Integrative Medicine Research 14 (2025) 101100

Contents lists available at ScienceDirect

Integrative Medicine Research

journal homepage: www.elsevier.com/locate/imr

Education Article

Methods for identifying health status from routinely collected health data: An overview

Mei Liu [a,b,c,d,1], Ke Deng [a,c,d,1], Mingqi Wang [a,c,d], Qiao He [a,c,d], Jiayue Xu [a,c,d], Guowei Li [e,f,g], Kang Zou [a,c,d], Xin Sun [a,c,d,h,*], Wen Wang [a,c,d,*]

[a] Institute of Integrated Traditional Chinese and Western Medicine, Chinese Evidence-based Medicine and Cochrane China Center, West China Hospital, Sichuan University, Chengdu, China
[b] Hospital of Chengdu University of Traditional Chinese Medicine, Chengdu, China
[c] NMPA Key Laboratory for Real World Data Research and Evaluation in Hainan, Chengdu, China
[d] Sichuan Center of Technology Innovation for Real World Data, Chengdu, China
[e] Department of Health Research Methods, Evidence and Impact, McMaster University, Hamilton, ON, Canada
[f] Center for Clinical Epidemiology and Methodology, Guangdong Second Provincial General Hospital, Guangzhou, Guangdong, China
[g] Biostatistics Unit, Research Institute at St. Joseph's Healthcare Hamilton, Hamilton, ON, Canada
[h] West China School of Public Health and West China Fourth Hospital, Sichuan University, Chengdu, China

ARTICLE INFO

Key words:
Routinely collected health data
Health status
Machine learning algorithms
Rule-based algorithms

ABSTRACT

Routinely collected health data (RCD) are currently accelerating publications that evaluate the effectiveness and safety of medicines and medical devices. One of the fundamental steps in using these data is developing algorithms to identify health status that can be used for observational studies. However, the process and methodologies for identifying health status from RCD remain insufficiently understood. While most current methods rely on International Classification of Diseases (ICD) codes, they may not be universally applicable. Although machine learning methods hold promise for more accurately identifying the health status, they remain underutilized in RCD studies. To address these significant methodological gaps, we outline key steps and methodological considerations for identifying health statuses in observational studies using RCD. This review has the potential to boost the credibility of findings from observational studies that use RCD.

1. Introduction

Routinely collected health data (RCD) refers to data generated from various clinical practice settings, including electronic medical records and administrative claims.[1] These rich data provide a great opportunity for evaluation of the effectiveness and safety of drugs and medical devices, diagnostic and prognostic studies, helping to guide clinical care and health policy decision-making.[2-6] In the field of integrative medicine researches, RCD can provide vast information include detailed information on diagnosis, traditional Chinese medicine (TCM) medications, or laboratory test results.[4] In recent years, RCD has increasingly been used to investigate the effectiveness, utilization patterns, and post-marketing safety of TCM medication.[7-10] However, the methodological quality of such studies has often been insufficient, significantly impacting the successful translation and application of the research findings.[11]

Given that RCDs are typically collected for administrative or healthcare purposes,[1,12,13] they may pose methodological challenges when utilized for research purposes. One of the fundamental steps in leveraging RCD is identifying patients with certain health status that are of interest for research questions. For example, a researcher may want to identify patients diagnosed with type 2 diabetes, patients suspected with sepsis, or prescriptions of anti-depression medications. A multitude of methodologies can be employed to identify health statuses, ranging from simplistic approaches like utilizing ICD codes to more sophisticated techniques such as novel computational methods, notably machine learning algorithms.[14]

However, the application of inappropriate methods for health status identifying can pose a significant risk of misclassification, leading to potential inaccuracies and erroneous conclusions.[15-17] For example, a study published in *JAMA* showed a notable variation in sepsis incidence, ranging from 2 % to 12 %, when different algorithms were

* Corresponding authors at: Institute of Integrated Traditional Chinese and Western Medicine, Chinese Evidence-based Medicine and Cochrane China Center, West China Hospital, Sichuan University, Chengdu, China.
E-mail addresses: sunxin@wchscu.cn (X. Sun), wangwen@wchscu.cn (W. Wang).
[1] The authors contributed equally to this work as co-first authors.

https://doi.org/10.1016/j.imr.2024.101100
Received 19 September 2024; Received in revised form 1 November 2024; Accepted 13 November 2024
Available online 15 November 2024

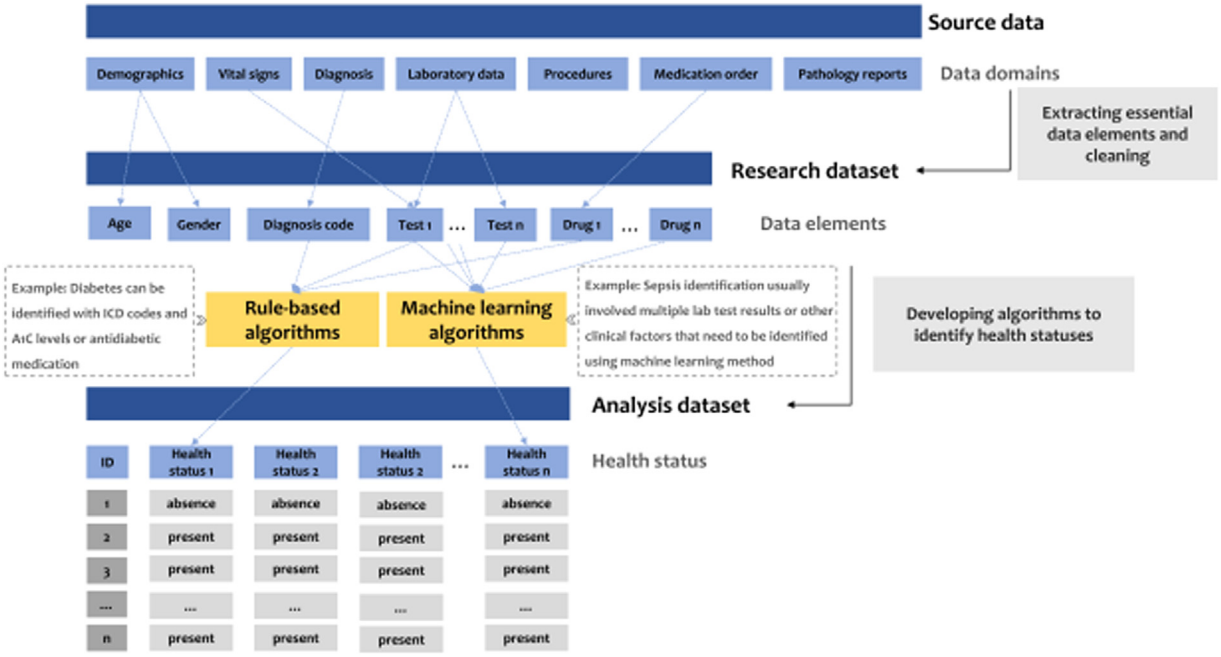

Fig. 1. illustration of key steps for identifying health statuses from RCD.
Rule-based approaches involve the use of expert-defined logic rules, such as those based on specific indicators like ICD codes, medication prescription, or specific laboratory test results. These rules often utilize Boolean logic (e.g., AND, OR, NOT) and other logical criteria to identify and classify features. The machine learning approaches may involve models such as random forest, decision trees, regression, naïve bayes, which are trained on a dataset to learn patterns and make classification predictions.

applied.[18] Nonetheless, many studies using RCD often overlook the requisite scrutiny of the rationale or suitability of the algorithms employed, frequently relying solely on ICD codes for health status identification.[19,20] Yet, the utilization of ICD codes for identifying specific diseases may present challenges due to factors such as unrecorded diagnostic findings or the absence of specific diagnostic codes (e.g., pseudogout).[21-24] Furthermore, while some researchers assert that machine learning methods may offer enhanced accuracy in identifying some specific health status, such as sepsis,[25] these methods were seldom used in observational studies using RCD.[19]

Thus, we believe it is important for all stakeholders to understand when and how to employ appropriate algorithms. In this review, we outline the key steps involved in employing algorithms to identify health status in observational studies using RCD. Additionally, we provide an overview of the pipelines, strengths, and weaknesses of various methods used to develop identification algorithms, with the goal of supporting future research.

2. The importance and general steps for health status identification in RCD studies

In contrast to randomized controlled trials, where health statuses of participants are prospectively collected with uniform definition, RCD presents a challenge as health statuses cannot be directly obtained. RCD encompasses diverse healthcare data domains, such as demographic characteristics, vital signs, diagnoses, laboratory test results, medication prescriptions, and pathology reports. However, these data were typically collected for administrative purposes and lacking uniform definition.

Consequently, to achieve a uniform definition, researchers must undertake additional steps to extract relevant information from the diverse data sources available in RCD. This involves data pre-processing to construct a research dataset, and developing algorithms to identify or categorize health status by employing standardized definitions. Fig. 1 illustrates the key steps for identifying health statuses from RCD.

2.1. Constructing research dataset

RCD were originally collected during the care of patients and were integrated in to Clinical Data Warehouses (CDW). The source data in CDW includes a large variety of information, ranging from structured data—such as ICD codes for diagnosis, procedure codes and laboratory data—to free-text clinical narratives and imagines. The source data often exhibit characteristics like noise, sparsity, and incomplete records.[26,27] To effectively use RCD for clinical research, researchers must undertake rigorous data curation, which may include the integration of data from multiple healthcare institutions. This process also entails the selective extraction of pertinent data elements that are aligned with specific research objectives, followed by the cleansing and standardization of these elements to assemble a coherent research dataset.[28,29]

A desirable research dataset should include a comprehensive set of data elements that satisfy the research objectives.[29] For example, in studies on diabetes, the research dataset may at least include elements such as ICD codes for diabetes and its comorbidity, prescriptions for antidiabetic medications, levels of glucose, and so forth.[30-33] Therefore, it is crucial for researchers to evaluate the accessibility and quality of the essential data elements from the selected source data.[28,29]

2.2. Identifying health status

Identifying health status from the research dataset is a fundamental step to answer a research question. For example, researchers may employ algorithms to ascertain if a patient has sepsis, and this determination could be pivotal for assessing their eligibility for inclusion, or ascertaining the presence or absence of outcomes of interest within the target population, or establishing baseline characteristics of the study cohort.

However, the meta-data within the research dataset usually cannot be directly used to identify the health status. For example, the absence of a diagnosis code for a patient usually does not automatically imply the absence of a particular disease.[14,34,35] For instance, previous stud-

ies found that over 40 % of patients with Alzheimer's disease were not identified by the specific diagnosis code (i.e., 692.0 or 698.1),[36,37] and over 60 % of patients with sepsis were not coded.[38] Thus, researches should dedicatedly develop the algorithms that can accurately identify the health status.

Identifying health status typically involves developing computer-executable algorithms that combined multiple data elements.[35] For example, in identifying pulmonary arterial hypertension (PAH), an algorithm using only ICD-9 codes 416.0 and 416.8 yielded a positive predictive value (PPV) of only 9.3 %, while integrating information on two or more classes of PAH specific medications increased the PPV to 66.9 %.[39] Therefore, to address a specific clinical question, it may be necessary to select various data elements to define the health status and incorporate them into algorithms.

2.3. Relationship between research dataset and health statuses identification

Both establishing a research dataset and identifying health status are vital steps for conducting studies using RCD.[40] The construction of a research dataset entails the extraction and refinement of a comprehensive pool of data, often without strict adherence to a predefined research question. Conversely, the identification of health statuses serves as a crucial means of structuring the research dataset to align with the analytical requirements. Unlike dataset construction, health status identification demands a well-defined research inquiry and standardized definitions. By applying identification algorithms developed using multiple data elements from research dataset, researchers can effectively select participants meeting inclusion criteria or attribute specific health status values. Subsequently, the research dataset undergoes further refinement to transition into the analytical dataset.

While the process of constructing the research dataset has been extensively discussed elsewhere,[29] in this paper, we focus primarily on introducing the methods for identifying health status.

3. Methods for developing algorithms for health status identification

In general, two primary methods are commonly employed in the development of identification algorithms: rule-based methods and machine learning methods.[41-43] Rule-based methods involve integrating data elements through logical rules derived by experts, while machine learning methods utilize models to classify whether a patient falls into the category of cases or controls.[41] Each method may be suited to different scenarios with its own strengths and weaknesses.[44] In order to assist researchers in selecting the most suitable method and enhancing the accuracy of the identification algorithms, we have compiled details regarding the applicability, strengths, and limitations of both methods.

3.1. Rule-based methods

3.1.1. Application scenarios

The rule-based algorithms are particularly suitable for identifying health statuses that are characterized by simple disease or well-defined diagnoses.[44,45] Examples include diabetes,[46,47] atrial fibrillation (AF),[48] and cirrhosis,[49,50] all of which have been successfully identified using rule-based algorithms with good accuracy.

The suitability of rule-based algorithms for identifying the status of these conditions lies in their well-established nature in clinical settings. The accuracy or completeness of diagnosis codes for these conditions is often higher. In addition, these diseases also can be confirmed by the objective laboratory test results or medications. Therefore, they can be effectively identify using structured data elements (such as diagnosis code results of laboratory test, vital signs, and prescription orders)[13,16,51] and logic rules, such as Boolean operations (e.g., OR, AND, NOT), comparative operations (e.g., setting a threshold for laboratory test results), and aggregative functions (e.g., CHADS2).[43,44]

For instance, determining whether a patient has diabetes can involve logical assessments using ICD-9 codes related to diabetes, levels of glucose or hemoglobin A_{1c} and antidiabetic medications.[52] Similarly, algorithms that employ logical combinations of data elements—such as the frequency of electrocardiograms with AF, electrical cardioversion, antiarrhythmic use, novel oral anticoagulant use, and the ICD-9 codes for AF—can assist in determining whether a patient has AF.[48]

3.1.2. Key points for development

Rule-based algorithms exhibit significant variability. Different algorithms can be developed for the same disease based on varying combinations of data elements and rules, resulting in differing levels of accuracy.[39,53] For example, algorithms for identifying type 2 diabetes can be developed using ICD codes from outpatient or inpatient setting, or combining elements such as elevated fasting glucose or hemoglobin A_{1c}. However, algorithms relying on solely outpatient diagnoses typically show lower accuracy compared to those based on inpatient diagnoses, and incorporating laboratory tests may help enhance the positive predictive values at the expense of sensitivity.[34] Understanding how different data elements affect algorithm accuracy is essential for researchers. Table 1 displays considerations for utilizing various types of data elements in developing identification algorithms.

Regardless of the data elements used, the logical constraints (rules) should adhere to healthcare guidelines or expert opinions.[54] In the development of rule-based algorithms, the active participation of domain clinicians is crucial important. Clinicians can provide invaluable insights into disease diagnosis, medication usage trends, and characteristics for distinguishing similar conditions.

3.1.3. Strengths and weakness

Rule-based methods have been the most commonly employed techniques for developing identification algorithms, showcasing several strengths.[41] These algorithms inherently provide interpretability, assisting in standard health status definition and enabling effective classification.[44] Additionally, rule-based algorithms offer flexibility through the easy addition or removal of data elements, enabling adjustments to algorithm accuracy to meet specific research needs. Different health status may have distinct requirements for algorithm accuracy (e.g., sensitivity, specificity). For instance, when the health status was used to determine the occurrence of outcomes among participants, the algorithms may require high specificity.[55,56]

However, these strengths diminish when dealing with complex conditions. The development of rule-based algorithms heavily relies on human customization.[44] When dealing with health statuses that involve high-dimensional data elements, the effort and time required for developing the algorithms can be significant, potentially extending beyond six months.[41,44]

3.2. Machine learning based methods

3.2.1. Application scenarios

Machine learning methods are particularly valuable for identification tasks involving complex diseases, vague diagnoses, subjective judgement,[57] as well as scenarios where the relationship between input data and the health status (e.g., presence or not presence of health event of interest) is non-linear or too complex to be formulated by simple logic rules or functions.[57]

Take primary hyperparathyroidism (PHPT) as an example. The clinical diagnosis for PHPT relies on multiple laboratory test results including PTH, calcium, and vitamin D. However, the trend of these indices may vary and relies more on the subjective judgement by clinicians. For example, some cases may exhibit hypercalcemia alongside inappropriately normal concentrations of parathyroid hormone (PTH), while others may show a normal rage of calcium concentration but elevated

Table 1
Considerations when using different type of data elements for developing rule-based algorithms.

Data elements	Aspects for Considerations	Explanation
Diagnosis codes	Sources of codes	The accuracy of diagnostic codes varies across different sources; generally, diagnostic codes derived from outpatient settings exhibit lower accuracy compared to those from inpatient settings.
	Versions of codes	The ICD coding version is periodically updated; researchers should be mindful that for the codes may vary when the data selected periods are changed.
Laboratory data	Order context	Routine tests typically have fewer instances of missing data than non-routine tests, making them more suitable for algorithm development.
	Types of specimens	When utilizing laboratory data, it is crucial to consider the specificity of the sample type concerning the disease under identification. For instance, arterial blood testing may provide more accurate insights into myocardial ischemia compared to venous blood.
	Repeated measures of laboratory results	When utilizing laboratory data, the choice of utilization the repeated measures—whether using the median, maximum, most recently, or the timing trend—partly depends on the definition or features of health events to be identified.
Medication order	Therapeutic indication	When utilizing medication information, it is essential to consider issues related to off-label drug use. For instance, antiepileptic order records do not exclusively indicate the presence of epilepsy, as these drugs are frequently used for preventing migraines or managing chronic pain as well.

ICD: international classification of diseases.

PTH levels.[58] Moreover, the PTH may be undermeasured in over 60 % of the PHPT cases.[59] This complexity makes it challenging to develop rule-based algorithms using these data elements. Yash and colleagues developed machine learning algorithms using preoperative calcium, vitamin D, creatinine, PTH as inputs, achieving the accuracy of 95.2 %.[60] Interestingly, even when excluding PTH, which had the most missing values among predictors, the accuracy of the identification algorithms did not significantly decrease.[60]

The case of sepsis also illustrates the application of machine learning methods. Sepsis is a complex syndrome involving organ dysfunction resulting from infection.[61] Several rule-based algorithms have been developed to identify sepsis, but they tend to exhibit low sensitivity or precision due to the fuzzy pathobiology of sepsis, which is not easily captured by rules.[62,63] For this condition, machine learning methods have demonstrated higher accuracy for identification.[64,65]

3.2.2. Key points for development

The development of identification algorithms using machine learning includes supervised or unsupervised methods. Supervised learning relies on labeled data to identify the status of health event of interest, whereas unsupervised methods uncover patterns without labeled health event status. Interpreting the resulting classification of health status derived from unsupervised learning is challenging, as there may be no clear ground truth for the assigned groups.[41] Most existing machine learning applications in health status identification rely on supervised learning.[57] However, the accuracy of these algorithms heavily relies on the quality of training labels, which can be obtained through various strategies such as manual chart review, linkage to external databases, expert-defined rules, and pathology reports.

In addition to the quality of training labels, the applicability of the model plays a crucial role in ensuring accurate identification. For instance, the regression or tree model can provide the clear relationship between input data and the health events but limited by their simplicity. Neural networks models can handle complex or non-linear relationships (e.g., dynamic time-series data or time-dependent relationships) but are limited in interpretability.[14,64] These considerations are vital throughout the algorithm development process to achieve precise disease identification within EHR data.

3.2.3. Strengths and weakness

Machine learning methods have the potential to effectively process noisy, high-dimensional data and uncover hidden patterns. This approach is particularly suitable for complex situations where the disease to be identified has a fuzzy boundary with similar diseases, as it does not require the health event to be well pre-specified.[14,57] Additionally, machine learning methods are less affected by data quality issues compared to rule-based methods.[66]

However, developing a machine learning algorithm comes with several limitations. Firstly, obtaining high-quality training labels can be challenging, often requiring a significant number of manual chart reviews, which is time-consuming and demands expertise from domain experts.[41] Secondly, machine learning algorithms, especially deep learning methods, can be challenging to interpret, as they often function as "black boxes".[67] The lack of transparency raises uncertainty about how the algorithm makes decisions and how its classification outcomes align with the definitions of health statuses in the research questions. Another limitation lies in the portability of the model. If the model was trained on poorly representative samples, its performance may not be satisfactory when applied to the entire dataset, despite achieving good performance on the training data.

4. The future

We see many efforts have been doing to assist the identification of health statuses.

Firstly, there are publicly available platforms that have collected a variety of developed and validated algorithms.[68-70] Although the portability of algorithms across different databases is challenging, the data elements, logic rules, or machine learning models within these methods serve as valuable references for algorithms development.

Secondly, the utilization of unstructured data (e.g., free texts written by clinician) can help improve the performance of identification algorithms, which may require pre-processing through natural language processing (NLP) techniques[16,71] or large language model (LLM).[72] Studies have found that NLP have the potential to facilitate the automated transformation of free texts written by clinicians into structured codes.[73]

Thirdly, ongoing standardization efforts in electronic health record (EHR) data, such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMET CT), provide promising future for enhancing data structuring and facilitating more effective information utilization.

These efforts hold the potential to significantly advance the identification of health statuses and facilitate studies on drug safety, exploration of risk factors, and analysis of healthcare utilization patterns. Furthermore, they provide a path for rapid translation of research findings in healthcare delivery.

5. Practical example—an implication in integrative medicine research

The integrated TCM and Western Medicine (WM) approach has been widely adopted for treating acute pancreatitis (AP) in China.[4] To investigate the treatment effect of integrative therapy, we established a research database comprising 44,033 AP patients. In this study, health statuses of interest include TCM syndromes of included patients, the presence or absence of sepsis, and whether patients received the treatment of chaiqin chengqi decoction (CQCQD).

To identify patients who received CQCQD, we employed the rule-based algorithms, since this information was well-structured and suitable to such an approach. For identifying patients with sepsis, we opted for a machine learning approach due to the diagnostic complexity of sepsis, which involves multiple clinical factors that cannot be easily captured by simple logic rules. To enhance the classification performance, we implemented rigorous labeling strategies and chose appropriate models.[56]

With respect to TCM syndrome, the information is generally embedded within unstructured free-text documents, such as admission summaries and progress notes, making it challenging to directly apply for clinical research.[74] To address this issue, we employed NLP technology to assist in extracting relevant TCM syndrome information from these texts and encoding it into a structured format. Then, the presence or absence of a specific TCM syndrome were identified using rule-based algorithms composed of the codes we have assigned.

6. Conclusions

In light of the exponential increase in the use of RCD in clinical studies, efforts to enhance the quality of these studies are undoubtedly necessary. Identifying health statuses represent a fundamental step in implementing studies using RCD. The choice of suitable methods for developing identification algorithms may vary depending on the characteristics of health events to be identified. The careful selection of an appropriate algorithm for identifying health statuses emerge as a pivotal step in reducing resources waste or enhancing the accuracy of algorithms to be constructed.

This review provides application scenarios, strengths and weakness of different methods for developing identification algorithms. As the structured nature of RCD evolves and clinical recognition or diagnosis methods for diseases advance, the most suitable methods may shift. Researchers need to adopt a dynamic perspective, considering these changes in the context of clinical questions. Regardless of the method used to develop the algorithm, thorough validation is imperative to provide further insights into the robustness of study results.

Author's contributions

All authors contributed to the conceptualization, design, interpretation and reviewing the manuscript. ML wrote the original draft, XS and WW revised it and acquired fundings.

Declaration of competing interest

The authors declare that they have no conflicts of interest.

Funding

This study was supported by National Natural Science Foundation of China (Grant No. 82225049, 72104155), Special fund for traditional Chinese medicine of Sichuan Provincial Administration of Traditional Chinese Medicine (Grant No. 2024zd023), and the 1·3·5 Project for Disciplines of Excellence, West China Hospital, Sichuan University (Grant No. ZYGD23004).

Ethical statement

No ethical approval was required as this study did not involve human participants or laboratory animals.

Data availability

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

References

1. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med*. 2015;12(10):e1001885.
2. 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(23):2293–2297.
3. Janssen A, Shah K, Keep M, Shaw T. Community perspectives on the use of electronic health data to support reflective practice by health professionals. *BMC Med Inform Decis Mak*. 2024;24(1):226.
4. Deng L, Chen Z, Zhu P, Hu C, Jin T, Wang X, et al. Effects of integrated traditional Chinese and Western medicine for acute pancreatitis: A real-world study in a tertiary teaching hospital. *J Evid Based Med*. 2024;17(3):575–587.
5. Peng L, Zhang K, Li Y, Chen L, Gao H, Chen H. Real-world evidence of traditional chinese medicine (TCM) treatment on cancer: a literature-based review. *Evid Based Complement Alternat Med*. 2022;2022:7770380.
6. Xinyao J, Yifan Z, Keyi W, Wentai P, Chunyang W, Hui W, et al. Post-marketing safety surveillance and re-evaluaiton of Shu-Xue-Ning injection: a real-world study based on 30,122 cases. *Front Pharmacol*. 2023;14:1194367.
7. Lyu J, Liu Y, Liu F, Liu G, Gao Y, Wei R, et al. Therapeutic effect and mechanisms of traditional Chinese medicine compound (Qilong capsule) in the treatment of ischemic stroke. *Phytomedicine*. 2024;132:155781.
8. Chang Y, Zhang W, Xie Y, Xu X, Sun R, Wang Z, et al. Postmarketing safety evaluation: depside salt injection made from Danshen (Radix Salviae Miltiorrhizae). *J Tradit Chin Med*. 2014;34(6):749–753.
9. Wang W, He Q, Wang MQ, Xu JY, Ji P, Zhang R, et al. Effects of tanreqing injection on ICU mortality among icu patients receiving mechanical ventilation: time-dependent cox regression analysis of a large registry. *Chin J Integr Med*. 2023;29(9):782–790.
10. Zhang X, Wang M, Wang W, Li L, Sun X. Utilization of traditional Chinese medicine in the intensive care unit. *Chin Med*. 2021;16(1):84.
11. Xu J, Wu W, Jia J, Du L, Wang W, Sun X. Methodology quality was inadequate for observational studies investigating drug safety of Chinese patent medicine using real-world data: A cross-sectional survey. *J Evid Based Med*. 2024;17(3):483–485.
12. Langan SM, Schmidt SA, Wing K, Ehrenstein V, Nicholls SG, Filion KB, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ (Clinical research ed)*. 2018;363:k3532.
13. Nissen F, Quint JK, Morales DR, Douglas IJ. How to validate a diagnosis recorded in electronic health records. *Breathe (Sheffield, England)*. 2019;15(1):64–68.
14. Casey JA, Schwartz BS, Stewart WF, Adler NE. Using Electronic Health Records for Population Health Research: A Review of Methods and Applications. *Annu Rev Public Health*. 2016;37(1):61–81.
15. Prada-Ramallal G, Takkouche B, Figueiras A. Bias in pharmacoepidemiologic studies using secondary health care databases: a scoping review. *BMC Med Res Methodol*. 2019;19(1):53.
16. Weinstein EJ, Ritchey ME, Lo Re V. Core concepts in pharmacoepidemiology: Validation of health outcomes of interest within real-world healthcare databases. *Pharmacoepidemiol Drug Saf*. 2022;32(1):1–8.
17. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. *BMJ (Clinical research ed)*. 2010;341:c4226.
18. Rhee C, Dantes R, Epstein L, Murphy DJ, Seymour CW, Iwashyna TJ, et al. Incidence and Trends of Sepsis in US Hospitals Using Clinical vs Claims Data, 2009-2014. *Jama*. 2017;318(13):1241–1249.
19. Wang W, Liu M, He Q, Wang M, Xu J, Li L, et al. Validation and impact of algorithms for identifying variables in observational studies of routinely collected data. *J Clin Epidemiol*. 2023;166:111232.
20. Hemkens LG, Benchimol EI, Langan SM, Briel M, Kasenda B, Januel JM, et al. The reporting of studies using routinely collected health data was often insufficient. *J Clin Epidemiol*. 2016;79:104–111.
21. Johnson AEW, Aboab J, Raffa JD, Pollard TJ, Deliberato RO, Celi LA, et al. A Comparative Analysis of Sepsis Identification Methods in an Electronic Database. *Crit Med*. 2018;46(4):494–499.
22. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *J Am Medical Inform Assoc*. 2012;19(e1):e162–e169.
23. Bartels CM, Singh JA, Parperis K, Huber K, Rosenthal AK. Validation of administrative codes for calcium pyrophosphate deposition: a Veterans Administration study. *J Clin Rheumatol*. 2015;21(4):189–192.
24. Fryer K, Reid CN, Chaphalkar C, Marshall J, Szalacha L, Johnson K, et al. Development of a 5-Step electronic medical record-based algorithm to identify patients with opioid use disorder in pregnancy. *J Registry Manag*. 2024;51(2):69–74.

25. Saria S, Henry KE. Too many definitions of sepsis: can machine learning leverage the electronic health record to increase accuracy and bring consensus? *Criti Care Med*. 2020;48(2):137–141.
26. Knevel R, Liao KP. From real-world electronic health record data to real-world results using artificial intelligence. *Ann Rheum Dis*. 2022.
27. Honerlaw J, Ho YL, Fontin F, Murray M, Galloway A, Heise D, et al. Centralized interactive phenomics resource: an integrated ny phenomics knowledgebase for health data users. *J Am Med Inform Assoc*. 2024;31(5):1126–1134.
28. Wang W, Zhu S, He Q, Zhang R, Kang Y, Wang M, et al. Developing a registry of healthcare-associated infections at intensive care units in west china: study rationale and patient characteristics. *Clin Epidemiol*. 2019;11:1035–1045.
29. Wen W, Pei G, Jing W, Jianwei X, Xiaoning H, Ming H, et al. Technical guidance for developing research databases using existing health and medical data. *J Evid Based Med*. 2019;19:763–770.
30. Zhong VW, Obeid JS, Craig JB, Pfaff ER, Thomas J, Jaacks LM, et al. An efficient approach for surveillance of childhood diabetes by type derived from electronic health record data: the SEARCH for Diabetes in Youth Study. *J Am Med Inform Assoc*. 2016;23(6):1060–1067.
31. Zhong VW, Juhaeri J, EJ Mayer-Davis. Trends in Hospital admission for diabetic ketoacidosis in adults with Type 1 and Type 2 diabetes in england, 1998-2013: a retrospective cohort study. *Diabetes Care*. 2018;41(9):1870–1877.
32. Sauder KA, Dabelea D, Bailey-Callahan R, Kanott Lambert S, Powell J, James R, et al. Targeting risk factors for type 2 diabetes in american indian youth: the tribal turning point pilot study. *Pediatr Obes*. 2018;13(5):321–329.
33. Zhong VW, Juhaeri J, Cole SR, Shay CM, Gordon-Larsen P, Kontopantelis E, et al. HbA(1C) variability and hypoglycemia hospitalization in adults with type 1 and type 2 diabetes: A nested case-control study. *J Diabetes Complicat*. 2018;32(2):203–209.
34. Lanes S, Brown JS, Haynes K, Pollack MF, Walker AM. Identifying health outcomes in healthcare databases. *Pharmacoepidemiol Drug Saf*. 2015;24(10):1009–1016.
35. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc*. 2016;23(e1):e20–e27.
36. Hsu DY, Dalal P, Sable KA, Voruganti N, Nardone B, West DP, et al. Validation of international classification of disease ninth revision codes for atopic dermatitis. *Allergy*. 2017;72(7):1091–1095.
37. Dizon MP, Yu AM, Singh RK, Wan J, Chren MM, Flohr C, et al. Systematic review of atopic dermatitis disease definition in studies using routinely collected health data. *Br J Dermatol*. 2018;178(6):1280–1287.
38. Liu B, Hadzi-Tosev M, Liu Y, Lucier KJ, Garg A, Li S, et al. Accuracy of international classification of diseases, 10th revision codes for identifying sepsis: a systematic review and meta-analysis. *Critical Care Explor*. 2022;4(11):e0788.
39. Gillmeyer KR, Lee MM, Link AP, Klings ES, Rinne ST, Wiener RS. Accuracy of algorithms to identify pulmonary arterial hypertension in administrative data: a systematic review. *Chest*. 2019;155(4):680–688.
40. Verheij RA, Curcin V, Delaney BC, McGilchrist MM. Possible Sources of Bias in Primary Care Electronic Health Record Data Use and Reuse. *J Med Internet Res*. 2018;20(5):e185.
41. Richesson RL, Sun J, Pathak J, Kho AN, Denny JC. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif Intell Med*. 2016;71:57–61.
42. Basile AO, Ritchie MD. Informatics and machine learning to define the phenotype. *Expert Rev Mol Diagn*. 2018;18(3):219–226.
43. Mo H, Thompson WK, Rasmussen LV, Pacheco JA, Jiang G, Kiefer R, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J Am Med Inform Assoc*. 2015;22(6):1220–1230.
44. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Ann Rev. Biomed Data Sci*. 2018;1:53–68.
45. Guralnik E. Utilization of electronic health records for chronic disease surveillance: a systematic literature review. *Cureus*. 2023;15(4):e37975.
46. Nichols GA, Schroeder EB, Karter AJ, Gregg EW, Desai J, Lawrence JM, et al. Trends in diabetes incidence among 7 million insured adults, 2006–2011: the supreme-dm project. *Am J Epidemiol*. 2014;181(1):32–39.
47. Lingren T, Thaker V, Brady C, Namjou B, Kennebeck S, Bickel J, et al. Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers. *Appl Clin Inform*. 2016;7(3):693–706.
48. Khurshid S, Keaney J, Ellinor PT, Lubitz SA. A simple and portable algorithm for identifying atrial fibrillation in the electronic medical record. *Am J Cardiol*. 2016;117(2):221–225.
49. Khalifa A, Obeid JS, Gregoski MJ, Rockey DC. Accurate identification of patients with cirrhosis and its complications in the electronic health record. *Dig Dis Sci*. 2023;68(6):2360–2369.
50. Dahiya M, Eboreime E, Hyde A, Rahman S, Sebastianski M, Carbonneau M, et al. International classification of diseases codes are useful in identifying cirrhosis in administrative databases. *Dig Dis Sci*. 2022;67(6):2107–2122.
51. Wilcox AB. Leveraging electronic health records for phenotyping. In: Payne PRO, Embi PJ, eds. *Translational Informatics: Realizing the Promise of Knowledge-Driven Healthcare*. London: Springer London; 2015:61–74.
52. Kho AN, Hayes MG, Rasmussen-Torvik L, Pacheco JA, Thompson WK, Armstrong LL, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J Am Med Inform Assoc*. 2012;19(2):212–218.
53. Culpepper WJ, Marrie RA, Langer-Gould A, Wallin MT, Campbell JD, Nelson LM, et al. Validation of an algorithm for identifying MS cases in administrative health claims datasets. *Neurology*. 2019;92(10) e1016-e28.
54. Shivade C, Raghavan P, Fosler-Lussier E, Embi PJ, Elhadad N, Johnson SB, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*. 2013;21(2):221–230.
55. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. *Jclinical epidemiology*. 2012;65(3) 343-49 e2.
56. Wang W, Jin YH, Liu M, He Q, Xu JY, Wang MQ, et al. Guidance of development, validation, and evaluation of algorithms for populating health status in observational studies of routinely collected data (DEVELOP-RCD). *Mil Med Res*. 2024;11(1):52.
57. Wong J, Horwitz MM, Zhou L, Toh S. Using machine learning to identify health outcomes from electronic health record data. *Current Epidemiol Rep*. 2018;5(4):331–342.
58. Bilezikian JP, Bandeira L, Khan A, Cusano NE. Hyperparathyroidism. *Lancet (London, England)*. 2018;391(10116):168–178.
59. Press DM, Siperstein AE, Berber E, Shin JJ, Metzger R, Monteiro R, et al. The prevalence of undiagnosed and unrecognized primary hyperparathyroidism: a population-based analysis from the electronic medical record. *Surgery*. 2013;154(6):1232–1237 discussion 37-8.
60. Somnay YR, Craven M, McCoy KL, Carty SE, Wang TS, Greenberg CC, et al. Improving diagnostic recognition of primary hyperparathyroidism with machine learning. *Surgery*. 2017;161(4):1113–1121.
61. Schlapbach LJ, Watson RS, Sorce LR, Argent AC, Menon K, Hall MW, et al. International consensus criteria for pediatric sepsis and septic shock. *Jama*. 2024.
62. Angus DC, Seymour CW, Coopersmith CM, Deutschman CS, Klompas M, Levy MM, et al. A framework for the development and interpretation of different sepsis definitions and clinical criteria. *Critical Care Med*. 2016;44(3):e113–e121.
63. Rhee C, Dantes RB, Epstein L, Klompas M. Using objective clinical data to track progress on preventing and treating sepsis: CDC's new 'Adult Sepsis Event' surveillance strategy. *BMJ Qual Saf*. 2019;28(4):305–309.
64. Liu S, Fu B, Wang W, Liu M, Sun X. Dynamic sepsis prediction for intensive care unit patients using xgboost-based model with novel time-dependent features. *IEEE J Biomed Health Inform*. 2022;26(8):4258–4269.
65. Taneja I, Reddy B, Damhorst G, Dave Zhao S, Hassan U, Price Z, et al. Combining biomarkers with emr data to identify patients in different phases of sepsis. *Sci Reports*. 2017;7(1):10800.
66. Jamshidi A, Pelletier JP, Martel-Pelletier J. Machine-learning-based patient-specific prediction models for knee osteoarthritis. *Nat Rev Rheumatol*. 2019;15(1):49–60.
67. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *Jama*. 2018;319(13):1317–1318.
68. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inf Assoc*. 2016;23(6):1046–1052.
69. Ostropolets A, Ryan P, Hripcsak G. Phenotyping in distributed data networks: selecting the right codes for the right patients. *AMIA Annual Symposium proceedings AMIA Symposium*; 2022:826–835.
70. Thayer DS, Mumtaz S, Elmessary MA, Scanlon I, Zinnurov A, Coldea AI, et al. Creating a next-generation phenotype library: the health data research UK Phenotype Library. *JAMIA Open*. 2024;7(2) ooae049.
71. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *Jallergy Clin Immunol*. 2020;145(2):463–469.
72. Guevara M, Chen S, Thomas S, Chaunzwa TL, Franco I, Kann BH, et al. Large language models to identify social determinants of health in electronic health records. *NPJ Digit Med*. 2024;7(1):6.
73. Dong H, Falis M, Whiteley W, Alex B, Matterson J, Ji S, et al. Automated clinical coding: what, why, and where we are? *NPJ Digit Med*. 2022;5(1):159.
74. Tian D, Chen W, Xu D, Xu L, Xu G, Guo Y, et al. A review of traditional Chinese medicine diagnosis using machine learning: Inspection, auscultation-olfaction, inquiry, and palpation. *Comput Biol Med*. 2024;170:108074.