BMC Medical Informatics and Decision Making

RESEARCH Open Access

The validity of electronic health data for measuring smoking status: a systematic review and meta-analysis

Md Ashiqul Haque[1], Muditha Lakmali Bodawatte Gedara[1], Nathan Nickel[1], Maxime Turgeon[2] and Lisa M. Lix[1*]

Abstract

Background Smoking is a risk factor for many chronic diseases. Multiple smoking status ascertainment algorithms have been developed for population-based electronic health databases such as administrative databases and electronic medical records (EMRs). Evidence syntheses of algorithm validation studies have often focused on chronic diseases rather than risk factors. We conducted a systematic review and meta-analysis of smoking status ascertainment algorithms to describe the characteristics and validity of these algorithms.

Methods The Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines were followed. We searched articles published from 1990 to 2022 in EMBASE, MEDLINE, Scopus, and Web of Science with key terms such as validity, administrative data, electronic health records, smoking, and tobacco use. The extracted information, including article characteristics, algorithm characteristics, and validity measures, was descriptively analyzed. Sources of heterogeneity in validity measures were estimated using a meta-regression model. Risk of bias (ROB) in the reviewed articles was assessed using the Quality Assessment of Diagnostic Accuracy Studies-2 tool.

Results The initial search yielded 2086 articles; 57 were selected for review and 116 algorithms were identified. Almost three-quarters (71.6%) of algorithms were based on EMR data. The algorithms were primarily constructed using diagnosis codes for smoking-related conditions, although prescription medication codes for smoking treatments were also adopted. About half of the algorithms were developed using machine-learning models. The pooled estimates of positive predictive value, sensitivity, and specificity were 0.843, 0.672, and 0.918 respectively. Algorithm sensitivity and specificity were highly variable and ranged from 3 to 100% and 36 to 100%, respectively. Model-based algorithms had significantly greater sensitivity ($p = 0.006$) than rule-based algorithms. Algorithms for EMR data had higher sensitivity than algorithms for administrative data ($p = 0.001$). The ROB was low in most of the articles (76.3%) that underwent the assessment.

Conclusions Multiple algorithms using different data sources and methods have been proposed to ascertain smoking status in electronic health data. Many algorithms had low sensitivity and positive predictive value, but the data source influenced their validity. Algorithms based on machine-learning models for multiple linked data sources have improved validity.

Keywords Algorithms, Electronic health records, Review, Routinely collected health data, Validation study

*Correspondence:
Lisa M. Lix
lisa.lix@umanitoba.ca
Full list of author information is available at the end of the article

Background

Electronic health databases, including electronic medical records (EMRs) and administrative data, contain routinely-collected information that is widely used for health research [1, 2] even though they were not originally intended for this purpose. EMRs typically include a diverse amount of information about the patient, including medical history, family history, immunization status, laboratory test results, and radiology images [3]. Administrative data also include large amounts of information, including insurance enrollment dates, inpatient and outpatient contacts, and vital statistics [4]. Data quality is an important consideration when using electronic health databases for research, given that the data are used for secondary purposes.

Smoking is responsible for more than 8 million deaths worldwide each year [5] and is the leading cause of preventable diseases and premature deaths [6]. Smoking is a significant risk factor for cancers, cardiovascular diseases, and respiratory diseases. Valid measurement of smoking status contributes to accurate estimates from risk prediction models and other outcome studies about these diseases [7]. Valid smoking status measures can also aid in accurately estimating disease trends at the population level.

Population-based surveys are typically used to capture information about smoking status. However, they are expensive to conduct and are not always conducted on a routine basis [8]. Routinely collected administrative data often contain indirect information about smoking status, such as diagnosis codes for related diseases (e.g., chronic bronchitis) and substance use disorders. EMRs capture information on smoking status through free-text information about one's health history as well as diagnosis codes. Therefore, many studies have investigated the validity of electronic health databases, including administrative databases and EMRs, for capturing information on smoking status. For example, a study based on Medicare claims data reported smoking status ascertainment algorithms with limited sensitivity but very high specificity [9]. In a different study, sensitivity estimates of algorithms for EMRs within an integrated healthcare system varied widely by years of data used [10]. However, the positive predictive value (PPV) consistently remained high.

To date, there have been few, if any studies that have systematically examined validation studies about smoking status in electronic health databases. Summary information about smoking status ascertainment algorithms might be used to develop recommendations about the optimal algorithm(s) to use and opportunities for further research. The latter is particularly timely, given increasing interest in novel approaches to mine new information from electronic health databases using machine-learning methods [11–13].

Given this background, the purpose of our study was to synthesize information about smoking status algorithms developed for electronic health databases. The objectives were to describe smoking status algorithm characteristics, methods to construct the algorithms, and estimates of their validity. A systematic review methodology was used to provide a comprehensive summary of the algorithms to ascertain smoking status [14]. Meta-analysis [15] of algorithm validity measure estimates was conducted to assess the potential sources of heterogeneity in them.

Methods

We used the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guideline for this review [16] (see Additional file 4). This guideline is widely recognised for ensuring rigorous and consistent reporting in systematic reviews and meta-analyses [17].

Data sources and search strategy

EMBASE, MEDLINE, Scopus, and Web of Science were searched from 1990 to November 22, 2022. The target sources were English-language peer-reviewed journal articles. We excluded review articles. The search strategy was developed by the research team in consultation with an experienced university librarian. The search strategy was based on three concepts, each built with specific sets of keywords. The concepts are validity measures (valid*, quality, accuracy, sensitivity, specificity), electronic health data (electronic medical record*, electronic health record*, administrative health data, administrative data, health care data, administrative billing record*, administrative claims data, claims data, hospital data, hospital discharge data, medicare data, medicaid data), and smoking status (smok*, tobacco). The * at the end of the words valid, record, and smok indicates the use of truncation to capture variant endings. The keywords within each set were connected with OR and the concepts were connected with AND. Article titles, abstracts, and keywords were reviewed to identify potentially relevant articles. The detailed search strategy implemented for each database is available in Additional file 1.

Study selection

The titles, abstracts, and keywords of the selected articles were uploaded to Rayyan [18] for deduplication and screening for inclusion. An article was included if it reported the results of a validation study for one or more smoking status ascertainment algorithms developed for EMRs or administrative health data (e.g., hospital records, physician claims, prescription drug records).

There were no restrictions on geography of the data or population characteristics (e.g., age, sex). To ensure an acceptable level (>80%) [19] of agreement between the two reviewers who conducted the screening, two rounds of abstract and title screening training were undertaken; each training session was conducted on a random sample of 10% of the identified articles. Both reviewers independently screened all the articles. Inter-reviewer agreement was assessed with Cohen's kappa [20]. Disagreements on study selection for full-text review were resolved by consensus. Articles were retained when there was uncertainty regarding the eligibility to be included based on title, abstract, and keywords alone. The final decision about the inclusion of an article in this study was made with full text review of the articles identified based on titles, abstracts, and keywords screening. The reference lists of the selected articles were searched for additional articles.

Data extraction

Two reviewers independently extracted data from two randomly selected articles in a training session to maintain high inter-reviewer reliability [21]. Disagreements on data extraction were resolved by consensus and discussion between the reviewers. The remainder of the articles were equally distributed to the two reviewers for data extraction. We extracted information from the selected studies about article characteristics, algorithm characteristics, and algorithm validation estimates.

Article characteristics included year of publication, geographical data source, whether data from multiple jurisdictions were used, and journal discipline. The latter was determined based on subject terms from the United States National Library of Medicine catalog in PubMed.

Algorithm characteristics included International Statistical Classification of Diseases and Related Health Problems (ICD) codes, procedure or intervention codes, data source, data structure, and the use of a predictive model to develop the algorithm. The algorithm data source was categorized as EMR or administrative data. Data structure was classified as structured (e.g., diagnosis codes), unstructured text (e.g., clinical notes), or both structured and unstructured. Algorithms were classified as model-based or deterministic/rule-based on the basis of the method of construction [22]. Model-based algorithms implemented predictive statistical and/or machine-learning models (e.g., support vector machine). Rule-based approaches relied on measuring the type and frequency of diagnosis/billing codes in the records of an individual.

Information about the validation data source and validity measures (e.g., sensitivity, specificity) was also extracted from the articles. For model-based algorithms, the validity measure estimates for test data were extracted. The validation source was classified as self-reported data (e.g., survey), chart review data (e.g., patient charts reviewed to extract smoking status based on assessments by clinical or domain-knowledge experts), and clinical data (e.g., blood test results). The reported validity measures and their respective estimates and 95% confidence intervals (CI) were recorded. If estimates were reported for more than one sub-group or category (e.g., by demographic characteristics, by years of data), only the overall value of the validity measure was extracted. Finally, we assessed whether the Standards for Reporting Diagnostic Accuracy (STARD) criterion [23] about the number of measures recommended for reporting were fulfilled.

Statistical analysis

We analyzed the extracted data at both the study level and algorithm level. At the study level, data about the articles were descriptively analyzed using frequencies and percentages. At the algorithm level, we conducted descriptive analyses overall, and then stratified by algorithm characteristics. The distributions of algorithm validity measures were visually summarized using boxplots; the median and interquartile range (IQR) were used to describe the data.

Sources of heterogeneity in algorithm validity estimates, including algorithm characteristics and article characteristics, were examined using a three-level meta-regression model [24]. The first level accounts for sampling error, the second level examines algorithm characteristics, and the third level considers article characteristics. The structure of this model is depicted in Fig. 1. The random deviations at each level were assumed to follow a normal distribution with zero mean and constant variance.

Standard errors for the reported estimates were calculated from CIs or number of positive/negative cases in the validation data [25]. Likelihood ratio (LR) tests were conducted to compare three-level models against their two-level counterparts [26]. Null models were fitted to calculate pooled estimate and variance in the reported validity measures. To further investigate the sources of heterogeneity in these estimates, predictor variables were included in the model. At level 2, data source (1 = EMRs, 0 = Administrative) and the use of a predictive model were included in the model (1 = Yes, 0 = No) to account for algorithm characteristics. The variables included at level 3 were article characteristics: reference standard (1 = Chart review/clinical data, 0 = Self-report), clinical population (1 = Yes (e.g., HIV patients), 0 = No), study population age (1 = Restricted to only a specific age-group (e.g., 15–45 years [27]), 0 = all ages), and country of data origin (1 = US,0 = non-US). Data structure was not

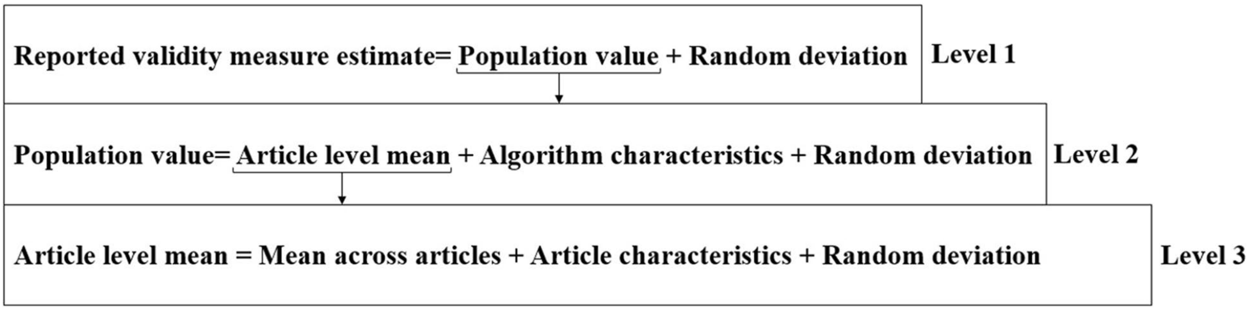
Reported validity measure estimate= Population value + Random deviation
Level 1
Population value= Article level mean + Algorithm characteristics + Random deviation
Level 2
Article level mean = Mean across articles + Article characteristics + Random deviation
Level 3

Fig. 1 Three-level meta-regression model

included in the model since it has a strong association with data source; administrative data are generally structured while EMRs can be either structured or unstructured. Test of residual heterogeneity was conducted to find if the heterogeneity not explained by the models is significant or not [28]. To assess model fit, reduction in variance estimates for the models with predictors relative to the initial random-effects pooling models [28] were calculated. An R package metafor [29] was used to conduct the meta-analysis.

Risk of bias assessment

The articles included in the meta-analysis underwent a risk of bias (ROB) assessment, utilizing the Quality Assessment of Diagnostic Accuracy Studies-2 (QUADAS-2) tool [30]. This tool comprises four domains that evaluate patient selection, index test, reference standard, and flow of patients through the study and timing of the index test(s). Each domain includes specific signaling questions to aid in determining the ROB. For each article, the ROB was evaluated as high, low, or unclear for each domain. A low ROB was assigned if all signaling questions within a domain were answered affirmatively (i.e., best practices were followed). Conversely, a high or unclear ROB was assigned if any signaling question received a negative or unclear response. To ensure accuracy, two reviewers independently performed the ROB assessment on a 5% random training sample of eligible articles. Discrepancies between the reviewers were addressed by a third reviewer, to reach consensus. The remaining articles were then evenly distributed between the two reviewers for ROB assessment.

A sensitivity analysis of the meta-regression models was conducted to assess robustness of the synthesized results. The sensitivity analysis excluded the articles with the presence of high/unclear ROB in any of the four domains of QUADAS-2. A publication bias test was conducted by regressing the null model residuals on their corresponding variance estimates [31].

Results

Search results

As shown in Fig. 2, a total of 4335 articles were retrieved from the literature search. After removing duplicates, the titles and abstracts of 2086 studies were screened for study inclusion. The screening process left 70 articles for full-text review. Cohen's kappa for study inclusion/exclusion was 0.97 (95% [CI]: 0.93, 1.00). After full-text review, 20 articles were removed. An additional seven articles were included after the review of reference lists of the remaining 50 articles. Therefore, a total of 57 articles were included in our systematic review (see Additional file 2).

Article characteristics

Only one (1.8%) of the included articles was published before 2000. The majority of the articles (32, 61.4%) were published after 2014 (Table 1). In most of the articles (44, 77.2%), algorithms were constructed using US data. A large number (37, 64.9%) of articles used clinical population (e.g., lung cancer patients) data. Very few articles (8, 14.0%) reported the use of data from more than one jurisdiction. Overall, the largest number of articles (24, 42.1%) were published in medical informatics/electronic data journals. The majority of the studies (27, 47.4%) validated algorithms using self-report data, followed by (26, 45.6%) chart review data, and (4, 7.0%) clinical data (e.g., serum cotinine in the blood). Only three (5.3%) articles reported that the study population included exclusively a single biological sex group.

Characteristics of the identified algorithms

The 57 articles reported on validity estimates for 116 algorithms. Overall, 50 (43.1%) algorithms used ICD codes; of this number only 10 used the 10th revision (e.g., tobacco dependence syndrome, personal history of tobacco use disorder) of this classification system and the remainder used the 9th revision (e.g., tobacco use

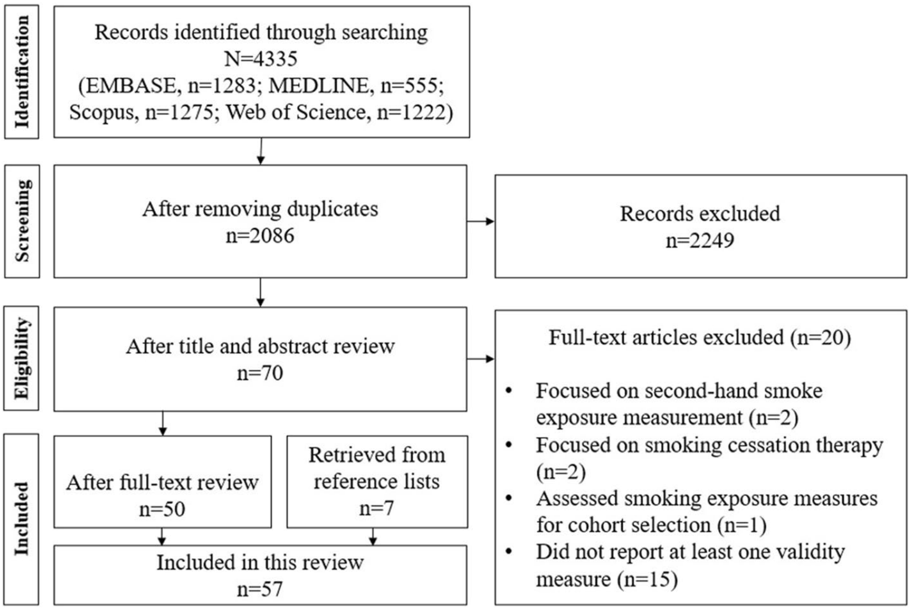
Identification
Records identified through searching
N=4335
(EMBASE, n=1283; MEDLINE, n=555;
Scopus, n=1275; Web of Science, n=1222)
Screening
After removing duplicates
n=2086
Records excluded
n=2249
Eligibility
After title and abstract review
n=70
Full-text articles excluded (n=20)
• Focused on second-hand smoke exposure measurement (n=2)
• Focused on smoking cessation therapy (n=2)
• Assessed smoking exposure measures for cohort selection (n=1)
• Did not report at least one validity measure (n=15)
Included
After full-text review
n=50
Retrieved from reference lists
n=7
Included in this review
n=57

Fig. 2 Flowchart of the study selection process

Table 1 Summary of article characteristics (*n* = 57)

Characteristic	%	Article Reference
Year of publication		
≥ 2015	61.4	[9, 12, 27, 32–63]
< 2015	38.6	[10, 64–84]
Geographical location of data source		
US	77.2	[9, 10, 27, 33–40, 42–48, 50–55, 57, 59–61, 66–73, 75–83]
Australia	5.3	[64, 65, 74]
UK	3.5	[49, 81]
Canada	3.5	[41, 56]
Other	10.5	[12, 32, 46, 58, 62, 84]
Clinical population		
Yes	64.9	[27, 33, 35, 36, 38, 39, 41–48, 52, 53, 55–58, 60–63, 66–68, 70, 72–74, 76–78, 82, 83]
No	35.1	[10, 12, 32, 34, 37, 40, 49–51, 54, 59, 64, 65, 69, 71, 75, 79–81, 84]
Data from multiple jurisdictions		
Yes	14.0	[43, 44, 48, 50, 68, 70, 72, 73]
No	86.0	[9, 10, 12, 27, 32–42, 45–47, 49, 51–67, 71, 74–84]
Journal discipline		
Medical informatics/Electronic data	42.1	[27, 34, 41, 45, 47–55, 57, 60, 66, 67, 73, 74, 76, 78–80, 83]
Medicine/Clinical	29.8	[35, 36, 38–40, 43, 44, 46, 56, 58, 59, 64, 65, 70, 71, 75, 77]
Public health and epidemiology	10.5	[9, 12, 32, 37, 42, 81]
Health services	8.8	[62, 63, 68, 72, 82]
Substance-related disorders	5.3	[10, 61, 69]
Biomedical	3.5	[33, 84]

complicating pregnancy, toxic effect of tobacco). Only 11 (9.5%) algorithms used procedure or intervention codes such as advisement to quit and screening for tobacco use followed by an intervention (i.e., smoking cessation program).

Almost three-quarters of the algorithms (83, 71.6%) were constructed using EMR data; the remaining 33 (28.4%) algorithms were constructed using administrative data. Nearly half of the algorithms (54, 46.5%) used structured data such as diagnosis codes, while unstructured EMR data were used to construct 41 (35.3%) algorithms. Only 21 (18.1%) algorithms were based on both structured and unstructured data.

More than half of the algorithms (61, 52.6%) were developed using rule-based methods, such as the presence of any tobacco-related ICD code or a procedure/intervention code in any data source, or the presence of any smoking-related information in inpatient records and/or outpatient medical claims within a defined period of time. The model-based algorithms ($n=55$) were almost exclusively (53, 96.4%) developed using EMR data. Largest number of the model-based algorithms (24, 43.6%) were developed using natural language processing methods. Specifically, these algorithms were developed by extracting smoking-related information from EMRs and constructing features relevant to smoking status (e.g., former smoker, current smoker), frequency of smoking (e.g., number of cigarettes per day), and temporal information relevant to date or duration (e.g., smoked for 10 years). A total of 16 (29.1%) model-based algorithms were developed using support vector machine models. The remainder (15, 27.3%) used statistical or machine-learning models, such as logistic regression, naïve Bayes, Bayesian networks, neural networks, deep learning methods, and decision trees.

Validity measures

Algorithm validity measures reported are depicted in Fig. 3. The number of validation measures reported per algorithm had a median value of 3.0 (IQR = 2). The STARD recommendation of four measures was met for slightly less than half (45.7%) of the identified algorithms. The three most common validity measures were sensitivity (80, 68.9%), specificity (61, 52.6%), and PPV (58, 50.0%). Area under the receiver operating characteristic (ROC) curve (10, 8.6%), true positives (12, 10.3%), and true negatives (14, 12.1%) were the least reported validity measures. The median (IQR) for PPV, sensitivity, and specificity were 88.3% (14.5%), 77.5% (36.5%), and 97.0% (12.0%) respectively.

Figure 4 indicates a negatively skewed distribution of specificity estimates for algorithms constructed using administrative data or EMRs, but not for sensitivity and PPV. The median PPV (91.0%) and sensitivity (86.0%) of

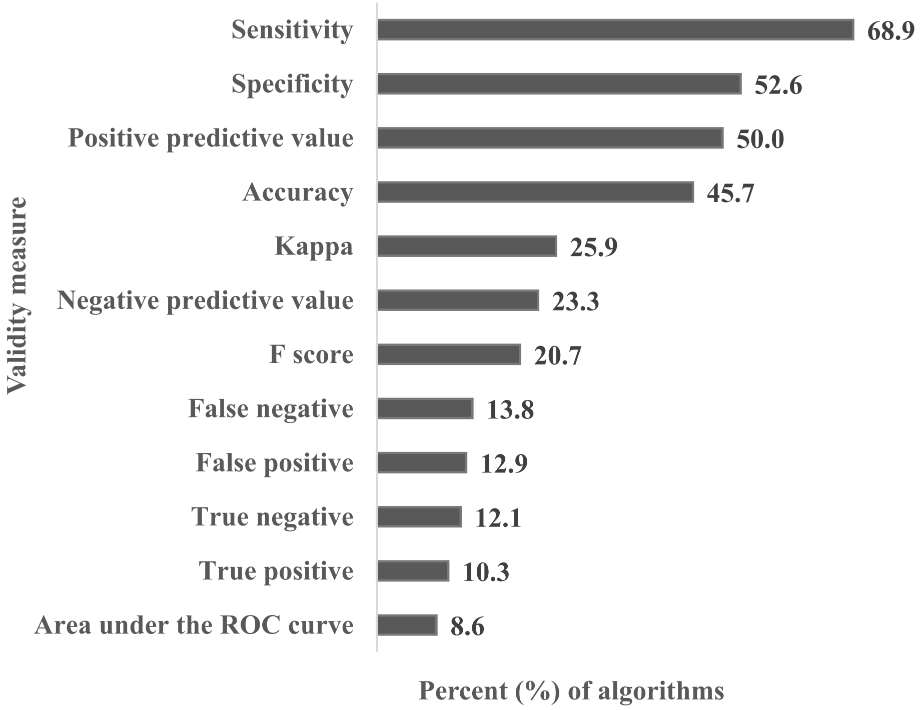

Fig. 3 Percent (%) of smoking status algorithms characterized by validity measures ($n=116$)

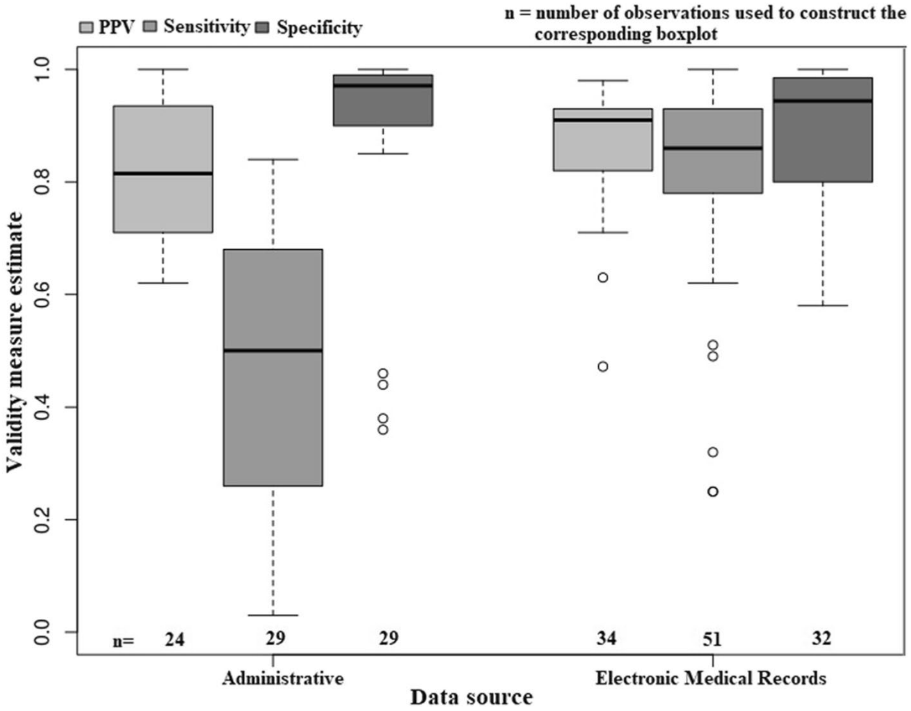

Fig. 4 Distribution of selected algorithm validity measures, stratified by algorithm data source. Note: The centre horizontal line within the box represents the median (50th percentile); upper and lower bounds of the box indicate the 25th and 75th percentiles; dashed lines connect the maximum and minimum values; circles represent outliers. PPV = positive predictive value

the algorithms based on EMR data was higher than for administrative data (PPV = 81.5%, sensitivity = 50.0%). However, median specificity (94.4%) estimate was lower for EMR data than for administrative data (97.1%). For EMR data, variability in estimates of PPV (IQR = 10.5%) and sensitivity (IQR = 15.0%) were lower than for administrative data (IQR for PPV = 23.0% and sensitivity = 42.0%). However, variation in estimates of specificity for EMR data was about twice that of administrative data (IQR 15.2 and 9.0%, respectively).

Figure 5 shows that the distribution of sensitivity was negatively skewed irrespective of data structure, while this was not the case for PPV and specificity. Median PPV and sensitivity estimates for algorithms based on unstructured data (PPV = 91.0%, sensitivity = 88.0%, specificity = 92.5%) were higher than the estimates for algorithms based on structured data (PPV = 80.0%, sensitivity = 62.5%, specificity = 97.1%). Algorithms developed using both structured and unstructured data had PPV and sensitivity estimates similar to algorithms based on unstructured data alone (PPV = 91.0%, sensitivity = 93.0%, specificity = 74.0%). The estimated validity of algorithms based on unstructured data showed less variation (IQR for PPV = 8.2%, sensitivity = 11.0%, and specificity = 9.5%) than algorithms based on structured data (IQR for PPV = 22.0%, sensitivity = 46.5%, and specificity = 7.5%) or that used a combination of structured and unstructured data (IQR for PPV = 11.0%, sensitivity = 10.0%, and specificity = 23.9%).

Regardless of the approach (model-based or rule-based) used to create an algorithm, the distributions of sensitivity and specificity estimates were highly skewed; most of the observations were below their respective average estimates (Fig. 6). Mean sensitivity was 87.3% for model-based algorithms and 57.1% for rule-based algorithms, while mean specificity was 83.9% for model-based algorithms and 90.5% for rule-based algorithms. The median sensitivity for model-based algorithms and rule-based algorithms were 90.0 and 63.0%, respectively. The median PPV and specificity of the algorithms based on predictive models were 87.5 and 88.0% respectively, while algorithms based on deterministic methods had median PPV and specificity of 89.5 and 97.0%, respectively. Variation in PPV (IQR = 9.2%) and sensitivity (IQR = 12.8%) estimates of model-based algorithms was lower than for rule-based algorithms (PPV IQR = 20.5%, sensitivity IQR = 47.5%). Variation in specificity estimates (IQR = 25.4%) was higher for model-based algorithms than rule-based algorithms (IQR = 9.0%).

Estimates of PPV (47 from 20 articles), sensitivity (69 from 34 articles), and specificity (57 from 28 articles) with necessary information to calculate standard error

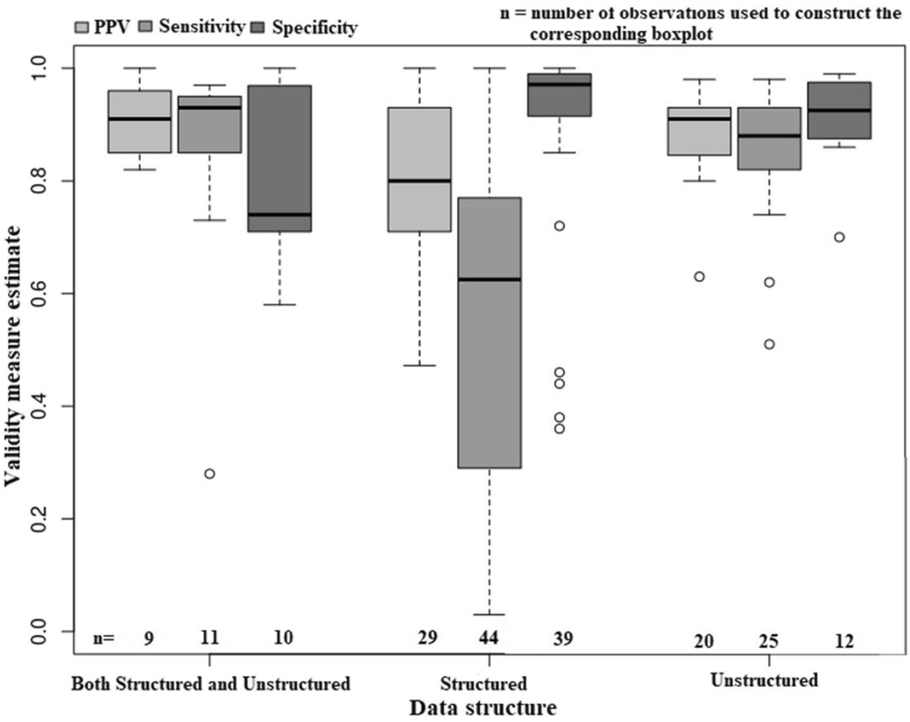

Fig. 5 Distribution of selected algorithm validity measures, stratified by algorithm data structure. Note: The centre horizontal line within the box represents the median (50th percentile); upper and lower bounds of the box indicate the 25th and 75th percentiles; dashed lines connect the maximum and minimum; circles represent outliers. PPV = positive predictive value

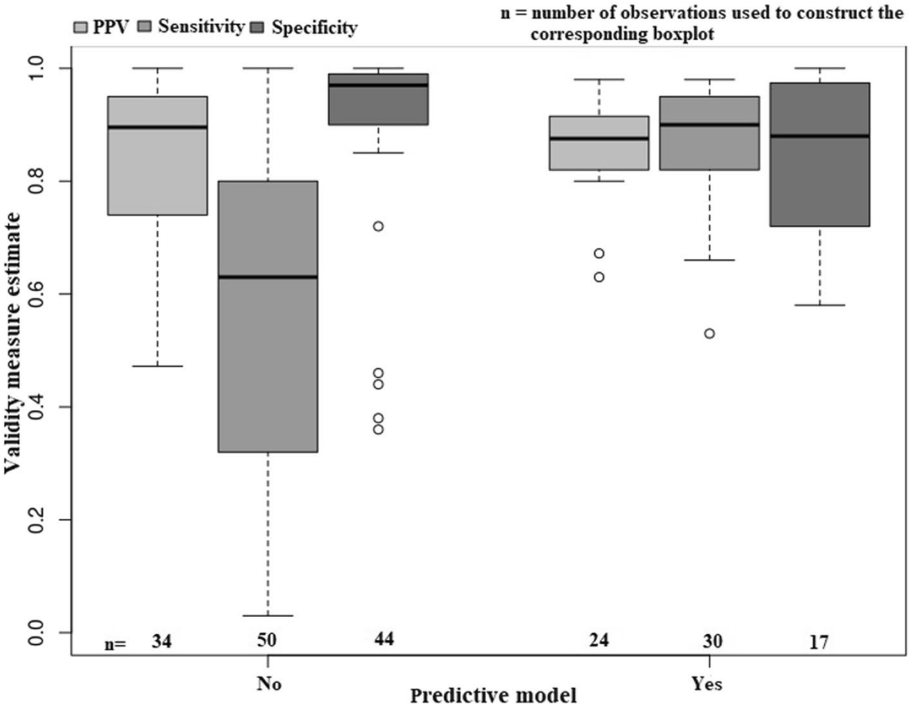

Fig. 6 Distribution of selected algorithm validity measures, stratified by use of predictive model in algorithm construction. Note: The centre horizontal line within the box represents the median (50th percentile); upper and lower bounds of the box indicate the 25th and 75th percentiles; dashed lines connect the maximum and minimum with the box; circles represent outliers. PPV = positive predictive value

(e.g., CI) were included in the meta-analysis. Three-level model provided significantly better fits compared to two-level models for all three measures PPV (LR = 7.75; $p = 0.005$), sensitivity (LR = 10.180; $p = 0.001$), and specificity (LR = 7.63; $p = 0.006$). The pooled estimates from the null three-level meta-regression models were 84.0% (95% [CI]: 79.0–90.0), 67.0% (95% [CI]: 59.0–76.0), and 92.0% (95% [CI]:87.0–97.0) for PPV, sensitivity, and specificity, respectively. The estimated variances of PPV, sensitivity, and specificity were 0.02, 0.07, and 0.03, respectively. The largest portion of estimated variance was attributable to level 2 in case of sensitivity (64.1%) and specificity (79.7%) followed by level 3 (sensitivity = 35.8%, specificity = 20.3%). However, level 3 (54.8%) accounted for the largest share of estimated variance in PPV followed by level 2 (45.2%). The residual regressions tests (deviation of the intercept from zero) did not indicate the presence of publication bias in the reported estimates of PPV ($p = 0.753$), sensitivity ($p = 0.471$), and specificity (0.124).

Algorithms for EMRs had significantly higher ($p = 0.001$) sensitivity estimates than algorithms for administrative data (Table 2). The model-based algorithms produced significantly higher sensitivity estimates than their rule-based counterparts ($p = 0.006$). The articles including only individuals with certain clinical conditions (e.g., HIV patients) had lower estimates of specificity than articles that used data for general populations ($p = 0.020$). The articles that used data from the US reported significantly larger specificity estimates than the articles based on non-US data ($p = 0.002$). None of the algorithm characteristics and article characteristics included in the model were significantly associated with PPV estimates.

The tests of residual heterogeneity in the models for all three measures suggest statistically significant amounts of unexplained heterogeneity in their estimates ($p < 0.0001$). The models with predictors had lower variances compared to the initial random-effects pooling model by 41.1 and 16.0%, respectively, for sensitivity and specificity. However, the estimated PPV variance remained unchanged in the null model and the model with the predictors. The predictors included in our models were able to explain the variability in sensitivity and specificity better than the respective null models. However, the predictors did not add any value when explaining variation in the PPV estimates.

Table 2 Meta-regression model parameter estimates (SE) for validity measures of PPV, sensitivity and specificity

Variable	**Validity measure**					
	PPV		**Sensitivity**		**Specificity**	
	Parameter estimate (SE)	***P*-value**	**Parameter estimate (SE)**	***P*-value**	**Parameter estimate (SE)**	***P*-value**
Data source						
EMR	−0.076 (0.073)	0.304	**0.251 (0.076)**	0.001	−0.057 (0.054)	0.297
Administrative	Ref		Ref		Ref	
Predictive model						
Yes	−0.017 (0.076)	0.826	**0.195 (0.069)**	0.006	−0.094 (0.052)	0.078
No	Ref		Ref		Ref	
Reference standard						
Chart review/clinical data	0.095 (0.072)	0.198	0.086 (0.082)	0.297	−0.025 (0.059)	0.672
Self-report	Ref		Ref		Ref	
Clinical population						
Yes	0.108 (0.063)	0.144	−0.096 (0.081)	0.239	**– 0.140 (0.058)**	0.020
No	Ref		Ref		Ref	
Study population age						
Restricted	0.133(0.067)	0.054	−0.016(0.062)	0.793	0.035(0.045)	0.441
All ages	Ref		Ref		Ref	
Country of data origin						
US	−0.006(0.072)	0.937	– 0.083(0.079)	0.304	**0.195(0.059)**	0.002
non-US	Ref		Ref		Ref	

Boldface font denotes a statistically significant estimate; *EMR* electronic medical record, *SE* standard error, *PPV* positive predictive value

Table 3 Risk of bias (ROB) assessment results [9, 10, 12, 33–38, 40–45, 48, 50, 51, 53–55, 57, 58, 60, 63, 64, 66, 68–70, 72–75, 77, 78, 80, 83]

Article Reference	**ROB domains**			
	Patient selection	**Index test**	**Reference standard**	**Flow and timing**
[9, 12, 33, 34, 36–38, 40–43, 45, 48, 51, 53, 57, 60, 63, 64, 68–70, 72–74, 77, 78, 80, 83]	⊕	⊕	⊕	⊕
[44, 55, 58, 75]	⊕	⊕	?	⊕
[35, 66]	?	⊕	⊕	⊕
[50]	⊕	⊖	⊕	⊕
[54]	?	⊕	⊕	⊖
[10]	⊕	⊕	⊕	⊖

Judgment on bias

⊕ Low risk ⊖ High risk ? Unclear risk

The results of the ROB assessment are summarized in Table 3. In total, 38 articles from the meta-regression models were evaluated for ROB using the QUADAS-2 tool. The majority of these articles (29, 76.3%) were deemed to have a low ROB across all four domains. High ROB was observed in three articles, with one in the index test domain and two in the flow and timing domain. Additionally, ROB was unclear in four articles in the reference standard domain and three articles in the patient selection domain.

The PPV, sensitivity, and specificity meta-regression models, respectively, had two articles (containing 4 algorithms), nine articles (containing 14 algorithms), and five articles (containing 5 algorithms) with high/unclear ROB. A sensitivity analysis that removed these articles with potential bias produced results consistent with the primary meta-analysis (Additional file 3).

Discussion

Smoking status is an important covariate in many disease risk prediction models and trends in smoking status are of interest in epidemiologic studies [85–87]. Electronic health databases can be leveraged to develop prediction models [88] and surveillance estimates that include smoking status information. Validation studies are important to assess the quality of electronic health data to ascertain a variety of individual characteristics, including smoking status [89]. Validation studies of electronic health data sources have been synthesized for chronic diseases such as diabetes [90], cancer [91], and social determinants of health (e.g., ethnicity, occupation) [92], but the validity of smoking status algorithms constructed from electronic health databases is a gap in the literature.

We found that a large number of validation studies for smoking status algorithms used electronic health data from the US; a similar trend was reported in systematic reviews for other validation studies, such as for comorbidity indices [93] and kidney disease [94]. Canadian provinces and territories collect comprehensive administrative health data [95]. Additionally, many provinces collect EMR data through the Canadian Primary Care Sentinel Surveillance Network [96] or other activities [97]. However, this study identified only two articles that used Canadian data. Similarly, we found very few articles used data from European countries, including the Scandinavian countries, which also have comprehensive administrative and EMR data [98, 99]. The vast majority of the algorithms were constructed with data from a single jurisdiction, potentially limiting the transportability of these algorithms across jurisdictions. At the same time, this finding is not unexpected, given that it can be challenging to find a comparable validation data source in more than one jurisdiction. A large number of studies used medical chart review to validate the algorithms, a result comparable to a systematic review of algorithms for obesity [100].

The median number of validity measures reported per algorithm was below the STARD recommendation of at least four measures. Overall, the sensitivity estimates were lower than the estimates of PPV and specificity. This finding is in line with a study that reviewed validation studies of algorithms to identify obesity in administrative databases [100]. Similar to other reviews focusing on chronic conditions [101], infections [102], and nonmedical opioid use [103]; considerable variation was found in the estimates for the reported algorithm validity measures such as PPV, sensitivity, and specificity overall and by selected algorithm characteristics. The model-based algorithms tended to have less variable PPV and sensitivity estimates than the rule-based algorithms. This suggests the model-based algorithms had more consistent performance in terms of accurately predicting positive cases (i.e., PPV) and capturing true positives (i.e., sensitivity). However, the model-based algorithms were more variable in specificity estimates compared to the rule-based algorithms. This may suggest the complexity of model-based algorithms affected specificity in different study-specific scenarios. Model-based smoking status ascertainment algorithms had better performance than rule-based algorithms, which is in contrast to findings from a rheumatoid arthritis validation study that utilised these two approaches and found similar predictive performance [104]. Nevertheless, another study demonstrated that model-based algorithms can improve sensitivity and specificity estimates over the estimates from rule-based algorithms for classifying carotid endarterectomy indication using physician claims and hospital discharge data [105]. Residual confounding due to data source may remain in the summarized relationship between validity measures and the use of a predictive model. This is particularly noteworthy, because a majority of model-based algorithms were constructed using EMR data. These data capture more detailed and comprehensive clinical information compared to administrative data, resulting in better performance than the rule-based algorithms that relied on administrative data. The model for sensitivity suggested significant difference in algorithms for EMRs and administrative data, in contrast to findings from an earlier study to predict binary outcomes such as 1 year mortality and hospital readmission [106]. The specificity model indicated considerably lower estimates for algorithms belonging to articles focusing only on clinical population and higher estimates for articles using US data. Class imbalance may be partially responsible for these findings on specificities [107]. No significant difference was detected in the results of PPV model. Overall, the fitted models with predictors did not adequately explain heterogeneity in validity measure estimates. This finding may be attributed to factors identified as sources of heterogeneity, such as the use of alternative coding methods and misclassified diagnosis by examining medications prescribed for different purposes, in the studies included in our meta-analysis [108]. Many articles in our meta-analysis had low ROB, but a review of algorithms for neurodevelopmental disorders had contrasting results [109]. The key findings from the meta-analysis did not change after a sensitivity analysis that excluded articles considering the ROB. A similar result was observed in a meta-analysis of articles using machine-learning to predict the spread of breast cancer to armpit lymph nodes [110]. Our analysis did not find strong evidence of publication bias in the pooled estimates of validity measures from null models. This finding should be interpreted with caution considering the limitations of linear regression [111]. For example, the relationship between residuals and variances may be non-linear, which can lead to inaccurate assessment of publication bias.

Strengths and limitations

The strengths of this systematic review and meta-analysis include the breadth of citation databases that we searched, the wide variety of article characteristics, and the detailed analysis of extracted information at the algorithm level. We identified statistically significant sources of variation in the estimates of sensitivity (data source and use of predictive model) and specificity (patient characteristics and country of data origin). However, we recognize that this study is not without limitations. Articles published in languages other than English and grey literature, such as government reports and graduate dissertations, were excluded. These exclusions may affect the generalizability of our findings. To mitigate this gap, the reference lists of the included articles were searched for additional articles. A large portion of the variation in the reported estimates of PPV, sensitivity, and specificity remained unexplained in the meta-regression models. To evaluate the reliability of the results of these models, a sensitivity analysis was performed incorporating the findings from the ROB assessment.

Conclusions

Evidence syntheses of algorithm validation studies have often focused on chronic or infectious disease case ascertainment and social determinants of health [90–92]. This study contributes to the body of literature about validation studies and examines a relatively unexplored area of behavioral risk factor algorithms for electronic health databases.

We found that numerous algorithms have been developed to identify smoking status in electronic health databases. The identified algorithms vary in terms of data source, data structure, and methods of construction. In general, the algorithms had high specificity and low sensitivity when predicting smoking status, although there is evidence that sensitivity can be improved by using EMR data and predictive models to construct the algorithms.

A number of opportunities exist to develop algorithms to measure smoking status using population-based electronic health data. For example, combining multiple data sources, including EMR and administrative data, may produce algorithms with high sensitivity [112]. The breadth of the longitudinal information [22] available in the electronic health databases can be utilized to develop algorithms. Methods such as longitudinal discriminant analysis [113] and semiparametric mixed-effects model [114] can be used to construct algorithms based on longitudinal data. The application of ensemble machine learning classification models and use of large language models (LLMs) remained unexplored in this line of research. Ensemble machine learning involves combining individual models to improve overall predictive performance. For example, random forest ensemble classifiers [115, 116] may be used to identify smoking status in electronic health databases. These classifiers may have reduced potential over-fitting of the model and improve performance measures relative to decision trees [117]. LLMs are trained deep-learning models that understands and generates text in a human-like fashion [118]. These models can be deployed to identify smoking status from text-based unstructured EMR data.

Abbreviations

EMR	Electronic medical record
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
CI	Confidence interval
ICD	International Statistical Classification of Diseases and Related Health Problems
STARD	Standards for Reporting Diagnostic Accuracy
IQR	Interquartile range
ROB	Risk of bias
PPV	Positive predictive value
ROC	Receiver operating characteristic

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02416-3.

Additional file 1. Search strategy.

Additional file 2. Article Characteristics.

Additional file 3. Sensitivity analysis.

Additional file 4. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist.

Acknowledgements

Not applicable.

Authors' contributions

LML: Conceived the study, prepared the analysis plan, conducted the analysis, and prepared the draft manuscript. AH: Conceived the study, prepared the analysis plan, performed the literature search, screening for study inclusion/exclusion, and risk of bias assessment, conducted the analysis, and prepared the draft manuscript. MLBG: Prepared the analysis plan, performed the literature search, screening for study inclusion/exclusion, and risk of bias assessment, and reviewed the manuscript. NN: Prepared the analysis plan and reviewed the manuscript. MT: Prepared the analysis plan and reviewed the manuscript. All authors approved the final version of the manuscript.

Funding

Funding for this study was provided by the Canadian Institutes of Health Research (grant # MOP-142404). MAH and MLBG are supported by funding from the Visual and Automated Disease Analytics (VADA) Program at the University of Manitoba. LML is supported by a Tier 1 Canada Research Chair in Methods for Electronic Health Data Quality (CRC-2017-00186).

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

This study does not involve human participants. Hence, ethics approval is not required.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Community Health Sciences, University of Manitoba, Winnipeg, MB, Canada. [2]Department of Statistics, University of Manitoba, Winnipeg, MB, Canada.

Received: 19 June 2023 Accepted: 3 January 2024
Published online: 02 February 2024

References

1. Cowie MR, Blomster JI, Curtis LH, Duclaux S, Ford I, Fritz F, Goldman S, Janmohamed S, Kreuzer J, Leenay M, Michel A. Electronic health records to facilitate clinical research. Clin Res Cardiol. 2017;106:1–9.
2. Lee S, Xu Y, D'Souza AG, Martin EA, Doktorchik C, Zhang Z, Quan H. Unlocking the potential of electronic health records for health research. Int J Popul Data Sci. 2020;5(1):1123.
3. Kierkegaard P. Electronic health record: wiring Europe's healthcare. Comput Law Secur Rev. 2011;27(5):503–15.
4. Harbaugh CM, Cooper JN. Administrative databases. Semin Pediatr Surg. 2018;27(6):353–60.
5. World Health Organization. Tobacco fact sheet from WHO providing key facts and information on surveillance. https://www.who.int/news-room/fact-sheets/detail/tobacco. Accessed 10 Apr 2022.
6. Canadian Lung Association. Smoking and tobacco statistics. https://www.lung.ca/lung-health/lung-info/lung-statistics/smoking-and-tobacco-statistics. Accessed 10 Apr 2022.
7. Barrett JK, Sweeting MJ, Wood AM. Dynamic risk prediction for cardiovascular disease: an illustration using the ARIC study, vol. 36. Handbook of Statistics; 2017. p. 47–65.
8. Kelsey JL, Kelsey C, Whittemore AS, Whittemore P, Evans AS, Thompson WD, et al. Methods in observational epidemiology. Oxford University Press; 1996. p. 458.
9. Desai RJ, Solomon DH, Shadick N, Iannaccone C, Kim SC. Identification of smoking using Medicare data—a validation study of claims-based algorithms. Pharmacoepidemiol Drug Saf. 2016;25(4):472–5.
10. Chen LH, Quinn V, Xu L, Gould MK, Jacobsen SJ, Koebnick C, Reynolds K, Hechter RC, Chao CR. The accuracy and trends of smoking history documentation in electronic medical records in a large managed care organization. Subst Use Misuse. 2013;48(9):731–42.
11. Chowdhury M, Cervantes EG, Chan WY, Seitz DP. Use of machine learning and artificial intelligence methods in geriatric mental health research involving electronic health record or administrative claims data: a systematic review. Front Psychiatry . 2021;12:738466.

12. Groenhof TK, Koers LR, Blasse E, de Groot M, Grobbee DE, Bots ML, Asselbergs FW, Lely AT, Haitjema S, van Solinge W, Hoefer I. Data mining information from electronic health records produced high yield and accuracy for current smoking status. J Clin Epidemiol. 2020;118:100–6.
13. Yadav P, Steinbach M, Kumar V, Simon G. Mining electronic health records (EHRs): a survey. ACM Comput Surv. 2018;50(6):1–40.
14. Caldwell PH, Bennett T. Easy guide to conducting a systematic review. J Paediatr Child Health. 2020;56(6):853–6.
15. Deeks JJ, Higgins JP, Altman DG, Cochrane Statistical Methods Group. Analysing data and undertaking meta-analyses. In: Cochrane handbook for systematic reviews of interventions. John Wiley & Sons, Ltd; 2019. p. 241–84.
16. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. BMJ. 2015;349:g7647.
17. PRISMA Statement organization. PRISMA Endorsers http://www.prismastatement.org/Endorsement/PRISMAEndorsers?AspxAutoDetectCookieSupport=1. Accessed 16 May 2023.
18. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Syst Rev. 2016;5:1–10.
19. Belur J, Tompson L, Thornton A, Simon M. Interrater reliability in systematic review methodology: exploring variation in coder decision-making. Sociol Methods Res. 2021;50(2):837–65.
20. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22(3):276–82.
21. Lange RT. Inter-rater reliability. In: Kreutzer JS, DeLuca J, Caplan B, editors. Encyclopedia of clinical neuropsychology. New York, NY: Springer; 2011. p. 1348.
22. Feely A, Lim LS, Jiang D, Lix LM. A population-based study to develop juvenile arthritis case definitions for administrative health data using model-based dynamic classification. BMC Med Res Methodol. 2021;21(1):1–3.
23. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, Lijmer JG, Moher D, Rennie D, De Vet HC, Kressel HY. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. Clin Chem. 2015;61(12):1446–52.
24. Weisz JR, Kuppens S, Ng MY, Eckshtain D, Ugueto AM, Vaughn-Coaxum R, Jensen-Doss A, Hawley KM, Krumholz Marchette LS, Chu BC, Weersing VR. What five decades of research tells us about the effects of youth psychological therapy: a multilevel meta-analysis and implications for science and practice. Am Psychol. 2017;72(2):79.
25. Wallis S. Binomial confidence intervals and contingency tests: mathematical fundamentals and the evaluation of alternative methods. J Quant Linguist. 2013;20(3):178–208.
26. Glover S, Dixon P. Likelihood ratios: a simple and flexible statistic for empirical psychologists. Psychon Bull Rev. 2004;11(5):791–806.
27. Wang Y, Sohn S, Liu S, Shen F, Wang L, Atkinson EJ, Amin S, Liu H. A clinical text classification paradigm using weak supervision and deep representation. BMC Medical Inform Decis Mak. 2019;19:1–3.
28. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. Doing meta-analysis with R: a hands-on guide. CRC Press; 2021.
29. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Softw. 2010;36(3):1–48.
30. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM, QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–36.
31. Doleman B, Freeman SC, Lund JN, Williams JP, Sutton AJ. Funnel plots may show asymmetry in the absence of publication bias with continuous outcomes dependent on baseline risk: presentation of a new publication bias test. Res Synth Methods. 2020;11(4):522–34.
32. Chung WS, Kung PT, Chang HY, Tsai WC. Demographics and medical disorders associated with smoking: a population-based study. BMC Public Health. 2020;20:1–8.
33. Wang L, Ruan X, Yang P, Liu H. Comparison of three information sources for smoking information in electronic health records. Cancer Informat. 2016;15:CIN-S40604.
34. Harris DR, Henderson DW, Corbeau A. Improving the utility of tobacco-related problem list entries using natural language processing. In: In: American Medical Informatics Association Annual Symposium Proceedings; 2020. p. 534.
35. Regan S, Meigs JB, Grinspoon SK, Triant VA. Determinants of smoking and quitting in HIV-infected individuals. PLoS One. 2016;11(4):e0153103.
36. Melzer AC, Pinsker EA, Clothier B, Noorbaloochi S, Burgess DJ, Danan ER, Fu SS. Validating the use of veterans affairs tobacco health factors for assessing change in smoking status: accuracy, availability, and approach. BMC Med Res Methodol. 2018;18:1–10.
37. Huo J, Yang M, Shih YC. Sensitivity of claims-based algorithms to ascertain smoking status more than doubled with meaningful use. Value Health. 2018;21(3):334–40.
38. Luck J, Larson AE, Tong VT, Yoon J, Oakley LP, Harvey SM. Tobacco use by pregnant Medicaid beneficiaries: validating a claims-based measure in Oregon. Prev Med Rep. 2020;19:101039.
39. Etzioni DA, Lessow C, Bordeianou LG, Kunitake H, Deery SE, Carchman E, Papageorge CM, Fuhrman G, Seiler RL, Ogilvie J, Habermann EB. Concordance between registry and administrative data in the determination of comorbidity: a multi-institutional study. Ann Surg. 2020;272(6):1006–11.
40. McVeigh KH, Lurie-Moroni E, Chan PY, Newton-Dame R, Schreibstein L, Tatem KS, Romo ML, Thorpe LE, Perlman SE. Generalizability of indicators from the New York city macroscope electronic health record surveillance system to systems based on other EHR platforms. eGEMs. 2017;5(1):25.
41. Marrie RA, Tan Q, Ekuma O, Marriott JJ. Development of an indicator of smoking status for people with multiple sclerosis in administrative data. Mult Scler J–Exp, Transl Clin. 2022;8(1):20552173221074296.
42. Floyd JS, Blondon M, Moore KP, Boyko EJ, Smith NL. Validation of methods for assessing cardiovascular disease using electronic health data in a cohort of veterans with diabetes. Pharmacoepidemiol Drug Saf. 2016;25(4):467–71.
43. Calhoun PS, Wilson SM, Hertzberg JS, Kirby AC, McDonald SD, Dennis PA, Bastian LA, Dedert EA, Mid-Atlantic VA, Workgroup MIRECC, Beckham JC. Validation of veterans affairs electronic medical record smoking data among Iraq-and Afghanistan-era veterans. J Gen Intern Med. 2017;32:1228–34.
44. Mu Y, Chin AI, Kshirsagar AV, Bang H. Data concordance between ESRD medical evidence report and Medicare claims: is there any improvement? PeerJ. 2018;6:e5284.
45. LeLaurin JH, Gurka MJ, Chi X, Lee JH, Hall J, Warren GW, Salloum RG. Concordance between electronic health record and tumor registry documentation of smoking status among patients with cancer. JCO Clin Cancer Inform. 2021;5:518–26.
46. Caccamisi A, Jørgensen L, Dalianis H, Rosenlund M. Natural language processing and machine learning to enable automatic extraction and classification of patients' smoking status from electronic medical records. Ups J Med Sci. 2020;125(4):316–24.
47. Palmer EL, Higgins J, Hassanpour S, Sargent J, Robinson CM, Doherty JA, Onega T. Assessing data availability and quality within an electronic health record system through external validation against an external clinical data source. BMC Medical Inform Decis Mak. 2019;19(1):1–9.
48. Golden SE, Hooker ER, Shull S, Howard M, Crothers K, Thompson RF, Slatore CG. Validity of veterans health administration structured data to determine accurate smoking status. Health Inform J. 2020;26(3):1507–15.
49. Atkinson MD, Kennedy JI, John A, Lewis KE, Lyons RA, Brophy ST. Development of an algorithm for determining smoking status and behaviour over the life course from UK electronic primary care records. BMC Medical Inform Decis Mak. 2017;17(1):1–2.
50. Reps JM, Rijnbeek PR, Ryan PB. Supplementing claims data analysis using self-reported data to develop a probabilistic phenotype model for current smoking status. J Biomed Inform. 2019;97:103264.
51. Ni Y, Bachtel A, Nause K, Beal S. Automated detection of substance use information from electronic health records for a pediatric population. J Am Med Inform Assoc. 2021;28(10):2116–27.
52. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. J Biomed Inform. 2015;58:S128–32.
53. Urbain J. Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models. J Biomed Inform. 2015;58:S143–9.

54. McVeigh KH, Newton-Dame R, Chan PY, Thorpe LE, Schreibstein L, Tatem KS, Chernov C, Lurie-Moroni E, Perlman SE. Can electronic health records be used for population health surveillance? Validating population health metrics against established survey data. eGEMs. 2016;4(1):1267.
55. Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. J Biomed Inform. 2015;58:S111–9.
56. Gauthier MP, Law JH, Le LW, Li JJ, Zahir S, Nirmalakumar S, Sung M, Pettengell C, Aviv S, Chu R, Sacher A. Automating access to real-world evidence. JTO Clin Res Rep. 2022;3(6):100340.
57. O'Brien EC, Mulder H, Jones WS, Hammill BG, Sharlow A, Hernandez AF, Curtis LH. Concordance between patient-reported health data and electronic health data in the ADAPTABLE trial. JAMA Cardiol. 2022;7(12):1235–43.
58. Alhaug OK, Kaur S, Dolatowski F, Småstuen MC, Solberg TK, Lønne G. Accuracy and agreement of national spine register data for 474 patients compared to corresponding electronic patient records. Eur Spine J. 2022;31(3):801–11.
59. Teng A, Wilcox A. Simplified data science approach to extract social and behavioural determinants: a retrospective chart review. BMJ Open. 2022;12(1):e048397.
60. McGinnis KA, Skanderson M, Justice AC, Tindle HA, Akgün KM, Wrona A, Freiberg MS, Goetz MB, Rodriguez-Barradas MC, Brown ST, Crothers KA. Using the biomarker cotinine and survey self-report to validate smoking data from United States veterans health administration electronic health records. JAMIA Open. 2022;5(2):ooac040.
61. McGinnis KA, Justice AC, Tate JP, Kranzler HR, Tindle HA, Becker WC, Concato J, Gelernter J, Li B, Zhang X, Zhao H. Using DNA methylation to validate an electronic medical record phenotype for smoking. Addict Biol. 2019;24(5):1056–65.
62. Maier B, Wagner K, Behrens S, Bruch L, Busse R, Schmidt D, Schühlen H, Thieme R, Theres H. Comparing routine administrative data with registry data for assessing quality of hospital care in patients with myocardial infarction using deterministic record linkage. BMC Health Serv Res. 2016;16(1):1–9.
63. Nickel KB, Wallace AE, Warren DK, Ball KE, Mines D, Fraser VJ, Olsen MA. Modification of claims-based measures improves identification of comorbidities in non-elderly women undergoing mastectomy for breast cancer: a retrospective cohort study. BMC Health Serv Res. 2016;16:1–2.
64. Havard A, Jorm LR, Lujic S. Risk adjustment for smoking identified through tobacco use diagnoses in hospital data: a validation study. PLoS One. 2014;9(4):e95029.
65. Lujic S, Watson DE, Randall DA, Simpson JM, Jorm LR. Variation in the recording of common health conditions in routine hospital data: study using linked survey and administrative data in New South Wales, Australia. BMJ Open. 2014;4(9):e005768.
66. Wiley LK, Shah A, Xu H, Bush WS. ICD-9 tobacco use codes are effective identifiers of smoking status. J Am Med Inform Assoc. 2013;20(4):652–8.
67. McGinnis KA, Brandt CA, Skanderson M, Justice AC, Shahrir S, Butt AA, Brown ST, Freiberg MS, Gibert CL, Goetz MB, Kim JW. Validating smoking data from the Veteran's affairs health factors dataset, an electronic data source. Nicotine Tob Res. 2011;13(12):1233–9.
68. Kim HM, Smith EG, Stano CM, Ganoczy D, Zivin K, Walters H, Valenstein M. Validation of key behaviourally based mental health diagnoses in administrative data: suicide attempt, alcohol abuse, illicit drug abuse and tobacco use. BMC Health Serv Res. 2012;12(1):1–9.
69. Lee JD, Delbanco B, Wu E, Gourevitch MN. Substance use prevalence and screening instrument comparisons in urban primary care. Subst Abus. 2011;32(3):128–34.
70. Jollis JG, Ancukiewicz M, DeLong ER, Pryor DB, Muhlbaier LH, Mark DB. Discordance of databases designed for claims payment versus clinical information systems: implications for outcomes research. Ann Intern Med. 1993;119(8):844–50.
71. Steffen MW, Murad MH, Hays JT, Newcomb RD, Molella RG, Cha SS, Hagen PT. Self-report of tobacco use status: comparison of paper-based questionnaire, online questionnaire, and direct face-to-face interview—implications for meaningful use. Popul Health Manag. 2014;17(3):185–9.
72. Borzecki AM, Wong AT, Hickey EC, Ash AS, Berlowitz DR. Identifying hypertension-related comorbidities from administrative data: what's the optimal approach? Am J Med Qual. 2004;19(5):201–6.
73. Bui DD, Zeng-Treitler Q. Learning regular expressions for clinical text classification. J Am Med Inform Assoc. 2014;21(5):850–7.
74. Khor R, Yip WK, Bressel M, Rose W, Duchesne G, Foroudi F. Practical implementation of an existing smoking detection pipeline and reduced support vector machine training corpus requirements. J Am Med Inform Assoc. 2014;21(1):27–30.
75. DeJoy S, Pekow P, Bertone-Johnson E, Chasan-Taber L. Validation of a certified nurse-midwifery database for use in quality monitoring and outcomes research. J Midwifery Womens Health. 2014;59(4):438–46.
76. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Medical Inform Decis Mak. 2006;6(1):1–9.
77. Longenecker JC, Coresh J, Klag MJ, Levey AS, Martin AA, Fink NE, Powe NR. Validation of comorbid conditions on the end-stage renal disease medical evidence report: the CHOICE study. J Am Soc Nephrol. 2000;11(3):520–9.
78. Meystre SM, Deshmukh VG, Mitchell J. A clinical use case to evaluate the i2b2 Hive: predicting asthma exacerbations. AMIA Ann Symp Proc. 2009;2009:442–6.
79. Clark C, Good K, Jezierny L, Macpherson M, Wilson B, Chajewska U. Identifying smokers with a medical extraction system. J Am Med Inform Assoc. 2008;15(1):36–9.
80. Savova GK, Ogren PV, Duffy PH, Buntrock JD, Chute CG. Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc. 2008;15(1):25–8.
81. Mant J, Murphy M, Rose P, Vessey M. The accuracy of general practitioner records of smoking and alcohol use: comparison with patient questionnaires. J Public Health. 2000;22(2):198–201.
82. Yeager DS, Krosnick JA. The validity of self-reported nicotine product use in the 2001–2008 National Health and nutrition examination survey. Med Care. 2010;48:1128–32.
83. Liu M, Shah A, Jiang M, Peterson NB, Dai Q, Aldrich MC, et al. A study of transportability of an existing smoking status detection module across institutions. AMIA Ann Symp Proc. 2012;2012:577–86.
84. Figueroa RL, Soto DA, Pino EJ. Identifying and extracting patient smoking status information from clinical narrative texts in Spanish. In: In: 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE; 2014. p. 2710–3.
85. Teramukai S, Okuda Y, Miyazaki S, Kawamori R, Shirayama M, Teramoto T. Dynamic prediction model and risk assessment chart for cardiovascular disease based on on-treatment blood pressure and baseline risk factors. Hypertens Res. 2016;39(2):113–8.
86. Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, Lassale CM, Siontis GC, Chiocchia V, Roberts C, Schlüssel MM. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416.
87. Chang JT, Meza R, Levy DT, Arenberg D, Jeon J. Prediction of COPD risk accounting for time-varying smoking exposures. PLoS One. 2021;16(3):e0248535.
88. Cadarette SM, Wong L. An introduction to health care administrative data. Can J Hosp Pharm. 2015;68(3):232.
89. Hoeven LR, Bruijne MC, Kemper PF, Koopman MM, Rondeel JM, Leyte A, Koffijberg H, Janssen MP, Roes KC. Validation of multisource electronic health record data: an application to blood transfusion data. BMC Medical Inform Decis Mak. 2017;17(1):1–10.
90. Rahimi AK, Canfell OJ, Chan W, Sly B, Pole JD, Sullivan C, Shrapnel S. Machine learning models for diabetes management in acute care using electronic medical records: a systematic review. Int J Med Inform. 2022;162:104758.
91. Conderino S, Bendik S, Richards TB, Pulgarin C, Chan PY, Townsend J, Lim S, Roberts TR, Thorpe LE. The use of electronic health records to inform cancer surveillance efforts: a scoping review and test of indicators for public health surveillance of cancer prevention and control. BMC Medical Inform Decis Mak. 2022;22(1):1–3.
92. Cook LA, Sachs J, Weiskopf NG. The quality of social determinants data in the electronic health record: a systematic review. J Am Med Inform Assoc. 2022;29(1):187–96.

93. Sharabiani MT, Aylin P, Bottle A. Systematic review of comorbidity indices for administrative data. Med Care. 2012;50(12):1109–18.
94. Vlasschaert ME, Bejaimal SA, Hackam DG, Quinn R, Cuerden MS, Oliver MJ, Iansavichus A, Sultan N, Mills A, Garg AX. Validity of administrative database coding for kidney disease: a systematic review. Am J Kidney Dis. 2011;57(1):29–43.
95. Lucyk K, Lu M, Sajobi T, Quan H. Administrative health data in Canada: lessons from history. BMC Medical Inform Decis Mak. 2015;15(1):1–6.
96. Birtwhistle R, Keshavjee K, Lambert-Lanning A, Godwin M, Greiver M, Manca D, Lagacé C. Building a pan-Canadian primary care sentinel surveillance network: initial development and moving forward. J Am Board Fam Med. 2009;22(4):412–22.
97. Tu K, Mitiku TF, Ivers NM, Guo H, Lu H, Jaakkimainen L, Kavanagh DG, Lee DS, Tu JV. Evaluation of electronic medical record administrative data linked database (EMRALD). Am J Manag Care. 2014;20(1):e15–21.
98. Hess DT. The Danish National Patient Register. Surg Obes Relat Dis. 2016;12(2):304.
99. Rusk N, The UK. Biobank. Nat Methods. 2018;15(12):1001.
100. Samadoulougou S, Idzerda L, Dault R, Lebel A, Cloutier AM, Vanasse A. Validated methods for identifying individuals with obesity in health care administrative databases: a systematic review. Obes Sci Pract. 2020;6(6):677–93.
101. McBrien KA, Souri S, Symonds NE, Rouhi A, Lethebe BC, Williamson TS, Garies S, Birtwhistle R, Quan H, Fabreau GE, Ronksley PE. Identification of validated case definitions for medical conditions used in primary care electronic medical record databases: a systematic review. J Am Med Inform Assoc. 2018;25(11):1567–78.
102. Barber C, Lacaille D, Fortin PR. Systematic review of validation studies of the use of administrative data to identify serious infections. Arthritis Care Res. 2013;65(8):1343–57.
103. Canan C, Polinski JM, Alexander GC, Kowal MK, Brennan TA, Shrank WH. Automatable algorithms to identify nonmedical opioid use using electronic data: a systematic review. J Am Med Inform Assoc. 2017;24(6):1204–10.
104. Kroeker K, Widdifield J, Muthukumarana S, Jiang D, Lix LM. Model-based methods for case definitions from administrative health data: application to rheumatoid arthritis. BMJ Open. 2017;7(6):e016173.
105. Van Gaal S, Alimohammadi A, Yu AY, Karim ME, Zhang W, Sutherland JM. Accurate classification of carotid endarterectomy indication using physician claims and hospital discharge data. BMC Health Serv Res. 2022;22(1):1–9.
106. Zeltzer D, Balicer RD, Shir T, Flaks-Manov N, Einav L, Shadmi E. Prediction accuracy with electronic medical records versus administrative claims. Med Care. 2019;57(7):551–9.
107. Van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. J Am Med Inform Assoc. 2022;29(9):1525–34.
108. Coleman N, Halas G, Peeler W, Casaclang N, Williamson T, Katz A. From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. BMC Fam Pract. 2015;16(1):1–8.
109. O'Donnell S, Palmeter S, Laverty M, Lagacé C. Accuracy of administrative database algorithms for autism spectrum disorder, attention-deficit/hyperactivity disorder and fetal alcohol spectrum disorder case ascertainment: a systematic review. Health Promot Chronic Dis Prev Canada: Res, Policy Pract. 2022;42(9):355.
110. Chen C, Qin Y, Chen H, Zhu D, Gao F, Zhou X. A meta-analysis of the diagnostic performance of machine learning-based MRI in the prediction of axillary lymph node metastasis in breast cancer patients. Insights Imaging. 2021;12:1–2.
111. Furuya-Kanamori L, Xu C, Lin L, Doan T, Chu H, Thalib L, Doi SA. P value–driven methods were underpowered to detect publication bias: analysis of Cochrane review meta-analyses. J Clin Epidemiol. 2020;118:86–92.
112. Al-Azazi S, Singer A, Rabbani R, Lix LM. Combining population-based administrative health records and electronic medical records for disease surveillance. BMC Medical Inform Decis Mak. 2019;19(1):1–2.
113. Hughes DM, El Saeiti R, García-Fiñana M. A comparison of group prediction approaches in longitudinal discriminant analysis. Biom J. 2018;60(2):307–22.
114. Arribas-Gil A, De la Cruz R, Lebarbier E, Meza C. Classification of longitudinal data through a semiparametric mixed-effects model based on lasso-type estimators. Biometrics. 2015;71(2):333–43.
115. Miled ZB, Haas K, Black CM, Khandker RK, Chandrasekaran V, Lipton R, Boustani MA. Predicting dementia with routine care EMR data. Artif Intell Med. 2020;102:101771.
116. Jauk S, Kramer D, Großauer B, Rienmüller S, Avian A, Berghold A, Leodolter W, Schulz S. Risk prediction of delirium in hospitalized patients using machine learning: an implementation and prospective evaluation study. J Am Med Inform Assoc. 2020;27(9):1383–92.
117. James G, Witten D, Hastie T, Tibshirani R. Tree-based methods. In: James G, Witten D, Hastie T, Tibshirani R, editors. An introduction to statistical learning: with applications in R. New York, NY: Springer; 2013. p. 303–35.
118. Thirunavukarasu AJ, Ting DS, Elangovan K, Gutierrez L, Tan TF, Ting DS. Large language models in medicine. Nat Med. 2023;29(8):1930–40.

Publisher's Note