Francis Nissen[1], Jennifer K. Quint[2], Daniel R. Morales[3], Ian J. Douglas[1]

francis.nissen@lshtm.ac.uk

https://www.linkedin.com/in/francisnissen/

Doing science

How to validate a diagnosis recorded in electronic health records

Cite as: Nissen F, Quint JK, Morales DR, *et al*. How to validate a diagnosis recorded in electronic health records. *Breathe* 2019; 15: 64–68.

Why validate diagnoses in electronic health records?

Over the last decades, the adoption of electronic health records (EHR) by health services worldwide has facilitated the construction of large population-based patient databases. These routinely generated longitudinal records have an enormous potential for epidemiological and clinical research [1–3]. EHR contain information on the health of an individual and are an electronic version of a patient's medical history. This contrasts with administrative claims data, whose main purpose is administration of reimbursement of medical services to healthcare providers. Due to the immense size of EHR, they can offer high statistical power and can often be representative of a population. Linkage between different EHR can further improve the completeness of the data. However, the primary *raison d'être* of most of these EHR is for clinical, administrative or audit purposes, which is a major challenge to their use for health research. Data elements that would be useful for research can therefore be wrongly classified, insufficiently specified or missing. Misclassified data can lead to systematic measurement errors. Missing data can lead to selection bias and counteract the statistical power provided by the magnitude of EHR [4].

Measuring data validity is therefore needed to establish whether the values "make sense" [5, 6]. When considering answering a research question using EHR, a researcher should always contemplate the following question: are we measuring what we are intending to measure?

The size of EHR does not resolve these data validity issues, and could even magnify these problems [7]. Big sample sizes can equally lead to big inferential errors if the validity of data is poor [8]. Whether the codes in an EHR database accurately capture the target condition (and thus minimise measurement errors) strongly affects the reliability of subsequent observational studies [9, 10]. The data validation process for research purposes is crucial to draw valid inference from these databases [4, 7, 11, 12]. In prospective cohort studies where one collects the data solely for the purpose of epidemiological research, application of stricter definitions for exposures, covariates and outcomes is possible, and specific tests or treatments that are not part of routine clinical practice can be requested, which is not the case when using EHR data. In addition, the validity of different diseases or conditions in EHR can vary within and between datasets. Some diseases (such as asthma) might be coded using combinations of diagnoses and/or less specific symptoms, whereas the validity of diagnoses with very specific symptoms (such as tension pneumothorax) is likely to be higher.

Algorithm construction

EHR databases generally store diagnostic information using codes selected from a structured medical

@ERSpublications

Systematic measurement errors in electronic health record databases can lead to large inferential errors. Validation techniques can help determine the degree of these errors and therefore aid in the interpretation of findings. http://ow.ly/iHQ630np4xU

Table 1 *Test measures*

	Reference standard		Outcome	Test measures
	Positive	Negative		
Diagnostic algorithm				
Positive	True positives correctly identified (A)	False positives (B)	Total identified as positive	PPV=A/(A+B)
Negative	False negatives (C)	True negatives correctly identified (D)	Total identified as negative	NPV=D/(C+D)
Outcome	True positives	True negatives		
Test measures	Sensitivity=A/(A+C)	Specificity=D/(B+D)		

dictionary. An algorithm consisting of a combination of codes can be constructed in order to identify all events of a target condition from the EHR database. These algorithms can consist of one or more diagnostic codes, or can include several other parameters, including medications, test results and disease symptoms. Additional inclusion or exclusion criteria can be included in the algorithm, such as age, sex or exclusion of other diseases [6]. These algorithms can be constructed manually or using machine-learning methods, to automate algorithm generation [13].

In general, if there are more parameters in an algorithm, it will identify fewer false positives. However, this comes with a trade-off, as the total identifiable population with the target condition will also decrease as fewer patients will fulfil all parameters, and the algorithm may only detect severe cases of the target condition. Furthermore, if two conditions have many overlapping parameters (for example, this is the case in asthma and chronic obstructive pulmonary disease (COPD)) [14], these parameters may not be very useful in differentiating between the two diseases and the inclusion of further parameters that can differentiate between the two conditions, or an exclusion based on the diagnostic code of the second disease, may be necessary. This process of code and parameter selection is therefore not always straightforward, hence the importance of considering their validity. The validity of an algorithm for a target condition within a database can be measured using separate test measures, which will be discussed in this article.

To avoid confusion, the word "validation" is frequently used in a variety of disciplines, including medicine and psychology, to measure the accuracy of an instrument or test. Examples of these are the degree to which evidence supports the interpretation of biomarkers or questionnaires for disease diagnosis. In most literature on healthcare databases, including this article, the term "validation" refers only to the reliability of coding.

Test measures

In research using EHR, the validity of codes and algorithms are quantified using diagnostic accuracy test measures, which relate what is recorded in the data to a recognised reference standard. The most commonly used and practical of these measures are the positive predictive value (PPV), the negative predictive value (NPV), the sensitivity and specificity. These test measures can be used to quantify the validity of an algorithm, and are a core concept of both epidemiology and instrument validation. The chosen test measure depends on the scope of the study. An overview of how these test measures are calculated is provided in table 1.

The PPV is the proportion of identified individuals with the target condition that truly have the target condition. The PPV is arguably the most practical test measure to validate an algorithm, as it can be measured using only a small sample of the population and reflects how accurate an algorithm is in identifying individuals with a target condition in EHR databases [15]. Similarly, the NPV is the proportion of individuals identified as negative that truly did not have the target condition. The NPV is useful, for example, to assess whether a control group that has been categorised as unexposed was truly unexposed.

Sensitivity measures the proportion of all individuals with a target condition that the algorithm identified correctly. An algorithm with a high sensitivity would detect a high proportion of all individuals with the target condition. The specificity measures the proportion of individuals that do not have the target condition that the algorithm correctly identified as negative. An algorithm with a high specificity would detect a high proportion of all individuals without the target condition. The sensitivity and specificity are important measures of the impact of missing data in the EHR data. If an algorithm fails to identify many individuals with a certain condition due to a low sensitivity, this can lead to selection bias. The specificity is important to consider when constructing control groups without the target condition. The prevalence of the target condition (or an estimate thereof) is required to calculate sensitivity or specificity values.

Sample validation techniques

There is no one-size-fits-all method of assessing the validity of algorithms in EHR. The optimal technique depends on the nature of the studied EHR database, the study question that needs answering and the way in which the diagnostic algorithm was constructed.

There are multiple ways to test the validity of these diagnostic algorithms in EHR. In a systematic review on validation studies in the General Practice Research Database, a large primary care UK EHR database that later evolved into the Clinical Practice Research Datalink (CPRD), Herrett *et al*. [16] divided the methods of the included studies into internal and external validation methods. External validation methods require dependable external reference standards (often referred to as "gold standard"), while internal validation methods do not require this external reference standard but will therefore not be able to quantify the discussed standard test measures.

The remainder of this section outlines a non-exhaustive list of eight common techniques to validate diagnostic algorithms with references to examples, ranked loosely from most to least resource-intensive. There are other possible techniques, including studying the completeness, plausibility, uniformity and time patterns of the data, which are not described in detail with examples in this study [5]. Not all these validation techniques are necessarily implementable in each database and some techniques are resource-intensive, while others provide only an indication of the validity. The choice of techniques is dependent on the database and access to data. Table 2 provides an overview of the test measures that can be calculated or estimated with the discussed techniques.

Manual validation of physical records

This technique determines if the EHR reflect the physical chart of the patient by manually going through a sample of clinical notes. Historically, this method was used to test EHR reliability when EHR database systems were being implemented. If the physical records accurately reflect the patient's status, this is a reliable way to test the validity of diagnostic codes. Weaknesses of this approach include the considerable time investment, and that these physical records may not always be available any more, as they tend to be phased out in favour of digital records. This technique also relies on the examination of physical records by someone who is usually not the treating physician, which can lead to misinterpretation. This technique is commonly used [17, 18].

Questionnaires for healthcare practitioners or patients

One can assess the "true" disease status of a patient by sending out questionnaires to either the patient or the healthcare professional responsible for their care. A questionnaire with appropriate design that can reliably ascertain the disease status of the individual patient is necessary for this technique. This technique can provide a reliable measure of the validity of an algorithm but can be resource-intensive, and the option may not be available in all databases. In addition, the clinician may be using the EHR database to look up the patient diagnoses, and patients or clinicians may be less likely to answer in more complicated cases. The validity of asthma and COPD recording in the CPRD-Global Initiative for Chronic Obstructive Lung Disease (CPRD-GOLD) was assessed using this technique [19, 20].

Validation of machine-learning algorithms

If the diagnostic algorithms were created using machine-learning techniques, it is possible to

Table 2 *Test measures that it is possible to calculate or estimate by each validation technique*

Technique	PPV	NPV	Sensitivity	Specificity
Manual validation of physical records	Can be calculated	Can be calculated	Can be calculated	Can be calculated
Questionnaires for healthcare practitioners or patients	Can be calculated	Can be calculated		
Validation of machine-learning algorithms	Can be calculated	Can be calculated	Can be calculated	Can be calculated
Comparison to an external database (complete overlap)	Can be calculated	Can be calculated	Can be calculated	Can be calculated
Comparison of rates in a comparable population			Estimate only	Estimate only
Internal validation using additional parameters	Estimate only	Estimate only		
Internal validation using free text in the database	Estimate only	Estimate only		
Sensitivity analyses using restrictive algorithms				

validate the algorithms within the database and create algorithms with high sensitivity values by varying the imbalance ratio between positive cases and negative cases. An example of this process was described by Afzal *et al*. [21]. Large amounts of data are usually required to derive, train and test the algorithm.

Comparison to an external database

If an independent secondary database is available, this can be a reliable and reasonably fast way to validate diagnostic algorithms. However, if the second database is not representative of the same population as the first database, results may not be generalisable. This technique was used by Edgren *et al.* [22] to compare data in a blood donation database to nationwide population and health registers in Denmark and Sweden.

Comparison of rates in a comparable population

A quick way to assess the credibility of recording of diagnostic coding is by comparing the prevalence or rates of the diagnostic code to the same measure in a comparable population. This technique is limited, as it can only provide rough estimates. In addition, if the over- and under-diagnosis rates of a disease are similar (*i.e.* there is systematic error affecting both), this technique will miss both measurement errors. This technique was used by Hansell *et al.* [23] to explore patterns in asthma and COPD morbidity and mortality.

Internal validation using additional parameters

This technique is most useful in algorithms consisting only of the diagnostic code. In essence, this method checks whether the patients who have received the diagnostic code also received treatment for that condition, have symptoms of the condition or were tested for the target condition. For example, the diagnosis of acute myocardial infarction was strengthened in a study by Andersohn *et al.* [24] by including only cases that also had codes for tests or treatments. This method requires a good degree of completeness in the data and certain data parameters not present in all data sources.

Internal validation using free text in the database

Similar to the previous technique, this technique checks whether the diagnosis is confirmed in available free text in the database. This is only available if the database offers free-text records, and the number of databases offering free-text access to researchers is declining due to confidentiality concerns. Yang *et al.* [25] confirmed colorectal cases by looking at free text in the database for confirmation of the cases.

Sensitivity analyses using restrictive algorithms

This method tests the soundness of study results using different diagnostic algorithms, so is an aggregate measure of both study analysis and diagnostic results. By comparing the baseline characteristics or measures of effects of a study using a broader algorithm (fewer parameters) and a narrower algorithm (more parameters), it is possible to assess whether results are driven by the inclusion or exclusion of patients in whom the diagnosis may be less certain. For example, the recording of vitamin D supplementation was assessed using sensitivity analyses in a study on cancer survival by Jeffreys *et al*. [26]. In this study, the analysis was restricted to women over the age of 60 years (as they received free vitamin D prescriptions at pharmacies and thus were less likely to obtain over-the-counter drugs); no difference in results was found.

Summary

When using large EHR databases for epidemiological or clinical research, it is paramount to be aware of the possibility of systematic measurement errors, as this can lead to large inferential errors. Validation studies help determine the degree of systematic measurement error and therefore aid in the interpretation of findings. Validation of diagnosis algorithms can help researchers by making their research in EHR more credible by quantifying the correctness of the data.

Affiliations

Francis Nissen[1], Jennifer K. Quint[2], Daniel R. Morales[3], Ian J. Douglas[1]

[1]Dept of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, UK. [2]National Heart and Lung Institute, Imperial College, London, UK. [3]Division of Population Health and Genomics, University of Dundee, Dundee, UK.

Conflict of interest

F. Nissen reports grants from GlaxoSmithKline (PhD scholarship) and Novartis (ERS fellowship in industry), outside the submitted work. J.K. Quint reports grants from the Medical Research Council, the British Lung Foundation, the Health Foundation and the Wellcome Trust, and grants and personal fees from GlaxoSmithKline, Insmed, Bayer and Boehringer Ingelheim, all outside the submitted work. D.R. Morales has nothing to disclose. I.J. Douglas reports grants from GlaxoSmithKline during the conduct of the study, as well as grants and share holdings from GlaxoSmithKline outside the submitted work.

References

1. Coorevits P, Sundgren M, Klein GO, *et al.* Electronic health records: new opportunities for clinical research. *J Intern Med* 2013; 274: 547–560.
2. Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS Med* 2015; 12: e1001885.
3. Langan SM, Benchimol EI, Guttmann A, *et al.* Setting the RECORD straight: developing a guideline for the REporting of studies Conducted using Observational Routinely collected Data. *Clin Epidemiol* 2013; 5: 29–31.
4. Ehrenstein V, Nielsen H, Pedersen AB, *et al.* Clinical epidemiology in the era of big data: new opportunities, familiar challenges. *Clin Epidemiol* 2017; 9: 245–250.
5. van Hoeven LR, de Bruijne MC, Kemper PF, *et al.* Validation of multisource electronic health record data: an application to blood transfusion data. *BMC Med Inform Decis Mak* 2017; 17: 107.
6. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013; 20: 144–151.
7. 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.
8. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci* 2014; 7: 342–346.
9. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. *BMJ* 2010; 341: c4226.
10. Wong M, Day NE. Validation studies in epidemiology: the relative precision of different designs. *J Epidemiol Biostat* 2000; 5: 331–337.
11. Gange SJ, Golub ET. From smallpox to big data: the next 100 years of epidemiologic methods. *Am J Epidemiol* 2016; 183: 423–426.
12. Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am J Public Health* 2005; 95: Suppl. 1, S144–S150.
13. Afzal Z, Engelkes M, Verhamme KM, *et al.* Automatic generation of case-detection algorithms to identify children with asthma from large electronic health record databases. *Pharmacoepidemiol Drug Saf* 2013; 22: 826–833.
14. Nissen F, Morales DR, Mullerova H, *et al.* Concomitant diagnosis of asthma and COPD: a quantitative study in UK primary care. *Br J Gen Pract* 2018; 68: e775–e782.
15. Brenner H, Gefeller O. Use of the positive predictive value to correct for disease misclassification in epidemiologic studies. *Am J Epidemiol* 1993; 138: 1007–1015.
16. Herrett E, Thomas SL, Schoonen WM, *et al.* Validation and validity of diagnoses in the General Practice Research Database: a systematic review. *Br J Clin Pharmacol* 2010; 69: 4–14.
17. Donahue JG, Weiss ST, Goetsch MA, *et al.* Assessment of asthma using automated and full-text medical records. *J Asthma* 1997; 34: 273–281.
18. Wu ST, Sohn S, Ravikumar KE, *et al.* Automated chart review for asthma cohort identification using natural language processing: an exploratory study. *Ann Allergy Asthma Immunol* 2013; 111: 364–369.
19. Quint JK, Müllerova H, DiSantostefano RL, *et al.* Validation of chronic obstructive pulmonary disease recording in the Clinical Practice Research Datalink (CPRD-GOLD). *BMJ Open* 2014; 4: e005540.
20. Nissen F, Morales DR, Mullerova H, *et al.* Validation of asthma recording in the Clinical Practice Research Datalink (CPRD). *BMJ Open* 2017; 7: e017474.
21. Afzal Z, Schuemie MJ, van Blijderveen JC, *et al.* Improving sensitivity of machine learning methods for automated case identification from free-text electronic medical records. *BMC Med Inform Decis Mak* 2013; 13: 30.
22. Edgren G, Hjalgrim H, Tran TN, *et al.* A population-based binational register for monitoring long-term outcome and possible disease concordance among blood donors and recipients. *Vox Sang* 2006; 91: 316–323.
23. Hansell A, Hollowell J, McNiece R, *et al.* Validity and interpretation of mortality, health service and survey data on COPD and asthma in England. *Eur Respir J* 2003; 21: 279–286.
24. Andersohn F, Suissa S, Garbe E. Use of first- and second-generation cyclooxygenase-2-selective nonsteroidal antiinflammatory drugs and risk of acute myocardial infarction. *Circulation* 2006; 113: 1950–1957.
25. Yang CC, Jick SS, Jick H. Statins and the risk of idiopathic venous thromboembolism. *Br J Clin Pharmacol* 2002; 53: 101–105.
26. Jeffreys M, Redaniel MT, Martin RM. The effect of pre-diagnostic vitamin D supplementation on cancer survival in women: a cohort study within the UK Clinical Practice Research Datalink. *BMC Cancer* 2015; 15: 670.