TRUST: Clinical Text Retrieval and Use towards Scientific Rigor and Transparent Process

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Sunyang Fu

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Advisor: Hongfang Liu, PhD, Co-advisor: Yuk Sham, PhD

December, 2021

Acknowledgements

This dissertation was made possible by the contributions and support of many people. First and foremost, I would like to thank my advisor, Dr. Hongfang Liu. I am truly grateful, honored, and lucky to have the opportunity to work with her. Under Dr. Liu's influence, I started my research topic on real-world pragmatics challenges related to clinical research informatics and natural language processing. I truly learned a lot from her unique perspective and vision. Dr. Liu is very patient in terms of the development of my skill sets. I enjoyed every piece of opportunities to working with her, which include interesting discussions of informatics-related challenges, detailed guidance on scientific writing, and wholeheartedly supported my exploration of various research topics. From Dr. Liu, I learned that there is no shortcut for research, and being a successful researcher need to have a persistent mindset.

I want to express my sincere gratitude to my thesis committee members: Dr. Yuk Sham, Dr. Chih-Lin Chi, Dr. Rui Zhang, and Dr. Hongfang Liu. My committee members represent strong and diverse scientific background and informatics expertise. Their mentorship and scientific feedback not only guides me in the right direction but also ensure the thesis topics were comprehensively covered. I am very appreciative of the guidance and advice.

I want to thank Dr. Yuk Sham and Dr. Chad Myers's leadership and contribution to the BICB program. I feel very honored to be part of the BICB family.

I am grateful to all my current and former Mayo Clinic colleagues and collaborators. First, I want to thank my scientific mentors and advisors Dr. Sunghwan Sohn, Dr. Jungwei Fan, Dr. Jenny St Sauver, and Mr. Michael Lin. I am appreciative for their guidance and expertise in the areas of clinical NLP, biomedical terminology, and clinical and translational science. I want to thank my awesome colleagues and mentors (current

and former) in the ADVANCE lab: Dr. Yanshan Wang, Dr. Feichen Shen, Dr. Liwei Wang, Dr. Sijia Liu, Dr. Sungrim Moon, Dr. Majid Rastegar-Mojarad, Dr. Nansu Zong, Dr. Huan He, Dr, Jun Jiang, Dr. Yiqing Zhao, Luke Carlson, and Donna M Ihrke. I've learned a lot about how to be a scientist and informatician from working with all of you. I also want to express my genuine gratitude to my clinical collaborators and mentors: Dr. Patrick Luetmer, Dr. Sandeep Pagali, Dr. Hilal Maradit Kremers, Dr. Lester Leung, and Dr. David Kent. This work cannot be successful without your domain knowledge and expertise. Last but not least, I want to thank my peer mentors Dr. Kevin Peterson and Andrew Wen. Kevin has been providing an enormous amount of supports during my Ph.D. journey. From coursework study to exam preparation, the advice and guidance was an invaluable asset for my Ph.D. journey. I also want to thank Andrew Wen, who has been a tremendous helpful to my research, technical skills, and scientific writing.

I am incredibly thankful for all the support from my family and friends. I want to thank my parents for emphasizing education, exploration, and continuous improvement. Their support was instrumental to the completion of my education. I want to thank my grandpa, my inspirational role model, for the earnest edification. I want to thank my friend Yuxuan Zhang, who has had a strong influence on my early education and has been providing continuous inspiration to my curiosity and innovation. To my wife Jiawei, you are ineffable to the completion of this work.

I want to thank numerous funding agencies, including the National Institutes of Health, the National Institute of Aging, the National Center for Advancing Translational Sciences, and the Mayo Clinic Foundation for Medical Education and Research.

Dedication

To Jiawei and Kabi.

Abstract

Rapid proliferation and adoption of the electronic health record (EHR) has led to seamless integration of clinical research into practice, and has facilitated healthcare decision-making through enabling accurate and timely supply of health information. Leveraging this supply of information, the Institute of Medicine envisioned the concept of continuously Learning Health Systems (LHS) in 2007, with the aim of first deriving knowledge from routine care data and then translating such knowledge into evidence-based clinical practice. To achieve such a vision, it is critical to have a robust data and informatics infrastructure with the following properties: 1) high-throughput and real-time methods for data retrieval, extraction and analysis, 2) transparent and reproducible processes to ensure scientific rigor in clinical research, and 3) implementable and generalizable scientific findings.

There are many approaches to the derivation of knowledge from care data, one of which is through the use of chart review: a common, albeit manual, approach to practice-based knowledge discovery. Traditionally, chart review is performed by manually reviewing patient medical records. As a significant portion of clinical information is represented in textual format, this manual approach can be time-consuming and costly. With the implementation of EHRs, chart review can be automated by extracting data from structured fields systematically and leveraging natural language processing (NLP) techniques to extract information from text. Rigorous development and evaluation of NLP algorithms for a specific chart review task requires, however, data abstraction and annotation (i.e., the manual creation of a gold standard clinical corpus to evaluate the developed NLP algorithm). In EHR-based settings, there is, however, a lack of standard processes or best practices for creating such a corpus due to the heterogeneity of institutional EHR systems and process variation between single and multi-site research settings.

Recent advancement in healthcare AI identifies the need for detailed data provenance for data used in the training and validation of AI models. Secondary use of EHR for clinical research leveraging AI technologies such as NLP therefore requires the documentation of the provenance information relating to the process used for the retrieval and

organization of the raw data used as well as the extraction and annotation of training data. We thus define this process as clinical **T**ext **R**etrieval and **U**se towards **S**cientific rigor and **T**ransparent (TRUST) process. As EHR-based research becomes increasingly integrated into clinical care, it is important to have a systematic understanding of the TRUST process, its corresponding utilization when developing informatics tools and methods, as well as its overall impact on research reproducibility.

In this work, we propose a multi-phase method to develop informatics frameworks and best practices to ensure reproducible TRUST processes for single and multi-site studies. In the following chapters, we propose: 1) a definition of reproducibility in the context of the secondary use of EHRs, 2) methods to assess various levels of data heterogeneity caused by differing EHR systems and inter-institutional variations, 3) approaches to examine the implication of data heterogeneity to reproducibility, 4) steps to develop frameworks, best practices, and reporting standards conforming to the TRUST process, and 5) an application of the TRUST process in a real-world case study.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009, which provides incentives to institutions demonstrating aggressive applications and "meaningful use" of EHR systems, has heavily incentivized recent development in health information technology (HIT) [1, 2]. These efforts enable the conceptual strategy of continuously Learning Health Systems (LHS) [3], which focuses on routinely capturing clinical data, converting data into new knowledge, and applying evidence-based knowledge to improve the quality of care on the basis of rigorous research [3, 4, 5]. To achieve such a vision, it is critical to have a robust data and informatics infrastructure with the following properties: 1) high-throughput and real-time methods for data extraction and analysis [6], 2) transparent and reproducible processes to ensure scientific rigor in clinical research [7], and 3) implementable and generalizable scientific findings [8, 9].

There are many approaches to the derivation of knowledge from clinical data, one of which being the chart review, a common, albeit manual, approach to practice-based knowledge discovery [10, 11]. As a significant portion of clinical information is represented in textual format, execution of such a manual approach can be both time-consuming and costly [12, 13, 14]. EHR implementation can enable automation of chart review by extracting data from structured fields systematically and leveraging artificial intelligence (AI) technologies such as natural language processing (NLP) to extract information from text. The field of research related to this topic, commonly referred to as information extraction (IE), itself a sub-domain of NLP, seeks to address the challenges of computationally extracting information from free-text narratives [15, 16].

The validity and portability of NLP models are however dependent on the data from which they are derived. Consequently, models cannot be trusted without a good understanding of the data with which they are fed. The initial step of developing NLP-based IE models requires an annotated clinical corpus, production of which is traditionally referred to as corpus annotation: a task of covering any descriptive or analytic notations applied to raw language data [17]. In EHR-based research, this corpus annotation process is embedded within an existing research workflow, which typically includes, but is not limited to feasibility assessment, research protocol development, cohort screening, data retrieval, curation, organization and use [7, 18]. EHR data is known to suffer from data quality issues [19, 20]. Unlike data being prospectively collected in a controlled environment such as clinical trials, EHR systems are primarily designed for patient care and data documentation patterns can therefore be easily affected by numerous contextual factors. Furthermore, the data generation and curation process can be long-lasting, iterative, and complex, particularly due to the heterogeneity in EHR systems implementations at an institutional level, as well as process variation between single and multi-site research settings [21, 22, 23]. These issues can cause incomplete medical records, documentation errors, and misinterpretations. As a result, systematic bias, measurement error, and misclassification can occur, which will ultimately impact the overall reproducibility of any produced research [19, 20].

Recent advancement in healthcare AI identifies the need for detailed data provenance for data used to train and validate AI models. Secondary use of EHR for clinical research leveraging AI technologies such as NLP therefore requires the documentation of the provenance information that captures the retrieval process and organization of any raw data used as well as the extraction and annotation process for the derived training data. We thus define this process as clinical **T**ext **R**etrieval and **U**se towards **S**cientific rigor and **T**ransparent (TRUST) process. As EHR-based research becomes increasingly integrated into clinical care, it is important to have a systematic understanding of the TRUST process, its corresponding utilization when developing informatics tools and methods, as well as its overall impact on overall research reproducibility. In this work, we propose a multi-phase method to develop informatics frameworks and best practices to ensure reproducible TRUST processes for single and multi-site studies. In the following chapters, we propose: 1) a definition of reproducibility in the context of the

secondary use of EHRs, 2) methods to assess various data heterogeneity caused by differing EHR systems and institutional variations, 3) approaches to examine the implication of data heterogeneity to reproducibility, 4) steps to develop frameworks, best practices, and reporting standards conforming to the TRUST process, and 5) applications of the TRUST process to a real-world case study.

1.1 Motivation

Clinical research reproducibility is crucial in the field of clinical and translational science as the findings of a single limited study must first successfully be generalized for broader patient care before it becomes useful in clinical practice [8]. Many barriers to reproducibility however exist, including 1) the complex processing of collecting, transforming, extracting, and organizing of EHR data and 2) the imperfect interactions and pragmatics between the human and non-human actors (e.g., information, system, and stakeholder) within the heterogeneous EHR environment [24].

Heterogeneity of EHR System and Institutional Variation Many of the challenges faced by the secondary use of EHR data originate from the voluminous, complex, and dynamic nature of the data being documented, transformed, and represented within heterogeneous and institution-specific EHR systems [23, 21]. EHR systems are a vital part of health IT infrastructure that 1) contain a patient's medical record (e.g., diagnosis, history, medications, treatment plans, etc.), 2) provides access to clinical tools for decision making, 3) streamlines provider workflow and communication, and 4) provides digital management of health information [25]. EHRs data contain digitized patient medical records represented through a variety of different formats such as 1) structured (e.g. billing, medication), 2) semi-structured (e.g. patient provided information), 3) unstructured (e.g. clinical notes, radiology reports) and d) binary files (e.g. medical imaging files). Depending on the institutional data infrastructure, different data modalities can be generated, managed, and stored as part of differing Health IT systems or clinical data warehouses. A significant amount of effort and steps are required in order to locate, retrieve and pre-process EHR data into a specific format or representation [26]. Information loss and degradation can occur during this process when mapping EHR data to a specific study or data model [27]. On one hand, the EHR system itself has a

significant impact on the form and format of clinical data. In the context of secondary use, data quality issues may have already occurred when the data is generated. As the primary functions of EHR systems are centered around patient-oriented care management and billing rather than for research purposes, the objective and priority for documentation and reporting of clinical information may vary significantly by providers and care settings [28, 27]. This lack of consistency in documentation practices can result in multiple sets of patient data sharing differing definitions that reside in disconnected "silos". This inherent latent documentation bias and information variability in the EHR can ultimately result in information misinterpretation, ambiguous data definitions, incomplete medical records, recording errors (e.g., missing or redundant documentation), and non-standard data representations [29, 26, 19]. This issue can be further exacerbated in multisite studies, where variations in EHR system implementations, ETL processes, care practice, and patient populations, and values and definitions of EHR data across multiple institutions can suffer from representational information quality issues. Madigan et al. systematically assessed the variability of 10 different clinical databases. The study discovered that 40% of the results from observational database studies vary significantly [23]. To address the issue of data heterogeneity caused by EHR systems and institutional variations, we develop and implement informatics methods to assess data quality variations caused by different EHR systems and institutional variations.

Unstructured Text A well-known challenge in EHR-based studies is that much of the detailed patient information is embedded within clinical narratives [30, 16], with the usage of NLP being a proposed solution. In that vein, NLP-assisted chart review automatically extracts information from unstructured text in the EHR. Researchers have used NLP systems to identify clinical syndromes and common biomedical concepts from clinical notes [31], radiology reports [32], operative reports [33], pathology reports [34], and microbiology reports [35]. Different clinical NLP systems have been developed at different institutions, including MedTagger [36], MedLEE [37], cTAKES [38], MetaMap [39], KnowledgeMap [40] and HiTEX [41]. Text data is however known for suffering from various quality issues. For example, the documentation functionality such as templates, copy and paste, auto-documentation, and transcription built in to EHR systems can affect the EHR-specific syntactic and semantic definition for any text

data contained therein. As a result, NLP systems developed from heterogeneous clinical notes may suffer from issues of portability, i.e., a measurement of the feasibility and implementability of deploying a system to a different institution [42]. For example, studies have reported various performance degradations when the original NLP system is deployed to a different institution or for different studies [43, 42]. Currently, existing methods of addressing EHR-related data quality issues were primary concerned with structured data [44, 45, 28, 46], as opposed to unstructured. Our work aims to fill this gap by primarily focusing on building pipelines and best practices to help establish data provenance practices and reporting guidelines for clinical text.

Human Factor and Cognitive Bias The process of documenting, collecting, and using EHR data can be task-specific, iterative, and complex, often involving a multitude of stakeholders with diverse backgrounds in terms of training and domain expertise [47, 48]. During the secondary use of EHRs, stakeholders regularly make choices that directly affect the definition of the study cohort, data, disease measurement, and outcome. Variations in the decision-making practice can affect the certainty about the value of different options [49, 50, 51]. On one hand, the validity and reliability of research decisions are dependent on the rigor and explicitness of consensus activities and team-science discussion, as suggested by studies finding that the reliability of clinical choices [52], SNOMED-CT coding [53], and interpretation of radiology images and reports [54] can be imperfect. On the other hand, the traceability of EHR data is dependent on the presence, utilization, and quality of abstraction protocol. Morris et al found a substantial variation in clinicians' compliance, interpretation, and execution with respect to a clinical protocol [52]. A review conducted by Gilbert et al. on three emergency medicine journals discovered that only 11% (95% CI, 7% to 15%) of chart review studies reported the use of standardized abstraction forms [12]. This finding suggests that the current state of EHR-based research lacks standardized methods and best practices to help 1) improve people's decision-making and interpretation, 2) enforce explicit discussion of ambiguous definitions, and 3) incentivize detailed protocols specifying the common conventions and standards. Our work aims to address this gap by proposing methods to establish best practices, guidelines, and process frameworks based on real-world case studies and literature.

1.2 Research Summary

The methods of developing the TRUST process can be summarized into four phases: 1) conceptualizing the definition of reproducibility in EHR-based settings, 2) real world pragmatics assessment of the heterogeneity of institutional and EHR variance, 3) formulation of an informatics process with best practices and reporting standards, and 4) implementation and evaluation of a real-world NLP-assisted chart review study. Our approach leveraged multiple real-world investigations (single and multi-site) and large-scale literature reviews. Here, we briefly present the main research objectives, corresponding experiments, and findings below.

1.2.1 Latent Data Quality Impacted by EHR System Heterogeneity

- **Hypothesis**: The heterogeneous EHR systems can affect information quality and variability.
  - **Experiment**: We retrospectively assessed and compared the variability of data produced from two different EHR systems (Epic and GE Centricity). We considered three dimensions of data quality in the context of EHR-based AI modeling in three translational stages: model development (data completeness), model deployment (data variability), and model implementation (data timeliness). The case study was conducted based on the detection and prediction of post-surgical complications.
  - **Results**: We discovered a consistent level of data completeness across two EHR systems with an exception for the lab data. On the other hand, there was a high syntactic variation suggested by the corpus statistics. There was also a moderate difference in the semantic type and frequency of document sections. Textual similarity revealed a consistent pattern for roughly half of the concept-section pairs. High language variation was found for the sections of Secondary Diagnosis (Abscess) and Past Medical/Surgical History (Anemia). The data timeliness of clinical notes documentation for Epic was improved when compared with Centricity.

1.2.2 Latent Data Quality Impacted by Institutional and Process Variation

- **Hypothesis**: The variability of institutional data abstraction and annotation process can affect information quality and variability.
  - **Experiment**: Data abstraction and annotation was conducted at Mayo Clinic and Tufts Medical Center. The assessment of EHR heterogeneity was conducted using screening ratio, inter-rater reliability measurement, corpus statistics, clinical document similarity, and prevalence ratio.
  - **Results**: We discovered a significant variation in the patient populations, neuroimaging reporting, EHR systems, and abstraction processes across the two sites. There was also a variation regarding neuroimaging reporting where TMC was lengthy, standardized, and descriptive while Mayo's reports were short and definitive with more textual variations. Furthermore, differences in the EHR system, technology infrastructure, and data collection process were identified.

1.2.3 The Implication of Latent Information Quality to Reproducibility

- **Hypothesis**: Information quality dimensions have implications for reproducibility.
  - **Experiment**: To investigate heterogeneity involved in the process for the secondary use of EHRs and its implications for reproducibility, we formulated three components: 1) a conceptual process of information collection, extraction, organization, and representation, 2) information quality metrics for quantifying process feasibility and clinical outcome variability, and 3) a downstream implication through a case simulation.
  - **Results**: Based on the IQ assessment and case study implementation, we discovered various barriers to reproducibility, such as inconsistent information documentation patterns across settings, information loss during ETL processes, and variable levels of information resource accessibility. In addition, the accessibility IQ has direct implication to clinical outcome variability.

1.2.4 Methodological Standard and Best Practices of Reproducibility

- **Question**: Can we formulate an informatics framework to help mitigate issues of data quality and improve process transparency and standardization?
  - **Experiment**: A multi-phase methods approach was used to determine the components needed for the framework. The process includes the following steps: 1) literature review, 2) standards adoption, 3) prototyping, 4) expert evaluation, and 6) finalization. The development of the proposed framework was designed after a review of the existing guidelines and best practices, including Corpus Annotation Schemes; Fundamentals of clinical trials; and Research data management.
  - **Results**: The framework summarizes the linear process of extracting or reviewing information from EHRs and assembling a data set for various research needs. The processes consider important action items (data quality and EHR-related heterogeneity assessment methods) and documentation checklists to identify, evaluate and mitigate variations across settings and assist data provenance.

1.2.5 Research Standard and Metadata of Reproducibility

- **Hypothesis**: Evidence of reproducible methods and best practices can be systematically extracted from literature.
  - **Experiment**: We developed and evaluated natural language processing algorithms to extract the reporting patterns and data abstraction methodologies from EHR-based clinical research. Post-extracted findings were analyzed using logistic regression to test the incremental significance of the use of methodology throughout the years. In addition, authorship and affiliation information for each publication was retrieved from PubMed API and qualitatively analyzed. The goal of this analysis was to understand whether there was a variation regarding the reporting patterns given the first author's training background.
  - **Results**: Our investigation discovered an upward trend of reporting research

methodologies, good practices, and the utilization of informatics-related tools and methods for EHR-based clinical research. Despite these findings, the methodologic standards were still consistently under-reported. We also discovered high variation regarding clinical research reporting.

1.2.6 Applications to Real-World EHR-based Study

- **Hypothesis**: The adoption of the TRUST process can enhance process transparency and research reproducibility.
  - **Experiment**: The proposed TRUST process was implemented for a real-world EHR-based clinical research focusing on developing NLP algorithms for the ascertainment of delirium status. The post-implementation was evaluated through the RITE (Reproducible, Implementable, Transparent, Explainable) criteria.
  - **Results**: Guided by the TRUST process, we adopted the standardized evidence-based framework CAM to develop and evaluate NLP algorithms to identify the occurrence of delirium from EHRs. The evaluation demonstrated high performance in identifying patients with delirium using clinical notes in an expeditious and cost-effective manner. The case implementation emphasizes the importance of ensuring both 'process correctness' and 'result correctness'.

1.3 Research Contributions

Our work proposed a TRUST (clinical **T**ext **R**etrieval and **U**se towards **S**cientific rigor and **T**ransparent) process that facilitates the assessment of the EHR-related latent heterogeneity and documentation of the provenance information that captures the process of the retrieval and organization of raw data as well as the extraction and annotation of training data. The proposed framework is motivated by our desire to address real-world data management and reproducibility-related challenges faced by researchers and strengthen clinical research informatics processes. As the majority of previous reproducibility research has been focusing on the stage of dissemination, we tackle research reproducibility from the perspective of the research life cycle as a whole, which allows

us to address a different type of issues related to reproducibility (e.g., implementation quality, people-driven interpretation, semantic interoperability, reporting standard, etc).

To better understand and mitigate study outcome variability caused by various heterogeneous factors, we proposed and applied a standardized set of informatics (qualitative and quantitative) methods to examine the heterogeneity of EHR systems, institutions, people, and processes. Besides outcome variability, another important criterion of reproducibility is implementation feasibility. To address the current issue of lack of standard processes and best practices for conducting EHR-based studies, we followed a multi-phase process to produce and validate internal and external standards and evidence for developing best practices. We conducted large-scale literature reviews and several real-world case studies based on multiple institutions. The proposed best practices cover an end-to-end TRUST process from the perspective of implementation quality, people-driven decision interpretation and making, documentation (data catalog and provenance), and project management.

This work involves several multi-institutional efforts to provide pragmatic evidence to develop reproducibility standards and best practices. The ESPRESSO (Effectiveness of Stroke Prevention in Silent Stroke) is a multi-site EHR-based research project with the goal to identify individuals with silent brain infarctions (SBI) at Tufts Medical Center (TMC) and Mayo Clinic. This allows us to have real-world pragmatic evidence for framework development and early end-user engagement for usability assessment. Our work is also part of the open-source project under the Open Health Natural Language Processing (OHNLP) consortium (http://www.ohnlp.org), which creates an interoperable, scalable and usable NLP ecosystem.

1.4 Outline

The outline for the following chapters is presented below:

- Chapter 2 introduces the background and related work of EHR-based reproducibility research
- Chapter 3 conducts an informatics assessment on EHR-related information quality and variability from the perspective of EHR system variation.

- Chapter 4 conducts an informatics assessment on EHR-related information quality and variability from the perspective of institution variation.
- Chapter 5 conducts the evaluation of post framework implementation to examine the relationship between latent information quality and reproducibility.
- Chapter 6 illustrates the development process of an informatics framework and best practices aiming to address process reproducibility.
- Chapter 7 illustrates the development process of reporting standard and research metadata of reproducibility.
- Chapter 8 summarizes the implementation of the TRUST process in a real-world case study of clinical NLP.

With permission from the publishers, the following chapters included previously published materials. Sunyang Fu's contribution to these chapters includes study design and conceptualization, data collection, data analysis, methodology development, software and algorithm development, initial draft writing, revising, and editing.

1. **Fu S**, Wen A, Schaeferle GM, Wilson PM, Demuth G, Liu S, Ruan X, Liu S, Storlie C, Liu H. Assessment of Data Quality Variability across Two EHR Systems through a Case Study of Post-Surgical Complications. InAMIA Annual Symposium Proceedings 2022. American Medical Informatics Association.
2. **Fu S**, Leung LY, Raulli AO, Kallmes DF, Kinsman KA, Nelson KB, Clark MS, Luetmer PH, Kingsbury PR, Kent DM, Liu H. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. BMC medical informatics and decision making. 2020 Dec;20(1):1-2.
3. **Fu S**, Leung LY, Wang Y, Raulli AO, Kallmes DF, Kinsman KA, Nelson KB, Clark MS, Luetmer PH, Kingsbury PR, Kent DM. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. JMIR medical informatics. 2019;7(2):e12109.
4. **Fu S**, Wen A, Pagali S, Zong N, Sohn S, Fan J, Liu H. The Implication of Latent Information Quality to the Reproducibility of Secondary Use of Electronic Health Records. InMedInfo 2022.

5. **Fu S**, Chen D, He H, Liu S, Moon S, Peterson KJ, Shen F, Wang L, Wang Y, Wen A, Zhao Y. Clinical concept extraction: a methodology review. Journal of Biomedical Informatics. 2020 Aug 6:103526.

6. **Fu S**, Carlson LA, Peterson KJ, Wang N, Zhou X, Peng S, Jiang J, Wang Y, Sauver JS, Liu H. Natural Language Processing for the Evaluation of Methodological Standards and Best Practices of EHR-based Clinical Research. AMIA Summits on Translational Science Proceedings. 2020;2020:171.

7. **Fu S**, Lopes GS, Pagali SR, Thorsteinsdottir B, LeBrasseur NK, Wen A, Liu H, Rocca WA, Olson JE, St Sauver J, Sohn S. Ascertainment of delirium status using natural language processing from electronic health records. The Journals of Gerontology: Series A. 2020 Oct 30.

Chapter 2

Background and Related Work

This chapter presents relevant background and related work of reproducibility research in the secondary use of EHRs. Section 2.1 describes a brief history of the definition of reproducibility and our approach to addressing reproducibility in the context of EHRs. Section 2.2 summarizes existing methods and approaches for enhancing EHR-based reproducibility issues.

2.1 Definition of Reproducibility

Reproducibility is the foundation of trusted discoveries and science advancement [55]. However, the terminology of reproducibility has been broadly defined across numerous domains and science (Table **??**). One of the early efforts in attempting to address reproducibility that involved digital computers was carried by Claerbout and Karrenbach in 1992 [56]. Their work introduced two related terminologies: reproducibility and replicability. Based on Claerbout and Karrenbachin's definition, reproducibility focus on being able to follow the same computational steps and execution to obtain the same result that does not require an expert; and replicability focuses on running new software to achieve "similar enough" results [56, 57]. In 2016, the Association for Computing Machinery (ACM) adopted the definitions from the experimental science community to distinguish the variation (original versus new) in experimental design and study team [58]. In the same year, a unique view of reproducibility was introduced by Goodman et al. The authors proposed to specify the word 'reproducibility' with descriptors for the underlying

construct of the measuring objective, which introduced method reproducibility, results reproducibility and inferential reproducibility [59]. In the informatics community, McIntosh et al applied the conceptual definition of empirical reproducibility to measure the information availability for re-run the experiment [60]. More recently, the National Academies of Sciences, Engineering, and Medicine (NASEM) defined reproducibility as computational reproducibility of using the same data, computational steps, methods, and code, and conditions of analysis to obtain consistent results.

Rather than debating for the single definition of reproducibility, we use the fitness-for-use [61] approach to study reproducibility in the context of secondary use of EHRs. This approach advocates the view that quality measures for reproducibility should be relative, context-specific and outcome-driven. In the context of EHR-based observational clinical research - cohort, cross-sectional, and case-control studies - the knowledge discovered from the study will be 1) disseminated through research publication, 2) compared, interpreted, and synthesized by systematic reviews, and 3) replicated by other study teams or institutions. In the context of learning health systems, the cyclical cycle of knowledge conversion requires the ability to re-run the experiments regularly for the continuous discovery of the latest evidence to improve the quality of care. Both scenarios can involve single or multiple institutions. As shown in the table 2.1, different scenarios demand different experiment replication designs and needs.

Table 2.1: Scenarios and Needs of Study replication

	EHR-based Clinical Research		**Learning Health System**	
	Single Institution	Multiple Institutions	Single Institution	Multiple Institutions
Need to be replicated by the original team	No	No	Yes	Yes
Need to be replicated by the a different team	Yes	Yes	No	Yes
Need to be replicated in the same institution	No	No	Yes	Yes
Need to be replicated in a different institution	Yes	Yes	No	Yes

In order to satisfy the needs from the above two scenarios, we proposed the **RITE** criteria, which comprised of four dimensions of reproducibility related measuring objectives: **R**eproducible, **I**mplementable, **T**ransparent, and **E**xplainable. The RITE criteria not only emphasize result reproducibility but also process transparency and implementability; since the variability and explainability of the result are dependent on the process (Figure 2.1). Even in a situation that does not require the process or results to be replicated, all important steps and details need to be documented and made available to ensure the traceability of the process and explainability of the result. We thus focus on *activities of replicating the prior process of information generation and information use of EHRs to assess the technical soundness of the process.* Two metrics of reproducibility were defined: 1) feasibility of completing the original steps (i.e., implementation feasibility) and 2) variability of results (i.e., clinical outcome variability).

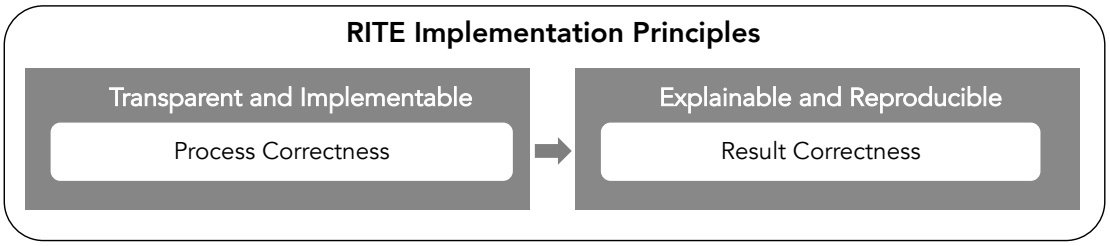


Figure 2.1: RITE Implementation Principles

Table 2.2: Comparison of Reproducibility Related Terminologies and Definitions

Terminology	**Year**	**Definition**	**Author**
Reproducibility	1992	"Running the same software on the same input data and obtaining the same results"	Claerbout and Karrenbach [56]
	2016	"The measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using artifacts which they develop completely independently."	ACM [62]
	2016	"Methods - provide sufficient detail about procedures and data so that the same procedures could be exactly repeated" "Results - obtain the same results from an independent study with procedures as closely matched to the original study as possible" "Inferential - draw the same conclusions from either an independent replication of a study or a reanalysis of the original study"	Goodman [59]
	2017	"The provision of sufficient methodological detail about a study so it could, in theory, or in actuality, be exactly repeated by investigators."	Denaxase et al [26]
	2019	"Obtaining consistent computational results using the same input data, computational steps, methods, and code, and conditions of analysis."	NASEM [63]

Terminology	Year	Definition	Author
Replicability	1992	"Writing and then running new software based on the description of a computational model or method provided in the original publication, and obtaining results that are similar enough"	Claerbout and Karrenbach [56]
	2016	"The measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, this means that an independent group can obtain the same result using the author's own artifacts."	ACM [62]
Repeatability	2016	"The measurement can be obtained with stated precision by the same team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same location on multiple trials. For computational experiments, this means that a researcher can reliably repeat her own computation."	ACM [62]
Generalizability	2019	"Results of a study apply in other contexts or populations that differ from the original one"	NASEM [63]

2.2 Existing Methods of Enhancing Reproducibility

Current efforts for enhancing the reproducibility of secondary use of EHRs can be summarized into following areas: data standards and interoperability [45, 28, 46, 64, 65, 66, 67], data pre-processing and analytics pipeline [68, 69, 70, 26, 71], data quality and heterogeneity assessment [72, 73, 74], reproducibility and replicability assessment [27], data and code repository [75], research design and implementation best practices [26, 52, 76, 77], and reporting guideline and regulation [50, 78, 79]. The topic

of data standards and interoperability has been widely studied due to the recent advancement in the standard community. For example, the OMOP Common Data Model (CDM) developed by the Observational Health Data Sciences and Informatics (OHDSI) community has been explored to promote reporting standards of clinical research [79]. More recently, the OHDSI community launched the Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND) research initiative. 10 guiding principles were recommended to address study bias, P hacking, and reporting bias [50]. Researchers also attempt to adopt a fitness-for-use approach and establish new standards to improve data capture, reuse, and sharing across organizations: Chow et al developed a nursing information model with repeatable steps and processes to enable the semantic interoperability of relevant and contextual nursing data [66]. Pan and Cimino created of knowledge representation of disease and database structure to facilitate clinical data retrieval through improving representational information quality of EHRs [67]. From the infrastructure and tooling perspective, the Critical Care Health Informatics Collaborative (CCHIC) developed a scalable EHR processing pipeline for extracting, linking, normalizing, curating, and anonymizing EHR data to enhance the data availability of multi-center EHRs [71]. The pipeline leveraged a common data model, validation of data provenance and de-identification [71]. In the clinical decision support (CDS) community, Wright et al identified 10 unique best practices in the area of CDS governance and content management by conducting specific site visits [80]. Asha et al conducted surveys and interviews followed by a template organizing method for identifying best practices []. In addressing research reproducibility, McIntosh et al. leveraged multi-phase approaches to extract recommendations and practices for improving reproducibility from publications for the development of a reproducibility framework [60]. Existing reporting guidelines such as RECORD (REporting of studies Conducted using Observational Routinely-collected health Data) and STROBE (The Strengthening the Reporting of Observational Studies in Epidemiology) serve as important standards for enhancing the reporting of clinical research. However, the reporting guideline for EHR and informatics-related methodologies are still missing. For example, reporting criteria for EHR-related data quality issues, EHR-specific biases, and implementation details of EHR-derived phenotyping algorithms.

In summary, the majority of existing research in the field focuses on reporting standards [81, 60], database standardization and cataloging [71], and system-level specification [82]. More importantly, the aforementioned works primarily focused on structured data. Our work differs from previous studies, which address reproducibility from the perspective of *process*, i.e., information quality life cycle (i.e., information collection and information use), and primarily focus on unstructured data.

Chapter 3

Latent Data Quality Impacted by EHR Systems Heterogeneity

3.1 Overview

NLP-assisted chart review offers unique opportunities for achieving real-time clinical decision support, risk management, and personalized patient monitoring [83, 84]. However, NLP models cannot be trusted without a good understanding of the data being fed into them. Consequently, the validity and portability of models are dependent on the data on which it is derived. EHR data is known to suffer from several data quality issues [20, 85]. As illustrated in Figure 3.1, since the primary functions of EHR systems are centered around patient-oriented care management and billing rather than for research purposes, the objective and priority of data documentation and reporting may vary significantly by providers and care settings [81]. The EHR system itself has a significant impact on the form and format of clinical data. Built-in documentation functionality such as templates, copy and paste, auto-documentation, and transcription can affect the EHR-specific syntactic and semantic definition for any data contained therein [86, 87]. These EHR-system-related specific factors can cause data heterogeneity, which measures the variability of information quality and semantic definition across heterogeneous data sources [7]. If NLP models are trained on data that cannot be reproduced due to a high level of variability, models may suffer the issues of portability and generalizability.

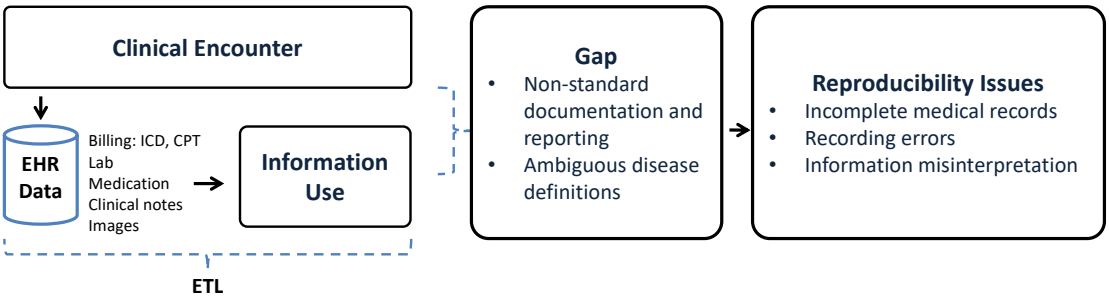


Figure 3.1: Issues of Reproducibility in the Context of EHRs

This chapter presents an investigation of the variation of data quality issues caused by the heterogeneous EHR systems. As EHR system functionality and information documentation patterns are deeply embedded within the clinical workflow and practice, we examined the quality of data for the given context (i.e., fitness for use). We considered three dimensions of data quality in the context of EHR-based modeling for three distinct transnational phases: model development (data completeness), model deployment (data variability), and model implementation (data timeliness). The data quality-related measurements were implemented in a real-world study of post-surgical complications (PSC) that comprised a wide range of clinical modalities collected from three stages of surgery (pre-operative, intra-operative, and post-operative). To the best of our knowledge, this is the first study that compares data heterogeneity of two EHR systems using the case matching design. We believe the pragmatic informatics methods presented by the study can be considered as potential data quality assessment methods for the implementation and translation of NLP models.

3.2 Methods

Study Setting Study Setting This study was approved by the Mayo Clinic Institutional Review Board. In May 2018, Mayo Clinic completed a large EHR migration and workflow standardization. The effort for Mayo Clinic Rochester campus included the conversion of the GE Centricity/LastWord EHR system (Centricity) to Epic EHR system (Epic). This migration offers an ideal scenario to study the difference between two EHR systems because confounding factors from inter-institutional variation can be mitigated due to

the entire study being conducted within a single institution. In addition, we used the case matching design to account for potential confounders contributed by patient population variation. Bins were created for the age variable with a fixed range of 5 years. We performed exact matching for age, sex, and type of surgery (Table 1). Two study cohorts with colorectal surgery performed as the primary procedure performed at Mayo Clinic Rochester were retrospectively constructed. Each cohort contains a total of 811 patients. (Table 3.1).

Table 3.1: Matching Criteria of Two EHRs

	GE Centricity (Pre-migration)	**Epic (Post-migration)**
Study Period	2017-01-01 - 2018-01-01	2019-01-01 - 2020-01-01
Total matched patients	811	811
Matching criteria	Age, sex, type and outcome of surgery (CPT: 44140, 44141, 44143-44147, 44150 - 44153, 44155 - 44158, 44160, 44204 – 44208, 44210 - 44212, 45110 - 45114, 45116 ,45119, 45123, 45395, 45397)	

Study Variables All study anticipates are part of the ACS National Surgical Quality Improvement Program (ACS NSQIP®) based on the Mayo Clinic Rochester campus. The program conducts a monthly evaluation of a sample (approximately 20%) of the colon and rectal surgery (CRS) practice based on standard procedure sampling methodology [88]. The NSQIP variables were defined by the ACS NSQIP abstraction guidelines and can be summarized into three stages: pre-operative, intra-operative, and post-operative (including post-hospitalization) (Figure 3.1). The key variables include patient demographic, comorbidities, preoperative labs (90 days before surgery), clinical, intraoperative elements, and postoperative occurrences/complications for 30 days after surgery.

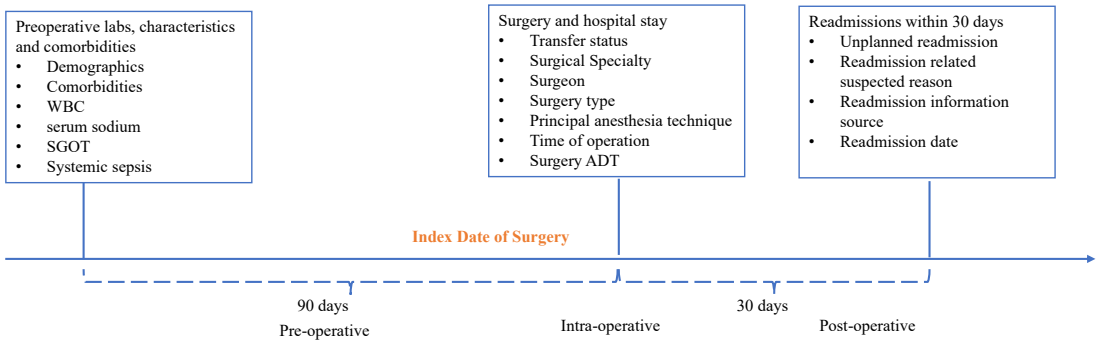


Figure 3.2: Study Timeline and Example Variables in Three Stages of Colon and Rectal Surgery

Data collection Definitions of data collection and abstraction were standardized and aggregated with 18 other participating institutions across the nation [89]. The structured data consists of 102 unique variables falling under categories of demographic data, patient-provided information (PPI), symptoms, comorbidities, physiologic measurements, laboratory tests, observational assessments, and operative factors. The data was retrieved from the Mayo Unified Data Platform (UDP) using an R-based application programming interface (API). The UDP is an enterprise data warehouse that loads data directly from the Mayo Clinic EHR. Patient comorbidities were found from ICD-9 and 10 codes recorded within one year of surgery. PPI was measured and collected at the time of admission. Symptoms, physiologic values, laboratory tests, and observational factors were abstracted from two weeks before surgery until the start of surgery. In addition to the 102 variables abstracted from EHR, another 16 were generated using NLP as a Service [30], a Mayo Clinic internal natural language processing (NLP) platform for extracting medical information from unstructured text. This system was developed based on an open-source NLP framework MedTaggerIE [36]. In total, 118 variables were created for the final data set.

Measurements of EHR Variability To examine the potential variability of information quality caused by two EHR systems, we consider three dimensions of data quality: data completeness, data variability (syntactic, semantic, textual and lexicon), and data timeliness, as listed in Table 8.3.

Table 3.2: Definitions of Dimensions in the Context of EHRs

Dimensions	**Definition**
Data completeness	A record contains all observations, all desired types of data, and a specified frequency of data over time
Data variability (syntactic)	The structure (or syntax) of exchanged data
Data variability (semantic)	The meaning (or semantics) of exchanged data
Data timeliness	The measurement of time expectation for accessibility and availability of data

Data Completeness As suggested by Juran and Weiskopf et al, data completeness needs to be viewed as context-dependent and fitness for use [90, 44]. We used the NSQIP as the reference standard to assess the data completeness. Each clinical encounter was defined as a colorectal surgery period using surgery operation date as the index date (Figure 3.1). The completeness of data is measured by the presence of a reference standard given all observations made about a patient [44]. We used the rate of missing (RoM) to calculate the presence of information frequently found in CRS patients. A missing event for pre-operative variables was defined as the absence of the information within 90 days prior to the surgery index date. The perioperative duration was calculated using admit and discharge date. The duration for post-operative variables was 30 days after the surgery. To further understand and measure the RoM variation, we organized the variables by seven unique data sources including admit, discharge, and transfer status (ADT), billing code, patient demographics, vital signs, laboratory result, clinical note, and surgery information. Three different stages (pre-operative, intra-operative, and post-operative) were also assigned to each variable. McNemar's test was performed to determine the statistically significant difference in the data completeness between Centricity and Epic [91].

Data Variability The health level seven (HL7) messages of unstructured clinical notes within one month before and after the surgery date were retrieved for the two matched cohorts. The HL7 Clinical Document Architecture (CDA) is a standard XML format for the syntactic representation of clinical documents based on the Reference Information Model (RIM) [92]. The general structure of a CDA document is

comprised of 1) document header or metadata information such as document date, document creator, and service location, and 2) narrative text in the body of the document. Based on the definition proposed by Elkin et al and Sohn et al, syntactic variability was examined by comparing meta-structure, documentation sections of the HL7 messages, and calculating corpus statistics [42, 93]. The following metrics were considered: tokens/section, tokens/document, tokens/patient, sections/document, sections/patient, and documents/patient15. The statistically significant difference between the two sites was determined using Wilcoxon signed-rank test [94].

The semantics variability was examined by comparing the number of PSC concepts per patient across two EHR systems. The PSC concepts were extracted by an existing NLP algorithm30. Since the original algorithm was developed and evaluated using the Centricity data only, we conducted corpus annotation and NLP refinement on 100 patients with roughly 1200 Epic clinical notes (within one month before and after surgery index date). Corpus annotation is a process of marking interpretative linguistic and predefined clinical concepts. The annotation was conducted by the same annotator (DMI) who participated in the previous study and had gone through training and consensus development. The same annotation guideline, annotation software (MAE), and schema were applied. The 100 patients were randomly split into 50 training and 50 test sets. The out-of-box (i.e., directly applied with no refinement) precision, recall, and f1-score for NLP were 0.72, 0.84, and 0.79, respectively. After the refinement of the training data, the final performance on test data was 0.92 in f1-score. Two versions of the NLP algorithm were applied separately to two cohorts (Centricity and Epic).

Furthermore, we assessed the variability of semantic textual similarity (STS) of the positive mentions extracted from the NLP algorithm. We measured the sentence pair textual similarity using the averaged value of three surface lexical similarities, which include a string-matching algorithm proposed by Ratcliff and Obershelp, cosine similarity of two-word vector space, and Levenshtein distance [95, 96, 97]. The method was utilized and evaluated in the 2018 BioCreative/OHNLP clinical semantic textual similarity challenge [96]. A high similarity sentence pair is determined when the average score was greater or equal to 0.40. Based on the concept distribution, we examine the textual similarity of two frequent concepts - Anemia and Abscess and the two least frequent concepts – Purulent Drain and Wound Infection (with minimal 50 sentence pairs

per section). We calculate both intra-EHR similarity (i.e., comparison within the same EHR system) and inter EHR similarity between Epic and Centricity. The distributions of the unique clinical expressions were visualized using histogram charts.

Data Timeliness In the era of achieving real-time clinical decision support and prospective risk detection, information timeliness becomes an important quality criterion since the timeliness of the model is dependent on data. Data timeliness was defined as the time expectation of whether information can be accessible given each patient encounter [98]. The analysis of data timeliness for structured data was focused on lab variables due to their high prevalence and importance to the prediction of PSC [99]. We retrieved the lab result record date (i.e., the date when a record loaded to the source system) and compared it with the patient encounter date. For unstructured data, we retrospectively collected and measured the time spent on the documentation of clinical notes for each patient visit (i.e., comparison of note date on source system and encounter date). To simplify the measure, we define timely information as the data that can be accessed within 24 hours of a CRS-related clinical encounter.

3.3 Results

Data completeness The overall comparison of RoM across the two EHRs was illustrated in Figure 5.3. Among a total of 118 variables studied, the median rate of missing for Centricity is 0.011 (1st IQR 0, 3rd 0.71), whereas it was 0.007 (1st IQR 0, 3rd IQR 0.665) for Epic. We observed a high RoM among the intraoperative variables (green dots) compared with the postoperative variables (red dots). There was no significant pattern discovered for the comparison of measurement and temporal variables. A zero to mild difference was discovered for both highly complete variables (RoM < 0.1) and highly incomplete variables (RoM > 0.85). On the other hand, there was a high variation among variables with RoM between 0.1 to 0.85 across two EHRs (Figure 2- Area of High Heterogeneity).

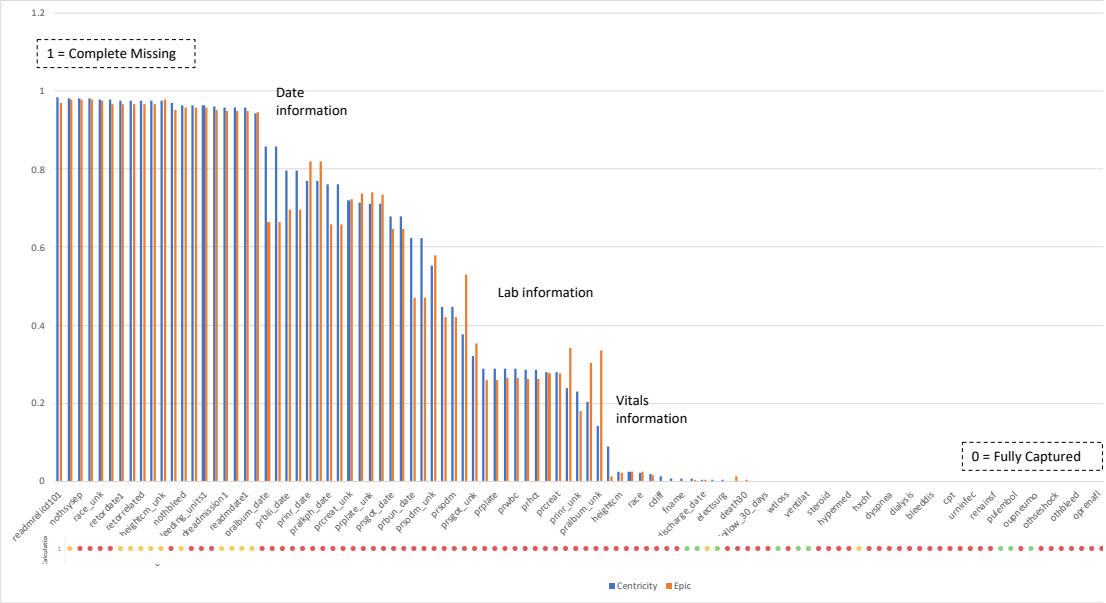


Figure 3.3: Comparison of the Information Completeness Across Two EHRs

Table 3.3 provides the aggregated comparison of RoM by operation stage and data source. No significant difference was found for the data collected in the intra-operative and postoperative stages. However, there was a significant differences in the level of RoM for unknown and lab-related variables. Based on the McNemar's test, variables with significant difference in the level of RoM were Serum albumin ($p<0.001$), BUN ($p<0.001$), Bilirubin ($p<0.001$), Alkaline phosphatase ($p<0.001$), C. Diff ($p<0.001$), Transfer status ($p=0.002$), International Normalized Ratio (INR) of PT values ($p=0.012$).

Table 3.3: Comparison of Information Completeness by Operation Stage and Data Source

Stage	Original Data Source	Total No. of Variables	No. of Significant Variables (%)
Preoperative	ADT	4	0 (0)
	Billing (D)	16	0 (0)
	CN	26	1 (4)
	Demo	6	2 (33)
	Lab	36	16 (44)
	Vitals	4	0 (0)
Intraoperative	Billing (P)	1	0 (0)
	Surgery	11	1 (9)
Postoperative	ADT	10	0 (0)
	Billing (P)	3	0 (0)

Variables were organized by the stage of surgery and summarized by the original data source. Statistically significant difference was determined by paired McNemar's test of the data completeness between Centricity and Epic, a significant variable was defined as $p<0.05$; Abbreviation: ADT: admit, discharge, and transfer status, CN: clinical notes, Billing (D): diagnosis code, Billing (P): procedure code.

Data Variability The comparison of the corpus statistics between Centricity and Epic was provided in Table 3.4. We observed a larger number of clinical documents, tokens, and sections (total) in Epic compared with Centricity. Because the total number of documents and sections for Epic has increased, the number of sections/patient and documents/patient were higher than Centricity. On the other hand, the median of the number of tokens/patient for Epic was lower than Centricity despite the fact that the total number of documents and tokens were higher. Based on the Wilcoxon signed-rank test, all five corpus statistics metrics were found to be significant for the comparison between two EHRs. Overall, it is evident that two systems have different ways of organizing clinical documents.

Table 3.4: Example of Language Variation between Two Data Sources

Original data source	**Centricity**	**Epic**	***p*-value**
No. of patients	811	811	
No. of documents	18,648	30,476	
No. of tokens (Total)	8,273,327	11,383,088	
No. of sections (Total)	94,645	116,399	
No. of sections (Unique)	64	47	
No. of tokens/section, median (IQR)	29 (95)	64 (90)	<0.001
No. of tokens/document, median (IQR)	243 (440)	229 (328)	<0.001
No. of tokens/patient, median (IQR)	35,927 (42,828)	26,744 (36,367)	<0.001
No. of sections/document, median (IQR)	4 (5)	3 (5)	0.0012
No. of sections/patient, median (IQR)	81 (69)	94 (92)	<0.001
No. of documents/patient, median (IQR)	15 (14)	26 (26)	<0.001

**IQR: interquartile range. Statistically significant difference was determined by paired Wilcoxon signed-rank test.*

Based on the semantic mapping and analysis of the document sections across two EHRs, there was a high similarity of the clinical document sections across the two systems (Figure 3.4). Among the total 94,645 sections in Centricity, the top three sections were "Impression/Report/Plan" (14,203), "Chief Complaint/Reason for Visit" (7,106), and "Physical Examination" (5,853). The three most prevalent sections for Epic were Impression/Report/Plan" (20,392), "Procedure Information" (8,267), and "Physical Examination" (6,628). Among the top 15 sections, 9 sections were matched, including 'Impression/Report/Plan', 'Chief Complaint/Reason for Visit', 'Physical Examination', 'History of Present Illness', 'Vital Signs', 'Subjective', 'Diagnosis', 'Procedure Information', and 'Social History'.

Centricity		Epic	
Impression/Report/Plan	14203	**Impression/Report/Plan**	20392
Chief Complaint/Reason for Visit	7106	**Procedure Information**	8267
Physical Examination	5853	**Physical Examination**	6628
History of Present Illness	4925	**Chief Complaint/Reason for Vis**	5733
Vital Signs	4363	**Subjective**	5478
Subjective	3853	**History of Present Illness**	5422
Diagnosis	3850	**Vital Signs**	4209
Current Medications	3254	**Diagnosis**	3714
Referral Source	2918	Labs	2650
Allergies	2879	Past Medical/Surgical History	2481
Administrative	2867	System Reviews	2073
Past Medical/Surgical History	2470	Family History	2013
Pre Procedure Information	2406	**Social History**	1702
Procedure Information	2405	Anesthesia	1545
Social History	2147	Patient Disposition	1488

Venn Diagram Comparison of Section Headers

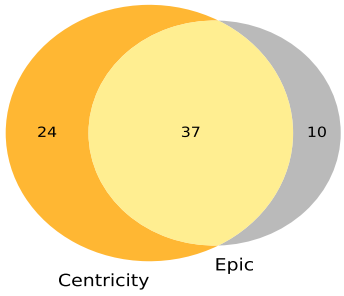


Figure 3.4: Semantic Mapping and Comparison of Document Sections across two EHRs

The overall summary statistics of the number of PSC-related clinical concepts extracted by NLP was provided in Table 3.5. Overall, the concept/document ratios and concept/patient ratios for Epic (blue and orange columns) were lower for all concept types and significantly lower for Anemia, Abscess, Cellulitis and Painful incision; this pattern indicates that Centricity has potentially higher semantic breadth in the context of PSC.

Table 3.5: Semantic Concept Distribution of Two EHRs

	Concept/Document Ratio	**Concept/Document Ratio**	**Concept/Patient Ratio**	**Concept/Patient Ratio**
PSC Concept	Centricity	Epic	Centricity	Epic
Anemia	0.1058	0.0524	1.8947	1.1690
Abscess	0.0813	0.0218	1.4562	0.4856
Cellulitis	0.0575	0.0097	1.0292	0.2158
Painfulincision	0.0201	0.0017	0.3592	0.0371
Purulentdrain	0.0076	0.0013	0.1356	0.0293
Woundinfec	0.0063	0.0015	0.1126	0.0339
Wounddehiscence	0.0020	0.0006	0.0365	0.0136
Fascialdehiscence	0.0017	0.0008	0.0302	0.0170
Abdomtender	0.0010	0.0007	0.0188	0.0165
Infecabdom	0.0009	0.0004	0.0162	0.0097
Fever	0.0008	0.0001	0.0146	0.0017
Reopen	0.0001	0.0001	0.0010	0.0015

Based on the concept distribution from Table 3.5, we further examined the textual similarity of two most frequent concepts, anemia and abscess, and two least frequent concepts, purulent drain and wound infection. Figure 3.5 presents the summarized

textual similarity scores for intra and inter EHR comparison. Since the document section plays an important role in contextual information, this analysis was further stratified by document sections. Compared with Epic, the intra-EHR textual similarity of Centricity was higher for all disease categories and the majority of the sections. On the other hand, Epic yielded a substantially higher similarity under the 'Secondary Diagnoses' section. Among the section dimension, the similarity difference of 'Diagnosis', 'Past Medical/Surgical History', 'Secondary Diagnosis', and 'History of Present Illness' was substantial. For most clinical concepts and document sections, there was no substantial drop in inter-EHR similarity.

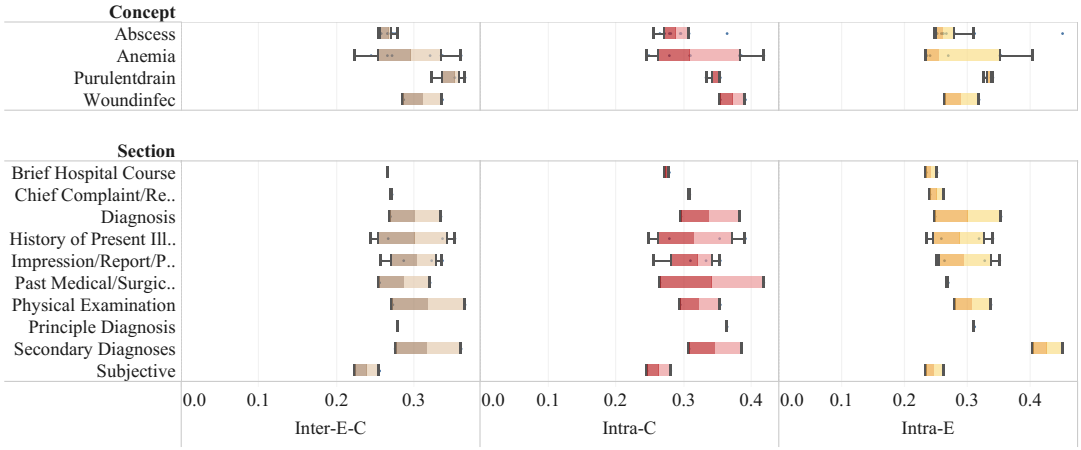


Figure 3.5: Comparison of the Textual Similarity between Centricity and Epic. High similarity: greater or equal to 0.40, Abbreviation: Intra-C: Intra-Centricity, Intra-E: Intra-Epic, Inter-E-C, Inter-Epic-Centricity

The distribution of the unique clinical expressions related to abscess and anemia (Figure 3.6) revealed two completely opposite patterns: Epic has more standardized language for abscess, whereas Centricity is lengthy and descriptive. On the other hand, the language representing anemia for Epic was more variant than Centricity. For example, the expression of "Anemia Posthemorrhagic Acute (Blood Loss Anemia)" was repetitively documented for more than 30% of the total sample size. The varying similarity patterns affirm that the characteristics and patterns of data are context-dependent.

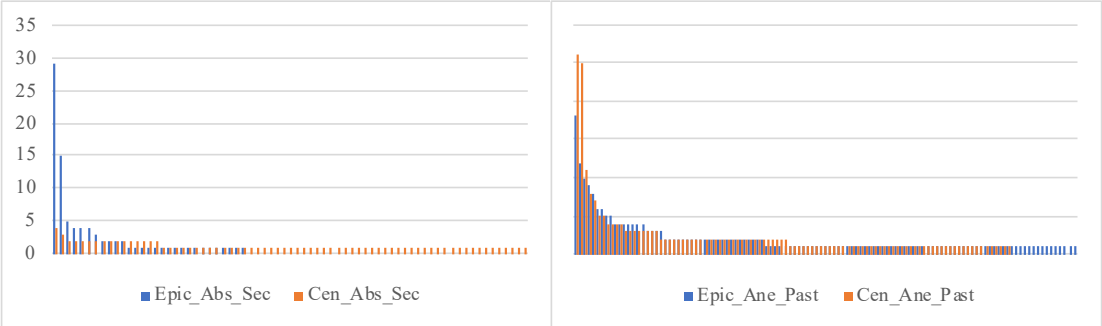


Figure 3.6: Comparison of the Concept Distribution between Epic and Centricity. Figure (left): distribution of abscess-related concepts under the section of Secondary Diagnosis, figure (right): distribution of anemia-related concepts under the section of past medical/surgical history between Epic (blue) and Centricity (orange). X-axis: unique clinical expressions related to abscess and anemia; Y-axis: frequency of expression. Bars skewed to the left: indication of high language repetition; bars scatter to the right: indication of low language repetition.

Data timeliness There was no delay of information found in the structured lab variables across two EHR systems. Amongst the total 811 patients who had CRS under the centricity EHR, there were a total of 1855 visits and 1673 instances of on-time documentation and 182 instances of delayed documentation. For the other 811 matched patients under Epic EHR, there were a total of 3260 encounters, of which 44 encounters had documentation delay. The delayed documentation rate for Centricity and Epic cohorts were 0.098 and 0.013, respectively.

3.4 Discussion

The translation of EHR-derived findings into routine clinical care faces challenges in the form of various data quality issues caused by the heterogeneity of EHR systems. To better understand this barrier, we retrospectively assessed the variability of data from two EHR systems in the context of PSC. We discovered a consistent level of data completeness across two EHR systems with an exception for lab data. To further understand Epic's significant improvement of capturing lab data in the context of CRS, we investigated the workflow difference between two EHRs. We learned that after EHR migration,

there was a process change with how laboratory tests can be ordered. The migration enables the primary providers to order the laboratory test directly through the Epic EHRs, which may explain the lower RoM score. Conversely, there was a high syntactic variation suggested by the corpus statistics. There was also a moderate-high difference in the semantic type and frequency of document sections. Textual similarity revealed a consistent pattern for roughly half of the concept-section pairs. High language variation was found for the sections of Secondary Diagnosis (Abscess) and Past Medical/Surgical History (Anemia). The data timeliness of clinical notes documentation for Epic was improved when compared with Centricity. The improvement of information timeliness from the Epic system suggests a potential higher utilization of auto- or assisted documentation. However, confirmation of this finding requires additional on-site evaluation, which we have left to a future study.

The validity and reliability of clinical data are crucial for the development of robust, safe, and scalable NLP models. However, data is often being viewed as the least incentivized aspect by ML researchers [100]. Dealing with data can indeed be challenging; for example, data curation and wrangling can be time-consuming and tedious, especially in the context of secondary use of EHRs, where researchers have limited control of how the data is documented and standardized. On the other hand, because these latent factors (e.g., variant patterns of documentation) may introduce systematic bias and measurement error, a solid understanding of how data is documented, defined, and collected is required prior to adoption of any predictive models relying upon it. Solid data understanding can promote a good data curation plan and solutions for mitigating potential biases or confounders prior to model development and re-deployment. The transparency of information documentation has a direct implication to the explainability, implementability, and ultimately the trust of the models derived from the data. Based on our case study investigation on CRS patients, the EHR system plays an important role in how data is documented, defined, and organized. In a situation when a model will be translated to care practice or deployed to a different environment, proper data quality assessment needs to be conducted including the comparison of data characteristics and variability between the destination environment and the development environment.

Our investigation confirmed that the quality of data needs to be viewed from the context of data being generated and documented. For example, the results from Figure 3.5

discovered high similarity patterns in two EHR systems under different contextual factors (disease-section combination). Chart review was conducted to confirm expressions with high textual similarity were associated with the use of documentation templates. Although the use of templates may enhance the documentation standardization, the clinician's reasoning process may be eliminated. The direct implication to machine learning models may be a varying level of contextual knowledge loss36. The varying results from the analysis also strongly indicated a proper model re-training, refinement, and re-evaluation are needed.

Our study has several limitations. Since the study was conducted on a single-case scenario, the generalizability of the findings is limited by the scope of the study. We aim to expand our investigation on multiple different institutions with diverse case scenarios as part of future work. Furthermore, we plan to leverage qualitative methods to study the workflow of data documentation and transformation across multiple EHR systems.

3.5 Conclusion

To better understand the potential data heterogeneity caused by different EHR systems, we proposed and applied a standardized set of informatics methods to retrospective assess the variability of data quality contributed by two EHR systems. We discovered a varying level of data quality across two EHR systems, for which the quality of data is context-specific and closely related to the documentation workflow and the functionality of individual EHR systems. We recommend that data understanding should be equally incentivized as model development.

Chapter 4

Latent Data Quality Impacted by Institutional and Process Variation

4.1 Overview

Findings of a single site study must be able to be independently validated at different sites. It is very challenging to validate an EHR-based study, particularly due to the variation in institution-specific choice, design, and implementation of health IT infrastructure and applications, process and workflow, policy and guideline, and social-demographics characteristics. In this chapter, we assess the heterogeneity of institutional and process variations based on a multi-site EHR-based comparative effectiveness study of stroke prevention. The study involves multiple steps to generate a corpus for the development of complex phenotype algorithms. The heterogeneity of healthcare institutions, EHR systems, documentation, and process variation in case identification was assessed quantitatively and qualitatively.

4.2 Methods

To assess the data variation caused by institutional differences, we conduct a second case study based on two healthcare institutions: Mayo Clinic and Tufts Medical Center. This study is designed as an EHR-based comparative effectiveness study aiming to estimate the preventive therapies on the risk of future stroke and dementia in patients with

incidentally-discovered brain infraction [101, 7]. The study has been approved by the Mayo Clinic and Tufts Medical Center institutional review boards.

Study Site Mayo Clinic is a tertiary care, nonprofit, academic medical center. Mayo Clinic is a referral center with major campuses in three regions of the U.S. including Rochester, Minnesota; Jacksonville, Florida; and Phoenix/Scottsdale, Arizona as well as Mayo Clinic Health System locations that serve more than 70 communities in Iowa, Wisconsin, and Minnesota. The organization attends to nearly 1.2 million patients each year, who come from throughout the United States and abroad. The Saint Mary's (1,265 licensed beds) and Rochester Methodist (794 beds) campuses are two main hospitals located in Rochester, Minnesota. Tufts Medical Center is similarly a tertiary care, nonprofit, academic medical center that is located in Boston, MA and is the principal teaching hospital of the Tufts University School of Medicine. The 415 licensed bed medical center provides comprehensive patient care across a wide variety of disciplines with disease-specific certifications through the Joint Commission as a Comprehensive Stroke Center and transplant center. TMC is the referral center for the WellForce network serving communities throughout Eastern Massachusetts and New England (Maine, New Hampshire, Vermont, Rhode Island). The medical center is actively engaged in clinical research and medical education with ACGME-accredited residencies and fellowships.

Disease Silent brain infarction (SBI) is the presence of one or more brain infarcts, presumed to be due to vascular occlusion, found by neuroimaging in patients without clinical manifestations of stroke [102, 103, 104]. It is more common than a stroke and can be detected in 20% of healthy elderly people [102, 103, 104]. Early detection of SBI may prompt efforts to mitigate the risk of stroke by offering preventative treatment plans. In addition to SBI, white matter disease (WMD) or leukoaraiosis is another common finding in neuroimaging of older patients. SBI and WMD are related, but it is unclear whether they result from the same, independent, or synergistic processes [105, 106]. Since SBIs and WMDs are usually incidentally detected, there are no related International Classification of Diseases (ICD) codes in the structured fields of EHRs to facilitate large-scale screening. Instead, the findings are usually recorded in neuroimaging reports, so NLP techniques offer an opportunity to systematically identify SBI and WMD cases in EHRs.

To assess the data variability across two institutions, we conducted retrospective data

abstraction and annotation from two EHRs and examined the variations that occurred during the process.

Protocol development

A screening protocol was co-developed by the two institutions using procedure codes, diagnosis codes, and problem lists. The protocol included ICD-9 and ICD-10 codes to identify non-incidental clinical events. The codes were expanded with the corresponding descriptions to enable us to perform a text search. The initial criteria were developed by a vascular neurologist at TMC and were evaluated by two neuroradiologists and one internist. The inclusion criteria were defined as individuals with neuroimaging scans between 2009 and October 2015. The exclusion criteria included patients with clinically-evident stroke, transient ischemic attack (TIA), and dementia any time before or up to 30-days after the imaging exam. TIA was considered an exclusion criterion as TIA is sometimes incorrectly assigned on occasion by clinicians as the diagnosis in the setting of transient neurologic symptoms and positive evidence of brain infarction on neuroimaging. Dementia was an exclusion criterion because of a projected future application of the NLP algorithm in identifying patients for comparative effectiveness studies or clinical trials for which both stroke and dementia could be outcomes of interest. The systematic reviews suggested that the U.S. population over 50-years old had a high average prevalence of SBI [44]. By identifying a large cohort of patients with SBIs, age restriction was applied to exclude individuals 50-years of age or younger at the time of the first neuroimaging scan.

Data collection

At TMC, the data was aggregated and retrieved from three EHR systems: General Electric Logician, eClinicalWorks, and Cerner Soarian. The EHRs in TMC were implemented in 2009 with 1,031,159 unique patient records. At Mayo Clinic, the data was retrieved from the Mayo Unified Data Platform (UDP), an enterprise data warehouse that loads data directly from Mayo EHRs. Mayo EHR was implemented in 1994. Currently, there are 9,855,533 unique patient records. To allow data sharing across the sites, we de-identified the data by applying the de-identification tool DEID [107], a Java-based software that automatically removes protected health information (PHI) in

neuroimaging reports with manual verification where an informatician, an abstractor and a statistician manually reviewed all the output from DEID.

Cohort screening

At Mayo Clinic, an NLP system, MedTagger [108], was utilized to capture mentions from the exclusion list in the clinical notes. As the system has a regular expression component, language variations such as spelling and abbreviations were able to be captured. Structured ICD-9 and ICD-10 codes were obtained by an informatician from the UDP. A clinician and an abstractor manually compared the screened cohort with the EHRs to ensure the validity of the screening algorithm.

At TMC, due to infrastructure limitations, this process was conducted through a manual chart review. To ensure reproducibility, we carefully documented each step of the workflow. Briefly, a vascular neurologist and three research assistants conducted manual chart review in order to determine whether individuals were included or excluded appropriately at each step. This process was performed using a list of free text exclusion criteria associated with the exclusionary ICD-9 and ICD-10 codes. It involved review of the full text of any discharge summaries associated with the encounter during which the neuroimaging scan was obtained in Cerner Soarian, if present, as well as review of the neuroimaging scan indication in the neuroimaging report.

Each site randomly selected 500 eligible reports to form the raw corpus for guideline development and corpus annotation. The cohort consisted of 1400 reports with 400 duplications for double reading. Among the total 400 double-read reports, 5 reports were removed because of invalid scan types. The remaining 395 reports were comprised of 207 from Mayo and 188 from TMC.

Guideline development

A baseline guideline was created by a vascular neurologist based on domain knowledge and published literature. To develop the annotation guideline, 40 reports pooled from the two institutions were annotated by two neuroradiologists and one neurologist using the baseline guideline. Inter-annotator agreement (IAA) was calculated and a consensus was organized to finalize the guideline, which included task definitions, annotation instructions, annotation concepts, and examples.

Corpus annotation

The annotation processes consist of two tasks: neuroimaging report annotation and neuroimage interpretation. Neuroimaging report annotation is the process of reading and extracting SBI and WMD related sentences or concepts from text documents. Neuroimage interpretation is the process of identifying SBIs or WMDs from CT or MRI images. Figure 4.1 provides an example of two tasks.

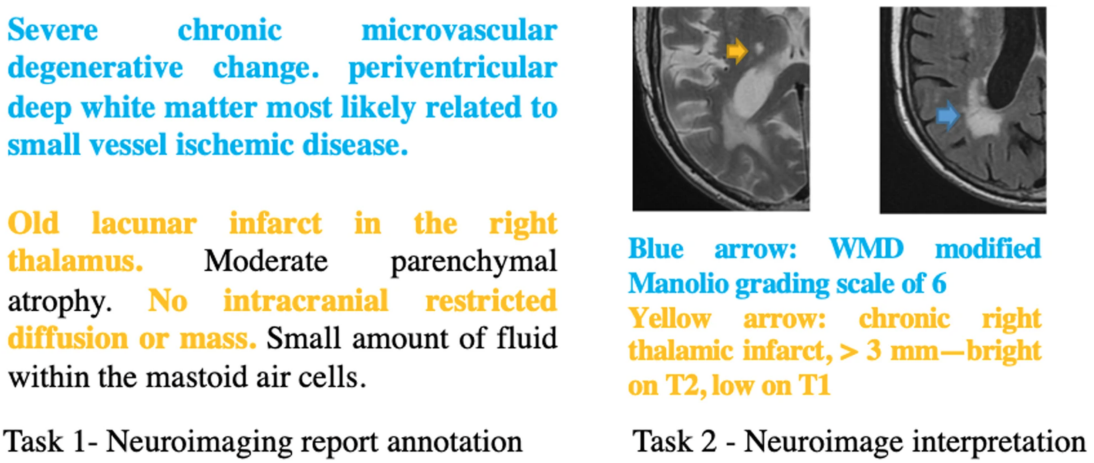


Figure 4.1: Example of neuroimaging report annotation (left) and neuroimage interpretation (right) for SBI (yellow) and WMD (blue)

Neuroimaging report annotation The purpose of the annotation task was to annotate the findings of SBI and WMD lesions in both the body (Findings) and summary (Impression and Assessment) sections of neuroimaging reports. The annotation was organized into two iterations. The first iteration extended from the finalization of the process guideline until the midpoint when half of the reports were annotated. The goal of the first iteration was to identify new problems that were not captured in the sample data. After the first iteration, all problematic cases were reviewed by the two senior clinicians, and the guidelines were updated. The second iteration of annotation then commenced using the updated guidelines. Several consensus meetings were organized to resolve all disagreements after the annotation process was completed. All conflicting cases were adjudicated by the two senior clinicians. All of the issues encountered during

the process were documented.

The annotation team was formed with members from both institutions. Two third-year residents from Mayo and two first-year residents from TMC performed the annotation. The experts for quality control were two senior radiologists from Mayo and one senior internist and one vascular neurologist from TMC. We used Multi-document Annotation Environment (MAE) [109], a Java-based natural language annotation software package, to conduct the annotation.

Prior to annotation, training was conducted for all four annotators including one online video session and two on-site training sessions. The online video provided demonstrations and instructions on how to download, install, and use the annotation software. The on-site training conducted by two neuroradiologists contained initial annotation guideline walkthroughs, case studies, and practice annotations. The same clinicians supervised the subsequent annotation process.

Neuroimage interpretation To assess the validity of the corpus, we obtained a balanced sample of images with and without SBI from the annotated neuroimaging reports. From each site, 81 neuroimages were de-identified and reformatted to remove institution-specific information and then pooled together for the sample group. We invited four attending neuroradiologists, two from each site, to read grade the imaging exams. Each exam was graded twice by two neuroradiologists independently. The image reading process followed the proposed best practices including guideline development, image extraction form, training, and consensus building. The level of agreements between the research-grade reading of the neuroimages and the corresponding annotation of the reports was calculated.

Assessment of heterogeneity

The screening ratio was calculated on the post screened cohort. Cohen's kappa [110] and F-measure [111] were adopted to measure the IAA during the annotation and image reading processes. Corpus statistics were used to measure the variations in clinical documentation across institutions. The analysis compared corpus length, number of SBI and WMD concepts, number of documents with SBI and WMD concepts, and distribution of SBI related concept mentions. Document similarity was calculated by comparing the cosine similarity between two vectors created by term frequency-inverse

document frequency (tf-idf), where each corpus was represented by a normalized tf-idf vector [42]. Age-specific prevalence of SBI and WMD were calculated and compared with the literature. To analyze the cohort characteristics between Mayo and TMC, Student's t-test was performed for continuous variables. Comparison of categorical variables was calculated with frequency tables with Fisher's exact test.

Qualitative assessments were conducted to evaluate the abstraction process and an assessment protocol was created to facilitate the post abstraction interview. The protocol was designed to focus on three main areas: 1) evaluation of the abstraction process, 2) language patterns in the reports, and 3) abstraction techniques. Four back-to-back interviews were conducted with the four abstractors following the guidelines of Contextual Interview (CI) suggested by Rapid Contextual Design [112]. Each interview was conducted by an informatician and lasted approximately 30 min. Questions and issues raised by each annotator during the two iterations of annotation were collected and qualitatively assessed. The data were then classified into six categories: data, modifier, medical concept, annotation rules, linguistic, and others.

4.3 Results

Neuroimaging report annotation The average inter-annotator agreements across 207 Mayo reports and 188 TMC reports on SBI and WMD were 0.87 and 0.98 in kappa score and 0.98 and 0.99 in F-measure, respectively. Overall, both Mayo and TMC annotators achieved high inter-annotator agreements.

Neuroimage interpretation The average inter-annotator agreement among four neuroradiologists was 0.66 in kappa score and 0.83 in F-measure. The average agreement between neuroimaging interpretation and corpus annotation was 0.68 in kappa score and 0.84 in F-measure. The result suggested high corpus validity outcomes.

Institutional variation The process of screening eligible neuroimaging reports across two institutions was variant. At Mayo, 262,061 reports were obtained from Mayo EHR based on the CPT inclusion criteria. 4015 reports were randomly sampled for cohort screening. 749 were eligible for annotation after applying the ICD exclusion criteria (structured and unstructured). At TMC, 63,419 reports were obtained from TMC EHR based on CPT inclusion criteria. 12,092 reports remained after applying the ICD

exclusion criteria (structured). 1000 reports were randomly selected for text screening, a method of identifying eligible patients using NLP techniques to extract eligibility criteria from patient clinical notes. 773 reports were eligible for annotation. Among the total 1522 eligible (Mayo 749, TMC 773) neuroimaging reports, 1000 (Mayo 500, TMC 500) reports were randomly selected.

The prevalence of SBI and WMD for Mayo and TMC patients at age of 50, 60, 70, and 80 is listed in Table 4.1. Despite the variation, the results were consistent with the published literature, between 10 and 20% [102, 103], and the number increased with age in both computed tomography (CT) and magnetic resonance imaging (MRI) as anticipated.

Table 4.1: The prevalence of SBI and WMD for Mayo and TMC patients at age of 50, 60, 70 and 80

	SBI				WMD			
	CT Scan (%)		MRI Scan (%)		CT Scan (%)		MRI Scan (%)	
Age	Mayo	TMC	Mayo	TMC	Mayo	TMC	Mayo	TMC
>=50	12.5	7.4	11.3	7.7	28.7	55.0	69.2	51.7
>=60	16.0	9.4	14.0	9.7	35.1	65.9	75.3	60.2
>=70	23.5	11.4	20.2	12.2	47.1	80.7	84.6	65.3
>=80	26.3	18.4	26.5	20.8	52.6	94.7	85.3	66.7

The average age of Mayo and TMC patients 65 and 66, respectively. The number of female patients in the Mayo and TMC cohort were 243 and 274, respectively. We found a moderate variation in the presence of SBI and WMD and a high variation in the WMD grading. A significant variation in the missing documentation of WMD grading between Mayo and TMC was found ($p = 0.0024$). Table 4.2 summarizes the cohort characteristics across two institutions.

Table 4.2: Attributes of SBI and WMD for Mayo and TMC patients

Variables	**Mayo (n =500)**	**TMC (n =500)**	**p Value**
Age (mean)	65 (±10.6)	66 (±9.7)	0.1197
Gender (female)	243 (48.6)	274 (54.8)	0.0576
SBI	57 (11.4)	38 (7.6)	0.0516
Acuity			
Acuity/subacute	6 (1.2)	6 (1.2)	1.0000
Chronic	44 (8.8)	29 (5.8)	0.0882
Non-specified	7 (1.4)	3 (0.6)	0.3407
Location			
Lacunar/subcortical	27 (5.4)	10 (2.0)	0.0065
Cortical/juxtacortical	9 (1.8)	13 (2.6)	0.5188
Both	0 (0)	3 (0.6)	0.2492
Non-specified	21(4.2)	12 (2.4)	0.1558
WMD	291 (58.2)	264 (52.8)	0.9800
WMD grading			
Mild	191 (38.2)	154 (30.8)	0.0165
Mild/moderate	21 (4.2)	0 (0.0)	7.6963e-7
Moderate	42 (8.4)	45 (9.0)	0.8226
Moderate/severe	2 (0.4)	0 (0)	0.4995
Severe	8 (1.6)	11 (2.2)	0.6443
No mention of quantification	27 (5.4)	54 (10.8)	0.0024

EHR system variation There was a high variation in the EHR system vendors, the number of EHR systems per site, and the extract, transform, and load (ETL) processes for the different EHR systems between Mayo and TMC. At TMC, the data was obtained directly from three EHR systems: General Electric Logician, eClinicalWorks, and Cerner Soarian. The data retrieval process involved different abstraction processes due to the different interface design and data transfer capabilities. At Mayo Clinic, there was an ETL process to aggregate the data from Mayo EHRs to the enterprise data warehouse. Since data could be linked and transferred through direct queries, the abstraction process was less variant.

Documentation variation There was variation between Mayo and TMC in expressing SBI and WMD in neuroimaging reports. Corpus statistics identified the three

most frequent expressions of negated infarction in neuroimaging reports (Table 4.3). In the TMC reports, "no acute territorial infarction" is a common phrase to describe negated SBI concepts. This expression was never discovered in Mayo reports. When describing the grade measure for WMDs, definitive expressions such as "mild", "moderate" and "severe" were used by Mayo physicians. On the other hand, TMC physicians used more descriptive expressions in describing the grade measure for WMDs. In regards to documentation styles, TMC used a template-based reporting method whereas Mayo did not adopt any reporting schemas. The average numbers of tokens per document on Mayo and TMC reports were 217 and 368, respectively. The corpus similarity between TMC and Mayo Clinic radiology reports was 0.82 and suggested a potential moderate-to-high semantic similarity. Overall, Mayo's reports are definitive and varied, whereas TMC reports are lengthy, standardized and descriptive.

Table 4.3: Example of Language Variation between Two Data Sources

Mayo – Non-SBI	**TMC – Non-SBI**
No restricted diffusion.	There is no acute territorial infarct.
No focal masses, focal atrophy, or foci of restricted water diffusion.	No acute territorial infarct.
No evidence for acute ischemia on the diffusion weighted images.	There is no decreased diffusion to indicate an acute infarct.
Mayo – WMD	**TMC – WMD**
Mild leukoaraiosis	There are scattered foci of hypodensity in the subcortical and periventricular white matter, a non-specific finding but likely reflecting the sequela of chronic microangiopathy
Minimal leukoaraiosis	Areas of white matter hypodensity are a non-specific finding but may represent the sequela of chronic microangiopathy
Moderate leukoaraiosis	There are multiple foci of t2 flair hyperintensity in the periventricular, deep and subcortical white matter, a non-specific finding but likely reflecting the sequela of chronic microangiopathy

Process variation The process map of the ESPRESSO data abstraction is illustrated in Figure. 4.2 – Part I. The map provides an overview of the relationship and interaction between people and technology in the context of the data abstraction process. The analysis suggested that the variations of EHR systems and technology infrastructures between the two sites have resulted in differences in the number of processing steps, experts, and duration (Figure. 4.2 – Part II).

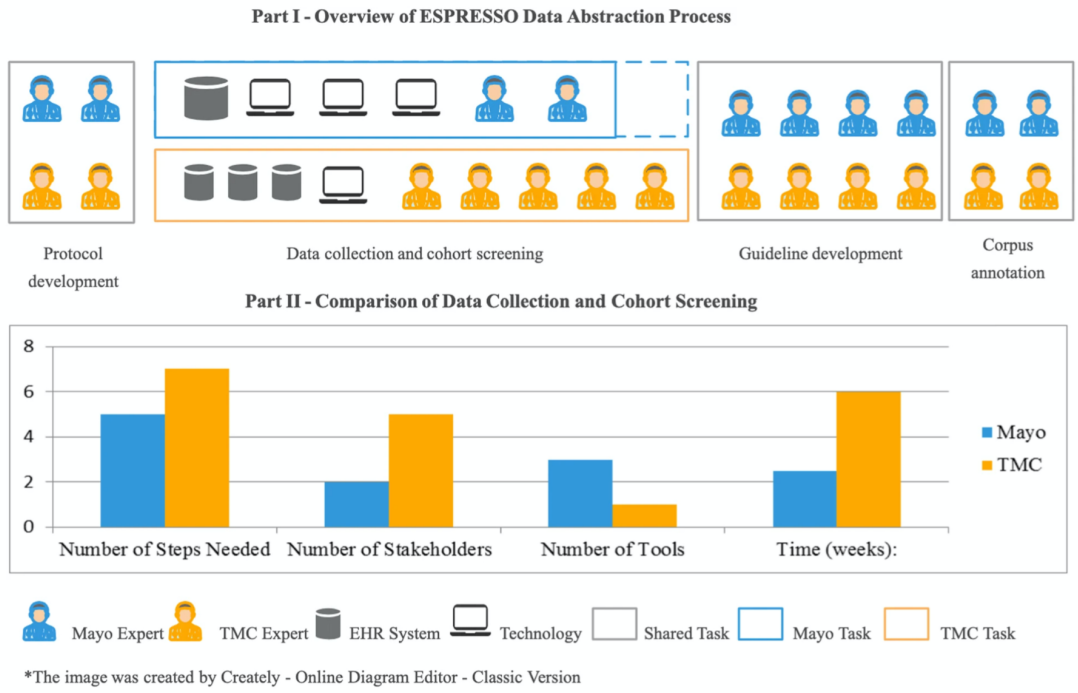


Figure 4.2: Overview of ESPRESSO Data Abstraction Process. Total Annotation Issues during Two Iterations

4.4 Discussion

We conducted a multi-site EHR-based case study in the implementation of the ESPRESSO project to assess the impact of EHR heterogeneity for clinical research. The case study discovered significant variation regarding patient population, neuroimaging reporting, EHR systems, and abstraction processes. Despite the variation, the evaluation of the final corpus yielded high reliability and validity.

The assessment through the ESPRESSO discovered a high variation in the reported prevalence of SBIs between Mayo and TMC. There are two potential reasons for the low prevalence of SBIs in TMC. First, the two locations have different patient sociodemographic characteristics at the two locations. Although both Mayo and TMC are referral centers, Mayo may have a larger proportion of patients who are referred from distant locations whereas TMC may have predominantly local and regional referrals. Second, low SBI prevalence may be due to the different documentation priorities during the routine practice. For further investigation, a qualitative assessment was utilized to learn how clinicians report neuroimaging interpretations. Based on the analysis of cohort characteristics between Mayo and TMC (Table 3) and the post abstraction interview, we discovered a portion of SBIs were under-documented by TMC neuroradiologists due to their historical perceptions of potentially low clinical significance for SBIs. For example, the descriptions about the clinical utility of reporting on small and presumably asymptomatic brain lesions that could represent infarcts were very uncertain. Compared with TMC, the wording describing SBIs on the Mayo reports was more definitive.

Although the average kappa score on the Mayo reports was lower than the TMC reports, the score still reflected an exceptionally high agreement between all annotators. We believe this was achieved by a well-designed process. During guideline development, we found that variation could be reduced by adding an instruction manual to the guidelines. Due to the large number of reports that were assigned to each resident, the de-identified reports were equally distributed to individuals as a "take home" assignment. The instruction manual helped to guide annotation activities, such as suggesting the number of reports that needed to be annotated per day. One of the most commonly raised issues was the lack of precise modifier definitions for WMD. To reduce the abstraction variation caused by different interpretation of modifiers, we created a normalization mapping schema. For example, the level of grading for WMDs was explicitly defined to be mild, mild/moderate, moderate, moderate/severe, and severe.

The qualitative assessment of the annotation process (Figure. 1 - process 5 - box 2) identified that medical concepts (i.e. mention of SBI and WMD) and modifiers (i.e. acuity and location) were the primary issues during the first iteration of annotation. Additional training was offered to address the primary issues experienced during the first iteration of annotation and thus, decreased the occurrence of issues during the second

iteration (Fig. 3). All four annotators noted that with the combination of training and comprehensive annotation guidelines, annotation time was shortened, effort redundancy was reduced, and annotation consistency was improved.

4.5 Limitations and future work

Since the study was conducted on two sites with one case scenario, the generalizability of the process is limited by the scope of the study. Our next step is to expand our investigation on pragmatic clinical trials by incorporating more sites and case scenarios. Furthermore, we plan to develop a standardized process framework for EHR-based clinical research to ensure the validity, reliability, reproducibility and transparency of research findings.

Chapter 5

Implications of Latent Data Quality to Reproducibility

5.1 Overview

The various heterogeneous factors identified in Chapter 3 and 4 have both direct and indirect impacts on the quality of data and implementation. To better understand the downstream impact of these issues to reproducibility and outcomes of clinical research, we conducted an investigation of real-world clinical research of aging under the Rochester Epidemiology Project (REP). We first summarized various heterogeneous factors into four dimensions of information quality (IQ): Intrinsic IQ, Accessibility IQ, Representational IQ, and Contextual IQ. The implication of IQ caused by the heterogeneous EHR environment and variations in the process of information collection and use to reproducibility was examined using both real-world evidence and a simulation study.

5.2 Materials and Methods

Assessment Methods To investigate heterogeneity involved in the process for the secondary use of EHRs and its implications for reproducibility, we formulated three components: 1) a conceptual process of information collection, extraction, organization, and representation, 2) information quality metrics for quantifying process feasibility and clinical outcome variability, and 3) a downstream implication through a case simulation.

TRUST Process The TRUST process stands for clinical **T**ext **R**etrieval and **U**se towards **S**cientific rigor and **T**ransparent. To define the TRUST process, we adapted an IQ life cycle model as a conceptual representation of IQ through a sequence of processes [113]. The information quality models [114, 113], based on the definition of "fitness for use", as a middle layer conceptual representation to help organize and define the context. In the context of secondary use of EHRs, we view information as the subject of heterogeneous network interactions between human and non-human actors [115]. The barriers to achieving reproducibility were modeled as the gaps between the ideal state of multi-level interaction and imperfect pragmatics (i.e., implementation quality). This view helps to capture the dynamic interactions between users (e.g., data abstractor, informatician) and information in a sequential order, providing additional context useful for studying reproducibility. Four stages were considered in the TRUST process: planning, information generation, information use, and information dissemination. During information generation, four sub-processes were identified: documentation (primary use), collection, annotation, extraction, representation, and organization (Figure 5.1). Although the study focuses on the secondary use, we included the stage of information documentation during patient care as quality issues have already occurred.

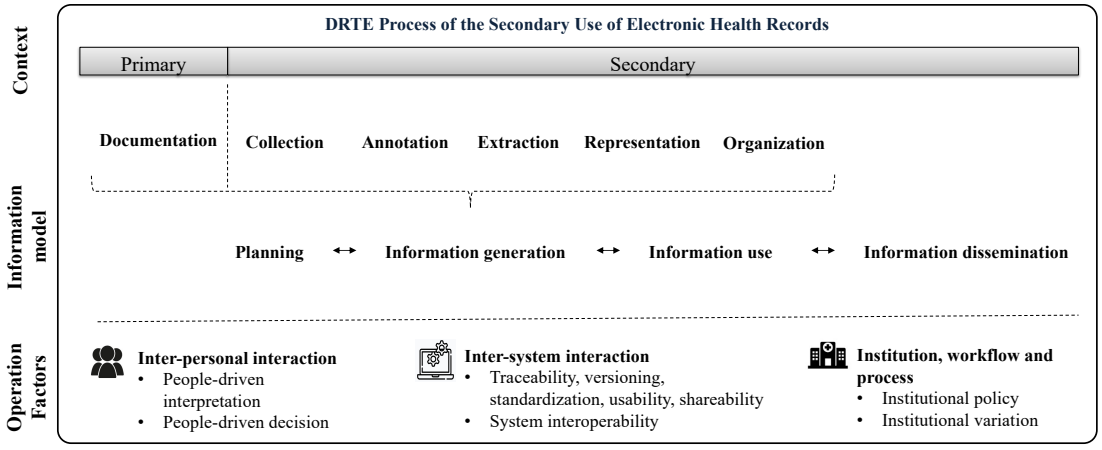


Figure 5.1: TRUST Process in the Context of EHR)

IQ Assessment The TRUST process defines the scope of reproducibility measurements. To assess the technical soundness of the process, there are two metrics to be

measured 1) feasibility of completing the original steps (i.e., implementation feasibility) and 2) variability of results (i.e., clinical outcome variability). To quantify the previously defined metrics, we considered four IQ surrogate measurements related to reproducibility, including intrinsic IQ, accessibility IQ, representational IQ, and contextual IQ [116]. The definition of each measurement is provided in Table 5.1.

Table 5.1: Definition of Information Quality

Measurement	Definition	Primary Implication
Intrinsic IQ	The quality of information is a measurement based on its own right (i.e., information validity).	O
Accessibility IQ	The information is accessible and available.	O, I
Representational IQ	The information is interpretable and interoperable (syntax and semantics), and easy to manipulate.	O, I
Contextual IQ	The information is measured within the context of the task at hand. (i.e., completeness, timeliness)	O

*O: clinical outcome variability, I: implementation feasibility

IQ Implication to Reproducibility Model simulation can be applied to illustrate the downstream impact of IQ on reproducibility. This simulation can create synthetic medical record data based on detailed analyses of a real observational database. As suggested by Hum et al, the agent-based simulation (model) can be used to model the complex interactions between patient and provider through simulating individual characteristics of providers [29]. It is important to note that the model is not a perfect representation of real world scenarios but exaggerates certain aspects for the given condition. In our study, the model was used to simulate the implication of poor accessibility IQ to clinical outcome variability.

5.3 Case Study

Study Setting The study was approved by the Mayo Clinic Institutional Re- view Board (IRB) and the Olmsted Medical Center IRB. The study population consisted of participants of the Mayo Clinic Biobank [117]. The Mayo Clinic Biobank is an institutional resource comprised of volunteers who have donated biological specimens, provided risk factor data, and have given permission to access clinical data from their EHRs for clinical research studies. Participants were contacted as part of a pre-scheduled medical examination at Mayo Clinic sites between April 2009 and September 2015. All participants were 18 years or older at the time of consent. Approximately 57,000 participants have been enrolled, of which 24,224 were 65 years of age or older at the time of consent.

Disease The case study was conducted with an objective to ascertain delirium status from electronic health records. Based on the definition of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) and in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), delirium is a syndrome with symptoms of acute onset, cognitive impairment, fluctuating course, attentional and awareness deficits, and psychomotor and circadian changes [118]. Delirium is underreported and not every patient has a formal assessment for delirium diagnosis [119]. Most patients present with encephalopathy, confusion, and alternation of mental status as the main symptom. Diagnosing delirium is typically based on a combination of mental status assessment, physical, and neurological exams. The Confusion Assessment Method (CAM) is the most widely used bedside clinical assessment tool for the diagnosis of delirium.

Due to the variability of diagnostic methods (i.e., no singular, conclusive diagnostic test) in the clinical setting, the documentation patterns of delirium-related findings can be variable. As suggested by prior studies, the following definition was applied to determine patients' delirium status (Table 5.2): the presence of International Classification of Diseases (ICD) codes for delirium, the presence of a nursing flowsheet documentation of their assessment of delirium, and the presence of CAM definition based on information extracted from clinical notes (CAM-NLP) [120].

Table 5.2: Definition of EHR-derived Measures for Ascertaining Delirium Status

EHR-derived measures	**Definition**
CAM-NLP	Definitive: number of unique CAM criteria >= 3
	Possible: 2 <= number of unique CAM criteria < 3
ICD	Delirium ICD-9: 290.11, 290.3, 290.41, 291.0, 292.81, 293.0, 293.1, 293.89, 293.9, 300.11, 437
	Delirium ICD-10: F05, R41, F10.231, F10.921
	Encephalopathy ICD-9: 348.30
	Encephalopathy ICD-10: G93.40, G93.41, G93.49, G92, G94, G31.2
Flowsheet	CAM-ICU, B-CAM

Appreciation: CAM-ICU: Confusion Assessment Method for the Intensive Care Unit, B-CAM: modified CAM-ICU for non-critically ill patients

Implementation Workflow To better understand the implication of IQ in reproducibility with respect to secondary use of EHR data, we describe the pragmatic implementation process aiming to capture the dynamic interactions between user, system, and information. Post thematic analysis was applied to understand multi-dimensional user-system–information interactions. The TRUST process of ascertaining delirium status from EHRs can be summarized into the following: information collection, information extraction, and information organization and representation. Information collection involves retrospectively retrieving various data topics through either human-assisted manual data abstraction or an automatic extract, transform and load (ETL) process. Before the data transformation process began, we utilized an i2b2 (Informatics for Integrating Biology and the Bedside) data warehouse to help obtain patient's demographic status. i2b2 is an interactive informatics platform that is widely used for patient cohort identification [121]. The system is based on an internal Datamart of i2b2 Ontology for data standardization. For privacy and security purposes, information accessibility was limited to summary statistics and patient identifiers. Additional data including appointment (admit, discharge, and transfer), diagnosis, flowsheet, and clinical notes were automatically retrieved through customized Structured Query Language (SQL) from the Mayo Unified Data Platform (UDP). UDP is an enterprise data warehouse that loads

data directly from Mayo EHRs. Information extraction is a sub-task of natural language processing (NLP) aiming to automatically extract structured information from unstructured text. We applied our previously developed and validated NLP system to extract CAM-related features from patient clinical notes [120]. For information representation and organization, we used patient id, encounter id, and encounter date to link patients across EHRs. All data was normalized from visit level to patient level.

IQ Assessment Intrinsic IQ was assessed using agreements on case and non- case ascertainment between three EHR-derived measures (CAM, ICD, and Flowsheet) using unweighted Cohen's kappa, sensitivity, specificity, and f1-score. The accessibility IQ was evaluated through analyzing the information accessibility and shareability (system and method) based on two settings: intra- institution (i.e., study occurs in the same organization) and inter-institution (i.e., multi-site collaboration). The evaluation process was done independently by two informaticians and adjudicated by a third informatics researcher. We defined four- levels of accessibility: direct access (level 1), adaptive access (level 2), partial access (level 3), and no access (level 4) as follows:

- Level 1: information resources can be directly shared and used with no information loss.
- Level 2: information resources can be directly shared; site-specific adaptation needed for information use.
- Level 3: information resources cannot be shared without usage agreements or de-identification; site-specific adaptation needed for information use
- Level 4: information resources cannot be shared and re-used.

To illustrate the level of information loss caused by representational IQ, we retrospectively analyzed the inconsistency in data representation and identified the presence of information loss during the TRUST process. Lastly, as suggested by Weiskopf et al, information completeness can be measured in the following four dimensions: documentation, breadth, density, and predictive [44]. Due to the underdiagnosed nature of delirium in clinical practice, we focused on documentation completeness ratio to assess contextual IQ, defined as a record containing all observations made about a patient. Patients with positive delirium status should contain all enough CAM features for satisfying the

diagnosis definition. We calculated the CAM missing rate for positive delirium patients diagnosed by ICD and examine the documentation pattern of delirium-related findings in ICU and non-ICU settings.

IQ Implication to Reproducibility To illustrate the implication of IQ issues caused by the imperfect interactions between users and information based on the situated scenarios, we conducted a simulation to show the effect of the performance variability in case ascertainment of delirium. The one common scenario in clinical research is the systematic bias or measurement error caused by imperfect EHR-derived phenotypes [19]. Thus, we focus on simulating the effects of poor accessibility IQ to clinical outcome variability. The hold-out method was applied by providing individual EHR-derived measures at a time and observing the associated outcomes. Logistic regression was used to model for each outcome variable while adjusting for age and sex. The Odds Ratios (OR) are reported for these models. We compared the model using true disease status with the model using simulated misclassified outcomes caused by the latent effects

5.4 Results

Intrinsic IQ The agreement between ICD, flowsheet, and CAM-NLP are provided in Table 5.3. The agreement between ICD and CAM-NLP was moderate-high (k = 0.61). Agreement between ICD and flowsheet was moderate (k = 0.41). Similarly, the agreement between CAM-NLP and flowsheet was moderate (k = 0.42). Although CAM-NLP yielded the highest agreement, no single data type comprehensively represented the delirium status. There was a strong indication that a data quality assessment should be conducted prior to the information use.

Table 5.3: Agreements between ICD, flowsheet and NLP

EHR Data	Sen	Spe	F1	kappa
ICD – CAM-NLP	0.56	0.97	0.67	0.61
ICD - Flowsheet	0.54	0.91	0.47	0.41
CAM-NLP - Flowsheet	0.72	0.86	0.50	0.42

Abbreviation: Sen: sensitivity, Spe: specificity

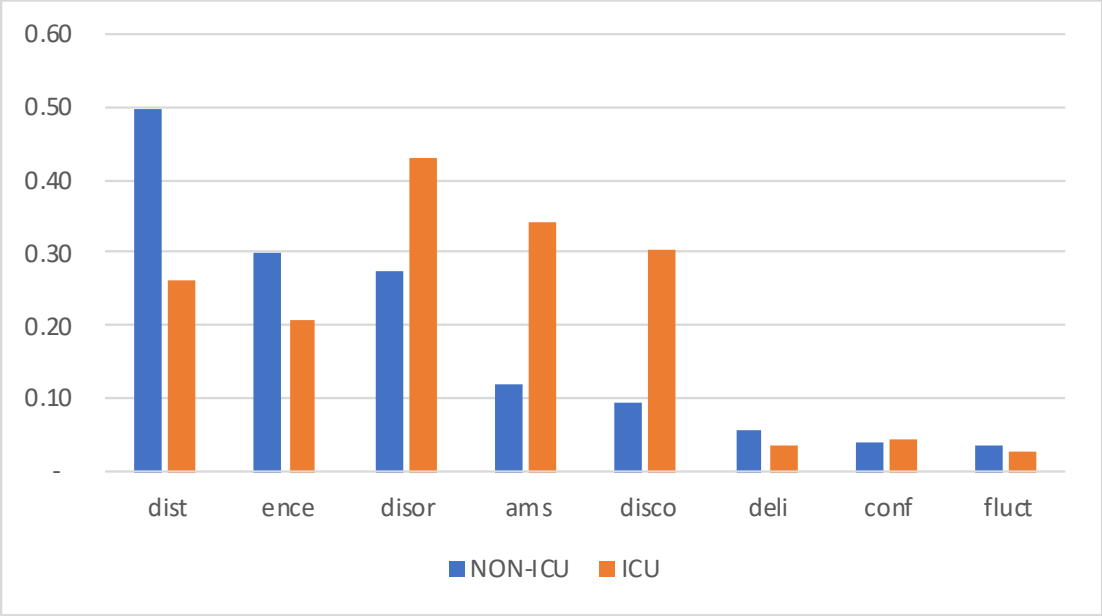
Accessibility IQ The workflow analysis indicated a variant level of accessibility IQ

in both intra-institutional and inter-institutional settings (Figure 5.2). We observed that user-level resources (i.e., information resources generated by users) had better accessibility under the inter-institutional setting. Site-specific ETL infrastructures may help to promote internal data accessibility. For example, the integration of multiple EHRs from Mayo Clinic Health Systems allows direct information access (also see Figure 3). On the other hand, there were greater issues with information and system-level accessibility under the inter-institutional setting due to privacy and regulatory issues.

		Intra	Inter
I	ICD	1	2
I	Clinical note	1	3
I	Flowsheet	1	3
U	Data analytics script	1	3
U	SQL script	1	2
U	Data linkage script	1	3
S	Screening tool	1	3
S	ETL infrastructure	1	3
S	NLP system	2	3

Figure 5.2: Level of Accessibility IQ in Two Settings

Contextual IQ The assessment indicated a high level of variation in the documentation and reporting of CAM feature documentation for delirium patients in ICU and non-ICU settings. We found the concepts of disorganized thinking, encephalopathy, and delirium are more likely to be documented in the non-ICU environment. Alternatively, the concepts of disoriented, altered mental status and disconnected yielded a much higher documentation completeness ratio (Figure 5.3).



Blue bar: Non-ICD setting, orange bar: ICU setting, x-axis: documentation completeness index: the higher index indicates higher completeness; abbreviation: dist: disorganized thinking, ence: encephalopathy, disor: disoriented, ams: altered mental status, disco: disconnected, deli: delirium, conf: confusion, fluct: fluctuation

Figure 5.3: Information Completeness of CAM Feature Documentation in ICU and Non-ICU Settings

Representational IQ Figure 5.4 shows two types of ETL processes (structured and unstructured data) based on the case study implementation. The ETL process for structured data involves information transformation from a non-relational database to a relational database, a direct copy of said information to create a duplicated research datamart, and a complete ETL process. We identified that the amount of information loss is lower if the information transformation happens within two databases that were developed by the same company and share the same data standards such as Chronicles and EHR Clarity. During the direct information copy from Epic Clarity to the MS SQL server and UDP integration layer to i2b2 datamart, we observed a greater information loss due to the incomplete view of both data syntax and semantic standards. For unstructured data, we observed information loss due to the syntactic structure of

a document (e.g., tabular data) when converting RTF to plain text. Similarly, not all information was extracted during information transformation for structured metadata elements (e.g., patient id, provider id, doc id).

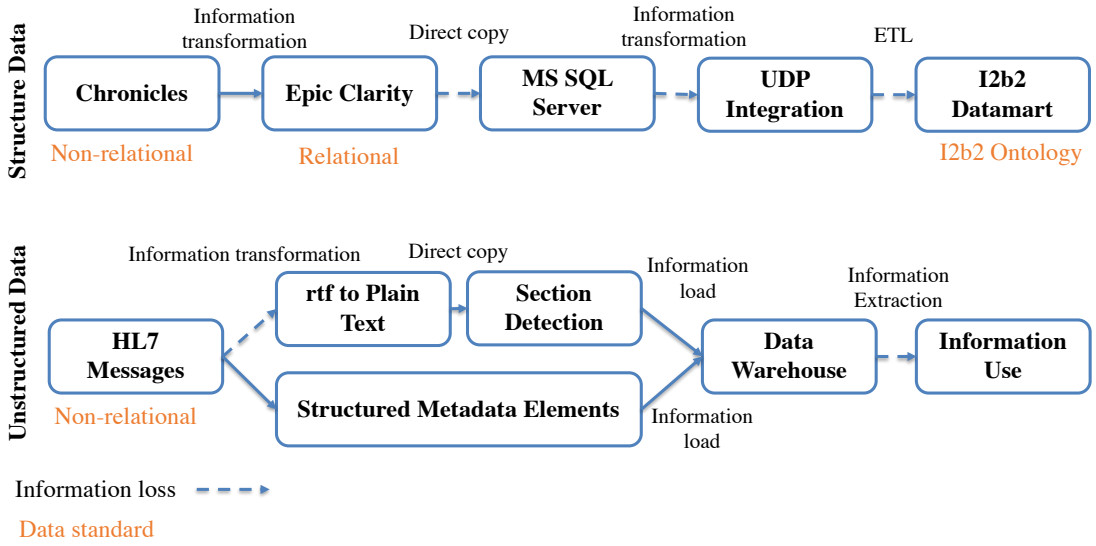
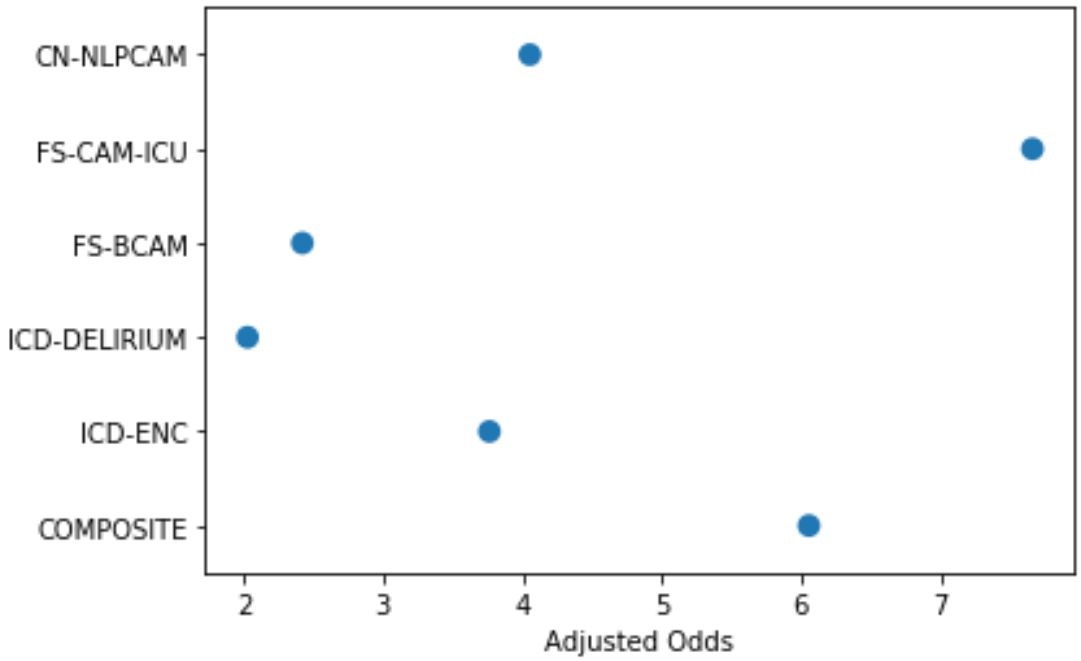


Figure 5.4: ETL processes for Structured and Unstructured Data

Accessibility IQ Implication for Reproducibility The simulation of five different information sources on the outcome of delirium patients indicated high variability in the estimated odds ratios. As illustrated in Figure 5.5, directly applying any single information resource may not accurately reflect the true disease status. When inaccurate or incomplete information sources are used as the gold standard for downstream applications, bias, errors, or misclassification can occur. The experiment demonstrated that IQ has a significant effect on reproducibility.



*The presented result is not to provide any clinical indications but rather demonstrate the variability caused by information quality

Figure 5.5: Odds Ratio for All-cause Mortality at Discharge for Delirium Cohorts with Simulated EHR-derived measures

5.5 Discussion

Based on the IQ assessment and case study implementation, we discovered various barriers to reproducibility, such as inconsistent information documentation patterns across settings, information loss during ETL processes, and variable levels of information resource accessibility. One key recommendation for the informatics community is to place a higher value on the information resources prior to the information use. Due to the heightened downstream impact of EHRs, information and implementation quality carry an equivalent significance with downstream applications (e.g., machine learning models). We thus believe it is important to promote transparent, high-throughput and accessible data infrastructures and implementation best practices (e.g., data quality assessment,

reporting standard) aiming for process standardization. Our study was limited due to only involving a single secondary EHR use application as a case study, and as such the generalizability of our findings (e.g., ETL process, barriers to reproducibility) is limited by the scope of the study.

5.6 Conclusion

Reproducibility is crucial for the secondary use of EHRs. We applied a multi-phase method to investigate heterogeneity in the processes involved the secondary use of EHRs and its implications for reproducibility. We discovered that four types of IQ measurements suggested that barriers to reproducibility occurred for all stages of secondary use of EHR data

Chapter 6

Methodological Standard and Best Practices of Reproducibility

6.1 Overview

This chapter aims to address various data quality issues that occur during the implementation and execution of the TRUST process. We focus on gathering best practices of implementation strategies from existing literature and real-world case studies based on chapters 3, 4 and 5. A process framework was proposed to capture important steps needed for conducting TRUST.

6.2 Background and Related Work

Manual chart review is a common way for researchers to investigate the disease causal path and to predict health outcomes [122]. It has been widely applied to clinical research, epidemiology, and quality assessment [10, 12, 123, 124]. Despite its popularity, manual chart review has been criticized for lack of research reproducibility [12], data reliability and validity [125], and scalability [14]. First, clinical research reproducibility is crucial in the field of medicine so that the findings of a single limited study may be translated to broader patient care [126, 127]. However, it is not easy to reproduce an EHR-based study, due to the heterogeneity of EHR systems, the complexity of the research team, and the difficulty of capturing every essential piece of information needed to reproduce

the study [128, 27]. Second, the validity and reliability of clinical data manually drawn from EHR can be questioned, because the individuals who performed the review may not have a detailed protocol specifying the common conventions and standards for annotations [22, 7]. Studies have reported differences in health care quality measured by manual review [129, 130, 131, 132], and indicate a need for more clinical involvement in the capture and validation. Data reliability in EHR-based research can be measured by the agreement between several data abstractors. Failure to address inter-rater reliability can lead to inconsistent results and unreproducible discoveries [125]. Furthermore, the gap between structured and unstructured clinical research data creates barriers to routinely aggregate and analyze data. As much of the detailed patient information is embedded in clinical narratives, it is a very time- consuming and costly process to extract information from clinical records manually.

NLP-assisted chart review has been seen as a scalable way to extract information. However, it has been shown that NLP algorithms developed in one institution for a study may not perform well when reused in the same institution or deployed to a different institution or for different studies [133, 22, 134, 135]. Mehrabi et al. [43], Liu et al. [136] and Carroll et al. [137] examined the system performance and discovered that customization was essential to achieve desirable performance when the system is deployed to different institutions. Wagholikar et al. evaluated the performance of an NLP tagging system from two sites. The performance degrades when the tagger was ported to a different hospital [133]. To address the issue, efforts have been made to standardize NLP development by creating annotated lexical resources and normalizing data elements. Savova et al. created a gold standard for anaphoric relations from a cross-institutional corpus [138]. Albright et al. built the first syntactically and semantically annotated clinical corpus [139]. South et al. developed a clinical corpus from manually annotated EHRs to identify phenotypic concepts [140]. To address the heterogeneity of EHR, Sohn et al. assessed the clinical documentation variations across two hospitals [42]. To enhance the transparency of NLP development, Scuba et al. developed an ontology-based web tool [141], and the open-source consortium Open Health Natural Language Processing (OHNLP) was initiated [142].

However, even with sharable corpora, online tools, and open source initiatives, an NLP system may still have degraded performance when the system is reused by others.

One challenge is that NLP-enabled clinical research involves an intangible, iterative, and complex process; the validity and reliability of data abstraction and annotation are often ignored. Lack of data abstraction protocols, heterogeneity of EHRs, and unstandardized practices are common issues during the process [42]. One potential solution is to standardize the process by establishing best practices. As NLP-enabled clinical research studies become increasingly integrated into clinical care, a universal framework is needed to ensure every step of the research is consistent, reliable, and valid. In this paper, we propose a new framework to facilitate automated data abstraction and cohort annotation when conducting single- and multi-site EHR-based clinical research.

6.3 Methods

The development of EHR-based TRUST framework was followed by a multi-phase process (Figure 6.1) The process includes following steps: 1) conceptualization, 2) adoption of external standard, 3), design and prototyping, and 4) expert review.

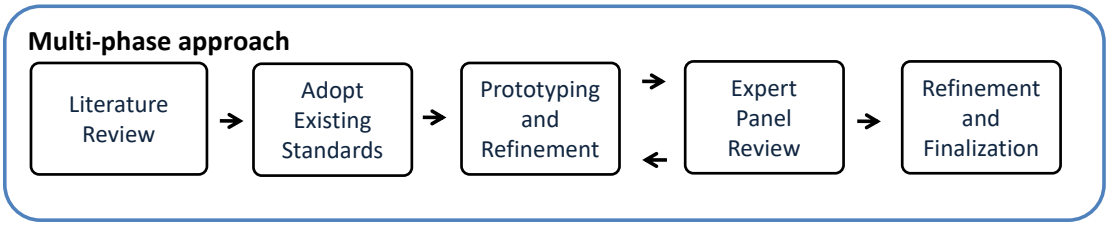


Figure 6.1: A Multi-phase Approach for the Development of TRUST Framework

Conceptualization Based on the findings from Chapter2, we conceptualize the definition of reproducibility in the context of secondary use of EHRs. Instead of pursuing an exact universal definition, we propose to view reproducibility as a generic quality measure with the fundamental meaning of the capability for reiterating the experiment for achieving valid results. We view the target experiment as an ideal state. To satisfy the definition of reproducibility, the new study should be performed within a satisfactory variability and avoid any deviation from the original state. This would require all actors and non-actors within the process to be adherent to the original intent. We introduce the concept of 'process reproducibility' that shares the view of reproducibility is dynamic which can be applied to various of contexts and objects. In the context of EHR, we focus

on activities of replicating the prior process of information generation and information use of EHRs to assess the technical soundness of the process.

Adoption of External Standards A literature review was conducted to identify existing methods and best practices of the TRUST process. A search was conducted, retrieving EHR-based concept extraction articles that were written in English and published from January 2009 through June 2019. Literature databases surveyed included Ovid MEDLINE In-Process and Other Non-Indexed Citations, Ovid MEDLINE, Ovid EMBASE, Scopus, Web of Science, and the ACM Digital Library. The implementations of search patterns were consistent across the different databases. The search query was designed and implemented by an experienced librarian (LJP) as: ("clinic" or "clinical" or "electronic health record" or "electronic health records" or "electronic medical record" or "electronic medical records" or "electronic patient record" or "electronic patient records" or "EHR" or "EMR" or "EPR" or "ATR") AND ("information extraction" OR "concept extraction" OR "named entity extraction" OR "named entity recognition" OR "text mining" OR "natural language processing") AND (NOT "information retrieval"). A total of 10,441 articles were retrieved from five libraries, of which 6,686 articles were found to be unique. The articles were then filtered manually based on the title, abstract, and method sections to keep articles with EHR-based clinical information extraction from English text. After this screening process, 928 articles were considered for subsequent categorization. We conducted an additional manual review to keep articles with methodology descriptions focusing on clinical concept extraction. The final inclusion criteria for the target papers are as follows 1) using concept extraction methods, 2) applied to EHR data in English, 3) providing a methodological contribution via: a) presenting novel methods for clinical concept extraction, including introducing a new model, data processing framework, NLP pipeline, etc., or b) applying existing methods to a new domain or task. Articles without full text or methodology descriptions were excluded. Following this screening process, 228 articles were selected and categorized based on the methods used. A comprehensive full-text review of all 228 studies was performed by the study team. In addition to the retrieved literature, the following standards were also considered: Corpus Annotation Schemes [17]; Fundamentals of clinical trials [143]; and Research data management [144].

Design and Prototyping To prototype the framework, we considered the pragmatic implementation process, best practices, and lessons learned from Chapters 3, 4, and 5 as the real world evidence and help improve the design of the framework. Specifically, we focused on the implementation strategies and best practices learned from the case studies.

- Identification of Silent Brain Infarction Events The ESPRESSO (Effectiveness of Stroke Prevention in Silent Stroke) study is an EHR-based study aiming to estimate the comparative effectiveness of preventive therapies on the risk of future stroke and dementia in patients with incidentally-discovered brain infraction. SBIs are commonly detected as incidental findings in patients without clinical manifestations of stroke via neuroimaging. Descriptions of these events are frequently documented in radiology reports as text, rendering NLP an ideal tool to assist in the identification of SBI cases. The TRUST process was initially implemented at Mayo Clinic and Tuft Medical Center and later replicated at Kaiser Permanente Southern California (KPSC) health system. The process involved five different EHR systems, across an interdisciplinary team of neurologist, radiologists, informaticians, statisticians, study coordinators, students and residents.

- Characterizing Chronic Pain Episodes in Clinical Text Clinical text contains rich information about chronic pain, but no systematic appraisal has been performed to assess the electronic health record (EHR) narratives for these patients. We applied the TRUST process to characterize individual episodes of chronic pain and analyze EHR notes for a stratified cohort of adults with known chronic pain. An iterative consensus development and an episode-centered approach were applied to annotating chronic pain based on input from clinical domain experts.

Post-implementation interviews were conducted to retrospectively collect user-level feedback. An interview protocol was created to focus on three main areas: 1) abstraction process, 2) human factors, and 3) tooling. Four back-to-back interviews were conducted with the four abstractors following the guidelines of Contextual Interview (CI) suggested by Rapid Contextual Design. Each interview was conducted by an informatician and lasted approximately 30 min. Questions and issues raised by each annotator during the two iterations of annotation were collected and qualitatively assessed.

Expert Review The framework was reviewed by three domain experts including two biomedical informatics faculties and one epidemiology faculty at Mayo Clinic. A Delphi consensus technique was utilized during the review. The initial framework was independently and blindly reviewed by each reviewer. A post-review interview was organized to understand the feedback and suggestions. Items were further refined after the discussion. The process continued until a consensus was reached.

6.4 Current Practice of Clinical Corpus Annotation

Based on our review, we determined that the typical tasks involved in this process included annotation task formulation, data collection, and cohort screening, annotation guideline development, annotation training, annotation production, and corpus adjudication [145, 146]. Formulation of a task in a practice setting involved definition of points of interest, execution of a literature review, consultation with domain experts, and identification of study stakeholders such as abstractors and annotators with specialized knowledge. This step was then followed by definition and/or creation of a study population or cohort. For example, a concept extraction application of geriatric syndromes was developed based on the cohort of 18,341 people who were 1) 65 years or older, 2) received health insurance coverage between 2011 and 2013 and 3) were enrolled in a regional Medicare Advantage Health Maintenance Organization [147]. Once the cohort definition was finalized, data was screened and retrieved. Studies have found the usefulness of leveraging open-source informatics technologies such as i2b2 or customized application programming interface and SQL queries for automatic screening and retrieval [121]. Subsequent to dataset definition and creation, the development of a detailed annotation guideline specifying the common conventions and standards was necessary, ensuring that definitions created are scientifically valid and robust. Notably, this step involves prototyping a baseline guideline, performing trial annotation runs, calculating inter-annotator agreement (IAA), and consensus building.

Based on the analysis of 18 articles with exclusive discussions of the gold standard development process, 4 (22%) studies used a single annotator, 10 (56%) studies reported of using two annotators and one adjudicator, 3 (17%) studies reported of using multiple

annotators but did not specify the number, and 1 (6%) study with no mention of annotator. The median, minimal and maximal number of annotated clinical documents were 251, 100 and 8,321 respectively. Most studies choose 200 to 600 documents as the study data size. Among these, 30% to 60% were randomly sampled for double annotation and IAA assessment. Process iteration was used to save the annotation cost and increase effectiveness. For example, Mayer et al. reported having two annotators perform the initial annotation on 15 documents for training and consensus development. During the second iteration, another 30 new documents were applied and IAA was calculated. The process was repeated until the IAA reached to 0.85 [148].

6.5 Implementation Steps of the TRUST Process

The TRUST Process summarizes the linear process of extracting or reviewing information from EHRs and assembling a data set for various research needs. The processes consider important action items and documentation checklists to identify, evaluate and mitigate variations across sites. Depending on the study design, the order of processes and selection of activities can be altered.

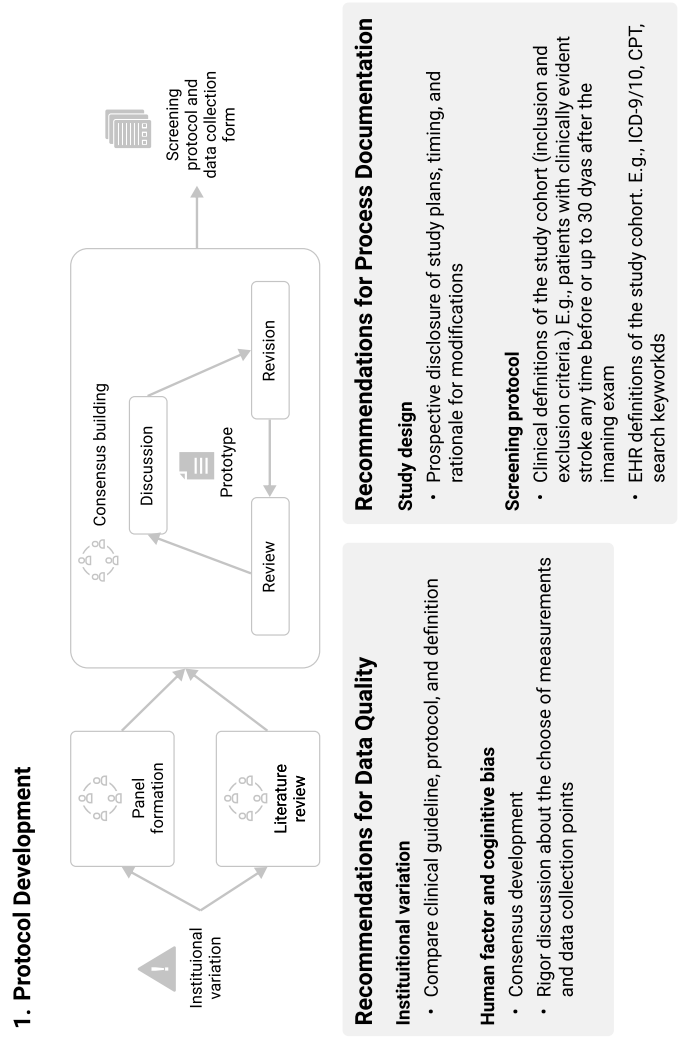


Figure 6.2: TRUST process - Protocol Development

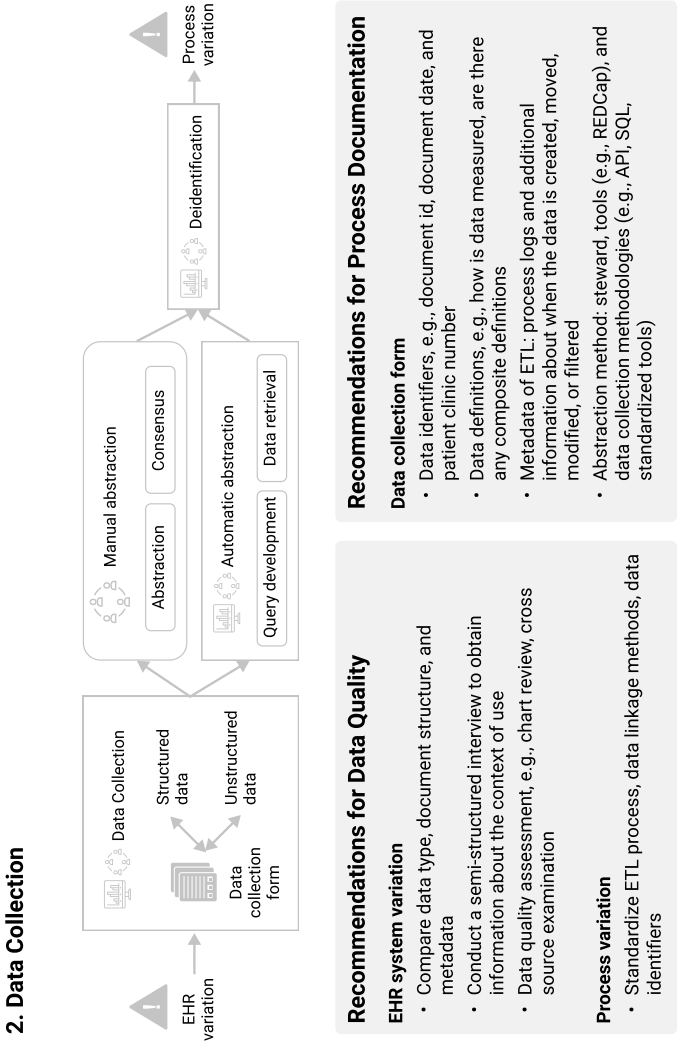


Figure 6.3: TRUST process - Data Collection

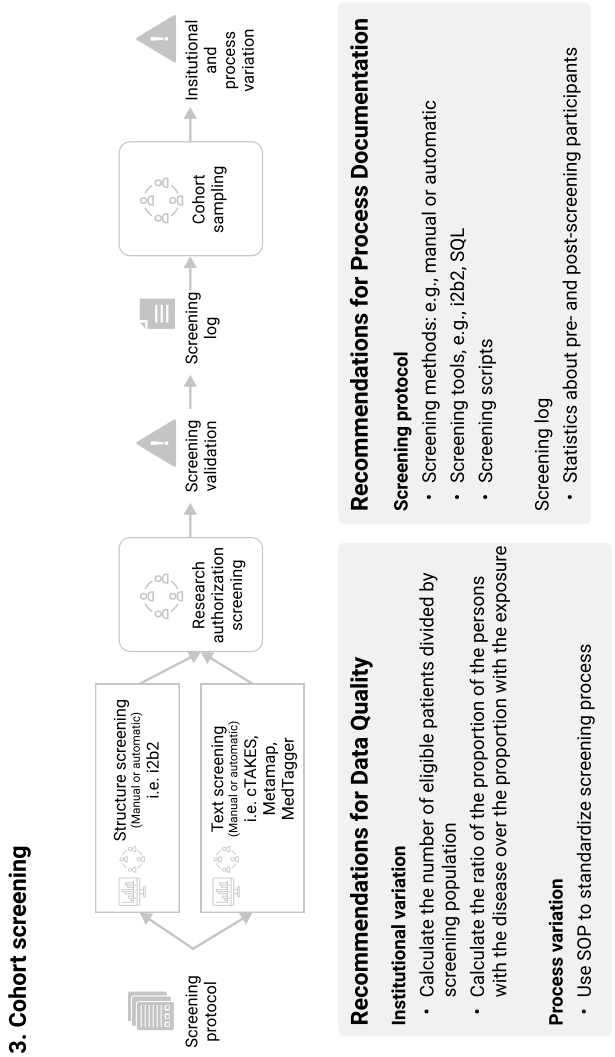


Figure 6.4: TRUST process - Cohort Screening

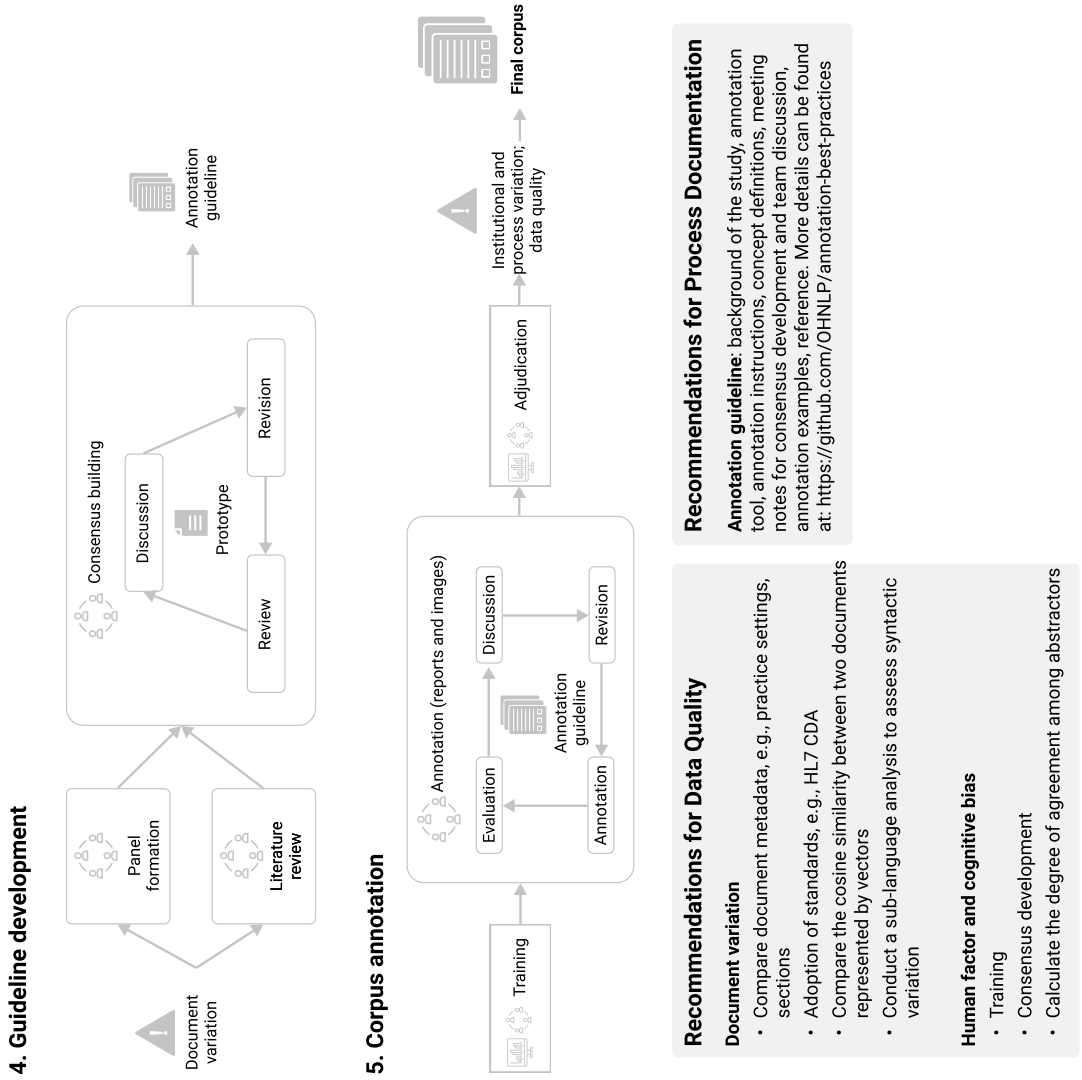


Figure 6.5: TRUST process - Guideline Development and Corpus Annotation

6.6 Recommended Implementation Best Practices

Institutional variation It is inevitable to encounter variabilities across different institutions. Being aware of the degree of variation can help estimate biases and prevent inaccurate study conclusions. Thus, it is always helpful to apply informatics techniques

to capture and assess the variation to ensure transparent and informed EHR-based clinical research.

Documentation plan and checklist A comprehensive documentation plan for a study allows interventions aimed at process replication and error prevention to be designed into the data abstraction. The plan should explicitly mention what, where, and when to document experimental elements such as protocols, guidelines, codes, operations manuals, and process workflows. Ensuring adequate time is devoted to documentation is critical in order to prevent details from being overlooked or omitted. A documentation checklist ensures important study details are documented. Examples of important metadata elements are data identifier (i.e. document id, document date, and patient clinic number), cohort definition (i.e. inclusion and exclusion criteria), steward, and description of the data (when the data is created, moved, modified, or filtered). During data abstraction, process logs, tools, data definitions, and methodologies need to be carefully analyzed and explicitly stated.

Concept definitions and protocol co-development To ensure data validity, the variables of the study should be strictly defined. Standardized terminology codes, such as ICD, SNOMED-CT, CPT-4, or RxNorm are useful for describing observable medical characteristics. Protocol co-development and consensus building helped reduce institutional and process variance in our study (Figs. 1 and 3). Particularly, having a well-represented expert panel (from all sites) for developing and evaluating inclusion and exclusion criteria and annotation guidelines helped the creation of high-quality protocol documents.

Annotation study design Determining the appropriate number of annotators and the size of corpus for annotation is critical and often challenged by the resources available. In general, the process requires at least two annotators with (clinical) domain expertise to independently perform the annotation [48], [49], [50]. Having only one annotator in the study is not recommended since data validity and reliability cannot be measured and ensured [50]. For multi-site studies, the process requires at least two annotators from both sites in order to help estimate inter-institutional and intra-institutional variation [51].

Abstraction and annotation training Proper training and education can help reduce process inconsistency and increase transparency, especially for a cross disciplinary

team. When the training sessions were applied, a shared understanding of rigorous experimental design, research standards, and objective evaluation of data was ensured. The training materials generated included an initial annotation guideline walk-through; demos and instructions on how to download, install, and use the annotation software; case studies; and practice annotations. Annotation production is typically organized into several iterations with a significant amount of overlap in the data annotated by each individual annotator to ensure the ability to determine IAA. Finally, in cases where annotators disagree, conflicts were adjudicated by subject matter experts. All the issues encountered during the gold standard creation process were documented. Some example training activities included discussing the overall study goal, going through the contents of the annotation guideline and definitions of interest, and practicing using the annotation tool (i.e. allowing people to work on a sample of 5–10 notes).

Process iteration and consensus building A consensus reaching process is an iterative and dynamic process for building agreement on any potential issues and disagreements. A consensus meeting should be organized when developing screening protocols and annotation guidelines. Routine discussions ensure guidelines and protocols are scientifically valid and robust.

Adoption of appropriate informatics tools Successfully leveraging informatics techniques can improve process efficiency, data quality, and reproducibility. For example, automatic data retrieval techniques (such as application programming interface and structured query language) and cohort screening tools (such as i2b2 [54]) can enable a high-throughput data abstraction process. Using annotation tools ensures a standardized and reproducible annotation process. It is more important to choose an appropriate informatics solution than an advanced solution. In the study, we chose a light and standalone version of annotation software over an advanced web-based tool due to its high feasibility and efficiency. In situations that require extensive validation for processes, such as de-identification, human validations are needed after applying the informatics tools.

Chapter 7

Reporting Standards and Research Metadata of Reproducibility

7.1 Overview

This chapter aims to address issues related to reporting practices and standards that occur during the information dissemination of the TRUST process. We focus on gathering new reporting standards and research metadata of reproducibility from high-quality research studies. Based on the STROBE Statement, the newly discovered standards were provided as extended recommendations for EHR-based studies.

7.2 Related Work

Evidence has suggested that the methodologies of chart review lack standardization, scientific rigor, and reporting guidelines. For example, a review conducted by Gilbert et al. on three emergency medicine journals discovered that among all studies related to retrospective chart review, only 11% reported the use of an abstraction form and 4% reported inter-rater agreement [12]. Meanwhile, informatics tools have been developed to enhance the use of EHR data for supporting clinical and translational research. For example, Informatics for Integrating Biology and the Bedside (I2B2) is a patient privacy-preserving query tool for facilitating research feasibility assessment [149]. Once study

feasibility is determined, data capture tools such as Redcap, TELEforms® (Cardiff Software, Inc., Vista, CA), and Studytrax® (ScienceTRAX LLC, Macon, GA, USA) aim to facilitate patient information collection. Protocol management aids such as Protocol Builder® (Biomedical Research Alliance of New York) and questionnaire development tools like QDS™(NOVA Research Company, Bethesda, USA) ensure the document development process is more efficient. Additionally, many informatics methodologies, such as natural language processing (NLP), have been leveraged to perform chart review by automatically extracting clinical concepts from unstructured EHR data. Various types of EHR-based phenotype algorithms have been developed, ranging from drug-related adverse events [150] to individualized risk prediction [16]. However, even with the great advancement of clinical research informatics, there is a lack of systematic understanding of how the tools and methods are utilized and reported, as well as the impact on overall research quality. Our empirical analysis on clinical research ontologies and reporting standards found little-to-no informatics-related standards. This often manifests as the withholding of key methodological details such as data abstraction methods, protocols, processes, and definitions11. Several pragmatic evaluations of high-profile clinical journals have shown that only 11% to 25% of projects can be replicated12-14. The issues often appear as the inability to reproduce research data. Mobley et al. surveyed faculty and trainees at MD Anderson Cancer Center, and discovered that 50% of respondents had experienced issues with data reproducibility in cancer-related research [151]. The consequences of invalid methodologic processes and unreproducible results in biomedical research can be serious, such as preventing clinical knowledge translation, wasting scientific resources, and delaying treatment timing [127].

To ensure valid, transparent, and reproducible clinical research, a growing number of informatics related efforts have been reported. Several clinical research ontologies have been developed to ensure scientific standards, including Ontology of Clinical Research (OCRe) [152] and Biomedical Resource Ontology (BRO) [153]. Leveraging existing ontologies, Kong et al. further expanded the schematic representation of clinical research using Conceptual Model Representation (CMP) to aid the development of clinical research databases [154]. Sahoo et al. created an informatics framework that allows detailed research data elements to be systematically mapped and represented [155]. By

combining NLP techniques, Valdez et al. were able to build an ontology-based clinical research knowledgebase for evaluating research studies and enhancing study reproducibility [156]. Ross et al. systematically analyzed eligibility criteria in clinical trials using heuristic rules and logic [157]. Our project aims to enhance the current informatics solution by demonstrating a methodologic development process (corpus development, sub-language analysis, and modeling) that uses NLP to discover the reporting patterns of EHR-based observational studies. Our investigation is focused on evaluating the trends, variability, utilization and adoption of EHR-based data abstraction related methodologies. Existing clinical research ontologies and research reporting standards were leveraged to help define important data elements.

7.3 Methods

Data Selection The Rochester Epidemiology Project is a National Institutes of Health-funded research infrastructure that collates and indexes health care information from virtually all sources of medical care available to residents of Olmsted County, Minnesota [158]. It has maintained a comprehensive medical records linkage system for over half a century, which makes it an ideal resource for conducting population-based studies. These data have been utilized by investigators throughout the country, resulting in more than 2,000 publications on a wide range of health care topics from top clinical journals and conferences including JAMA, NEJM, and Lancet. Our study investigated all articles from the REP publication registry between 1995 and 2016. In total, 1,543 articles were retrieved; 321 were removed due to their non-convertible PDF format or no full text being available. The final data set was comprised of 1,279 articles.

Guideline We adopted existing guidelines gathered from the Reporting Guidelines for Health Research EQUATOR Network24 including RECORD (REporting of studies Conducted using Observational Routinely-collected health Data) and STROBE (The Strengthening the Reporting of Observational Studies in Epidemiology) as the baseline guideline [159, 160]. Additional methodologic strategies were borrowed from research by Boyd et al. and Horwitz et al., including training, abstraction forms, meetings, monitoring, and testing of interrater agreement [161, 162]. In consultation with the above standards and guidelines, we defined the use of EHRs to include the following

processes: feasibility assessment, cohort identification (case selection), and data retrieval. Reporting categories were provided in Table 7.2:

Table 7.1: Definition of EHR-based Reporting Category

Reporting Category	**Definition**
Participants	The methods of study population selection (such as codes or algorithms used to identify subjects), validation of the codes or algorithms used to select the population, data linkage process, participant follow-up and matching.
Variables	The methods of the classification, assessment, and validation of variables (exposures, outcomes, confounders, and effect modifiers)
Data sources	The methods of data assessment (reliability and validity) and data collection (development, training, validation, and administration, such as blinding)

Annotation Based on the baseline guideline, we randomly sampled 200 articles (from 1,279) for manual review. 71 out of the 200 reports were randomly sampled and double read to determine interrater reliability. We defined two objectives for this process. The first objective was to assess the reporting cohesiveness to the existing standards. Each article was annotated for the presence or absence of methodologic standards provided in the guideline. The second objective was to identify additional important activities that were not captured by the current standards, such as recently proposed best practices and methodologies for the use of EHRs.

The annotation process was conducted according to Corpus Annotation Schemes [17], including organizing training sessions, developing annotation guidelines, multi-phase annotation, evaluation, and adjudication. Four annotators (N.W., J.J, X.Z, and S.P) were given initial one-hour training. Questions raised from the training exercise were used to refine the baseline guideline. In the first week, each annotator annotated four to eight papers (two papers for every batch). After each batch, the inter-annotator agreement (IAA) was calculated using F-measure (2* precision * recall/(precision + recall)). The matched cases were determined by comparing the bipartite set alignment for two annotated sentences using Kuhn-Munkres algorithm [163]. A consensus meeting was

organized to resolve disagreements and annotation issues. The process continued until a high agreement was reached. Over the next three to four weeks, weekly batches of a total of 200 papers were annotated. Each document was independently annotated by two annotators. After the weekly assignments were completed, we computed the IAA, resolved disagreements, and clarified the guidelines.

The final gold standard annotations were created by combining the individual experts' annotations followed by adjudication of the mismatches. The jointly annotated training notes were added to the gold standard but excluded from the final IAA computation. The annotation tool for this project is Multi-document Annotation Environment (MAE), a Java-based natural language annotation software package [109] .

Natural Language Processing Based on the above methodologic standards and gold standard corpus, we developed an NLP algorithm to automate the manual process. The infrastructure for the NLP system was adopted from the existing open source NLP framework MedTaggerIE [36], a resource-driven open-source Unstructured Information Management Architecture (UIMA) [164]-based IE framework. The NLP algorithm was developed through three steps: 1) prototype system development based on existing knowledge and standards, information theory algorithms, and expert knowledge, 2) formative system development using a training dataset and manual case review for iterative refinement, and 3) final system evaluation using a test dataset. A total of 200 annotated articles were divided into a training set (n=100) and a test set (n=100).

The initial step was to exclude irrelevant information by segmenting the information into different sections. Since MedTaggerIE contains a built-in section detector through terminology lookup [165], we only needed to modify the dictionary to match with sections for clinical research. Based on STROBE, we included sections that are related to study method, study design, data (collection), case definition, and participants (cohort).

Concept extraction is a knowledge-driven annotation and indexing process to identify phrases referring to concepts of interest in an unstructured text. We leveraged Pointwise Mutual Information (PMI), to provide heuristic ranks of n-gram features (an adjoining sequence of n items from a given sample of sentences) for keyword prototyping. Table 1 lists the top 15 n-gram features that were automatically generated from the algorithm.

Table 7.2: Example of N-gram Features with High PMI Scores

Reporting Category	**Top Uni- and Bigram Features with the Highest PMI Score***	**Example Sentence**
Participant	records; medical; medical records; residents; identified; reviewed; records of; review; complete; case; county residents; were identified; subjects; were reviewed; nurse	"We reviewed charts to identify cases of PJP, cross referenced with the REP database using diagnostic codes for PJP and the Mayo Clinic and Olmsted Medical Center databases."
Data source	medical records; data; records of; reviewed; information; was collected; records linkage; used; database; nurse; system; selected; from the; obtained; by trained	"We used the REP database to retrieve all medical records for residents of Olmsted county who had an established diagnosis of any of the subtypes of CLE"

*Keywords in second column were only for system prototyping and had not been manually curated

Due to a high textual similarity between participant and data source, additional patterns needed to be identified in order to accurately distinguish the two classes. Thus, we analyzed the syntactic patterns by parsing the dependencies of each sentence. We used the Stanford CoreNLP 3.9.235 and integrated it into the existing MedTagger UIMA framework. We found the majority of methodologic events can be modeled into concepts separated by semantic connectors. For example, Figure 7.1 shows the syntactic structure of sentence objective (event confirmation) and methodologic event (medical record review) were connected through the case making element and the adjectival modifier.

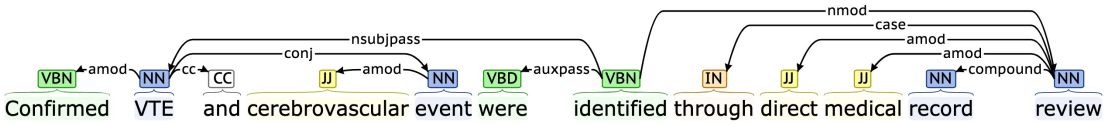


Figure 7.1: Parsing Structure of Case Ascertainment

Context Detector The assertion and temporal expressions were handled by the MedTaggerIE context detector. The assertion of each concept includes certainty (i.e., positive, negative, and possible) along with experiencer (i.e., patient, associated with someone else), while temporality identifies historical or present. For example, from the sentence "Data were collected from a random sample using questionnaires," "Data" would be extracted as a data concept and "collected" would be extracted as a methodology concept, along with corresponding assertion status "positive", temporality "past".

Normalization and Summarization After keywords or phrases were extracted from a sentence, they were normalized to a specific concept. As an example, the phrases "patient record" and "medical record" were normalized to the concept of "data". The normalized categories were processed by the rule engine, a series of conditional clauses including "and", "or" and "not" independently. A summary of these concepts, keywords, and modifiers are listed in Table **??**.

Table 7.3: Keywords for Concept Extraction

NLP Concept	Keywords*
Population selection general	assemble(d); categorize(d); choose(chosen); classify(classfied); construct(ed); contact(ed); draw(drawn); determine(d); establish(ed); screen(ed); select(ed); recruit(ed); invit(ed); sample(d); abstract(ed); cohort screen(ing); complete(d) abstraction
Population selection specific	search(ed); review(ed); identify(ied); exclude(d); include(d); confirm(ed)/ascertain(ed); avoid(ed);
Measurement	measurement(s)/ assessment(s) performed; assessed; measure(d); determine(d); evaluate(d); search(ed); review(ed); identify(ied);
Data collection	retrieve(d); collecte(d); obtaine(d); retrieve(d); contact(ed); interview(ed); complete abstraction; data collection/abstraction; questionnaire(s)/survey(ies) was/were designed/used/created/mailed;
Validation	validate(d); validation; confirm(ed); verify(verified); verification; ensure(d); agreement(s); agreement measure(s); accuracy; inter/intra-rader/annotator/observer agreement(s); agreement(s) between; IAA; test retest; gold standard; kappa; reliability; validity; (doubly; double; triply; triple; quadruply; quadruple) + (review/read/exam/assess/measure) + (twice; multiple times); Consensus; disagreement(s) resolved;
Data linkage	linking; link(ed); data linkage; linkage system; indexing; cross referenc(ed/ing); cross match(ed);
Follow up	follow(ed) up; follow(ed) up through; follow up period; follow(ed) for
Matching	match(ed) with; matching; match(ed) (subject) to; matched pair with; matched(ing) on; matched in (characteristics); matched for;
Cohort-related	cohort (of); sub(-)cohort; population; participant(s); patient(s); control(s); case(s); resident(s); child; children; man; men; woman; women; subject(s); adult(s); volunteer(s); person(s); survey respondent(s); comparison group(s)

NLP Concept	Keywords*
Study team/abstractor	abstractor(s); specialist(s); fellow(s); RTP(s); research temporary professional(s); intern(s); author(s); reviewer(s); operator(s)
Eligibility-related	eligible; eligibility; ineligible; criteria; criterion; inclusion criteria; exclusion criteria; included; excluded; screening protocol; screening
EHR-related	medical record(s); information; data; record(s); characteristic(s); chart(s); sample(s); questionnaire.?; database; computerized diagnostic index; EHR(s); EMR(s); electronic medical record(s); electronic health record(s); survey(s);
Terminology	diagnostic; diagnostic code; ICD(s); international classification of diseases; CPT(s); current procedural terminology; Berkson code (s or ing) (REP cohort only); symptom(s); factor(s);

*Keywords should be connected through wild card regular expression for fuzzy matching; additional refinement is required when used for different data sources.

The following patterns were used to identify the common expressions for each class. We used square brackets "[]" to represent each concept group, parentheses "()" for direct keywords, curly brackets for "" typed dependencies, "|" for the conjunction or, and "&" for the conjunction and. The expressions for dependencies were followed by the Stanford Typed Dependencies Manual35, where "auxpass" represents passive auxiliary, "case" represents case-marking elements, "mark" represents a marker which introduces a clause subordinate to another clause, and "amod" as an adjectival modifier. Textbox 1 provides the logic rules for nine different events related to the use of EHRs for clinical research.

Table 7.4: Example of Language Variation between Two Data Sources

Reporting Category	**Methodologic Events**	**NLP Rules**
Participants	Study population selection (study event)	[Cohort-related] AND auxpass AND [Population selection general] AND preposition case-marking element AND [Population selection specific]; [Population selection general | specific] AND preposition marker AND [Population selection specific] AND ([Cohort-related | Eligibility]); [Data] AND auxpass AND [Population selection specific] AND preposition case-marking element AND [Study team/abstractor] AND preposition marker AND [Population selection general | specific]
	Screening validation (validation of screening protocol, procedure, inclusion and exclusion criteria)	[Population selection concepts] AND auxpass AND [Validation]; [Cohort-related] AND [Study population selection] AND [Validation]
	Data linkage process (study event)	([EHR related] | [Cohort-related]) AND [Data linkage]
	Participant follow-up (cohort study only)	[Cohort related] AND auxpass AND [Follow up]; [Follow up] AND [Cohort related]
	Matching (matched studies only)	[Matching] AND [Cohort related]
Variables	Measurement and classification of variables (study and clinical event)	[UMLS Dictionary (medical concept | procedure)] AND [Measurement]
	Validation of variables (confirmation of subject has certain characteristics)	[UMLS Dictionary (medical concept | procedure)] AND [Measurement] AND [Validation]

Reporting Category	Methodologic Events	NLP Rules
Data sources	Data collection (study event)	[EHR related] AND auxpass AND [Data collection] AND preposition case-marking element; [Study team/abstractor] AND [Data collection] AND [EHR-related]; [Population selection general] AND preposition marker AND [Data collection] AND [EHR-related]
	Data quality assessment (validation of data collection tools, frameworks, protocols and methods)	[Data collection] AND [Validation]

After section selection and sentence detection, the final test corpus consisted of 1220 sentences. Each sentence was pre-annotated with either one of the nine categories listed in Table 4 or "other". The final system was evaluated on all tasks using F-measure.

Analysis of Methodologic Reporting Patterns We applied the algorithm to the entire REP cohort from 1995 to 2016. Each sentence was categorized into the above nine categories. Once the category was determined, we conducted a sublanguage analysis to identify how each activity was conducted. Briefly, we identified the top five commonly used case-marking elements and markers including "using", "through", "with use of", "with", "via" for identifying the key methodologic expressions. The expressions that cannot be identified automatically were assessed through manual review. Furthermore, we applied the research practice framework proposed by Boyd et al., Horwitz et al., and Gilbert et al. to evaluate the quality of the reported methods through the identification of the following six activities: screening/data collection protocol, training, blinding, inter-observer agreement, team meetings, and supervision. To understand the usage of informatics tools and methods, a trend analysis was conducted using least-squares regression to test the incremental significance of the use of methodology throughout the years. Finally, we randomly sampled 40 articles to conduct an authorship and affiliation analysis through manual review. The goal of this analysis is to understand whether there would be a variation regarding the reporting patterns given the first author's training

background. As the focus of our study is observational research, we classified each author into epidemiology background and other using the affiliation information from PubMed. We then defined the positive outcome as the satisfaction of at least three activities from the framework (Boyd et al., Horwitz et al., and Gilbert et al.) and the negative outcome as less than three criteria were discovered.

7.4 Results and Discussion

Performance of Annotation and NLP Three articles were removed due to no full-text found. The averaged inter-annotator agreement (IAA) of manual reviewing of 71 articles is 0.863 in F-measure. The evaluation of the NLP system on the test articles is provided in Table 7.5. We found the identification of variables was the most challenging task. Many expressions either lacked context or used very specific terminology such as directly referring to the inventor's name (e.g. Morris, a type of clinical rating scale assessment for dementia). The MedTaggerIE dictionary look up from UMLS Metathesaurus was able to provide additional context information such as medical concept and assessment type, however, the performance of the system on this task is capped by the comprehensiveness of the dictionary. Despite this limitation, the system achieved a moderate-high performance over nine different tasks.

Table 7.5: Performance of IAA and NLP System of Nine Different Tasks

Reporting Category	Methodologic Events	NLP (F-measure)
Participants	Study population selection	0.716
	Screening validation	0.866
	Data linkage process	0.900
	Participant follow-up (cohort study only)	0.888
	Matching (matched studies only)	0.955
Variables	Measurement and classification of variables	0.780
	Validation of variables	0.759
Data sources	Data collection	0.850
	Data quality assessment	0.967

Analysis of Methodologic Reporting Patterns Our analysis showed that manual chart review was the most popular method reported for study population selection (51.92%) and case validation (7.97%) and the second most popular method for data collection (4.14%). We found electronic retrieval (i.e. query) was the most popular method for data collection (6.18%). However, there were a large number of articles that did not specify what methods they used for various tasks, e.g. only 49% of articles mentioned activities related to data collection. We believed this was due to the lack of reporting standards. For example, the expression like "all clinical variables were either obtained electronically or from patient records" described a potential data collection activity that was conducted. However, there were no related expressions discussing how exactly the data were collected, when this activity happened, or who conducted the abstraction. Furthermore, even among the sentences with abstractors mentioned, 77% use the pronoun "we" as an unspecified expression for the entire method sections.

Table 7.6: Summary of Methodologic Events Among 1279 Articles from REP

Sections	Methodologic Events*	Number of Articles (n=1279)
Participants	Study population selection	90.30% (1155)
	Chart review	51.92% (664)
	Database Query	30.6% (391)
	Standard terminology	7.43% (95)
	Cohort screen tools	1.49% (19)
	Computer-based algorithms	1.41% (18)
	Natural language processing	1.09% (14)
	Validation	22.28% (285)
	Chart review	7.97% (102)
	Existing criteria	5.47% (70)
	Questionnaires/survey/interview	0.70% (9)
	Data linkage	26.11% (334)
	Participant follow-up	49.57% (634)
	Matching	21.11% (270)
Variables	Measurement/Assessment of variables	35.26% (451)
	Clinical intervention/criteria	29.48% (377)
	Questionnaires/survey/interview	7.35% (94)
	Chart review	1.41% (18)
	Computer-based algorithms	0.47% (6)
	Validation	9.85% (126)
	Questionnaires/survey/interview	1.96% (25)
	Existing criteria	0.16% (2)
Data sources	Data collection	49.18% (629)
	Electronic retrieval	6.18% (79)
	Manual chart review	4.14% (53)
	Survey/questionnaire/interview	1.40% (18)
	Electronic data capture tools	0.70% (9)
	Data quality assessment	0.63% (8)

In assessing the use of methodologic standards, 5% (61) reported the use of a screening/data collection protocol, 24.0% (146) reported training for data abstraction, 6% (74) reported the abstractors were blinded, 4.5% (57) tested the inter-observer agreement,

1.5% (19) reported that team meetings were organized for consensus building, and 0.8% (10) mentioned supervision activities by senior researchers. In comparison with the study conducted by Gilbert et al. in 1996, we found an increasing number of studies reported the use of good methodologic practices when dealing with EHRs. However, no single methodologic standard had an adoption rate of 25% or greater among the 1279 articles. Our author and affiliation analysis showed that papers with the first author of epidemiology background were more likely to report good practices (Figure 2). However, the result was not significant (p-value = 0.118).

The trend analysis showed a significantly increased number of articles reported using informatics-related methods (i.e. electronic data capture, phenotype algorithms, etc.). Figure 3 shows an upward trend of using informatics methods since 1998 (p-value $<$ 0.0001). Among these articles, we were able to identify 11 different phenotype algorithms from 14 articles that used computer-based algorithms for case ascertainment (Table 5).

Table 7.7: Computer Based Case Ascertainment Algorithms

Computer Based Case Ascertainment Algorithms	**Articles**
Interstitial lung disease	[166]
Myocardial infarction	[167, 168, 169]
Osteoarthritis	[170]
Fracture risk assessment	[171]
Antineutrophil cytoplasmic autoantibody –associated vasculitis	[172]
Vertebral deformities	[173]
Nonalcoholic fatty liver disease	[174]
White matter hyperintensity volume	[175]
Herpes zoster	[176, 177]
Cause of death	[178]
Heart failure	[179, 180]

Reporting Standards Based on the gap analysis of the STROBE statement, we identified the following recommendations for the reporting of EHR-based clinical research (Table 7.8). The recommendations included specific activities related to EHR-based information quality assessment and information provenance that involve additional informatics activities such as electronic data retrieval, cohort screening, assessment methods

of study variable, definition and data sources, etc.

Table 7.8: Reporting Standard of EHR-based Chart Review Research (Item number based on STROBE Statement)

Section	**Item Number**	**Reporting Recommendation**
Introduction		
Background and rationale	2	Existing clinical criteria, validity of clinical criteria, EHR data source used (e.g. free text, billing), methods of EHR-derived phenotyping (e.g. code-based method, text-based method, hybrid), validity of EHR-based methods, reported epidemiology ratio (e.g. incidence, prevalence)
Methods		
Participants	6	Methods of selection of participants: Clinical criteria for cohort definition (e.g. inclusion, exclusion criteria) EHR related methods for cohort screening Evaluation method for cohort screening (e.g. chart review) Validity of cohort screening methods
Variables	7	Definitions of all clinical variables (e.g. outcome, exposures, confounders), clinical criteria, validation methods
Data sources and measurement	8	Describe sources of EHRs; Details of methods of assessment (measurement); Describe comparability of assessment methods if there is more than one group
Bias	9	Report EHR related bias (e.g. query bias, documentation bias)

7.5 Conclusion

In summary, our study demonstrated a process of using informatics to discover research reporting patterns and methodologic events from a series of papers that used the REP cohort. Our investigation discovered an upward trend of reporting research methodologies, good practices, and the utilization of informatics-related tools and methods for EHR based clinical research. Despite these findings, the methodologic standards

were still consistently under-reported. We also discovered high variation regarding clinical research reporting. Developing process frameworks, ontology models and reporting guidelines for the given context are recommended for future work.

Chapter 8

Application to Real-World EHR-based Study

8.1 Overview

This chapter presents an end-to-end implementation of the TRUST process leveraging the proposed implementation guide and best practices based on a real-world EHR-based clinical study. We first present the process of design, development, evaluation, and deployment of a clinical NLP algorithm of delirium. We then qualitatively and quantitatively examine the validity and reproducibility of the process based on the RITE criteria.

8.2 Background and Motivation

Delirium is a syndrome with symptoms that present as confusion and is characterized by an acute change in mental status, fluctuating course, lack of attention, and disorganized thinking or altered level of consciousness [181]. Delirium is common in hospitalized older adults [182, 118], with prevalence in the postoperative setting ranging between 21% and 35% [183], and in intensive care unit patients between 60% and 85% [184]. Delirium has been associated with multiple predisposing and precipitating risk factors, including infections, use of specific medications, and the presence of a wide range of other chronic conditions [185]. In particular, persons with dementia are at high risk of delirium, and

delirium has been associated with subsequent cognitive impairment [186] and additional adverse outcomes [187], including longer hospital stays, increased likelihood of nursing home placement, and an increased risk of death [118, 188, 189, 190, 191].

Delirium is underdiagnosed in clinical practice and is not routinely coded for billing [119]. A study of patients undergoing elective surgery indicated that delirium was given an International Classification of Diseases, 9th Revision (ICD-9) code in only 3% of patient records [182]. Causes of delirium are multi-factorial, but it has been estimated that 30%–40% of cases may be preventable [118]. However, further research is necessary to identify the best ways to detect, prevent, and manage delirium. Such research is currently limited by challenges in identifying persons with delirium from electronic health records (EHRs). Inouye et al. [182] have developed methods for identifying persons with delirium from chart review of medical records. However, manual chart review is time-consuming and costly to extract information from clinical notes for large patient populations. Natural language processing (NLP) has been adopted to computationally extract clinical information from EHRs for a wide range of applications ranging from advancing EHR-based clinical research [192, 193] to supporting clinical decision-making [83, 194]. Different NLP frameworks have been developed to convert clinical narratives into structured data, including MedLEE [195], MetaMap [39], KnowledgeMap [40], MedTagger [36], and cTAKES [38]. In this study, we adopted the MedTagger framework with domain-specific customizability to develop and validate an NLP algorithm to identify the occurrence of delirium using clinical notes derived from the Mayo Clinic EHR.

8.3 Materials and Methods

Study Implementation This study was approved by the Mayo Clinic Institutional Review Board and the Olmsted Medical Center Institutional Review Board. We followed our previously proposed assessment methods (chapter 3, 4, 5), frameworks (chapter 6, 7) and best practices[1] to conduct the study, which include sequentially following the TRUST process, adopting templates of screening protocol and annotation guideline, referring best practices in the procedure guideline and complying with documentation checklist.

[1] `https://github.com/OHNLP/annotation-best-practices`

Evaluation Criteria To assess the technical soundness of the process, we considered the FAIR data principle. The FAIR data principle (Findable, Accessible, Interoperable, and Reusable) is comprised of a set of guiding principles to guide the reusability of digital assets [196]. In addition, we applied our previously proposed RITE implementation principles (Reproducible, Implementable, Transparent, and Explainable). The RITE principles, presented in the figure 8.2, not only emphasize result reproducibility but also process transparency and implementability; since the variability and explainability of the result are dependent on the process. The implementation details are presented in the section 8.4.

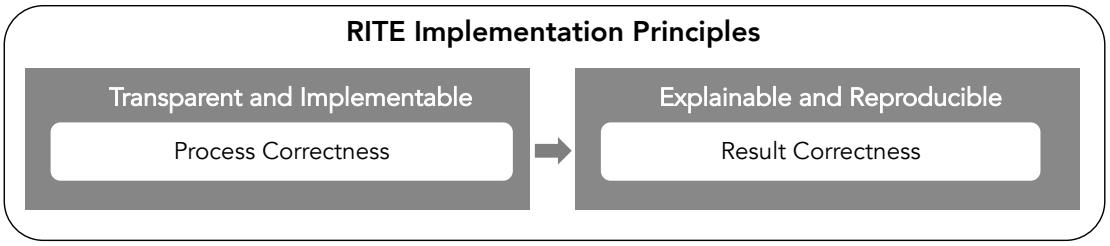


Figure 8.1: RITE Implementation Principles

8.4 Process Correctness

The overview of the implementation of the TRUST process for the development of the delirium algorithm is presented in Figure 8.2.

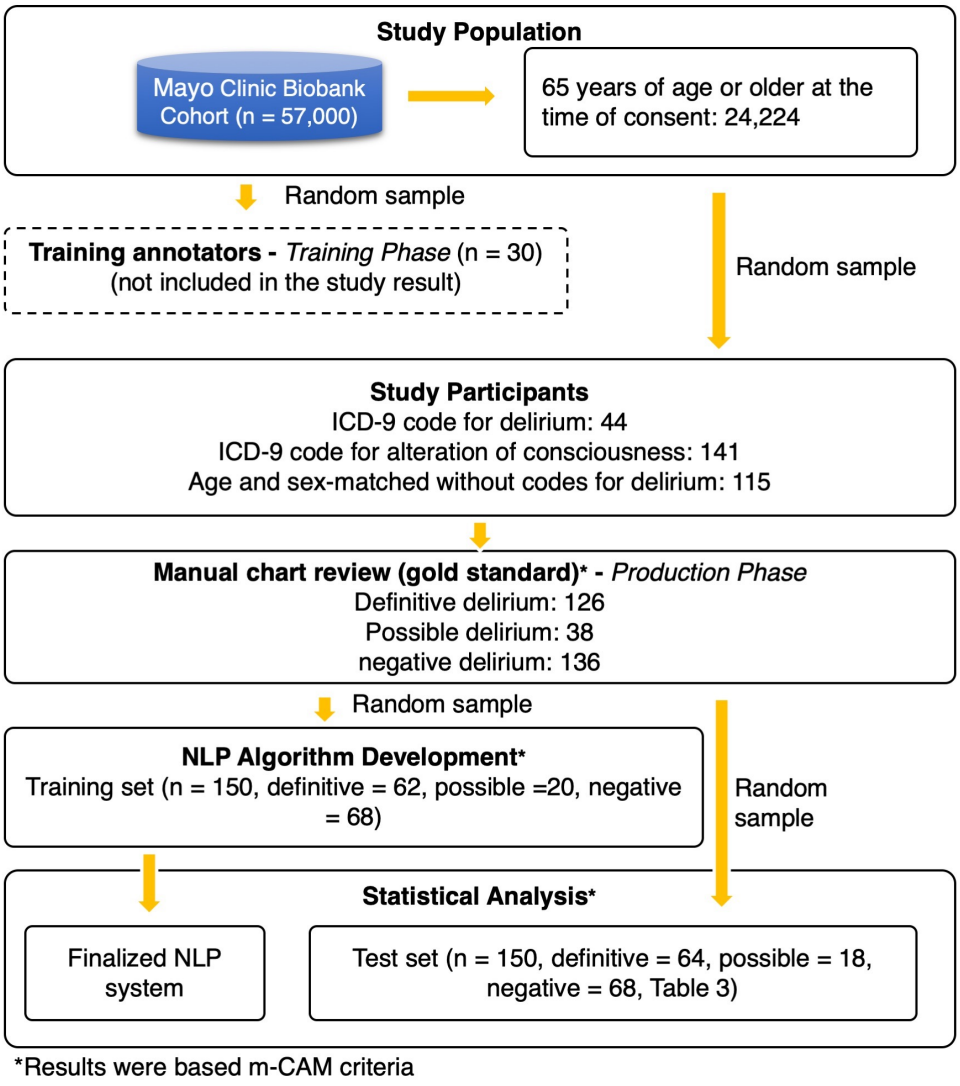


Figure 8.2: Workflow of Cohort Screening and Sampling for the Corpus Annotation and NLP Development

Study Population The study population consisted of participants of the Mayo Clinic Biobank [197]. The Mayo Clinic Biobank is an institutional resource comprised of volunteers who have donated biological specimens, provided risk factor data, and have given permission to access clinical data from their EHRs for clinical research studies. Participants were contacted as part of a prescheduled medical examination at Mayo

Clinic sites between April 2009 and September 2015. All participants were 18 years or older at the time of consent. Approximately 57 000 participants have been enrolled, and 24 224 of these participants were 65 years of age or older at the time of consent. Among these participants, we identified all persons who received an ICD-9 code for delirium or alteration of consciousness after date of enrollment (N = 731; ICD-9 codes: 290.3, 290.41, 291.0, 292.81, 293.0, 293.1, 348.30, and 780.09). We randomly sampled persons (N = 300) from this population for the annotation guideline development and gold standard identification phases of the study (details below). Among 300 randomly selected individuals aged 65 and older, 48.3% were female. A total of 615 visit records were noted for these 300 individuals of which 247 were inpatient records, 55 observation records, 186 emergency records, and 127 outpatient records. Among the 247 inpatient records, 89 visits were ICU records. Half of the persons from the sample population (n = 150) were used to develop NLP algorithms to identify occurrences of delirium, and the remaining sample (n = 150) was set aside to evaluate the performance of the NLP algorithm.

Annotation Guideline Development Delirium is diagnosed based on a constellation of established clinical symptoms. Therefore, our primary criterion for identification of delirium was an explicit mention of a delirium episode (eg, "Patient experienced acute post-operative delirium this morning") documented in the clinical notes. If there was no clear diagnosis or mention of delirium, then we identified delirium based on whether symptoms documented in the clinical notes satisfied the "Confusion Assessment Method" (CAM) criteria [182].

Briefly, CAM has 4 features that are used to facilitate the diagnosis of delirium, including (A) acute onset and fluctuating course, (B) inattention, (C) disorganized thinking, and (D) altered level of consciousness. Each feature was further represented by a list of specific delirium-related concepts, such as deteriorating mental status, drowsiness, mumbling gibberish, impaired orientation, and encephalopathy [182]. For example, the CAM feature "Disorganized thinking" was represented by several expressions, including "mumbling gibberish," "rambling speech," "unclear flow of ideas," etc. Figure 8.3 summarizes CAM features.

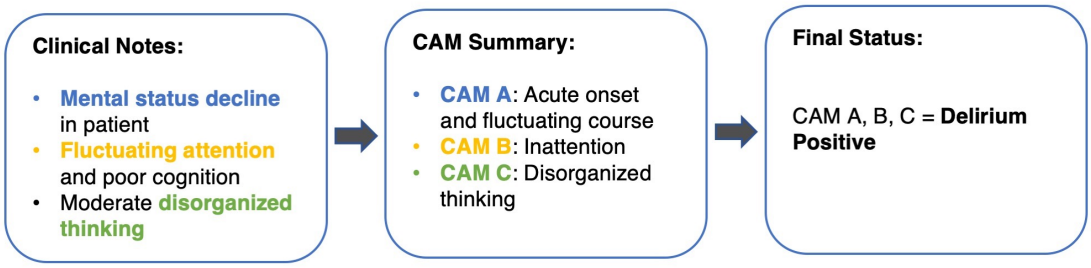


Figure 8.3: Comparison of Original CAM and the Modified CAM. CAM = Confusion Assessment Method.

In our study, we developed 2 versions of the criteria: the original CAM and the modified CAM (mCAM) (Table8.1). For the original CAM, we operationalized the results as "definitive delirium" or "no delirium." "Definitive delirium" status was achieved when the medical records described symptoms that match criteria A and B and either C or D of the CAM criteria within 1 month. The average duration of time from the first symptom to the last symptom was 13.8 days (range: 0–28 days). Most persons with delirium met the criteria within 48 hours (38%).

Table 8.1: Confusion Assessment Method (CAM)

A: acute onset and fluctuating course	**B:** inattention
Do the abnormal behaviors?	Does the patient:
Come and go	Have difficulty focusing attention
Fluctuate during the day	Become easily distracted
Increase/decrease in severity	Have difficulty keeping track of what is said
C: disorganized thinking	**D:** altered level of consciousness
Is the patient's thinking?	What is the patient's level of consciousness?
Disorganized	Alert (normal)
Incoherent	Vigilant (hyper-alert)
	Lethargic (drowsy but easily roused)
	Stuporous (difficult to rouse)
	Comatose (unrousable)
Original CAM:	*Modified CAM:*
Definitive: A and B and (C or D)	Definitive: At least 3 unique CAM criteria
	Possible: Any 2 criteria and does not meet
	the definitive criteria as above

Two Mayo geriatricians and one palliative care physician (S.P., Z.X., B.T.) helped to define the review criteria and noted that symptoms of delirium may be poorly documented (missing information related to delirium features and concepts) [198]. Therefore, we created mCAM to address this issue. Using the mCAM criteria, definitive delirium status was defined as when the medical records describe symptoms that matched 3 of 4 CAM criteria (eg, CAM B + C + D or A + C + D). Possible delirium status was defined as when symptoms matched exactly 2 CAM criteria.

Corpus Annotation Corpus annotation is the process of manual chart review, marking interpretative linguistic (eg, syntax, negation) or predefined clinical information (eg, delirium-related concepts) to a corpus that can be used for NLP algorithm development and evaluation [17, 18, 7]. There were 2 phases involved in our process: (a) training phase to be familiar with the annotation process and refine annotation guidelines and (b) production phase to create the gold standard for NLP algorithm development and evaluation.

Training phase One geriatrician and one psychologist (S.P. and G.S.L.) annotated a random sample of records obtained from 15 patients who had an ICD-9 code for delirium or alteration of consciousness and another 15 patients who did not have code to identify documentation of delirium and/or keywords and terms related to the 4 CAM components. In the training phase, annotators reviewed the full medical records from 30 days prior to the date of the ICD-9 code through 30 days following receipt of the initial diagnosis code. They identified delirium-related terms previously described by Puelle et al. [199] for the identification of delirium from medical records as well as additional keywords associated with episodes of delirium observed in the sample records. Discrepancies between reviewers were discussed and resolved, and annotation criteria were updated.

Production phase A new sample of 44 patients with an ICD-9 code for delirium and 141 patients with an ICD-9 code for the alteration of consciousness was matched by age ($\pm$1 year) and sex to 115 patients without a code for delirium (Figure 1). All 300 patient records within $\pm$ 30 days anchored by ICD-9 delirium diagnosis (total 8761 documents) were double-annotated by 2 reviewers (G.S.L. and D.I.) using the final annotation guidelines. Interannotator agreement (IAA) was calculated at the patient-level delirium status (eg, definitive delirium, no delirium) and for each concept (eg, confusion). All conflicting cases were adjudicated by a geriatrician and palliative care physician with expertise in geriatrics (S.P. and B.T.). The results of the final adjudicated annotation served as the gold standard for the development and test of the NLP algorithm.

NLP Algorithm Development The NLP algorithm was developed to automate EHR chart review to identify patients with delirium based on 2 definitions of delirium (CAM and mCAM). To identify a patient's delirium status, the NLP algorithm screens the clinical notes to extract direct mention of delirium (physician's diagnosis) and delirium-related clinical concepts that match the CAM criteria. Each concept was then normalized into a standard form based on the CAM instruments. For example, "unresponsiveness" and "decreased responsiveness" were normalized into "disconnected." Furthermore, the normalized CAM instruments were mapped into the CAM Features A, B, C, and D. As an example in Figure 8.4, a patient who experienced acute altered mental status (CAM Feature A), inattention (CAM Feature B), and disorganized thinking (CAM Feature C) was considered "definitive" delirium.

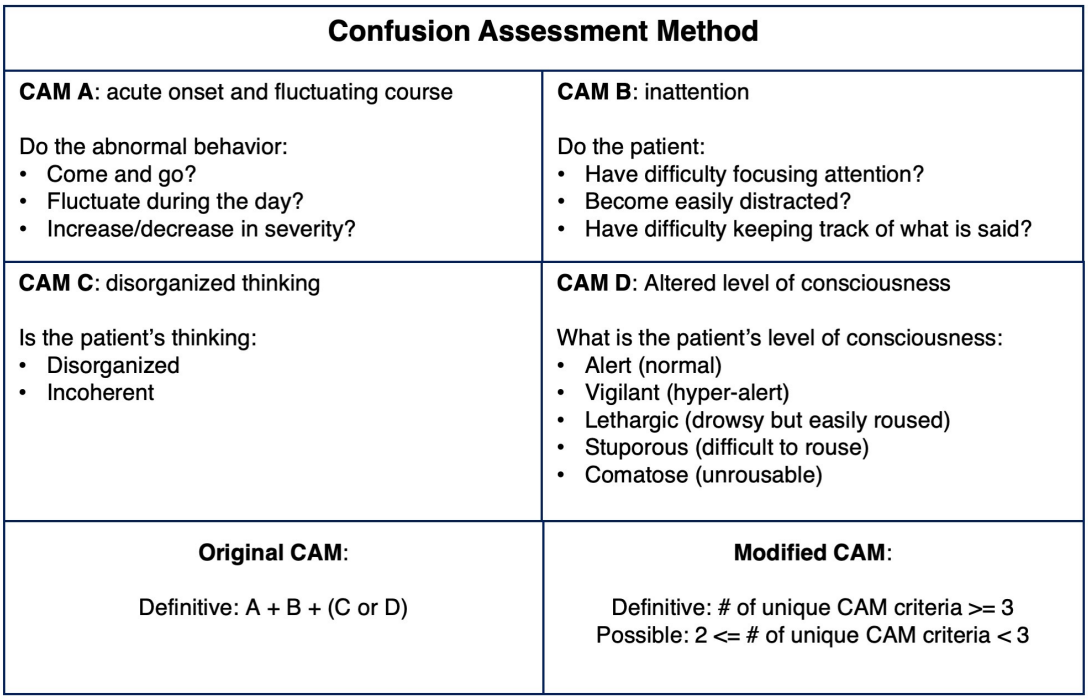


Figure 8.4: An Example for Detecting Delirium Status Based on CAM

To implement the algorithm, we adopted the open-source NLP pipeline MedTaggerIE [36], an open-source unstructured information management architecture–based information extraction framework. This system separates task-specific NLP knowledge engineering (ie, CAM criteria) from the generic routine NLP, which enables words and phrases containing clinical information (ie, keywords relevant to CAM features) to be directly coded by subject matter experts. The tool has been utilized in various clinical NLP tasks and adopted by multiple studies of phenotyping algorithm development [191, 200]. Additionally, we utilized the Mayo Clinic big data NLP platform [30], a distributed parallel computing environment to support data sets of extremely large volume which integrates NLP with existing EHR data stores. This enabled us to execute the NLP algorithm without manually retrieving large sets of documents prior to execution.

Figure 8.5 shows the NLP workflow. The generic NLP includes sentence segmentation, tokenization, temporal status detection (eg, present, history), and assertion detection (eg, negation, possible, hypothetical). The task-specific NLP includes the detection of keywords relevant to delirium in the text using regular expressions and normalized to specific delirium concepts. The summarization component applies heuristic rules (ie, CAM criteria) for assigning the delirium status.

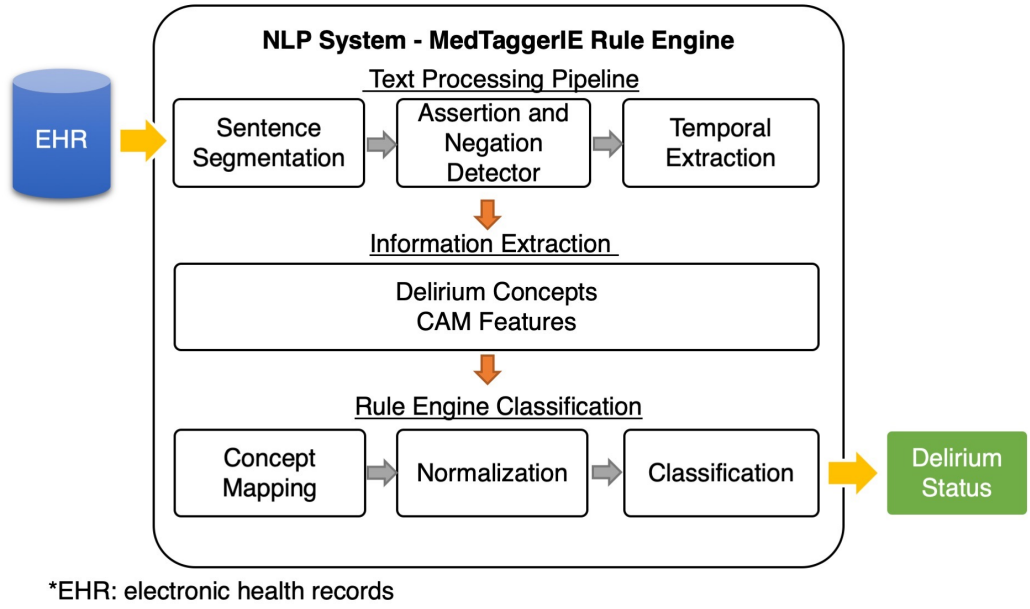


Figure 8.5: Architecture of the NLP

The NLP algorithms were developed in the following 3 steps: (a) prototype algorithm development based on CAM, (b) formative algorithm development using the training data after the acceptable performance was reached (accuracy >0.95), and (c) final algorithm evaluation on the independent test set. The algorithm was applied and refined on the training data. Incorrect cases were manually reviewed by 2 domain experts (G.S.L. and D.K.I.) and iteratively refined until all issues were resolved.

NLP Algorithm Evaluation The manual annotation of delirium status by 2 annotators was assessed by F1-score [111]. F1-score (eqn (1)) is a well-established metric in the information retrieval and machine learning community. It measures both positive

predictive value (precision) and sensitivity (recall) of the test object. The performance of the algorithm was assessed by using sensitivity (eqn (2)), specificity (eqn (3)), and accuracy (eqn (4)), to assess concordance between delirium status identified using the NLP algorithms and delirium status identified via manual chart review (gold standard).

F1-score = 2 × (Precision × Recall)/(Precision + Recall) (1)

Sensitivity = True positive/(True positive + False negative) (2)

Specificity = True negative/(False positive + True negative) (3)

Accuracy = (True positive + True negative)/(Total positive + Total negative) (4)

NLP Algorithm Deployment To further compare the effectiveness between the NLP algorithm and ICD-9, we applied NLP-CAM and NLP-mCAM to screen all hospitalized patients who visited Mayo Clinic Rochester from April 2009 to September 2015. About 134 910 patients aged 65 years or older who were hospitalized at the Mayo Clinic in Rochester Minnesota were identified. For the purposes of comparison, we calculated the prevalence of positive delirium cases based on ICD-9, NLP-CAM, and NLP-mCAM.

8.5 Results Correctness

Annotation Guideline As the guideline was developed and used, some attributes were refined to make annotations more informative. For example, subcategories such as injury and trauma were added to the cause attribute to make the selections better fit the data. Other attributes were dropped due to scarce mentions in the corpus. Examples of dropped attributes include pain trend, which was intended to summarize whether pain was increasing, decreasing, or staying the same, and referral, which would identify a clinician referring a patient to another service.

Interannotator Agreement Among the 300 patients, 8761 clinical documents were double-reviewed, and 7515 delirium-related concepts were annotated. The IAA of patient-level delirium status (N = 300) between 2 annotators in F1-score was 0.94. Agreement between the 2 annotators at the concept level (ie, whether 2 annotators identify the same delirium-related terms, eg, Figure 8.3, box 1, mental status decline; N = 7515) in F1-score was 0.87. Overall, there were high agreements between the 2 annotators at both the delirium status and individual delirium concept levels.

Corpus Availability To enhance interoperability, shareability, and reusability, the

finalized annotation corpus was converted to the BioC format, a common interchange and extensible mark-up language format to represent, store and exchange text data [201]. Figure 8.6 presents a mock-up patient with identified CAM concepts. Here, the patient presents all CAM features within three different documents during a single episode of delirium. All CAM concepts were further standardized and mapped to SNOMED CT (Systematized Nomenclature Of Medicine–Clinical Terms) Identifier.

```
{
  "patient_1": {
    "patient_id": 123123,
    "delirium_status": "definitive",
    "document_1": {
      "document_id": 1234567,
      "document_date": 20150614,
      "cam_b": [
        {
          "concept_1": {
            "name": "inattention",
            "spans": "23-34",
            "SCTID": "22058002"
          }
        },
        {
          "concept_2": {
            "name": "inattention ",
            "spans": "67-78",
            "SCTID": "22058002"
          }
        }
      ],
      "cam_a": [
        {
          "concept_3": {
            "name": "fluctuating_course",
            "spans": "120-138",
            "SCTID": "255341006"
          }
        }
      ]
    },
    "document_2": {
      "document_id": 1234568,
      "document_date": 20150615,
      "cam_d": [
        {
          "concept_4": {
            "name": "altered_mental_status",
            "spans": "320-341",
            "SCTID": "419284004"
          }
        }
      ]
    },
    "document_3": {
      "document_id": 1234569,
      "document_date": 20150616,
      "cam_c": [
        {
          "concept_5": {
            "name": "disorganized_thinking",
            "spans": "23-44",
            "SCTID": "736319003"
          }
        }
      ]
    }
  }
}
```


Figure 8.6: Illustration of annotated delirium concepts exported into JavaScript Object Notation (JSON) format.

NLP Performance The NLP-CAM algorithm demonstrated a sensitivity, specificity, and accuracy of 0.919, 1.000, and 0.967, respectively, at identifying delirium compared to the gold standard (Table 8.2). The NLP-mCAM algorithm demonstrated a

sensitivity, specificity, and accuracy of 0.827, 0.913, and 0.827, respectively, at identifying definite, possible, and no delirium compared to the gold standard (Table 8.3).

Table 8.2: 2 × 2 Contingency Table of NLP-CAM

		Gold Standard		
		Delirium	**No Delirium**	***Total***
NLP	Delirium	57	0	57
	No Delirium	5	88	93
	Total	62	88	150
		Sensitivity = 0.919	Specificity = 1.000	Accuracy = 0.967

Table 8.3: 3 × 3 Contingency Table of NLP-mCAM

		Gold Standard			
		Definitive Delirium	**Possible Delirium**	**No Delirium**	***Total***
NLP	Delirium	60	4	2	66
	Possible Delirium	1	8	10	19
	No Delirium	3	6	56	65
	Total	64	18	68	150

NLP System Availability To enhance system sharability and transparency, we released and maintained the NLP system under the Open Health Natural Language Processing (OHNLP) Consortium, which can be accessed through[2] under the code-exchange and version control platform Git. The NLP system contains two components: 1) a generic MedTagger framework and 2) delirium algorithms, which are separated part from the main program. This architecture design can allow the algorithms to be easily plugged into the main program for better sharability and customizability. Additional user instructions including syntactic formats, such as text input, pre-processing instructions, and system directories and system-level instructions and requirements were maintained in a delirium-specific repository[3]

[2] `https://github.com/OHNLP/MedTagger`
[3] `https://github.com/OHNLP/AgingNLP/tree/master/delirium`

Downstream Application When the NLP-CAM algorithm was applied to the clinical notes of patients who were aged 65 or older and hospitalized at the Mayo Clinic in Rochester Minnesota between April 2009 and September 2015, 12 651 (9.4%) patients were identified as having delirium. The NLP-mCAM algorithm yielded 20 611 (15.3%) definite delirium cases and 10 762 (8.0%) possible cases. About 5490 (4.1%) of the sample population were identified as having delirium through ICD-9 screening. The results were consistent with the published literature, between 15% and 20% for patients with age greater or equal to 65 (4,35). These results reflect the ability of the NLP-based phenotyping algorithm to identify more likely delirium cases than conventional code-based screening methods.

8.6 Discussion

In this study, we implemented the TRUST process to develop and evaluate two NLP algorithms (NLP-CAM and NLP-mCAM) in the task of automatically identifying patients with delirium from clinical notes. The evaluation demonstrated that even without a definitive diagnosis of delirium, using descriptive terminology consistent with and meeting standard CAM criteria. The implementation of both algorithms demonstrated high performance in classifying delirium. In addition, when applied to a large group of randomly selected patients, the NLP algorithms were able to identify more patients with delirium compared to structured data (ICD codes). These results suggest that these algorithms may have high sensitivity to capture occurrences of delirium by mining relevant keywords and concepts in clinical free text. Compared with conventional manual chart review, NLP provides more systematic and scalable solutions in identifying clinical concepts. In general, human annotators have a lower test–retest reliability score (eg, missing true positives due to human error) and manual chart review is impractical to do on a large number of documents. NLP can solve these issues. However, the performance of NLP algorithms is affected by the quality of EHR documentation (eg, presence, absence, and consistency of information required for delirium) because NLP algorithms are based on the records in EHR documents.

Error Analysis Most disagreements resulted from asymmetric annotation presence (ie, one annotator identified something the other annotator did not). In resolving these

disagreements, we determined that some represented under-annotation and others over-annotation due to human interpretation bias. We learned that asymmetric presence of annotations usually emerged due to unclear or inadequate extensional definition of the attribute to be annotated. The typical scenario was that an annotator did not realize that an entity should be annotated. For example, "paranoid thoughts" can be annotated as disorganized thinking, but such qualification was not apparent unless specifically named in the guidelines. The inconsistency could be rooted in differential interpretation or domain literacy, which were compensated by the iteratively refined guidelines through patching of inclusion criteria as annotators gained more experience.

We noticed that the accuracy of the algorithm was lower when classifying the delirium based on the modified CAM criteria compared with the original CAM criteria. This may be due to the introduction of the "possible" category. Our modified CAM changed the task from a 2-class to a 3-class classification problem. This implementation causes a relatively lower performance measure than NLP-CAM even though NLP-mCAM is able to identify more number of total delirium cases (definite and possible). During the NLP evaluation, we performed an error analysis to identify the most common causes of errors. We found that extracting "disorganized thinking" related concepts from clinical notes was a major challenge. The same concept can be expressed in many different ways. For example, the same concept can be expressed through different behaviors, such as speech, cognitive process, decision making, and patient reactions. Therefore, the NLP algorithms needed to be both accurate and generalizable in capturing these concepts. To address this problem, we used relaxed word distance, that is, allowing number of words between 2 anchor words, to help fine-tune the rules in multiple iterations (36). This process is, however, time-consuming. In the future, we will explore advanced machine learning techniques to capture the common meaning of the various expressions to aid the development effort.

Implications for Research The process of manually abstracting clinical concepts for delirium ascertainment is time-consuming, costly, and non-scalable. NLP algorithms are distinctive in their ability to extract critical information from free text in EHRs. NLP techniques offer a sophisticated way of handling free text with high levels of accuracy, allowing efficient mining of unstructured data for broad applications. The NLP-CAM

algorithm was developed strictly based on the original definition of CAM with the objective of achieving a high degree of precision. Researchers may find NLP-CAM to be helpful for identifying delirium with high confidence. However, we recognize that delirium symptoms are likely to be poorly documented in the clinical notes [198]. Thus, strictly applying the original definition may not be sensitive enough to capture highly likely or possible cases. We therefore modified the criteria definition to adapt to the real-world EHR. Instead of strictly following the CAM criteria, we developed a modified definition for definite cases and added an additional "possible" category. The modified NLP-mCAM algorithm also had a good performance in identifying definite delirium cases and also identified a significant number of possible cases. Depending on the research study, investigators may wish to include only definite cases or may want to include possible cases. The NLP-CAM and NLP-mCAM algorithms therefore offer investigators the flexibility to apply either algorithm depending on the needs of the study.

Implications for Clinical Practice The NLP algorithms also have many potential clinical applications especially when it comes to proactively identify patients at high risk of delirium based on prior history, or flagging hospitalized patients, who per clinical documentation are showing signs of delirium in real time. Because delirium is underreported and not all patients have a formal assessment for delirium diagnosis, the use of NLP algorithms on routine EHRs can facilitate the early detection of delirium. This can be achieved by integrating the NLP algorithms into clinical workflow through application programming interface technologies, which allows outputs from NLP to be delivered to clinicians by mobile applications or EHR system (eg, EPIC). Such clinical decision tools could facilitate the implementation of preventive measures to reduce the incidence of delirium (through risk factors) and institute early intervention strategies to avoid escalating symptoms and associated complications of delirium.

Limitations Our study has several limitations. The algorithms initially were developed after a review of clinical notes at a single institution, with the structure tailored to a specific EHR system. Although we have successfully demonstrated the external validity of other NLP algorithms on EHRs in another hospital setting [42, 202] (37,38), additional work is necessary to demonstrate the portability of these algorithms to other institutions and EHRs. The NLP algorithms do not intend to replace a formal delirium assessment. Instead, the NLP algorithms can be used to automate manual chart review

based on CAM criteria. The system was developed based on manual chart review from EHRs, the success of NLP algorithms depends on the level of detail and accuracy of the medical records. If there is no documentation of the features and concepts about delirium (hypoactive and hyperactive), NLP cannot identify the cases. We also note that hyperactive delirium is more likely to be documented in the medical records than hypoactive delirium. Therefore, we expect to consistently underestimate the presence of hypoactive delirium.

8.7 Conclusion

Guided by the TRUST process, we adopted the standardized evidence-based framework CAM to develop and evaluate NLP algorithms to identify the occurrence of delirium from EHRs. Our NLP algorithms demonstrated excellent performance in identifying patients with delirium using clinical notes in an expeditious and cost-effective manner. These algorithms represent a promising alternative to manual chart review used in EHR-based delirium research projects and artificial intelligence-based clinical decision support.

Chapter 9

Conclusion and Future Work

9.1 Conclusion

As EHR-based research becomes increasingly integrated into clinical care, it is important to ensure reproducible processes for NLP development. Our work proposed a TRUST (clinical **T**ext **R**etrieval and **U**se towards **S**cientific rigor and **T**ransparent) process that facilitates the assessment of the EHR-related latent heterogeneity and documentation of the provenance information that captures the process of the retrieval and organization of raw data as well as the extraction and annotation of training data. We first conceptualized the common barriers to reproducibility. We then conduct a real-world pragmatics assessment of the heterogeneity of institutional and EHR variance. Based on the case implementation and external literature evidence, we formulate an informatics (TRUST) process with best practice and reporting standards. Lastly, we conduct an end-to-end real-world NLP-assisted chart review study. The main contributions of this work are described below.

Reproducibility in Context of EHR Reproducibility is an important quality criterion for the secondary use of electronic health records (EHRs). However, multiple barriers to reproducibility are embedded in the heterogeneous EHR environment. Traditional methods of addressing research reproducibility are primarily focused on enhancing reporting standards or system specifications. The barriers to reproducibility can occur at all stages of information collection, extraction, transformation, and use. Unlike the

deterministic use of information, such as clinical trials, the objective of information collection and use for the secondary use of EHR data can be ad-hoc. To address that, we first conceptualized the problem and scope of reproducibility in the context of EHRs. We purpose to study EHRs from the perspective of a dynamic process instead of a static view. We thus introduced and defined the TRUST process to capture the nature of information exchange in EHR settings. The second challenge in the secondary use of EHR data is the heterogeneous nature of the problem. This includes various quality issues of data, implementation, documentation, and reporting. We tackle this by conceptualizing a multi-actor interaction model to capture the dynamic multi-level interactions occurring during information use (e.g., inter-personal, inter-system, and cross-institutional). The multi-actor interaction model combined with the TRUST process defines the scope of reproducibility measurements with the following two metrics: 1) feasibility of completing the original steps (i.e., implementation feasibility) and 2) variability of outcome (i.e., clinical outcome variability). We believe the proposed representation offers a holistic view of multi-level interactions in the information quality life cycle. This view can be used as a conceptual framework when considering reproducibility in the context of the secondary use of EHRs.

Heterogeneity assessment To better understand and mitigate study outcome variability caused by various heterogeneous factors, we proposed and applied a standardized set of informatics (qualitative and quantitative) methods to examine the heterogeneity of EHR systems, institutions, people, and processes. The first set of methods were focused on data quality, which includes data completeness, data variability, data timeliness, data validity, and data reliability. Specifically, our approach of assessing data quality is context-dependent and fitness for use. The second set of methods were focused on contextual variability. We studied four potential macro factors: EHR system, institution, people behavior, and process. To link these heterogeneous factors and issues of information quality, we adopted four unique dimensions of information quality: Intrinsic IQ, Accessibility IQ, Representational IQ, and Contextual IQ. Based on the IQ assessment and case study implementation, we discovered various barriers to reproducibility, such as inconsistent information documentation patterns across settings, information loss during ETL processes, and variable levels of information re- source accessibility.

Best practices and reporting standard Besides outcome variability, another important criterion of reproducibility is implementation feasibility. To address the current issue of lack of standard processes and best practices for conducting EHR-based studies, we followed a multi-phase process to produce and validate internal and external standards and evidence for developing best practices. Evidence was based on large-scale literature reviews and several real-world case studies based on multiple institutions. The proposed best practices cover an end-to-end TRUST process from the perspective of implementation quality, people-driven decision interpretation and making, documentation (data catalog and provenance), and project management. To address these issues of little-to-no informatics-related standards in existing clinical research ontologies and reporting standards, we develop novel NLP-assisted literature mining techniques to discover the reporting patterns and data abstraction methodologies for EHR-based clinical research. Our investigation discovered an upward trend of reporting EHR-related research methodologies, good practice, and the use of informatics-related methods. Despite that, the overall ratio of reporting/adoption of methodologic standards was still low. There was also a high variation regarding clinical research reporting. This investigation also produced a new set of reporting standards for EHR-based Chart Review Research (Table 7.8)

9.2 Limitations

This work has several limitations. First, our study aims to propose a standard set of informatics methods for leveraging EHRs for clinical research. The generalizability of these methods is limited by the scope of the case studies (e.g., number of institutions involved) as well as existing evidence from the literature. The unstructured EHRs although serve as valuable sources for research, there are inherent limitations in the nature of the data source. For example, the missing documentation was discovered when detecting SBIs cases from neuroimages compared with neuroimaging reports. Thirdly, although NLP is a powerful solution for high throughput clinical concept extraction, it is still very expensive to develop high-quality NLP solutions. Re-using existing NLP systems may also suffer portability issues. More cost-effective and 'ecologically efficient' strategies are needed for future clinical NLP research.

9.3 Future Work

Building on top of this work, we have identified the following potential future directions to continuously ensure the validity, reliability, and reproducibility of research findings

Living evidence synthesis A novel approach to further enhance and expand upon our existing methods of generating best practices is to incorporate a living interactive evidence synthesis framework to present 'live' evidence of current practice and reporting standards of EHR-based chart review research. The framework will incorporate natural language processing techniques, ontology, common data elements, and human-in-the-loop concepts to achieve systematic extracting and real-time management of reproducible best practices for the secondary use of EHRs.

Community engagement and knowledge dissemination The dissemination of the proposed framework and best practices is crucial to achieving community-level awareness and good practices in the secondary use of EHRs for clinical research. We thus plan to seek potential collaboration with leverage existing multi-center consortiums such as All of Us, OHNLP, N3C, CD2H for promoting the development and adoption of best practices for trusted science advancement and discovery in the informatics community. With the support from these collaboration centers, we can prospectively implement our framework into EHR-based observational studies to compare the effectiveness. The newly generated evidence will be further synthesized for developing a new set of guidelines.

Tooling Existing best practices can be integrated into tools to enhance the standardization and provenance of ETL processes. The ultimate comparability and consistency of clinical data sets derived from heterogeneous clinical data sources can be enhanced by the adoption of existing and emerging standards such as standard formats, common data elements, ontology, and controlled vocabularies. The traceability, validity, and reproducibility of clinical data sets can be enhanced by the detailed documentation and logging of the ETL process. Built upon our ongoing informatics effort to support data management and digital curation, we will develop a suite of digital curation methods and tools focusing on the standardization and provenance of ETL processes.

References

[1] Annetine C Gelijns and Sherine E Gabriel. Looking beyond translation—integrating clinical research with medical practice. *N Engl J Med*, 366(18):1659–61, 2012.

[2] MD Arnold Milstein. Code red and blue–safely limiting health care's gdp footprint. *The New England journal of medicine*, 368(1):1, 2013.

[3] C. P. Friedman, A. K. Wong, and D. Blumenthal. Achieving a nationwide learning health system. *Sci Transl Med*, 2(57):57cm29, 2010.

[4] C. P. Friedman, J. C. Rubin, and K. J. Sullivan. Toward an information infrastructure for global health improvement. *Yearb Med Inform*, 26(1):16–23, 2017.

[5] Bright I Nwaru, Charles Friedman, John Halamka, and Aziz Sheikh. Can learning health systems help organisations deliver personalised care? *BMC medicine*, 15(1):1–8, 2017.

[6] Vinod C Kaggal, Ravikumar Komandur Elayavilli, Saeed Mehrabi, Joshua J Pankratz, Sunghwan Sohn, Yanshan Wang, Dingcheng Li, Majid Mojarad Rastegar, Sean P Murphy, Jason L Ross, et al. Toward a learning health-care system–knowledge delivery at the point of care empowered by big data and nlp. *Biomedical informatics insights*, 8:BII–S37977, 2016.

[7] Sunyang Fu, Lester Y Leung, Anne-Olivia Raulli, David F Kallmes, Kristin A Kinsman, Kristoff B Nelson, Michael S Clark, Patrick H Luetmer, Paul R Kingsbury, and David M Kent. Assessment of the impact of ehr heterogeneity for clinical

research through a case study of silent brain infarction. *BMC medical informatics and decision making*, 20:1–12, 2020.

[8] Marguerite VB Dresser, Lisa Feingold, Susan L Rosenkranz, and Kathryn L Coltin. Clinical quality measurement: comparing chart review and automated methodologies. *Medical care*, pages 539–552, 1997.

[9] Genevieve B Melton and George Hripcsak. Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association*, 12(4):448–457, 2005.

[10] R. E. Gearing, I. A. Mian, J. Barber, and A. Ickowicz. A methodology for conducting retrospective chart review research in child and adolescent psychiatry. *J Can Acad Child Adolesc Psychiatry*, 15(3):126–34, 2006.

[11] J. Frankovich, C. A. Longhurst, and S. M. Sutherland. Evidence-based medicine in the emr era. *N Engl J Med*, 365(19):1758–9, 2011.

[12] E. H. Gilbert, S. R. Lowenstein, J. Koziol-McLain, D. C. Barta, and J. Steiner. Chart reviews in emergency medicine research: Where are the methods? *Ann Emerg Med*, 27(3):305–8, 1996.

[13] Ralph Grishman, Silja Huttunen, and Roman Yangarber. Information extraction for enhanced access to disease outbreak reports. *Journal of biomedical informatics*, 35(4):236–246, 2002.

[14] H. Xu, M. Jiang, M. Oetjens, E. A. Bowton, A. H. Ramirez, J. M. Jeff, M. A. Basford, J. M. Pulley, J. D. Cowan, X. M. Wang, M. D. Ritchie, D. R. Masys, D. M. Roden, D. C. Crawford, and J. C. Denny. Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin. *Journal of the American Medical Informatics Association*, 18(4):387–391, 2011.

[15] Jim Cowie and Wendy Lehnert. Information extraction. *Communications of the ACM*, 39(1):80–91, 1996.

[16] Y. Wang, K. Zheng, H. Xu, and Q. Mei. Clinical word sense disambiguation with interactive search and classification. *AMIA Annu Symp Proc*, 2016:2062–2071, 2016.

[17] Geoffrey Leech. Corpus annotation schemes. *Literary and linguistic computing*, 8(4):275–281, 1993.

[18] Diego Mollá and María Elena Santiago-Martínez. Creation of a corpus for evidence based medicine summarisation. *The Australasian medical journal*, 5(9):503, 2012.

[19] Jiayi Tong, Jing Huang, Jessica Chubak, Xuan Wang, Jason H Moore, Rebecca A Hubbard, and Yong Chen. An augmented estimation procedure for ehr-based association studies accounting for differential misclassification. *Journal of the American Medical Informatics Association*, 27(2):244–253, 2020.

[20] Nicole Gray Weiskopf and Chunhua Weng. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1):144–151, 2013.

[21] K Bretonnel Cohen, Jingbo Xia, Christophe Roeder, and Lawrence E Hunter. Reproducibility in natural language processing: a case study of two r libraries for mining pubmed/medline. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation*, volume 2016, page 6. NIH Public Access, 2016.

[22] Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'avolio, Guergana K Savova, and Ozlem Uzuner. Overcoming barriers to nlp for clinical text: the role of shared tasks and the need for additional creative solutions, 2011.

[23] David Madigan, Patrick B Ryan, Martijn Schuemie, Paul E Stang, J Marc Overhage, Abraham G Hartzema, Marc A Suchard, William DuMouchel, and Jesse A Berlin. Evaluating the impact of database heterogeneity on observational study results. *American journal of epidemiology*, 178(4):645–651, 2013.

[24] Nicole G Weiskopf, Suzanne Bakken, George Hripcsak, and Chunhua Weng. A data quality assessment guideline for electronic health record data reuse. *Egems*, 5(1), 2017.

[25] What is an electronic health record (ehr)? `https://www.healthit.gov/faq/what-electronic-health-record-ehr`. Accessed: 2021-08-29.

[26] Spiros Denaxas, Arturo Gonzalez-Izquierdo, Maria Pikoula, Kenan Direk, Natalie Fitzpatrick, Harry Hemingway, and LIam Smeeth. Methods for enhancing the reproducibility of observational research using electronic health records: preliminary findings from the caliber resource. In *2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 506–508. IEEE, 2017.

[27] Meredith N Zozus, Rachel L Richesson, Anita Walden, Jessie D Tenenbaum, and William Edward Hammond. Research reproducibility in longitudinal multi-center studies using data from electronic health records. *AMIA Summits on Translational Science Proceedings*, 2016:279, 2016.

[28] Steven G Johnson, Stuart Speedie, Gyorgy Simon, Vipin Kumar, and Bonnie L Westra. A data quality ontology for the secondary use of ehr data. In *AMIA Annual Symposium Proceedings*, volume 2015, page 1937. American Medical Informatics Association, 2015.

[29] R Stanley Hum and Samantha Kleinberg. Replicability, reproducibility, and agent-based simulation of interventions. In *AMIA Annual Symposium Proceedings*, volume 2017, page 959. American Medical Informatics Association.

[30] A. Wen, S. Fu, S. Moon, M. El Wazir, A. Rosenbaum, V. C. Kaggal, S. Liu, S. Sohn, H. Liu, and J. Fan. Desiderata for delivering nlp to accelerate healthcare ai advancement and a mayo clinic nlp-as-a-service implementation. *NPJ Digit Med*, 2(1):130, 2019.

[31] A. Stubbs, C. Kotfila, H. Xu, and O. Uzuner. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/uthealth shared task track 2. *J Biomed Inform*, 58 Suppl:S67–77, 2015.

[32] Ewoud Pons, Loes MM Braun, MG Myriam Hunink, and Jan A Kors. Natural language processing in radiology: a systematic review. *Radiology*, 279(2):329–343, 2016.

[33] C. C. Wyles, M. E. Tibbo, S. Y. Fu, Y. S. Wang, S. Sohn, W. K. Kremers, D. J. Berry, D. G. Lewallen, and H. Maradit-Kremers. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. *Journal of Bone and Joint Surgery-American Volume*, 101(21):1931–1938, 2019.

[34] Gerard Burger, Ameen Abu-Hanna, Nicolette de Keizer, and Ronald Cornet. Natural language processing in pathology: a scoping review. *Journal of clinical pathology*, 69(11):949–955, 2016.

[35] Sunyang Fu, Cody C Wyles, Douglas R Osmon, Martha L Carvour, Elham Sagheb, Taghi Ramazanian, Walter K Kremers, David G Lewallen, Daniel J Berry, Sunghwan Sohn, et al. Automated detection of periprosthetic joint infections and data elements using natural language processing. *The Journal of Arthroplasty*, 36(2):688–692, 2021.

[36] Hongfang Liu, Suzette J Bielinski, Sunghwan Sohn, Sean Murphy, Kavishwar B Wagholikar, Siddhartha R Jonnalagadda, KE Ravikumar, Stephen T Wu, Iftikhar J Kullo, and Christopher G Chute. An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings*, 2013:149, 2013.

[37] C. Friedman, P. O. Alderson, J. H. Austin, J. J. Cimino, and S. B. Johnson. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc*, 1(2):161–74, 1994.

[38] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *J Am Med Inform Assoc*, 17(5):507–13, 2010.

[39] A. R. Aronson and F. M. Lang. An overview of metamap: historical perspective and recent advances. *J Am Med Inform Assoc*, 17(3):229–36, 2010.

[40] Joshua C Denny, Plomarz R Irani, Firas H Wehbe, Jeffrey D Smithers, and Anderson Spickard III. The knowledgemap project: development of a concept-based medical school curriculum database. In *AMIA Annual Symposium Proceedings*, volume 2003, page 195. American Medical Informatics Association.

[41] Qing T Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N Murphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics and decision making*, 6(1):1–9, 2006.

[42] Sunghwan Sohn, Yanshan Wang, Chung-Il Wi, Elizabeth A Krusemark, Euijung Ryu, Mir H Ali, Young J Juhn, and Hongfang Liu. Clinical documentation variations and nlp system portability: a case study in asthma birth cohorts across institutions. *Journal of the American Medical Informatics Association*, 25(3):353–359, 2018.

[43] Saeed Mehrabi, Anand Krishnan, Alexandra M Roch, Heidi Schmidt, DingCheng Li, Joe Kesterson, Chris Beesley, Paul Dexter, Max Schmidt, Mathew Palakal, et al. Identification of patients with family history of pancreatic cancer-investigation of an nlp system portability. *Studies in health technology and informatics*, 216:604, 2015.

[44] Nicole G Weiskopf, George Hripcsak, Sushmita Swaminathan, and Chunhua Weng. Defining and measuring completeness of electronic health records for secondary use. *Journal of biomedical informatics*, 46(5):830–836, 2013.

[45] Efthymios Chondrogiannis, Vassiliki Andronikou, Anastasios Tagaris, Efstathios Karanastasis, Theodora Varvarigou, and Masatsugu Tsuji. A novel semantic representation for eligibility criteria in clinical trials. *Journal of biomedical informatics*, 69:10–23, 2017.

[46] Zhisheng Huang, Frank van Harmelen, Annette ten Teije, and Kathrin Dentler. Knowledge-based patient data generation. In *Process support and knowledge representation in health care*, pages 83–96. Springer, 2013.

[47] Anouchka Seesaghur, Natalia Petruski-Ivleva, Victoria Banks, Jocelyn Ruoyi Wang, Pattra Mattox, Edwin Hoeben, Joe Maskell, David Neasham, Shannon L Reynolds, and George Kafatos. Real-world reproducibility study characterizing patients newly diagnosed with multiple myeloma using clinical practice research datalink, a uk-based electronic health records database. *Pharmacoepidemiology and drug safety*, 30(2):248–256, 2021.

[48] W. W. Yim, A. J. Wheeler, C. Curtin, T. H. Wagner, and T. Hernandez-Boussard. Secondary use of electronic medical records for clinical research: Challenges and opportunities. *Converg Sci Phys Oncol*, 4(1), 2018.

[49] Steve G Langer, George Shih, Paul Nagy, and Bennet A Landman. Collaborative and reproducible research: goals, challenges, and strategies. *Journal of digital imaging*, 31(3):275–282, 2018.

[50] Martijn J Schuemie, Patrick B Ryan, Nicole Pratt, RuiJun Chen, Seng Chan You, Harlan M Krumholz, David Madigan, George Hripcsak, and Marc A Suchard. Principles of large-scale evidence generation and evaluation across a network of databases (legend). *Journal of the American Medical Informatics Association*, 27(8):1331–1337, 2020.

[51] Marilyn Chow, Murielle Beene, Ann O'Brien, Patricia Greim, Tim Cromwell, Donna DuLong, and Diane Bedecarré. A nursing information model process for interoperability. *Journal of the American Medical Informatics Association*, 22(3):608–614, 2015.

[52] Alan H Morris, James Orme Jr, Jonathon D Truwit, Jay Steingrub, Colin Grissom, Kang H Lee, Guoliang L Li, B Taylor Thompson, Roy Brower, Mark Tidswell, et al. A replicable method for blood glucose control in critically ill patients. *Critical care medicine*, 36(6):1787–1795, 2008.

[53] Michael F Chiang, John C Hwang, C Yu Alexander, Daniel S Casper, James J Cimino, and Justin Starren. Reliability of snomed-ct coding by three physicians using two terminology browsers. In *AMIA Annual Symposium Proceedings*, volume 2006, page 131. American Medical Informatics Association.

[54] Lester Y Leung, Sunyang Fu, Patrick H Luetmer, David F Kallmes, Neel Madan, Gene Weinstein, Vance T Lehman, Charlotte H Rydberg, Jason Nelson, Hongfang Liu, et al. Agreement between neuroimages and reports for natural language processing-based detection of silent brain infarcts and white matter disease. *BMC neurology*, 21(1):1–5, 2021.

[55] Monya Baker. 1,500 scientists lift the lid on reproducibility. *Nature News*, 533(7604):452, 2016.

[56] Jon F Claerbout and Martin Karrenbach. Electronic documents give reproducible research a new meaning. In *SEG technical program expanded abstracts 1992*, pages 601–604. Society of Exploration Geophysicists, 1992.

[57] Hans E Plesser. Reproducibility vs. replicability: a brief history of a confused terminology. *Frontiers in neuroinformatics*, 11:76, 2018.

[58] B Rous. The acm task force on data, software, and reproducibility in publication, 2018.

[59] Steven N Goodman, Daniele Fanelli, and John PA Ioannidis. What does research reproducibility mean? *Science translational medicine*, 8(341):341ps12–341ps12, 2016.

[60] L. D. McIntosh, A. Juehne, C. R. H. Vitale, X. Liu, R. Alcoser, J. C. Lukas, and B. Evanoff. Repeat: a framework to assess empirical reproducibility in biomedical research. *BMC Med Res Methodol*, 17(1):143, 2017.

[61] Tiffany C Veinot, Charles R Senteio, David Hanauer, and Julie C Lowery. Comprehensive process model of clinical information interaction in primary care: results of a "best-fit" framework synthesis. *Journal of the American Medical Informatics Association*, 25(6):746–758, 2018.

[62] Association for Computing Machinery. Artifact review and badging, 2016.

[63] National Academies of Sciences Engineering, Medicine, et al. *Reproducibility and replicability in science*. National Academies Press, 2019.

[64] Peter D Stetson, Frances P Morrison, Suzanne Bakken, Stephen B Johnson, and eNote Research Team. Preliminary development of the physician documentation quality instrument. *Journal of the American Medical Informatics Association*, 15(4):534–541, 2008.

[65] Kathy Giannangelo. Making the connection between standard terminologies, use cases, and mapping. *Health Information Management Journal*, 35(3):8–12, 2006.

[66] Marilyn Chow, Murielle Beene, Ann O'Brien, Patricia Greim, Tim Cromwell, Donna DuLong, and Diane Bedecarré. A nursing information model process for interoperability. *Journal of the American Medical Informatics Association*, 22(3):608–614, 2015.

[67] Xuequn Pan and James J Cimino. Locating relevant patient information in electronic health record data using representations of clinical concepts and database structures. In *AMIA Annual Symposium Proceedings*, volume 2014, page 969. American Medical Informatics Association, 2014.

[68] Alistair EW Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2018.

[69] Ligang Luo, Liping Li, Jiajia Hu, Xiaozhe Wang, Boulin Hou, Tianze Zhang, and Lue Ping Zhao. A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. *BMC medical informatics and decision making*, 16(1):1–14, 2016.

[70] Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934, 2020.

[71] Steve Harris, Sinan Shi, David Brealey, Niall S MacCallum, Spiros Denaxas, David Perez-Suarez, Ari Ercole, Peter Watkinson, Andrew Jones, and Simon Ashworth. Critical care health informatics collaborative (cchic): Data, tools and methods for reproducible research: A multi-centre uk intensive care database. *International journal of medical informatics*, 112:82–89, 2018.

[72] Hossein Estiri, Kari A Stephens, Jeffrey G Klann, and Shawn N Murphy. Exploring completeness in clinical data research networks with dqe-c. *Journal of the American Medical Informatics Association*, 25(1):17–24, 2018.

[73] Hossein Estiri and Kari Stephens. Dqe-v: a database-agnostic framework for exploring variability in electronic health record data across time and site location. *eGEMs*, 5(1), 2017.

[74] Dingcheng Li, Cory M Endle, Sahana Murthy, Craig Stancl, Dale Suesse, Davide Sottara, Stanley M Huff, Christopher G Chute, and Jyotishman Pathak. Modeling and executing electronic health records driven phenotyping algorithms using the nqf quality data model and jboss® drools engine. In *AMIA annual symposium proceedings*, volume 2012, page 532. American Medical Informatics Association, 2012.

[75] David A Springate, Evangelos Kontopantelis, Darren M Ashcroft, Ivan Olier, Rosa Parisi, Edmore Chamapiwa, and David Reeves. Clinicalcodes: an online clinical codes repository to improve the validity and reproducibility of research using electronic medical records. *PloS one*, 9(6):e99825, 2014.

[76] Selen Bozkurt, Eli M Cahan, Martin G Seneviratne, Ran Sun, Juan A Lossio-Ventura, John PA Ioannidis, and Tina Hernandez-Boussard. Reporting of demographic data and representativeness in machine learning models using electronic health records. *Journal of the American Medical Informatics Association*, 27(12):1878–1884, 2020.

[77] Mark J Pletcher, Valerie Flaherman, Nader Najafi, Sajan Patel, Robert J Rushakoff, Ari Hoffman, Andrew Robinson, Russell J Cucina, Charles E McCulloch, Ralph Gonzales, et al. Randomized controlled trials of electronic health

record interventions: design, conduct, and reporting considerations. *Annals of Internal Medicine*, 172(11_Supplement):S85–S91, 2020.

[78] Lucinda S Orsini, Marc Berger, William Crown, Gregory Daniel, Hans-Georg Eichler, Wim Goettsch, Jennifer Graff, John Guerino, Pall Jonsson, Nirosha Mahendraratnam Lederer, et al. Improving transparency to build trust in real-world secondary data studies for hypothesis testing—why, what, and how: recommendations and a road map from the real-world evidence transparency initiative. *Value in Health*, 23(9):1128–1136, 2020.

[79] Yiqing Zhao, Yanshan Wang, Henry Wang, Benjamin Yan, Feichen Shen, Kevin J Peterson, Walter A Rocca, Jennifer St Sauver, and Hongfang Liu. Annotating cohort data elements with ohdsi common data model to promote research reproducibility. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1310–1317. IEEE, 2018.

[80] Adam Wright, Shobha Phansalkar, Meryl Bloomrosen, Robert A Jenders, Anne M Bobb, John D Halamka, Gilad Kuperman, Thomas H Payne, Sheila Teasdale, Allen J Vaida, et al. Best practices in clinical decision support. *Applied clinical informatics*, 1(03):331–345, 2010.

[81] Shirley V Wang, Olga V Patterson, Joshua J Gagne, Jeffrey S Brown, Robert Ball, Pall Jonsson, Adam Wright, Li Zhou, Wim Goettsch, and Andrew Bate. Transparent reporting on research using unstructured electronic health record data to generate 'real world'evidence of comparative effectiveness and safety. *Drug safety*, 42(11):1297–1309, 2019.

[82] William Digan, Aurélie Névéol, Antoine Neuraz, Maxime Wack, David Baudoin, Anita Burgun, and Bastien Rance. Can reproducibility be improved in clinical natural language processing? a study of 7 clinical nlp suites. *Journal of the American Medical Informatics Association*, 28(3):504–515, 2021.

[83] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald. What can natural language processing do for clinical decision support? *J Biomed Inform*, 42(5):760–72, 2009.

[84] Jia Xu, Pengwei Yang, Shang Xue, Bhuvan Sharma, Marta Sanchez-Martin, Fang Wang, Kirk A Beaty, Elinor Dehan, and Baiju Parikh. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. *Human genetics*, 138(2):109–124, 2019.

[85] Taxiarchis Botsis, Gunnar Hartvigsen, Fei Chen, and Chunhua Weng. Secondary use of ehr: data quality issues and informatics opportunities. *Summit on Translational Bioinformatics*, 2010:1, 2010.

[86] Genna R Cohen, Charles P Friedman, Andrew M Ryan, Caroline R Richardson, and Julia Adler-Milstein. Variation in physicians' electronic health record documentation and potential patient harm from that variation. *Journal of general internal medicine*, 34(11):2355–2367, 2019.

[87] Rui Zhang, Serguei Pakhomov, Bridget T McInnes, and Genevieve B Melton. Evaluating measures of redundancy in clinical texts. In *AMIA annual symposium proceedings*, volume 2011, page 1612. American Medical Informatics Association, 2011.

[88] Clifford Y Ko, Bruce L Hall, Amy J Hart, Mark E Cohen, David B Hoyt, et al. The american college of surgeons national surgical quality improvement program: achieving better and safer surgery. *The Joint Commission Journal on Quality and Patient Safety*, 41(5):199–AP1, 2015.

[89] Angela M Ingraham, Karen E Richards, Bruce L Hall, and Clifford Y Ko. Quality improvement in surgery: the american college of surgeons national surgical quality improvement program approach. *Advances in surgery*, 44(1):251–267, 2010.

[90] Joseph M Juran and Joseph A De Feo. *Juran's quality handbook: the complete guide to performance excellence*. McGraw-Hill Education, 2010.

[91] Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[92] Robert H Dolin, Liora Alschuler, Sandy Boyer, Calvin Beebe, Fred M Behlen, Paul V Biron, and Amnon Shabo. Hl7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association*, 13(1):30–39, 2006.

[93] Peter L Elkin, David Froehling, Brent A Bauer, Dietlind Wahner-Roedler, STrent Rosenbloom, Kent Bailey, Steven H Brown, et al. Aequus communis sententia: defining levels of interoperability. In *Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems*, page 725. IOS Press, 2007.

[94] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[95] Paul E Black. Ratcliff/obershelp pattern recognition. *Dictionary of algorithms and data structures*, 17, 2004.

[96] Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. Medsts: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72, 2020.

[97] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001.

[98] David Loshin. *Master data management*. Morgan Kaufmann, 2010.

[99] David Chen, Naveed Afzal, Sunghwan Sohn, Elizabeth B Habermann, James M Naessens, David W Larson, and Hongfang Liu. Postoperative bleeding risk prediction for patients undergoing colorectal surgery. *Surgery*, 164(6):1209–1216, 2018.

[100] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[101] Lester Y Leung, Paul KJ Han, Christine Lundquist, Gene Weinstein, David E Thaler, and David Kent. Clinicians' perspectives on incidentally discovered silent brain infarcts–a qualitative study. *PloS one*, 13(3):e0194971, 2018.

[102] Sarah E Vermeer, William T Longstreth Jr, and Peter J Koudstaal. Silent brain infarcts: a systematic review. *The Lancet Neurology*, 6(7):611–619, 2007.

[103] Jonathon P Fanning, Allan J Wesley, Andrew A Wong, and John F Fraser. Emerging spectra of silent brain infarction. *Stroke*, 45(11):3461–3471, 2014.

[104] Jonathon P Fanning, Andrew A Wong, and John F Fraser. The epidemiology of silent brain infarction: a systematic review of population-based cohorts. *BMC medicine*, 12(1):1–11, 2014.

[105] Yaojing Chen, Ailin Wang, Jinfu Tang, Dongfeng Wei, Peng Li, Kewei Chen, Yongyan Wang, and Zhanjun Zhang. Association of white matter integrity and cognitive functions in patients with subcortical silent lacunar infarcts. *Stroke*, 46(4):1123–1126, 2015.

[106] John Conklin, Frank L Silver, David J Mikulis, and Daniel M Mandell. Are acute infarcts the cause of leukoaraiosis? brain mapping for 16 consecutive weeks. *Annals of neurology*, 76(6):899–904, 2014.

[107] John Aberdeen, Samuel Bayer, Reyyan Yeniterzi, Ben Wellner, Cheryl Clark, David Hanauer, Bradley Malin, and Lynette Hirschman. The mitre identification scrubber toolkit: design, training, and assessment. *International journal of medical informatics*, 79(12):849–859, 2010.

[108] S. Liu, H. Liu, V. Chaudhary, and D. Li. An infinite mixture model for coreference resolution in clinical notes. *AMIA Jt Summits Transl Sci Proc*, 2016:428–37, 2016.

[109] Amber Stubbs. Mae and mai: lightweight annotation and adjudication tools. In *Proceedings of the 5th linguistic annotation workshop*, pages 129–133, 2011.

[110] Jacob %J Educational Cohen and psychological measurement. A coefficient of agreement for nominal scales. 20(1):37–46, 1960.

[111] Cornelis Joost Van Rijsbergen. *The geometry of information retrieval*. Cambridge University Press, 2004.

[112] Karen Holtzblatt, Jessamyn Burns Wendell, and Shelley Wood. *Rapid contextual design: a how-to guide to key techniques for user-centered design*. Elsevier, 2004.

[113] Geir Inge Hausvik, Devinder Thapa, and Bjørn Erik Munkvold. Information quality life cycle in secondary use of ehr data. *International Journal of Information Management*, 56:102227, 2021.

[114] Liping Liu and Lauren Chi. Evolutional data quality: A theory-specific view. In *ICIQ*, pages 292–304, 2002.

[115] Kathrin M Cresswell, Allison Worth, and Aziz Sheikh. Actor-network theory and its role in understanding the implementation of information technology developments in healthcare. *BMC medical informatics and decision making*, 10(1):1–11, 2010.

[116] Yang W Lee, Diane M Strong, Beverly K Kahn, and Richard Y Wang. Aimq: a methodology for information quality assessment. *Information & management*, 40(2):133–146, 2002.

[117] Janet E Olson, Euijung Ryu, Kiley J Johnson, Barbara A Koenig, Karen J Maschke, Jody A Morrisette, Mark Liebow, Paul Y Takahashi, Zachary S Fredericksen, Ruchi G Sharma, et al. The mayo clinic biobank: a building block for individualized medicine. In *Mayo Clinic Proceedings*, volume 88, pages 952–962. Elsevier, 2013.

[118] Sharon K Inouye, Rudi GJ Westendorp, and Jane S Saczynski. Delirium in elderly people. *The Lancet*, 383(9920):911–922, 2014.

[119] Simone RF Ritter, Anne F Cardoso, Marina MP Lins, Thayana LV Zoccoli, Marco Polo D Freitas, and Einstein F Camargos. Underdiagnosis of delirium in the elderly in acute care hospital settings: lessons not learned. *Psychogeriatrics*, 18(4):268–275, 2018.

[120] Sunyang Fu, Guilherme S Lopes, Sandeep R Pagali, Bjoerg Thorsteinsdottir, Nathan K LeBrasseur, Andrew Wen, Hongfang Liu, Walter A Rocca, Janet E Olson, and Jennifer St Sauver. Ascertainment of delirium status using natural language processing from electronic health records. *The Journals of Gerontology: Series A*, 2020.

[121] S. N. Murphy, G. Weber, M. Mendis, V. Gainer, H. C. Chueh, S. Churchill, and I. Kohane. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc*, 17(2):124–30, 2010.

[122] Ashish K Jha, Gilad J Kuperman, Jonathan M Teich, Lucian Leape, Brian Shea, Eve Rittenberg, Elisabeth Burdick, Diane Lew Seger, Martha Vander Vliet, and David W Bates. Identifying adverse drug events: development of a computer-based monitor and comparison with chart review and stimulated voluntary report. *Journal of the American Medical Informatics Association*, 5(3):305–314, 1998.

[123] David W Baker, Stephen D Persell, Jason A Thompson, Neilesh S Soman, Karen M Burgner, David Liss, and Karen S Kmetik. Automated review of electronic health records to assess quality of care for outpatients with heart failure. *Annals of internal medicine*, 146(4):270–277, 2007.

[124] M Weiner, TE Stump, CM Callahan, JN Lewis, and CJ McDonald. Pursuing integration of performance measures into electronic medical records: beta-adrenergic receptor antagonist medications. *BMJ Quality & Safety*, 14(2):99–106, 2005.

[125] Vassar Matt and Holzmann Matthew. The retrospective chart review: important methodological considerations. *Journal of educational evaluation for health professions*, 10, 2013.

[126] David Baker, Katie Lidster, Ana Sottomayor, and Sandra Amor. reporting standards fall short. *Nature*, 492(7427):41–41, 2012.

[127] C Glenn Begley, Alastair M Buchan, and Ulrich Dirnagl. Institutions must do their part for reproducibility: the funding to verified good institutional practice, and robust science will shoot up the agenda. *Nature*, 525(7567):25–28, 2015.

[128] Arjun K Manrai, Chirag J Patel, Nils Gehlenborg, Nicholas P Tatonetti, John PA Ioannidis, and Isaac S Kohane. Methods to enhance the reproducibility of precision medicine. In *Biocomputing 2016: Proceedings of the Pacific Symposium*, pages 180–182. World Scientific, 2016.

[129] Stephen D Persell, Jennifer M Wright, Jason A Thompson, Karen S Kmetik, and David W Baker. Assessing the validity of national quality measures for coronary

artery disease using an electronic health record. *Archives of Internal Medicine*, 166(20):2272–2277, 2006.

[130] Amanda Parsons, Colleen McCullough, Jason Wang, and Sarah Shih. Validity of electronic health record-derived quality measurement for performance monitoring. *Journal of the American Medical Informatics Association*, 19(4):604–609, 2012.

[131] Karen S Kmetik, Michael F O'Toole, Heidi Bossley, Carmen A Brutico, Gary Fischer, Sherry L Grund, Bridget M Gulotta, Mark Hennessey, Stasia Kahn, Karen M Murphy, et al. Exceptions to outpatient quality measures for coronary artery disease in electronic health records. *Annals of internal medicine*, 154(4):227–234, 2011.

[132] Eve A Kerr, Dylan M Smith, Mary M Hogan, Sarah L Krein, Leonard Pogach, Timothy P Hofer, and Rodney A Hayward. Comparing clinical automated, medical record, and hybrid data sources for diabetes quality measures. *The Joint Commission journal on quality improvement*, 28(10):555–565, 2002.

[133] K. Wagholikar, M. Torii, S. Jonnalagadda, and H. Liu. Feasibility of pooling annotated corpora for clinical concept extraction. *AMIA Jt Summits Transl Sci Proc*, 2012:38, 2012.

[134] Richard Eckart De Castilho and Iryna Gurevych. A broad-coverage collection of portable nlp components for building shareable analysis pipelines. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT*, pages 1–11, 2014.

[135] Jeffrey P Ferraro, Ye Ye, Per H Gesteland, Peter J Haug, Fuchiang Tsui, Gregory F Cooper, Rudy Van Bree, Thomas Ginter, Andrew J Nowalk, and Michael Wagner. The effects of natural language processing on cross-institutional portability of influenza case detection for disease surveillance. *Applied clinical informatics*, 8(02):560–580, 2017.

[136] Mei Liu, Anushi Shah, Min Jiang, Neeraja B Peterson, Qi Dai, Melinda C Aldrich, Qingxia Chen, Erica A Bowton, Hongfang Liu, Joshua C Denny, et al. A study

of transportability of an existing smoking status detection module across institutions. In *AMIA Annual Symposium Proceedings*, volume 2012, page 577. American Medical Informatics Association, 2012.

[137] Robert J Carroll, Will K Thompson, Anne E Eyler, Arthur M Mandelin, Tianxi Cai, Raquel M Zink, Jennifer A Pacheco, Chad S Boomershine, Thomas A Lasko, Hua Xu, et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1):e162–e169, 2012.

[138] G. K. Savova, W. W. Chapman, J. Zheng, and R. S. Crowley. Anaphoric relations in the clinical narrative: corpus creation. *J Am Med Inform Assoc*, 18(4):459–65, 2011.

[139] D. Albright, A. Lanfranchi, A. Fredriksen, W. F. th Styler, C. Warner, J. D. Hwang, J. D. Choi, D. Dligach, R. D. Nielsen, J. Martin, W. Ward, M. Palmer, and G. K. Savova. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *J Am Med Inform Assoc*, 20(5):922–30, 2013.

[140] Brett R South, Shuying Shen, Makoto Jones, Jennifer Garvin, Matthew H Samore, Wendy W Chapman, and Adi V Gundlapalli. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. In *BMC bioinformatics*, volume 10, page S12. BioMed Central.

[141] William Scuba, Melissa Tharp, Danielle Mowery, Eugene Tseytlin, Yang Liu, Frank A Drews, and Wendy W Chapman. Knowledge author: facilitating user-driven, domain content development to support clinical information extraction. *Journal of biomedical semantics*, 7(1):1–11, 2016.

[142] James Masanz, Serguei V Pakhomov, Hua Xu, Stephen T Wu, Christopher G Chute, and Hongfang Liu. Open source clinical nlp–more than any single system. *AMIA Summits on Translational Science Proceedings*, 2014:76, 2014.

[143] Lawrence M Friedman, Curt Furberg, David L DeMets, David Reboussin, and Christopher B Granger. *Fundamentals of clinical trials*, volume 3. Springer, 1998.

[144] Carly Strasser. Research data management. *National Information Standards Organization*, 2015.

[145] S. T. Wu, C. I. Wi, S. Sohn, H. Liu, and Y. J. Juhn. Staggered nlp-assisted refinement for clinical annotations of chronic disease events. In N. Calzolari, K. Choukri, H. Mazo, A. Moreno, T. Declerck, S. Goggi, M. Grobelnik, J. Odijk, S. Piperidis, B. Maegaard, and J. Mariani, editors, *10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 426–429. European Language Resources Association (ELRA).

[146] H. S. Chase, L. R. Mitrani, G. G. Lu, and D. J. Fulgieri. Early recognition of multiple sclerosis using natural language processing of the electronic health record. *BMC Med Inform Decis Mak*, 17(1):24, 2017.

[147] T. Chen, M. Dredze, J. P. Weiner, L. Hernandez, J. Kimura, and H. Kharrazi. Extraction of geriatric syndromes from electronic health record clinical notes: Assessment of statistical natural language processing methods. *JMIR Med Inform*, 7(1):e13039, 2019.

[148] Jeanmarie Mayer, Shuying Shen, Brett R South, Stephane Meystre, F Jeff Friedlin, William R Ray, and Matthew Samore. Inductive creation of an annotation schema and a reference standard for de-identification of va electronic clinical notes. In *AMIA Annual Symposium Proceedings*, volume 2009, page 416. American Medical Informatics Association.

[149] Shawn N Murphy, Michael E Mendis, David A Berkowitz, Isaac Kohane, and Henry C Chueh. Integration of clinical and genetic data in the i2b2 architecture. In *AMIA Annual Symposium Proceedings*, volume 2006, page 1040. American Medical Informatics Association, 2006.

[150] Casey Lynnette Overby, Chunhua Weng, Krystl Haerian, Adler Perotte, Carol Friedman, and George Hripcsak. Evaluation considerations for ehr-based phenotyping algorithms: A case study for drug-induced liver injury. *AMIA Summits on Translational Science Proceedings*, 2013:130, 2013.

[151] Aaron Mobley, Suzanne K Linder, Russell Braeuer, Lee M Ellis, and Leonard Zwelling. A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. *PloS one*, 8(5):e63221, 2013.

[152] Ida Sim, Samson W Tu, Simona Carini, Harold P Lehmann, Brad H Pollock, Mor Peleg, and Knut M Wittkowski. The ontology of clinical research (ocre): an informatics foundation for the science of clinical research. *Journal of biomedical informatics*, 52:78–91, 2014.

[153] Jessica D Tenenbaum, Patricia L Whetzel, Kent Anderson, Charles D Borromeo, Ivo D Dinov, Davera Gabriel, Beth Kirschner, Barbara Mirel, Tim Morris, Natasha Noy, et al. The biomedical resource ontology (bro) to enable resource discovery in clinical and translational research. *Journal of biomedical informatics*, 44(1):137–145, 2011.

[154] Y Megan Kong, Carl Dahlke, Qun Xiang, Yu Qian, David Karp, and Richard H Scheuermann. Toward an ontology-based framework for clinical research databases. *Journal of biomedical informatics*, 44(1):48–58, 2011.

[155] Satya S Sahoo, Joshua Valdez, and Michael Rueschman. Scientific reproducibility in biomedical research: provenance metadata ontology for semantic annotation of study description. In *AMIA Annual Symposium Proceedings*, volume 2016, page 1070. American Medical Informatics Association, 2016.

[156] Joshua Valdez, Matthew Kim, Michael Rueschman, Vimig Socrates, Susan Redline, and Satya S Sahoo. Provcare semantic provenance knowledgebase: evaluating scientific reproducibility of research studies. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1705. American Medical Informatics Association, 2017.

[157] Jessica Ross, Samson Tu, Simona Carini, and Ida Sim. Analysis of eligibility criteria complexity in clinical trials. *Summit on Translational Bioinformatics*, 2010:46, 2010.

[158] Jennifer L St. Sauver, Brandon R Grossardt, Barbara P Yawn, L Joseph Melton III, and Walter A Rocca. Use of a medical records linkage system to enumerate a

dynamic population over time: the rochester epidemiology project. *American journal of epidemiology*, 173(9):1059–1068, 2011.

[159] Eric I Benchimol, Liam Smeeth, Astrid Guttmann, Katie Harron, David Moher, Irene Petersen, Henrik T Sørensen, Erik von Elm, Sinéad M Langan, and RECORD Working Committee. The reporting of studies conducted using observational routinely-collected health data (record) statement. *PLoS medicine*, 12(10):e1001885, 2015.

[160] Erik Von Elm, Douglas G Altman, Matthias Egger, Stuart J Pocock, Peter C Gøtzsche, and Jan P Vandenbroucke. The strengthening the reporting of observational studies in epidemiology (strobe) statement: guidelines for reporting observational studies. *Bulletin of the World Health Organization*, 85:867–872, 2007.

[161] NF Boyd, JL Pater, AD Ginsburg, and RE Myers. Observer variation in the classification of information from medical records. *Journal of Chronic Diseases*, 32(4):327–332, 1979.

[162] Ralph I Horwitz and C Yu Eunice. Assessing the reliability of epidemiologic data obtained from medical records. *Journal of chronic diseases*, 37(11):825–831, 1984.

[163] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[164] David Ferrucci and Adam Lally. Uima: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, pages 1–26, 2004.

[165] J. C. Denny, 3rd Spickard, A., K. B. Johnson, N. B. Peterson, J. F. Peterson, and R. A. Miller. Evaluation of a method to identify and categorize section headers in clinical documents. *J Am Med Inform Assoc*, 16(6):806–15, 2009.

[166] Tim Bongartz, Carlotta Nannini, Yimy F Medina-Velasquez, Sara J Achenbach, Cynthia S Crowson, Jay H Ryu, Robert Vassallo, Sherine E Gabriel, and Eric L Matteson. Incidence and mortality of interstitial lung disease in rheumatoid arthritis: a population-based study. *Arthritis & Rheumatism*, 62(6):1583–1591, 2010.

[167] Jens P Hellermann, Steven J Jacobsen, Margaret M Redfield, Guy S Reeder, Susan A Weston, and Véronique L Roger. Heart failure after myocardial infarction: clinical presentation and survival. *European journal of heart failure*, 7(1):119–125, 2005.

[168] Yariv Gerber, Susan A Weston, Jill M Killian, Steven J Jacobsen, and Véronique L Roger. Sex and classic risk factors after myocardial infarction: a community study. *American heart journal*, 152(3):461–468, 2006.

[169] Véronique L Roger, Jill Killian, Mary Henkel, Susan A Weston, Tauqir Y Goraya, Barbara P Yawn, Thomas E Kottke, Robert L Frye, and Steven J Jacobsen. Coronary disease surveillance in olmsted county objectives and methodology. *Journal of clinical epidemiology*, 55(6):593–601, 2002.

[170] Sherine E Gabriel, Cynthia S Crowson, and W Michael O'Fallon. A mathematical model that improves the validity of osteoarthritis diagnoses obtained from a computerized diagnostic database. *Journal of clinical epidemiology*, 49(9):1025–1029, 1996.

[171] JA Kanis, Olof Johnell, Anders Odén, Helena Johansson, and EFRAX McCloskey. Frax™ and the assessment of fracture probability in men and women from the uk. *Osteoporosis international*, 19(4):385–397, 2008.

[172] Richard Watts, Suzanne Lane, Thomas Hanslik, Thomas Hauser, Bernhard Hellmich, Wenche Koldingsnes, Alfred Mahr, Mårten Segelmark, Jan W Cohen-Tervaert, and David Scott. Development and validation of a consensus methodology for the classification of the anca-associated vasculitides and polyarteritis nodosa for epidemiological studies. *Annals of the rheumatic diseases*, 66(2):222–227, 2007.

[173] Alvise Berti, Divi Cornec, Cynthia S Crowson, Ulrich Specks, and Eric L Matteson. The epidemiology of antineutrophil cytoplasmic autoantibody–associated vasculitis in olmsted county, minnesota: A twenty-year us population–based study. *Arthritis & rheumatology*, 69(12):2338–2350, 2017.

[174] LJ Melton, DE Wenger, EJ Atkinson, SJ Achenbach, TH Berquist, BL Riggs, G Jiang, and R Eastell. Influence of baseline deformity definition on subsequent vertebral fracture risk in postmenopausal women. *Osteoporosis international*, 17(7):978–985, 2006.

[175] Limor Raz, M Jayachandran, Nirubol Tosakulwong, Timothy G Lesnick, Samantha M Wille, Matthew C Murphy, Matthew L Senjem, Jeffrey L Gunter, Prashanthi Vemuri, Clifford R Jack, et al. Thrombogenic microvesicles and white matter hyperintensities in postmenopausal women. *Neurology*, 80(10):911–918, 2013.

[176] Barbara P Yawn, Patricia Saddier, Peter C Wollan, Jennifer L St Sauver, Marge J Kurland, and Lina S Sy. A population-based study of the incidence and complication rates of herpes zoster before zoster vaccine introduction. In *Mayo Clinic Proceedings*, volume 82, pages 1341–1349. Elsevier, 2007.

[177] Hyo Jin Kwon, Duk Won Bang, Eun Na Kim, Chung-Il Wi, Barbara P Yawn, Peter C Wollan, Brian D Lahr, Euijung Ryu, and Young J Juhn. Asthma as a risk factor for zoster in adults: a population-based case-control study. *Journal of Allergy and Clinical Immunology*, 137(5):1406–1412, 2016.

[178] Francisco Lopez-Jimenez, Steven J Jacobsen, Guy S Reeder, Susan A Weston, Ryan A Meverden, and Véronique L Roger. Prevalence and secular trends of excess body weight and impact on outcomes after myocardial infarction in the community. *Chest*, 125(4):1205–1212, 2004.

[179] Alanna M Chamberlain, Margaret M Redfield, Alvaro Alonso, Susan A Weston, and Véronique L Roger. Atrial fibrillation and mortality in heart failure: a community study. *Circulation: Heart Failure*, 4(6):740–746, 2011.

[180] Suzette J Bielinski, Jyotishman Pathak, David S Carrell, Paul Y Takahashi, Janet E Olson, Nicholas B Larson, Hongfang Liu, Sunghwan Sohn, Quinn S Wells, Joshua C Denny, et al. A robust e-epidemiology tool in phenotyping heart failure with differentiation for preserved and reduced ejection fraction: the electronic medical records and genomics (emerge) network. *Journal of cardiovascular translational research*, 8(8):475–483, 2015.

[181] Karin J Neufeld and Christine Thomas. Delirium: definition, epidemiology, and diagnosis. *Journal of Clinical Neurophysiology*, 30(5):438–442, 2013.

[182] Sharon K Inouye, Linda Leo-Summers, Ying Zhang, Sidney T Bogardus Jr, Douglas L Leslie, and Joseph V %J Journal of the American Geriatrics Society Agostini. A chart-based method for identification of delirium: validation compared with interviewer ratings using the confusion assessment method. 53(2):312–318, 2005.

[183] P. B. Ryan, M. J. Schuemie, E. Welebob, J. Duke, S. Valentine, and A. G. Hartzema. Defining a reference set to support methodological research in drug safety. *Drug Saf*, 36 Suppl 1(1):S33–47, 2013.

[184] B. T. Pun and E. W. Ely. The importance of diagnosing and managing icu delirium. *Chest*, 132(2):624–36, 2007.

[185] Y. Yang, X. Zhao, T. Dong, Z. Yang, Q. Zhang, and Y. Zhang. Risk factors for postoperative delirium following hip fracture repair in elderly patients: a systematic review and meta-analysis. *Aging Clin Exp Res*, 29(2):115–126, 2017.

[186] Evelyn Parrish. Delirium superimposed on dementia. *Nurs Clin North Am*, 54(4):541–50, 2019.

[187] RA Diwell, DH Davis, Victoria Vickerstaff, and EL Sampson. Key components of the delirium syndrome and mortality: greater impact of acute change and disorganised thinking in a prospective cohort study. *BMC geriatrics*, 18(1):1–8, 2018.

[188] Jelle W Raats, Wilbert A Van Eijsden, Rogier MPH Crolla, Ewout W Steyerberg, and Lijckle van der Laan. Risk factors and outcomes for postoperative delirium after major surgery in elderly patients. *PloS one*, 10(8):e0136071, 2015.

[189] Dan K Kiely, Edward R Marcantonio, Sharon K Inouye, Michele L Shaffer, Margaret A Bergmann, Frances M Yang, Michael A Fearing, and Richard N Jones. Persistent delirium predicts greater mortality. *Journal of the American Geriatrics Society*, 57(1):55–61, 2009.

[190] Jorge IF Salluh, Han Wang, Eric B Schneider, Neeraja Nagaraja, Gayane Yenokyan, Abdulla Damluji, Rodrigo B Serafim, and Robert D Stevens. Outcome of delirium in critically ill patients: systematic review and meta-analysis. *bmj*, 350, 2015.

[191] J. Witlox, L. S. Eurelings, J. F. de Jonghe, K. J. Kalisvaart, P. Eikelenboom, and W. A. van Gool. Delirium in elderly patients and the risk of postdischarge mortality, institutionalization, and dementia: a meta-analysis. *JAMA*, 304(4):443–51, 2010.

[192] S. Fu, L. Y. Leung, Y. Wang, A. O. Raulli, D. F. Kallmes, K. A. Kinsman, K. B. Nelson, M. S. Clark, P. H. Luetmer, P. R. Kingsbury, D. M. Kent, and H. Liu. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. *JMIR Med Inform*, 7(2):e12109, 2019.

[193] Feifan Liu, Chunhua Weng, and Hong Yu. Natural language processing, electronic health records, and clinical research. In *Clinical Research Informatics*, pages 293–310. Springer, 2012.

[194] Henk Harkema, Wendy W Chapman, Melissa Saul, Evan S Dellon, Robert E Schoen, and Ateev Mehrotra. Developing a natural language processing application for measuring the quality of colonoscopy procedures. *Journal of the American Medical Informatics Association*, 18(Supplement_1):i150–i156, 2011.

[195] C. Friedman and G. Hripcsak. Evaluating natural language processors in the clinical domain. *Methods Inf Med*, 37(4-5):334–44, 1998.

[196] Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9, 2016.

[197] Janet E Olson, Euijung Ryu, Kiley J Johnson, Barbara A Koenig, Karen J Maschke, Jody A Morrisette, Mark Liebow, Paul Y Takahashi, Zachary S Fredericksen, and Ruchi G Sharma. The mayo clinic biobank: a building block for

individualized medicine. In *Mayo Clinic Proceedings*, volume 88, pages 952–962. Elsevier.

[198] Carol Hope, Nicollete Estrada, Charlene Weir, Chia-Chen Teng, Kavitha Damal, and Brian C Sauer. Documentation of delirium in the va electronic health record. *BMC research notes*, 7(1):1–6, 2014.

[199] M. R. Puelle, C. M. Kosar, G. Xu, E. Schmitt, R. N. Jones, E. R. Marcantonio, Z. Cooper, S. K. Inouye, and J. S. Saczynski. The language of delirium: Keywords for identifying delirium from medical records. *J Gerontol Nurs*, 41(8):34–42, 2015.

[200] C. A. McCarty, R. L. Chisholm, C. G. Chute, I. J. Kullo, G. P. Jarvik, E. B. Larson, R. Li, D. R. Masys, M. D. Ritchie, D. M. Roden, J. P. Struewing, W. A. Wolf, and Merge Team e. The emerge network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics*, 4(1):13, 2011.

[201] Donald C Comeau, Rezarta Islamaj Doğan, Paolo Ciccarese, Kevin Bretonnel Cohen, Martin Krallinger, Florian Leitner, Zhiyong Lu, Yifan Peng, Fabio Rinaldi, Manabu Torii, et al. Bioc: a minimalist approach to interoperability for biomedical text processing. *Database*, 2013, 2013.

[202] Chung-Il Wi, Sunghwan Sohn, Mir Ali, Elizabeth Krusemark, Euijung Ryu, Hongfang Liu, and Young J Juhn. Natural language processing for asthma ascertainment in different practice settings. *The Journal of Allergy and Clinical Immunology: In Practice*, 6(1):126–131, 2018.