Drug Safety (2019) 42:1297–1309
https://doi.org/10.1007/s40264-019-00851-0

LEADING ARTICLE

Transparent Reporting on Research Using Unstructured Electronic Health Record Data to Generate 'Real World' Evidence of Comparative Effectiveness and Safety

Shirley V. Wang[1] · Olga V. Patterson[2,3] · Joshua J. Gagne[1] · Jeffrey S. Brown[4] · Robert Ball[5] · Pall Jonsson[6] · Adam Wright[7] · Li Zhou[7] · Wim Goettsch[8,9] · Andrew Bate[10]

Published online: 26 August 2019

Abstract
Research that makes secondary use of administrative and clinical healthcare databases is increasingly influential for regulatory, reimbursement, and other healthcare decision-making. Consequently, there are numerous guidance documents on reporting for studies that use 'real-world' data captured in administrative claims and electronic health record (EHR) databases. These guidance documents are intended to improve transparency, reproducibility, and the ability to evaluate validity and relevance of design and analysis decisions. However, existing guidance does not differentiate between structured and unstructured information contained in EHRs, registries, or other healthcare data sources. While unstructured text is convenient and readily interpretable in clinical practice, it can be difficult to use for investigation of causal questions, e.g., comparative effectiveness and safety, until data have been cleaned and algorithms applied to extract relevant information to structured fields for analysis. The goal of this paper is to increase transparency for healthcare decision makers and causal inference researchers by providing general recommendations for reporting on steps taken to make unstructured text-based data usable for comparative effectiveness and safety research. These recommendations are designed to be used as an adjunct for existing reporting guidance. They are intended to provide sufficient context and supporting information for causal inference studies involving use of natural language processing- or machine learning-derived data fields, so that researchers, reviewers, and decision makers can be confident in their ability to evaluate the validity and relevance of derived measures for exposures, inclusion/exclusion criteria, covariates, and outcomes for the causal question of interest.

1 Introduction

There is widespread interest in making greater use of the rich clinical information contained in administrative claims and electronic health records (EHRs) to generate 'real-world' evidence (RWE) on the safety, effectiveness, or value (cost effectiveness) of medical interventions [1–11]. Insurance claims data provide longitudinal records of patient contact with the healthcare system via billing codes and are routinely used to generate RWE. Similarly, research conducted with EHR databases can leverage the rich clinical history contained in patient medical records (e.g., free-text notes, reports, laboratory results, problem lists, etc.). However, these data are collected for administrative and clinical purposes, not research. Therefore, it is critically important to

Key Points

Existing guidance for reporting on research using healthcare databases does not differentiate between structured and unstructured information contained in the electronic health records (EHRs), registries, or other healthcare data sources.

Context and supporting detail for natural language processing- and machine learning-derived fields is extremely important to allow causal inference researchers as well as decision makers (e.g., health technology assessment, payers, regulators) to evaluate whether derived phenotypes, outcomes, or other clinical events are relevant to the question they seek to address.

We provide recommendations on reporting to increase transparency about the process of making unstructured text-based data usable for causal inference, pharmacoeconomic evaluations, and utilization studies.

✉ Shirley V. Wang
swang1@bwh.harvard.edu

Extended author information available on the last page of the article

use algorithms that validly capture patient phenotypes and clinical events when making secondary use of these data to generate evidence on the comparative safety and effectiveness of medical interventions.

Because research that makes secondary use of administrative and clinical healthcare databases is becoming increasingly influential for regulatory, reimbursement, and other healthcare decision-making, numerous guidance documents on reporting for observational studies [12, 13] as well as specific reporting recommendations for studies that make secondary use of such data have been developed [14–18]. These guidance documents are intended to improve transparency, reproducibility, and the ability to evaluate validity and relevance of design and analysis decisions. However, existing guidance for reporting on research using healthcare databases do not differentiate between structured and unstructured information contained in the EHR, registries, or other healthcare data sources. While unstructured text is convenient and readily interpretable in clinical practice, it can be difficult to use for investigation of causal inference questions regarding comparative effectiveness or safety, as well as pharmacoeconomics or utilization research, until the data have been cleaned and algorithms applied to extract relevant information to structured fields for analysis [19].

There is a growing body of research focused on developing natural language processing (NLP) and machine learning (ML) methods to facilitate classification and identification of computable phenotypes to define exposures, inclusion/exclusion criteria, covariates, and outcomes from EHRs in a way that is accurate and reliable for research [20]. Many academic organizations, health systems, and commercial organizations routinely use NLP- and ML-derived data from the EHR to support research and clinical practice. NLP and free-text analysis approaches have been used for tasks including extraction of clinical concepts such as smoking status and other risk factors [21–23], identification of medication discrepancies [24–26], detection of potential medicinal effects in spontaneous reporting systems [26–30], and evaluation of drug–disease [31] relationships.

Large distributed data networks such as PCORnet [32] and the US Food and Drug Administration's (FDA) Sentinel program [33, 34], which focus on conducting comparative effectiveness and safety research, are moving toward supplementing administrative claims data in their common data models with computable phenotypes and clinical events derived from unstructured data using NLP and ML [35–39]. Other distributed data networks such as EU-ADR and Asian-DURG (Asian Drug Utilization Research Group) independently implement common protocols across the network. These distributed networks have found that research results can vary due to underlying differences in data sources as well as processes for extracting information from unstructured data [40–42].

There are several important considerations when NLP or ML algorithms are reused to extract phenotypes or clinical events across data systems, time, and purposes. First, clinical documentation practices and jargon may vary across healthcare facilities. High-performing NLP and ML algorithms within one EHR system may use contextual information that is not applicable in other systems where clinical documentation processes and norms differ. Second, while NLP and ML are extremely useful tools that facilitate measurement of exposure, exclusion criteria, covariates, and outcomes, additional design and analysis methods must be applied for causal inference. For example, while temporality of exclusion/covariate assessment windows and follow-up relative to cohort entry is critical for causal inference research, timing with respect to cohort entry may not be a major consideration for studies focused on developing NLP or ML algorithms to classify patients, events, or notes. Context and supporting detail for NLP- and ML-derived fields is extremely important for causal inference researchers as well as decision makers (e.g., health technology assessment, payers, regulators) to evaluate whether derived phenotypes, outcomes, or other clinical events are relevant to the question they seek to address.

The general process to extract and validate information from the EHR is straightforward, typically starting with design of the study and establishing criteria for defining the reference standard. We deliberately use the term 'reference standard' rather than 'ground truth' or 'gold standard' to emphasize that determinations made by human reviewers are not the fixed truth; rather, they are the reviewer's interpretation based on inherently limited EHR data that were generated to document clinical care (usually in the context of an agreed upon standard such as a published case definition). After laying out the study design and criteria for the reference standard, researchers generally acquire access to EHR data, create a labeled corpus of data (where the reference standard is determined by human reviewers), then develop and evaluate NLP or ML algorithms. While the general process is straightforward, the details can be quite complex. Important scientific details are often not publicly reported [43]. Greater clarity at each step in this process would facilitate reviewer and decision-maker evaluation of the validity as well as relevance of NLP and ML outputs that are reused in different research investigations (e.g., a library of phenotypes).

The goal of this paper is to increase transparency for healthcare decision makers by providing general recommendations for reporting on steps taken to make unstructured text-based data usable for causal comparative effectiveness and safety research. Similar principles apply to audio or image-based information extraction but are beyond the scope of this article. These recommendations are intended to be used as an adjunct for existing reporting guidance such

as those produced by a joint task force between two professional societies focused on research using healthcare databases, the European Network of Centres for Pharmacoepidemiology and Pharmacovigilance, the US FDA, RECORD (REporting of studies Conducted using Observational Routinely-collected health Data), TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis), and others [14–18]. The level of detail recommended is intended to provide causal inference researchers reusing NLP- or ML-derived data fields as well as healthcare decision makers with sufficient context and supporting details for the derivation process that they can be confident in the validity and relevance of the derived measures for exposures, inclusion/exclusion criteria, covariates, and outcomes for their research questions. The recommendations focus on transparent reporting of NLP- and ML-related research, not on best practices for doing NLP or ML.

This paper was written by a wide range of stakeholders who generate or make decisions based on RWE involving information extracted from unstructured EHR text, including academics (pharmacoepidemiologists and informaticians), regulators, and industry and health technology assessors. We outline nine items that we would like to see in publications and other public reports of NLP- or ML-related research to effectively transfer knowledge about how the evidence was generated. These items are not all-encompassing; for example, reporting on research ethics, conflicts of interest, and Institutional Review Board approval is also important. We anticipate that greater transparency on the nine items outlined would increase our ability to evaluate the quality, relevance, and validity of information extracted from EHRs used to support generation of RWE. We note that greater transparency does not equate to study quality; rather, transparency makes it possible for reviewers to *assess* study quality. We provide examples of transparent reporting in each of these areas from published studies (Table 1).

2 Reporting Recommendations

2.1 Data

2.1.1 Describe Characteristics of Data Acquired by Investigators (Including Sources and Types of Data)

Providing information on the types of data available and acquired by investigators provides context for the study, algorithm performance, and generalizability. For example, are the data based on narrative reports only or do the researchers also utilize information from problem lists, laboratory results, ordered or dispensed medications, and other structured fields?

EHR data may go through multiple transformations before reaching a human reviewer, being fed through an NLP system or to a ML algorithm. There can be loss of fidelity from the original EHR when reviewers see an extensible markup language (XML) export of the medical chart from the EHR or review scanned images of medical record reports that have been converted to a machine-readable format with optical character recognition software. De-identification software can also introduce noise. For example, de-identification software that scrambles names can scramble references to diseases that appear to be names (e.g., Crohn's, Parkinson's, Hashimoto's disease).

2.1.2 Describe Transformations Performed on Data Received by Investigators to Pre-process or Clean Data

After acquiring access to EHR data for a research study, investigators may perform additional data transformations to clean or otherwise pre-process data prior to conducting the research of interest. Detailing the steps taken to create a corpus of documents for human review or compatible with NLP or ML software would be helpful to understand the extent of data manipulation to pre-process and clean the EHR data. This may include information such as software used, and whether there was any validation of fidelity at each step of the transformation.

2.2 Methods

2.2.1 Provide Key Details of Study Design to Identify the Study Cohort and/or Sampling Frame

Clarity about the study cohort or sampling frame is necessary to understand the context of NLP and ML performance metrics. For example, when developing algorithms for identifying outcome, it is critical that the events occur after the start of follow-up. It is important to describe which patients were included, the timeframe for the performance evaluation, and how charts were sampled. The types of details to report are covered in reporting guidelines such as those produced by RECORD [15] and a joint task force between the International Society of Pharmacoepidemiology and the International Society of Pharmacoeconomics and Outcomes Research [14]. High-level details may be summarized in a design diagram [44].

2.2.2 Provide the Criteria Used by Chart Reviewers to Determine the Reference Standard for Health Events or Conditions of Interest

Chart reviewers are given criteria or annotation instructions that guide how they determine the reference standard, i.e.,

Table 1 Meta-data about phenotypes or clinical events derived from unstructured data to publicly report in appendices of publications, reports, or data dictionaries of shared research databases

Reporting recommendation		Example
A. Data		
1	Describe characteristics of data acquired by investigators (including sources and types of data)	*Data characteristics:* "All data from this study were extracted from the publicly available MIMIC-III database. MIMIC-III contains de-identified data, per Health Insurance Portability and Accountability Act (HIPAA) privacy rules, on over 58,000 hospital admissions in the intensive care units (ICU) at Beth Israel Deaconess Medical Center from June 2001 to October 2012 … The unstructured clinical notes include discharge summaries (n = 52,746), nursing progress notes (n = 812,128), physician notes (n = 430,629), electrocardiogram (ECG) reports (n = 209,058), echocardiogram reports (n = 45,794), and radiology reports (n = 896,478). We excluded clinical notes that were related to any imaging results (ECG_Report, Echo_Report, and Radiology_Report). We extracted notes from MIMIC-III with the following data elements: patient identification number (SUBJECT_ID), hospital admission identification number (HADM_IDs), intensive care unit stay identification number (ICUSTAY_ID), note type, note date/time, and note text." [56] *Comparing SQL-based extraction of data from an EHR to a data warehouse vs. manual extraction of the same variables:* "For electronic data abstraction, we investigated variables extracted from Clarity™, the relational data repository from Epic™ [an EHR software platform], whose contents are populated from the production EHR on a nightly basis. Clarity™ allows execution of complex queries returning large sets of data. We extracted data for the same set of neonates, using their Medical Record Numbers (MRNs), along with associated data from 1,444 linked maternal records …To identify discrepancies, a subset of randomly selected charts was manually reviewed using the production EHR. Using Stata™ version 11, electronically-extracted and [production EHR] content was compared for accuracy …" [47]
2	Describe transformations performed on data received by investigators to pre-process or clean data	*Digitizing charts with optical character recognition:* "We first used the ABBYY Fine Reader Optical Character Recognition software to digitize these 62 typewritten charts. The digitization process involved scanning, high-level error-checking, and editing of the digitized subset (N = 62) of the 122 medical records from the original study as well as the creation of a database of electronic and searchable records … For each chart, the ABBYY-scanned sections were categorized, consistently named, and attached in text format … The text files were organized into 12 different sections: allergies, clinic notes, discharge notes, ED forms, extraction forms, flowsheet data, follow-up plan, inpatient (IP) notes, labs, orders, visit notes, and vitals." [38] *Pre-processing tweets:* "Before analysis, the data were preprocessed to remove the URL, screen handles (@username), retweet indicators, and non-ascii characters. Data were further normalized by removing capital letters, numbers, punctuations, and whitespaces from the tweets. Terms were filtered out to remove single characters like "d," "e," which do not convey any meaning about the topics in the corpus, and top words like "and," "so," etc were removed for the classification stage. Each tweet was represented as a feature vector of the words present in the tweet using unigrams." [57]

Table 1 (continued)

Reporting recommendation		Example
B. Methods		
3	Provide key details of study design to identify the study cohort and/or sampling frame	*Active comparator cohort study comparing new initiators of ACEIs and β-blockers on risk of angiedema:* "We used an inception cohort design to identify patients 18 years or older with an outpatient dispensing of an oral formulation of the following medications as a single ingredient or as combination products with non-study drugs between January 1, 2001, and December 31, 2010: (1) an ACEI (benazepril, captopril, enalapril, fosinopril, lisinopril, moexipril, quinapril, perindopril, ramipril, or trandolapril), (2) an ARB (candesartan, eprosartan, irbesartan, losartan, olmesartan, telmisartan, or valsartan), (3) aliskiren, or (4) a β-blocker (acebutolol, atenolol, bisoprolol, carvedilol, labetalol, metoprolol, nebivolol, pindolol, propranolol, or timolol), used as a common reference group. We refer to the dispensing date of the first prescription of any of the study drug as the index date. To be eligible for the study, these patients must also have met each of the following criteria during the 183-day period preceding the index date: (1) continuous health plan enrollment with pharmacy and medical benefits, (2) no prescription for any other study drug, and (3) no diagnosis of angioedema in any care setting." [58]
4	Provide criteria used by chart reviewers to determine the reference standard for health events or conditions of interest	*Automated extraction of DVT events from narrative radiology reports:* "The 4000 [radiology] reports were initially coded by a clinical expert … Positive radiology reports for a DVT were those where a thrombus was identified in the proximal deep veins of the lower extremities (eg, external iliac, common femoral, deep femoral, or popliteal veins), in the deep distal veins of the lower extremities (eg, peroneal and posterior tibial veins), or in the deep veins of the upper extremities (eg, brachial, radial, ulnar, axillary, subclavian). Negative cases included those where no thrombus was identified or where a thrombus was identified in a superficial vein of the lower extremity (eg, saphenous), in a superficial vein of the upper extremity (eg, cephalic), or in a perforating vein of the lower extremity but not extending into a deep vein. Radiologic examinations finding evidence of chronic thrombosis were coded as negative." [59]
5	Describe which data were accessible to human reviewers versus software (if different)	*NLP system to capture breast cancer recurrence:* "Clinical documents for all patients were obtained from Group Health EHR systems. They included all available machine-readable pathology reports, progress notes, and radiology reports during patient-specific follow-up periods. Paper reports and reports scanned as images in the EHR were available to the … human review[er] and so are reflected in our reference standard. However, they could not be processed by our NLP system." [48]

Table 1 (continued)

Reporting recommendation		Example
6	Provide full description of the NLP or ML algorithms, including details on inputs and outputs (for primary, secondary, and sensitivity analyses)	*NLP-derived attributes for ML algorithm to identify patients with RA:* "Twenty-one attributes of the patients' medical records were generated for RA …The details of these attributes can be found in supplementary table 1 … We applied the published logistic regression model to the 21 attributes … At Northwestern, the attributes … [were retrieved] using the HITEx NLP system with a set of customized regular expression queries (see supplementary table 2). At Vanderbilt, NLP was performed using KMCI, which was applied without customization to identify UMLS concepts with clinical note section tagging (using SecTag) and negation. The concepts were selected by hand from a list automatically generated by finding related concepts (using relationships such as parent–child found in the UMLS MRREL file) around each of the key terms, such as 'Rheumatoid Arthritis' (see supplementary table 3) … The betas from the logistic regression models, after using the lasso method to select the most influential attributes, are shown in supplementary table 1" [60] *Text mining with ML for adverse event detection:* "The unstructured narratives were tokenized, lemmatized, and part-of-speech tagged using Stanford CoreNLP … We incorporated the number of sentences and average number of words, nouns, verbs, adjectives, and adverbs per sentence as non-semantic features. We additionally counted the number of redacted terms in each narrative, which typically corresponded to dates detailing the time course of the adverse event." [26] *Automated inference of patient problems:* "A total of 17 problem inference rules were developed … The complete list and description of these rules is provided in online appendix A … Final rules used coded and free-text problem recognition and one or more of the following to infer patient problems: (a) related billing codes, (b) related medications, and/or (c) related laboratory data or vital signs." [61]
C. Results		
7	Describe the study population	Detailed by other reporting guidance documents. This may include an attrition table showing numbers of patients as eligibility criteria are applied as well as tables for baseline characteristics and outcome frequencies
8	Provide concordance/inter-rater reliability performance metrics	*Automated medication discrepancy detection:* "Medication matches and discrepancies between the clinical notes and the prescription lists were double annotated by two annotators using the Knowtator plug-in for Protégé. Both annotators [were] English speakers (one clinical research nurse and one with an associate of applied science degree in health information science), with at least one year of clinical text annotation experience … After the annotation process, differences between the annotators' decisions were resolved under the supervision of an annotation manager (bachelor's degree with more than four years of experience in clinical text annotation) and the inter-annotator agreement (IAA) was calculated using F-value to measure the agreement." [24]

Table 1 (continued)

Reporting recommendation		Example
9	Provide multiple measures of performance for algorithms in training and validation data (for primary, secondary, and sensitivity analyses)	*Classification of anaphylaxis:* "… standard metrics of recall (ie, sensitivity), precision (ie, positive predictive value), and F-measure (harmonic mean of recall and precision) were calculated based on [reviewer] classification of charts conducted in the Mini-Sentinel project." [38] *Automated inference of patient problems:* The goal was to automatically populate problem lists in EHRs. The corpus included a random sample of 100,000 patient records at a single teaching hospital. The patients were sampled from a pool of patients with at least one progress note recorded in the EHR within the 2 years prior to the study. To estimate sensitivity, specificity, PPV, and NPV, the authors made two assumptions: (1) patients with a problem documented on the problem list were true positives; and (2) patients with no laboratory tests, billing diagnosis or procedure codes, or medications related to a specific medical problem were true negatives. The status for patients with at least one laboratory test, diagnosis or procedure code, or medication related to the problem but no problem list entry was considered unknown. The authors evaluated various rules for defining whether a problem was present and conducted a chart review on a random sample of patients where the rule classified the problem as present and a random sample where the rule classified the problem as absent. Sensitivity, specificity, PPV, and NPV were estimated by inverse probability weighting based on sampling fraction "To validate the final version of each rule and to guard against over-fitting of the rules against the training set … we drew a second random sample of 100 000 patients from the same population as the initial sample but excluding patients in the initial sample. For each of the final rules, the same classification and chart review process was carried out, and sensitivity, specificity, and positive and negative predictive values were calculated using the same procedure[s]." [61]

ACEI ACE inhibitor, *ARB* angiotensin receptor blocker, *DVT* deep vein thrombosis, *ED* emergency department, *EHR* electronic health record, *HITEx* Health Information Text Extraction, *KMCI* KnowledgeMap Concept Identifier, *MIMIC-III* Medical Information Mart for Intensive Care, *ML* machine learning, *NLP* natural language processing, *NPV* negative predictive value, *PPV* positive predictive value, *RA* rheumatoid arthritis, *UMLS* unified medical language system

whether a patient has or does not have a given phenotype or clinical event. Laying out these criteria is particularly important when the phenotype or clinical event is complex or documentation in the EHR may be ambiguous. Reporting these criteria, the level of training provided to reviewers, whether they are domain/clinical experts, and the process for adjudication when reviewers provide discordant assessments will provide important transparency on what exactly the reference standard is measuring. Clarity on the criteria used to determine the reference standard from EHR documentation is important for determining the relevance of the information extracted by NLP or ML. The criteria used may be more relevant for some study questions than others.

Without clear reporting on the criteria to define the reference standard, the relevance of derived phenotypes for the question of interest and the degree of misclassification of important study variables (e.g., exposure, exclusion criteria, covariates, outcomes) can be masked. For example, a researcher could use an algorithm that identifies 'diabetic' patients from free-text EHR notes for a comparative evaluation of anti-diabetic drugs. However, without detail on how the phenotype was defined and possible limitations of this definition, it will not be apparent to reviewers or decision makers whether the algorithm is expected to be sensitive or specific, whether it identifies new-onset versus established diabetes mellitus, how the algorithm considers timing (if at all), or how it distinguishes between type 1, type 2, and gestational diabetes [45, 46]. Publishing the criteria used by chart reviewers would be an important step toward increasing transparency.

2.2.3 Describe Which Data were Accessible to Human Reviewers Versus Software (If Different)

When EHR data viewable by human reviewers and software are comprehensive and aligned, this allows evaluation of the 'efficacy' of the NLP system or ML algorithm at classifying phenotypes or identifying clinical events in ideal conditions. For example, when the reference standard and NLP are based on fully aligned data sources, the annotator agreement is viewed as the ceiling of the possible NLP system performance. However, when human reviewers and software have access to a mix of overlapping and non-overlapping portions of the EHR due to logistical or other considerations, evaluation focuses on the 'effectiveness' of the NLP or ML algorithms in practice. As an example of the latter, there are situations where a human reviewer may be able to access and review the entire medical record for presence or absence of a condition, whereas only structured data and a subset of notes from the EHR are accessible for the NLP or ML algorithm.

Providing information on the types of EHR data available to human reviewers versus software provides important context on the quality and completeness of data used to derive the reference standard. Incomplete access to the full patient record or misalignment of data available to human reviewers and/or software can inform assessment of performance, validity, and generalizability of NLP or ML algorithms.

2.2.4 Provide Full Description of the Natural Language Processing or Machine Learning Algorithms, Including Details on Inputs and Outputs (for Primary, Secondary, and Sensitivity Analyses)

We expand on TRIPOD [18] recommendations for reporting on prediction model specification to address the additional layer of complexity that comes with use of unstructured data in NLP and ML algorithms.

For full analytic reproducibility, sharing of code and data is encouraged. However, there are often privacy and intellectual property considerations that prevent sharing of data, data derivatives, or code. Additionally, even if data and code can be shared, without supplemental metadata in clear, natural language, the complexity of NLP systems and ML algorithms may lack transparency for many decision makers and other stakeholders. In the absence of the ability to share data and code, key details to increase understanding about behind-the-scenes decisions built into NLP systems and ML algorithms can still be reported to facilitate evaluation of validity and appropriateness for a given research question.

We advocate for details of inputs and outputs, which can be provided in different formats, e.g., an input–output process diagram, a de-identified sample chart containing highlights for relevant portions at each step of the process, and/or a natural language summary of each step taken by the code. Some details that would be helpful for increasing transparency and reproducibility of information extraction systems include the names and version of any software packages used, a citation or appendix with ontologies used to map clinical concepts (e.g., RxNorm, SNOMED-CT, homegrown), choices for the inputs and tuning parameters included in NLP systems (e.g., negation, pruning, word sense disambiguation, word order, conjunctions) or ML, as well as details of the outputs (e.g., algorithm, rule, model, coefficients, R objects). Naming the algorithm or software is a start, but describing or listing out specifications for the algorithm, configuration settings, and computing environments will be important for reviewers and other investigators to understand how variables are derived from unstructured data.

2.3 Results

2.3.1 Describe the Study Population

The rationale for reporting on characteristics of the study population is described in numerous other reporting guidance documents [14–18]. This includes items such as an attrition table (showing patient numbers as eligibility criteria are applied), baseline characteristics of the derived population, as well as the number and timing of outcomes of interest. It allows the investigator and reviewers to describe and assess whether the frequency of a derived variable is consistent with expectation (e.g., that the outcome incidence or a covariate prevalence looks approximately correct). The same rationale applies in studies that develop or use derived information from NLP and ML algorithms.

2.3.2 Provide Concordance/Inter-Rater Reliability Performance Metrics

Human review can be subjective and fallible. If human reviewers have poor agreement, this would appropriately lower confidence in the reliability of the reference standard used to train NLP systems and ML algorithms [47]. Measures of inter-rater reliability include kappa, intra-class correlation coefficients, F-value, and others [48].

2.3.3 Provide Multiple Measures of Performance for Algorithms in Training and Test/Validation Data (for Primary, Secondary, and Sensitivity Analyses)

Providing performance metrics for how well NLP or ML algorithms correctly identify phenotypes of clinical events is necessary to evaluate the degree of anticipated misclassification and whether the performance is fit-for-purpose in a given study. To make such assessments, in addition to reporting the positive predictive value (PPV; given the algorithm assessed a condition is present, the probability that it actually is present, also known as 'precision'), it is important to provide other metrics such as negative predictive value (NPV; given the algorithm assessed that a condition is not present, the probability it is actually not present), sensitivity (the proportion of true positives that are correctly identified by the algorithm, also known as 'recall'), and specificity (the proportion of true negatives that are correctly identified by the algorithm) [49]. To avoid overly optimistic assessments of performance, these metrics should be reported for a sample of test (validation) data that was not used to train the algorithms.

Both PPV and NPV are dependent on the prevalence of the condition being evaluated [49]. If the sampling frame from the underlying cohort is known, sampled cases and controls can be weighted by the sampling fraction when estimating performance metrics [50]. While sensitivity and specificity are not dependent on prevalence of the condition, these metrics could vary in different populations.

The importance of having high specificity versus high sensitivity for an algorithm can vary depending on how the measurement will be used. For example, when using an algorithm for ischemic stroke as an exclusion criterion, high sensitivity may be more important than a high PPV to ensure the study population does not include patients with prior stroke. In contrast, when conducting a comparative study evaluating risk of ischemic stroke as an outcome using a relative measure of effect, a high specificity may take precedence.

Other metrics that may be relevant in classification studies include the c-statistic (area under the receiver operating characteristic curve) and integrated discrimination improvement statistic [51]. Calibration measures may also be relevant for prediction studies [52].

3 Discussion

New evidence is useful to decision makers when it reduces decision uncertainty. Confidence in the credibility, quality, and therefore impact of RWE that relies on information extracted from unstructured data would be improved with more transparent data provenance and research processes [14, 53]. Clear reporting on processes, protocols, and other scientific decisions would facilitate reproducibility of NLP and ML methods as well as assessment of validity and relevance when applied in other studies with different data sources and populations [14, 53].

We have highlighted the importance of reporting nine types of meta-data when computable phenotypes or clinical events derived with NLP and ML are used in one-off research studies or stored as reusable structured elements in a data warehouse for research purposes. This information could be publicly shared via internet links, citations, or appendices in peer-reviewed publications and reports. Future studies would be able to retain the chain of data provenance by citing relevant meta-data.

In this paper, we have not looked to address the issue of selecting which algorithm(s) to use when defining a phenotype or clinical event from unstructured data. Nor have we resolved an important question that may arise for distributed data networks or other multi-user data warehouses regarding the relative utility of (1) storing NLP and ML coded elements from unstructured data in relational tables as a shared resource versus (2) preserving the unstructured native data and storing a library of NLP and ML algorithms to derive desired elements. Although one-off research projects can develop, evaluate, and select algorithms that are tuned to perform well for their purpose,

when populating a data model for distributed data networks that are a shared resource for multiple researchers, it may be difficult to identify a one-size-fits-all algorithm for many clinical elements. The 'best' could vary depending on many factors, including the specific use case, the hospital system, and temporal changes.

For example, the high performance of an algorithm in one hospital system may not be readily transportable to another, and performance of an algorithm within the same system could deteriorate over time as practice patterns and the EHR system change. Multiple versions of algorithms for similar concepts could exist over time and in health systems within data warehouses used by multiple researchers. The algorithms applied by any member of a distributed data network to define phenotypes or clinical events may or may not perform well in EHR data held by other members. Thus, it would be valuable for research conducted in distributed data networks to maintain clear documentation of the information extraction process and have date-stamped evaluation of performance of these algorithms in the data systems in which they will be applied.

Subtle differences in NLP and ML algorithms and their performance in different populations could have substantive impact on research findings [54]. In a research environment that is increasingly focused on distributed data networks and reuse of derived data elements from previously developed NLP or ML algorithms, clarity on how these data elements were created and demonstration of validity in different data systems will be critical to the credibility and value of information extracted from distributed EHR networks to generate RWE.

If the recommended meta-data were provided, it would reduce decision-maker uncertainty regarding reproducibility, generalizability, and robustness of evidence in different studies, data sources, and populations. For example, providing decision makers with clarity on which algorithms were used to identify outcomes across studies (or sites) and how the reference standard criteria for those outcomes were defined would allow them to determine whether they are comparing 'apples to apples' or 'apples to oranges'. Demonstrated value in decreasing decision uncertainty could increase acceptance of using evidence from unstructured healthcare data to inform decision-making. Furthermore, the availability of the recommended meta-data may circumvent the need for decision maker to have access to privacy-protected patient data.

4 Conclusion

Our reporting recommendations are designed to bolster trust and encourage appropriate use of information extracted from NLP or ML to support RWE generation. Given manuscript word limits, the information could be provided in detailed technical appendices or separate publication of a protocol. Indeed, some journals are already moving toward more stringent requirements for reporting on data, analytics, design, and other elements of the research process [55]. It may be idealistic to expect all recommended items to be publicly reported for every study. Nevertheless, we hope to increase recognition of what is currently missing and what, if made transparent, would better support and inform decision-maker evaluation of quality of measurement, validity, and relevance of RWE from administrative and clinical healthcare databases [38, 47, 48, 56–61].

Compliance with Ethical Standards

Conflict of Interest Dr Shirley V Wang has received salary support on investigator-initiated grants from Novartis Pharmaceuticals Corporation, Boehringer Ingelheim, and J&J to Brigham and Women's Hospital, and was a consultant to Aetion, Inc., all for unrelated work. Dr Olga V. Patterson receives research grants from the following for-profit organizations: Amgen Inc., Anolinx LLC, AstraZeneca Pharmaceuticals LP, Genentech Inc., Genomic Health, Inc., Gilead Sciences Inc., HITEKS Solutions Inc., Merck & Co., Inc., Northrop Grumman Information Systems, Novartis International AG, PAREXEL International Corporation, and Shire PLC through the University of Utah or Western Institute for Biomedical Research. Dr Patterson also receives research funding from the following federal and non-profit organizations: Agency for Healthcare Research and Quality, Brigham and Women's Hospital, Centers for Disease Control and Prevention, Department of Defense, Department of Veterans Affairs, Intermountain Healthcare, National Heart, Lung, and Blood Institute, National Institute on Alcohol Abuse and Alcoholism, National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institute of General Medical Sciences, National Institute of Standards and Technology, National Library of Medicine, National Science Foundation, Patient Centered Outcomes Research Institute, and RAND Corporation. Dr Joshua J. Gagne has received salary support from grants from Novartis Pharmaceuticals Corporation and Eli Lilly and company to Brigham and Women's Hospital and is a consultant to Aetion, Inc. and to Optum, Inc., all for unrelated work. Dr Andrew Bate is an employee and shareholder of Pfizer. The views expressed in this paper are those of Dr Bate and may not necessarily reflect those of Pfizer. Dr Robert Ball is an author of US Patent 9,075,796, "Text mining for large medical text datasets and corresponding medical text classification using informative feature selection". Dr Li Zhou has received research funding from the Agency of Healthcare Research and Quality (AHRQ): R01HS022728 and CRICO/RMF. Dr Jeffrey S Brown, Dr Pall Jonsson, Dr Adam Wright, and Dr Wim Goettsch have no conflicts of interest that are directly relevant to the content of this article. The views expressed in this article are the personal views of the authors and may not be understood or quoted as being made on behalf of or reflecting the position of the US Food and Drug Administration or the National Institute for Health and Care Excellence.

Funding This study was supported by funds from the Division of Pharmacoepidemiology and Pharmacoeconomics and Brigham and Women's Hospital.

Ethical Approval Not applicable; no data were analyzed.

Patient Consent No patient contact or data were involved.

References

1. National Academies of Sciences, Engineering, and Medicine; Health and Medicine Division; Board on Health Sciences Policy; Forum on Drug Discovery, Development, and Translation. Real-world evidence generation and evaluation of therapeutics: proceedings of a workshop. Washington, DC: National Academies Press (US). 2017.
2. Psaty BM, Breckenridge AM. Mini-Sentinel and regulatory science—big data rendered fit and functional. N Engl J Med. 2014;370(23):2165–7.
3. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc. 2014;21(4):578–82.
4. Oliveira JL, Lopes P, Nunes T, Campos D, Boyer S, Ahlberg E, et al. The EU-ADR web platform: delivering advanced pharmacovigilance tools. Pharmacoepidemiol Drug Saf. 2013;22(5):459–67.
5. Collaborators A, Andersen M, Bergman U, Choi N-K, Gerhard T, Huang C, et al. The Asian Pharmacoepidemiology Network (AsPEN): promoting multi-national collaboration for pharmacoepidemiologic research in Asia. Pharmacoepidemiol Drug Saf. 2013;22(7):700–4.
6. Behrman RE, Benner JS, Brown JS, McClellan M, Woodcock J, Platt R. Developing the Sentinel system—a national resource for evidence development. N Engl J Med. 2011;364(6):498–9.
7. Suissa S, Henry D, Caetano P, Dormuth CR, Ernst P, Hemmelgarn B, et al. CNODES: the Canadian network for observational drug effect studies. Open Med. 2012;6(4):e134–40.
8. Trifiro G, Coloma PM, Rijnbeek PR, Romio S, Mosseveld B, Weibel D, et al. Combining multiple healthcare databases for postmarketing drug and vaccine safety surveillance: why and how? J Intern Med. 2014;275(6):551–61.
9. Engel P, Almas MF, De Bruin ML, Starzyk K, Blackburn S, Dreyer NA. Lessons learned on the design and the conduct of post-authorization safety studies: review of 3 years of PRAC oversight. Br J Clin Pharmacol. 2017;83(4):884–93.
10. Eichler H-G, Hurts H, Broich K, Rasi G. Drug regulation and pricing—can regulators influence affordability? N Engl J Med. 2016;374(19):1807–9.
11. Makady A, Ham RT, de Boer A, Hillege H, Klungel O, Goettsch W, et al. Policies for use of real-world data in health technology assessment (HTA): a comparative study of six HTA agencies. Value Health. 2017;20(4):520–32.
12. von Elm E, Altman DG, Egger M, Pocock SJ, Gotzsche PC, Vandenbroucke JP, et al. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296.
13. Vandenbroucke JP, von Elm E, Altman DG, Gøtzsche PC, Mulrow CD, Pocock SJ, et al. Strengthening the reporting of observational studies in epidemiology (STROBE): explanation and elaboration. PLoS Med. 2007;4(10):e297.
14. Wang SV, Schneeweiss S, Berger ML, Brown J, de Vries F, Douglas I, et al. Reporting to improve reproducibility and facilitate validity assessment for healthcare database studies V1.0. Pharmacoepidemiol Drug Saf. 2017;26(9):1018–32.
15. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies conducted using observational routinely-collected health Data (RECORD) statement. PLoS Med. 2015;12(10):e1001885.
16. EMA. ENCePP guide on methodological standards in pharmacoepidemiology. London: EMA; 2014.
17. Us FDA. Guidance for industry and FDA staff: best practices for conducting and reporting pharmacoepidemiologic safety studies using electronic healthcare data. Rockville: US FDA; 2013.
18. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350:g7594.
19. Wong A, Plasek JM, Montecalvo SP, Zhou L. Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges. Pharmacotherapy. 2018;38(8):822–41.
20. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc. 2016;23(5):1007–15.
21. Weiss LS, Zhou X, Walker AM, Ananthakrishnan AN, Shen R, Sobel RE, et al. A case study of the incremental utility for disease identification of natural language processing in electronic medical records. Pharm Med. 2018;32(1):31–7.
22. Walker AM, Zhou X, Ananthakrishnan AN, Weiss LS, Shen R, Sobel RE, et al. Computer-assisted expert case definition in electronic health records. Int J Med Inform. 2016;86:62–70.
23. Khalifa A, Meystre S. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. J Biomed Inform. 2015;58:S128–32.
24. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, et al. An end-to-end hybrid algorithm for automated medication discrepancy detection. BMC Med Inform Decis Mak. 2015;15:37.
25. White RW, Wang S, Pant A, Harpaz R, Shukla P, Sun W, et al. Early identification of adverse drug reactions from search log data. J Biomed Inform. 2016;59:42–8.
26. Han L, Ball R, Pamer CA, Altman RB, Proestel S. Development of an automated assessment tool for MedWatch reports in the FDA adverse event reporting system. J Am Med Inform Assoc. 2017;24(5):913–20.
27. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform. 2015;53:196–207.
28. Strandell J, Caster O, Bate A, Norén N, Edwards IR. Reporting patterns indicative of adverse drug interactions. Drug Saf. 2011;34(3):253–66.
29. Botsis T, Buttolph T, Nguyen MD, Winiecki S, Woo EJ, Ball R. Vaccine adverse event text mining system for extracting features from vaccine safety reports. J Am Med Inform Assoc. 2012;19(6):1011–8.
30. Wunnava S, Qin X, Kakar T, Kong X, Rundensteiner EA, Sahoo SK, et al. One size does not fit all: an ensemble approach towards information extraction from adverse drug event narratives. In: Proceedings of the 11th international joint conference on biomedical engineering systems and technologies, vol 5. HEALTHINF. 2018. p. 176–188
31. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease—drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc. 2008;15(1):87–98.
32. Califf RM. The patient-centered outcomes research network: a national infrastructure for comparative effectiveness research. N C Med J. 2014;75(3):204–10.
33. Ball R, Robb M, Anderson SA, Dal Pan G. The FDA's sentinel initiative—a comprehensive approach to medical product surveillance. Clin Pharmacol Ther. 2016;99(3):265–8.
34. US FDA. Safety: FDA's sentinel initiative. http://www.fda.gov/Safety/FDAsSentinelInitiative/ucm2007250.htm. Accessed Jan 2016.
35. Duke-Margolis Center for Health Policy. Discussion guide. Improving the efficiency of outcome validation in the Sentinel System. Washington, DC: Duke-Margolis Center for Health Policy. 2018.

36. PCORNet. PTNP-CCR. PCORnet common data model. 2016. http://www.pcornet.org/pcornet-common-data-model/. Accessed Jan 2016.
37. Brown JB, N; Curtis, L; Raebel, MA; Haynes, K, Rosofsky, R. Sentinel common data model. 2017. https://www.sentinelinitiative.org/sentinel/data/distributed-database-common-data-model/sentinel-common-data-model. Accessed 14 June 2018.
38. Ball R, Toh S, Nolan J, Haynes K, Forshee R, Botsis T. Evaluating automated approaches to anaphylaxis case classification using unstructured data from the FDA sentinel system. Pharmacoepidemiol Drug Saf. 2018;27(10):1077–84.
39. Seninel. Surveillance tools. Health outcome of interest validations and literature reviews. https://www.sentinelinitiative.org/sentinel/surveillance-tools/validations-lit-review. Accessed 11 Dec 2018.
40. Huerta C, Abbing-Karahagopian V, Requena G, Oliva B, Alvarez Y, Gardarsdottir H, et al. Exposure to benzodiazepines (anxiolytics, hypnotics and related drugs) in seven European electronic healthcare databases: a cross-national descriptive study from the PROTECT-EU Project. Pharmacoepidemiol Drug Saf. 2016;25(Suppl. 1):56–65.
41. Lai ECC, Stang P, Yang YHK, Kubota K, Wong ICK, Setoguchi S. International multi-database pharmacoepidemiology: potentials and pitfalls. Curr Epidemiol Rep. 2015;2(4):229–38.
42. Pratt N, Andersen M, Bergman U, Choi N-K, Gerhard T, Huang C, et al. Multi-country rapid adverse drug event assessment: the Asian Pharmacoepidemiology Network (AsPEN) antipsychotic and acute hyperglycaemia study. Pharmacoepidemiol Drug Saf. 2013;22(9):915–24.
43. Wang S, Verpillat P, Rassen J, Patrick A, Garry E, Bartels D. Transparency and reproducibility of observational cohort studies using large healthcare databases. Clin Pharmacol Ther. 2016;99(3):325–32.
44. Schneeweiss S, Rassen JA, Brown JS, Rothman KJ, Happe L, Arlett P, et al. Graphical depiction of longitudinal study designs in health care databases. Ann Intern Med. 2019;170(6):398–406.
45. Datta-Nemdharry P, Thomson A, Beynon J. Opportunities and challenges in developing a cohort of patients with type 2 diabetes mellitus using electronic primary care data. PloS One. 2016;11(11):e0162236.
46. Reeves D, Springate DA, Ashcroft DM, Ryan R, Doran T, Morris R, et al. Can analyses of electronic patient records be independently and externally validated? The effect of statins on the mortality of patients with ischaemic heart disease: a cohort study with nested case–control analysis. BMJ Open. 2014;4(4):e004952.
47. Shiloach M, Frencher SK, Steeger JE, Rowell KS, Bartzokis K, Tomeh MG, et al. Toward robust information: data quality and inter-rater reliability in the American College of Surgeons National Surgical Quality Improvement Program. J Am Coll Surg. 2010;210(1):6–16.
48. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol. 2012;8(1):23–34.
49. Rothman KJ, Greenland S, Lash TL. Modern epidemiology. Philadelphia: Lippincott, Williams & Wilkins; 2008.
50. van Zaane B, Vergouwe Y, Donders ART, Moons KGM. Comparison of approaches to estimate confidence intervals of post-test probabilities of diagnostic test results in a nested case–control study. BMC Med Res Methodol. 2012;12:166.
51. Pencina MJ, D'Agostino RB, Massaro JM. Understanding increments in model performance metrics. Lifetime Data Anal. 2013;19(2):202–18.
52. Demler OV, Paynter NP, Cook NR. Tests of calibration and goodness-of-fit in the survival setting. Stat Med. 2015;34(10):1659–80.
53. Berger ML, Sox H, Willke R, Brixner D, Eichler H-G, Goettsch W, 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;20(8):1003–8
54. Requena G, Huerta C, Gardarsdottir H, Logie J, González-González R, Abbing-Karahagopian V, et al. Hip/femur fractures associated with the use of benzodiazepines (anxiolytics, hypnotics and related drugs): a methodological approach to assess consistencies across databases from the PROTECT-EU project. Pharmacoepidemiol Drug Saf. 2016;25(Suppl 1):66–78.
55. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015;348(6242):1422–5.
56. Sarmiento RF, Dernoncourt F. Improving patient cohort identification using natural language processing. In: MIT Critical Data, editor. Secondary analysis of electronic health records. Springer, Cham. 2016, pp 405–17.
57. Miller M, Banerjee T, Muppalla R, Romine W, Sheth A. What are people tweeting about zika? An exploratory study concerning its symptoms, treatment, transmission, and prevention. JMIR Public Health Surveill. 2017;3(2):e38.
58. Toh S, Reichman ME, Houstoun M, Ross Southworth M, Ding X, Hernandez AF, et al. Comparative risk for angioedema associated with the use of drugs that target the renin-angiotensin-aldosterone system. Arch Intern Med. 2012;172(20):1582–9.
59. Tian Z, Sun S, Eguale T, Rochefort CM. Automated extraction of VTE events from narrative radiology reports in electronic health records: a validation study. Med Care. 2017;55(10):e73–80.
60. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc. 2012;19(e1):e162–9.
61. Wright A, Pang J, Feblowitz JC, Maloney FL, Wilcox AR, Ramelson HZ, et al. A method and knowledge base for automated inference of patient problems from structured data in an electronic medical record. J Am Med Inform Assoc. 2011;18(6):859–67.

Affiliations

Shirley V. Wang[1] · Olga V. Patterson[2,3] · Joshua J. Gagne[1] · Jeffrey S. Brown[4] · Robert Ball[5] · Pall Jonsson[6] · Adam Wright[7] · Li Zhou[7] · Wim Goettsch[8,9] · Andrew Bate[10]

1 Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont St Suite 303, Boston, MA 02120, USA

2 Division of Epidemiology, Internal Medicine, School of Medicine, University of Utah, Salt Lake City, UT, USA

3 VA Salt Lake City Health Care System, Salt Lake City, UT, USA

4 Department of Population Medicine, Harvard Pilgrim Health Care Institute and Harvard Medical School, Boston, MA, USA

5 Food and Drug Administration, Silver Spring, MD, USA

6 National Institute for Health and Care Excellence, London, UK

7 Department of General Internal Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

8 The National Healthcare Institute (ZIN), Diemen, The Netherlands

9 Division of Pharmacoepidemiology and Clinical Pharmacology, Utrecht University, Utrecht, The Netherlands

10 Pfizer, London, UK