Published in final edited form as:
Pharmacoepidemiol Drug Saf. 2023 January ; 32(1): 1–8. doi:10.1002/pds.5537.

Core Concepts in Pharmacoepidemiology: Validation of Health Outcomes of Interest Within Real-World Healthcare Databases

Erica J Weinstein[1,2], **Mary Elizabeth Ritchey**[3,4], **Vincent Lo Re III**[1,2]

[1]Division of Infectious Diseases, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2]Center for Pharmacoepidemiology Research and Training, Center for Clinical Epidemiology and Biostatistics, and Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3]Med Tech Epi, LLC, Philadelphia, PA, USA

[4]Center for Pharmacoepidemiology and Treatment Science, Rutgers University, New Brunswick, New Jersey, USA

Abstract

Real-world healthcare data, including administrative and electronic medical record databases, provide a rich source of data for the conduct of pharmacoepidemiologic studies but carry the potential for misclassification of health outcomes of interest (HOIs). Validation studies are important ways to quantify the degree of error associated with case-identifying algorithms for HOIs and are crucial for interpreting study findings within real-world data. This review provides a rationale, framework, and step-by-step approach to validating case-identifying algorithms for HOIs within healthcare databases. Key steps in validating a case-identifying algorithm within a healthcare database include: 1) selecting the appropriate health outcome; 2) determining the reference standard against which to validate the algorithm; 3) developing the algorithm using diagnosis codes, diagnostic tests or their results, procedures, drug therapies, patient-reported symptoms or diagnoses, or some combinations of these parameters; 4) selection of patients and sample sizes for validation; 5) collecting data to confirm the HOI; 6) confirming the HOI; and 7) assessing the algorithm's performance. Additional strategies for algorithm refinement and methods to correct for bias due to misclassification of outcomes are discussed. The review concludes by discussing factors affecting the transportability of case-identifying algorithms and the need for ongoing validation as data elements within healthcare databases, such as diagnosis codes, change over time or new variables, such as patient-generated health data, are included in these data sources.

Keywords

Validation; methods; electronic health records; database; algorithm; misclassification

Correspondence to: Vincent Lo Re III, MD, MSCE, Division of Infectious Diseases, Department of Medicine, Division of Epidemiology, Department of Biostatistics, Epidemiology, and Informatics, Center for Clinical Epidemiology and Biostatistics, 836 Blockley Hall, 423 Guardian Drive, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104-6021, vincentl@pennmedicine.upenn.edu, Phone: (215) 573-5964; Fax: (215) 573-5315.

Introduction

Case-identifying algorithms are typically employed by pharmacoepidemiology researchers to identify health outcomes of interest (HOIs) within real-world electronic healthcare data. Based on knowledge of clinical workflow and the data available, such algorithms may consist of a combination of one or more diagnosis codes, laboratory test codes or results, procedural codes, drug therapies, or patient-reported symptoms or diagnoses.[1] However, case-identifying algorithms may not always accurately identify the HOI, and this could lead to misclassification of these events in future studies. In analyses evaluating associations between medications and endpoints of interest, outcome misclassification due to poor validity of an algorithm could lead to biased estimates of treatment effect.

Although many case-identifying algorithms can be found in the literature, information regarding the performance of such algorithms is often lacking.[2] Researchers should assess the accuracy of an algorithm that they have developed against a reference standard to ensure that it accurately ascertains the HOI.[3] Validation of the algorithm can help to quantify the potential degree of outcome misclassification in a study and is critical to drawing valid conclusions from analyses using them within a given database.[4]

Despite the importance of studies validating the accuracy of case-identifying algorithms in healthcare databases, there is very little methodologic guidance for their design and conduct.[5] This Core Concepts in Pharmacoepidemiology review provides an overall framework for the development and validation of case-identifying algorithms for HOIs within healthcare databases (Figure 1). We will first discuss considerations for selecting appropriate HOIs to study. We will then present options for reference standards against which to validate algorithms. Next, we will review approaches to developing algorithms to identify HOIs within databases and consider selection methods and sample sizes for validation. We will then examine ways to confirm the HOI and determine algorithm performance. Statistical approaches for handling bias due to outcome misclassification after the algorithm's performance has been assessed will be reviewed. We will conclude by discussing the transportability of algorithms to different databases and future research directions, including the use of patient-generated data.

Approach to Validating HOIs Within Real-World Healthcare Databases

Select an Appropriate HOI—When selecting a particular HOI to evaluate within a database, investigators should consider the disease's natural history and clinical characteristics. To reduce the likelihood of misclassifying HOIs, investigators should focus on ascertaining severe, acute events that prompt presentation for medical care and for which the date of onset is clear.[1] Indolent conditions or diseases with a more gradual onset may be difficult to ascertain, and their date of onset could be inaccurate. For example, if a researcher wanted to evaluate end-stage liver disease as an endpoint in a cohort study, selecting cirrhosis as the HOI to ascertain would not be ideal, because this condition is often initially clinically silent and would not have a definitive date of onset. Instead, hepatic decompensation would be a more appropriate outcome to ascertain end-stage liver disease, since this condition is characterized by clinically apparent complications of decompensated cirrhosis, such as ascites, spontaneous bacterial peritonitis, variceal hemorrhage, or hepatic encephalopathy, which are overt diagnoses that prompt patients to seek medical care.[6] HOIs selected for case-identifying algorithms should therefore be severe enough to cause clinical presentation and have a well-defined date of onset to minimize potential for outcome misclassification.

Determine the Reference Standard Against Which to Validate the Algorithm—After selecting the HOI, the next step in validating an HOI is to determine the reference standard against which to measure algorithm performance. The reference standard represents the best available method to determine the presence or absence of the target HOI. Possible reference standards include medical records, disease registries, or survey results about the presence of the disease completed by healthcare providers or patients.[3,7] The reference standard selected will depend on the type of HOI under study, availability of disease registries, and access to medical records, providers, or patients.

In some instances, the reference standard may be imperfect for classifying the target HOI. Use of an imperfect reference standard can lead to bias in estimates of the accuracy of a case-identifying algorithm. Several approaches have been suggested to address such situations.[8] Researchers could utilize an expert panel of clinicians who have experience diagnosing and managing the disease.[8] Alternatively, statistical methods could be used to impute missing data on the reference standard.[9] Finally, estimates of accuracy could be corrected based on external evidence about the degree of imperfection of the reference standard.[10]

Develop the Case-Identifying Algorithm—Researchers should review the published literature to assess whether prior studies have developed and validated algorithms to identify the HOI in the same or a different data source. If an algorithm has been developed, researchers should consider whether the existing algorithm could be applicable to their research question and database. Even when not directly applied, previously developed algorithms can serve as a valuable starting point for introducing refinements to enhance algorithm accuracy.

Prior to constructing an algorithm, consultation with clinicians experienced in diagnosing and treating the HOI is imperative. A thorough understanding of the condition, including its signs/symptoms, typical presentation, methods and criteria for diagnosis, and approaches to treatment, is important. This clinical information is crucial for considering which diagnoses, procedures, laboratory results, and/or therapies to include within a case-identifying algorithm.

Consultation with individuals who have expertise in the database in which the algorithm will be applied is also important. Such collaborations can aid understanding of the variables available, methods of data capture, clinical practice patterns, and reimbursement policies, all of which can change over time.[11] Collaboration with database experts ensures appropriate use of data elements during the time period of interest.

Researchers may choose to develop a case-identifying algorithm based on one or more administrative codes within a structured classification system, such as diagnosis or procedure codes, or based on diagnostic tests (or their results), drug therapies, or patient-reported symptoms or diagnoses.[7] Attention should be paid to the description and specificity of codes used. Some conditions or procedures may not have explicitly defined codes, and their use within an algorithm could compromise its validity.

When using diagnosis codes, the study team will need to determine if ambulatory (outpatient) diagnoses, hospital diagnoses, or both (if available) should be included in the algorithm. If outpatient diagnoses are included, consideration should be given to the number and time between diagnoses required to meet the algorithm. Outpatient diagnoses often reflect the clinician's understanding of the reason for care at the time of a patient's presentation, but the clinician's diagnostic suspicion may change as more information becomes available. Requiring two or more outpatient diagnoses within an algorithm could help to exclude events that are suspected (e.g., "rule-out diagnoses") but not confirmed at follow-up visits.[1]

In contrast, diagnosis codes associated with hospital care may more accurately reflect a patient's clinical condition than outpatient diagnoses, since these diagnoses are typically recorded because of diagnostic procedures or tests during the hospitalization. Moreover, hospital diagnoses often undergo internal hospital oversight, and depending on the database, may reflect review by health insurers to ensure proper reimbursement. If hospital diagnosis codes are included in an algorithm, researchers will need to consider whether only discharge diagnoses (final reason for hospitalization) will be used or if other in-hospital diagnoses (e.g., by specialists) will be considered for inclusion. If discharge diagnoses are employed, investigators will need to consider whether only principal/primary diagnoses will be included or if secondary/supporting diagnoses will also be eligible for inclusion. The choice will depend upon whether the investigators seek to enhance capture of potential events, but at the cost of algorithm positive predictive value (PPV; proportion of persons identified by the algorithm who are confirmed to have the HOI) or specificity (proportion of persons who do not have the HOI that the algorithm correctly identified as negative; Box 1).[12,13]

There are many examples of approaches to developing case-identifying algorithms in the published literature. To identify hepatic decompensation events in electronic medical record data, one team of researchers constructed an algorithm based on ≥1 principal or contributory hospital discharge diagnosis or ≥2 outpatient diagnoses of ascites, spontaneous bacterial peritonitis, or variceal hemorrhage.[6] They also explored algorithms based on laboratory abnormalities suggestive of hepatic decompensation (total bilirubin ≥5.0 g/dL; albumin ≤2.0 g/dL; international normalized ratio ≥1.7 off oral anticoagulation).[6] As another example, an algorithm to ascertain hospitalizations for serious infections identified patients with a primary or secondary hospital discharge diagnosis of serious infection, and within 7 days prior to admission, required an outpatient or emergency department infection diagnosis or outpatient prescription for an antimicrobial drug.[14,15] The rationale for this algorithm was to enhance the likelihood of identifying serious infections that developed in the outpatient setting and precipitated hospital admission.

When using electronic medical record databases, one must consider whether “unstructured” or “narrative” data, such as free-form text within clinician or operative notes, should be included within an algorithm. Over the last decade there has been considerable growth in natural language processing (NLP), an automated method of processing of text based on sets of rules, which permits identification of diagnoses or procedures within clinical notes.[16] Liao et al. demonstrated that by combining diagnosis codes and narrative data in a rheumatoid arthritis algorithm, patients with this disease could be identified with a PPV=94%, compared to a PPV=88% when diagnosis codes alone were used.[17] Thus, inclusion of unstructured text in medical records captured by NLP could help to improve algorithm performance.

Select Persons for Validation—When selecting a sample of cases for validation, one should ensure that sampling is not somehow related to the HOI. Additionally, it is important to recognize that PPV estimates depend on the prevalence of the outcome in the population studied.[13] Thus, researchers should ensure that the prevalence of the HOI in the sampled population is the same as that of the target population. Depending on the HOI, algorithm, and data source, it might also be important to ensure that relevant sub-diagnoses, geographic regions, or type of care setting are represented in sampled cases (see Box 2 for example).[15]

Validation studies often do not confirm the absence of the HOI among persons identified as not having the condition via the algorithm, hence assuming perfect specificity. To address this issue, researchers could take a random sample of persons who were not identified by the algorithm over the time-period of interest and determine what proportion had the HOI based on the reference standard, allowing calculation of the algorithm’s specificity and negative predictive value (NPV; proportion of persons identified by the algorithm as not having the outcome who are confirmed not to have the HOI; Box 1). Calculating algorithm specificity and NPV is particularly important when the disease of interest is common.[1] Alternatively, conditional or risk sampling, such as randomly sampling people who meet some, but not all, of the algorithm requirements, can be performed to enrich for individuals not identified as cases, but who are more likely to be misclassified.[18]

When considering the sample size of a validation study, one must ensure that the sample is large enough to provide the desired precision for the estimates of algorithm accuracy. Validation studies are often not designed to estimate all performance characteristics. PPV and sensitivity (proportion of persons with the HOI that the algorithm correctly identified; Box 1) are the most reported measures.[3] For rare HOIs, specificity and NPV are expected to be close to 100%. The number of validated cases determines the precision of the estimates of algorithm accuracy. Typically, the number of persons sampled is based on the desired width of the 95% confidence interval (CI) around the anticipated estimate of the performance characteristic of interest. For example, if researchers seek a PPV ≥80% with a 95% CI of +/−10%, then approximately 75 patients who meet the algorithm would be required to achieve the desired precision.

Collect Relevant Data to Confirm the HOI—The data needed to confirm the HOI should be reviewed with clinicians who have expertise in its diagnosis and management, preferably within the healthcare system represented by the data source. If the reference standard involves confirmation of the HOI by clinicians or patients, questionnaires should be designed in collaboration with qualitative researchers to ensure that the survey is valid and reliable. If the validation involves linkage to a medical record or disease registry, a variable present in both the study and validation data sources must be identified to permit linkage so that the HOI can be confirmed.

When medical record review is required for validation, a structured abstraction form should be developed to collect the necessary clinical data from relevant chart components to allow confirmation of the diagnosis. The period for record review relative to the date the potential HOI is identified by the algorithm should be specified (e.g., +/− 90 days). Researchers may consider extracting hospital discharge summaries (for confirmation of hospital events) or specialist consultation notes from hospital or ambulatory settings and include these along with the structured forms to facilitate HOI determination.

Employing dedicated data abstractors is important, particularly when HOI confirmation should be blinded to information such as treatment status in validating endpoints for safety assessments. Data abstractors should receive training on the components of the medical record that should be reviewed, time periods over which to review records, and variables to abstract. Insufficient training of data abstractors can lead to improper abstraction of data. To ensure the validity of data abstraction, a second trained individual should separately review the medical records of a sample of patients whose charts have already undergone abstraction, and agreement between the individuals should be assessed (e.g., via *kappa* statistic).[19]

For the study validating the hepatic decompensation algorithm, the investigators abstracted relevant data onto structured forms that collected information from: 1) abdominal imaging reports (presence/quantity of ascites); 2) laboratory results (peritoneal fluid cell count/differential and culture); 3) endoscopic reports (presence of varices, signs of active variceal bleeding); and 4) hepatologist notes (chronic liver disease etiology, diagnosis of variceal bleeding).[6]

Confirm the HOI—Confirming HOIs identified by case-identifying algorithms can be undertaken via either internal or external validation. The decision to use internal versus external validation depends on the HOI, available reference standard, and planned application of the algorithm.[20,21]

Internal validation assesses whether the algorithm accurately identifies the diagnosis recorded by the healthcare provider.[21] This type of validation may be sufficient when confirmation of a primary care-based diagnosis is required for research purposes. An example is in assessing the performance of a diagnostic coding-based algorithm for a rash. The ability of the algorithm to identify a rash could be confirmed by documentation in the physical exam section of the medical record or by response to a questionnaire for the managing clinician.

In contrast, external validation assesses the performance of an algorithm compared to one of the following reference standards: 1) clinician adjudication (e.g., single clinician, multiple clinicians, or panel of experts) of the HOI by applying formal diagnostic criteria to abstracted medical record data, 2) linkage to a disease registry that records the diagnosis, or 3) clinician response to a questionnaire confirming the diagnosis.[21] This more classic method of validation is typically employed for studies required by regulators for newly marketed treatments (e.g., post-authorization safety study) or for complex health outcomes, such as hepatic decompensation.[6]

For external validation studies, formal adjudication of the HOI by clinical experts is recommended. A formal definition for the HOI should be developed based on clinical criteria used to diagnose the condition in practice and current management guidelines. Structured abstraction forms, discharge summaries, and other relevant clinical information can be reviewed by adjudicators to determine if criteria for the outcome definition(s) are met. Having clinician adjudicators arbitrate the outcome using a formal definition permits a systematic approach to case arbitration. In circumstances when definite criteria to meet the definition for the HOI might not be available (e.g., symptoms needed to confirm a diagnosis might not be recorded in the medical record because the clinician was too busy to record them; laboratory results might not have been obtained due to patient refusal), it can be beneficial to create definitions for probable (likely true, but not certain) or possible (could be true) events. The inclusion of probable or possible events allows for evaluation of the accuracy of the algorithm across a spectrum of clinical case presentations. Using too stringent of a case definition could result in exclusion of true cases and inadvertently lead to poor algorithm performance (e.g., low PPV).

The accuracy of the adjudication process can be enhanced by having at least two clinical experts independently adjudicate the outcome. In instances of disagreement, a third adjudicator can be enlisted to arbitrate the event. Such an approach was employed in the study evaluating the validity of the hepatic decompensation algorithms.[6] In situations where an adjudication panel of experts is used, an odd number of adjudicators should serve on the panel to ensure a majority decision on classification of the event.

Assess the Algorithm Performance —The performance of a case-identifying algorithm refers to the accuracy with which it identifies the HOI compared to the reference standard. Analogous to evaluating a diagnostic test, parameters such as PPV, NPV, sensitivity, and specificity - and their precision (i.e., 95% CI) - are determined (Box 1).[12,13] These performance characteristics quantify the validity of the algorithm. Sensitivity and specificity can vary across different settings and populations but are not influenced meaningfully by prevalence,[12] whereas PPV and NPV are dependent on sensitivity and specificity as well as the prevalence of the outcome in the study population.[13]

The study team must consider the intended use of a particular algorithm and prioritize the accuracy measure most important to the target study. In many pharmacoepidemiology studies in which the objective is to compare the safety of medical therapies and where the HOIs are expected to occur in a small proportion of patients, the PPV is often the prioritized algorithm characteristic.[3,22] In these studies, ensuring that only persons who truly develop the HOI are identified in the study is important to minimize outcome misclassification and promote confidence that outcomes identified by the algorithm are true events.[23]

Algorithms for rare outcomes typically have a high NPV, but are prone to lower PPV; thus, false-positive errors are the focus. Specificity is important when classifying outcomes, and imperfect specificity will bias the relative effect estimate even if the sensitivity is 100%. Furthermore, a high specificity is often important but not alone sufficient to ensure a high PPV.[23] When the goal of a study is to estimate incidence, sensitivity should be prioritized. Thus, researchers should consider how the algorithm will be employed in future pharmacoepidemiology studies.

Considerations in the Interpretation and Application of Algorithm Performance

There has been little literature on the optimal cut-offs of algorithm performance characteristics. In general, validation studies typically seek algorithm PPVs exceeding 80% to ensure a minimal degree of misclassification,[3] but the acceptable threshold of algorithm PPV remains unclear. There is also a need for guidance on acceptable cut-offs for sensitivity, specificity, and NPV, depending on an algorithm's use. Due to this limited guidance, acceptable thresholds of the parameter(s) of interest for a validation study should be determined and justified by the research team prior to conduct. Although not typically implemented as part of the validation of HOI algorithms, a data-driven determination of algorithm performance could be achieved by borrowing the methodology of diagnostic test development and identifying an optimal cut-off value using receiver-operating characteristic curve analysis.[24]

If an algorithm's accuracy is found to be below a pre-specified threshold, refinement of the algorithm may be helpful. For example, if an algorithm consisting of diagnosis codes alone has poor accuracy, evaluation of the accuracy of individual diagnosis codes and removal of poorly performing codes may be beneficial. Alternatively, statistical methods can be used to correct for imperfect algorithm accuracy explicitly or to quantify the bias due to algorithm misclassification. Several likelihood-based methods exist to reduce bias arising from algorithms subject to misclassification.[25,26] Electronic medical record-derived probabilistic phenotypes that return a predicted probability of being an outcome can be used within subsequent studies and minimize bias in estimated association parameters.[27–29] Quantitative bias analysis can be implemented to assess the magnitude of bias of an exposure-outcome association caused by outcome misclassification by an algorithm.[30,31] Quantifying the potential bias from imperfect algorithm accuracy and estimating its impact on effect measures can provide reassurance about the soundness of a study's results.

Transportability of a Case-Identifying Algorithm to Different Databases

The performance of a case-identifying algorithm can vary across databases, populations, and time periods. This can be due to differences in the prevalence of the HOI across the settings, differences in comorbidities between the populations, or changes in diagnostic approaches over time.[1] When assessing whether an algorithm might be transportable to a different database, pharmacoepidemiologists should initially consider if the algorithm's components are available in the different database and whether the prevalence of the HOI in that database might be similar to that of the data source in which the algorithm was originally validated.[32] Changes in coding over time should also be considered. One notable contemporary change in diagnosis coding is the transition from ICD-9 to ICD-10 systems, which took place at different times in many countries. Similarly, ICD-11 has been developed and will be employed in the future. Mapping ICD-9 to ICD-10 diagnoses, and eventually future iterations of these codes, might be valuable when developing algorithms.

To assess the transportability of a case-identifying algorithm in a new setting, whether it be a new coding system, database, or study population, additional validation studies are needed within the new setting. Such studies will quantify the extent of misclassification and reveal whether notable variation in accuracy exists when the algorithm is transported across settings, populations, coding systems, or time periods.

Newer Approaches to Enhance the Validity of Case-Identifying Algorithms

As new structured and unstructured clinically generated real-world data emerge, the field of pharmacoepidemiology must consider approaches to increase quality and standardization processes to make these data amenable to pharmacoepidemiologic research. Machine learning and NLP have become increasingly common to improve case-identifying accuracy, by incorporating unstructured or qualitative data elements from electronic medical record data to identify patients with HOIs.[16,33] Case-identifying methods incorporating free-text data from radiology reports, operative reports, pathology reports, or other sources offer increased granularity of clinical data for pharmacoepidemiology research. Additionally, the availability of patient-generated data from home monitoring devices (e.g., pulse oximeters, pacemakers, glucometers), wearables, mobile applications, social media, and other sources is increasing, and incorporating these data sources into case-identifying algorithms, and studies validating their use, will be important in the future.[34]

Conclusions

Validation of case-identifying algorithms to determine the degree of outcome misclassification is important to drawing valid inferences from studies using real-world healthcare databases. To develop an accurate case-identifying algorithm for an HOI, one must have a thorough understanding of the structure and nuances of the database as well as the patterns of clinical care relative to the natural history, diagnosis, and management of the HOI in the study population. To assess an algorithm’s accuracy, researchers must conduct a thorough evaluation of its performance, following the steps outlined above. An algorithm’s accuracy can vary across databases, time periods, populations, and settings. Researchers should determine the transportability of an algorithm to a different database with a validation study when considering whether it can be used in that setting.

Funding Statement:

The review was supported, in part, by a research grant from the National Institutes of Health/National Institute of Allergy and Infectious Diseases (T32-AI-055435).

REFERENCES

1. Lanes S, Brown JS, Haynes K, Pollack MF, Walker AM. Identifying health outcomes in healthcare databases. Pharmacoepidemiol Drug Saf. 2015;24(10):1009–1016. [PubMed: 26282185]

2. van Walraven C, Bennett C, Forster AJ. Administrative database research infrequently used validated diagnostic or procedural codes. J Clin Epidemiol. 2011;64(10):1054–1059. [PubMed: 21474278]
3. Benchimol EI, Manuel DG, To T, Griffiths AM, Rabeneck L, Guttmann A. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. J Clin Epidemiol. 2011;64(8):821–829. [PubMed: 21194889]
4. Ehrenstein V, Petersen I, Smeeth L, et al. Helping everyone do better: a call for validation studies of routinely recorded health data. Clin Epidemiol. 2016;8:49–51. [PubMed: 27110139]
5. Lash TL, Olshan AF. Epidemiology announces the “Validation Study” submission category. Epidemiology. 2016;27(5):613–614. [PubMed: 27388372]
6. Lo Re V 3rd, Lim JK, Goetz MB, et al. Validity of diagnostic codes and liver-related laboratory abnormalities to identify hepatic decompensation events in the Veterans Aging Cohort Study. Pharmacoepidemiol Drug Saf. 2011;20(7):689–699. [PubMed: 21626605]
7. Nissen F, Quint JK, Morales DR, Douglas IJ. How to validate a diagnosis recorded in electronic health records. Breathe (Sheff). 2019;15(1):64–68. [PubMed: 30838062]
8. Reitsma JB, Rutjes AW, Khan KS, Coomarasamy A, Bossuyt PM. A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. J Clin Epidemiol. 2009;62(8):797–806. [PubMed: 19447581]
9. Harel O, Zhou XH. Multiple imputation for correcting verification bias. Stat Med. 2006;25(22):3769–3786. [PubMed: 16435337]
10. Hadgu A, Dendukuri N, Hilden J. Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. Epidemiology. 2005;16(5):604–612. [PubMed: 16135935]
11. Lo Re V 3rd. Validation of health outcomes of interest in healthcare databases. In: Girman CJ, Ritchey ME, eds. Pragmatic randomized clinical trials: Using primary data collection and electronic health records. Cambridge, MA: Elsevier Science; 2021.
12. Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. BMJ. 1994;308(6943):1552. [PubMed: 8019315]
13. Altman DG, Bland JM. Diagnostic tests 2: Predictive values. BMJ. 1994;309(6947):102. [PubMed: 8038641]
14. Saine ME, Gizaw M, Carbonari DM, et al. Validity of diagnostic codes to identify hospitalizations for infections among patients treated with oral anti-diabetic drugs. Pharmacoepidemiol Drug Saf. 2018;27(10):1147–1150. [PubMed: 29250905]
15. Lo Re V, 3rd, Carbonari DM, Jacob J, et al. Validity of ICD-10-CM diagnoses to identify hospitalizations for serious infections among patients treated with biologic therapies. Pharmacoepidemiol Drug Saf. 2021;30(7):899–909. [PubMed: 33885214]
16. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol. 2020;145(2):463–469. [PubMed: 31883846]
17. Liao KP, Cai T, Gainer V, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care Res (Hoboken). 2010;62(8):1120–1127. [PubMed: 20235204]
18. Gravel CA, Farrell PJ, Krewski D. Conditional validation sampling for consistent risk estimation with binary outcome data subject to misclassification. Pharmacoepidemiol Drug Saf. 2019;28(2):227–233. [PubMed: 30746841]
19. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–174. [PubMed: 843571]
20. Herrett E, Thomas SL, Schoonen WM, Smeeth L, Hall AJ. Validation and validity of diagnoses in the General Practice Research Database: a systematic review. Br J Clin Pharmacol. 2010;69(1):4–14. [PubMed: 20078607]
21. Nicholson A, Tate AR, Koeling R, Cassell JA. What does validation of cases in electronic record databases mean? The potential contribution of free text. Pharmacoepidemiol Drug Saf. 2011;20(3):321–324. [PubMed: 21351316]
22. Hall GC, Lanes S, Bollaerts K, Zhou X, Ferreira G, Gini R. Outcome misclassification: Impact, usual practice in pharmacoepidemiology database studies and an online aid to correct biased

estimates of risk ratio or cumulative incidence. Pharmacoepidemiol Drug Saf. 2020;29(11):1450–1455. [PubMed: 32860317]

23. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. J Clin Epidemiol. 2012;65(3):343-349 e342. [PubMed: 22197520]
24. Altman DG, Bland JM. Diagnostic tests 3: receiver operating characteristic plots. BMJ. 1994;309(6948):188. [PubMed: 8044101]
25. Magder LS, Hughes JP. Logistic regression when the outcome is measured with uncertainty. Am J Epidemiol. 1997;146(2):195–203. [PubMed: 9230782]
26. Lyles RH, Tang L, Superak HM, et al. Validation data-based adjustments for outcome misclassification in logistic regression: an illustration. Epidemiology. 2011;22(4):589–597. [PubMed: 21487295]
27. Cai B, Hennessy S, Lo Re V 3rd, Small DS. Epidemiologic research using probabilistic outcome definitions. Pharmacoepidemiol Drug Saf. 2015;24(1):19–26. [PubMed: 25257346]
28. Sinnott JA, Dai W, Liao KP, et al. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. Hum Genet. 2014;133(11):1369–1382. [PubMed: 25062868]
29. Hubbard RA, Tong J, Duan R, Chen Y. Reducing bias due to outcome misclassification for epidemiologic studies using EHR-derived probabilistic phenotypes. Epidemiology. 2020;31(4):542–550. [PubMed: 32282406]
30. Lash TL, Fox MP, MacLehose RF, Maldonado G, McCandless LC, Greenland S. Good practices for quantitative bias analysis. Int J Epidemiol. 2014;43(6):1969–1985. [PubMed: 25080530]
31. 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(12):1664–1674. [PubMed: 31365086]
32. Carroll RJ, Thompson WK, Eyler AE, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc. 2012;19(e1):e162–169. [PubMed: 22374935]
33. Majnaric LT, Babic F, O'Sullivan S, Holzinger A. AI and big data in healthcare: Towards a more comprehensive research framework for multimorbidity. J Clin Med. 2021;10(4).
34. Bourke A, Dixon WG, Roddam A, et al. Incorporating patient generated health data into pharmacoepidemiological research. Pharmacoepidemiol Drug Saf. 2020;29(12):1540–1549. [PubMed: 33146896]

Box 1.

Glossary of terms.

Term	Definition
Health outcome of interest (HOI)	A health state or condition of an individual, group of people, or population. Examples include hepatic decompensation, rash, or mortality.
Algorithm	Defined set of parameters used to classify the HOI. The simplest algorithm is a single criterion, such as a diagnosis code, to identify the HOI. The algorithm classifies anyone in the study population whose record includes the code during the specified time window as having the condition, and those without the code as not having the condition.
Validation study	Cases identified by the algorithm (or not) are compared with a reference standard and provide estimates on algorithm performance.
Misclassification	A systematic error in the way an individual is assigned to a different category than the one to which they should be assigned. In this context, the incorrect classification of an algorithm with respect to an individual's HOI due to an error in determination.
Sensitivity	Proportion of true positives correctly identified by the algorithm.
Specificity	Proportion of true negatives correctly identified by the algorithm.
Positive predictive value (PPV)	Proportion of persons identified by the algorithm who are confirmed to have the HOI.
Negative predictive value (NPV)	Proportion of persons not identified by the algorithm who are confirmed not to have the HOI.

Box 2.

Example illustrating sampling approach for validation of a health outcome of interest (hospitalization for serious infection) that accounts for sub-diagnoses.

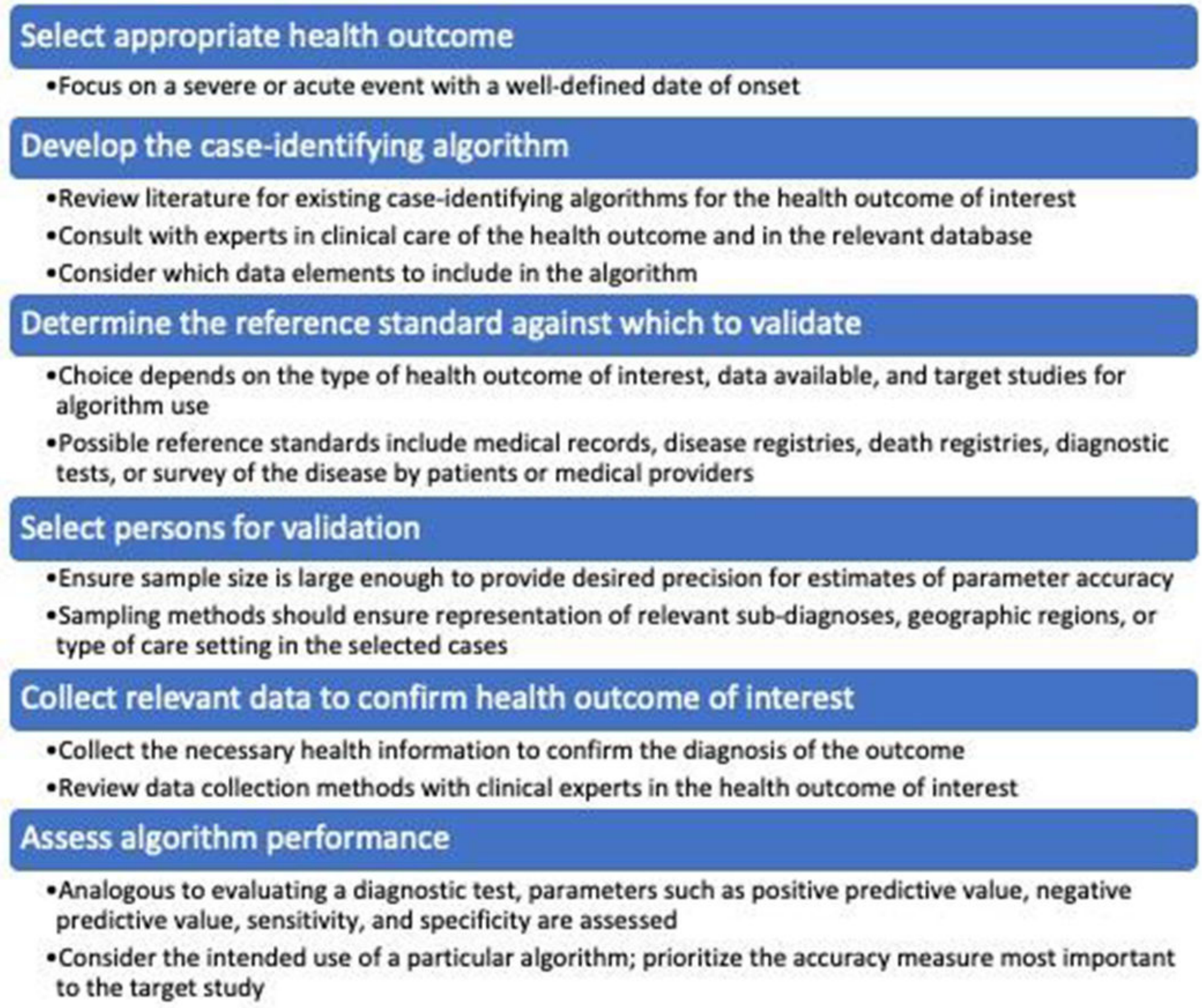
In a recent study examining the positive predictive value (PPV) of an International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM)-based algorithm to identify a hospitalization for a serious infection (defined by acute meningitis, acute osteomyelitis, bacteremia, gastrointestinal infection, acute pyelonephritis, pneumonia, or skin/soft tissue infection) within the United States Food and Drug Administration's Sentinel Distributed Database, the study team sought to ensure adequate samples of each type of serious infection for validation.[15] The study team noted that the prevalence of hospitalizations for the types of serious infections of interest varied widely, ranging from 5 hospitalizations/10,000 health plan members for acute meningitis to 207 hospitalizations/10,000 health plan members for bacteremia.

Random sampling without regard for the prevalence of the individual infection types would have yielded intolerably imprecise or incalculable PPV estimates for rare infections (i.e., acute meningitis, pyelonephritis, osteomyelitis). Consequently, the PPV of the overall ICD-10-CM-based algorithm for hospitalization for serious infection was determined based on a sampling strategy and weighting that accounted for the prevalence of each of the seven types of serious infections. Rare infections were over-sampled and common infections were under-sampled relative to their proportion in the source population. Once the target number of cases of each type of serious infection was determined based on its prevalence, stratified sampling within the type of infection was performed according to whether the participating Data Partner was a claims insurer or integrated health system to be reflective of the overall make-up of health plan members within the Sentinel Distributed Database. The investigators then corrected for the relative over- and under-sampling of the specific types of serious infections by weighting their PPV estimate by the prevalence of each infection type, down-weighing the PPVs of low prevalence infections and up-weighting the PPVs of the high prevalence infections.

In this way, the researchers accounted for the prevalence of the type of serious infection in the population in which the algorithm was validated and the population to which it was being applied.

Key Points

- There is the potential for misclassification of health outcomes of interest in pharmacoepidemiologic studies utilizing real-world healthcare databases.
- Key steps in developing and validating a case-identifying algorithm for a health outcome of interest within a healthcare database include: 1) selecting the appropriate health outcome; 2) determining the reference standard against which to validate the algorithm; 3) developing the algorithm using diagnosis codes, diagnostic tests or their results, procedures, drug therapies, patient-reported symptoms or diagnoses, or some combinations of these parameters; 4) selection of patients and sample sizes for validation; 5) collection of data to confirm the outcome; 6) confirmation of the outcome; and 7) assessment of the algorithm's performance.
- The accuracy of case-identifying algorithms can vary across time periods, populations, and health-care settings; consequently, it is imperative to re-assess the accuracy of the algorithm after a change in clinical practice or when transporting the algorithm to a new database.



Select appropriate health outcome

- Focus on a severe or acute event with a well-defined date of onset

Develop the case-identifying algorithm

- Review literature for existing case-identifying algorithms for the health outcome of interest
- Consult with experts in clinical care of the health outcome and in the relevant database
- Consider which data elements to include in the algorithm

Determine the reference standard against which to validate

- Choice depends on the type of health outcome of interest, data available, and target studies for algorithm use
- Possible reference standards include medical records, disease registries, death registries, diagnostic tests, or survey of the disease by patients or medical providers

Select persons for validation

- Ensure sample size is large enough to provide desired precision for estimates of parameter accuracy
- Sampling methods should ensure representation of relevant sub-diagnoses, geographic regions, or type of care setting in the selected cases

Collect relevant data to confirm health outcome of interest

- Collect the necessary health information to confirm the diagnosis of the outcome
- Review data collection methods with clinical experts in the health outcome of interest

Assess algorithm performance

- Analogous to evaluating a diagnostic test, parameters such as positive predictive value, negative predictive value, sensitivity, and specificity are assessed
- Consider the intended use of a particular algorithm; prioritize the accuracy measure most important to the target study

Figure 1:

Framework for validating health outcomes of interest within electronic healthcare data.