The Implication of Latent Information Quality to the Reproducibility of Secondary Use of Electronic Health Records

Sunyang Fu, M.H.I[a,d], **Andrew Wen, M.S.**[a], **Sandeep Pagali, M.D.**[c], **Nansu Zong, PhD**[a], **Jennifer St Sauver, Ph.D.**[b], **Sunghwan Sohn, Ph.D.**[a], **Jungwei Fan, Ph.D.**[a], **Hongfang Liu, Ph.D.**[a,d]

[a]Department of Artificial Intelligence and Informatics Research, Mayo Clinic, Rochester, Minnesota, USA

[b]Department of Quantitative Health Sciences, Mayo Clinic, Rochester, Minnesota, USA

[c]Department of Medicine, Division of Hospital Internal Medicine, Mayo Clinic, Rochester, MN, USA

[d]University of Minnesota – Twin Cities, Minneapolis, Minnesota, USA

Abstract

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. These barriers include complex processes for collecting and organizing EHR data and dynamic multi-level interactions occurring during information use (e.g., inter-personal, inter-system, and cross-institutional). To ensure reproducible use of EHRs, we investigated four information quality dimensions and examine the implications for reproducibility based on a real-world EHR study. Four types of IQ measurements suggested that barriers to reproducibility occurred for all stages of secondary use of EHR data. We discussed our recommendations and emphasized the importance of promoting transparent, high-throughput, and accessible data infrastructures and implementation best practices (e.g., data quality assessment, reporting standard).

Keywords

Reproducibility; EHR; Quality Improvement

Introduction

Rapid growth in the implementation of electronic health records (EHRs) has led to an unprecedented expansion in the availability of dense longitudinal datasets. This transformation holds great promise for the secondary use of EHRs to drive clinical research

*Address for correspondence Corresponding Author: Hongfang Liu, Ph.D., Liu.hongfang@mayo.edu.

through enriching patient information, integrating computable phenotype algorithms, and facilitating cohort exploration.

Research reproducibility is crucial in the field of biomedical informatics: knowledge-driven insights can be translated to broader patient care [1]. Reproducibility is however broadly defined [2], one pertinent definition being the ability to obtain consistent results when using the same data, methods, and tools [3]. One particular challenge for ensuring reproducibility in EHR-based research is data quality-related issues caused by non-standard data representations, ambiguous definitions, and missing or redundant documentation [4]. As illustrated in Figure 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 [5]. This lack of consistency in documentation practices can result in multiple sets of patient data sharing different definitions that resides in disconnected "silos" [6]. Furthermore, the care process is iterative and complex, involving multiple different stakeholders, which may further complicate the data creation and transformation process [7]. This inherent latent documentation bias and information variability in the EHR can ultimately result in incomplete medical records, recording errors, and information misinterpretation [8].

As EHR-derived findings become increasingly integrated into clinical research, methods of assessing reproducibility are needed to address information gaps (Figure 1- middlebox). The majority of existing research in the field focuses on reporting standards [5, 9], database standardization and cataloging [10], and system-level specification [11]. We propose to study reproducibility from the 'process' point of view (i.e., information collection and information use). To achieve reproducible EHR-based research, we need to ensure every step of the process is valid and reproducible.

In this paper, we proposed a multi-phase approach to examine potential barriers to reproducibility caused by information quality (IQ) -related issues. Specifically, we identified latent effects to IQ caused by the heterogeneous EHR environment and variations in the process of information collection and use. The study was based on a real-world study involving secondary use of EHR data for delirium status ascertainment.

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.

DIET Process—In this study, we refer to any information-related interactions in secondary use of EHRs as **D**igital **I**nformation **E**xtraction and **T**ransformation (DIET) process. To define the DIET process, we adapted an IQ life cycle model as a conceptual representation

of IQ through a sequence of processes [12]. Key activities contained within the IQ processes include information collection and information use. The advantage of IQ representation is that it captures the dynamic interactions between users (e.g., data abstractor, informatician) and information in a sequential order, providing additional context useful for studying reproducibility. In the context of EHR and clinical research, the DIET process includes information collection, extraction (if the study involves chart review or NLP), organization, and application.

IQ assessment—The DIET 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 [13]. The definition of each measurement is provided in Table 1.

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 [8]. 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.

Case Study

Study Setting—The study was approved by the Mayo Clinic Institutional Review Board (IRB) and the Olmsted Medical Center IRB. The study population consisted of participants of the Mayo Clinic Biobank [14]. 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, 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 [15]. Delirium is underreported and not every patient has a formal assessment for delirium diagnosis [16]. 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 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) [17].

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 DIET 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 & the Bedside) data warehouse to help obtain patient's demographics status. i2b2 is an interactive informatics platform that is widely used for patient cohort identification [18]. 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 [19]. We applied our previously developed and validated NLP system to extract CAM-related features from patient clinical notes [17]. 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 is 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 DIET process. Lastly, as suggested by Weiskopf et al, information completeness can be measured in the following four dimensions: documentation, breadth, density, and predictive [20]. Due to the underdiagnosed nature of delirium in clinical practice, we focused on documentation completeness ratio to assess contextual IQ, defined as a record that contains 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 [21]. 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

Results

Intrinsic IQ

The agreements between ICD, flowsheet, and CAM-NLP are provided in Table 3. Agreement between ICD and CAM-NLP was moderate-high (k = 0.61). Agreement between ICD and flowsheet was moderate (k = 0.41). Similarly, 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.

Accessibility IQ

The workflow analysis indicated a variant level of accessibility IQ in both intra-institutional and inter-institutional settings (Table 4). 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.

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.

Representational IQ

Figure 3 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).

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 4, 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.

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.

Conclusions

Reproducibility is crucial for the secondary use of EHRs. We applied a multi-phase method to investigate heterogeneity in the processes involved in 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

Acknowledgments

References

[1]. Dresser MV, Feingold L, Rosenkranz SL, and Coltin KL: 'Clinical quality measurement: comparing chart review and automated methodologies', Medical care, 1997, pp. 539–552 [PubMed: 9191700]

[2]. Stodden V: 'What scientific idea is ready for retirement', Edge, 2014

[3]. National Academies of Sciences, E., and Medicine: 'Reproducibility and replicability in science' (National Academies Press, 2019. 2019)

[4]. Seesaghur A, Petruski-Ivleva N, Banks V, Wang JR, Mattox P, Hoeben E, Maskell J, Neasham D, Reynolds SL, and Kafatos G: '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, 2021, 30, (2), pp. 248–256 [PubMed: 33174338]

[5]. Wang SV, Patterson OV, Gagne JJ, Brown JS, Ball R, Jonsson P, Wright A, Zhou L, Goettsch W, and Bate A: 'Transparent reporting on research using unstructured electronic health record data to generate 'real world'evidence of comparative effectiveness and safety', Drug safety, 2019, 42, (11), pp. 1297–1309 [PubMed: 31452075]

[6]. Chow M, Beene M, O'Brien A, Greim P, Cromwell T, DuLong D, and Bedecarré D: 'A nursing information model process for interoperability', Journal of the American Medical Informatics Association, 2015, 22, (3), pp. 608–614 [PubMed: 25656515]

[7]. Chiang MF, Hwang JC, Alexander CY, Casper DS, Cimino JJ, and Starren J: 'Reliability of SNOMED-CT coding by three physicians using two terminology browsers', in Editor (Ed.)^(Eds.): 'Book Reliability of SNOMED-CT coding by three physicians using two terminology browsers' (American Medical Informatics Association, 2006, edn.), pp. 131

[8]. Hum RS, and Kleinberg S: 'Replicability, Reproducibility, and Agent-based Simulation of Interventions', in Editor (Ed.)^(Eds.): 'Book Replicability, Reproducibility, and Agent-based Simulation of Interventions' (American Medical Informatics Association, 2017, edn.), pp. 959

[9]. McIntosh LD, Juehne A, Vitale CRH, Liu X, Alcoser R, Lukas JC, and Evanoff B: 'Repeat: a framework to assess empirical reproducibility in biomedical research',BMC Med Res Methodol, 2017, 17, (1), pp. 143 [PubMed: 28923006]

[10]. Harris S, Shi S, Brealey D, MacCallum NS, Denaxas S, Perez-Suarez D, Ercole A, Watkinson P, Jones A, and Ashworth S: '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, 2018, 112, pp. 82–89 [PubMed: 29500026]

[11]. Digan W, Névéol A, Neuraz A, Wack M, Baudoin D, Burgun A, and Rance B: 'Can reproducibility be improved in clinical natural language processing? A study of 7 clinical NLP suites', Journal of the American Medical Informatics Association, 2021, 28, (3), pp. 504–515 [PubMed: 33319904]

[12]. Hausvik GI, Thapa D, and Munkvold BE: 'Information quality life cycle in secondary use of EHR data', Int J Inf Manage, 2021, 56, pp. 102227

[13]. Lee YW, Strong DM, Kahn BK, and Wang RY: 'AIMQ: a methodology for information quality assessment', Information & management, 2002, 40, (2), pp. 133–146

[14]. Olson JE, Ryu E, Hathcock MA, Gupta R, Bublitz JT, Takahashi PY, Bielinski SJ, St Sauver JL, Meagher K, and Sharp RR: 'Characteristics and utilisation of the Mayo Clinic Biobank, a clinic-based prospective collection in the USA: cohort profile', BMJ Open, 2019, 9, (11), pp. e032707

[15]. Inouye SK, Westendorp RG, and Saczynski JSJTL: 'Delirium in elderly people', 2014, 383, (9920), pp. 911–922

[16]. Ritter SR, Cardoso AF, Lins MM, Zoccoli TL, Freitas MPD, and Camargos EF: 'Underdiagnosis of delirium in the elderly in acute care hospital settings: lessons not learned', Psychogeriatrics, 2018, 18, (4), pp. 268–275 [PubMed: 30133935]

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

[18]. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, and Kohane I: 'Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)', J Am Med Inform Assoc, 2010, 17, (2), pp. 124–130 [PubMed: 20190053]

[19]. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, Shen F, Wang L, Wang Y, and Wen A: 'Clinical concept extraction: a methodology review', Journal of Biomedical Informatics, 2020, pp. 103526 [PubMed: 32768446]

[20]. Weiskopf NG, Hripcsak G, Swaminathan S, and Weng C: 'Defining and measuring completeness of electronic health records for secondary use', Journal of biomedical informatics, 2013, 46, (5), pp. 830–836 [PubMed: 23820016]

[21]. Tong J, Huang J, Chubak J, Wang X, Moore JH, Hubbard RA, and Chen Y: 'An augmented estimation procedure for EHR-based association studies accounting for differential misclassification', Journal of the American Medical Informatics Association, 2020, 27, (2), pp. 244–253 [PubMed: 31617899]
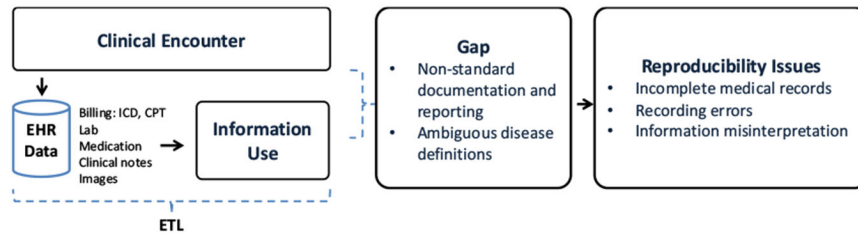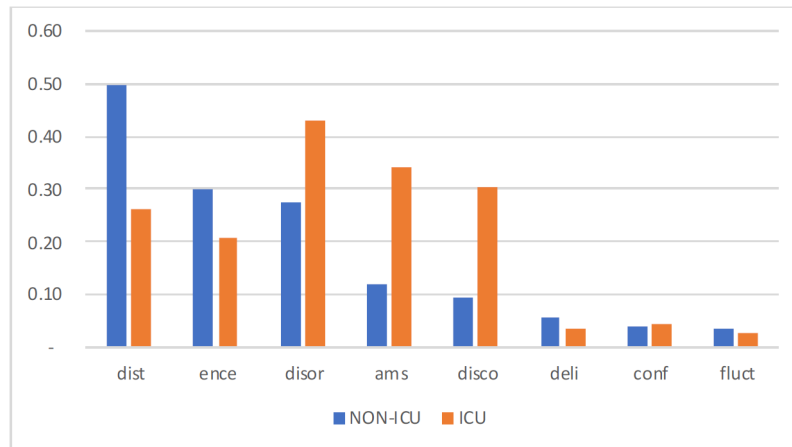
Figure 1 –.
Issues of Reproducibility in the Context of EHRs

Figure 2 –.
Information Completeness of CAM Feature Documentation in ICU and Non-ICU Settings
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
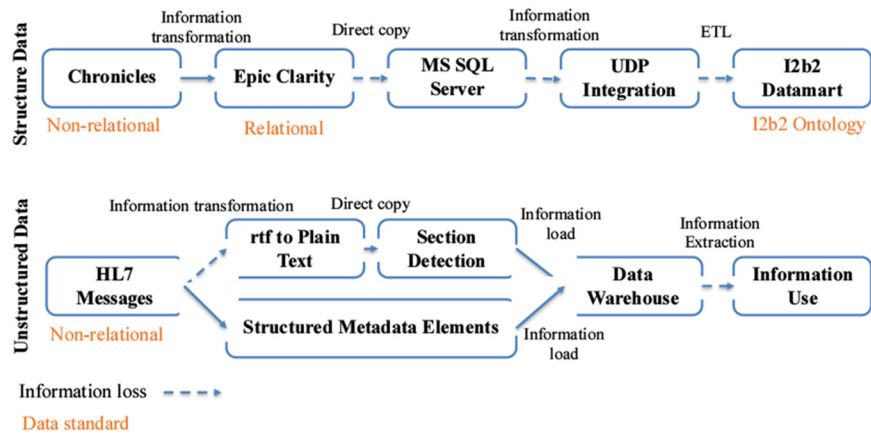
Figure 3 –.
ETL processes for Structured and Unstructured Data
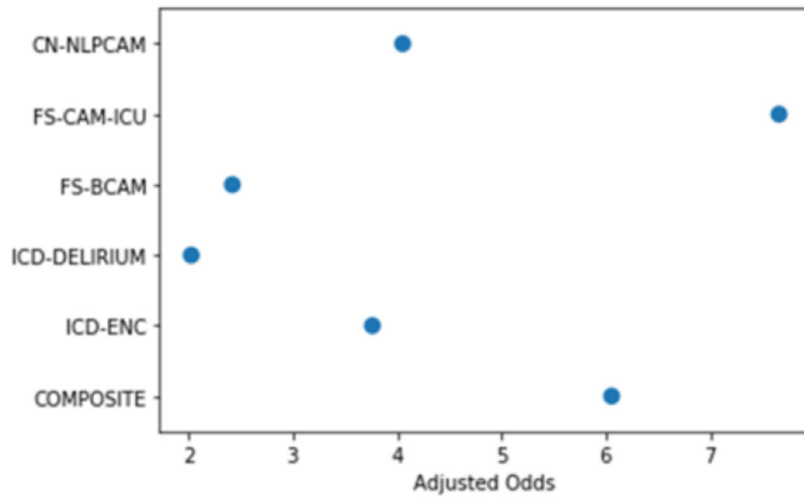
Figure 4 –.
Odds Ratio for All-cause Mortality at Discharge for Delirium Cohorts with Simulated EHR-derived measures
*The presented result is not to provide any clinical indications but rather demonstrate the variability caused by information quality

Table 1 –

Definition and Implication of IQ Measurement

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

Table 2 –

Definition of EHR-derived Measures for Ascertaining Delirium Status

EHR-derived measures	Definition
CAM-NLP	Definitive: # of unique CAM criteria >=3
	Possible: 2 <= # 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

Table 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

Table 4 –

Level of Accessibility IQ in Two Settings

		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

Abbreviation: 1: direct access, 2: adoptive access, 3: partial access, I: information level, U: user level, S: system level.