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 IBenchimol, MD, PhD7,8,9

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
Word count, excluding title page, abstract, references, boxes and figures: approx. 6,136
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.\textsuperscript{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;\textsuperscript{4} and patient-reported events.\textsuperscript{5} RCD-based studies are playing an increasingly important role in decisions by regulators, payers, and in health technology assessment,\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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,\textsuperscript{23-27} potentially leading to wrong conclusions, and, ultimately, suboptimal clinical practice.\textsuperscript{28} Although misclassification is inevitable in any epidemiologic study, regardless of data collection methods,\textsuperscript{28} the extent, the nature, and the impact of misclassification for any given study using RCD may be difficult to quantify.\textsuperscript{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”.\textsuperscript{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.\textsuperscript{34-36} There have been several comprehensive efforts to validate sources of RCD,\textsuperscript{37-43} while guidelines for assessing and reporting validation studies\textsuperscript{30} have helped improve reporting transparency.\textsuperscript{44-51} Nevertheless, literature on validating RCD remains fragmented,\textsuperscript{31 52-55} and scientists have repeatedly raised awareness about the need for better education in the realm of validation studies.\textsuperscript{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);\textsuperscript{57} with the data-quality-focused regulatory strategy of the European Medicines Agency,\textsuperscript{58} and with the RWE transparency initiative.\textsuperscript{59}

**Manuscript development**

This paper was conceptualised as a contribution of the ongoing efforts to improve conduct and reporting of observational studies,\textsuperscript{34,44,51,60,61} specifically, validation studies in RCD sources.\textsuperscript{32,34,52} 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.\textsuperscript{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.\textsuperscript{57} Following an earlier study on RWD validation,\textsuperscript{30} we use the term *health state* for any event of interest measured in a
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). 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.

Algorithms in routinely collected data

An algorithm is “a completely defined set of operations that will produce the desired outcome” (Box 1). 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); 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). 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. ICD-10-CM [clinical modification] vs. Read codes); and database maintenance routines (e.g., frequency of updates, plausibility checks). Each link in the chain of events from point of care to a research data point harbours a potential source of error (Figure 1).

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. 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). 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. 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.
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. Machine learning, increasingly used to aid development of RCD algorithms, is valuable as a screening tool for identifying candidate algorithm components for subsequent validation. 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.\textsuperscript{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,\textsuperscript{72} corresponding to the registry’s completeness with respect to each drug, as it captures only by-prescription dispensings.\textsuperscript{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\textsuperscript{74} as Sweden banned over-the-counter sales following a cardiovascular safety concern\textsuperscript{75}). Completeness is also affected by changes in diagnostic process (e.g., introduction of troponin as the main diagnostic biomarker of myocardial infarction\textsuperscript{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.\textsuperscript{30,76} These measures, covered in standard texts and tutorial articles,\textsuperscript{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)\textsuperscript{79}, diagnostic odds ratio (DOR)\textsuperscript{80}); indicators of performance (positive and negative predictive values (PPV and NPV)); and measures of agreement (kappa statistic\textsuperscript{77,81}). (Other sources have used, the term ‘bias parameters’\textsuperscript{56} or validity indices\textsuperscript{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).\textsuperscript{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.\textsuperscript{30,56} Many validation studies are not designed to estimate all validity measures.

Sensitivity and PPV are the most commonly reported measures\textsuperscript{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). 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.

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.\textsuperscript{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\textsuperscript{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.\textsuperscript{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.\textsuperscript{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 discussion56). 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
population, we could exclude from that population patients with a hospital diagnosis of venous thromboembolism or a record of arthroplasty. Although this exclusion criterion will remove some eligible patients, we may decide to pay this price for the sake of obtaining a valid estimation of comparative effectiveness and safety. The measures we take to increase the specificity of the eligibility criteria algorithm will come at a price of potential selection bias, if patients with hospital-diagnosed AF systematically differ from patients with AF initiating anticoagulants seen in all medical care settings. Patients seen in hospitals are likely to represent the most severe spectrum of the disease; this bias would be avoided by the sensitive eligibility algorithm, described above. Selection bias due to inclusion of patients with other indications could be avoided by keeping such patients in the study population and conducting a stratified analysis. With rare diseases, we may prioritise eligibility criteria algorithms with high PPV and specificity, even at a price of reduced precision, to avoid “contamination” of the study cohort with false-positives; alternatively one may opt for a hybrid approach by identifying potentially eligible patients based on RCD algorithms and subsequently eliminate false-positive cases by reviewing medical records.

RCD algorithms for exposure

RCD-based algorithms establishing patients’ treatment/drug exposure rely on routine records of prescriptions (e.g., in EHR data), dispensing (e.g., in pharmacy administrative data), or administrations (e.g., in hospitalization data) to measure treatment initiation, duration, dosage, discontinuation. In identifying new users of a drug, the length of the washout (treatment-free) period may produce severe misclassification between the status of new and prevalent user, which may lead to underestimation of risks of early side effects. Incorrect allocation of on-/off-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. 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) 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. 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.

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 infrastructure. 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. 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. 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, 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 to full-scale re-validation in the new setting (Figure 3).

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, 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. 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. 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. 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. 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” 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). For polytomous variables created by categorising continuous variables, the direction of measurement error depends on a category. 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.

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 range from simple calculations examining the impact of misclassification of one variable at a time to simultaneous assessment of the impact of errors of multiple study variables (bias analysis) to using imputation methods to correct for misclassification. 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. 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).\textsuperscript{118} Other techniques for correcting study results for measurement error include hierarchical semi-Bayes methods,\textsuperscript{114} or bootstrap imputation.\textsuperscript{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\textsuperscript{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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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”. 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.
### Box 1. Glossary of key terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition/elaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithm</td>
<td>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.</td>
</tr>
<tr>
<td>Alloyed gold standard</td>
<td>See Reference standard</td>
</tr>
<tr>
<td>Area under the curve (AUC)</td>
<td>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.</td>
</tr>
<tr>
<td>Completeness</td>
<td>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.</td>
</tr>
<tr>
<td>Confounder</td>
<td>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.</td>
</tr>
<tr>
<td>Covariate</td>
<td>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.</td>
</tr>
<tr>
<td>External validation</td>
<td>Validation of events of a health state identified with an RWD-based algorithm in a population external to a given study.</td>
</tr>
<tr>
<td>Diagnostic odds ratio (DOR)</td>
<td>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.</td>
</tr>
<tr>
<td>Gold standard</td>
<td>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.</td>
</tr>
<tr>
<td>Health state</td>
<td>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.</td>
</tr>
<tr>
<td>Information bias</td>
<td>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.</td>
</tr>
<tr>
<td>Internal validation</td>
<td>Validation of events of a health state identified with an RWD-based algorithm in a given study.</td>
</tr>
<tr>
<td>Measurement error</td>
<td>See Information bias.</td>
</tr>
</tbody>
</table>
| 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
<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negative predictive value (NPV)</td>
<td>Proportion of patients who truly negative for a health state among all those</td>
</tr>
<tr>
<td>Positive predictive value (PPV)</td>
<td>Proportion of patients who truly positive for a health state among all those</td>
</tr>
<tr>
<td>Real-world data (RWD)</td>
<td>Data that originate in the course of routine clinical practice, usually by</td>
</tr>
<tr>
<td>Real-world evidence (RWE)</td>
<td>Evidence generated using RWD.</td>
</tr>
<tr>
<td>Reference standard</td>
<td>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.</td>
</tr>
<tr>
<td>Receiver operating characteristic (ROC) curve</td>
<td>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.</td>
</tr>
<tr>
<td>Routinely collected health data (RCD)</td>
<td>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.</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>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.</td>
</tr>
<tr>
<td>Specificity</td>
<td>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.</td>
</tr>
<tr>
<td>Exposure</td>
<td>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.</td>
</tr>
<tr>
<td>Outcome</td>
<td>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.</td>
</tr>
<tr>
<td>Validation cohort</td>
<td>Study population used to evaluate measures of algorithm validity.</td>
</tr>
<tr>
<td>Validity (internal)</td>
<td>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.</td>
</tr>
<tr>
<td>Health state or event being validated</td>
<td>Country</td>
</tr>
<tr>
<td>--------------------------------------</td>
<td>---------</td>
</tr>
<tr>
<td>In-hospital chemotherapy treatment</td>
<td>Denmark</td>
</tr>
<tr>
<td>Start, duration and end of prescribed medicine</td>
<td>Finland</td>
</tr>
<tr>
<td>Cardiac interventions</td>
<td>Denmark</td>
</tr>
<tr>
<td>Osteonecrosis of the jaw</td>
<td>Denmark</td>
</tr>
<tr>
<td>Colorectal cancer recurrence</td>
<td>Denmark</td>
</tr>
<tr>
<td>Frailty in the elderly</td>
<td>United States</td>
</tr>
<tr>
<td>Condition</td>
<td>Country</td>
</tr>
<tr>
<td>---------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Hepatic encephalopathy</td>
<td>United States</td>
</tr>
<tr>
<td>Malignant melanoma</td>
<td>Denmark</td>
</tr>
<tr>
<td>Childhood-onset inflammatory bowel disease</td>
<td>Canada</td>
</tr>
<tr>
<td>Use of oral anticoagulants due to atrial fibrillation</td>
<td>France</td>
</tr>
<tr>
<td>Five events deemed to be important in studying risks and benefits of treatments:</td>
<td>Eight European countries</td>
</tr>
<tr>
<td>Condition</td>
<td>Country</td>
</tr>
<tr>
<td>-----------</td>
<td>---------</td>
</tr>
<tr>
<td>Hyponatremia</td>
<td>Denmark</td>
</tr>
<tr>
<td>Asthma</td>
<td>Canada</td>
</tr>
<tr>
<td>Serious infection</td>
<td>Denmark</td>
</tr>
<tr>
<td>Cardiovascular risk scores previously developed and validated in other sources of data: CHADS2 (congestive heart failure and stroke)</td>
<td>Denmark</td>
</tr>
<tr>
<td>Condition/Study</td>
<td>Country</td>
</tr>
<tr>
<td>----------------</td>
<td>---------</td>
</tr>
<tr>
<td>Heart failure, hypertension, age ≥ 75 years, diabetes mellitus, previous stroke/transient ischaemic attack</td>
<td>CHA2DS2-VASc</td>
</tr>
<tr>
<td>Registered cancer staging according to TNM at diagnosis</td>
<td>Canada</td>
</tr>
<tr>
<td>Multiple sclerosis (MS)</td>
<td>Japan</td>
</tr>
<tr>
<td>Lung cancer</td>
<td>United States</td>
</tr>
<tr>
<td>Multiple malignancies</td>
<td>Taiwan</td>
</tr>
<tr>
<td>Multiple malignancies</td>
<td>United States</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>----------------------</td>
</tr>
</tbody>
</table>

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 Nomasco Classification of Surgical Procedures
<table>
<thead>
<tr>
<th>Measure</th>
<th>Range</th>
<th>Description</th>
<th>Formulas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>[0, 1]</td>
<td>Proportion of persons with a given health state according to gold/reference standard who are classified as such by an algorithm</td>
<td>Health state status based on gold/reference standard Positive Negative</td>
</tr>
</tbody>
</table>
| Specificity                  | [0, 1]| Proportion of persons without a given health state according to gold/reference standard who are classified as such by an algorithm        | $Sensitivity = \frac{A}{(A + B)}$  
Specificity $= \frac{D}{(C + D)}$  
$PPV = \frac{A}{(A + C)}$  
$NPV = \frac{D}{(B + D)}$ |
| 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                      | Chuback et al and Benchimol et al provide formulae connecting sensitivity and specificity, PPV and NPV with prevalence of a health state.36,64                                                         |
| 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.              |                                                                                                                                                                                                          |
| 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 al36) 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) |
**Diagnostic odds ratio (DOR)**

A 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. 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.

\[
DOR = \frac{A}{B} (\text{True Positive}) / \frac{C}{D} (\text{False Positive})
\]

**Kappa statistic**

Kappa statistic is used to quantify interrater variability. 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, agreement may be qualified on a scale 'less than chance' to 'almost perfect' (Viera and Garrett 2005).

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

<table>
<thead>
<tr>
<th>Kappa Agreement</th>
<th>Health state status according to data source 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0</td>
<td>Positive</td>
</tr>
<tr>
<td>0.00-0.20</td>
<td>a, b, m_1</td>
</tr>
<tr>
<td>0.21-0.40</td>
<td>Slight</td>
</tr>
<tr>
<td>0.41-0.60</td>
<td>c, d, m_0</td>
</tr>
<tr>
<td>0.61-0.80</td>
<td>Fair</td>
</tr>
<tr>
<td>0.81-0.99</td>
<td>Moderate</td>
</tr>
<tr>
<td></td>
<td>Substantial</td>
</tr>
<tr>
<td></td>
<td>Almost perfect</td>
</tr>
</tbody>
</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

\[
\text{Observed agreement } p_{\text{obs}} = \frac{(a + d)}{n}
\]

\[
\text{Expected agreement } p_{\text{exp}} = \left[ \frac{(a_n)}{n} \times \frac{(m_1)}{n} \right] + \left[ \frac{(c_n)}{n} \times \frac{(m_0)}{n} \right]
\]

\[
\text{Kappa statistic } \kappa = \frac{(p_{\text{obs}} - p_{\text{exp}})}{(1 - p_{\text{exp}})}
\]
References


try: a review of content, data


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]


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.


101. ACCESS STUDY PLACEHOLDER - AVAILABLE IN JUNE 2021 FOR CITATION. PLACEHOLDER


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.

**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).*
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. 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 end 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, 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%. 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. 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.

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 = \( \frac{a+a^*}{a+a^*+b+b^*} \)

Sensitivity against reference standard = \( \frac{a+c^*}{a+b+c^*+d^*} \)

\( c^* \) 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

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

Clinical aspects
- Diagnostic process
- Treatment decisions
- Disease outcome

Evidence generation
- Reporting results
- Interpretation of results

Evidence utilisation
- Meta-analyses
- Guidelines

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
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)?

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

External validation

Internal validation

Gold standard imperfect ( alloyed, reference standard)

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)

Create reference standard cohort, link to RCD source

Gold standard perfect

Cohort identification perfect

Calculate sensitivity, specificity, PPV, NPV etc. with 95% CI, in subgroups

Algorithm accuracy metrics acceptable given the study question?

NO YES

Use algorithm

Use algorithm, but conduct sensitivity analysis

Refine or re-validate

Reconsider bespoke data collection
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 = \( \frac{a+a^*}{a+a^*+b+b^*} \)

Sensitivity against reference standard = \( \frac{a+c^*}{a+b+c^*+d^*} \)

\( c^* \) 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.