Target journal: BMJ /Article type: RESEARCH METHODS AND REPORTING

Validation of algorithms in studies based on routinely collected health data: general principles

Vera Ehrenstein, MPH, DSc,[1] Maja Hellfritzsch, MD, PhD,[2] Johnny Kahlert, PhD,[1] Sinéad M Langan, FRCP, MSc, PhD,[3] Hisashi Urushihara, MSc, DrPH,[4] Danica Marinac-Dabic, MD, PhD, MMSc, FISPE,[5] Jennifer L Lund, PhD,[6] Henrik Toft Sørensen, MD, PhD, DMSc, DSc,[1] Eric I Benchimol, MD, PhD[7,8,9]

Corresponding author: Dr. Vera Ehrenstein, Department of Clinical Epidemiology, Department of Clinical Medicine, Aarhus University, Olof Palmes Allé 43-45, Aarhus N, Denmark, ve@clin.au.dk

Author affiliations:

[1] Department of Clinical Epidemiology, Aarhus University, Aarhus, Denmark

[2] Clinical Pharmacology, Pharmacy and Environmental Medicine, Department of Public Health, University of Southern Denmark, Odense, Denmark

[3] Department of Non-communicable Disease Epidemiology, Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, United Kingdom

[4] Division of Drug Development & Regulatory Science Faculty of Pharmacy, Keio University, Tokyo, Japan

[5] Office of Clinical Evidence and Analysis, Center for Devices and Radiological, United States Food and Drug Administration, Silver Spring MD, USA

[6] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill NC, USA

[7] Division of Gastroenterology, Hepatology and Nutrition and Child Health Evaluative Sciences, SickKids Research Institute, The Hospital for Sick Children, Toronto, ON, Canada

[8] Department of Paediatrics and Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada

[9] ICES, Toronto, ON, Canada

29 **Word count,** excluding title page, abstract, references, boxes and figures**: approx. 6,136**

30

Abstract (140 words)

Clinicians, researchers, regulators and other decision-makers are increasingly relying on evidence based on routinely collected health data (RCD). This paper aims to systematise terminology, methods, and practical considerations relevant to the conduct of validation studies of RCD-based algorithms. First, we define algorithms in the context of RCD-based research and summarise measures of algorithm accuracy. Second, we discuss the concepts of gold standard and reference standard. Third, we offer considerations for determining study size, prioritising accuracy measures, and algorithm portability. Finally, we briefly discuss the use of validity data in interpreting results. We use published studies to illustrate all points. In all studies, some degree of information and other biases is inevitable, thus decisions regarding prioritising measures of algorithm accuracy and basis for those decisions should be transparently reported. Validation work should be an ongoing routine task of RCD source maintenance.

Key words

Accuracy; algorithm; data quality, epidemiology; information bias; measurement error; misclassification; observational studies; routinely collected health data; real-world data, real-world evidence, validity

Introduction

Real-world data (RWD) are data originating in the course of routine clinical practice, and the term is used by contrast with data collected in phase I-III interventional trials.[1-3] *Real-world evidence* (RWE) is evidence generated by analysing RWD. RWD is represented to a large extent by *routinely collected health data* (RCD), i.e., data accruing as a by-product of health care delivery, encompassing electronic health records; health administrative data; claims data, records of treatments, procedures, devices, or diseases;[4] and patient-reported events.[5] RCD-based studies are playing an increasingly important role in decisions by regulators, payers, and in health technology assessment,[1-3 7-14] as corroborated by the 2021 publication of the CONSORT-ROUTINE - extension for the reporting of randomised controlled trials conducted using cohorts and routinely collected data.[15] Rapid accumulation of RCD in diverse health care systems, coupled with expanding computing capabilities, predict the ever-strengthening role of RCD in generating evidence on disease prevention, aetiology, epidemiology, clinical course and quality of care, and on utilisation, benefits, risks, and costs of treatments and devices.[18-21]

The advantages of RCD/RWE –include accrual independent of research (and thus free of investigator's preconceptions), comparatively low cost, large number of observations, rapid availability for analysis, analysis-friendly structure, standard coding for many data types – afford precise results from diverse patient populations in routine care settings. An important disadvantage of the RCD is susceptibility to measurement error, introduced along the path from the point of care to a data point in an analytic dataset.[22] In RCD-based studies this causes *information bias*, i.e., systematic error resulting from imperfect correspondence between the true (health) event and its representation in a RCD record, e.g. misclassification. *Random* measurement errors on the level of data accrual translate into *systematic* error on the level of

study result,[23-27] potentially leading to wrong conclusions, and, ultimately, suboptimal clinical practice.[28] Although misclassification is inevitable in any epidemiologic study, regardless of data collection methods,[28] the extent, the nature, and the impact of misclassification for any given study using RCD may be difficult to quantify.[29] Large-scale RCD-based studies, yielding results that are precise but biased are especially troublesome, because large size and associated precision may create a false sense of trust among RWE consumers, if precision is mistaken for validity. Thus, "without taking data quality into account [..] the more the data, the surer we fool ourselves".[33] For RCD-based studies to provide high-quality evidence for clinical practice, data validity must be quantified and the quantitative information incorporated in the analysis and interpretation of study results.[34-36] There have been several comprehensive efforts to validate sources of RCD,[37-43] while guidelines for assessing and reporting validation studies[30] have helped improve reporting transparency.[44-51] Nevertheless, literature on validating RCD remains fragmented,[31 52-55] and scientists have repeatedly raised awareness about the need for better education in the realm of validation studies.[28 56]

This paper aims to systematise terminology, methods, and practical considerations relevant to validation studies of RCD-based algorithms. First, we define algorithms in the context of RCD and summarise measures of algorithm validity. Second, we discuss the concepts of gold standard and reference standard. Third, we consider prioritising validity measures, discuss portability of algorithms, and offer considerations for determining the size of a validation study. Finally, we discuss the role of validation information in interpreting study results. We use published RCD-based studies for illustration. Our intended audience includes RWE originators and consumers in academia, regulatory science, industry and other decision-makers involved in planning, conducting and interpreting RCD-based studies. This effort aligns with the Guidelines for Good

Pharmacoepidemiology Practices (GPP) maintained by the International Society for Pharmacoepidemiology (ISPE);[57] with the data-quality-focussed regulatory strategy of the European Medicines Agency,[58] and with the RWE transparency initiative.[59]

Manuscript development

This paper was conceptualised as a contribution of the ongoing efforts to improve conduct and reporting of observational studies,[3 44-51 60 61] specifically, validation studies in RCD sources.[32 34 52 54 62] The authors are ISPE members with clinical, academic, or regulatory affiliations in Europe, North America, and Asia and with collective expertise in clinical research, epidemiology, RCD, validation studies, and development of reporting guidelines.[30 47 48 50-52 63] Manuscript drafts were circulated among the authors and discussed via teleconferences or e-mail. At the face-to-face meeting at the ISPE Annual Meeting in Philadelphia, USA (25 August 2019), the authors provided input on the scope, structure, and terminology of the manuscript. After several subsequent drafting rounds, the manuscript was circulated to the ISPE membership for feedback and subsequently revised accordingly. The ISPE membership includes experts from clinical, academic, and industry sectors from the Americas, Africa, Asia, Oceania, and Europe.

Terminology

Throughout this paper, by validity we mean internal validity. External validity (generalisability) of study results is not a topic of this paper, except when portability of algorithms is discussed. Box 1 provides the glossary of the main terms used in this paper. Terminology may differ slightly from study to study; our aim was to use general, all-inclusive terms, while maintaining consistency with previous studies and the current guideline documents.[57] Following an earlier study on RWD validation,[30] we use the term *health state* for any event of interest measured in a

study. Depending on study aims, a given health state will take on a role of exposure (e.g., medicinal or surgical treatment or device whose safety or effectiveness is of interest), outcome/endpoint (e.g., treatment adverse event, disease relapse, mortality), or a covariate/confounder/subgroup (e.g., comorbidity).[30] For example, the RCD record of an antidiabetics dispensing may define exposure groups "antidiabetic treatment" in one study, the endpoint "diabetes onset" in another, and a "comorbid diabetes" indicator in the third.

BOX 1 HERE

Algorithms in routinely collected data

An *algorithm* is "a completely defined set of operations that will produce the desired outcome".[64] In RCD-based studies, algorithms are operational definitions of health states, used to classify persons with respect to presence/absence and attributes of those health states (e.g., illness and its severity or treatment and its duration). RCD-based research relies on algorithms to define inclusion and exclusion criteria; treatment/exposure status, duration, intensity; date of onset of a disease or a behaviour; and covariates for inclusion in the analyses. An individual's RCD-based record, and, by extension, any algorithm based on that record, is a product of the unique chain of objective and subjective events and practices. These include patient-level events and behaviours, such as symptoms and care-seeking thresholds (which may vary by sex, age, education, income);[65 66] characteristics of the health care system (e.g., universal vs. insurance-based coverage); referral patterns (e.g., whether or not specialist care is general-practitioner-triaged); clinical aspects (e.g., diagnostic process and treatment decisions); and disease outcome (e.g., fatality).[29] RCD-based algorithms depend on record-keeping systems and the specific purpose of RCD generation (e.g., health administrative data vs. discharge summaries vs. general-practitioner clinical notes); health sector feeding to a RCD source (e.g., primary vs. hospital care); level of

detail allowable by RCD (e.g., type and version of classification used to code diagnoses (SNOMED, International Classification of Diseases [ICD], 9th Revision vs. ICD-10 vs. -10 vs. ICD-10-CM [clinical modification] vs. Read codes); and database maintenance routines (e.g., frequency of updates, plausibility checks).[30 31] Each link in the chain of events from point of care to a research data point harbours a potential source of error (Figure 1).[22]

FIGURE 1 HERE

RCD-based algorithms are analogous to diagnostic tests in patient care: both must accurately classify – "diagnose" – individuals with respect to presence or absence of a given health state at a given point in time.[30] Similar to the diagnostic process, RCD-based algorithms may be based on rules (e.g., biomarker value cut-offs), on formal guidelines (e.g., result of a given diagnostic test), or on clinical features (e.g., a set of signs and symptoms).[67] An algorithm for a given health state may be as simple as a search for a single diagnostic code during a hospitalisation or as complex as a decision tree with time-dependent variables measured across multiple data sources.[31 67 68] A hospital discharge summary carrying the ICD-10 code I21 "Acute myocardial infarction" is an example of a simple algorithm that can be used to identify patients with acute myocardial infarction, albeit recorded posthoc. The subcodes may be used to classify patients according to the affected area of the heart muscle (I21.1 "Acute transmural myocardial infarction of inferior wall"). Figure 2 shows a comparison of incidence rates of pertussis, identified in three European countries using various candidate RWD-based algorithms and as defined by the European surveillance based on a standard case definition.

FIGURE 2 HERE

An example of a complex algorithm is a decision-tree based semi-automated procedure that establishes drug treatment start, duration and continuation based on patients' age, sex, purchasing history, standard packaging information, and expert input.[69] Machine learning, increasingly used to aid development of RCD algorithms, is valuable as a screening tool for identifying candidate algorithm components for subsequent validation.[31 67 68 70 71] The raised concepts apply regardless of algorithm-generating mechanism, be they static or dynamic, machine-leaning or human expertise based. Box 2 provides examples of RCD-based algorithms and of studies assessing their validity.

BOX 2 HERE

Researchers planning an RCD-based study should consider whether candidate algorithms are suitable for use in a given RCD source, based on the data source completeness and algorithm validity (discussed below). Meaningful development and application of RCD-based algorithms may require input from clinicians, database administrators, epidemiologists, statisticians, and data custodians, collectively familiar with local coding, referral and record-keeping practices, and overall data flow. Figure 3 exemplifies considerations regarding whether and to what extent to rely on RCD in a given study. For an RCD source be deemed suitable for a given study question, the investigator needs to ensure validity and reliability of the algorithms that will be used to identify study-specific health states. The investigator needs to consider whether a validated algorithm is available for the RCD source and if so, whether it can be used given the study question (see Prioritising validity measures); whether it applies to a given population and the study period (e.g., whether validity estimated for ICD-9 codes holds for ICD-10 codes). If a validated algorithm is unavailable, should the validation be performed on all or a sample of health states observed in the study at hand (internal validation) or should a separate – external –

validation effort be undertaken. Further decisions relate to the choice of gold/reference standards (described below), logistics of the validation, and incorporating algorithm validity metrics in study results and interpretation (Figure 3).

FIGURE 3 HERE

Completeness of a data source

Completeness of a data source is the proportion of all events of the study-relevant health states in the study target population captured in that data source.[32] For example, before planning a RCD-based study of safety of oral non-steroidal anti-inflammatory drugs (NSAIDs) using data on outpatient dispensings from the Danish National Prescription Registry, we need to understand the registry's completeness in capturing the NSAIDs dispensings in the population. In Denmark, the proportion of by-prescription dispensings is 66% for ibuprofen and 100% for diclofenac,[72] corresponding to the registry's completeness with respect to each drug, as it captures only by-prescription dispensings.[73] Completeness of an RCD source may be affected by changes in health policy (e.g., from <100% to 100% for diclofenac dispensings in the Swedish Prescribed Drug Register[74] as Sweden banned over-the-counter sales following a cardiovascular safety concern[75]). Completeness is also affected by changes in diagnostic process (e.g., introduction of troponin as the main diagnostic biomarker of myocardial infarction[65]); or by introduction of screening (e.g., screening for colorectal cancer will lead to an increase over time in the number of persons with a diagnosed colorectal cancer recorded in cancer registries). When planning RCD-based studies, completeness of candidate data sources needs to be assessed with respect to health states representing all study variables – exposures, outcomes, covariates, and subgroup identifiers.

Measures of algorithm validity

The RCD algorithms/diagnostic test analogy extends to the measures of validity.[30 76] These measures, covered in standard texts and tutorial articles,[20 21 77 78] are summarised in Box 3, for reference. Algorithm validity measures include sensitivity and specificity, and their derivatives (receiver operating characteristic (ROC)-curve, area under the curve (AUC)[79], diagnostic odds ratio (DOR)[80]); indicators of performance (positive and negative predictive values (PPV and NPV)); and measures of agreement (kappa statistic[77 81]). (Other sources have used, the term 'bias parameters'[56] or validity indices[78] to collectively describe the measures of algorithm validity). When a health state is rare, both specificity and NPV of an algorithm are less useful because they are expected to be close to 100%. Sensitivity and specificity of algorithms are generally, although not always, independent of prevalence of the health state in the underlying population (equivalent of pre-test probability for a diagnostic test). Crucially, both PPV (equivalent of post-test probability for a diagnostic test) and NPV depend on the prevalence of the health state of interest in the underlying population (PPV directly proportional, NPV inversely proportional to the prevalence).[79] A PPV estimated in a validation cohort in which the health state of interest is more prevalent than in the underlying source population, will be overly 'optimistic', meaning that the PPV applied to the data source will be lower. Simple calculations allow derivation of PPV and NPV from known sensitivity, specificity, and prevalence of the health state in a population.[30 56] Many validation studies are not designed to estimate all validity measures. Sensitivity and PPV are the most commonly reported measures[30 82] because estimation of specificity and NPV requires identification of a representative group of true-negative individuals, which may be challenging with routinely collected data due to requirement of exhausting labour.

BOX 3 HERE

Gold standard and reference standard

Sensitivity, specificity, PPV or NPV of an algorithm are estimated against a gold standard. Conceptually, a gold standard is a method that classifies individuals with respect to presence and absence of a given health state without errors, i.e., there are no false-positives and no false-negatives. As a true gold standard rarely exists, in practice validation studies rely on a "reference standard" or an "alloyed gold standard", i.e., a classification method inferior to the theoretical (but often non-existent) gold standard, but superior to the RCD algorithm being validated (Box 2 provides examples of gold and reference standards).[77] For the health state under validation, the reference standard provides case and non-case definitions, against which to compute the measures of validity and performance (Box 3). Figure 4 shows relations between the gold standard, the reference standard, and the algorithm. Setoguchi *et al.*, in a validation of haematological malignancies defined from health administrative data against a cancer registry as the reference standard illustrated dependence of validity measures on the completeness of the reference standard.[83]

FIGURE 4 HERE

Medical records are commonly used as a gold/reference standard in validation studies of RCD-based algorithms. When planning a validation study based on review of medical records, points to consider include:

- Look-back and look-forward period for record review relative to the date of the potential health state identified by the RCD algorithm, e.g., only information in 3 months before and after the RCD-recorded event will be considered in validation, to allow for data flow artefacts, e.g., delay of diagnosis recording relative to a true diagnosis date in administrative data;

- What types of key records should be sought to confirm the health state (e.g., glycated haemoglobin for diabetes, liver enzyme measurements for hepatotoxicity, bone scintigraphy for bone metastases etc.);
- Whether only records from specific contacts are of interest for validation, e.g., at specific clinical departments (see the osteonecrosis of the jaw example, in the description to Figure 3);
- Whether to flag prevalent vs. incident status of the health events under validation;
- Whether chart abstraction and case adjudication processes should or can be combined – e.g., conducted by an expert adjudicator or should a nurse conduct the abstraction and should the health state be subsequently adjudicated by an expert? De-coupling chart abstraction and adjudication may be necessary, for example, if adjudicators need to be blinded to certain information, such as treatment status in validating endpoints in safety assessments;
- How to quantify to resolve disagreements and quantify interrater variability, if chart abstraction/adjudication is done by more than one person.[30 81]
- Design of data collection instruments such as chart abstraction forms and software: in what order should the information be abstracted?
- What are the mechanisms for correcting entries if needed or unclear? Is it possible and if so, for how long, to revisit the abstraction source to edit information?
- Staff skills and potential need for study-specific training.

A pilot medical record review ahead of a large-scale effort will inform the process and help use resources efficiently.

Medical records must be viewed especially critically when used as the reference standard for adjudicating RCD-identified endpoints in safety studies of medicines, where the initial safety data originate from preapproval trials. Justifiably, a researcher may want to have the same case definitions for trial- and RCD-based events. However, the components of case definition available for closely monitored participants in clinical trials may be under-recorded or not recorded in the course of routine clinical care, especially if events are asymptomatic or insufficiently severe to trigger diagnostic activity or care seeking. For example, if QT prolongation is a safety concern in a prospective trial, patients' electrocardiogram (ECG) would be taken at baseline and monitored during the entire follow-up, regardless of severity or symptoms. In routine clinical practice, physicians will detect and document ECG results predominantly on symptomatic care-seeking patients, and generally not have information on the equivalent of the baseline ECG to gauge whether QT- prolongation is pre-existing or newly occurring (Berkson's bias[84]). In this case, the medical record is likely to have low sensitivity, and trial case definitions may need to be modified to the type of information typically available from medical records.[85] Furthermore, if physicians are on a higher alert to adverse events with new treatments, differential misclassification may ensue.

Another validation option in the absence of a suitable gold/reference standard is use of latent class modelling, whereby different RCD-based algorithms for a given health state are assigned probabilities of correctly classifying patients with respect to the health state status, through data-driven modelling. Prosser and colleagues used latent class modelling to validate three RCD-based algorithms to identify treated asthma in administrative data in Canada, with each algorithm based on a combination of frequency of physician visits and hospitalisation carrying qualifying diagnostic codes.[86] In addition to medical records, patient or physician interviews, or another

database are examples of gold/reference standards used in RCD validation studies (examples in Box 2).[30 31 77]

What measures of validity can be estimated in a validation study?

Types of validity measures estimable in a given validation study depend on the strategy used to assemble the study population, which typically falls into three scenarios.[56] In the first scenario, study population is assembled based on the RCD-algorithm defined health status – algorithm-positive and algorithm-negative are sampled from an RCD source and their RCD-record is compared with the gold-standard based case definition (e.g., medical records). This scenario allows estimation of PPV and NPV. In the second scenario, the study population consists of persons who are health-status positive and health-status negative according to the gold standard. Their RCD records are searched for the elements of the proposed RCD-based algorithm. This scenario enables estimation of sensitivity and specificity. In the third scenario, study population is assembled independent of either RCD- or reference-standard based status and allows estimation of sensitivity, specificity, PPV, and NPV (see Fox et al for a detailed discussion[56]). If RCD algorithms are assessed in data sources, none of which is superior, kappa statistic may be the only estimable metric. Bollaerts et al. derived analytical expressions connecting the observed (algorithm-based) prevalence of a health state and its four validity measures (sensitivity, specificity, PPV, and NPV). These expressions, and an associated web-based application allows derivation of unknown validity measures based on the observed prevalence and any two other measures.[78] A worked example of this application is provided in a study of Morkem et al.[87]

Prioritising validity measures

Ideally, we want to maximise all measures of algorithm validity, however, practically we must accept trade-offs between the measures: increasing algorithm sensitivity (allowing it to capturing more true-positive cases) is accompanied by a decreasing specificity (by capturing more false-positive cases) and vice versa.[64 83] The impact of imperfect validity measures on study findings depends on the role of an RCD-defined health state in a given analysis: inclusion/exclusion criterion, exposure, outcome, covariate, as illustrated in several examples, below.

RCD algorithms for eligibility criteria

Consider an observational study of comparative effectiveness and safety of oral anticoagulants among patients with atrial fibrillation (AF) as an example. Suppose we want to conduct this study using RCD from Scandinavian nationwide registries, whereby data on diagnoses originate from hospital contacts, and data on treatment, from outpatient dispensings.[88] In addition to AF, indications of oral anticoagulants include post-arthroplasty thromboprophylaxis and treatment of recurrent venous thromboembolism.[66] Patients' characteristics, treatment dosing and duration vary by indication. Therefore, we wish to ensure that the study population includes only patients treated for AF. As routine dispensing records rarely contain information on indication or prescribed dose, the RCD-based eligibility criteria must indirectly identify patients treated for AF. A drug utilisation study showed that AF is recorded among nearly 80% of patients who initiate anticoagulant.[66] Thus, we could first identify drug initiators and exclude patients with a record of arthroplasty or venous thromboembolism and assume that all remaining patients have AF.[89] A limitation of this approach is that some remaining patients will not have AF. An approach that alleviates, but not fully resolves this limitation, would require a hospital diagnosis of AF as an inclusion criterion: this will ensure that all patients have AF, but not that they are treated for it with anticoagulants. To further increase specificity of identifying the eligible

355 population, we could exclude from that population patients with a hospital diagnosis of venous
356 thromboembolism or a record of arthroplasty. Although this exclusion criterion will remove
357 some eligible patients, we may decide to pay this price for the sake of obtaining a valid
358 estimation of comparative effectiveness and safety. The measures we take to increase the
359 specificity of the eligibility criteria algorithm will come at a price of potential selection bias, if
360 patients with hospital-diagnosed AF systematically differ from patients with AF initiating
361 anticoagulants seen in all medical care settings.[90] Patients seen in hospitals are likely to represent
362 the most severe spectrum of the disease; this bias would be avoided by the sensitive eligibility
363 algorithm, described above. Selection bias due to inclusion of patients with other indications
364 could be avoided by keeping such patients in the study population and conducting a stratified
365 analysis. With rare diseases, we may prioritise eligibility criteria algorithms with high PPV and
366 specificity, even at a price of reduced precision, to avoid "contamination" of the study cohort
367 with false-positives; alternatively one may opt for a hybrid approach by identifying potentially
368 eligible patients based on RCD algorithms and subsequently eliminate false-positive cases by
369 reviewing medical records.[64]
370
371 *RCD algorithms for exposure*
372 RCD-based algorithms establishing patients' treatment/drug exposure rely on routine records of
373 prescriptions (e.g., in EHR data), dispensing (e.g., in pharmacy administrative data), or
374 administrations (e.g., in hospitalization data) to measure treatment initiation, duration, dosage,
375 discontinuation. In identifying new users of a drug, the length of the washout (treatment-free)
376 period may produce severe misclassification between the status of new and prevalent user, which
377 may lead to underestimation of risks of early side effects.[91] Incorrect allocation of on-/off-
378 treatment person-time will likewise bias studies of comparative effectiveness. Although routine

prescription and dispensing records are generally considered to be high quality, their correspondence to the true treatment status/adherence is difficult to validate. Therefore RWE on medication safety and effectiveness often include multiple sensitivity analyses to assess robustness of studies against different RCD-based algorithms and their assumptions.[51] Information on treatment absence may be of interest in studies aiming to explore whether some patients who may benefit from treatment remain untreated. If an RCD source has incomplete treatment records, there is likely a mixture of true and false negatives amongst the cohort of patients without an RCD-based treatment record. If under-treatment is of interest, the NPV of the RCD-based treatment record should be prioritised. In addition, whether the effect of exposure is prolonged for a predetermined period after one-time exposure to a drug or limited during the period when a patient is actually exposed to a drug should be determined considering the characteristics of the drug for each outcome health state.

RCD algorithms for outcomes

When RCD-based algorithms are used to define events to estimate their absolute risk or risk difference, such as adverse treatment effects, an algorithm with low sensitivity will lead to underestimation of the true risk, potentially providing false reassurance. However, the corresponding relative effect (risk ratio) will be unbiased no matter how low the sensitivity, provided a near-perfect specificity. This consideration is only true for a binary outcome. The same consideration will apply in studies of disease epidemiology for the purposes of planning health services and in studies comparing disease populations with different severity or among different medical-care settings. For example, an RCD-algorithm based study on epidemiology of diabetes when only secondary-care diagnoses are available (e.g., in Sweden or Denmark[92 93]) will underestimate the incidence and prevalence of diabetes (diagnostic delay) and overestimate the

average disease severity. Moreover, the time of disease onset will be incorrect for patients with diabetes seen in hospitals after being diagnosed in primary care. In the case of diabetes or other diseases with specific treatments, sensitivity of identification can be increased by use of treatment proxies or laboratory values.

RCD algorithms for confounders

Accurate RCD-based algorithms are important for confounding control in observational studies. Poorly measured confounding variables perform poorly in controlling for confounding and yield biased results. For example, obesity is an important confounder in many associations, yet many RCD sources do not capture obesity or body mass index with sufficient accuracy or completeness. For example, the ICD-10 based algorithm for overweight or obesity in the Danish National Patient Registry has a PPV close to 90%, but the registry is merely 11% complete in capturing patients with obesity.[94] If obesity is an important confounder, analyses adjusted for obesity using that algorithm will remain largely confounded in the analyses based on that RWD source, and researcher should consider using an external adjustment to evaluate whether some or all of the observed association is likely to be explained by the unmeasured confounding.[95]

In summary, decisions about prioritising validity measures imply balancing the risks and benefits of including false-positives vs. false-negatives that should be considered given the study aims.

Portability of algorithms

Variability of algorithm validity within and between RCD sources is likely to depend on several factors including patients' age (paediatric vs adult population), calendar time (changes of guidelines, diagnostics, classifications, referral patterns), disease severity (hospitalisation

patterns), health care funding scheme (single-payer vs. multiple-payers), accessibility (universal vs. choice-based), health care sector contributing to RCD (general practice vs. hospital care vs. both), and coding practices rooted from insurance system and policy, and changes in coding dictionaries and supporting IT infrastracture.[67] Therefore, algorithms validated in one setting (country, RCD source) cannot be assumed to have the same validity in a different setting (country, RCD source). As illustrated in the pertussis infection example (Figure 2), in multi-database studies, assuming that an algorithm is transportable may not be tenable, and must be critically assessed given the differences of the underlying health care systems, guidelines, and mechanism of RCD record generation.[96-100] In preparation for studies of effectiveness of the vaccines against the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Europe, background incidence rates of multiple potential adverse events have been estimated in using multiple and heterogeneous RWD sources in several European countries.[101] The variability of the observed background rates estimated based on common code lists underscores the importance of algorithm calibration and benchmarking using available evidence, and features of a given RWD source. These background rates should remain country-specific and are meant for detecting signals for subsequent formal evaluation. Differences in observed rates across multiple databases, in a process similar to triangulation,[102] may be used to glean the direction and the magnitude of measurement error associated with a given RWD source.

On the other hand, because of cost and effort associated with conducting a validation study, before embarking on one, researchers should search the relevant literature on whether a validated algorithms are already available for their given study aim (Figure 3). At the same time, with strategic investments in interoperability, creation of searchable and updated algorithm libraries, alignments and partnership consortia between the countries and relevant data sources the burden

associated with conducting validation studies can be dramatically reduced. Approaches to ensuring applicability range from calibration-refinement exercises[98] to full-scale re-validation in the new setting (Figure 3).[103]

Another aspect of algorithm portability is whether incident and/or prevalent events are of interest. This distinction is important in safety studies of association between a treatment and a side effect that is a chronic disease, whereby only incident events, i.e., events with onset following the medication initiation should be counted as a case. Such case definition would require the gold/reference standard to afford a sufficiently long lookback period to rule out prevalent conditions. For example, some patients with inflammatory bowel disease may have more than 8 years without disease-related records in administrative data,[104] suggesting that a lookback period of 1 year, commonly used in studies based on insurance claims, may not be sufficient to exclude prevalent cases.

Size of a validation study

If the aim of a validation study is to estimate a dichotomous measure of validity (Box 3), the number of the validated records determines precision of the resulting estimates. When the target validity measure is dichotomous, size of the study becomes the consideration of precision of a dichotomous proportion.[79 105] Precision considerations must be balanced against considerations of available human and financial resources and must account for the expected nonresponse rate (e.g., if medical records are unavailable). Further considerations include whether variability of the validity measures is expected to vary by subgroups, such as calendar period (e.g., record recency may determine classification version, diagnostic validity, treatment guidelines, referral or reimbursement practices; classifications); disease subtype or severity (e.g., subcodes); priority

and setting of the diagnosis (e.g., primary vs secondary, inpatient vs outpatient); or patients' characteristics (e.g., sex or comorbidity). If estimation of the validity measures is important in a given subgroup, the number of validated records should allow for the desired subgroup precision. For planning validation studies of complex algorithms, researchers should consider whether validity measures would be estimated for the overall algorithm or for each algorithm component (e.g., a group of diagnostic codes vs. a single diagnostic code), with potential subsequent algorithm refinement, such as removal of diagnostic codes with low validity.[106] The study size should include sufficient number of observations with a given algorithm component. For internal validation, the proportion of records selected for validation may depend on whether the condition is rare or common.[107] An adaptive validation study design using a Bayesian approach has been recently proposed for internal validation studies aiming to estimate PPV and NPV. The validation information accrues as the study is being conducted. The design defines decision rules for when the amount of validation accrued is sufficient to provide reliable inference, whereupon validation can be stopped.[108] The increasing availability of electronic medical charts in combination with machine learning and text mining may help undertake validation studies on larger scale in the future.

Interpreting study results in the light of information bias

Researchers may be tempted to dismiss misclassification bias (e.g., of exposure status) on the grounds that the obtained estimates are "conservative" i.e., that the true association is speculated to be (even) stronger than the one observed. In a safety study of a treatment, procedure, or device, nondifferential misclassification of exposure may mask an important safety signal. A true association masked by misclassification may provide false assurance about treatment safety, while a spurious association when none exists may cause unnecessary treatment withdrawal.

More generally the "nondifferential misclassification mantra"[109] refers to a statistical expectation under a narrow set of conditions (in which the exposure has an effect, the health state is dichotomous, and misclassification is non-differential).[24 25] For polytomous variables created by categorising continuous variables, the direction of measurement error depends on a category.[110] Finally, in RCD, all variables with rare exceptions, are measured with error, and estimation of joint effect of errors in all study variables cannot be assumed to be in a given direction without quantitative assessment.[2828 109 111]

If a validation study reveals that an RCD-based algorithm's validity measures are unsuitable given study aims and its role in the analysis, an alternative approach to identifying that health state should be considered (e.g., de-novo data collection) (Figure 3). If one proceeds with the use of an RCD-based algorithm, the potential impact of information bias on the interpretation of study findings should be considered and, ideally, quantified. Methods and software for such quantification[82 109 112-116] range from simple calculations examining the impact of misclassification of one variable at a time[113] to simultaneous assessment of the impact of errors of multiple study variables (bias analysis)[82 112] to using imputation methods to correct for misclassification.[117] The latter methods can be used to estimate the range of study results that could be observed under plausible ranges of validity measures or allow re-computation of what study result would have been expected had the measurement been perfect or better than the one available.[109] If sensitivity and specificity parameters are unknown, these methods can be used to evaluate the impact of a plausible range of different sets of sensitivities and specificities on study results. Predictive values can be used to conduct quantitative bias analysis to correct for misclassification of outcome variables and a research has shown that imperfect predictive values impact study results ranging from negligible to significant extent, producing incorrect

conclusions (e.g., misclassified data show presence of an exposure – outcome association, but corrected data do not).[118] Other techniques for correcting study results for measurement error include hierarchical semi-Bayes methods,[114] or bootstrap imputation.[115]

Interpretation of study findings warrants quantifying the amount and the potential impact of information bias due to imperfect algorithm validity and especially discuss and quantify plausible scenarios that are likely to invalidate study conclusions. Despite extensive theoretical developments and practical applications and examples[109 111 119] measurement error, especially differential and non-independent errors, remain neglected in interpreting study findings, and formal analysis of potential impact of measurement error in interpreting results remains an exception, rather than a rule.[28]

Summary

As the RCD landscape evolves, ensuring, quantifying, and updating algorithm validity information should become a standard maintenance task of all RCD-based research efforts. Information on algorithm validity needs to reflect up-to-date clinical knowledge, diagnostic accuracy, clinical guidelines, regulatory practice, insurance policies, coding conventions, granularity and completeness of recording, and data flow.[34 120-124] As a starting point for developing protocols and implementation of validation studies of RCD-based algorithms, we recommend use of the published checklist of reporting criteria for studies validating RCD-based algorithms, spelling out practical elements of conducting the validation study.[30] Formalising the conduct of validation studies will help improve quality of evidence of RCD-based research and foster trust between RCD researchers and research consumers, such as policy-makers and clinicians. We posit that if RCD-based research delivers valid evidence, accurate clinical and

administrative record-keeping at the point of health care delivery will be viewed by record-keepers as an investment in the quality of care rather than a bureaucratic task. In the words of the FDA Commissioner, "To enable greater adoption of RWE in clinical and regulatory decisions, we'll need to work with the healthcare system to change the way clinical information is collected. Ideally, we'd like to have a system where providers have the right incentives to enter clinically relevant information into EMRs at the point of care".[125] Despite their advantages, RCD cannot be a default solution to any research question. For example, for studies of rare diseases with centralised clinical management at a few treatment centres, a study that uses primary data collection, alone or in combination with RCD, may be more efficient and valid than a purely RCD-based study, as using RCD may be like seeking a needle in a haystack because of the dependence of some validity measures on diseases prevalence. Limited information on confounders is another potentially important limitation of RCD.

The principles described here are broadly applicable for all studies relying on routinely collected health data, including observational studies of diseases, benefit-risk assessments of interventions, as well as pragmatic randomised trials harnessing routinely collected health data to measure patients' characteristics, risk factors, and/or outcomes in routine clinical practice. As RCD are being increasingly used to make important clinical, regulatory and policy decisions about drugs, procedures, devices and other healthcare interventions, we hope to have raised awareness about the importance of quantifying measurement error around RCD-based algorithms when generating real-world evidence.

Funding

Funding to support this manuscript development was provided by the International Society for Pharmacoepidemiology (ISPE). A draft of the manuscript was made available for review to the ISPE members and subsequently revised in response to that review.

Acknowledgements

This manuscript is endorsed by the International Society for Pharmacoepidemiology (ISPE). We gratefully acknowledge review of the manuscript by ISPE members. We thank Ms. Helle Vester, MSc, for excellent administrative support.

Disclosures

Dr. Ehrenstein is a salaried employee of Aarhus University, which receives institutional research funding from various pharmaceutical companies and regulatory agencies, administered by Aarhus University.

Dr. Hellfritzsch has nothing to disclose.

Dr. Kahlert is a salaried employee of Aarhus University; institutional research funding from various pharmaceutical companies and regulatory agencies to and administered by Aarhus University.

Dr. Langan was supported by a Wellcome Senior Research Fellowship in Clinical Science (205039/Z/16/Z). Dr. Langan was also supported by Health Data Research UK (grant No. LOND1), which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust.

Dr. Urushihara is a salaried employee of Keio University, and has received research support from CAC Croit Corporation, Shionogi & Co., Ltd., and Senju Pharmaceutical Co., Ltd.

Dr. Marinac-Dabic has nothing to disclose.

Dr. Lund receives research support from AbbVie, Inc. and her spouse is a full-time, paid employee of GlaxoSmithKline.

Dr. Sørensen is a salaried employee of Aarhus University; institutional research funding from various pharmaceutical companies and regulatory agencies to and administered by Aarhus University.

Dr. Benchimol was supported by a New Investigator Award from the Canadian Institutes of Health Research, Canadian Association of Gastroenterology and Crohn's and Colitis Canada. Dr. Benchimol was also supported by the Career Enhancement Program of the Canadian Child Health Clinician Scientist Program. Dr. Benchimol has acted as a legal consultant for Hoffman La-Roche Limited and Peabody & Arnold LLP, has received consulting fees from McKesson Canada, and receives research funding from Crohn's and Colitis Canada for matters unrelated to this article.

609 **List of Boxes**

Box 1. Glossary of key terms

Term	Definition/elaboration
Algorithm	An operational definition of a *health state*, used to classify persons with respect to presence/absence and attributes of those health states. An algorithm may be as simple as a single diagnostic code or as complex as a set of decision rules. Implicitly, persons not meeting algorithm criteria are classified in the analysis as not having a given health state.
Alloyed gold standard	See *Reference standard*
Area under the curve (AUC)	Area under the ROC curve, used as a single measure of the discriminative performance of an algorithm based on cut-off values of a continuous variable. See Box 3 for details.
Completeness	Completeness of a data source is the proportion of all events of the study-relevant health states in the study target population captured in that data source.
Confounder	A confounder can be conceptualized as a common cause of the exposure status (eg treatment decisions) and the outcome development. Thus, a confounder cannot be a causal intermediate between the exposure and the outcome. In a given study, a characteristic that fulfils the above criteria will exert confounding if it is 1) unequally distributed across exposure categories AND b) predicts the study outcome. Confounder is a special case of a covariate.
Covariate	A characteristic of the members of the study population besides the exposure and the outcome that may be of interest. This term is often used for predictors of the outcome of interest. By contrast with a confounder, covariate is need not be associated the exposure of interest.
External validation	Validation of events of a health state identified with an RWD-based algorithm in a population external to a given study.
Diagnostic odds ratio (DOR)	The ratio of the odds of algorithm positivity in those with a health state of interest to the odds of algorithm positivity in those without the health state of interest.
Gold standard	In an ideal sense, a method that classifies individuals with respect to a health state of interest without errors. True gold standards rarely exist (see *Reference standard*), as no method of measurement is error-free. In practice, the quality of gold standards may be time-dependent. For example, new diagnostic tests may, paradoxically, perform better than methods previously considered gold standards.
Health state	A generic term used in this paper for any event of interest measured in a study. Depending on study aims, the same health state may play a role of exposure, outcome, covariate, or a subgroup indicator.
Information bias	Discrepancy between the true value of a given trait or characteristic (height, weight, disease status) and its measured value. Errors may be inherent in a measurement instrument (eg biased scale used to measure weight) or be human (eg data entry errors). *Information bias* is often used synonymously with *Measurement error* and *Misclassification*.
Internal validation	Validation of events of a health state identified with an RWD-based algorithm in a given study.
Measurement error	See Information bias.
Misclassification	Incorrect classification by an *algorithm* of an individual's health state, resulting from *Measurement error* (eg, a patient with a body mass index in a

	normal range may be classified as obese as a result of a measurement error of their height or weight).
Negative predictive value (NPV)	Proportion of patients who truly negative for a health state among all those who are classified as negative by an algorithm. See Box 3 for details.
Positive predictive value (PPV)	Proportion of patients who truly positive for a health state among all those who are classified as positive by an algorithm. See Box 3 for details.
Real-world data (RWD)	Data that originate in the course of routine clinical practice, usually by contrast with data originating in phase I-III clinical trials.
Real-world evidence (RWE)	Evidence generated using RWD.
Reference standard	A method or a data source that is expected to classify individuals with respect to a health state of interest better than the algorithm being validated. Alternative term: alloyed gold standard.
Receiver operating characteristic (ROC) curve	For algorithms that classify patients into health states (eg., presence/absence/severity) a plot of sensitivity (or true positive proportion) against 1-specificity (or false-positive proportion), Box 2.
Routinely collected health data (RCD)	Data that accrue routinely as a by-product of health care delivery or administration of the health care system. Routinely collected data may also include non-health characteristics such as income, education, and employment, if important in studying health states. RCD are a type of RWD.
Sensitivity	Proportion of persons with a given health state according to gold/reference standard who are classified as positive by an (RWD) algorithm. See Box 2 for details.
Specificity	Proportion of persons without a given health state according to gold/reference standard who are classified as negative by an (RWD) algorithm. See Box 2 for details.
Exposure	A health state that, in a given study, explicitly or implicitly, defines patients' characteristic that is equivalent to treatment allocation in randomized trials. In an RWD -based study, exposure status may correspond to treatment status (medicinal, surgical, device), but may also correspond to a patient's characteristic (comorbidity, age, sex, socioeconomic status). Conceptually, exposure is the independent variable.
Outcome	A health state that, in a given study, plays a role of an outcome of interest. In an RWD -based study, it could be any outcome of treatment (benefit or risk), disease recurrence (e.g., recurrent malignancy, readmission for a myocardial infarction), or death. Conceptually, exposure is the dependent variable.
Validation cohort	Study population used to evaluate measures of algorithm validity.
Validity (internal)	Correspondence between an estimate and the true value of a parameter. In epidemiologic studies, a valid (unbiased) estimate may estimate occurrence of a health event, or an association between exposure and outcome.

610

Box 2. Examples of RCD-based algorithms and validation studies of RCD-based algorithms conducted with different types of gold/reference standard

Health state or event being validated	Country	RCD source(s)	RCD-based algorithm	Gold/reference standard	Validation study
In-hospital chemotherapy treatment	Denmark	Danish National Patient Registry	Agent-specific treatment codes	In-hospital pharmacy production system and chart review	By comparing RCD data with the reference standards, and vice-versa, for 50 randomly selected patients with colorectal cancer having nodal involvement, the kappa, sensitivity, specificity, positive and negative predictive values were estimated for the RCD-based algorithm for receipt of any chemotherapy and for specific treatments (Lund et al 2013[126])
Start, duration and end of prescribed medicine	Finland	Finnish Prescription Registry	Different approaches to estimation of drug use periods based on prescription data. Three fixed methods were included and one data-driven method (PRE2DUP)	Expert review of purchase history	For 200 randomly selected purchase histories of warfarin, bisoprolol, simvastatin, risperidone and mirtazapine in patients with Alzheimer's disease, two experts reviewed purchase histories and evaluated the RCD-based algorithms with regards to their ability to provide purchases and duration. (Tanskanen A et al 2017[127])
Cardiac interventions	Denmark	Danish National Patient Registry	Surgical and procedure codes according to the NSCP classification	Chart review	Charts from randomly sampled patients identified using the RCD-based algorithm were reviewed and positive predictive values for each intervention were calculated (Adelborg et al 2016[128]).
Osteonecrosis of the jaw	Denmark	Danish National Patient Registry	ICD-10 codes recorded at departments of oral and maxillofacial surgery in cancer patients	Chart review and an independent sample of confirmed cases	By assessing the occurrence of osteonecrosis of the jaw according to the chart among patients with a positive algorithm, the positive predictive value of the RCD-based algorithm was obtained. Sensitivity of the algorithm was estimated as the proportion of confirmed cases with a positive algorithm (Ehrenstein et al 2015[129])
Colorectal cancer recurrence	Denmark	Danish National Patient Registry and the Danish Pathology Registry	Combination of codes compatible with metastatic disease, chemotherapy administration and cancer recurrence in patients previously diagnosed with non-metastatic colorectal cancer	Independent sample of patients with known recurrence status	By comparing the algorithm-based and the true recurrence status, sensitivity, specificity, positive and negative predictive values were calculated (Lash et al 2015[130]).
Frailty in the elderly	United States	Administrative claims	Medicare claims-based algorithm of dependency in activities of daily living (or dependency) developed as a proxy for frailty	A reference standard measure of phenotypic frailty	We examined the discriminative ability of the claims-based algorithm to predict phenotypic frailty using a receiver operating characteristic (ROC) curve and estimated the sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) of the claims-based algorithm relative to the reference standard using a range of dichotomous cut-off values (Cuthbertson et al 2018[131]).

Hepatic encephalopathy	United States	Administrative claims	ICD-10 codes and treatment proxies after switch to ICD-10 in 2015 rendered algorithms validated using ICD-9 invalid	Independent prospectively acquired sample containing both true positive and true negative cases, with case status established by diagnostic criteria.	Data from the independent sample were cross-tabulated with various ICD-10 and treatment-proxy based algorithms, computing sensitivity, specificity, PPV, NPV, and area under the curve (AUC) (Tapper et al 2019[132]).
Malignant melanoma	Denmark	Danish Cancer Registry and Danish Pathology Registry	ICD-10 code of malignant melanoma and ICDO-10 for topography and morphology as registered in the Cancer Registry	Another RCD source: Danish Pathology Registry	In a sample identified and characterized by the algorithm, the positive predictive values of malignant melanoma diagnosis, topography and morphology were calculated by comparing with data registered in the Pathology registry. Using a sample from the Pathology Registry, the completeness and the sensitivity of the melanoma diagnosis in the Cancer Registry was estimated (Pedersen et al 2018[133]).
Childhood-onset inflammatory bowel disease	Canada	Health administrative databases in Ontario, Canada collectively including hospital discharge data, billing claims for all physician services and demographic data including region of residence).	Various combinations of physician office and procedure billings with hospital records	(1) Another RCD source: clinical database of patients with childhood-onset inflammatory bowel disease (2) Chart review and an independent sample of confirmed cases and controls	A population-based clinical database of patients with IBD diagnosed aged <15 years was used to define cases, and patient information was linked to health administrative data to compare the sample of patients with known childhood-onset inflammatory bowel disease was used to test the diagnostic accuracy (sensitivity, specificity, positive predictive value and negative predictive value) of various patterns of healthcare use as registered in health administrative registries. The most accurate algorithm was further validated with chart data of children medical practices. The algorithm was validated in a sample of patients with IBD diagnosed <18 years derived from chart reviews at various medical practices. (Benchimol et al 2009[134]).
Use of oral anticoagulants due to atrial fibrillation	France	French national health insurance system database linked to the French hospital discharge database	A logistic model based on 14 independent covariates included based on their ability to discriminate between atrial fibrillation and other treatment indications. Developed in a training sample of oral anticoagulant users with known treatment indication	Sample of oral anticoagulant users with known treatment indication (validation sample)	Patients with a known treatment indication were randomly divided into a training sample (50% of patients) and a validation sample (the remaining 50%). The training sample was used to develop the algorithm and estimate logistic regression coefficients, while the validation sample was used to select the best model based on an area under ROC curve (c-index). The predictive accuracy of the final model was assessed by determining discrimination (c-statistics) and calibration, both evaluated on the validation sample (Billionnet et al 2017[135]).
Five events deemed to be important in studying risks and benefits of treatments:	Eight European countries	Multi-sector, multi-lingual RCD from eight European countries with	The Unified Medical Language System was used to identify concepts and corresponding codes	Benchmarking using published rates	An iterative process of computing incidence rates of the five events, and obtaining feedback from database holders to explore any disparities/difference from expected/published rates (Coloma et al 2013[98]).

acute myocardial infarction; acute renal failure; anaphylactic shock; bullous eruption; rhabdomyolysis		heterogeneous classifications			
Hyponatremia	Denmark	Danish National Patient Registry	ICD-10 codes recorded at discharge	Another RCD: database collecting laboratory tests performed in hospital-based diagnoses. One sodium value <135 mmol/L measured at any time during hospitalisation confirmed the diagnosis.	Example of a validation study on a large scale, possible when one database can be plausibly assumed to be a reference standard (i.e., lab values are superior to ICD-10 codes). Study population included 819 701 patients of all ages admitted to hospital during the study period (2 186 642 hospitalisations). Estimated sensitivity, specificity, positive predictive value and negative predictive value for ICD-10 codes for hyponatraemia overall and for cut-off points for increasing hyponatraemia severity. An important advantage (Holland-Bill et al 2014[136]).
Asthma	Canada	Provincial health administrative data in British Columbia, Canada, containing billing records of physician visits, hospital discharge summaries, and prescription drug purchases.	Three algorithms based on combinations of ICD-9 codes and purchases of asthma medications	Latent class modelling based definition of treated asthma	Each algorithm component is associated with an estimable probability of assigning a given patient to a treatment asthma category, allowing for a range of different asthma definitions, depending on a given study aim, to prioritise high sensitivity for health services planning, or a need to restrict analysis to asthma cases requiring hospitalization (Prosser et al 2008[86]).
Serious infection	Denmark	Danish National Patient Registry	ICD-10 diagnoses for infections identified in primary position of inpatient hospital encounters	Chart review	Case definition was imported by adopting/mapping an ICD-9 codes validated in the US to ICD-10 codes as well as criteria for presence of a serious infection. Charts were reviewed by medical doctors. Data on presence of criteria of any or preselected specific infections (pneumonia, sepsis) were collected as well as information on a chart reviewer's clinical judgement, whether an infection was present and if so, what specific infection it was (Holland-Bill et al 2014[103]).
Cardiovascular risk scores previously developed and validated in other sources of data: $CHADS_2$ (congestive	Denmark	Linked data from Danish population registries	ICD-8/ICD-10 codes for congestive heart failure and stroke; ICD-8/ICD-10 codes and/or treatment proxies (ATC codes) for	Expected 1-, 5-, and 10-year risk of thromboembolism	Used time-to event analysis – Cox regression – to estimate c-statistic to assess the predictive capability of each score and its components to predict the recorded occurrence of thromboembolism (Olesen et al 2011[137])

heart failure, hypertension, age≥75 years, diabetes mellitus, previous stroke/transient ischaemic attack) CHA2DS2-VASc (congestive heart failure, hypertension, age category ≥75 years, diabetes mellitus, previous, stroke/transient ischaemic attack, vascular disease)			diabetes and hypertension from outpatient dispensing data		
Registered cancer staging according to TNM at diagnosis	Denmark	Danish Cancer Registry	TNM tumour stage at diagnosis recorded in a cancer registry	Expectation that the data source must have complete data on TNM staging on all primary tumours captured in a national cancer registry	Computed proportion of records with missing TNM stage based on the expectation that the data source must have complete data on TNM staging (Deleruan et al 2012[138])
Multiple sclerosis (MS)	Canada	Ontario Electronic Medical Record Administrative data Linked Database (EMRALD)	Multiple algorithms based on combinations of cumulative patient profile, prescription and physicians billing codes were tested.	EMR search followed by chart review.	To establish reference standards, the 943 potential MS cases were identified among 73,003 enrolees of the database based on only information in the EMR (cumulative patient profile, free text entries for multiple sclerosis-related phrases, and physicians billing codes) but not the linked administrative data. Random sampling was not feasible since it was likely to yield a very small number of cases. Chart review of potential cases was conducted by trained chart abstractors (Krysko et al 2015[139]).
Stevens-Johnson Syndrome (SJS)/Toxic Epidermal Necrolysis (TEN)	Japan	A single educational hospital electronic medical records	Combination of ICD-10 codes, treatment/test/procedural proxies for SJS/TEN and other similar diseases for exclusion in a single	Another data source: disease registry at division of dermatology	This study is to develop and optimise the algorithm based on claims data to make differential diagnosis between SJS/TEN and other similar diseases. Expert dermatologist confirmed cases in the registry. The patients who had the ICD-10 codes of differential diagnoses were identified in the EHR of division of dermatology. They were tested against 1214 algorithms composed of all combinations of six

			hospital administrative data.		algorithm components to calculate validity and performance measures including sensitivity, specificity, PPV, NPV and diagnostic odds ratio (Fukasawa et al 2019[106])
Lung cancer	United States	The HealthCore Integrated Research Environment (HIRE)-Oncology clinical database that were linked with the HealthCore Integrated Research Database (HIRD).	Modified Duh Algorithm based on the first-line chemotherapy regimens for NSCLC in Healthcare Common Procedure Coding System and ICD-10, with exclusion criteria included procedures, surgeries, and chemotherapies administered to patients with SCLC as recommended by the 2006 NCCN Guidelines.	Another data source: Lung cancer patients identified in the HIRE-Oncology database based on a diagnosis of lung cancer with a molecular classification of NSCLC or SCLC.	Performance of NSCLC case-finding algorithm to distinguish NSCLC and SCLC cases was compared with the control algorithm using sensitivity, specificity, PPV, NPV, accuracy, diagnostic odds ratio (DOR), and area under the curve (AUC) for ROC. The control algorithm included all tests and first-line treatments based on the complete list of the American Cancer Society[140]and 2016 NCCN[141] recommended treatments for NSCLC and SCLC were combined and no exclusion criteria (Turner et al, 2017[142]).
Multiple malignancies	Taiwan	The National Health Insurance (NHI) database and the National Cancer Registry (NCR) in Taiwan	Cancer patients in the NHI database were identified based on the catastrophic illness certificates for medical co-payment waivers, which were recorded in the Registry of Catastrophic Illness Patients. Cancer types including all cancers and top 10 cancers were based on ICD-9.	Another data source: the National Cancer Registry (NCR), a mandatory population-based cancer registry%, respectively.	Deterministic linkage by the unique personal identification assigned to each Taiwanese resident was conducted, which also allows the linkage to civil registration, birth and death registries, and the NCR. Sensitivity, specificity, PPV, and NPV for all cancer and each cancer were estimated using the NCR as reference standard and survival recorded in the NHI were calculated. (Kao et al, 2017[143])
Multiple malignancies	United States	Medicare claims	Four candidate algorithms validated for cancers of lung, colorectal, breast, lymphoma and leukaemia: Definition 1: Combination of diagnosis and procedures on the same day or within the same hospitalisation Definition 2: Two diagnoses of a specific cancer within 2 months Definition 3: Definition 1	Another RCD source: Pennsylvania State Cancer Registry, with certified high-quality and completeness of data on malignancies	Cross-tabulation of the claims-based algorithms with the cancer registry data to assess sensitivity, specificity and PPV among 157,310 Medicare beneficiaries in 1997-2000. In addition, the study evaluated algorithm's ability to distinguish between incident and recurrent malignancies and evaluated the impact of imperfect completeness of the references standard in capturing haematological malignancies (Setoguchi et al, 2007[83]).

			or Definition 2 Definition 4: One diagnosis of cancer		

Abbreviations not explained in the table: ICD-9 International Classification of Diseases, 9th Revision; ICD-10 International Classification of Diseases, 10th Revision; ATC Anatomical Therapeutic Chemical; TNM Tumour, Nodes, Metastases; SCLC small cell lung cancer; NSCLC non-small cell lung cancer; EHR electronic health records; NSCP Nomesco Classification of Surgical Procedures

Box 3. Measures of algorithm validity

Measure	Range	Description	Formulas
Sensitivity	[0, 1]	Proportion of persons with a given health state according to gold/reference standard who are classified as such by an algorithm	
Specificity	[0, 1]	Proportion of persons without a given health state according to gold/reference standard who are classified as such by an algorithm	
Positive predictive value (PPV)	[0, 1]	Proportion of patients who truly have the health state among all those who are classified as positive by the algorithm	
Negative predictive value (NPV)	[0, 1]	Proportion of patients who truly do not have the health state among all those who are classified as negative by the algorithm. Chuback et al and Benchimol et al provide formulae connecting sensitivity and specificity, PPV and NPV with prevalence of a health state.[30 64]	

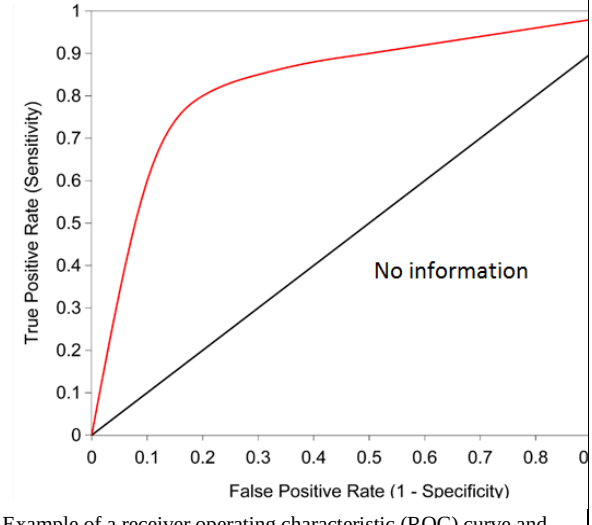
		Health state status based on gold/reference standard: Positive	Negative
Health state status based on algorithm	Positive	A (True positive)	C (False negative)
	Negative	B (False positive)	D (True negative)

$$Sensitivity = \frac{A}{(A+B)} =$$

$$Specificity = \frac{D}{(C+D)} =$$

$$PPV = \frac{A}{(A+C)} =$$

$$NPV = \frac{D}{(B+D)} =$$

Receiver operating characteristic (ROC) curve	Definition of an RCD-based algorithm may be based on a cut-off value of a measured continuous variable (eg, hyponatremia based on serum potassium levels in the study by Holland-Bill et al[136]) Generally, more extreme cut-off values of a continuous variable leads to increase in proportion of both false-positives, but also true-positives. For algorithms that classify patients into health states (eg., presence/absence/severity) ROC is a plot of sensitivity (or true positive proportion) on the x-axis against (1-specificity) or false-positive proportion on the y-axis. For a (hypothetical) perfect algorithm, sensitivity=specificity=1.[79]	
Area under the curve (AUC)	Area under the curve (AUC) is derived from ROC curve and is a single measure, frequently used to indicate performance of diagnostic tests, or algorithms, as they are used to 'diagnose' presence of a health state in the study population. With varying thresholds for case definition, the points (1-Sensitivity, Specificity) are plotted on the plain with 1-Sensitivity on the x-axis and Specificity on the y-axis to construct the curve. An algorithm with sensitivity = specificity=1 ROC=1; an algorithm that would classify patients not better than a coin toss would have a AUC=0.5.[79]	

Example of a receiver operating characteristic (ROC) curve and area under the curve (AUC). (Sørensen and Vandenbroucke (in press).[144])

<table>
<tr>
<td>Diagnostic odds ratio (DOR)</td>
<td>[0, ∞)</td>
<td>An odds ratio of dichotomous tests in diagnostic application and frequently used for meta-analysis of diagnostic tests. For the purposes of validation studies, DOR can be calculated as the ratio of the odds of positivity in those with a health state of interest relative to the odds of positivity in those without the health state of interest.[80] DOR does not depend on prevalence of health state, with higher values indicating better discriminatory test performance and used in combination with sensitivity and specificity. A value of 1 means that the test does not discriminate between cases and non-cases. Higher DOR correspond to higher probability of an algorithm to be positive among true cases than in non-cases of a given health state.</td>
<td>$$DOR = \frac{A\ (True\ Positive)}{B\ (False\ Negative)} \Big/ \frac{C\ (False\ Positive)}{D\ (True\ Negative)}$$</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>[-1, 1]</td>
<td>Kappa statistic is used to quantify interrater variability.[144] In evaluating RCD-based algorithm, it can be used to quantify agreement between two algorithms or data sources, none of which can be considered a better (a reference standard) relative to the other.

Based on the value of kappa statistic, agreement may be qualified on a scale ‘less than chance’ to ‘almost perfect (Viera and Garrett 2005[81].)

Kappa — Agreement
< 0 — Less than chance
0.01–0.20 — Slight
0.21–0.40 — Fair
0.41–0.60 — Moderate
0.61–0.80 — Substantial
0.81–0.99 — Almost perfect</td>
<td>
<table>
<tr><td></td><td></td><td>Health state status according to data source 1
Positive</td><td>Negative</td><td></td></tr>
<tr><td>Health state status according to data source 2</td><td>Positive</td><td>a</td><td>b</td><td>m_1</td></tr>
<tr><td></td><td>Negative</td><td>c</td><td>d</td><td>m_0</td></tr>
<tr><td></td><td></td><td>n_1</td><td>n_0</td><td>n</td></tr>
</table>
Expected agreement between two data sources/algorithm, none of which is superior to the other in classifying a given health state into present/absent
$$Observed\ agreement\ p_{obs} = \frac{(a+d)}{n}$$
$$Expected\ agreement\ p_{exp} = \left[\left(\frac{n_1}{n}\right) \times \left(\frac{m_1}{n}\right)\right] + \left[\left(\frac{n_0}{n}\right) \times \left(\frac{m_0}{n}\right)\right]$$
$$Kappa\ statistic\ \kappa = \frac{(p_{obs} - p_{exp})}{(1 - p_{exp})}$$
</td>
</tr>
</table>

615

References

1. Makady A, de Boer A, Hillege H, et al. What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. *Value Health* 2017;20(7):858-65. doi: https://doi.org/10.1016/j.jval.2017.03.008
2. Corrigan-Curay J, Sacks L, Woodcock J. Real-World Evidence and Real-World Data for Evaluating Drug Safety and Effectiveness. *JAMA* 2018;320(9):867-68. doi: 10.1001/jama.2018.10136 [published Online First: 2018/08/15]
3. Berger ML, Sox H, Willke RJ, et al. Good practices for real-world data studies of treatment and/or comparative effectiveness: Recommendations from the joint ISPOR-ISPE Special Task Force on real-world evidence in health care decision making. *Pharmacoepidemiol Drug Saf* 2017;26(9):1033-39. doi: 10.1002/pds.4297 [published Online First: 2017/09/16]
4. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-World Evidence - What Is It and What Can It Tell Us? *N Engl J Med* 2016;375(23):2293-97. doi: 10.1056/NEJMsb1609216 [published Online First: 2016/12/14]
5. Administration UFaD. Real-World Evidence 2019 [Available from: https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence accessed 16 October 2019.
6. Booth CM, Karim S, Mackillop WJ. Real-world data: towards achieving the achievable in cancer care. *Nature Reviews Clinical Oncology* 2019;16(5):312-25. doi: 10.1038/s41571-019-0167-7
7. Administration UFaD. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices 2017 [Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/use-real-world-evidence-support-regulatory-decision-making-medical-devices accessed 16 October 2019.
8. Administration UFaD. Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics 2019 [Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/submitting-documents-using-real-world-data-and-real-world-evidence-fda-drugs-and-biologics-guidance accessed 16 October 2019.
9. Plueschke K, McGettigan P, Pacurariu A, et al. EU-funded initiatives for real world evidence: descriptive analysis of their characteristics and relevance for regulatory decision-making. *BMJ Open* 2018;8(6):e021864. doi: 10.1136/bmjopen-2018-021864
10. Pharmaceuticals and Medical Devices Agency. The guidelines for the conduct of pharmacoepidemiological studies using medical information databases, etc., in evaluation of drug safety. 1 ed. Tokyo2014 [Available from: https://www.pmda.go.jp/files/000147250.pdf accessed 31 January 2020.
11. Pharmaceutical Evaluation Division and Pharmaceutical Safety Division. Basic Principles on the utilization of health information databases for Post-Marketing Surveillance of Medical Products Tokyo2017 [Available from: https://www.pmda.go.jp/files/000218531.pdf accessed 31 January 2020.
12. Ministry of Health Labour and Welfare. Amendment of Ministerial Ordinance Concerning Good Post-Marketing Study Practice [in Japanese] Tokyo2017 [Available from: https://elaws.e-gov.go.jp/search/elawsSearch/elaws_search/lsg0500/detail?lawId=416M60000100171 accessed 31 January 2020.
13. European Medicines Agency. Medical Devices [Available from: https://www.ema.europa.eu/en/human-regulatory/overview/medical-devices accessed 24 February 2020.
14. Strom BL, Kimmel SE, Hennessy S. Pharmacoepidemiology. Sixth edition. ed. Hoboken, NJ: Wiley-Blackwell 2020.
15. Kwakkenbos L, Imran M, McCall SJ, et al. CONSORT extension for the reporting of randomised controlled trials conducted using cohorts and routinely collected data (CONSORT-ROUTINE): checklist with explanation and elaboration. *BMJ* 2021;373:n857. doi: 10.1136/bmj.n857 [published Online First: 2021/05/01]
16. About the Food and Drug Administration (FDA) Sentinel Initiative [Available from: https://www.sentinelinitiative.org/about#section-1592851590618 accessed 11 May 2021.
17. (EMA) EMA. Technical workshop on real-world metadata for regulatory purposes 2021 [Available from: https://www.ema.europa.eu/en/events/technical-workshop-real-world-metadata-regulatory-purposes#documents-section accessed 19 May 20021.
18. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. *N Engl J Med* 2019;380(14):1347-58. doi: 10.1056/NEJMra1814259 [published Online First: 2019/04/04]
19. Eichler HG, Bloechl-Daum B, Broich K, et al. Data Rich, Information Poor: Can We Use Electronic Health Records to Create a Learning Healthcare System for Pharmaceuticals? *Clin Pharmacol Ther* 2019;105(4):912-22. doi: 10.1002/cpt.1226 [published Online First: 2018/09/05]

20. Fletcher RH, Fletcher SW. Clinical epidemiology : the essentials. 4th ed. Philadelphia: Lippincott Williams & Wilkins 2005.
21. Weiss NS. Clinical epidemiology : the study of the outcome of illness. 3rd ed. Oxford ; New York: Oxford University Press 2006.
22. Maclure M, Schneeweiss S. Causation of bias: the episcope. *Epidemiology* 2001;12(1):114-22. [published Online First: 2001/01/04]
23. Bosco JL, Silliman RA, Thwin SS, et al. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *J Clin Epidemiol* 2010;63(1):64-74. doi: 10.1016/j.jclinepi.2009.03.001 [published Online First: 2009/05/22]
24. Jurek AM, Greenland S, Maldonado G, et al. Proper interpretation of non-differential misclassification effects: expectations vs observations. *Int J Epidemiol* 2005;34(3):680-7. doi: 10.1093/ije/dyi060 [published Online First: 2005/04/02]
25. Rothman KJ, Greenland S, Lash TL. Chapter 9. Validity in Epidemiologic Studies. In: Rothman KJ, Greenland S, Lash TL, eds. Modern epidemiology. 3rd ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins 2008:x, 758 p.
26. Manuel DG, Rosella LC, Stukel TA. Importance of accurately identifying disease in studies using electronic health records. *BMJ* 2010;341:c4226. doi: 10.1136/bmj.c4226 [published Online First: 2010/08/21]
27. Greenland S. Basic methods for sensitivity analysis of biases. *Int J Epidemiol* 1996;25(6):1107-16. [published Online First: 1996/12/01]
28. van Smeden M, Lash TL, Groenwold RHH. Reflection on modern methods: five myths about measurement error in epidemiological research. *Int J Epidemiol* 2020;49(1):338-47. doi: 10.1093/ije/dyz251 [published Online First: 2019/12/11]
29. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *BMJ* 2018;361:k1479. doi: 10.1136/bmj.k1479 [published Online First: 2018/05/02]
30. Benchimol EI, Manuel DG, To T, et al. Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *J Clin Epidemiol* 2011;64(8):821-9. doi: 10.1016/j.jclinepi.2010.10.006 [published Online First: 2011/01/05]
31. Lanes S, Brown JS, Haynes K, et al. Identifying health outcomes in healthcare databases. *Pharmacoepidemiol Drug Saf* 2015;24(10):1009-16. doi: 10.1002/pds.3856 [published Online First: 2015/08/19]
32. Sørensen HT, Baron JA. Registries and medical databases. In: Olsen J, Greene N, Saracci R, et al., eds. Teaching epidemiology : a guide for teachers in epidemiology, public health and clinical medicine
Fourth edition. ed. Oxford, United Kingdom ; New York, NY, United States of America: Oxford University Press 2015.
33. Meng XL. Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox, and the 2016 Us Presidential Election. *Ann Appl Stat* 2018;12(2):685-726. doi: 10.1214/18-Aoas1161sf
34. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. *Int J Epidemiol* 1996;25(2):435-42. [published Online First: 1996/04/01]
35. Schneeweiss S, Eichler HG, Garcia-Altes A, et al. Real World Data in Adaptive Biomedical Innovation: A Framework for Generating Evidence Fit for Decision-Making. *Clin Pharmacol Ther* 2016;100(6):633-46. doi: 10.1002/cpt.512 [published Online First: 2016/10/21]
36. Franklin JM, Glynn RJ, Martin D, et al. Evaluating the Use of Nonrandomized Real-World Data Analyses for Regulatory Decision Making. *Clin Pharmacol Ther* 2019;105(4):867-77. doi: 10.1002/cpt.1351 [published Online First: 2019/01/14]
37. Jick SS, Kaye JA, Vasilakis-Scaramozza C, et al. Validity of the general practice research database. *Pharmacotherapy* 2003;23(5):686-9. [published Online First: 2003/05/14]
38. Ludvigsson JF, Andersson E, Ekbom A, et al. External review and validation of the Swedish national inpatient register. *BMC Public Health* 2011;11(1):450. doi: 10.1186/1471-2458-11-450 [published Online First: 2011/06/11]
39. Carnahan RM. Mini-Sentinel's systematic reviews of validated methods for identifying health outcomes using administrative data: summary of findings and suggestions for future research. *Pharmacoepidemiol Drug Saf* 2012;21 Suppl 1:90-9. doi: 10.1002/pds.2318
40. Springate DA, Kontopantelis E, Ashcroft DM, et al. ClinicalCodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. *PLoS One* 2014;9(6):e99825. doi: 10.1371/journal.pone.0099825 [published Online First: 2014/06/19]
41. Schmidt M, Schmidt SA, Sandegaard JL, et al. The Danish National Patient Registry: a review of content, data quality, and research potential. *Clin Epidemiol* 2015;7:449-90. doi: 10.2147/CLEP.S91125 [published Online First: 2015/11/26]

42. Mini-Sentinel. [Available from: http://www.mini-sentinel.org/ accessed 7 December 2015.
43. Koram N, Delgado M, Stark JH, et al. Validation studies of claims data in the Asia-Pacific region: A comprehensive review. *Pharmacoepidemiol Drug Saf* 2019;28(2):156-70. doi: 10.1002/pds.4616
44. von Elm E, Altman DG, Egger M, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 2007;335(7624):806-8. doi: 10.1136/bmj.39335.541782.AD [published Online First: 2007/10/20]
45. Vandenbroucke JP, von Elm E, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): explanation and elaboration. *Ann Intern Med* 2007;147(8):W163-94. doi: 10.7326/0003-4819-147-8-200710160-00010-w1 [published Online First: 2007/10/17]
46. Nicholls SG, Langan SM, Sorensen HT, et al. The RECORD reporting guidelines: meeting the methodological and ethical demands of transparency in research using routinely-collected health data. *Clin Epidemiol* 2016;8:389-92. doi: 10.2147/CLEP.S110528
47. Benchimol EI, Langan S, Guttmann A, et al. Call to RECORD: the need for complete reporting of research using routinely collected health data. *J Clin Epidemiol* 2013;66(7):703-5. doi: 10.1016/j.jclinepi.2012.09.006
48. 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. doi: 10.2147/clep.s36885 [published Online First: 2013/02/16]
49. Nicholls SG, Quach P, von Elm E, et al. The REporting of Studies Conducted Using Observational Routinely-Collected Health Data (RECORD) Statement: Methods for Arriving at Consensus and Developing Reporting Guidelines. *PLoS One* 2015;10(5):e0125620. doi: 10.1371/journal.pone.0125620
50. 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(10):e1001885. doi: 10.1371/journal.pmed.1001885 [published Online First: 2015/10/07]
51. Langan SM, Schmidt SA, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). *BMJ* 2018;363:k3532. doi: 10.1136/bmj.k3532 [published Online First: 2018/11/16]
52. 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. doi: 10.2147/CLEP.S104448 [published Online First: 2016/04/26]
53. Chun DS, Lund JL, Sturmer T. Pharmacoepidemiology and Drug Safety's special issue on validation studies. *Pharmacoepidemiol Drug Saf* 2019;28(2):123-25. doi: 10.1002/pds.4694 [published Online First: 2019/02/05]
54. Lash TL, Olshan AF. EPIDEMIOLOGY Announces the "Validation Study" Submission Category. *Epidemiology* 2016;27(5):613-4. doi: 10.1097/EDE.0000000000000532 [published Online First: 2016/07/09]
55. Nissen F, Quint JK, Morales DR, et al. How to validate a diagnosis recorded in electronic health records. *Breathe (Sheff)* 2019;15(1):64-68. doi: 10.1183/20734735.0344-2018 [published Online First: 2019/03/07]
56. Fox MP, Lash TL, Bodnar LM. Common misconceptions about validation studies. *Int J Epidemiol* 2020 doi: 10.1093/ije/dyaa090
57. International Society for Phamacoepidemiology. Guidelines for Good Pharmacoepidemiology Practices (GPP) 2015 [Available from: https://www.pharmacoepi.org/resources/policies/guidelines-08027 accessed 23 May 2021.
58. Hines PA, Janssens R, Gonzalez-Quevedo R, et al. A future for regulatory science in the European Union: the European Medicines Agency's strategy. *Nat Rev Drug Discov* 2020;19(5):293-94. doi: 10.1038/d41573-020-00032-0 [published Online First: 2020/04/03]
59. Orsini LS, Monz B, Mullins CD, et al. Improving transparency to build trust in real-world secondary data studies for hypothesis testing-Why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Pharmacoepidemiol Drug Saf* 2020;29(11):1504-13. doi: 10.1002/pds.5079 [published Online First: 2020/09/15]
60. Wang SV, Schneeweiss S, Berger ML, et al. Reporting to Improve Reproducibility and Facilitate Validity Assessment for Healthcare Database Studies V1.0. *Pharmacoepidemiol Drug Saf* 2017;26(9):1018-32. doi: 10.1002/pds.4295 [published Online First: 2017/09/16]
61. Greenfield S. Making Real-World Evidence More Useful for Decision Making. *Value Health* 2017;20(8):1023-24. doi: 10.1016/j.jval.2017.08.3012 [published Online First: 2017/10/02]
62. Nicholls SG, Langan SM, Benchimol EI. Routinely collected data: the importance of high-quality diagnostic coding to research. *CMAJ* 2017;189(33):E1054-e55. doi: 10.1503/cmaj.170807 [published Online First: 2017/08/23]

63. Nicholls SG, Langan SM, Benchimol EI, et al. Reporting transparency: making the ethical mandate explicit. *BMC Med* 2016;14:44. doi: 10.1186/s12916-016-0587-5
64. Chubak J, Pocobelli G, Weiss NS. Tradeoffs between accuracy measures for electronic health care data algorithms. *J Clin Epidemiol* 2012;65(3):343-49 e2. doi: 10.1016/j.jclinepi.2011.09.002
65. Schmidt M, Jacobsen JB, Lash TL, et al. 25 year trends in first time hospitalisation for acute myocardial infarction, subsequent short and long term mortality, and the prognostic impact of sex and comorbidity: a Danish nationwide cohort study. *BMJ* 2012;344:e356. doi: 10.1136/bmj.e356 [published Online First: 2012/01/27]
66. Vinter N, Linder M, Andersen M, et al. Classification and characteristics of on-label and off-label apixaban use in Denmark and Sweden. *Pharmacoepidemiol Drug Saf* 2019;28(6):867-78. doi: 10.1002/pds.4778 [published Online First: 2019/04/18]
67. Shivade C, Raghavan P, Fosler-Lussier E, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* 2014;21(2):221-30. doi: 10.1136/amiajnl-2013-001935 [published Online First: 2013/11/10]
68. Wong J, Horwitz MM, Zhou L, et al. Using machine learning to identify health outcomes from electronic health record data. *Curr Epidemiol Rep* 2018;5(4):331-42. doi: 10.1007/s40471-018-0165-9 [published Online First: 2018/12/18]
69. Tanskanen A, Taipale H, Koponen M, et al. From prescription drug purchases to drug use periods - a second generation method (PRE2DUP). *BMC Med Inform Decis Mak* 2015;15:21. doi: 10.1186/s12911-015-0140-z [published Online First: 2015/04/19]
70. Kirby JC, Speltz P, Rasmussen LV, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc* 2016;23(6):1046-52. doi: 10.1093/jamia/ocv202 [published Online First: 2016/03/31]
71. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA* 2018;319(13):1317-18. doi: 10.1001/jama.2017.18391 [published Online First: 2018/03/14]
72. Pottegard A, Schmidt SAJ, Wallach-Kildemoes H, et al. Data Resource Profile: The Danish National Prescription Registry. *Int J Epidemiol* 2017;46(3):798-98f. doi: 10.1093/ije/dyw213 [published Online First: 2016/10/30]
73. Schmidt M, Hallas J, Friis S. Potential of prescription registries to capture individual-level use of aspirin and other nonsteroidal anti-inflammatory drugs in Denmark: trends in utilization 1999-2012. *Clin Epidemiol* 2014;6:155-68. doi: 10.2147/CLEP.S59156 [published Online First: 2014/05/30]
74. Sweden bans over-the-counter sales of painkiller over heart failure fears 2019 [Available from: https://www.thelocal.se/20191105/sweden-bans-over-the-counter-sales-of-painkiller-over-heart-failure-fears accessed 28 August 2020.
75. Schmidt M, Sorensen HT, Pedersen L. Diclofenac use and cardiovascular risks: series of nationwide cohort studies. *BMJ* 2018;362:k3426. doi: 10.1136/bmj.k3426 [published Online First: 2018/09/06]
76. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. *BMJ* 2001;323(7305):157-62. doi: 10.1136/bmj.323.7305.157 [published Online First: 2001/07/21]
77. West SL, Ritchey ME, Poole C. Chapter 12. Validity of Pharmacoepidemiologic Drug and Diagnosis Data. In: Strom BL, Kimmel SE, Hennessy S, eds. Textbook of Pharmacoepidemiology. 5th ed. Chichester, West Sussex, UK: Wiley-Blackwell 2013:p.
78. Bollaerts K, Rekkas A, De Smedt T, et al. Disease misclassification in electronic healthcare database studies: Deriving validity indices-A contribution from the ADVANCE project. *PLoS One* 2020;15(4):e0231333. doi: 10.1371/journal.pone.0231333 [published Online First: 2020/04/23]
79. Bewick V, Cheek L, Ball J. Statistics review 13: receiver operating characteristic curves. *Crit Care* 2004;8(6):508-12. doi: 10.1186/cc3000 [published Online First: 2004/11/30]
80. Glas AS, Lijmer JG, Prins MH, et al. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;56(11):1129-35. doi: 10.1016/s0895-4356(03)00177-x [published Online First: 2003/11/15]
81. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med* 2005;37(5):360-3. [published Online First: 2005/05/11]
82. Fox MP, Lash TL, Greenland S. A method to automate probabilistic sensitivity analyses of misclassified binary variables. *Int J Epidemiol* 2005;34(6):1370-6. doi: 10.1093/ije/dyi184 [published Online First: 2005/09/21]
83. Setoguchi S, Solomon DH, Glynn RJ, et al. Agreement of diagnosis and its date for hematologic malignancies and solid tumors between medicare claims and cancer registry data. *Cancer Causes Control* 2007;18(5):561-9. doi: 10.1007/s10552-007-0131-1 [published Online First: 2007/04/21]
84. Berkson J. Limitations of the application of fourfold table analysis to hospital data. *Biometrics* 1946;2(3):47-53. [published Online First: 1946/06/01]

85. Ording AG, Cronin-Fenton D, Ehrenstein V, et al. Challenges in translating endpoints from trials to observational cohort studies in oncology. *Clin Epidemiol* 2016;8:195-200. doi: 10.2147/CLEP.S97874 [published Online First: 2016/06/30]
86. Prosser RJ, Carleton BC, Smith MA. Identifying persons with treated asthma using administrative data via latent class modelling. *Health Serv Res* 2008;43(2):733-54. doi: 10.1111/j.1475-6773.2007.00775.x [published Online First: 2008/03/29]
87. Morkem R, Handelman K, Queenan JA, et al. Validation of an EMR algorithm to measure the prevalence of ADHD in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). *BMC Med Inform Decis Mak* 2020;20(1):166. doi: 10.1186/s12911-020-01182-2 [published Online First: 2020/07/22]
88. Comparative effectiveness and safety of non-vitamin K oral anticoagulants and warfarin in non-valvular atrial fibrillation - a cohort study in 3 Nordic countries. ESC Congress; 2019; Paris, France.
89. Nielsen PB, Skjoth F, Sogaard M, et al. Effectiveness and safety of reduced dose non-vitamin K antagonist oral anticoagulants and warfarin in patients with atrial fibrillation: propensity weighted nationwide cohort study. *BMJ* 2017;356:j510. doi: 10.1136/bmj.j510
90. Hellfritzsch M, Pottegard A, Haastrup SB, et al. Cohort selection in register-based studies of direct oral anticoagulant users with atrial fibrillation: An inevitable trade-off between selection bias and misclassification. *Basic Clin Pharmacol Toxicol* 2020;127(1):3-5. doi: 10.1111/bcpt.13423 [published Online First: 2020/05/05]
91. Riis AH, Johansen MB, Jacobsen JB, et al. Short look-back periods in pharmacoepidemiologic studies of new users of antibiotics and asthma medications introduce severe misclassification. *Pharmacoepidemiol Drug Saf* 2015;24(5):478-85. doi: 10.1002/pds.3738 [published Online First: 2015/01/21]
92. Schmidt M, Schmidt SAJ, Adelborg K, et al. The Danish health care system and epidemiological research: from health care contacts to database records. *Clin Epidemiol* 2019;11:563-91. doi: 10.2147/CLEP.S179083 [published Online First: 2019/08/03]
93. Bakken IJ, Ariansen AMS, Knudsen GP, et al. The Norwegian Patient Registry and the Norwegian Registry for Primary Health Care: Research potential of two nationwide health-care registries. *Scand J Public Health* 2019;0(0):1403494819859737. doi: 10.1177/1403494819859737 [published Online First: 2019/07/11]
94. Gribsholt SB, Pedersen L, Richelsen B, et al. Validity of ICD-10 diagnoses of overweight and obesity in Danish hospitals. *Clin Epidemiol* 2019;11:845-54. doi: 10.2147/CLEP.S214909 [published Online First: 2019/10/02]
95. Sturmer T, Glynn RJ, Rothman KJ, et al. Adjustments for unmeasured confounders in pharmacoepidemiologic database studies using external information. *Med Care* 2007;45(10 Supl 2):S158-65. doi: 10.1097/MLR.0b013e318070c045 [published Online First: 2007/10/25]
96. Valkhoff VE, Coloma PM, Masclee GM, et al. Validation study in four health-care databases: upper gastrointestinal bleeding misclassification affects precision but not magnitude of drug-related upper gastrointestinal bleeding risk. *J Clin Epidemiol* 2014;67(8):921-31. doi: 10.1016/j.jclinepi.2014.02.020 [published Online First: 2014/05/06]
97. Gini R, Schuemie M, Brown J, et al. Data Extraction and Management in Networks of Observational Health Care Databases for Scientific Research: A Comparison of EU-ADR, OMOP, Mini-Sentinel and MATRICE Strategies. *EGEMS (Wash DC)* 2016;4(1):1189. doi: 10.13063/2327-9214.1189 [published Online First: 2016/03/26]
98. Avillach P, Coloma PM, Gini R, et al. Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. *J Am Med Inform Assoc* 2013;20(1):184-92. doi: 10.1136/amiajnl-2012-000933 [published Online First: 2012/09/08]
99. Ehrenstein V, Huang K, Kahlert J, et al. Abstract. *Pharmacoepidemiol Drug Saf* 2017;26(S2):3-636. doi: 10.1002/pds.4275
100. Forns J, Cainzos-Achirica M, Hellfritzsch M, et al. Validity of ICD-9 and ICD-10 codes used to identify acute liver injury: A study in three European data sources. *Pharmacoepidemiol Drug Saf* 2019;28(7):965-75. doi: 10.1002/pds.4803 [published Online First: 2019/06/07]
101. ACCESS STUDY PLACEHOLDER - AVAILABLE IN JUNE 2021 FOR CITATION. PLACEHOLDER
102. Lawlor DA, Tilling K, Davey Smith G. Triangulation in aetiological epidemiology. *Int J Epidemiol* 2016;45(6):1866-86. doi: 10.1093/ije/dyw314 [published Online First: 2017/01/22]
103. Holland-Bill L, Xu H, Sorensen HT, et al. Positive predictive value of primary inpatient discharge diagnoses of infection among cancer patients in the Danish National Registry of Patients. *Ann Epidemiol* 2014;24(8):593-7, 97 e1-18. doi: 10.1016/j.annepidem.2014.05.011 [published Online First: 2014/08/03]
104. Benchimol EI, Guttmann A, Mack DR, et al. Validation of international algorithms to identify adults with inflammatory bowel disease in health administrative data from Ontario, Canada. *J Clin Epidemiol* 2014;67(8):887-96. doi: 10.1016/j.jclinepi.2014.02.019 [published Online First: 2014/04/30]

105. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. *J Clin Epidemiol* 2005;58(8):859-62. doi: 10.1016/j.jclinepi.2004.12.009 [published Online First: 2005/07/16]
106. Fukasawa T, Takahashi H, Kameyama N, et al. Development of an electronic medical record-based algorithm to identify patients with Stevens-Johnson syndrome and toxic epidermal necrolysis in Japan. *PLoS One* 2019;14(8):e0221130. doi: 10.1371/journal.pone.0221130 [published Online First: 2019/08/14]
107. Holcroft CA, Spiegelman D. Design of validation studies for estimating the odds ratio of exposure-disease relationships when exposure is misclassified. *Biometrics* 1999;55(4):1193-201. doi: 10.1111/j.0006-341x.1999.01193.x [published Online First: 2001/04/21]
108. Collin LJ, MacLehose RF, Ahern TP, et al. Adaptive Validation Design: A Bayesian Approach to Validation Substudy Design With Prospective Data Collection. *Epidemiology* 2020;31(4):509-16. doi: 10.1097/EDE.0000000000001209 [published Online First: 2020/06/03]
109. Lash TL, Fox MP, Fink AK. Applying quantitative bias analysis to epidemiologic data (https://sites.google.com/site/biasanalysis/). Dordrecht ; New York: Springer 2009.
110. van Walraven C, Hart RG. Leave 'em alone - why continuous variables should be analyzed as such. *Neuroepidemiology* 2008;30(3):138-9. doi: 10.1159/000126908 [published Online First: 2008/04/19]
111. Hall GC, Lanes S, Bollaerts K, et al. 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):1450-55. doi: 10.1002/pds.5109 [published Online First: 2020/08/30]
112. Lash TL, Fox MP, MacLehose RF, et al. Good practices for quantitative bias analysis. *Int J Epidemiol* 2014;43(6):1969-85. doi: 10.1093/ije/dyu149 [published Online First: 2014/08/01]
113. Rothman KJ. Episheet [Available from: http://krothman.hostbyet2.com/episheet.xls accessed 26 October 2019.
114. MacLehose RF, Bodnar LM, Meyer CS, et al. Hierarchical Semi-Bayes Methods for Misclassification in Perinatal Epidemiology. *Epidemiology* 2018;29(2):183-90. doi: 10.1097/ede.0000000000000789 [published Online First: 2017/11/23]
115. Walraven CV. A comparison of methods to correct for misclassification bias from administrative database diagnostic codes. *Int J Epidemiol* 2018;47(2):605-16. doi: 10.1093/ije/dyx253 [published Online First: 2017/12/19]
116. Chu H, Wang Z, Cole SR, et al. Sensitivity analysis of misclassification: a graphical and a Bayesian approach. *Ann Epidemiol* 2006;16(11):834-41. doi: 10.1016/j.annepidem.2006.04.001 [published Online First: 2006/07/18]
117. van Walraven C. Bootstrap imputation with a disease probability model minimized bias from misclassification due to administrative database codes. *J Clin Epidemiol* 2017;84:114-20. doi: 10.1016/j.jclinepi.2017.01.007 [published Online First: 2017/02/09]
118. Newcomer SR, Xu S, Kulldorff M, et al. A primer on quantitative bias analysis with positive predictive values in research using electronic health data. *J Am Med Inform Assoc* 2019;26(12):1664-74. doi: 10.1093/jamia/ocz094 [published Online First: 2019/08/01]
119. Funk MJ, Landi SN. Misclassification in administrative claims data: quantifying the impact on treatment effect estimates. *Curr Epidemiol Rep* 2014;1(4):175-85. doi: 10.1007/s40471-014-0027-z [published Online First: 2015/06/19]
120. 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(1):4-14. doi: 10.1111/j.1365-2125.2009.03537.x [published Online First: 2010/01/19]
121. Centers for Medicare and Medicaid Services. [Available from: https://www.cms.gov/Newsroom/MediaReleaseDatabase/Press-releases/2014-Press-releases-items/2014-07-31.html accessed 9 December 2015.
122. ICD-11 [Available from: https://icd.who.int/en/ accessed 06 May 2019.
123. Miksad RA, Abernethy AP. Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality. *Clin Pharmacol Ther* 2018;103(2):202-05. doi: 10.1002/cpt.946 [published Online First: 2017/12/08]
124. WHO. ICD-11: Classifying disease to map the way we live and die 2019 [Available from: https://www.who.int/health-topics/international-classification-of-diseases accessed 06 May 2019.
125. Gottlieb S. Remarks to the National Academy of Sciences on the Impact of Real World Evidence on Medical Product Development 2017 [Available from: https://www.fda.gov/newsevents/speeches/ucm576519.htm accessed 02 July 2021.

126. Lund JL, Froslev T, Deleuran T, et al. Validity of the Danish National Registry of Patients for chemotherapy reporting among colorectal cancer patients is high. *Clin Epidemiol* 2013;5:327-34. doi: 10.2147/CLEP.S49773 [published Online First: 2013/09/17]
127. Tanskanen A, Taipale H, Koponen M, et al. Drug exposure in register-based research-An expert-opinion based evaluation of methods. *PLoS One* 2017;12(9):e0184070. doi: 10.1371/journal.pone.0184070 [published Online First: 2017/09/09]
128. Adelborg K, Sundboll J, Munch T, et al. Positive predictive value of cardiac examination, procedure and surgery codes in the Danish National Patient Registry: a population-based validation study. *BMJ Open* 2016;6(12):e012817. doi: 10.1136/bmjopen-2016-012817 [published Online First: 2016/12/13]
129. Ehrenstein V, Gammelager H, Schiodt M, et al. Evaluation of an ICD-10 algorithm to detect osteonecrosis of the jaw among cancer patients in the Danish National Registry of Patients. *Pharmacoepidemiol Drug Saf* 2015;24(7):693-700. doi: 10.1002/pds.3786 [published Online First: 2015/05/15]
130. Lash TL, Riis AH, Ostenfeld EB, et al. A validated algorithm to ascertain colorectal cancer recurrence using registry resources in Denmark. *Int J Cancer* 2015;136(9):2210-15. doi: 10.1002/ijc.29267
131. Cuthbertson CC, Kucharska-Newton A, Faurot KR, et al. Controlling for Frailty in Pharmacoepidemiologic Studies of Older Adults: Validation of an Existing Medicare Claims-based Algorithm. *Epidemiology* 2018;29(4):556-61. doi: 10.1097/EDE.0000000000000833 [published Online First: 2018/04/06]
132. Tapper EB, Korovaichuk S, Baki J, et al. Identifying Patients With Hepatic Encephalopathy Using Administrative Data in the ICD-10 Era. *Clin Gastroenterol Hepatol* 2019 doi: https://doi.org/10.1016/j.cgh.2019.12.017
133. Pedersen SA, Schmidt SAJ, Klausen S, et al. Melanoma of the Skin in the Danish Cancer Registry and the Danish Melanoma Database: A Validation Study. *Epidemiology* 2018;29(3):442-47. doi: 10.1097/ede.0000000000000802 [published Online First: 2018/01/18]
134. Benchimol EI, Guttmann A, Griffiths AM, et al. Increasing incidence of paediatric inflammatory bowel disease in Ontario, Canada: evidence from health administrative data. *Gut* 2009;58(11):1490-7. doi: 10.1136/gut.2009.188383 [published Online First: 2009/08/05]
135. Billionnet C, Alla F, Berigaud E, et al. Identifying atrial fibrillation in outpatients initiating oral anticoagulants based on medico-administrative data: results from the French national healthcare databases. *Pharmacoepidemiol Drug Saf* 2017;26(5):535-43. doi: 10.1002/pds.4192 [published Online First: 2017/03/16]
136. Holland-Bill L, Christiansen CF, Ulrichsen SP, et al. Validity of the International Classification of Diseases, 10th revision discharge diagnosis codes for hyponatraemia in the Danish National Registry of Patients. *BMJ Open* 2014;4(4):e004956. doi: 10.1136/bmjopen-2014-004956 [published Online First: 2014/04/25]
137. Olesen JB, Lip GY, Hansen ML, et al. Validation of risk stratification schemes for predicting stroke and thromboembolism in patients with atrial fibrillation: nationwide cohort study. *BMJ* 2011;342:d124. doi: 10.1136/bmj.d124
138. Deleuran T, Sogaard M, Froslev T, et al. Completeness of TNM staging of small-cell and non-small-cell lung cancer in the Danish cancer registry, 2004-2009. *Clin Epidemiol* 2012;4:39-44. doi: 10.2147/CLEP.S33315 [published Online First: 2012/09/01]
139. Krysko KM, Ivers NM, Young J, et al. Identifying individuals with multiple sclerosis in an electronic medical record. *Mult Scler* 2015;21(2):217-24. doi: 10.1177/1352458514538334 [published Online First: 2014/06/21]
140. Sacher AG, Le LW, Lau A, et al. Real-world chemotherapy treatment patterns in metastatic non–small cell lung cancer: Are patients undertreated? *Cancer* 2015;121(15):2562-69. doi: 10.1002/cncr.29386
141. Ettinger DS, Wood DE, Akerley W, et al. NCCN Guidelines Insights: Non-Small Cell Lung Cancer, Version 4.2016. *J Natl Compr Canc Netw* 2016;14(3):255-64. doi: 10.6004/jnccn.2016.0031 [published Online First: 2016/03/10]
142. Turner RM, Chen Y-W, Fernandes AW. Validation of a Case-Finding Algorithm for Identifying Patients with Non-small Cell Lung Cancer (NSCLC) in Administrative Claims Databases. *Frontiers in Pharmacology* 2017;8(883) doi: 10.3389/fphar.2017.00883
143. Kao W-H, Hong J-H, See L-C, et al. Validity of cancer diagnosis in the National Health Insurance database compared with the linked National Cancer Registry in Taiwan. *Pharmacoepidemiology and drug safety* 2018;27(10):1060-66. doi: 10.1002/pds.4267
144. Vandenbroucke JP, Sorensen HT. Chapter **. Clinical Epidemiology. In: Rothman KJ, Greenland S, Lash TL, eds. Modern Epidemiology. 4th ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins (in press).

145. Gini R, Dodd CN, Bollaerts K, et al. Quantifying outcome misclassification in multi-database studies: The case study of pertussis in the ADVANCE project. *Vaccine* 2020;38 Suppl 2:B56-B64. doi: 10.1016/j.vaccine.2019.07.045 [published Online First: 2019/11/05]
146. Lohse SR, Farkas DK, Lohse N, et al. Validation of spontaneous abortion diagnoses in the Danish National Registry of Patients. *Clin Epidemiol* 2010;2:247-50. doi: 10.2147/CLEP.S13815 [published Online First: 2010/12/15]
147. Acquavella J, Ehrenstein V, Schiodt M, et al. Design and methods for a Scandinavian pharmacovigilance study of osteonecrosis of the jaw and serious infections among cancer patients treated with antiresorptive agents for the prevention of skeletal-related events. *Clin Epidemiol* 2016;8:267-72. doi: 10.2147/CLEP.S107270 [published Online First: 2016/08/09]
148. Xue F, Ma H, Stehman-Breen C, et al. Design and methods of a postmarketing pharmacoepidemiology study assessing long-term safety of Prolia(R) (denosumab) for the treatment of postmenopausal osteoporosis. *Pharmacoepidemiol Drug Saf* 2013;22(10):1107-14. doi: 10.1002/pds.3477 [published Online First: 2013/07/17]
149. Gammelager H, Svaerke C, Noerholt SE, et al. Validity of an algorithm to identify osteonecrosis of the jaw in women with postmenopausal osteoporosis in the Danish National Registry of Patients. *Clin Epidemiol* 2013;5:263-7. doi: 10.2147/CLEP.S45226 [published Online First: 2013/08/16]
150. Schiodt M, Larsson Wexell C, Herlofson BB, et al. Existing data sources for clinical epidemiology: Scandinavian Cohort for osteonecrosis of the jaw - work in progress and challenges. *Clin Epidemiol* 2015;7:107-16. doi: 10.2147/CLEP.S71796 [published Online First: 2015/02/07]
151. Abstract 2436. Pharmacovigilance cohort study of osteonecrosis of the jaw and serious infection among cancer patients treated with denosumab or zoledronic acid in Denmark, Norway, and Sweden. ICPE All Access Virtual Event, September 2020; 2020; Virtual.
152. Abstract 3835. A multinational European study of anaphylaxis among recipients of intravenous Iron. ICPE All Access Virtual Event, September 2020; 2020; Virtual.

Figure legends

Figure 1 Point of care to RWD data point and back

Figure 1 Incidence rates (per 100,000 person-years) of pertussis infection identified in RWD in three countries using various RWD-based algorithms (horizontal bars) and using a gold standard (the dashed lines), i.e., national surveillance data reported to the European Centre for Disease Prevention and Control. For RWD-based composite algorithms, the grey bar represents cases detected only by the left-hand component (indicated in the label before the key Boolean operator 'OR'); the black bar represents cases detected by both components; the white bar represents cases detected by the right-hand component (indicated in the label after the key Boolean operator word 'OR'). Reproduced from Gini et al. 2020 *(permitted given full attribution).*[145]

Figure 2 An example of a decision process when considering validation studies/algorithm selection. As an example of a ready-to-use algorithm, consider ICD codes for spontaneous abortion recorded in the Danish National Patient Registry. International classifications of diseases have specific diagnostic codes for spontaneous abortion, and a validation study revealed that >95% of ICD records corresponded to a spontaneous abortion record in medical chart(reference standard), with slight variation by period, and type of hospital.[146] Therefore, the ICD-10 based algorithm is suitable for constructing cohorts of women with spontaneous abortions or for studying relative associations between an exposure and the outcome of spontaneous abortion. Use of the algorithm for studies evaluating absolute risks or risk differences for the outcome of spontaneous abortions will need evidence (or assumption) of high sensitivity of the algorithm and high completeness of the data source. The two were not examined in the validations study, but the assumptions may be defensible using the knowledge that most clinically apparent spontaneous abortions in Denmark are seen in hospital settings. (Very early events may not be clinically apparent even to the affected woman and are therefore generally difficult to quantify.) On the other and of the spectrum is osteonecrosis of the jaw – a side effect of antiresorptive therapy, whose risk is an important parameter of the therapy safety profile – exemplifies a condition without a clear-cut algorithm. Antiresorptive therapy is used in osteoporosis (low dose) in bone malignancies (high dose), and risk of osteonecrosis of the jaw is dose dependent. Although ICD-10 has a potentially useful code M87.1 "Osteonecrosis due to drugs", consistency of its use was unclear, including attribution to the specific agent. Furthermore, majority of the osteonecrosis cases affect bones other than the jaw. In evaluating suitability of the Danish National Patient Registry's for estimating risks of osteonecrosis of the jaw in postauthorisation safety studies of antiresorptive agents,[147];[148] a candidate algorithm consisting of a list ICD-10 codes originating from departments of oral and maxillofacial surgery

(to ensure jaw localization) was validated (external validation). The PPV of the candidate algorithm ranged from 20% in patients with osteoporosis to 42% in patients with cancer, and had a sensitivity of 73%.[129 149] *Because the studies' main objective was estimation of absolute risks of osteonecrosis of the jaw, the PPV and the sensitivity were deemed unsuitable for this purpose, cases of osteonecrosis of the jaw for the study were identified directly by at the departments of oral and maxillofacial surgery.*[150 151] *In a multinational European postauthorisation safety study of anaphylaxis following the use of intravenous iron preparations, the initial RWD-based algorithm for anaphylaxis originated from earlier US studies, whose PPV was estimated in the European setting using medical chart review; notably, application of a US-based algorithm in the European setting yielded a lower-than-expected risk of anaphylaxis, potentially indicative of differences in diagnostic or recording practices leading to record generation in the two types of setting.*[152]

Figure 3 Gold standard and reference standard

A: True positive, B: False negative, C: False positive, D: True negative

Inner circle represents positive reference standard.

a, b*, c*, d* represent misclassifications with reference standard*

Unbiased sensitivity = (a+a)/(a+a*+b+b*)*

Sensitivity against reference standard = (a+c)/(a+b+c*+d*)*

c and d* are consequently eliminated from sensitivity calculation through chart review against reference standard cases.*

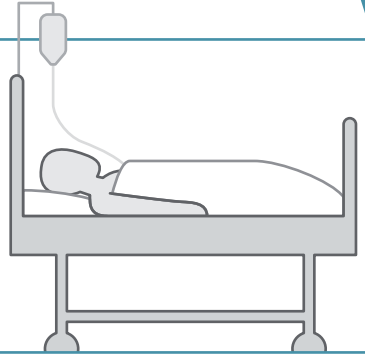
if (a:b=a:b*) holds, no bias. Otherwise, bias in sensitivity unavoidable.*

It is necessary to review the charts of all non-cases against reference standards (a+b*+c+d),*

Remarks to the National Academy of Sciences on the Impact of Real World Evidence on Medical

Product Development to determine the numbers of a and b*, possibly taking a tremendous effort.*

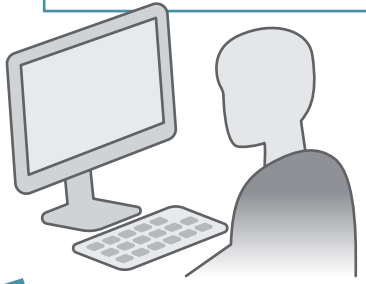
Characteristics of health care system and referral pattern
- Universal coverage
- Insurance-based coverage
- Employment-based coverage
- Access to specialist care



Clinical aspects
- Diagnostic process
- Treatment decisions
- Disease outcome

Generation of RCD/health sector
- Administrative data
- Discharge summaries
- Clinical notes

Type of data/updates
- Structured/coded
- Level of clinical detail
- Database maintenance/ updates/cleaning

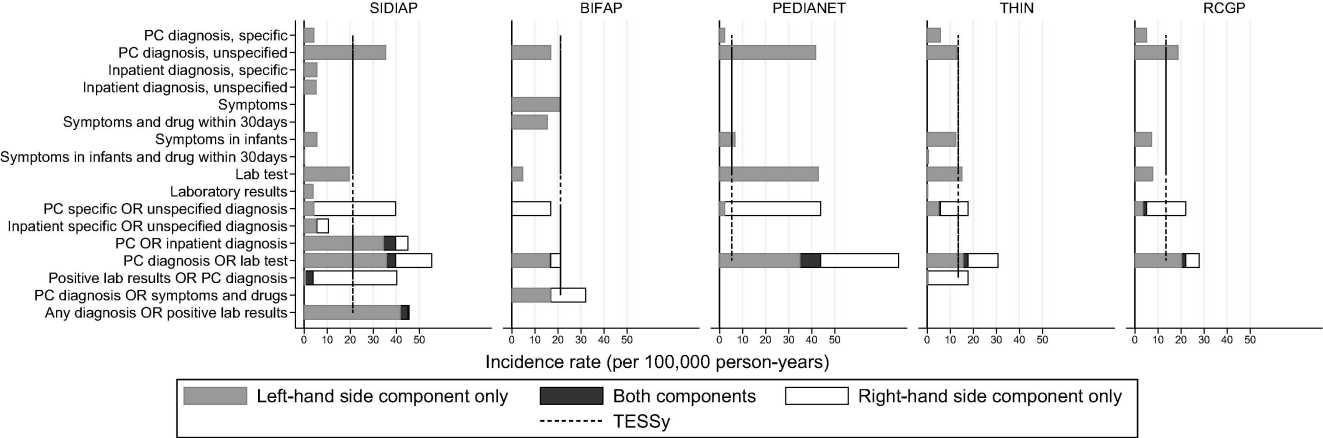


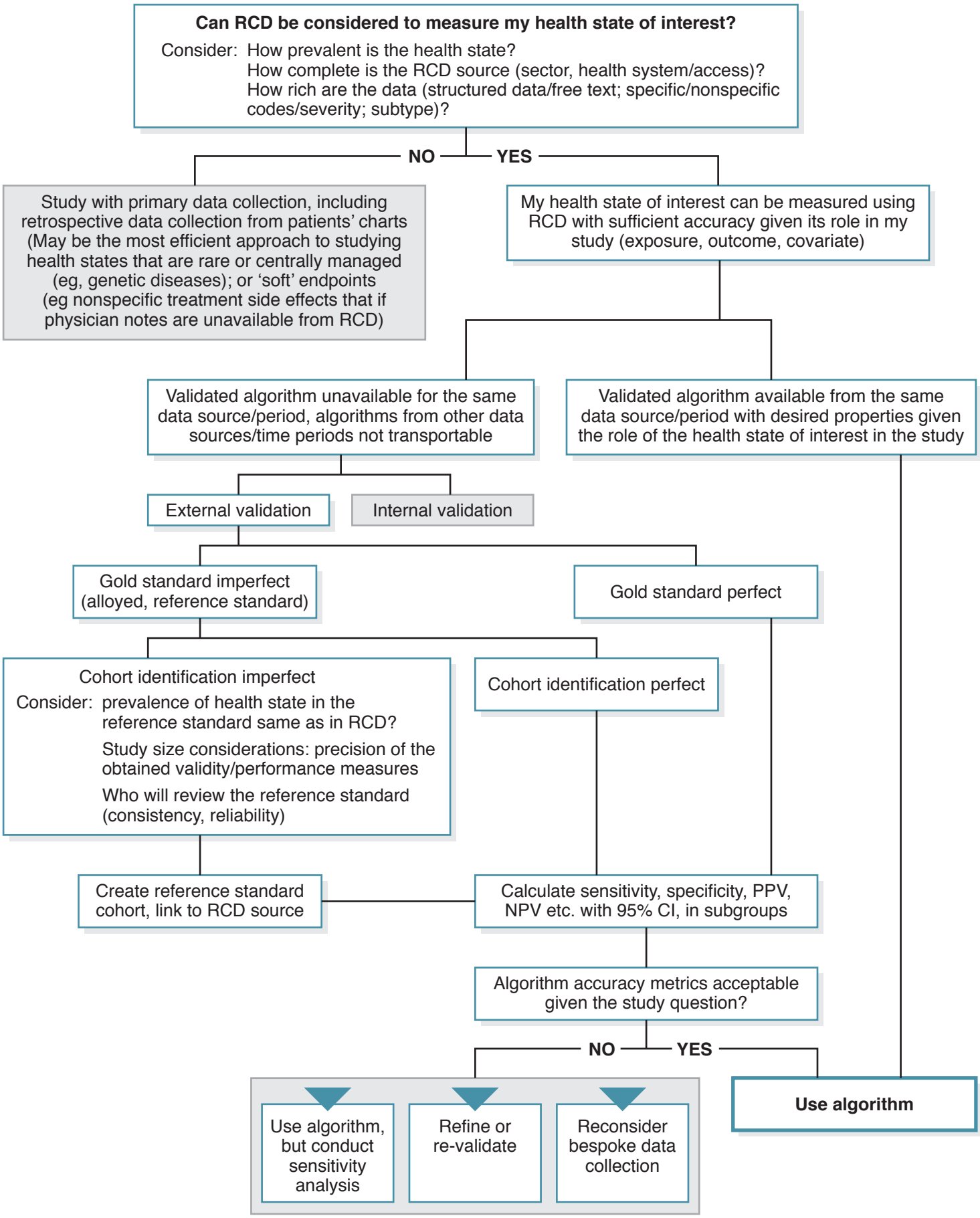
Study conduct
- Study population, design, setting, period
- Choice of RCD algorithms
- Epidemiologic expertise
- Clinical expertise

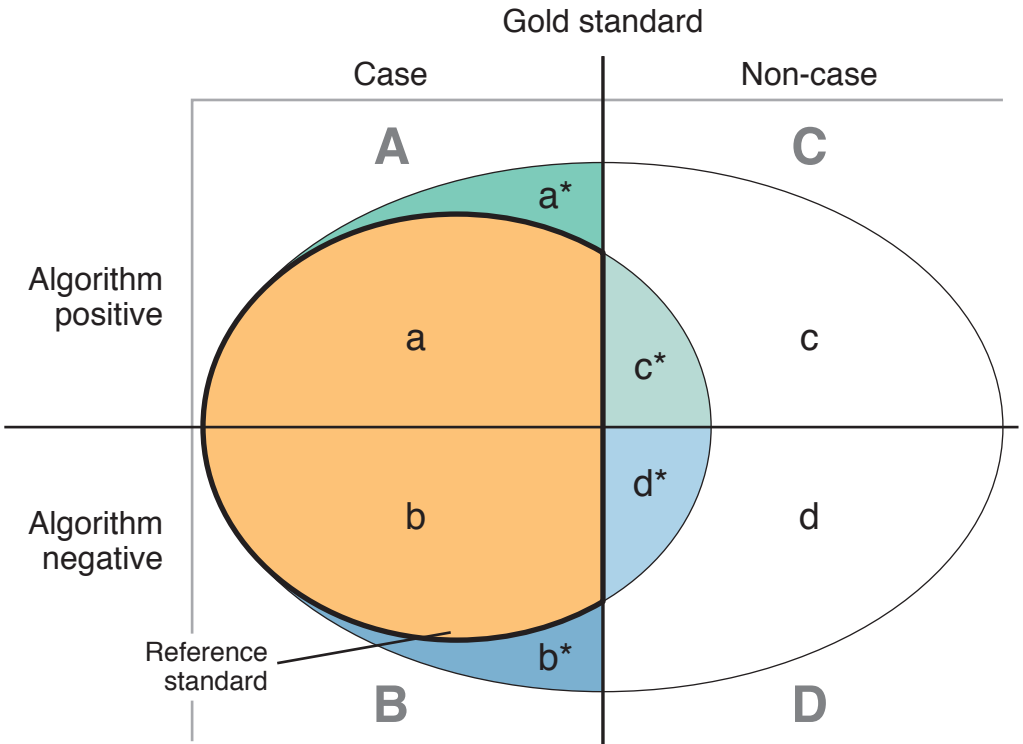
Evidence utilisation
- Meta-analyses
- Guidelines

Evidence generation
- Reporting results
- Interpretation of results

Patient-level events:
- Symptoms
- Care-seeking thresholds
- Age
- Sex
- Education
- Income


SIDIAP
BIFAP
PEDIANET
THIN
RCGP
PC diagnosis, specific
PC diagnosis, unspecified
Inpatient diagnosis, specific
Inpatient diagnosis, unspecified
Symptoms
Symptoms and drug within 30days
Symptoms in infants
Symptoms in infants and drug within 30days
Lab test
Laboratory results
PC specific OR unspecified diagnosis
Inpatient specific OR unspecified diagnosis
PC OR inpatient diagnosis
PC diagnosis OR lab test
Positive lab results OR PC diagnosis
PC diagnosis OR symptoms and drugs
Any diagnosis OR positive lab results
0 10 20 30 40 50
Incidence rate (per 100,000 person-years)
Left-hand side component only
Both components
Right-hand side component only
TESSy


Can RCD be considered to measure my health state of interest?
Consider: How prevalent is the health state?
How complete is the RCD source (sector, health system/access)?
How rich are the data (structured data/free text; specific/nonspecific codes/severity; subtype)?
NO
YES
Study with primary data collection, including retrospective data collection from patients' charts (May be the most efficient approach to studying health states that are rare or centrally managed (eg, genetic diseases); or 'soft' endpoints (eg nonspecific treatment side effects that if physician notes are unavailable from RCD)
My health state of interest can be measured using RCD with sufficient accuracy given its role in my study (exposure, outcome, covariate)
Validated algorithm unavailable for the same data source/period, algorithms from other data sources/time periods not transportable
Validated algorithm available from the same data source/period with desired properties given the role of the health state of interest in the study
External validation
Internal validation
Gold standard imperfect (alloyed, reference standard)
Gold standard perfect
Cohort identification imperfect
Consider: prevalence of health state in the reference standard same as in RCD?
Study size considerations: precision of the obtained validity/performance measures
Who will review the reference standard (consistency, reliability)
Cohort identification perfect
Create reference standard cohort, link to RCD source
Calculate sensitivity, specificity, PPV, NPV etc. with 95% CI, in subgroups
Algorithm accuracy metrics acceptable given the study question?
NO
YES
Use algorithm, but conduct sensitivity analysis
Refine or re-validate
Reconsider bespoke data collection
Use algorithm



A: True positive

B: False negative

C: False positive

D: True negative

Inner circle represents positive reference standard.

Misclassifications with reference standard: a*, b*, c*, d*

Unbiased sensitivity = (a+a*)/(a+a*+b+b*)

Sensitivity against reference standard
= (a+c*)/(a+b+c*+d*)

c* and d* are consequently eliminated from sensitivity calculation through chart review against reference standard cases.

if (a:b=a*:b*) holds, no bias. Otherwise, bias in sensitivity unavoidable.

However, it necessitates reviewing all non-reference standards (a*+b*+c+d) against chart to determine the numbers of a* and b*, possibly taking a tremendous effort.