. doi: 10.1093/rheumatology/kead298

9. Xiong A, Yang G, Song Z, Xiong C, Liu D, Shuai Y, et al. Rituximab in the treatment of immune-mediated necrotizing myopathy: a review of case reports and case series. *Ther Adv neurological Disord*. (2021) 14:1756286421998918. doi: 10.1177/1756286421998918

10. Pinal-Fernandez I, Parks C, Werner JL, Albayda J, Paik J, Danoff SK, et al. Longitudinal course of disease in a large cohort of myositis patients with autoantibodies recognizing the signal recognition particle. *Arthritis Care Res*. (2017) 69:263–70. doi: 10.1002/acr.22920

11. Suzuki S, Nishikawa A, Kuwana M, Nishimura H, Watanabe Y, Nakahara J, et al. Inflammatory myopathy with anti-signal recognition particle antibodies: case series of 100 patients. *Orphanet J rare diseases*. (2015) 10:61. doi: 10.1186/s13023-015-0277-y

12. Chen S, Yang J, He D, Fu J, Lai X, Zhao B, et al. Anti-SRP immune-mediated necrotizing myopathy responsive to ofatumumab: a case report. *Front Immunol*. (2023) 14:1301109. doi: 10.3389/fimmu.2023.1301109

13. Cui BB, Tian YR, Ma XY, Yin G, Xie Q. Belimumab for immune-mediated necrotizing myopathy associated with anti-SRP antibodies: A case report and retrospective review of patients treated with anti-B-cell therapy in a single center and literature. *Front Immunol*. (2021) 12:777502. doi: 10.3389/fimmu.2021.777502

14. Kruse RL, Albayda J, Vozniak SO, Lawrence CE, Goel R, Lokhandwala PM, et al. Therapeutic plasma exchange for the treatment of refractory necrotizing autoimmune myopathy. *J Clin apheresis*. (2022) 37:253–62. doi: 10.1002/jca.21968

15. Heo YA. Efgartigimod: first approval. *Drugs*. (2022) 82:341–8. doi: 10.1007/s40265-022-01678-3

16. Howard JF Jr., Bril V, Vu T, Karam C, Peric S, Margania T, et al. Safety, efficacy, and tolerability of efgartigimod in patients with generalised myasthenia gravis (ADAPT): a multicentre, randomised, placebo-controlled, phase 3 trial. *Lancet Neurol*. (2021) 20:526–36. doi: 10.1016/S1474-4422(21)00159-9

17. Werner JL, Christopher-Stine L, Ghazarian SR, Pak KS, Kus JE, Daya NR, et al. Antibody levels correlate with creatine kinase levels and strength in anti-3-hydroxy-3-methylglutaryl-coenzyme A reductase-associated autoimmune myopathy. *Arthritis rheumatism*. (2012) 64:4087–93. doi: 10.1002/art.v64.12

18. Tiniakou E, Pinal-Fernandez I, Lloyd TE, Albayda J, Paik J, Werner JL, et al. More severe disease and slower recovery in younger patients with anti-3-hydroxy-3-methylglutaryl-coenzyme A reductase-associated autoimmune myopathy. *Rheumatol (Oxford England)*. (2017) 56:787–94. doi: 10.1093/rheumatology/kew470

19. Meyer A, Troyanov Y, Drouin J, Oligny-Longpré G, Landon-Cardinal O, Hoa S, et al. Statin-induced anti-HMGCR myopathy: successful therapeutic strategies for corticosteroid-free remission in 55 patients. *Arthritis Res Ther*. (2020) 22:5. doi: 10.1186/s13075-019-2093-6

20. Mercuri E, Cini C, Pichiecchio A, Allsop J, Counsell S, Zolkipli Z, et al. Muscle magnetic resonance imaging in patients with congenital muscular dystrophy and Ullrich phenotype. *Neuromuscul Disord*. (2003) 13:554–8. doi: 10.1016/S0960-8966(03)00091-9

21. Morrow JM, Matthews E, Raja Rayan DL, Fischmann A, Sinclair CD, Reilly MM, et al. Muscle MRI reveals distinct abnormalities in genetically proven non-dystrophic myotonias. *Neuromuscular disorders: NMD*. (2013) 23:637–46. doi: 10.1016/j.nmd.2013.05.001

22. Ruiz-Arruza I, Barbosa C, Ugarte A, Ruiz-Irastorza G. Comparison of high versus low-medium prednisone doses for the treatment of systemic lupus erythematosus patients with high activity at diagnosis. *Autoimmun Rev*. (2015) 14:875–9. doi: 10.1016/j.autrev.2015.05.011

23. Aggarwal R, Rider LG, Ruperto N, Bayat N, Erman B, Feldman BM, et al. 2016 American College of Rheumatology/European League Against Rheumatism criteria for minimal, moderate, and major clinical response in adult dermatomyositis and polymyositis: An International Myositis Assessment and Clinical Studies Group/Paediatric Rheumatology International Trials Organisation Collaborative Initiative. *Ann rheumatic Dis*. (2017) 76:792–801. doi: 10.1136/annrheumdis-2017-211400

24. Szczesny P, Barsotti S, Nennesmo I, Danielsson O, Dastmalchi M. Screening for anti-HMGCR antibodies in a large single myositis center reveals infrequent exposure to statins and diversiform presentation of the disease. *Front Immunol*. (2022) 13:866701. doi: 10.3389/fimmu.2022.866701

25. Hou Y, Shao K, Yan Y, Dai T, Li W, Zhao Y, et al. Anti-HMGCR myopathy overlaps with dermatomyositis-like rash: a distinct subtype of idiopathic inflammatory myopathy. *J neurol*. (2022) 269:280–93. doi: 10.1007/s00415-021-10621-7

26. Konno S, Uchi T, Kihara H, Sugimoto H. Real-world case series of efgartigimod for Japanese generalized myasthenia gravis: well-tailored treatment cycle intervals contribute to sustained symptom control. *Biomedicines*. (2024) 12:1214. doi: 10.3390/biomedicines12061214

27. van Steen C, Celico L, Spaepen E, Hagenacker T, Meuth SG, Ruck T, et al. Efgartigimod and ravulizumab for treating acetylcholine receptor auto-antibody-positive (AChR-ab+) generalized myasthenia gravis: indirect treatment comparison. *Adv Ther*. (2024) 41:2486–99. doi: 10.1007/s12325-024-02856-3

28. Nelke C, Schroeter CB, Theissen L, Preusse C, Pawlitzki M, Räuber S, et al. Senescent fibro-adipogenic progenitors are potential drivers of pathology in inclusion body myositis. *Acta neuropathologica*. (2023) 146:725–45. doi: 10.1007/s00401-023-02637-2

29. Broome CM, McDonald V, Miyakawa Y, Carpenedo M, Kuter DJ, Al-Samkari H, et al. Efficacy and safety of the neonatal Fc receptor inhibitor efgartigimod in adults with primary immune thrombocytopenia (ADVANCE IV): a multicentre, randomised, placebo-controlled, phase 3 trial. *Lancet (London England)*. (2023) 402:1648–59. doi: 10.1016/S0140-6736(23)01460-5

30. Sivadasan A, Bril V. Clinical efficacy and safety of efgartigimod for treatment of myasthenia gravis. *Immunotherapy*. (2023) 15:553–63. doi: 10.2217/imt-2022-0298

31. Ulrichts P, Guglietta A, Dreier T, van Bragt T, Hanssens V, Hofman E, et al. Neonatal Fc receptor antagonist efgartigimod safely and sustainably reduces IgGs in humans. *J Clin Invest*. (2018) 128:4372–86. doi: 10.1172/JCI97911

32. Goebeler M, Bata-Csörgő Z, De Simone C, Didona B, Remenyik E, Reznichenko N, et al. Treatment of pemphigus vulgaris and foliaceus with efgartigimod, a neonatal Fc receptor inhibitor: a phase II multicentre, open-label feasibility trial. *Br J Dermatol*. (2022) 186:429–39. doi: 10.1111/bjd.v186.3

33. Howard JF Jr., Bril V, Vu T, Karam C, Peric S, De Bleecker JL, et al. Long-term safety, tolerability, and efficacy of efgartigimod (ADAPT+): interim results from a phase 3 open-label extension study in participants with generalized myasthenia gravis. *Front neurol*. (2023) 14:1284444. doi: 10.3389/fneur.2023.1284444

34. Zhang C, Lin Y, Kuang Q, Li H, Jiang Q, Yang X. Case report: A highly active refractory myasthenia gravis with treatment of telitacicept combined with efgartigimod. *Front Immunol*. (2024) 15:1400459. doi: 10.3389/fimmu.2024.1400459

frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 22 November 2024
DOI 10.3389/fimmu.2024.1375915

OPEN ACCESS

EDITED BY
Alex Tsoi,
University of Michigan, United States

REVIEWED BY
Shi Xue Dai,
Guangdong Provincial People's Hospital, China
Richa Rai,
Icahn School of Medicine at Mount Sinai, United States

*CORRESPONDENCE
Juan-ni Zeng
575826199@qq.com

RECEIVED 24 January 2024
ACCEPTED 04 November 2024
PUBLISHED 22 November 2024

CITATION
Hu Y-z, Chen Z, Zhou M-h, Zhao Z-y, Wang X-y, Huang J, Li X-t and Zeng J-n (2024) Global and regional genetic association analysis of ulcerative colitis and type 2 diabetes mellitus and causal validation analysis of two-sample two-way Mendelian randomization. *Front. Immunol.* 15:1375915. doi: 10.3389/fimmu.2024.1375915

Global and regional genetic association analysis of ulcerative colitis and type 2 diabetes mellitus and causal validation analysis of two-sample two-way Mendelian randomization

Yan-zhi Hu[1], Zhe Chen[2], Ming-han Zhou[1], Zhen-yu Zhao[3], Xiao-yan Wang[1], Jun Huang[1], Xin-tian Li[1] and Juan-ni Zeng[1,4]*

[1]The Second Affiliated Hospital of Hunan University of Chinese Medicine, Changsha, China, [2]Department of Thoracic Surgery, The Second Xiangya Hospital, Central South University, Changsha, China, [3]College of Traditional Chinese Medicine, Hunan University of Chinese Medicine, Changsha, China, [4]Laboratory of Vascular Biology and Translational Medicine, Medical School, Hunan University of Chinese Medicine, Changsha, China

Background: Clinical co-occurrence of UC (Ulcerative Colitis) and T2DM (Type 2 Diabetes Mellitus) is observed. The aim of this study is to investigate the potential causal relationship between Ulcerative Colitis (UC) and Type 2 Diabetes Mellitus (T2DM) using LDSC and LAVA analysis, followed by genetic verification through TSMR, providing insights for clinical prevention and treatment.

Methods: Genetic loci closely related to T2DM were extracted as instrumental variables from the GWAS database, with UC as the outcome variable, involving European populations. The UC data included 27,432 samples and 8,050,003 SNPs, while the T2DM data comprised 406,831 samples and 11,914,699 SNPs. LDSC and LAVA were used for quantifying genetic correlation at both global (genome-wide) and local (genomic regions) levels. MR analysis was conducted using IVW, MR-Egger regression, Weighted median, and Weighted mode, assessing the causal relationship between UC and diabetes with OR values and 95% CI. Heterogeneity and pleiotropy were tested using Egger-intercept, MR-PRESSO, and sensitivity analysis through the "leave-one-out" method and Cochran Q test. Subsequently, a reverse MR operation was conducted using UC as the exposure data and T2DM as the outcome data for validation.

Results: Univariable and bivariable LDSC calculated the genetic correlation and potential sample overlap between T2DM and UC, resulting in rg = -0.0518, se = 0.0562, P = 0.3569 with no significant genetic association found for paired traits. LAVA analysis identified 9 regions with local genetic correlation, with 6negative and 3 positive associations, indicating a negative correlation between T2DM and UC. MR analysis, with T2DM as the exposure and UC as the outcome, involved 34 SNPs as instrumental variables. The OR values and 95% CI from IVW, MR-Egger, Weighted median, and Weighted mode were 0.917 (0.848~0.992), 0.949 (0.800~1.125), 0.881 (0.779~0.996), 0.834(0.723~0.962) respectively, with IVW P-value < 0.05, suggesting a negative causal relationship between T2DM and UC. MR-Egger regression showed an intercept of -0.004 with a standard error of

0.009, P = 0.666, and MR-PRESSO Global Test P-value > 0.05, indicating no pleiotropy and no outliers detected. Heterogeneity tests showed no heterogeneity, and the "leave-one-out" sensitivity analysis results were stable. With UC as the exposure and T2DM as the outcome, 32 SNPs were detected, but no clear causal association was found.

Conclusion: There is a causal relationship between T2DM and UC, where T2DM reduces the risk of UC, while no significant causal relationship was observed from UC to T2DM.

KEYWORDS

UC, T2DM, Mendelian randomization, LDSC analysis, LAVA analysis

1 Introduction

At present, the prevalence of chronic diseases such as diabetes, obesity, cardiovascular diseases, and inflammatory bowel disease poses a significant threat to global health. Diabetes Mellitus (DM) is a complex chronic illness characterized by glucose metabolic dysfunction caused by either absolute or relative insulin deficiency. The global incidence of diabetes is on the rise, with a total prevalence of 10.5% among adults aged 20-79 years, reaching 537 million diabetic patients worldwide by 2021, and it's estimated to increase to 783.2 million by 2045 (1). Type 2 Diabetes Mellitus (T2DM) is the most common form, accounting for approximately 90-95% of all cases. T2DM typically presents with various comorbidities and long-term complications, including cardiovascular diseases, retinopathy, nephropathy, and neurological disorders, which have garnered significant attention. Moreover, the emergence of new complications such as COVID-19, pulmonary fibrosis, and gastrointestinal diseases is increasingly common (2). The onset of T2DM is closely associated with genetic factors, aging, and unhealthy lifestyle habits. It's generally believed that the pathophysiology of T2DM is rooted in impaired insulin responsiveness, known as insulin resistance (IR) (3), coupled with inadequate insulin secretion. Research indicates that the development of insulin resistance is linked to endotoxemia, chronic inflammatory responses, short-chain fatty acid, and bile acid metabolism, with a notable imbalance in the gut microbiota of diabetic patients (4). This dysbiosis of gut microbiota, resulting from changes in microbial composition, bacterial metabolic activity, or local distribution, can trigger a decline in immune function, chronic inflammatory responses, and an imbalance in energy metabolism, leading to metabolic disorder and insulin resistance, ultimately contributing to the development of T2DM (5).

Ulcerative colitis (UC), a chronic, non-specific inflammatory bowel disease (6), has increasingly become a common and intractable condition in the digestive system. Its primary clinical manifestations include recurrent diarrhea with mucosal bloody stool, with or without abdominal pain. Inflammation and ulcers can appear in various sections of the large intestine, predominantly affecting the rectum and sigmoid colon, and occasionally the ileum, leading to backwash ileitis. This condition can cause anemia, liver disease, arthropathy, mucocutaneous diseases, and eye disorders. Severe cases may develop toxic megacolon, intestinal perforation, and cancer. UC, along with Crohn's disease (CD), is categorized as inflammatory bowel disease (IBD), frequently observed in individuals with a high-fat diet preference (7). Ulcerative colitis (UC) has now emerged as a pervasive global health challenge, with its epidemiological trends evolving continuously. Research highlights a rapid escalation in the incidence of UC within low to middle-income nations. The disease manifests with comparable frequency in both males and females, predominantly affecting individuals aged between 2 and 40 years. However, there's an increasing prevalence of UC in the population over 60 years of age, who account for 20% of newly diagnosed cases. These shifting patterns underscore the imperative need for refining and globalizing preventive and therapeutic strategies for UC, to effectively address its dynamic disease burden (8). Past studies have attributed the etiology of UC to genetic (9), environmental (10), dietary (11, 12), and psychological factors (13). The pathogenesis primarily involves genetic predisposition, gut microbiome imbalance, immune response irregularities, imbalance of pro-inflammatory and anti-inflammatory factors, aberrant signaling pathways, hypercoagulable blood state, intestinal epithelial cell apoptosis, necroptosis, long non-coding RNA, and proteomics. Theories such as "autophagy-cytokine-bacteria-UC" and "intestinal loop poisoning" have been proposed (14–23). Clinically, patients with coexisting UC and T2DM exhibit disease-related characteristics, higher hospitalization rates, increased risk of concurrent infections, and poorer prognosis (1). When T2DM coexists with UC, fluctuations in blood sugar levels, combined with intestinal lesions in patients, hinder the intake and absorption of nutrients and accelerate their loss. Particularly during active phases of UC, symptoms such as fever and diarrhea can increase the body's metabolism, leading to an insufficient supply of

nutrients. Consequently, patients may exhibit varying degrees of malnutrition, significantly impacting their physical health and quality of life.

Linkage Disequilibrium Score Regression (LDSC) is a statistical method used in genetic research (24), widely applied in Genome-Wide Association Studies (GWAS). Its primary purpose is to estimate the degree of genetic influence on specific traits or diseases, known as genetic correlation. LDSC's key feature is the use of linkage disequilibrium (LD) scores to correct associations between multiple genetic markers. LD describes the co-inheritance patterns of genetic markers (like Single Nucleotide Polymorphisms, SNPs) within a population. In GWAS, the abundance of genetic markers and their LD can cause statistical confusion, affecting accurate estimations of genetic correlation. LDSC analysis enables researchers to more precisely estimate the contribution of genetic variations to specific traits or disease risks. This method is significant in understanding the genetic background of complex traits, revealing genetic risk factors, and providing a theoretical foundation for personalized medicine and gene therapy.

LAVA (Local Analysis of Variant Association) refers to a statistical method or tool used in genetics and bioinformatics (25). It is designed to analyze genetic variations, such as Single Nucleotide Polymorphisms (SNPs), whether in localized regions or across the entire genome. The primary aim of this analysis is to identify associations between genetic variations and specific traits or diseases. Employed in Genome-Wide Association Studies (GWAS), researchers use LAVA to pinpoint genetic variations linked to particular traits or diseases. Focusing on localized areas, LAVA provides in-depth insights, aiding researchers in accurately locating specific variants or groups of variants contributing to disease risks or manifestations. LAVA typically involves the use of statistical algorithms and computational tools to process large genomic datasets and can be integrated with other bioinformatics methods to enhance the analysis and interpretation of genetic data.

Randomized Controlled Trials (RCTs), due to various constraints, are challenging to implement effectively in clinical settings. Observational experimental methods, influenced by confounding factors and reverse causality, tend to yield biased results with relatively low credibility. In 1986, Martijn B. Katan proposed that different alleles determine varying Apolipoprotein E subtypes, influencing cancer incidence rates through cholesterol level regulation, laying the groundwork for the concept of Mendelian Randomization (MR) (26). In 2004, Thomas and Conti introduced the use of genetic information as instrumental variables for causal inference in epidemiology (27). MR employs Single Nucleotide Polymorphisms (SNPs), or genetic variants, as instrumental variables. Based on Mendel's laws of inheritance, genetic variations are randomly distributed to offspring during meiosis and remain unchanged thereafter. This directional and invariant nature of MR reduces the influence of reverse causation and confounding factors, as compared to observational studies (28, 29). Two-sample Mendelian Randomization (TSMR), involving data from two independent databases, enhances sample size and the availability of exposure and outcome sources. Recently, with the release of numerous large-scale Genome-Wide Association Studies (GWAS), MR has become a viable method for assessing disease risk factors. To circumvent the limitations of observational studies, we use LDSC and LAVA analysis to explore the genetic correlation between T2DM and UC, followed by MR analysis for bidirectional causal verification, investigating potential mechanisms influencing this correlation. All original studies have received ethical approval, so additional ethical approval or informed consent for this research is not required. The process flowchart of the analysis is shown in Figure 1.

2 Materials and methods

2.1 Study design

This study employs LDSC (LD Score Regression) and LAVA (Local Analysis of Variant Association) to estimate the genetic correlation between Type 2 Diabetes Mellitus (T2DM) and Ulcerative Colitis (UC). Utilizing T2DM as the exposure factor, Single Nucleotide Polymorphisms (SNPs) significantly related to T2DM are used as instrumental variables (IVs), with UC as the outcome variable. The process involves reverse operation verification using the TwoSampleMR package in R for causal association analysis, including Cochran Q heterogeneity test, pleiotropy test, and sensitivity analysis to validate the results. The selection of IVs is based on three criteria: significant association with T2DM, irrelevance to UC, and exclusive influence on UC through T2DM. These criteria are independent and indispensable, determining the suitability of IVs for analysis.

2.2 Source of data

All data in this study are sourced from publicly available Genome-Wide Association Studies (GWAS) and the IEU GWAS database. We retrieved GWAS summary statistics, selecting SNPs significantly associated with T2DM as genetic instrumental variables from the IEU GWAS database. For Ulcerative Colitis (UC), GWAS summary statistics related to UC were selected from large-scale published GWAS meta-analyses, extracting gene outcome associations (30). The UC data includes a sample size of 27,432 individuals with 8,050,003 SNPs, while the T2DM data comprises 406,831 individuals with 11,914,699 SNPs.

2.3 LDSC analysis

To assess the shared genetic components between Type 2 Diabetes Mellitus (T2DM) and Ulcerative Colitis (UC), we conducted a global genetic correlation analysis using bivariate linkage disequilibrium (LD) score regression (LDSC), with values ranging from -1 to 1. LDSC estimates the heritability of individual traits or genetic correlation between traits by constructing a regression relationship between LD scores and GWAS test statistics. LD scores are calculated using European ancestry reference data from the 1000 Genomes Project, limited to 1.2 million well-qualified HapMap3 SNPs, excluding SNPs in the MHC region due to their complex LD patterns affecting genetic correlation estimates. To address unknown sample overlaps in LDSC analysis, we did not restrict the intercept

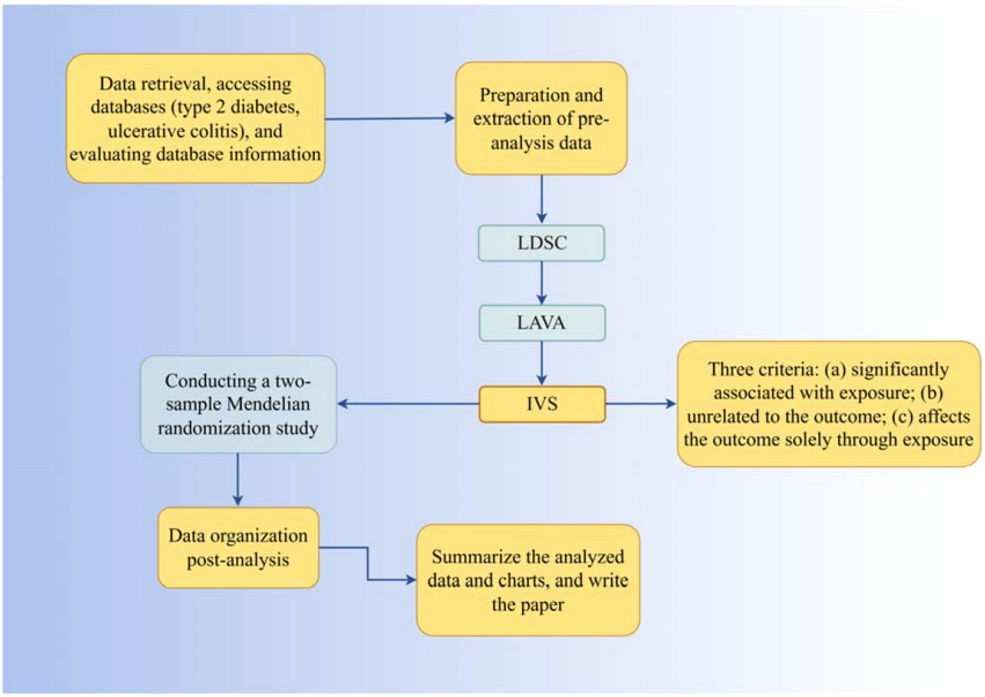


FIGURE 1
The process flowchart of the analysis.

term, using it to assess potential population stratification in individual trait GWAS or sample overlap between pairs of GWAS data. A significant threshold is determined as a *P*-value less than 0.05 after false discovery rate (FDR) correction (Benjamini-Hochberg method), equating to a q value < 0.05.

2.4 LAVA analysis

The global genetic correlation estimated by LDSC originates from the aggregate information of all variations in the genome. However, due to the complexity of genetic variations and their associations with diseases, different regions contribute variably in magnitude and direction to the genetic correlation. Moreover, significant disparities exist in the genetic correlation of two traits in different regions, particularly where opposing regional genetic correlations may neutralize each other. This can reduce the global genetic correlation between traits, obscuring potential pleiotropic effects. Therefore, we employ Local Analysis of Variant Association (LAVA) to estimate the genetic correlation between T2DM and UC in independent local regions of the genome (25). LAVA is conducted within 2,495 independent LD blocks previously delineated, with LD estimations based on the 1000G EUR reference. The significant threshold is set as a P-value less than 0.05 following false discovery rate (FDR) correction (Benjamini-Hochberg method), corresponding to a q value < 0.05.

Based on the chromosomal segments identified by LAVA analysis (details are provided in Supplementary Material), we conducted Bayesian colocalization analysis on diseases showing significant local genetic correlation after FDR multiple correction, in order to further clarify whether the two phenotypes share the same causal variant within a given region. Unfortunately, within the CHR segments provided by the LAVA analysis, no significant shared causal variant loci were observed between T2DM and UC (see Appendix for details), suggesting that although there may be some common genetic factors between T2DM and UC, they may be caused by different genetic variations on the studied chromosomal segments. That is, the two traits may be controlled by different regulatory regions of the same gene, appearing to be genetically related, but showing no significant shared genetic variation at the expression level, hence no colocalization signal was detected in this segment. Combined with the positive results of the MR analysis, it can be explained that the relationship between the studied traits is entirely due to the impact of exposure on the outcome. Of course, considering the robustness of the LAVA analysis results, future studies will focus on further explaining these results by increasing sample size, integrating other biological data, using more precise statistical methods, and possible experimental validation.

2.5 Selection and validation of IVs

For Type 2 Diabetes Mellitus (T2DM) exposure, the selection of Instrumental Variables (IVs) starts with T2DM's database. The steps to determine the included IVs are as follows: 1. Initially select SNPs that meet the significance threshold ($P < 5\times10^{-8}$); 2. Exclude SNPs in linkage disequilibrium (LD), mainly based on the distance and r^2 value between each SNP (r^2 < 0.01, distance > 10000kb);

3. Further remove palindromic SNPs from the determined SNP list, especially those with lower effect allele frequency (< 0.58) in the outcome, as it's challenging to discern the strand orientation of such SNPs. 4. Eliminate the influence of other confounding factors. Considering UC's complexity, it's crucial to account for common confounders like Irritable Bowel Syndrome, hyperlipidemia, body weight, and fatty liver, which might affect its occurrence as intermediate phenotypes. To avoid IVs affecting the outcome through common confounders, verify the SNP through the Pheno Scanner database (version 2: http://www.phenoscanner.medschl.cam.ac.uk/), delete SNP: rs11651052 (Prostate cancer), rs2844623 (Crohn's disease), and similarly, for the reverse scenario.

2.6 MR analytics

This study's statistical analysis is based on R software (version 4.3.0, R Foundation for Statistical Computing, Vienna, Austria). The focal analysis relies on the Two Sample MR (TSMR) R package developed by Gibran Hemani and colleagues (31, 32). We employed four methods to estimate effects: Inverse Variance Weighted method (IVW) (33), MR-Egger regression (34), Weighted median (35), and Weighted mode (36). The primary outcome measure is the Odds Ratio (OR), including a 95% Confidence Interval (CI). Statistical results encompass the overall effect size, standard error (yielding the final OR and 95% CI), and significance values, with a default two-sided test $P<0.05$ considered statistically significant. Scatter plots derived from statistical tables illustrate these results. MR-PRESSO and MR-Egger regression methods calculate the magnitude of pleiotropy, presented graphically via weighted linear regression, where the intercept's absolute value indicates the extent of pleiotropy; a pleiotropy $P>0.05$ is not statistically significant (37). Sensitivity analysis employs the "leave-one-out" approach from the R package, reanalyzing results after sequentially excluding individual SNPs and visualizing the impact of each SNP on outcomes via forest plots to assess result stability. Heterogeneity is tested using Cochran's Q test, with a $P>0.05$ indicating no significant heterogeneity, and results are presented in statistical tables.

3 Result

3.1 Genetic correlation analysis

We employed both LDSC and LAVA to quantify the pairwise genetic correlation at global (i.e., across the entire genome) and local (i.e., within specific genomic regions) levels. The LAVA analysis estimated local genetic correlations across 2,495 genomic regions.

Initially, we utilized bivariate LDSC to calculate the genetic correlation and potential sample overlap between T2DM and UC. After adjusting all P-values (FDR $P<0.05$), no significant genetic associations were found for the paired traits, with rg = -0.0518, se = 0.0562, $P = 0.35699$, as shown in Table 1. Subsequently, our LAVA analysis of local genetic correlations indicated, after FDR multiple adjustments, that 9 regions exhibited local genetic correlations for at least one pair of traits (Figure 2), with 33.33% positive and 66.67% negative correlations. Specifically, 6 regions were negatively and 3 positively significantly associated, leading us to conclude that LAVA results show a negative correlation between T2DM and UC incidence. The inheritance of disease is a multifaceted and intricate process, interwoven with a multitude of genetic and environmental interactions. Global analysis through LDSC, constrained by sample sizes, statistical methodologies, or the inherent genetic complexity of the diseases themselves, sometimes fails to significantly reveal correlations. In contrast, LAVA's local analysis, with its focus on specific genes or regions, possesses the capability to unearth more profound genetic mechanisms. Certain genetic effects may only be significant in specific gene areas or populations, nuances that global analysis might overlook, while local analysis can investigate these effects with greater precision. Moreover, global analysis might be influenced by sample bias, such as insufficient diversity or selective recruitment, potentially obscuring the true genetic associations. On the other hand, local analysis often employs more representative samples and more precise genetic markers. These factors could account for the negative findings in LDSC analysis versus the negative correlation in LAVA results.

3.2 Bidirectional MR analysis of UC by T2DM

3.2.1 Status of instrumental variables

Initially, with Type 2 Diabetes Mellitus (T2DM) as the exposure and Ulcerative Colitis (UC) as the outcome, we utilized R software to select genome-wide significant SNP loci according to our screening criteria. To mitigate the impact of common confounders via Instrumental Variables (IVs), we further validated these SNPs through the Pheno Scanner database, resulting in 34 SNPs as IVs. Similarly, with UC as the exposure and T2DM as the outcome, we identified 32 SNPs as IVs.

TABLE 1 LDSC results.

Genetic correlation				
Trait pair	Genetic correlation (SE)	P value for LDSC	Intercept (SE)	P value for Intercept
T2DM-UC	-0.0518 (0.0562)	0.3569	0.0112 (0.0067)	0.095

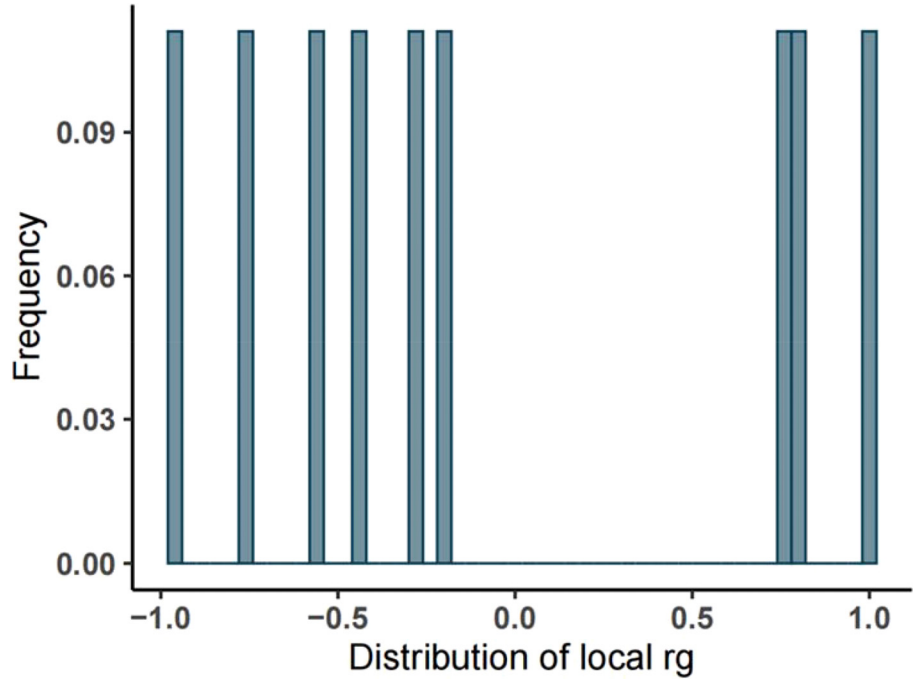


FIGURE 2
Frequency distribution of local heritability of T2DM and UC. (Note: The abscissa represents the heritability, the ordinate represents the frequency, and the rg ranges from -1 to 1).

3.2.2 MR analysis results

The Mendelian Randomization (MR) analysis for the relationship between Type 2 Diabetes Mellitus (T2DM) and Ulcerative Colitis (UC) was conducted using the TwoSampleMR package, employing methods such as Inverse Variance Weighted (IVW), MR-Egger regression, Weighted median, and Weighted mode. The results, detailed in Table 2, reveal Odds Ratios (OR) and 95% Confidence Intervals (CI) as follows: 0.917 (0.848~0.992), 0.949 (0.800~1.125), 0.881 (0.779~0.996), and 0.834(0.723~0.962). These findings indicate that having T2DM reduces the risk of developing UC, with *P*-values from the four tests being 0.031, 0.550, 0.042, and 0.018, respectively. The IVW result, significant at $P<0.05$, and the consistent direction of β values across all methods, validate the conclusion that T2DM lowers the risk of UC, suggesting a causal relationship. The MR-PRESSO result, with a *P*-value < 0.05, reinforces the robustness of this positive outcome, further substantiating the negative correlation found in the LAVA analysis.

The Mendelian Randomization (MR) analysis of Ulcerative Colitis (UC) on Type 2 Diabetes Mellitus (T2DM) was performed using the same methodology, with results presented in Table 3. The Odds Ratios (OR) and 95% Confidence Intervals (CI) are reported as 1.018 (0.991~1.046), 0.993 (0.911~1.083), 1.030 (0.995~1.065), and 1.039 (0.985~1.096). The P-values for the four tests are 0.187, 0.879, 0.096, and 0.171, respectively. With the IVW result being greater than 0.05, the difference is not statistically significant. Hence, the MR analysis suggests no evident causal relationship between the occurrence of UC and the development of T2DM.

3.2.3 Sensitivity analysis result

This study meticulously adhered to the selection criteria for instrumental variables, thus reducing the likelihood of false-negative results. For the MR analysis of T2DM's impact on UC, heterogeneity tests were conducted. The Q-values and Q*P*-values for IVW and MR-Egger were 19.933 (0.952) and 20.122 (0.962), respectively, both exceeding 0.05, indicating no significant heterogeneity. The results have been visualized in Figure 3.

The study employed MR-Egger regression's intercept to assess potential pleiotropy. The Egger-intercept value was -0.004, close to zero, with SE = 0.009 and $P = 0.666$, suggesting minimal pleiotropy.

TABLE 2 Results of MR Analysis of T2DM versus UC.

Method	BETA	SE	OR (95% CI)	P value
IVW	-0.086	0.040	0.917 (0.848~0.992)	0.031
MR-Egger	-0.053	0.087	0.949 (0.800~1.125)	0.550
Weighted median	-0.127	0.063	0.881 (0.779~0.996)	0.042
Weighted mode	-0.181	0.073	0.834(0.723~0.962)	0.018

TABLE 3 Results of MR Analysis of UC for T2DM.

Method	BETA	SE	OR (95% CI)	P value
IVW	0.018	0.014	1.018 (0.991~1.046)	0.187
MR-Egger	-0.065	0.044	0.993 (0.911~1.083)	0.879
Weighted median	-0.005	0.017	1.030 (0.995~1.065)	0.096
Weighted mode	-0.002	0.027	1.039 (0.985~1.096)	0.171

MR-PRESSO, supplementing the primary IVW results, showed a consistent direction of the Causal (beta effect value), with a *P*-value less than 0.05. The Global Test *P*value of 0.954 indicates no horizontal pleiotropy, affirming the MR results are free from multi-effect interference. Sensitivity analysis using the "Leave-one-out" method visualized the IVW results in Figure 4. After sequentially excluding individual SNPs, the remaining SNPs' IVW effect values showed no significant fluctuations, aligning closely with the red dot in the figure, and all *P*-values were above 0.05. This indicates the absence of SNPs with strong influence in the instrumental variables, confirming the stability and reliability of the IVW results. No outliers were detected in the MR-PRESSO process. The final MR results are visualized in Figure 5.

Based on the chromosomal segments identified by LAVA analysis (details are provided in Appendix), we conducted Bayesian colocalization analysis on diseases showing significant local genetic correlation after FDR multiple correction, in order to further clarify whether the two phenotypes share the same causal variant within a given region. Unfortunately, within the CHR segments provided by the LAVA analysis, no significant shared causal variant loci were observed between T2DM and UC (details see Supplementary Figures 1–9), suggesting that although there may be some common genetic factors between T2DM and UC, they may be caused by different genetic variations on the studied chromosomal segments. That is, the two traits may be controlled by different regulatory regions of the same gene, appearing to be genetically related, but showing no significant shared genetic variation at the expression level, hence no colocalization signal was detected in this segment. Combined with the positive results of the MR analysis, it can be explained that the relationship between the studied traits is entirely due to the impact of exposure on the outcome. Of course, considering the robustness of the LAVA analysis results, future studies will focus on further explaining these results by increasing sample size, integrating other biological data, using more precise statistical methods, and possible experimental validation.

4 Discussion

Significant progress in T2DM and UC comorbidity research has emerged, leveraging genomics and metabolomics. This work elucidates genetic and epigenetic links, and enhances our understanding of their epidemiology, pathogenesis, and therapeutic strategies.

The link between Ulcerative Colitis (UC) and Type 2 Diabetes Mellitus (T2DM) risk is inconsistent across studies. Jess et al.'s Danish cohort study and Kang's South Korean database analysis found an increased T2DM risk in UC patients (38–40), while a Taiwanese study did not (41). Surgical procedures, especially left-sided colon resections, may also heighten T2DM risk (42). These variations may stem from different study designs and populations. Our data hints at an inverse causal relationship from T2DM to UC, but not vice versa, as supported by LAVA and MR analyses. Both conditions share pathophysiological traits like gut microbiota disruption, epithelial barrier dysfunction, and inflammation (43, 44).

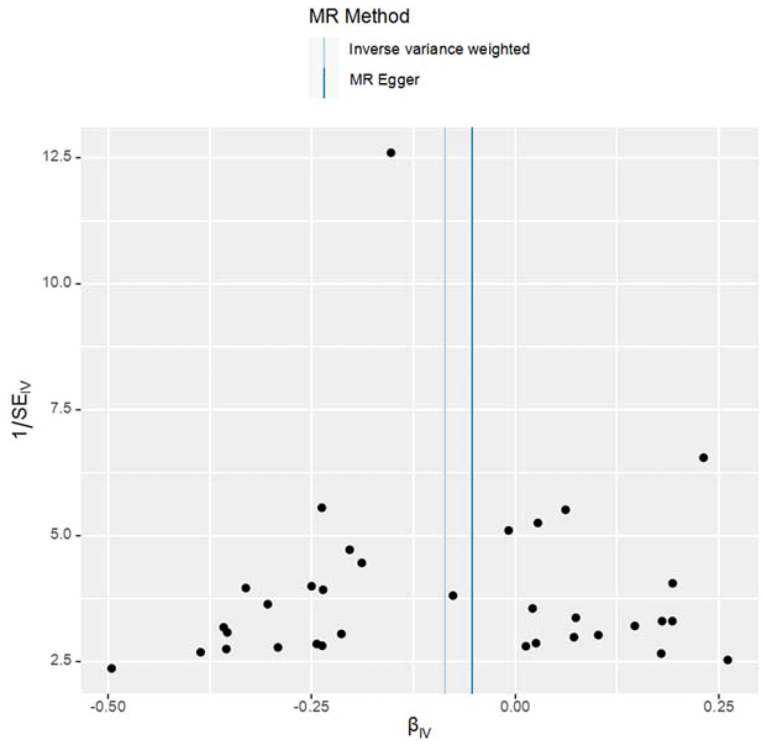


FIGURE 3
Funnel plot of the results of heterogeneity test for MR Method analysis.

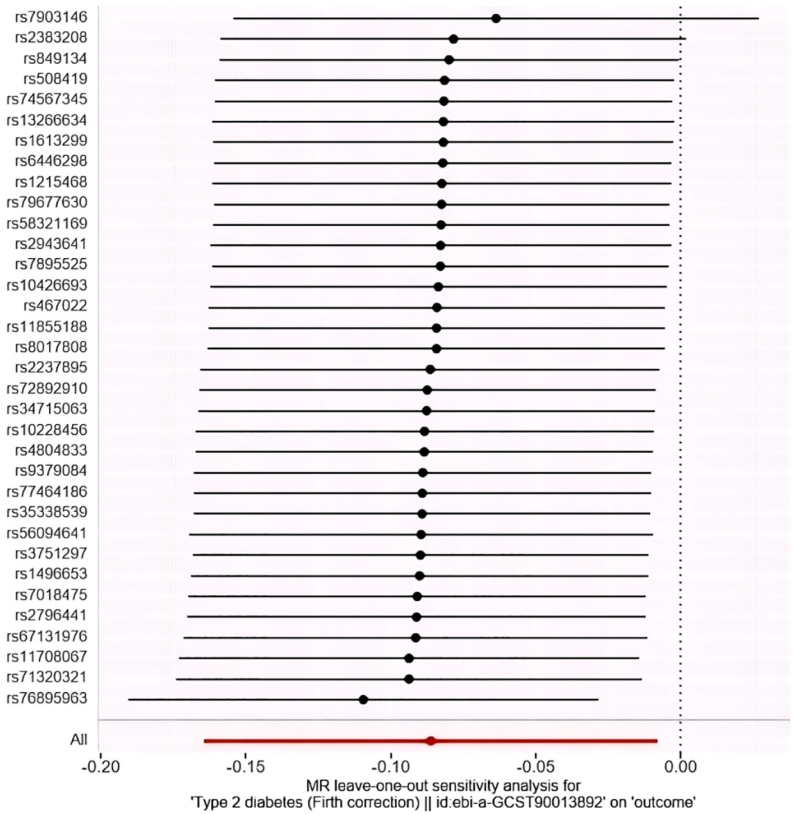


FIGURE 4
"Leave-one-out" method to visualize the results of IVW method. (Leave-one-out analysis refers to the MR Analysis after eliminating SNPS one by one, generally to see whether they are all significant, or whether the mean value is greater than/less than 0/1).

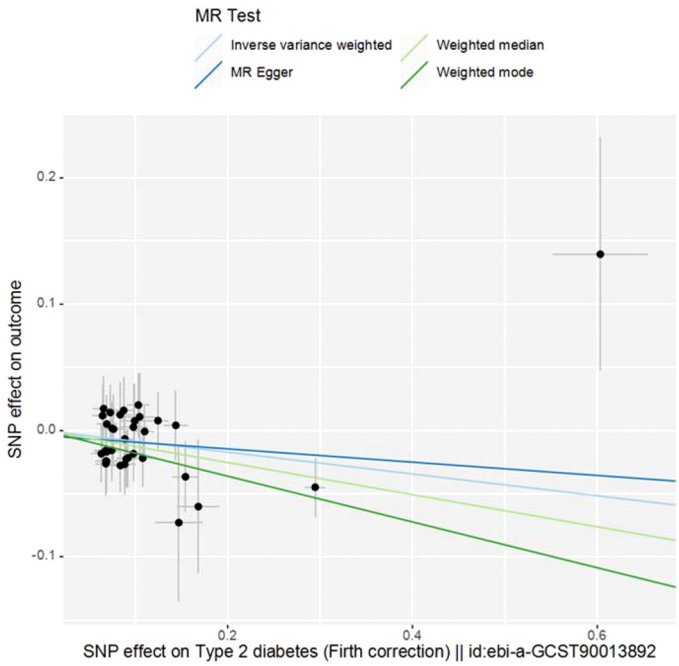


FIGURE 5
Scatter plot of MR Analysis. (Note: The X-axis represents the SNP effect on exposure, and the Y-axis represents the SNP effect on outcome. Slope less than 0, indicating that the exposure is a favorable factor for the outcome).

Our study, employing LDSC, LAVA, and MR methods, indicates a decreasing incidence of Ulcerative Colitis (UC) in patients with Type 2 Diabetes Mellitus (T2DM). Tseng CH's research highlights a dose-response relationship between metformin use and a diminished risk of UC, which could account for this observed trend (45). Metformin's potential to lower UC risk is thought to be mediated through its effects on improving insulin resistance, modulating gut microbiota, and reducing inflammation. It may alleviate intestinal inflammation in UC by suppressing pro-inflammatory cytokines and chemokines (46–49). Furthermore, metformin's ability to increase the presence of Akkermansia muciniphila, a bacterium associated with UC remission (50–53), suggests its contributory role in the lower incidence of UC among T2DM patients.

The use of Glucagon-Like Peptide-1 (GLP-1) may also play a role in the observed decrease in UC incidence among T2DM patients. GLP-1, a mild insulinotropic hormone, has diverse pharmacological effects, including stimulating insulin release in response to glucose, slowing stomach emptying, and reducing appetite (54). Modified GLP-1 receptor agonists, with improved potency and longer action, are effective in treating T2DM (55). Wenrui Wang's research indicates that GLP-1 can significantly reduce UC by inhibiting pro-inflammatory mediators, protecting against intestinal damage, and mitigating gut microbiota imbalance caused by DSS (56). Thus, GLP-1 treatment in T2DM patients might contribute to lower UC rates.

However, the impact of metformin and GLP-1 receptor agonists may involve gene-environment interactions. Genetic variants could affect UC risk in the context of drug exposure. Polygenic risk scores (PRS) might identify a shared genetic risk for T2DM and UC, with certain variants influencing medication responses. Since UC is immune-mediated and T2DM involves chronic inflammation, genetic factors could regulate immune responses, affecting both conditions. Metformin and GLP-1 receptor agonists may modify UC risk by altering immunoregulatory gene expression or function.

Irisin, a 112-amino acid peptide produced by skeletal muscle and derived from FNDC5, plays a pivotal role in a range of physiological responses and may mediate the connection between neurological health and physical activity (57). Huangfu Lixin's (58) research on UC patients found significantly reduced FNDC5 and Irisin levels in colonic tissue and serum, respectively, with Irisin inversely correlating with IL-12 and IL-23, echoing earlier findings of its link to inflammation (59, 60).

The study also identified significant gut microbiota imbalances in active UC, characterized by decreased lactobacillus and increased enterococcus, which correlated with inflammation severity. Irisin levels negatively associated with enterococcus and positively with lactobacillus, suggesting its role in UC pathogenesis.

In T2DM patients, higher FNDC5 levels were linked to older age and poor glycemic control (61). This proposes that elevated FNDC5 and Irisin's anti-inflammatory effects in T2DM may reduce cytokines, enhance gut microbiota health, and protect against UC by preventing dysbiosis and pathogen invasion.

Studies indicate that T2DM patients have significantly higher serum levels of TGF-β1 compared to non-diabetic individuals (62), Hefini conducted a study on the serum Transforming Growth Factor-β (TGF-β) levels in a cohort of 45 patients diagnosed with Type 2 Diabetes. The research revealed a substantial positive correlation between the onset of macroalbuminuria and the duration of diabetes. Furthermore, the analysis indicated that the serum TGF-β concentrations were substantially elevated in patients exhibiting macroalbuminuria (63). TGF-β, an anti-inflammatory cytokine primarily produced by activated T lymphocytes, B lymphocytes, and monocytes, promotes the synthesis and secretion of matrix proteins and epithelial repair (64), thereby potentially reducing the incidence of UC. Additionally, T2DM patients may adopt a healthier diet, opting for low-sugar and low-fat options, which could further decrease the potential risk of UC.

Addressing the needs of patients co-managing UC and T2DM presents a nuanced challenge in medical practice, owing to the intricacies in treatment and prognosis of these conditions. Stabilizing UC necessitates anti-inflammatory and immunomodulatory treatments, coupled with vigilant blood sugar level management. Moreover, prevention of cardiovascular diseases is critical, potentially entailing stricter control of blood pressure and lipid levels. Early identification of the interplay between UC and T2DM enables physicians to more accurately assess patient risk and tailor preventative and treatment strategies. Early interventions in high-risk groups, such as dietary improvements and increased physical activity, can significantly reduce disease risk. Comprehensive medical strategies, augmented by guidance on disease management, enhance patient adherence to treatment, thereby optimizing therapeutic outcomes.

5 Conclusion

In summary, this study utilized LDSC, LAVA, and TSMR analyses to explore the association between UC and T2DM. The results suggest a negative correlation, indicating that T2DM may reduce the risk of UC. This was further supported by genetic validation analysis. Factors contributing to this result include the use of metformin and GLP-1 in T2DM patients, increased Irisin secretion due to elevated serum FNDC5 levels, elevated serum TGF-β1, and dietary changes in T2DM patients. No significant causal association has been observed between UC and the risk of developing Type 2 Diabetes Mellitus. Based on publicly available GWAS data, this research avoids biases common in RCTs and observational studies, unlike previous observational studies. The findings are further supported by heterogeneity checks, with no evidence of heterogeneity or pleiotropy, and the "leave-one-out" sensitivity analysis confirms the reliability of the results. This research, unrestricted by ethical and financial constraints, provides insights into epidemiological etiology and may inform strategies for reducing the severity of UC in T2DM patients and aid in clinical treatment and risk prediction for patients with both conditions.

6 Deficiency and prospect

This study also has certain limitations, primarily including the following aspects: (1) The GWAS included in this study mainly comes from the European population, so the results may not necessarily match other ethnicities, which requires further GWAS

of more diverse ethnicities to validate the results or discover new applicable loci. (2) The GWAS data extracted for this analysis does not have stratified analysis results for gender, age, duration, etc., so specific information cannot be studied. Based on a large sample research design, obtaining more instrumental variables can enhance the reliability of the results, and subsequent research can delve into more studies on Asian populations. In the future, specific causal mechanisms between T2DM and UC can be further explored through experimental methods, including cellular biology factors, physicochemical factors, genetic factors, immune factors, etc. Addressing the needs of patients co-afflicted with Ulcerative Colitis (UC) and Type 2 Diabetes Mellitus (T2DM) presents a significant challenge in medical practice, stemming from the complexity inherent in the treatment and prognosis of both conditions. To maintain stability in UC, patients require anti-inflammatory and immunomodulatory therapies, while simultaneously managing glycemic levels. Additionally, the prevention of cardiovascular diseases is indispensable, potentially including stricter blood pressure and lipid control. Early identification of the interplay between UC and T2DM aids physicians in more accurately assessing patient risk and devising tailored prevention and treatment strategies. For high-risk patient groups, early interventions, such as dietary improvements and increased physical activity, can effectively mitigate disease risk. Comprehensive medical measures, coupled with guidance on disease management for patients, can enhance treatment adherence, thereby optimizing therapeutic outcomes.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/Supplementary Material.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

Y-zH: Writing – original draft, Writing – review & editing. ZC: Writing – original draft, Writing – review & editing. M-hZ: Writing – review & editing. Z-yZ: Writing – review & editing. X-yW: Writing – review & editing. JH: Writing – review & editing. X-tL: Writing – review & editing. J-nZ: Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The author (s) declare financial support was received for the research, authorship, and/orpublication of this article. This study was supported by Hunan University of Chinese Medicine (No.2022XYLH031) and the National Natural Science Foundation of China (No. 82274515).

Acknowledgments

We gratefully acknowledge all the studies and databases that made GWAS summary data available. We would like to thank all the individuals who participated in this study and the staff of the relevant departments for their valuable contributions. We are also grateful to Y-zH and Z-yZ for their equal contributions as co-first authors.

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fimmu.2024.1375915/full#supplementary-material

References

1. Ingelfinger JR, Jarcho JA. Increase in the incidence of diabetes and its implications. *New Engl J Med*. (2017) 376:1473–4. doi: 10.1056/nejme1616575

2. Sell H, Eckel J. Chemotactic cytokines, obesity and type 2 diabetes:*in vivo*and*in vitro*evidence for a possible causal correlation? *Proc Nutr Soc*. (2009) 68:378–84. doi: 10.1017/s0029665109990218

3. Svenson KL, Pollare T, Lithell H, Hällgren R. Impaired glucose handling in active rheumatoid arthritis: Relationship to peripheral insulin resistance. *Metabolism-clinical Exp*. (1988) 37:125–30. doi: 10.1016/s0026-0495(98)90005-1

4. Karstoft K, Pedersen BK. Exercise and type 2 diabetes: focus on metabolism and inflammation. *Immunol Cell Biol*. (2015) 94:146–50. doi: 10.1038/icb.2015.101

5. Johansen MY, MacDonald CS, Hansen KB, Karstoft K, Christensen R, Pedersen M, et al. Effect of an intensive lifestyle intervention on glycemic control in patients with type 2 diabetes. *JAMA*. (2017) 318:637–7. doi: 10.1001/jama.2017.10169

6. Yan PG, Li JN. The standard diagnosis and treatment of ulcerative colitis. *Chin J Internal Med*. (2021) 60:567–70. doi: 10.3760/cma.j.cn112138-20210316-00216

7. Du L, Ha C. Epidemiology and pathogenesis of ulcerative colitis. *Gastroenterol Clinics North America*. (2020) 49:643–54. doi: 10.1016/j.gtc.2020.07.005

8. Berre CL, Honap S, Peyrin-Biroulet L. Ulcerative colitis. *Lancet*. (2023) 402:571–84. doi: 10.1016/s0140-6736(23)00966-2

9. Cleynen I, Boucher G, Jostins L, Schumm L, Zeißig S, Ahmad T, et al. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet*. (2016) 387:156–67. doi: 10.1016/s0140-6736(15)00465-1

10. Preda C-M, Istratescu D. *Etiology of Ulcerative Colitis*. London, UK: IntechOpen (2022). doi: 10.5772/intechopen.106842

11. Shah A, Patel V, Jain M, Parmar G. Network pharmacology and systems biology in drug discovery. *null*. (2023), 231–52. doi: 10.1007/978-981-99-1316-9_10

12. Meyer A, Dong C, Casagrande C, Chan S, Huybrechts I, Nicolas G, et al. Food processing and risk of crohn's disease and ulcerative colitis: A european prospective cohort study. *Clin Gastroenterol Hepatol*. (2023) 21:1607–1616.e6. doi: 10.1016/j.cgh.2022.09.031

13. Sun H, Saeedi P, Karuranga S, Pinkepank M, Ogurtsova K, Duncan BB, et al. IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Res Clin Pract*. (2022) 183:109119–9. doi: 10.1016/j.diabres.2021.109119

14. Thorsteinsdottir S, Gudjonsson T, Nielsen OH, VainerB, SeidelinJB. Pathogenesis and biomarkers of carcinogenesis in ulcerative colitis. *Nat Rev Gastroenterol Hepatol*. (2011) 8:395–404. doi: 10.1038/nrgastro.2011.96

15. Barabási ALászló, Oltvai ZoltánN. Network biology: understanding the cell's functional organization. *Nat Rev Genet*. (2004) 5:101–13. doi: 10.1038/nrg1272

16. Paramsothy S, Kamm MA, Kaakoush NO, Walsh A, van den Bogaerde J, Samuel DB, et al. Multidonor intensive faecal microbiota transplantation for active ulcerative colitis: a randomised placebo-controlled trial. *Lancet*. (2017) 389:1218–28. doi: 10.1016/s0140-6736(17)30182-4

17. Yao W, Gong J, Zhu W, Tian H, Ding C, Gu L, et al. Pectin enhances the effect of fecal microbiota transplantation in ulcerative colitis by delaying the loss of diversity of gut flora. *BMC Microbiol*. (2016) 16:0–0. doi: 10.1186/s12866-016-0869-2

18. Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM, et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat Rev Drug Discovery*. (2015) 14:475–86. doi: 10.1038/nrd4609

19. Paolini GV, Shapland RHB, van Hoorn WP, Mason JS, Hopkins AL. Global mapping of pharmacological space. *Nat Biotechnol*. (2006) 24:805–15. doi: 10.1038/nbt1228

20. Mencher SK, Wang LG. Promiscuous drugs compared to selective drugs (promiscuity can be a virtue). *BMC Clin Pharmacol*. (2005) 5:0–0. doi: 10.1186/1472-6904-5-3

21. Hopkins AL, Mason JS, Overington JP. Can we rationally design promiscuous drugs? *Curr Opin Struct Biol*. (2006) 16:127–36. doi: 10.1016/j.sbi.2006.01.013

22. Flordellis C, Manolis AS, Hervé P, Karabinis A. Rethinking target discovery in polygenic diseases. *Curr Topics Medicinal Chem*. (2006) 6:1791–8. doi: 10.2174/156802606778194226

23. Dessalew N, Mikre W. On the paradigm shift towards multitarget selective drug design. *Curr Comput - Aided Drug Design*. (2008) 4:76–90. doi: 10.2174/157340908784533229

24. Bulik-Sullivan B, Finucane H, Anttila V, GusevA, Day FR, Loh P-R, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet*. (2015) 47:1236–41. doi: 10.1038/ng.3406

25. Werme J, Sluis Svd, Posthuma D, de Leeuw C. An integrated framework for local genetic correlation analysis. *Nat Genet*. (2022) 54:274–82. doi: 10.1038/s41588-022-01017-y

26. Shah A, Ghasemzadeh N, Zaragoza-Macias E, Patel R, Eapen DJ, Neeland IJ, et al. Sex and age differences in the association of depression with obstructive coronary artery disease and adverse cardiovascular events. *J Am Heart Assoc*. (2014) 3. doi: 10.1161/jaha.113.000741

27. Harshfield EL, Pennells L, Schwartz J, Willeit P, KaptogeSK, Bell S, et al. Association between depressive symptoms and incident cardiovascular diseases. *JAMA*. (2020) 324:2396–6. doi: 10.1001/jama.2020.23068

28. Lespérance François, Frasure-Smith N, Talajic M. Major depression before and after myocardial infarction. *Psychosomatic Med*. (1996) 58:99–110. doi: 10.1097/00006842-199603000-00001

29. Frasure-Smith N, Lespérance François, Talajic M. The impact of negative emotions on prognosis following myocardial infarction: Is it more than depression? *Health Psychol*. (1995) 14:388–98. doi: 10.1037/0278-6133.14.5.388

30. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi AT, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat Genet*. (2015) 47:979–86. doi: 10.1038/ng.3359

31. Virtamo J, Valkeila E, Alfthan G, Punsar S, Huttunen JK, Karvonen MJ, et al. Serum selenium and the risk of coronary heart disease and stroke. *Am J Epidemiol*. (1985) 122:276–82. doi: 10.1093/oxfordjournals.aje.a114099

32. Suadicani P, Hein HO, Gyntelberg F. Serum selenium concentration and risk of ischaemic heart disease in a prospective cohort study of 3000 males. *Atherosclerosis*. (1992) 96:33–42. doi: 10.1016/0021-9150(92)90035-f

33. Burgess S, Scott RA, Timpson NJ, Davey Smith G, Thompson SG. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur J Epidemiol*. (2015) 30:543–52. doi: 10.1007/s10654-015-0011-z

34. Bowden J, Smith GD, Burgess S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int J Epidemiol*. (2015) 44:512–25. doi: 10.1093/ije/dyv080

35. Bowden J, Smith GD, Haycock P, Burgess S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. *Genet Epidemiol*. (2016) 40:304–14. doi: 10.1002/gepi.21965

36. Hartwig FP, Smith GD, Bowden J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int J Epidemiol*. (2017) 46:1985–98. doi: 10.1093/ije/dyx102

37. Verbanck M, Chen CY, Neale BM, Do R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat Genet*. (2018) 50:693–8. doi: 10.1038/s41588-018-0099-7

38. Jess T, Jensen BW, Andersson M, Villumsen M, Allin KH. Inflammatory bowel diseases increase risk of type 2 diabetes in a nationwide cohort study. *Clin Gastroenterol Hepatol*. (2020) 18:881–888.e1. doi: 10.1016/j.cgh.2019.07.052

39. Wu X, Zhang Q, Meng Z, Gan H. Systematic review and meta-analysis of the association between inflammatory bowel disease and diabetes mellitus. *Chin J Postgraduate Med*. (2021) 44:748–54.

40. Kang ES, Han K, Chun J, Soh H, Park S, Im J, et al. Increased risk of diabetes in inflammatory bowel disease patients: A nationwide population-based study in korea. *J Clin Med*. (2019) 8:343–3. doi: 10.3390/jcm8030343

41. Lai S-W, Kuo Y-H, Liao K-F. Association between inflammatory bowel disease and diabetes mellitus. *Clin Gastroenterol Hepatol*. (2020) 18:1002–3. doi: 10.1016/j.cgh.2019.09.016

42. Allin KH, Agrawal M, Iversen A, Antonsen J, Villumsen M, Jess T. The risk of type 2 diabetes in patients with inflammatory bowel disease after bowel resections: A nationwide cohort study. *Gastro Hep Adv*. (2022) 1:777–84. doi: 10.1016/j.gastha.2022.06.007

43. Jurjus A, Eid A, Kattar SAl, Zeenny MN, Gerges-Geagea AG, Haydar H, et al. Inflammatory bowel disease, colorectal cancer and type 2 diabetes mellitus: The links. *BBA Clin*. (2016) 5:16–24. doi: 10.1016/j.bbacli.2015.11.002

44. Verdugo-Meza A, Ye J, Dadlani H, Ghosh S, Gibson DL. Connecting the dots between inflammatory bowel disease and metabolic syndrome: A focus on gut-derived metabolites. *Nutrients*. (2020) 12:1434–4. doi: 10.3390/nu12051434

45. Tseng CH. Metformin use is associated with a lower risk of inflammatory bowel disease in patients with type 2 diabetes mellitus. *J Crohns Colitis*. (2021) 15:64–73. doi: 10.1093/ecco-jcc/jjaa136

46. Viollet B, Guigas B, Garcia NS, LeclercJ, Foretz M, Andreelli F. Cellular and molecular mechanisms of metformin: an overview. *Clin Sci*. (2011) 122:253–70. doi: 10.1042/cs20110386

47. Al-Dwairi A, Alqudah M, Al-Shboul O, Alfaqih MA, Alomari D. Metformin exerts anti–inflammatory effects on mouse colon smooth muscle cells in vitro. *Exp Ther Med*. (2018). doi: 10.3892/etm.2018.6222

48. Xue Y, Zhang H, Sun X, Zhu M-J. Metformin improves ileal epithelial barrier function in interleukin-10 deficient mice. *PloS One*. (2016) 11:e0168670. doi: 10.1371/journal.pone.0168670

49. Deng J, Zeng L, Lai X, Li J, Liu L, Lin Q, et al. Metformin protects against intestinal barrier dysfunction via AMPKα1-dependent inhibition of JNK signalling activation. *J Cell Mol Med*. (2017) 22:546–57. doi: 10.1111/jcmm.13342

50. Russo E, Giudici F, Fiorindi C, Ficari F, Scaringi S, Amedei AA, et al. Immunomodulating activity and therapeutic effects of short chain fatty acids and

tryptophan post-biotics in inflammatory bowel disease. *Front Immunol*. (2019) 10:2754. doi: 10.3389/fimmu.2019.02754

51. Bian X, Wu W, Yang L, Lv LX, Wang Q, LiY, et al. Administration of akkermansia muciniphila ameliorates dextran sulfate sodium-induced ulcerative colitis in mice. *Front Microbiol*. (2019) 10:2259. doi: 10.3389/fmicb.2019.02259

52. Ryan P, Patterson E, Carafa I, Mandal R, Wishart DS, Dinan TG, et al. Metformin and dipeptidyl peptidase-4 inhibitor differentially modulate the intestinal microbiota and plasma metabolome of metabolically dysfunctional mice. *Can J Diabetes*. (2020) 44:146–155.e2. doi: 10.1016/j.jcjd.2019.05.008

53. Mccreight LJ, Bailey CJ, Pearson ER. Metformin and the gastrointestinal tract. *Diabetologia*. (2016) 59:426–35. doi: 10.1007/s00125-015-3844-9

54. Elahi D, Egan JM, Shannon RP, Meneilly GS, Khatri A, Habener JF, et al. GLP-1 (9-36) amide, cleavage product of GLP-1 (7-36) amide, is a glucoregulatory peptide. *Obesity*. (2008) 16:1501–9. doi: 10.1038/oby.2008.229

55. Müller T, Finan B, Bloom SR, D'Alessio DA, Drucker DJ, Flatt P, et al. Glucagon-like peptide 1 (GLP-1). *Mol Metab*. (2019) 30:72–130. doi: 10.1016/j.molmet.2019.09.010

56. Wang W. *Protective Effect and Mechanism of Glp-1 on Diabetes Mellitus Complicated With Chronic Colitis [D]*. Changchun, Jilin Province, China: Jilin University (2022).

57. Farghaly O. EXPERIMENTAL ULCERATIVE COLITIS: TGF-β AS A DIAGNOSTIC MARKER. *Al-Azhar J Pharm Sci*. (2020) 61:46–60. doi: 10.21608/ajps.2020.86014

58. Huang L. *The Role and Potential Therapeutic Mechanisms of Irisin in UC Mice*. Kaifeng, Henan Province, China: Henan University (2019).

59. Novelle MG, Contreras C, Romero-Picó A, López M, Diéguez C. Irisin, two years later. *Int J Endocrinol*. (2013) 2013:1–8. doi: 10.1155/2013/746281

60. Polyzos SA, Anastasilakis AD, Geladari E, Mantzoros CS. Irisin in patients with nonalcoholic fatty liver disease. *Metabolism-clinical Exp*. (2014) 63:207–17. doi: 10.1016/j.metabol.2013.09.013

61. Dulian K, Laskowski Radosław, Grzywacz T, Kujach S, Flis DJ, Smaruj M, et al. The whole body cryostimulation modifies irisin concentration and reduces inflammation in middle aged, obese men. *Cryobiology*. (2015) 71:398–404. doi: 10.1016/j.cryobiol.2015.10.143

62. Lima Mágero FR, Vasconcellos LF, Frazão L, Bandeira F. Irisin precursor FNDC5 and glycemic control in patients with type 2 diabetes. *J Endocrine Soc*. (2021). doi: 10.1210/jendso/bvab048.963

63. Li Q. Relationship between transforming growth factor beta 1 and peripheral neuropathy in type 2 diabetic. *Beijing Med J*. (2013).

64. John P, Yadla M. Noninvasive method of differentiating diabetic nephropathy and nondiabetic renal disease using serum bone morphogenetic protein-7 and transforming growth factor-beta 1 levels in patients with type-2 diabetes mellitus. *Saudi J Kidney Dis Transplant*. (2019) 30:1300–0. doi: 10.4103/1319-2442.275474

Frontiers in Immunology

Explores novel approaches and diagnoses to treat immune disorders.

The official journal of the International Union of Immunological Societies (IUIS) and the most cited in its field, leading the way for research across basic, translational and clinical immunology.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact