frontiers | Frontiers in Immunology

TYPE Original Research
PUBLISHED 10 April 2025
DOI 10.3389/fimmu.2025.1541203

OPEN ACCESS

EDITED BY
Anthony C. Y. Yau,
AliveDx, United Kingdom

REVIEWED BY
Walid Shalata,
Soroka Medical Center, Israel
Ved Vrat Verma,
National Institute of Cancer Prevention and Research (ICMR), India

*CORRESPONDENCE
Zhike Liu
liuzhike07@163.com
Siyan Zhan
siyan-zhan@bjmu.edu.cn

RECEIVED 07 December 2024
ACCEPTED 25 March 2025
PUBLISHED 10 April 2025

CITATION
Yang J, Wu Y, Guo J, Wang X, Gao X, Chen X, Zhang M, Yang J, Liu Z, Liu Y, Liu Z and Zhan S (2025) Development and validation of identification algorithms for five autoimmune diseases using electronic health records: a retrospective cohort study in China. *Front. Immunol.* 16:1541203. doi: 10.3389/fimmu.2025.1541203

Development and validation of identification algorithms for five autoimmune diseases using electronic health records: a retrospective cohort study in China

Junting Yang[1,2], Yunxiao Wu[1,2], Jinxin Guo[1,2], Xiaoxuan Wang[1,2], Xin Gao[1,2], Xin Chen[1,2], Mengdi Zhang[1,2], Jin Yang[3], Zuojing Liu[4], Yan Liu[5], Zhike Liu[1,2]* and Siyan Zhan[1,2,6,7]*

[1]Department of Epidemiology and Biostatistics, School of Public Health, Peking University, Beijing, China, [2]Key Laboratory of Epidemiology of Major Diseases (Peking University), Ministry of Education, Beijing, China, [3]Department of Endocrinology and Metabolism, Peking University Third Hospital, Beijing, China, [4]Department of Gastroenterology, Peking University Third Hospital, Beijing, China, [5]Department of Hematology, Peking University Third Hospital, Beijing, China, [6]Research Center of Clinical Epidemiology, Peking University Third Hospital, Beijing, China, [7]Center for Intelligent Public Health, Institute for Artificial Intelligence, Peking University, Beijing, China

Objective: This study aims to assess the identification algorithms for five autoimmune diseases—Hashimoto's thyroiditis, inflammatory bowel disease (IBD), primary immune thrombocytopenia (ITP), rheumatoid arthritis (RA), and type 1 diabetes (T1D)—using the Yinzhou Regional Health Information Platform (YRHIP) in China.

Methods: Diagnostic data was extracted from YRHIP's population registry (2010-2021), combining ICD-10 codes and Chinese medical terminology from outpatient, inpatient, and discharge records. Algorithms were validated through chart reviews, adhering to global clinical guidelines. Cases were adjudicated using electronic case report forms. We evaluated algorithm performance based on sensitivity and positive predictive value (PPV), with a 70% PPV threshold for optimization.

Results: Among all reviewed cases, we identified 136 cases for Hashimoto's thyroiditis, 65 for IBD, 76 for ITP, 130 for RA, and 43 for T1D. Algorithm performance varied across diseases: the final algorithm for Hashimoto's thyroiditis achieved optimal accuracy (sensitivity 97.44%, PPV 98.28%), followed by RA (sensitivity 100.00%, PPV 76.92%). Algorithms for IBD and ITP required synthesis of multiple data sources to achieve acceptable performance (IBD: sensitivity 79.66%, PPV 70.15%; ITP: sensitivity 62.50%, PPV 70.00%). For T1D, the final algorithm utilizing both admission and outpatient records yielded satisfactory results (sensitivity 84.09%, PPV 74.00%).

Conclusions: This study presents the first validated algorithms for identifying autoimmune diseases using EHR data in China, demonstrating satisfactory

performance (PPV >70%) across all diseases. Our findings demonstrate that a combination of data sources is crucial for accurate case identification in complex autoimmune conditions, providing an important methodological foundation for future real-world studies in Chinese populations.

KEYWORDS

Hashimoto's thyroiditis, inflammatory bowel disease (IBD), primary immune thrombocytopenia, rheumatoid arthritis (RA), type 1 diabetes (T1D), computable phenotype, algorithms, electronic health records (EHR)

1 Introduction

Autoimmune diseases (AD) constitute a group of conditions where the immune system mistakenly attacks the body's own tissues, leading to severe tissue/organ destruction or even fatal outcomes. Factors such as genetics, environmental changes, and vaccinations have been implicated in their etiology (1). Through rigorous epidemiological studies, it has now been shown that autoimmune diseases affect 3-10% of the global population (2, 3), suggesting an urgent demand for the control of these diseases. A study utilizing the Clinical Practice Research Datalink (CPRD) database explored the prevalence of 19 common autoimmune diseases, noting an increase from 7.7% in 2000 to 11.0% in 2019 (4). Further monitoring of the burden of these diseases in different regions and countries using high-quality real-world databases is essential.

As the demand for data-driven healthcare continues to rise, real-world studies (RWS) have become increasingly important in the context of regulatory decision-making. Nevertheless, a major challenge faced by RWS is the issue of data quality (5). The U.S. Food and Drug Administration's (FDA) Framework for Real-World Evidence Program underscores the necessity for reliable and relevant real-world data (RWD), with the accuracy of key variables being pivotal to data reliability (6). In the contexts of disease burden estimation, associative analyses, and vaccine safety surveillance, it is imperative to reduce biases stemming from misclassification. To maximize the efficacy of these studies, it is necessary to ensure the validity of outcome variables and possess an in-depth understanding of the definition algorithms (7, 8).

To address these challenges in disease identification and classification, computable phenotypes have emerged as a promising solution. These are meaningful health or disease characteristics extracted from raw data through data analysis and computational models (9). These algorithms include one or more structured data elements, such as ICD codes, laboratory test results, or medication prescriptions, which can identify specific disease risk factors or subtypes from electronic health records (EHR). This approach provides high-fidelity, personalized, and interpretable phenotype estimations, enabling researchers to extract valuable information from vast medical datasets, thereby enhancing diagnostic and therapeutic efficiency (10). Several studies have been conducted to develop and validate computable phenotypes for diseases such as hypertension, pulmonary hypertension, and adverse drug reactions in real-world databases (11, 12). However, there are few similar studies in China (13), and none for autoimmune diseases.

The Yinzhou Regional Health Information Platform (YRHIP) in Ningbo, China, has been validated as a reliable and robust real-world database (14, 15). It encompasses various types of RWD, such as EHR, registries, and electronic medical records (EMR). These data are converted into a structured format and linked to a unique national identifier or healthcare identifier. YRHIP has been extensively utilized in fields such as infectious disease surveillance, drug safety evaluation, chronic disease management, generating numerous high-quality studies and yielding significant societal benefits. Furthermore, we emphasize that assessing the accuracy of AD diagnostic records in the YRHIP is crucial for the active post-market surveillance of the HPV vaccine safety (16). In summary, this study aims to validate the performance of algorithms used to identify five specific ADs utilizing the YRHIP. The findings are intended to provide valuable insights for related real-world studies.

2 Methods

2.1 Data sources and study population

The YRHIP is a comprehensive electronic health information system established in 2005 by the Yinzhou District Center for Disease Control and Prevention. It is designed to enhance routine primary care services offered by local general practitioners. YRHIP retains the EHRs of the population in Yinzhou District, Ningbo, Zhejiang Province, which encompasses 976,409 permanent residents with valid healthcare coverage as identified by the end of 2019 (17). This database platform has progressively incorporated information from public health surveillance, hospital health information systems, maternal and child healthcare, immunizations, population screening, disease management, and other healthcare services. Since 2009, YRHIP has covered nearly all healthcare-related activities for residents from birth to death (16).

This study utilized a retrospective cohort design, utilizing the YRHIP database to compile registration information for all permanent residents documented in the EHRs maintained within the YRHIP. Data collection spanned from January 2010 to June 2021. The study specifically focused on healthcare visitations related to five ADs of interests: Hashimoto's thyroiditis, inflammatory bowel disease (IBD), primary immune thrombocytopenia (ITP), rheumatoid arthritis (RA), and type I diabetes (T1D). Mortality data were included to appropriately address censoring.

Inclusion criteria for the cohort were: a. All permanent residents with records in YRHIP from January 2010 to June 2021. Permanent residents are defined as those who have resided in the area for more than six months and have been registered in the EHR of YRHIP; b. Diagnosis during the observation period with one of the five ADs of interest. Exclusion criteria for the cohort were: a. Missing health record registration date in YRHIP; b. Missing unique personal identification code.

2.2 Identification of autoimmune diseases and incident cases

Within the cohort, cases were identified using disease diagnostic terms and/or International Classification of Diseases (ICD-10) codes (Table 1). We identified potential patients by assessing any disease diagnoses in outpatient records, emergency records, admission records, discharge records, and inpatient records. To ensure the integrity of clinical records used for chart review, this study focused exclusively on "incident cases" of admission for validation purposes. Incident cases included those patients who were admitted upon their initial diagnosis and patients admitted within one month following an outpatient diagnosis. A one-year washout period was applied, consistent with previous studies. Non-hospitalized patients were excluded due to insufficient data for validation. Notably, T1D was identified using terms such as "type I diabetes," "insulin-dependent diabetes," and ICD-10 code "E10".

2.3 Adjudication of autoimmune diseases and algorithms

In this study, we employed a chart review method for case verification, engaging two clinical physicians (at or above the department head level) to independently assess the cases based on clinical guidelines and standards. They provided a conclusion of "yes," "no," or "undeterminable," and also indicated whether the case was a previous one. Situations where insufficient data prevented a definitive conclusion were categorized as "undeterminable." If there was a disagreement between the two physicians, a third clinical physician participated in a discussion to reach a consensus.

Data collection was conducted using a standardized electronic case report form (eCRF), developed with reference to international clinical guidelines (Table 1), which were reviewed and confirmed by clinical experts at the department head level. The eCRF varied slightly between different diseases and primarily included definitive grading of diseases and suspected case information forms. The definitive grading incorporated criteria such as symptom manifestation, auxiliary examination results, and surgical outcomes. The case information form gathered essential data for determining definitive grading, including visiting information, diagnostic information and clinical symptoms, medical examination and laboratory testing, and the reviewed grading for suspected cases. These eCRFs were completed by trained clinical physicians or personnel with relevant medical expertise. (Appendix 1)

The identification algorithm was constructed using diagnostic terms and ICD-10 codes, with a screening condition of positive predictive value (PPV)≥70% to finalize the algorithm with the optimal sensitivity (13). The algorithm structure was broadly as follows (specific algorithms for each autoimmune disease are detailed in Table 2): a. The ICD-10 code related to the disease of interest; b. The Chinese terminology related to the disease of interest; c. Either the ICD-10 code or the Chinese terminology related to the disease of interest.

2.4 Statistical analysis

The performance of each identification algorithm was assessed by determining its sensitivity, and PPV, with 95% confidence intervals calculated using a binomial distribution. Categorical variables were expressed as frequencies (percentages), continuously normally distributed variables as mean ± standard deviation, and non-normally distributed continuous variables as medians and interquartile ranges. Statistical analyses were conducted using SAS® software (version 9.4; SAS Institute Inc., Cary, NC, USA).

3 Results

3.1 Case extraction and adjudication

The comprehensive extraction from the YRHIP yielded a total of 136 cases for Hashimoto's thyroiditis, 65 for IBD, 76 for ITP, 130 for RA, and 43 for T1D, with all cases meeting the criteria for further analysis without any being classified as indeterminate. Among these, the distribution of previous cases, eligible cases, and true cases varied by disease, reflecting the complexities in case identification within our retrospective cohort study (Table 2).

3.2 Performance of algorithms to identify autoimmune diseases

The performance assessment results of the identification algorithms were presented in Table 2. Overall, the table illustrated the varying sensitivity and PPV of the algorithms for different autoimmune diseases, with a notable emphasis on the importance of combining ICD-10 codes with diagnostic terms to achieve higher accuracy.

TABLE 1 Clinical guidelines and diagnostic criteria of autoimmune diseases.

Autoimmune diseases (English and Chinese)	ICD10	Reference guidelines	Diagnostic criteria
Hashimoto's thyroiditis	E06.3	the Chinese guidelines for the diagnosis and treatment of thyroiditis (18)	"2 and 5 and 6" and (or) "1 and 3 and 4" of Level 1 of certainty [a]
Inflammatory bowel disease	K51, K50	1. the Consensus on the Diagnosis and Treatment of Inflammatory Bowel Disease (Beijing, 2018) (19) 2. the World Gastroenterology Organization Practice Guidelines for the diagnosis and management of IBD in 2010 (20)	Level 1 of certainty
Primary immune thrombocytopenia	D69.3, D69.4	1. the Brighton Collaboration's guidelines for Thrombocytopenic Purpura (21) 2. the Pediatric Primary Immune Thrombocytopenia clinical Guidelines (2019 Edition) (22)	Level 1 of certainty
Rheumatoid arthritis	M0[56]\.[389]	1. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis (23) 2. 2010 ACR/EULAR rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative (24)	Meet one of the two criteria from reference guidelines [b]
Type I diabetes	E10	the Chinese Guidelines for the Diagnosis and Treatment of Type 1 Diabetes Mellitus (2013 edition) (25)	Level 1 of certainty

Specific details on clinical examinations and symptoms for the diagnostic criteria of each disease are provided in the electronic case report forms for each disease, available in Appendix 1. a. Considering that the definitive criteria for Hashimoto's thyroiditis involve pathological examinations such as biopsy, which are difficult to obtain, meeting certain essential clinical criteria is sufficient for diagnosis. b. The standards of the two clinical guidelines are not entirely identical, but their diagnostic capabilities are equivalent; therefore, meeting the criteria of either one is acceptable.

For Hashimoto's thyroiditis, the algorithm utilizing both ICD-10 codes and diagnostic terms demonstrated the highest accuracy, with a sensitivity of 97.44% (95%CI: 94.57%, 100.00%) and a PPV of 98.28% (95%CI: 95.91%, 100.00%). This indicated that the combined approach was highly effective in identifying true cases of the disease without a significant loss in PPV. Notably, the use of ICD-10 "E06.3" alone resulted in a sensitivity of 11.11% and a PPV of 100.00%, highlighting the limitations of using a single identifier. The final algorithm, which included multiple data sources, significantly improved sensitivity while maintaining a high PPV.

For IBD, the algorithm using outpatient diagnostic terms alone showed a high PPV of 100.00% but a low sensitivity of 18.64%. Conversely, the discharge diagnosis algorithm had a sensitivity of 59.32% and a PPV of 52.24%. The final algorithm, which combined both outpatient and inpatient data, achieved a sensitivity of 79.66% (95% CI: 69.39%, 89.93%) and a PPV of 70.15% (95% CI: 59.19%, 81.11%), reflecting the complexity of diagnosing IBD and the need for multidimensional evidence.

The ITP algorithm using only outpatient diagnostic terms showed a PPV of 100% but a sensitivity below 20%. The final algorithm, which synthesized multiple data sources, achieved a sensitivity of 62.50% (95% CI: 49.82%, 75.18%) and a PPV of 70.00% (95% CI: 57.30%, 82.70%), demonstrating that combining various data sources was necessary to improve the identification of ITP cases.

In the case of RA, the discharge diagnosis ICD-10 code algorithm showed a sensitivity of 100.00% and a PPV of 76.92% (95% CI: 63.70%, 90.15%), which was identical to the final algorithm's performance. This suggested that the inclusion of discharge diagnosis ICD-10 codes was sufficient for accurate identification of RA cases.

For T1D, the discharge diagnosis "E10 & diabetic ketoacidosis" algorithm had the highest sensitivity at 86.36% (95% CI: 76.22%, 96.50%) but a very low PPV, leading to its exclusion from the final algorithm. The final algorithm, requiring at least two outpatient records, resulted in a PPV of 74.00% (95% CI: 61.84%, 86.16%), demonstrating the need for multiple diagnostic records to ensure accurate identification of T1D cases.

4 Discussion

This is the first validation study and evaluation of identification algorithms for autoimmune diseases using EHR in China. Using the 10-year cohort data from YRHIP, we evaluated the performance of diagnostic terminology, ICD-10 codes, and related algorithms for five autoimmune diseases, with a PPV threshold of 70% to optimize sensitivity. The final algorithms demonstrated varying performance, with Hashimoto's thyroiditis and rheumatoid arthritis achieving excellent accuracy through single diagnostic terms or ICD-10 codes (PPV>70%), while diseases like ITP required more complex combinations yet still showed limited sensitivity.

4.1 Algorithm performance and optimization

Our evaluation revealed varying algorithm performance across different autoimmune diseases. While HT achieved the highest accuracy (sensitivity 97.44%, PPV 98.28%), followed by RA (sensitivity 100%, PPV 76.92%), IBD (sensitivity 79.66%, PPV

10.3389/fimmu.2025.1541203

TABLE 2 Performance of algorithms/computable phenotype for identification of autoimmune diseases.

Autoimmune diseases	Algorithms	Valid Cases	Confirmed Cases	SEN (95%CI)	PPV (95%CI)
Hashimoto's thyroiditis					
	1: Diagnosis of admission records	13	13	11.11 (5.42, 16.81)	100.00 (100.00,100.00)
	2: ICD10 of admission records	3	3	2.56 (0.00, 5.43)	100.00 (100.00,100.00)
	3: Diagnosis of discharge records	116	114	97.44 (94.57,100.00)	98.28 (95.91,100.00)
	4: ICD10 of discharge records	98	97	82.91 (76.08,89.73)	98.98 (96.99,100.00)
	5: Diagnosis of outpatient records	17	17	14.53 (8.14, 20.92)	100.00(100.00,100.00)
	6: ICD10 of outpatient records	17	17	14.53 (8.14, 20.92)	100.00(100.00,100.00)
	Final algorithms: 1-6	116	114	97.44(94.57,100.00)	98.28(95.91,100.00)
Inflammatory bowel disease					
	1: Diagnosis of admission records: "Crohn's disease"	6	6	10.17 (2.46,17.88)	100.00 (100.00,100.00)
	2: ICD10 of admission records: K50	0	0	NA	NA
	3: Diagnosis of admission records: "ulcerative colitis"	30	25	42.37 (29.76,54.98)	83.33 (70.00,96.67)
	4: ICD10 of admission records: K51	4	3	5.08 (0.00,10.69)	75.00 (32.57,100.00)
	5: Diagnosis of discharge records: "Crohn's disease"	18	13	22.03 (11.46,32.61)	72.22 (51.53,92.91)
	6: ICD10 of discharge records: K50	17	12	20.34 (10.07,30.61)	70.59 (48.93,92.25)
	7: Diagnosis of discharge records: "ulcerative colitis"	67	35	59.32 (46.79,71.86)	52.24 (40.28,64.20)
	8: ICD10 of discharge records: K51	69	35	59.32 (46.79,71.86)	50.72 (38.93,62.52)
	9: Diagnosis of outpatient records: "Crohn's disease"	19	16	27.12 (15.77,38.46)	84.21 (67.81,100.00)
	10: ICD10 of outpatient records: K50	19	16	27.12 (15.77,38.46)	84.21 (67.81,100.00)
	***11: Diagnosis of outpatient records: "ulcerative colitis	11	11	18.64 (8.71,28.58)	100.00 (100.00,100.00)
	***12: ICD10 of outpatient records: K51	43	32	18.64 (8.71,28.58)	74.42 (61.38,87.46)
	Final algorithms: 1, 3, 5, 6, 9-12	67	47	79.66 (69.39,89.93)	70.15 (59.19,81.11)
Primary immune thrombocytopenia					
	1: Diagnosis of admission records: "immune thrombocytopenia"	4	4	7.14 (0.40,13.89)	100.00 (100.00,100.00)
	2: ICD10 of admission records: D69.3	2	2	3.57 (0.00,8.43)	100.00 (100.00,100.00)
	3: Diagnosis of admission records: "idiopathic thrombocytopenia"	4	3	5.36 (0.00,11.25)	75.00 (32.57,100.00)
	4: ICD10 of admission records: D69.4	0	0	NA	NA
	5: Diagnosis of discharge records: "immune thrombocytopenia"	29	21	37.50 (24.82,50.18)	72.41 (56.15,88.68)
	6: ICD10 of discharge records: D69.3	33	23	41.07 (28.19,53.96)	69.70 (54.02,85.38)
	7: Diagnosis of discharge records: "idiopathic thrombocytopenia"	8	5	8.93 (1.46,16.4)	62.50 (28.95,96.05)
	8: ICD10 of discharge records: D69.4	12	5	8.93 (1.46,16.4)	41.67 (13.77,69.56)

(Continued)

10.3389/fimmu.2025.1541203

TABLE 2 Continued

Autoimmune diseases	Algorithms	Valid Cases	Confirmed Cases	SEN (95%CI)	PPV (95%CI)
Primary immune thrombocytopenia					
	9: Diagnosis of outpatient records: "immune thrombocytopenia"	31	22	39.29 (26.49,52.08)	70.97 (54.99,86.95)
	**10: ICD10 of outpatient records: D69.3	23	19	33.93 (21.53,46.33)	82.61 (67.12,98.10)
	**11: Diagnosis of outpatient records: "idiopathic thrombocytopenia"	10	8	14.29 (5.12,23.45)	80.00 (55.21,100.00)
	12: ICD10 of outpatient records: D69.4	6	5	8.93 (1.46,16.4)	83.33 (53.51,100.00)
	Final algorithms: 1-3, 5, 9-12	50	35	62.50 (49.82,75.18)	70.00 (57.3,82.7)
Rheumatoid arthritis					
	1: Diagnosis of admission records: "rheumatoid arthritis"	17	15	50.00 (32.11,67.89)	88.24 (72.92,100.00)
	2: ICD10 of admission records: M0[56]\.[389]	2	2	6.67 (0.00,15.59)	100.00 (100.00,100.00)
	3: Diagnosis of discharge records: "rheumatoid arthritis"	39	30	100.00 (100.00,100.00)	76.92 (63.70,90.15)
	4: ICD10 of discharge records: M0[56]\.[389]	36	29	96.67 (90.24,100.00)	80.56 (67.63,93.48)
	5: Diagnosis of outpatient records: "rheumatoid arthritis"	30	25	83.33 (70,96.67)	83.33 (70,96.67)
	6: ICD10 of outpatient records: M0[56]\.[389]	29	24	80.00 (65.69,94.31)	82.76 (69.01,96.51)
	Final algorithms: 1-6	39	30	100.00 (100.00,100.00)	76.92 (63.70,90.15)
Type I diabetes					
	1: Diagnosis of admission records: "Type I diabetes"	13	11	25.00 (12.21,37.79)	84.62 (65,100.00)
	2: Diagnosis of admission records: "autoimmune diabetes"	1	1	2.27 (0.00,6.68)	100.00 (100.00,100.00)
	3: ICD10 of admission records: E10	2	1	2.27 (0.00,6.68)	50.00 (0.00,100.00)
	4: Diagnosis of admission records: E10 & "diabetic ketosis"	2	1	2.27 (0.00,6.68)	50.00 (0.00,100.00)
	5: Diagnosis of discharge records: "Type I diabetes"	44	31	70.45 (56.97,83.94)	70.45 (56.97,83.94)
	6: Diagnosis of discharge records: "autoimmune diabetes"	0	0	NA	NA
	7: ICD10 of discharge records: E10	0	0	NA	NA
	8: Diagnosis of discharge records: E10 & "diabetic ketosis"	218	38	86.36 (76.22,96.50)	17.43 (12.4,22.47)
	**9: Diagnosis of outpatient records: "Type I diabetes"	20	14	31.82 (18.06,45.58)	70.00 (49.92,90.08)
	**10: ICD10 of outpatient records: E10	14	13	29.55 (16.06,43.03)	92.86 (79.37,100.00)
	**11: Diagnosis of outpatient records: "autoimmune diabetes"	2	2	4.55 (0.00,10.7)	100.00 (100.00,100.00)
	12: Diagnosis of outpatient records: ("Type I diabetes" or E10) & "diabetic ketosis"	15	14	31.82 (18.06,45.58)	93.33 (80.71,100.00)
	Final algorithms: 1, 2, 5, 10, 11, 12	50	37	84.09 (73.28,94.90)	74.00 (61.84,86.16)

Valid cases: cases identified by the algorithm. Confirmed cases: cases identified by the algorithm and validated as confirmed through chart review. Hashimoto's diagnosis: Hashimoto/Chiomoto, lymphocytic goiter. Hashimoto ICD10: E06.3. Hashimoto's final algorithm: The above algorithm is arbitrary. **: 2 or more records. ***: 3 or more records. $SEN = \frac{\text{Confirmed cases}}{\text{True cases}} \times 100\,\%$; $PPV = \frac{\text{Confirmed cases}}{\text{Valid cases}} \times 100\,\%$; Total number: The total number of cases identified using disease diagnostic terminology or ICD10. Ture cases: The number of patients confirmed as cases through case verification (gold standard).

70.15%), ITP (sensitivity 62.50%, PPV 70.00%), and T1D (sensitivity 84.09%, PPV 74.00%), most algorithms except HT demonstrated lower performance compared to international studies. This discrepancy can be attributed to several key factors. Most notably, established studies have employed more sophisticated algorithm construction approaches. The value of synthesizing multidimensional data is well-documented. For instance, Klompas et al (26) achieved remarkable accuracy in T1D identification (sensitivity 100.00%, 95% CI 96.00%-100.00%; validation cohort: sensitivity 97.00%, PPV 88.00%) by utilizing laboratory values, including plasma C-peptide and autoantibody levels, alongside diagnostic codes and prescription data. Similarly, RA algorithms combining specialist assessments with disease-specific medications reached PPV of 93.60% (27). In IBD studies, the Korean group reported superior performance (sensitivity 94.00%, PPV 97.90%) through therapeutic data synthesis (28), while our results (sensitivity 79.66%, PPV 70.15%) more closely resembled the Ontario study that relied solely on diagnostic information (sensitivity 79.40%) (28). The value of multiple diagnostic records has also been well established. Studies have shown that requiring two or more visits within six months improved ITP identification (PPV 93.00%) (29), with comparable benefits observed in IBD (30) and RA (31) algorithms. Despite these limitations in data dimensions, we achieved modest improvements through strategic data synthesis, combining outpatient, inpatient, and discharge records. This approach notably enhanced disease identification: for IBD, while single outpatient records showed perfect PPV but limited sensitivity (18.64%), our comprehensive algorithm improved sensitivity to 79.66%. Similar patterns emerged for ITP, where initial high PPV (82.61%-100.00%) but low sensitivity (<40%) improved to 62.50% after combining multiple data sources. Population characteristics further influenced algorithm performance. Previous T1D studies focused primarily on younger cohorts (<20 years), where T1D comprises ≥85% of diabetes cases, achieving higher accuracy (PPV 85.50% (32), 89.30% (33)). Our inclusion of the full age spectrum, particularly adults over 30 where type 2 diabetes predominates, introduced additional diagnostic challenges.

Recent advances in algorithm optimization have incorporated artificial intelligence and machine learning methodologie[s] (34). Notable achievements include convolutional neural network models for RA radiographic assessment (sensitivity 95.00%, specificity 94.00%) (35), and computer-aided diagnostic systems utilizing ultrasound features for HT detection (sensitivity 82.80%, PPV 88.90%) (36). Computational phenotyping approaches continue to evolve, as demonstrated by Zhou et al.'s EHR-based machine learning system achieving 92.29% accuracy in RA identification (37). These advanced methodologies, while promising, face implementation challenges including data accessibility, interpretability limitations, and generalizability constraints. Our structured data-based approach, though yielding moderate performance metrics, offers advantages in transparency and reproducibility.

Outcome validation and algorithm refinement fundamentally aim to address misclassification bias, with varying performance requirements across applications. Pharmacovigilance and vaccine safety studies require high sensitivity for comprehensive case capture, while risk assessment research may prioritize high PPV combined with nested case-control designs. Recent developments include quantitative bias analysis (QBA) for misclassification adjustment (38), with Newcomer et al. emphasizing exposure-stratified PPV assessment (39). Our algorithm selection process balanced these considerations, prioritizing practical utility despite sensitivity limitations in some disease categories.

It is critical to acknowledge the potential impacts of false positives and false negatives. False positives may increase healthcare utilization by triggering unnecessary diagnostic evaluations or overtreatment, thereby escalating medical costs. Conversely, false negatives risk missing individuals with genuine healthcare needs, potentially leading to delayed diagnoses and adverse clinical outcomes (40). However, the cumulative effects of these errors on downstream research and clinical decisions are highly context-dependent. Their interpretation requires careful consideration of the specific application scenarios to avoid overgeneralization. For instance: In disease burden estimation, a high false-positive rate could artificially inflate prevalence estimates, distorting public health prioritization (41). For vaccine/drug safety surveillance, tolerating a controlled level of false positives might enhance sensitivity in early signal detection, albeit at the cost of specificity (42). In comparative effectiveness or risk assessment studies, adjustments using predictive values (e.g., PPV) or sensitivity analyses are essential to mitigate bias in effect estimates (40).

4.2 Methodological limitations and potential improvements

Several methodological considerations warrant discussion. Our study utilized data from a single healthcare network (YRHIP) in eastern coastal China. While this network ensured comprehensive patient capture and reliable data quality, the single-center nature may limit result generalizability to other healthcare systems, particularly in regions with uneven resource distribution and varying clinical practices. Nevertheless, this study represents the first algorithm development for autoimmune disease identification using real-world data in China, establishing important theoretical and methodological foundations in this field. Future validation through multi-center databases could enhance algorithm applicability. Our algorithms primarily relied on diagnostic terminology and ICD-10 codes, aligning with typical EHR system structures and offering operational feasibility. However, for conditions requiring laboratory or imaging confirmation, such as RA, the inability to incorporate key biomarkers (rheumatoid factor, anti-cyclic citrullinated peptide antibodies) and imaging data may have limited sensitivity. Despite these constraints, our synthesis of outpatient, inpatient, and discharge records demonstrated meaningful improvements in algorithm performance, validating the importance of comprehensive data analysis approaches. Future studies could incorporate richer clinical data and employ advanced modeling methods to enhance diagnostic accuracy. The 2017 implementation of the real-name medical system introduced challenges in historical data linkage, potentially affecting both case

identification accuracy. While we employed washout periods to mitigate these effects, future improvements in data linkage technologies and governance are needed to enhance analysis quality and reduce potential biases. Based on these considerations, future research directions should prioritize: establishing multi-center research networks to validate algorithms across diverse healthcare settings; exploring advanced methodologies for combining multi-source clinical data. These efforts will contribute to improving early diagnosis and ultimately enhance patient outcomes.

Our study has important implications for global public health. The validated algorithms provide valuable tools for conducting real-world studies in autoimmune diseases, particularly in resource-limited settings where comprehensive clinical assessments may be challenging to implement. By improving case identification through routinely collected electronic health data, these algorithms can support large-scale epidemiological surveillance across diverse populations. Furthermore, our findings may contribute to enhancing clinical practice by facilitating earlier identification of autoimmune conditions, potentially addressing diagnostic delays that are commonly observed among patients worldwide. Research indicates that patients with autoimmune diseases often experience diagnostic journeys averaging over 4.5 years, with multiple physician consultations before receiving a definitive diagnosis (43). By refining case identification algorithms, our work could aid clinical decision support systems in identifying potential autoimmune conditions earlier in the diagnostic process. This may help reduce the global disease burden through more timely interventions, which is particularly relevant given the progressive nature of many autoimmune conditions, where early treatment can significantly influence long-term outcomes (44). Notably, the algorithms we developed are rule-based, utilizing ICD codes and diagnostic terms, which may serve as a foundational framework for future research on more advanced approaches, such as model-based case identification algorithms.

5 Conclusion

This study presents the first comprehensive validation of identification algorithms for five major autoimmune diseases using EHR data in China. By combining diagnostic terminology and ICD-10 codes, we developed algorithms with satisfactory performance (PPV >70%) across most diseases, demonstrating their utility in real-world applications. For diseases such as Hashimoto's thyroiditis and RA, high accuracy of the algorithms was achieved using single diagnostic terms or codes. However, the algorithms of more complex diseases like IBD and ITP need to improve sensitivity, highlighting the potential for further optimization by the inclusion of additional clinical biomarkers or advanced modeling techniques. These findings suggest two key areas for improvement: leveraging additional clinical parameters (such as laboratory biomarkers and imaging data) into algorithm development, and strengthening multi-center validation to enhance algorithm generalizability across diverse healthcare settings. The validated algorithms provide essential tools for future research in autoimmune disease epidemiology and pharmacovigilance studies, while emphasizing the importance of continued methodological refinement to support more accurate disease identification in real-world settings.

Data availability statement

The datasets presented in this article are not readily available because the dataset contains sensitive or private information, and confidentiality restrictions apply to ensure the protection of privacy. Requests to access the datasets should be directed to JTY, yangjunting@bjmu.edu.cn.

Author contributions

JTY: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. YW: Conceptualization, Investigation, Writing – original draft, Writing – review & editing. JG: Writing – review & editing. XW: Writing – review & editing. XG: Writing – review & editing. XC: Writing – review & editing. MZ: Writing – review & editing. JY: Writing – review & editing. ZJL: Writing – review & editing. YL: Writing – review & editing. ZKL: Conceptualization, Formal Analysis, Funding acquisition, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. SZ: Formal Analysis, Funding acquisition, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. Fund programs: Key Program of the National Natural Science Foundation of China (82330107); Funds for International Cooperation and Exchange of the National Natural Science Foundation of China (72361127500); Young Scientists Fund of the National Natural Science Foundation of China (82204175); Bill & Melinda Gates Foundation (INV-035024).

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.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

1. Miller FW. The increasing prevalence of autoimmunity and autoimmune diseases: an urgent call to action for improved understanding, diagnosis, treatment, and prevention. *Curr Opin Immunol*. (2023) 80:102266. doi: 10.1016/j.coi.2022.102266

2. Wang L, Wang FS, Gershwin ME. Human autoimmune diseases: a comprehensive update. *J Intern Med*. (2015) 278:369–95. doi: 10.1111/joim.2015.278.issue-4

3. Cao F, He YS, Wang Y, Zha CK, Lu JM, Tao LM, et al. Global burden and cross-country inequalities in autoimmune diseases from 1990 to 2019. *Autoimmun Rev*. (2023) 22:103326. doi: 10.1016/j.autrev.2023.103326

4. Conrad N, Misra S, Verbakel JY, Verbeke G, Molenberghs G, Taylor PN, et al. Incidence, prevalence, and co-occurrence of autoimmune disorders over time and by age, sex, and socioeconomic status: a population-based cohort study of 22 million individuals in the UK. *Lancet*. (2023) 401:1878–90. doi: 10.1016/S0140-6736(23)00457-9

5. Sherman RE, Anderson SA, Dal Pan GJ, Gray GW, Gross T, Hunter NL, et al. Real-world evidence - what is it and what can it tell us? *New Engl J Med*. (2016) 375:2293–7. doi: 10.1056/NEJMsb1609216

6. U.S. Food and Drug Administration. *Framework for FDA's Real-world evidence program* (2018). Available online at: https://www.fda.gov/downloads/ScienceResearch/SpecialTopics/RealWorldEvidence/UCM627769.pdf (Accessed December 08, 2023).

7. Stuurman AL, Sharan A, Jahagirdar S, Elango V, Riera-Montes M, Kashyap N, et al. WHO global vaccine safety multi-country collaboration project on safety in pregnancy: Assessing the level of diagnostic certainty using standardized case definitions for perinatal and neonatal outcomes and maternal immunization. *Vaccine X*. (2021) 9:100123. doi: 10.1016/j.jvacx.2021.100123

8. Ducharme R, Benchimol EI, Deeks SL, Hawken S, Fergusson DA, Wilson K. Validation of diagnostic codes for intussusception and quantification of childhood intussusception incidence in Ontario, Canada: a population-based study. *J Pediatr*. (2013) 163:1073–9.e3. doi: 10.1016/j.jpeds.2013.05.034

9. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Inform Assoc*. (2013) 20:117–21. doi: 10.1136/amiajnl-2012-001145

10. He T, Belouali A, Patricoski J, Lehmann H, Ball R, Anagnostou V, et al. Trends and opportunities in computable clinical phenotyping: A scoping review. *J BioMed Inform*. (2023) 140:104335. doi: 10.1016/j.jbi.2023.104335

11. Mcdonough CW, Babcock K, Chucri K, Crawford DC, Bian J, Modave F, et al. Optimizing identification of resistant hypertension: Computable phenotype development and validation. *Pharmacoepidemiol Drug Saf*. (2020) 29:1393–401. doi: 10.1002/pds.v29.11

12. Geva A, Gronsbell JL, Cai T, Cai T, Murphy SN, Lyons JC, et al. A computable phenotype improves cohort ascertainment in a pediatric pulmonary hypertension registry. *J Pediatr*. (2017) 188:224–31.e5. doi: 10.1016/j.jpeds.2017.05.037

13. Deng S, Liu Z, Yang J, Zhang L, Shou T, Zhu J, et al. Diagnostic validation and development of an algorithm for identification of intussusception in children using electronic health records of Ningbo city in China. *Expert Rev Vaccines*. (2023) 22:307–14. doi: 10.1080/14760584.2023.2189474

14. Liu X, Shen P, Zhang D, Sun Y, Chen Y, Liang J, et al. Evaluation of atherosclerotic cardiovascular risk prediction models in China: results from the CHERRY study. *JACC Asia*. (2022) 2:33–43. doi: 10.1016/j.jacasi.2021.10.007

15. Zhou X, Lee EWJ, Wang X, Lin L, Xuan Z, Wu D, et al. Infectious diseases prevention and control using an integrated health big data system in China. *BMC Infect Dis*. (2022) 22:344. doi: 10.1186/s12879-022-07316-3

16. Yang J, Welby S, Liu Z, Deng S, Liu G, Meng R, et al. Monitoring the safety of the adjuvanted human papillomavirus vaccine, HPV-16/18-AS04: Protocol for a cohort study using electronic health records in Yinzhou, China. *Hum Vaccines Immunotherapeutics*. (2024) 20:2378535. doi: 10.1080/21645515.2024.2378535

17. Wang HY, Ding GH, Lin H, Sun X, Yang C, Peng S, et al. Influence of doctors' perception on the diagnostic status of chronic kidney disease: results from 976 409 individuals with electronic health records in China. *Clin Kidney J*. (2021) 14:2428–36. doi: 10.1093/ckj/sfab089

18. Health Commission Of The People's Republic Of China N. National guidelines for diagnosis and treatment of thyroid cancer 2022 in China (English version). *Chin J Cancer Res*.. (2002) 34:131–50. doi: 10.21147/j.issn.1000-9604.2022.03.01

19. Inflammatory Bowel Disease Group, Chinese Society of Gastroenterology and Chinese Medical Association. Chinese consensus on diagnosis and treatment in inflammatory bowel disease (2018, Beijing). *J Dig Dis*. (2021) 22:298–317. doi: 10.1111/1751-2980.12994

20. Bernstein CN, Fried M, Krabshuis JH, Cohen H, Eliakim R, Fedail S, et al. World Gastroenterology Organization Practice Guidelines for the diagnosis and management of IBD in 2010. *Inflammation Bowel Dis*. (2010) 16:112–24. doi: 10.1002/ibd.21048

21. Wise RP, Bonhoeffer J, Beeler J, Donato H, Downie P, Matthews D, et al. Thrombocytopenia: case definition and guidelines for collection, analysis, and presentation of immunization safety data. *Vaccine*. (2007) 25:5717–24. doi: 10.1016/j.vaccine.2007.02.067

22. Hui D, Barbara A, McGill SC. *CADTH Health Technology Review [M]. Guidelines for Pediatric Immune Thrombocytopenia: Rapid Review*. Ottawa (ON: Canadian Agency for Drugs and Technologies in Health Copyright © 2022 Canadian Agency for Drugs and Technologies in Health (2022).

23. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum*. (1988) 31:315–24. doi: 10.1002/art.1780310302

24. Aletaha D, Neogi T, Silman AJ, Funovits J, Felson DT, Bingham CO 3rd, et al. 2010 rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Ann Rheum Dis*. (2010) 69:1580–8. doi: 10.1136/ard.2010.138461

25. Chinese Diabetes Society, Chinese Endocrinologist Association, Chinese Society of Endocrinology, Chinese Pediatric Society, Zhou ZG. Guidelines for the diagnosis and treatment of type 1 diabetes in China. *Clin Diabetes*. (2013) 7:6–21. doi: 10.3760/cma.j.cn115791-20220916-00474

26. Klompas M, Eggleston E, Mcvetta J, Lazarus R, Li L, Platt R. Automated detection and classification of type 1 versus type 2 diabetes using electronic health record data. *Diabetes Care*. (2013) 36:914–21. doi: 10.2337/dc12-0964

27. Almutairi K, Inderjeeth C, Preen DB, Keen H, Rogers K, Nossent J. The accuracy of administrative health data for identifying patients with rheumatoid arthritis: a retrospective validation study using medical records in Western Australia. *Rheumatol Int*. (2021) 41:741–50. doi: 10.1007/s00296-021-04811-9

28. Lee CK, Ha HJ, Oh SJ, Kim JW, Lee JK, Kim HS, et al. Nationwide validation study of diagnostic algorithms for inflammatory bowel disease in Korean National Health Insurance Service database. *J Gastroenterol Hepatol*. (2020) 35:760–8. doi: 10.1111/jgh.14855

29. Heden KE, Jensen AO, Farkas DK, Nørgaard M. Validity of a procedure to identify patients with chronic idiopathic thrombocytopenic purpura in the Danish National Registry of Patients. *Clin Epidemiol*. (2009) 1:7–10. doi: 10.2147/clep.s4832

30. Hutfless S, Jasper RA, Tilak A, Ghosh T, Kedia S, Liu S, et al. A systematic review of crohn's disease case definitions in administrative or claims databases. *Inflammation Bowel Dis*. (2023) 29:705–15. doi: 10.1093/ibd/izac131

31. Chung CP, Rohan P, Krishnaswami S, McPheeters ML. A systematic review of validated methods for identifying patients with rheumatoid arthritis using administrative or claims data. *Vaccine*. (2013) 31 Suppl 10:K41–61. doi: 10.1016/j.vaccine.2013.03.075

32. Lethebe BC, Williamson T, Garies S, McBrien K, Leduc C, Butalia S, et al. Developing a case definition for type 1 diabetes mellitus in a primary care electronic medical record database: an exploratory study. *CMAJ Open*. (2019) 7:E246–E51. doi: 10.9778/cmajo.20180142

33. Chi GC, Li X, Tartof SY, Slezak JM, Koebnick C, Lawrence JM. Validity of ICD-10-CM codes for determination of diabetes type for persons with youth-onset type 1 and type 2 diabetes. *BMJ Open Diabetes Res Care*. (2019) 7:e000547. doi: 10.1136/bmjdrc-2018-000547

34. Danieli MG, Brunetto S, Gammeri L, Palmeri D, Claudi I, Shoenfeld Y, et al. Machine learning application in autoimmune diseases: State of art and future prospectives. *Autoimmun Rev*. (2024) 23:103496. doi: 10.1016/j.autrev.2023.103496

35. Ahalya RK, Umapathy S, Krishnan PT, Joseph Raj AN. Automated evaluation of rheumatoid arthritis from hand radiographs using Machine Learning and deep learning techniques. *Proc Inst Mech Eng H*. (2022) 236:1238–49. doi: 10.1177/09544119221109735

36. Acharya UR, Sree SV, Krishnan MM, Molinari F, Zieleźnik W, Bardales RH, et al. Computer-aided diagnostic system for detection of Hashimoto thyroiditis on ultrasound images from a Polish population. *J Ultrasound Med*. (2014) 33:245–53. doi: 10.7863/ultra.33.2.245

37. Zhou SM, Fernandez-Gutierrez F, Kennedy J, Cooksey R, Atkinson M, Denaxas S, et al. Defining disease phenotypes in primary care electronic health records by a

machine learning approach: A case study in identifying rheumatoid arthritis. *PLoS One*. (2016) 11:e0154515. doi: 10.1371/journal.pone.0154515

38. Lash TL, Fox MP, Maclehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. *Int J Epidemiol*. (2014) 43:1969–85. doi: 10.1093/ije/dyu149

39. Newcomer SR, Xu S, Kulldorff M, Daley MF, Fireman B, Glanz JM. A primer on quantitative bias analysis with positive predictive values in research using electronic health data. *J Am Med Inform Assoc*. (2019) 26:1664–74. doi: 10.1093/jamia/ocz094

40. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol*. (2005) 58:323–37. doi: 10.1016/j.jclinepi.2004.10.012

41. Coloma PM, Schuemie MJ, Trifiro G, Gini R, Herings R, Hippisley-Cox J, et al. Combining electronic healthcare databases in Europe to allow for large-scale drug safety monitoring: the EU-ADR Project. *Pharmacoepidemiol Drug Saf*. (2011) 20:1–11. doi: 10.1002/pds.v20.1

42. Klompas M, Mcvetta J, Lazarus R, Eggleston E, Haney G, Kruskal BA, et al. Integrating clinical practice and public health surveillance using electronic medical record systems. *Am J Public Health*. (2012) 102 Suppl 3:S325–32. doi: 10.2105/AJPH.2012.300811

43. Molly Murray. *Guest Blog: A Major Health Crisis: The Alarming Rise of Autoimmune Disease*. Available online at: https://nationalhealthcouncil.org/blog/a-major-health-crisis-the-alarming-rise-of-autoimmune-disease National health council (2024). (Accessed December 08, 2024).

44. Wylezinski LS, Gray JD, Polk JB, Harmata AJ, Spurlock CF 3rd. Illuminating an invisible epidemic: A systemic review of the clinical and economic benefits of early diagnosis and treatment in inflammatory disease and related syndromes. *J Clin Med*. (2019) 8. doi: 10.3390/jcm8040493