Journal of the American Medical Informatics Association, 26(11), 2019, 1189–1194
doi: 10.1093/jamia/ocz119
Advance Access Publication Date: 12 August 2019
Research and Applications

Research and Applications

Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies

Tina Hernandez-Boussard,[1,2,3] Keri L Monda,[4,5] Blai Coll Crespo,[4] and Dan Riskin[1,3,6]

[1]Department of Medicine, Stanford University, Stanford, California, USA, [2]Department of Biomedical Data Science, Stanford University, Stanford, California, USA, [3]Department of Surgery, Stanford University School of Medicine, Stanford, California, USA, [4]The Center for Observational Research and Medical Affairs, Amgen, Inc., Thousand Oaks, California, USA, [5]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA and [6]Verantos Inc, Menlo Park, California, USA

Corresponding Author: Dan Riskin, MD, Verantos, Inc, 325 Sharon Park Dr. Suite 730, Palo Alto, CA 94025, USA; science@verantos.com

Received 15 January 2019; Revised 17 May 2019; Editorial Decision 12 June 2019; Accepted 19 June 2019

ABSTRACT

Objective: With growing availability of digital health data and technology, health-related studies are increasingly augmented or implemented using real world data (RWD). Recent federal initiatives promote the use of RWD to make clinical assertions that influence regulatory decision-making. Our objective was to determine whether traditional real world evidence (RWE) techniques in cardiovascular medicine achieve accuracy sufficient for credible clinical assertions, also known as "regulatory-grade" RWE.
Design: Retrospective observational study using electronic health records (EHR), 2010–2016.
Methods: A predefined set of clinical concepts was extracted from EHR structured (EHR-S) and unstructured (EHR-U) data using traditional query techniques and artificial intelligence (AI) technologies, respectively. Performance was evaluated against manually annotated cohorts using standard metrics. Accuracy was compared to pre-defined criteria for regulatory-grade. Differences in accuracy were compared using Chi-square test.
Results: The dataset included 10 840 clinical notes. Individual concept occurrence ranged from 194 for coronary artery bypass graft to 4502 for diabetes mellitus. In EHR-S, average recall and precision were 51.7% and 98.3%, respectively and 95.5% and 95.3% in EHR-U, respectively. For each clinical concept, EHR-S accuracy was below regulatory-grade, while EHR-U met or exceeded criteria, with the exception of medications.
Conclusions: Identifying an appropriate RWE approach is dependent on cohorts studied and accuracy required. In this study, recall varied greatly between EHR-S and EHR-U. Overall, EHR-S did not meet regulatory grade criteria, while EHR-U did. These results suggest that recall should be routinely measured in EHR-based studes intended for regulatory use. Furthermore, advanced data and technologies may be required to achieve regulatory grade results.

Key words: real world evidence, electronic health records, cardiovascular medicine, regulatory-grade, performance measures

BACKGROUND AND SIGNIFICANCE

With advances in modern medicine, average lifespan has expanded and patients have become more complex.[1] While randomized clinical trials (RCT) provide a foundation for clinical evidence, individual trials to assess treatment for a broad condition, such as hypertension or high cholesterol, may be less applicable to typical patients with multiple comorbidities.[2,3] Trials are often performed in highly controlled environments with narrow inclusion and exclusion criteria, which reduces their generalizability and external validity.[4,5] Highly protocolled care in an RCT may differ substantially from interventions in routine settings.

Based on these concerns over trial expense, lack of generalizability from selected patients to real world care, and lack of generalizability from protocolled care to real world care, there is an increasing drive to augment RCTs with real world data (RWD). RWD is information on medical interventions gathered from routine clinical care. These data may better reflect the general population seeking treatment for a particular condition rather than selective patients enrolled in an RCT.[6,7] RWD, whether claims data or extracted information from electronic health records (EHR), are often analyzed to produce real world evidence (RWE), assertions made using RWD. RWE may be clinical, related to medications, devices, or other interventions, and intended to guide future studies or to change practice.

Recognizing a need to reduce trial costs and augment traditional trials with real world learning, the United States Congress addressed RWE in the 21st Century Cures Act of 2016. Specifically, the Act required that the Food and Drug Administration create a pathway to allow RWE to support new drug indication and post-marketing surveillance starting in 2018.[8] In parallel to regulatory use of RWE, payers are increasingly demanding proof of real world effectiveness. Thus, insurance companies are increasingly demanding RWE to support reimbursement decisions.[9] Together, regulatory and reimbursement pathways are increasingly incorporating RWE and therefore the standard of care may soon be influenced by clinical assertions made using RWD.

The changing landscape of evidence from RCT to RWE has progressed rapidly. But, there are concerns over validity of clinical assertions as these data were not collected for research or regulatory purposes.[10,11] Concerns over EHR data accuracy have been highlighted in primary use applications.[12–14] Several companies have called for rigorous data assessment and setting of standards for use of RWE in regulatory settings.[15,16] Understanding the accuracy, quality, and availability of the underlying data and technology becomes critical to healthcare as they begin influence treatment decisions.

In this study, we aimed to 1) assess the occurrence of a predefined set of clinical concepts in the EHRs; 2) evaluate the accuracy of AI technologies when applied to clinical concepts in EHR-S and EHR-U; and 3) compare accuracies between traditional versus AI-based approaches using EHR structured and unstructured data. Our objective was to determine whether traditional real world evidence techniques are sufficiently accurate to support regulatory-grade EHR-based observational studies in cardiovascular medicine. We hypothesized that use of traditional query techniques (i.e. Standard Query Language (SQL)) on EHR structured data may be insufficient to support clinical assertions compared to more advanced approaches that leverage unstructured clinical text. Specifically, the use of problem, procedure, and other lists within the EHR matched by SQL query to a list of relevant codes may result in insufficient accuracy for some studies. This work can provide the preliminary evidence needed to set standards that will ensure regulatory-grade data quality and define best practices.

METHODS

In this observational, retrospective study, we assessed 10 840 clinical notes from a large academic medical center in the United States. This included a combination of outpatient and inpatient records drawn randomly from a multi-year experience. Criteria for record use included a problem list with at least one item and narrative text document length greater than a snippet of 1000 characters. The dataset included both EHR structured data (EHR-S) from problem list, medication list, and laboratory list, and EHR unstructured data (EHR-U) from clinical notes and other narrative text available in the EHR. The study was deemed exempt from the need for IRB approval.

Study population and cohorts

Cardiovascular medicine was chosen as a test case because it is an area of high cost, is the leading cause of death in the industrialized world, and was believed to represent a proxy for common medical care. A pre-defined feature list was selected based on common relevant conditions in cardiovascular medicine studies: Hyperlipidemia, hypercholesterolemia, coronary artery disease, diabetes mellitus, myocardial infarction, chronic kidney disease, stroke, dementia, cataract, coronary artery bypass graft, atorvastatin, pravastatin, rosuvastatin, simvastatin, LDL cholesterol, HDL cholesterol, and total cholesterol. These features represented a consensus list of important inclusion criteria, exclusion criteria, exposures, and outcomes based on a group of physicians. Of note, some cohorts may be used in different ways in different studies. For example, myocardial infarction may be used as inclusion criterion in one study, exclusion criterion in another study, and an outcome measure in a third study. Cohorts were selected as representative in cardiovascular medicine to provide a robust study set.

Study outcomes

The primary outcome was to assess whether the extracted data was "regulatory-grade". We define regulatory grade as "data sufficiently accurate to justify the clinical assertion." For this study, numeric thresholds were required and thus assumptions were made of how data skewness can lead to erroneous conclusions. A literature review of cardiovascular medicine found that assertions comparing study arms often seek a 10–20% benefit.[17,18] This suggests that a skew of 10% to 20% could result in inaccurate conclusions. Thus, the recall threshold was set at 85%. Precision is a different challenge, where inaccurate information is effectively added to the dataset. In this situation, there is less tolerance for error and a threshold of 90% was chosen. These criteria are intended as a first approximation of regulatory grade given that thresholds are highly dependent on study arm differences and potential skew in missing or erroneous data.

Precision was calculated as the proportion of patients correctly identified via the reference standard (see below) divided by the total number of patients identified in each cohort (true positive/(true positives + false positives). Recall was calculated as the proportion of patients correctly identified via the reference standard in each cohort (true positive/(true positives + false negatives). For example, if a patient is defined to have coronary artery disease in the gold standard, recall is determined based on whether or not the patient has coronary artery disease in their structured data and separately in their unstructured data. The F-score was also calculated as the weighted harmonic mean of the precision and recall. Optimal precision will lower recall, therefore the f-measure is used as a summary score across the two accuracy measures, precision and recall.

As additional measured variables, concept and patient occurrence were assessed. Concept occurrence is the sum of all occurrences of the concept, allowing for multiple occurrences per document. Patient occurrence is the number of patients that have at least one occurrence of the concept.

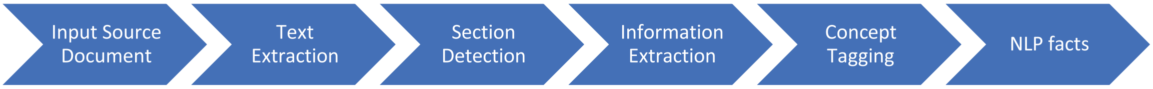

Figure 1. High-level NLP pipeline for clinical documentation.

Reference standard

Manual annotation was used to create a gold standard to validate automated assertion extraction from EHR-S and EHR-U. Two annotators manually labelled each concept and relevant meta-data for that concept. Each concept could include a single term (e.g. "hypertension") or a string of terms (e.g. "high blood pressure"). Meta-data included attributes that would change meaning for a clinical concept, such as negation and attribution to a subject other than the patient. Each annotator was a clinician, with a degree in medicine, nursing, or pharmacy and at least 5 years of clinical practice and at least one year of experience in clinical annotation.

For inter-operator reliability, an automated process was used to compare manual annotation between annotators on a daily basis early in the project and on a weekly basis later in the project. Any case where the two independent annotations did not agree was flagged as disagreement. In these cases, the disagreement was noted and the annotators brought together to discuss the cases and come to common agreement. For any cases where the two annotators disagree after discussion, a third annotator was engaged as a tie-breaker.

AI technologies

Artificial intelligence (AI) technologies were provided by Verantos, Inc. These included natural language processing (NLP) and machine learned inference.

The core of the AI is a deterministic NLP layer. This layer is built on top of the GATE NLP architecture.[19] The GATE architecture is used to construct a flexible pipeline for processing incoming text against English language syntactical rules augmented with a lexicon based on a clinical vocabulary. This pipeline is visually represented below and described in further detail below (Figure 1).

Text extraction

This stage of the NLP pipeline is responsible for extracting natural language text from various sources. For this study, text was extracted from fragments of narrative text in Clinical Document Architecture (CDA) XML documents which appear in HTML form. The open source Apache Tika library was deployed for this purpose.

Section detection

Attributing clinical text to the correct narrative section is important to add context in clinical concept interpretation. For example, a clinical concept appearing in a medical history section may indicate a past condition instead of an ongoing one. Section information is useful in disambiguation of abbreviations and acronyms. For example, the abbreviation *CP* in a past medical history section may favor cerebral palsy over chest pain depending on other features. Section detection was augmented using the techniques and vocabulary of SecTag.[20]

Information extraction and tagging

The steps in this stage are built using an infrastructure called ANNIE (A Nearly-New Information Extraction), which is part of the GATE NLP pipeline. Steps include: removal of special characters; tokenization; sentence splitter; POS tagger (tags tokens with part of speech tags such as adjectives, proper nouns, etc.); named entity recognition (matches tokens against an internal map of entities); and negation and subject tagging using the ConText library.[21]

Concept assignment

Phrases are matched against an internal database of clinical concepts to normalize identified information to known concepts. Mapping in this study included SNOMED-CT. RxNorm, and LOINC. Matching is done in multiple layers from deterministic for well-known text to probabilistic for inexact matches. Fuzzy matching, where required, is performed using approximate dictionary matching.

NLP output is fed into a machine learning system to identify clinically relevant patterns. A key component of the higher-level artificial intelligence is association rule mining. This is used for data-driven discovery of strong associations between clinical concepts. The noise filtering was achieved by measuring the support of concepts found in a clinical and measured against concept statistical attributes in the complete data set and filtering out or giving lower priority to concepts which were below a threshold. This results in a system that seeks patterns throughout the longitudinal patient record suggestive of a specific clinical concept, also known as clinical phenotyping.[22] For example, patterns such as chest pain, EKG changes, and troponin elevation may endorse a concept such as myocardial infarction when the content within the sentence boundary may simply state "evaluation for MI".

Statistical analyses

For the manual annotation, concordance between the annotators was assessed by the Kappa statistic. Kappa is a standard measure of concordance that controls for chance agreement. It pools all disagreements and may be affected by the frequency of cases. Kappas above .75 are considered excellent. Chi-square comparison of proportions between structured and unstructured data were used to identify significant differences between accuracy of EHR-S versus EHR-U datasets.

To assess a match between study arms and gold standard, each gold standard annotation was expanded to include the SNOMED-CT ontological "neighborhood" or module, which allowed the match to be closer to a clinical archetype instead of a specific SNOMED concept. For example, if the tested cohort was myocardial infarction and EHR-S contained ST elevation myocardial infarction (STEMI), EHR-S would be tested against the myocardial infarction SNOMED module. This would return a positive match since STEMI exists within the myocardial infarction SNOMED module. In this way, cohort matches were accurately identified despite discrepancies in granularity.

RESULTS

The records were drawn from an academic hospital. Patient demographics were representative of the US population. Table 1 shows results from the ten diseases of interest and the procedure of interest. Concept-level occurrence ranged from 194 for coronary artery

Table 1. Cohort identification of diseases and procedures stratified by EHR-S and EHR-U data[a]

Cohort	Occurrence		EHR-S			EHR-U		
	Concept	Patient	Recall (%)	Precision (%)	F1-score (%)	Recall (%)	Precision (%)	F1-score (%)
Hyperlipidemia	2471	837	65.2	99.3	78.7	98.2	99.4	98.8
Hypercholesterolemia	1899	478	55.1	98.0	70.5	90.4	98.8	94.4
Coronary artery disease	1427	465	67.5	99.4	80.4	94.6	96.2	95.4
Diabetes mellitus	4502	1377	80.6	97.9	88.4	97.0	92.6	94.8
Myocardial infarction	523	282	29.8	86.2	44.2	90.4	76.5	82.9
Chronic kidney disease	640	101	40.8	97.6	57.6	92.9	97.9	95.3
Stroke	693	307	36.5	97.2	53.0	95.7	79.6	87.0
Dementia	317	103	62.1	100.0	76.6	93.1	90.0	91.5
Cataract	240	85	28.6	100.0	44.4	96.1	94.9	95.5
CABG[b]	194	73	32.2	100.0	48.7	96.6	95.0	95.8

[a] All comparisons were significant at $P <. 0001$.

[b] Coronary artery bypass graft.

bypass graft to 4502 for diabetes mellitus. Patient-level occurrence for these concepts ranged from 73 for coronary artery bypass graft to 1377 for diabetes mellitus. In EHR-S, the minimum recall was 29.8% for myocardial infarction and the maximum was 80.6% for diabetes mellitus with respective F1-scores of 44.2 and 88.4. For EHR-U, the minimum recall was 90.4% for both myocardial infarction and hypercholesterolemia and the maximum was 98.2% for hyperlipidemia, F1-scores were 82.9, 94.4, and 98.8, respectively. All comparisons of proportions between EHR-S and EHR-U data were significant at $P < .0001$, with AI technologies applied to EHR-U outperforming traditional query techniques on EHR-S for each concept. We calculated the Cohen's kappa score between each pair of clinicians. The average of this score was 0.93.

Medications of interest also varied by data source (Table 2). Concept-level occurrence ranged from 586 for pravastatin to 2173 for rosuvastatin and patient-level occurrence ranged from 230 for pravastatin and 849 for rosuvastatin). In EHR-S, the minimum recall was 85.3% for both atorvastatin and simvastatin and the maximum was 94.1% for pravastatin with respective F1-scores of 92.0, 92.0, and 96.6. Similarly, for EHR-U the minimum recall was 97.9% for both atorvastatin and simvastatin and the maximum was 99.2% for rosuvastatin with respective F1-scores of 98.5, 98.5, and 99.3. All comparisons of proportions between EHR-S and EHR-U data were significant at $P < .0001$, again with AI technologies applied to EHR-U outperforming traditional query techniques on EHR-S for each concept.

Laboratory studies were not available for EHR-S. However, information regarding concept and patient occurrence as well as performance metrics for EHR-U are available in Table 3. The accuracy results are similar to disease and procedure concepts for EHR-U.

DISCUSSION

The goal of this study was to perform a rigorous quality assessment of RWD to understand the potential and limitations of RWE in regulatory decision-making. Using cardiovascular medicine as a test case, cohort identification in EHR-S data using traditional query techniques did not meet our definition of regulatory-grade. Recall, the ability to accurately identify all true cases, consistently fell below our set standard of 85% for disease and procedure identification. However, by applying AI technologies to EHR-U, cohort identification exceeded set standards for disease, procedure, medication, and laboratory studies. To our knowledge, this work provides the first evidence related to data standards and quality of RWE that is necessary to achieve regulatory-grade studies.

Defining regulatory-grade

In acknowledgement of increasing use of RWE to influence the standard of care, we define regulatory grade as "data sufficiently accurate to justify the clinical assertion." To support objective measurement, we proposed objective measurable criteria, specifically recall > 85% and precision > 90%. This definition is not intended to be a set standard for all types of study questions, but rather a starting benchmark to initiate discussion. A key point is that both missingness reflected by recall and errors reflected by precision are important when considering accuracy.

Precision versus recall

Measuring accuracy is labor intensive, but provides a true assessment of cohort precision and recall. In general, when accuracy is measured in studies today, precision is emphasized rather than recall. This is not because precision is more statistically relevant than recall, but rather because precision is easier to assess.[23] For example, in a study evaluating patients taking a cholesterol lowering drug after heart attack, a data scientist may pull all patients that meet these criteria in the structured data set (problem and medications lists). This may result in the identification of 300 patients from a cohort of one million patients. The data scientist next reviews the charts for the 300 patients and confirms that the correct conditions are referenced in the clinical narrative. This scenario will assess precision and ignore recall. However, recall is where the inaccuracy typically lies.

Although recall is frequently used in academic publications, it is rarely implemented in RWE studies because it is resource intensive. Continuing with the myocardial infarction example, it would be far more difficult to sample a portion of a million records to calculate false negatives for "myocardial infarction" than to confirm that out of 300 records the occurrence of "myocardial infarction" correlated with the patient having had a heart attack. For this reason, pharmaceutical companies and contract research organizations performing RWE studies often report precision and rarely assess recall. But, important clinical bias exists in the missed cases. Specifically, a patient with a mild heart attack may only have one clinical encounter with a physician. This may or may not result in "myocardial infarction" being added to the problem list. However, a patient with a severe heart attack will have multiple physicians and encounters during an

Table 2. Cohort identification of medications stratified by EHR-S and EHR-U data[a]

Cohort	Occurrence		EHR-S			EHR-U		
	Concept	Patient	Recall (%)	Precision (%)	F1-score (%)	Recall (%)	Precision (%)	F1-score (%)
Atorvastatin	1439	449	85.3	100.0	92.0	97.9	99.1	98.5
Pravastatin	586	230	94.1	99.1	96.6	99.2	98.3	98.8
Rosuvastatin	2173	849	91.4	99.5	95.3	99.2	99.4	99.3
Simvastatin	1439	449	85.3	100.0	92.0	97.9	99.1	98.5

[a]All comparisons were significant at $P < .0001$.

Table 3. Cohort identification of laboratory studies stratified by EHR-S and EHR-U data

Cohort	Occurrence		EHR-S			EHR-U		
	Concept	Patient	Recall (%)	Precision (%)	F1-score (%)	Recall (%)	Precision (%)	F1-score (%)
LDL cholesterol	475	243	NA	NA	NA	94.7%	100.0%	97.3%
HDL cholesterol	278	139	NA	NA	NA	95.7%	100.0%	97.8%
Total cholesterol	227	165	NA	NA	NA	94.0%	100.0%	96.9%

extended inpatient stay. The chance that "myocardial infarction" is added to the problem list is far higher for this patient than the patient with a mild attack and brief episode of care. Thus, if only precision is measured while recall is ignored, accuracy has not truly been tested and bias is likely. To assure that accuracy is properly measured, clinical research often focuses on the F1-score to harmonize recall and precision. The right balance depends on the clinical question being asked, i.e. more focused (higher precision) or broader (higher recall) searches.

EHR structured data vs EHR unstructured data

Similar to previous work, this study demonstrated that cohort identification from EHR structured data using standard query technologies may be insufficient for regulatory use. In addition, our study quantified the differences in accuracy between standard query technologies and more advanced methodologies. We found that the identification of concepts used for inclusion criteria, exclusion criteria, and outcomes differed significantly between EHR structured and unstructured data, with the exception of medication data which had comparable cohort identification accuracy between EHR-S and EHR-U data. Similarly, other studies have shown that integrating EHR structured and unstructured data improves clinical phenotyping, suggesting that advanced methodologies using unstructured data are necessary to improve performance impact. This limitation of EHR-S is concerning since data accuracy is rarely measured. In common datasets available today, including claims data and EHR structured data, the narrative text which allows accuracy assessment or data augmentation is often missing. Thus, there is no measure of cohort accuracy nor is there any way to independently assess accuracy as the underlying data are missing.

Our results suggest that EHR-U data analyzed with advanced technologies are needed to achieve regulatory-grade when using RWE. However, if only structured data are available or feasible to use, it is important to note the percent error in defining patient cohorts and understand the types of questions that can be feasibly answered with these data. Similarly, statistical models and predictive analytics should account for this classification error when determining confidence intervals and standard errors for predictive models. Therefore, both technology and expertise are critical in achieving high accuracy cohort extraction for regulatory decision making.

Review of specific concepts

Some results were surprising, even recognizing known inaccuracy in clinical documentation. Myocardial infarction notably had a recall of 29.8%. This was found to be due to a low rate of physicians placing heart attack on the problem list if the patient had experienced it in the past. A typical record would include "h/o MI" in the narrative text, but no suggestion of prior heart attack in the problem list. This highlights the discrepancy between clinical use of the problem list, where primarily new issues are highlighted, versus study use of the problem list, where any past heart attack may be a reasonable exclusion criterion. This discrepancy is extremely important when identifying patients with specific inclusion and exclusion criteria for secondary analyses, such as pragmatic clinical trials. Another surprise was medication recall below 100% for structured data. When evaluated after study completion, this was believed to be due to physicians treating patients in outpatient settings where the medication may have been prescribed by another physician who was not tracked within the same EHR. This is a common issue stemming from a lack of interoperability between healthcare settings and likely more common for tertiary care centers, where a patient may not receive their primary care and hence common medications for chronic diseases.

Limitations

This study has several limitations. The results are from a single healthcare system, which may not be generalizable to other settings. However, our cohort identification results are consistent with previous literature in other clinical domains.[12,13] The approach, including manual gold standard and definition of both precision and recall, was labor intensive and may not be repeatable across RWE needs. This study required expertise across multiple fields, drawn from academia, pharma, and technology. Expertise may become a gating factor in regulatory grade studies. This study assessed cohorts relevant to cardiovascular medicine and may not be generalizable to other clinical domains.

CONCLUSIONS

In summary, we document differences in obtained accuracy between EHR structured and unstructured data for clinical phenotyping in cardiovascular medicine. The clear learning from this study is that accuracy is heavily influenced by data and technology choices. These authors recommend that all real world evidence studies that may influence the standard of care, e.g. regulatory and reimbursement submissions, should include a data accuracy assessment of all key cohorts, including inclusion criteria, exclusion criteria, exposures, and outcomes. Expectations for both precision and recall for these cohorts should be defined within the study protocol in advance of collecting and evaluating data and these expectations for data accuracy should be consistent with anticipated effect size. In order to maintain credibility and advance science, pharma, academia, and vendors must not shy away from the hard work required to ensure data accuracy. As payers and regulatory agencies move forward with real world evidence to overcome cost and generalizability issues, understanding the benefits and limitations of different data and technologies is essential.

FUNDING

This work was supported by a research grant from Amgen, Inc. and the National Center For Advancing Translational Sciences of the National Institutes of Health under Award Number R44TR002437. Dr Hernandez-Boussard was supported in part by the National Cancer Institute of the National Institutes of Health under Award Number R01CA183962. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Amgen or the National Institutes of Health.

AUTHOR CONTRIBUTIONS

Dr Riskin had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. He attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. Dr Riskin affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned have been explained.

- Study concept and design: DR, THB.
- Acquisition of data: DR.
- Analysis and interpretation of data: DR, THB, KM
- Drafting of the manuscript: THB, DR
- Critical revision of the manuscript for important intellectual content: THB, KM, BC, DR
- Final Approval of the version to be published: DR, KM, BC, THB
- Statistical analysis: BC, THB.
- Administrative, technical or material support: DR.
- Study supervision: DR.

CONFLICT OF INTEREST STATEMENT

Dr Monda is an employee and stockholder of Amgen. Dr Blai is an employee and stockholder of Amgen. Dr Riskin is an employee and stockholder of Verantos.

REFERENCES

1. Warraich HJ, Hernandez AF, Allen LA. How medicine has changed the end of life for patients with cardiovascular disease. *J Am Coll Cardiol* 2017; 70 (10): 1276–89.
2. Jones WS, Roe MT, Antman EM, *et al*. The changing landscape of randomized clinical trials in cardiovascular disease. *J Am Coll Cardiol* 2016; 68 (17): 1898–907.
3. Ioannidis JP. Why most clinical research is not useful. *PLoS Med* 2016; 13 (6): e1002049.
4. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 2015; 16: 495.
5. Putting gender on the agenda. *Nature* 2010; 465 (7299): 665.
6. Sherman RE, Anderson SA, Dal Pan GJ, *et al*. Real-world evidence: what is it and what can it tell us? *N Engl J Med* 2016; 375 (23): 2293–7.
7. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. *JAMA* 2018; 320 (9): 867–8.
8. Congress. *21st Century Cures Act*. Washington, DC: Congress; 2016.
9. Willke RJ. Translating comparative effectiveness research evidence to real-world decision making: some practical considerations. *Decision Making in a World of Comparative Effectiveness Research*. Singapore: Adis; 2017: 105–116.
10. Hersh WR, Weiner MG, Embi PJ, *et al*. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care* 2013; 51 (8 Suppl 3): S30–7.
11. Kaplan RM, Chambers DA, Glasgow RE. Big data and large sample size: a cautionary note on the potential for bias. *Clin Transl Sci* 2014; 7 (4): 342–6.
12. Singer A, Yakubovich S, Kroeker AL, Dufault B, Duarte R, Katz A. Data quality of electronic medical records in Manitoba: do problem lists accurately reflect chronic disease billing diagnoses? *J Am Med Inform Assoc* 2016; 23 (6): 1107–12.
13. Luna D, Franco M, Plaza C, *et al*. Accuracy of an electronic problem list from primary care providers and specialists. *Stud Health Technol Inform* 2013; 192: 417–21.
14. Wright A, McCoy AB, Hickman TT, *et al*. Problem list completeness in electronic health records: a multi-site study and assessment of success factors. *Int J Med Inform* 2015; 84 (10): 784–90.
15. Miksad RA, Abernethy AP. Harnessing the power of real-world evidence (RWE): a checklist to ensure regulatory-grade data quality. *Clin Pharmacol Ther* 2018; 103 (2): 202–5.
16. Katkade VB, Sanders KN, Zou KH. Real world data: an opportunity to supplement existing evidence for the use of long-established medicines in health care decision making. *JMDH* 2018; 11: 295–304.
17. Schwartz GG, Steg PG, Szarek M, *et al*. Alirocumab and cardiovascular outcomes after acute coronary syndrome. *N Engl J Med* 2018; 379 (22): 2097–107.
18. McNeil JJ, Wolfe R, Woods RL, *et al*. Effect of aspirin on cardiovascular events and bleeding in the healthy elderly. *N Engl J Med* 2018; 379 (16): 1509–18.
19. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. *PLoS Comput Biol* 2013; 9 (2): e1002854.
20. Denny JC, Spickard A, Johnson KB, Peterson NB, Peterson JF, Miller RA. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc* 2009; 16 (6): 806–15.
21. Harkema H, Dowling JN, Thornblade T, Chapman WW. ConText: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *J Biomed Inform* 2009; 42 (5): 839–51.
22. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annu Rev Biomed Data Sci* 2018; 1 (1): 53–68.
23. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* 2015; 10 (3): e0118432.