

Comprehensive Fitness Assessment for Clinical Data: Standards Development and Reusable Tools

Hanieh Razzaghi, Kaleigh Wieand, Kimberley L. Dickinson, Michael G. Kahn, Jason Roy, Clair Blacketer, Dmitri A. Christakis, Christopher B. Forrest, Jane Greenberg, Harold P. Lehmann, Keith A. Marsolo, Jennifer Sciolla, Mark G. Weiner, Ncole G. Weiskopf, L. Charles Bailey

Submitted to: Journal of Medical Internet Research
on: April 23, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript.....	5
Supplementary Files.....	27
Multimedia Appendixes.....	28
Multimedia Appendix 1.....	28
Multimedia Appendix 2.....	28
Multimedia Appendix 3.....	28
Multimedia Appendix 4.....	28

Comprehensive Fitness Assessment for Clinical Data: Standards Development and Reusable Tools

Hanieh Razzaghi¹ PhD, MPH; Kaleigh Wieand¹ BSc; Kimberley L. Dickinson¹ MS; Michael G. Kahn² MD, PhD; Jason Roy³ PhD; Clair Blacketer⁴ MPH, PMP; Dmitri A. Christakis⁵ MD, MPH; Christopher B. Forrest¹ MD, PhD; Jane Greenberg⁶ PhD, MS; Harold P. Lehmann⁷ MD, PhD; Keith A. Marsolo⁸ PhD; Jennifer Sciolla⁹ MS, CTRS, CCLS; Mark G. Weiner¹⁰ MD; Ncole G. Weiskopf¹¹ PhD; L. Charles Bailey¹ MD, PhD

¹ Children's Hospital of Philadelphia Philadelphia US

² University of Colorado Anschutz Medical Campus Aurora US

³ Rutgers School of Public Health Piscataway US

⁴ Janssen Research & Development Titusville US

⁵ Seattle Children's Hospital Seattle US

⁶ Drexel University Philadelphia US

⁷ Johns Hopkins School of Medicine Baltimore US

⁸ Duke University School of Medicine Durham US

⁹ Nemours/Alfred I. DuPont Hospital for Children Wilmington US

¹⁰ Weill Cornell Medicine New York US

¹¹ Oregon Health & Science University Portland US

Corresponding Author:

Hanieh Razzaghi PhD, MPH

Children's Hospital of Philadelphia
3401 Civic Center Blvd
Philadelphia
US

Abstract

Background: Secondary use of clinical data offers the unprecedented opportunity to rapidly conduct large-scale research and improve patient care. However, incomplete understanding of data quality requirements for a study often causes significant delays in executing analyses and validating results. While multi-institutional networks have developed general data quality programs, they are closely tied to network characteristics and do not address many study-specific requirements. In turn, most investigators conduct per-study exploratory analyses, but these efforts are typically ad hoc and partially reported, which can hinder reproducible science and delay advances in patient care.

Objective: The objectives of this study were to develop a comprehensive yet extensible model that guides study-specific data quality assessment (SSDQA) and to construct reusable data quality modules, based on the model, that can be widely disseminated and adopted.

Methods: We reviewed current literature and studies, interviewed experts, and convened a group of scientists 14 times over 24 months to develop a strategy for effective and generalizable SSDQA. From this we built a model that provides guidance for improved SSDQA design and implementation, identifying reusable components and software development best practices. We created a set of executable data quality modules addressing multiple check requirements, packaged for dissemination using the R programming language. Modules include tabular output as well as high density data displays to improve the feasibility of more complete SSDQA.

Results: The data quality model integrates theoretical principles of data quality testing with pragmatic considerations of application to clinical data. Eighteen data quality modules were specified, incorporating a wide range of theory-based check categories and practice-based check types. The nine specifications covering most common SSDQA requirements were developed as reusable software packages. Study-specific user decisions for application of a module include: (a) target data element(s), (b) single- or multi- source analyses, (c) exploratory output or statistical anomaly detection; and (d) longitudinal or cross-sectional analyses. Consistent visualizations of results are provided across multiple checks, and tabular results are available for archiving

or further computation. We demonstrate that application of SSDQA modules to a cohort of pediatric patients diagnosed with sickle cell disease drawn from a multisite clinical research network produces informative results about data characteristics that may create analytic risk.

Conclusions: The study-specific data quality model builds on current practice, providing guidance for more complete and sound assessments. Reusable modules implementing model guidance can be readily deployed in large networks or smaller consortia, and provide detailed reporting of study-specific data quality in reproducible form. The flexibility allows investigators to tailor checks to their study requirements while retaining consistency of DQA methods. This approach fosters collaboration to identify data quality issues that can inform decisions about study design, as well as provide important context bearing on adoption of results.

(JMIR Preprints 23/04/2025:76398)

DOI: <https://doi.org/10.2196/preprints.76398>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a JMIR journal, my title and abstract will remain visible to all users.

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in a JMIR journal, my title and abstract will remain visible to all users.

Original Manuscript

Original Paper

Comprehensive Fitness Assessment for Clinical Data: Standards Development and Reusable Tools

Abstract

Background: Secondary use of clinical data offers the unprecedented opportunity to rapidly conduct large-scale research and improve patient care. However, incomplete understanding of data quality requirements for a study often causes significant delays in executing analyses and validating results. While multi-institutional networks have developed general data quality programs, they are closely tied to overall network characteristics and do not address many study-specific requirements. In turn, most investigators conduct per-study exploratory analyses, but these efforts are typically ad hoc and partially reported, which can hinder reproducible science and delay advances in patient care.

Objective: The objectives of this study were to develop a comprehensive yet extensible model that guides study-specific data quality assessment (SSDQA) and to construct reusable data quality modules, based on the model, that can be widely disseminated and adopted.

Methods: We reviewed current literature and studies, interviewed experts, and convened a group of scientists 14 times over 24 months to develop a strategy for effective and generalizable SSDQA. From this we built a model that provides guidance for improved SSDQA design and implementation, identifying reusable components and software development best practices. We created a set of executable data quality modules addressing multiple check requirements, packaged for dissemination using the R programming language. Modules include tabular output as well as high density data displays to improve the feasibility of more complete SSDQA.

Results: The data quality model integrates theoretical principles of data quality testing with pragmatic considerations of application to clinical data. Eighteen data quality modules were specified, incorporating a wide range of theory-based check categories and practice-based check types. The nine specifications covering most common SSDQA requirements were developed as reusable software packages. Study-specific user decisions for application of a module include: (a) target data element(s), (b) single- or multi- source analyses, (c) exploratory output or statistical anomaly detection; and (d) longitudinal or cross-sectional analyses. Consistent visualizations of results are provided across multiple checks, and tabular results are available for archiving or further computation. We demonstrate that application of SSDQA modules to a cohort of pediatric patients diagnosed with sickle cell disease drawn from a multisite clinical research network produces informative results about data characteristics that may create analytic risk.

Conclusion: The study-specific data quality model builds on current practice, providing guidance for more complete and sound assessments. Reusable modules implementing model guidance can be readily deployed in large networks or smaller consortia, and provide detailed reporting of study-specific data quality in reproducible form. The flexibility allows investigators to tailor checks to their study requirements while retaining consistency of DQA methods. This approach fosters collaboration to identify data quality issues that can inform decisions about study design, as well as provide important context bearing on adoption of results.

Keywords:

data quality; data fitness; electronic health records; research readiness; automation

Introduction

Secondary use of clinical data offers unprecedented opportunities to conduct large-scale research and derive insights into patient care. This is particularly true when using data from multi-institutional networks such as PEDSnet[1], PCORnet®[2], the National COVID Cohort Collaborative[3], AllofUs [4], or others[5-9]. However, clinical data are complex, and problems related to data quality[10-16] have significantly hindered researchers executing analyses. Many solutions have been proposed for identifying and resolving data quality issues in real-world data[17-32] through overarching data quality assessment (DQA) programs, including in multi-institutional networks, but large-scale efforts have often been siloed by network characteristics, differences in data models, or lack automation and reproducibility[33, 34]. Further, some types of data quality paradigms that impact electronic health record (EHR)-derived data are qualitatively different than those that impact primary research analyses, quality improvement efforts, or other topic- or domain-specific projects[35-37].

Broad network DQA has traditionally focused on overarching conformance, completeness, and plausibility components [27] rather than targeting specific clinical content, which can miss significant issues. For example, a network data quality metric may evaluate clinical procedures as a single group to ensure appropriate mappings and compute procedure records per patient. However, if a particular study relied on the presence of specific biopsy procedures, data extraction processes that map these procedures to a more general biopsy code or miss these mappings entirely would not be discovered during broader network data quality checks. This need must be balanced against the infeasibility of testing every possible topic of potential interest. As a result, context-aware data quality checks that are reusable across domains and research studies are needed to identify these types of issues. Further, transparency about data quality is an important step toward reproducible research. These types of checks are often referred to as assessments of *data fitness*, and are usually applied on a study-specific basis. There have been efforts spearheaded by the US Food and Drug Administration (FDA) and European Medicines Agency (EMA) that have developed guides to evaluating data fitness in real world data, which have considerably advanced the field[38-40]. These frameworks formalize study design assessments and data requirements, but lack specific tools or software to standardize and systematically report on findings in quantitative or visual formats.

Evaluations of current state in study- or topic- specific data quality assessment (SSDQA) for clinical data have consistently identified the need for more standard, applicable, automatable approaches[33, 34]. We describe here a model, incorporating prior art as well as extended consensus building and testing, that provides both a generalized approach to SSDQA development and reporting, and a set of reusable software modules addressing many common use cases. We tested using PEDSnet[1], the largest multi-institutional pediatric learning health network, to refine our model. Our goal was to bridge the gap between broad network DQA and ad hoc SSDQA by addressing the need for well-founded, context-aware data quality testing, and constructing reusable data quality metrics that can be widely adopted. We previously reported a first version of the conceptual framework[41], as well as testing initial constructs in a multi-institutional PCORnet® study evaluating blood pressure control in children with chronic kidney disease as a means of preserving kidney function[42, 43]. In that study, we tracked framework-derived data quality checks that were systematically applied to the study dataset, as well as resulting visualizations to aid in interpreting findings, for both participating institutions and public reporting.

Methods

Development of an Instantiated Model

Our previously reported data fitness framework served as a roadmap for the current work, drawing on thematic analysis of published papers as well as research conducted with PEDSnet[41]. Broadly, that framework consists of two phases to design a data quality *check*, which we refer to as any formal assessment to identify a data quality problem or issue. The first phase focuses on theoretical considerations that drive data quality such as clinical processes, the role of variables in analyses, and the science of previously published data quality approaches. The second operationalizes the concepts from these theoretical considerations to derive the requirements for assessment. For example, in a study of cancer patients, the first phase would require defining clinical concepts, weighing factors such as the tradeoff between a sensitive or specific definition, required follow up, or utilization requirements. The second phase would test operational implementation, such as data availability if a chemotherapeutic agent was integrated into the inclusion criteria. We build on this foundation here to develop a practical, reusable model that drives concrete DQ check development.

We augmented the literature review from our foundational work by updating our search using data quality terms in PubMed (“data quality”, “data” AND “quality” AND “EHR”, “data fitness” AND “EHR”, “data fitness” AND “real world data”); the updated results [25, 33-36, 44-51] were selected for their relevance in describing data quality frameworks, metadata, application, and methods. For included studies, the following topics were abstracted: (1) terms and metadata used to describe data quality in EHRs or clinical research networks; (2) methods to assess data quality; (3) level at which data quality was assessed; (4) domain specific context for the data quality assessment, if applicable; (5) reproducibility of process or results.

Because DQ results or exploratory analyses that check for data quality issues ad hoc are rarely published, it was important to learn directly from the experience of researchers about findings in data that led to significant issues during analyses or delayed research timelines. To address this need, we surveyed four established informatics investigators, representing the data quality testing approaches of several large secondary use research networks[2, 3, 52], using a common form (Supplemental Table 1) to assess a total of nine studies and 21 exemplar data quality issues.

We also convened a group of stakeholders and content experts (the development committee) to provide guidance about approaching DQA systematically (Supplemental Table 2). The committee met 14 times over the course of 24 months and provided feedback on both the theoretical underpinnings guiding the framework as well as specific model components, such as data quality term selection and user-driven requirements. During this time, multiple iterations of the model were reviewed with the committee, using both conceptual flow diagrams and concrete data quality checks developed from model principles. To test whether the abstract principles of the SSDQA model could be applied to real-world data, several meetings focused on interpretation of results when model-derived checks were tested on data from PEDSnet or PCORnet® studies. During this time, we also held individual interviews with several members of the committee to probe more deeply about personal experiences in applying data quality assessment to real-world clinical data and their recommendations for reusable checks that should be included in any instantiated model (Supplemental Table 3).

We combined these three sources of information to form a set of requirements that would drive the

development of a concrete model for DQA specific to evaluating data fitness for an intended purpose or study. We included both high-level strategies, or goals that an effective model must reflect, as well as specific design principles that motivate the process of check construction and application (Textbox 1).

Textbox 1. Implementation strategies and design principles to develop an instantiated SSDQA model.

DQ Implementation Strategies

- **Standardization:** The principles of the framework should be reusable across study-specific contexts, and code should be packaged to streamline implementation
- **Specificity:** DQ terms describing checks should be clearly paired with methods and output
- **Configurability:** Researchers should tailor modules to make decisions based on their own requirements without need for extensive coding
- **Accessibility:** Methods for DQA should range from simple exploratory visualizations to standardized statistical tests for outlier detection
- **Evident Utility:** The approach should be pragmatic and informative to the user, with results pointing to underlying characteristics of the data that may contribute to bias or confounding in their study.

DQ Design Principles

- Identify where DQA fits into the overall process of research design and procedures
- Determine how terms from the literature describing checks relate to each other in hierarchical structures to reduce redundancy in data quality approaches
- Differentiate between a DQ check, which measures a goal, and a DQ dimension, which provides additional information to evaluate the check
- Associate data quality *term* indicating check purpose with *method* and *output* to reduce ambiguity when applying checks and describing results
- Implement the most widespread checks as templates in reusable modules

DQ Check Implementation

Once the structure of the SSDQA model (*cf.* Figure 1) was established through consensus development in the committee, we applied the resulting Data Quality Module Check Development criteria to develop specifications for concrete checks, based on the strategies and design principles described above. These were iteratively refined through discussion between the study team and development committee. We selected the specifications that accounted for the most commonly used approaches to data fitness testing in our literature and study reviews, and implemented those check specifications. To ensure reproducibility and standardization in their use, we created code in R[53] according to software development principles such as modular code construction, parameterized functions to incorporate study-specific variability in user input, and reproducibility with extensive documentation. The implementations are compatible with both the Observational Medical Outcomes Partnership (OMOP) and PCORnet® data models.

We performed three rounds of testing to refine our model for SSDQA check implementation. First, we selected three different types of checks intended to measure *completeness* as initial cases for implementation. These checks measured clinical utilization, cohort medical complexity, and impact of altering variable definitions on clinical cohort definitions. Multiple dimensions, such as temporality, variable stratifications, and site-level benchmarking were included in the completeness analysis. We also tested various statistical and exploratory outputs (e.g., time-series, k-means, frequency distributions) and visualizations (e.g., heatmaps, treemaps, line graphs), to identify the

most informative. Second, we applied a set of checks drawn from multiple categories in the PRESERVE study, a large-scale observational study that integrates data from 15 health systems in PCORnet[®] to evaluate the effect of blood pressure control on kidney function outcomes in children with chronic kidney disease (CKD)[42]. We observed significant improvements in data quality as a result of our work[43]. From this process, we integrated modifications based on our learnings to improve our SSDQA model development. Finally, we executed a series of DQ checks, using final modules described above, across multiple sites in the PCORnet[®] network, containerizing our code to meet the network's practice of distributed operation, verifying that it was interoperable in a variety of technical environments.

Development Sample

For development purposes, we tested modules extensively using a cohort of patients with sickle cell disease (SCD), an uncommon condition in the general population but a major cause of both health burden and health care utilization in affected children. Examples of output in this manuscript use this cohort. Data were drawn from the harmonized dataset maintained by PEDSnet[1], a network of ten pediatric healthcare institutions collaborating to improve child health, which currently comprises over 14 million patients. Patients were included if they had evidence of at least one in-person encounter on or after January 1, 2012, and had their first visit of any kind associated with an SCD diagnosis on or after January 1, 2011. Patients who met these criteria were observed from the date of their first encounter until the date of their last or most recent encounter. Data were drawn from PEDSnet database version 2024-Aug, which yielded a cohort of 12,722 patients, of whom 99% had multiple occurrences of an SCD diagnosis, and for whom the median follow-up was 6.33 years [IQR 2.44-10.39].

All direct identifiers are replaced with study codes in the harmonized dataset, which constitutes a Limited Data Set as specified in the Health Insurance Portability and Accountability Act's Privacy Rule[54], and this work was determined to be Not Human Subjects Research by the Institutional Review Board at the Children's Hospital of Philadelphia.

Results

SSDQA Model

We synthesized theoretical and applied considerations into a concrete model (Figure 1) for constructing fitness-oriented data quality checks. The model provides consistent guidance for creating standardized modules that can be used to evaluate data fitness across a range of study-specific requirements. It delineates the processes of check development and execution through two interdependent processes, labeled on the left hand side of the graph: (1) SSDQA Development and Tools, which focuses on check development and backend automation, and (2) User Process, which highlights the interaction between the user and DQ modules. We describe these processes in the two sections below.

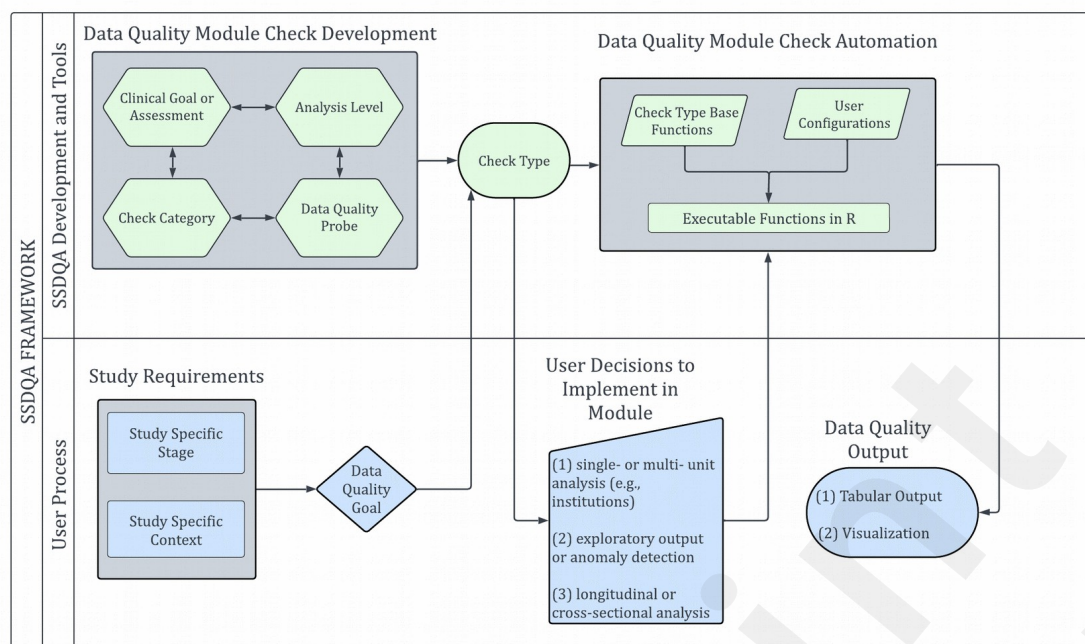


Figure 1. Study Specific data quality model for developing reusable modules to assess fitness across range of use cases

SSDQA Development and Tools

The development of a data quality module incorporates four separate but integrated components. The first is the Clinical Goal or Assessment, which describes the clinical truth that the user aims to evaluate. For example, an investigator might posit that patients in a study cohort should have confirmatory clinical data (e.g., end stage renal disease and evidence of dialysis) to establish basic face validity. The second is Check Category, which expresses the evaluated aspect of data quality, using widely adopted harmonized terms[55] such as completeness, plausibility, or conformance. The third is Data Quality Probe which, in contrast with the Clinical Goal, designates a data-centric purpose for a DQ check such as assessing eligibility criteria or mapping errors in the data. Finally, Analysis Level describes the application of the module across levels of aggregation including person-level, event-level, or visit-level analysis. The integration of these four components forms a Check Type, which is discoverable through these attributes; we have included in Supplemental Table 4 the metadata terms used to describe each Check Type in the current implementation.

Once a Check Type has been specified, application to data is mediated by a DQ Module, whose scope spans the set of user configurations (see below) required to execute a specific data quality check of that type. The module facilitates consistent application through a set of base functions that incorporate user choices via parameters, without requiring reimplementing or code editing and the consequent risk of divergent computation across uses.

User Process

Application of study-specific data quality must by definition be adapted to each user's study requirements. DQ checks are deployed to inform the user about study design and method choices, as well as to assess the validity of the dataset and identify potential biases or stratifications present in

the data. DQ requirements may vary based on the stage of the study or the background of the user. For example, data quality for a clinical researcher planning cohort selection will require a different set of evaluation measures than validation of primary outcomes, and testing at the design stage will differ from a statistician investigating an anomalous result during analysis. This is denoted as Study Specific Stage in Figure 1.

In parallel, Study Specific Context compels users to formally define the data constructs (e.g., *cohort*, *variables*, *concept sets*) around which to identify potential gaps in data quality or constraints in study design that require testing. This step drives the creation of “study-specific” requirements at the core of the SSDQA implementation model. Based on these criteria, users go on to select the appropriate Check Types to be used. For each Check Type, the user applies the associated module to the desired data elements, configuring it based on three decision points. First, execution can produce single- or multi- unit analysis. Most often, units are expressed as participating institutions, and multi-unit output facilitates comparisons among institutions while single-unit output permits more in-depth analysis within a particular institution. In other circumstances, it may be more correct to combine all institutions as a single unit of analysis, or to stratify by other characteristics such as clinical specialty. Second, the user determines whether the goal of this particular DQ analysis is to allow visual exploration, or to formally detect outliers. The former produces descriptive visualizations, while the latter computes quantitative metrics to differentiate potential outliers. Third, users can analyze their data longitudinally to understand changes over time, or cross-sectionally covering a specified time period. Following configuration, users apply the module to produce analytic results for human interpretation or further processing.

SSDQA Module Design

When implementing modules that translate abstract concepts into concrete checks, we balanced three considerations: 1) creating tools that could be adopted without extensive background in DQ methodology or programming effort, 2) covering related analyses so that users could fully evaluate specific requirements for their study, and 3) supporting reuse across a variety of similar cases. We selected the widely-used R programming language[53] and tidyverse[56] interface layer to reduce barriers to use, and maintained a consistent approach to functional interface and visualizations across modules. The modules are constructed to interact with data in the OMOP [52] or PCORnet[®] Common Data Models[57], which are widely used in health services research. All implemented modules are freely available on GitHub[58].

Module Selection and Definitions

Drawing on the model and principles described above, we specified a set of 18 DQ modules (Table 1) that cover foundational requirements for SSDQA. In order to provide guidance about intended use, we group the modules here based on different phases of a study. For example, during the *Cohort Identification* phase of a study, research teams might deploy the *Attrition Step* or *Sensitivity to Selection Criteria* modules. Many modules can be applied at multiple phases in practice; we have not duplicated descriptions to avoid redundant information.

Table 1. Data Quality Modules specified using the SSDQA framework

Module Name	Module Description
<i>Cohort Identification</i>	

Attrition Step	These checks assess each step of attrition criteria to identify anomalies in cohort construction.
Sensitivity to Selection Criteria	This module provides a set of checks that test the differences in demographics, follow-up time, utilization patterns, outcomes, and other key variables when alternative cohort definitions are used to define the patient population.
<i>Cohort Fitness</i>	
Patient-Facts	The completeness-focused checks in this module assess the availability of patient clinical data. The number of clinical events per year of follow-up for each patient in a cohort is computed and stratified by visit type.
Patient Event Sequencing	This module evaluates the plausibility of dates in relationship to clinical events. For example, a diagnosis for chronic kidney disease should precede a procedure order for dialysis. If events are not sequenced in the expected order, it may indicate inappropriate documentation or a patient receiving care at a different institution.
Patient Record Consistency	This module tests whether the clinical data in a patient's record is consistent and confirmatory (e.g., patients with leukemia should be receiving chemotherapy). This module is at the patient-level rather than at the visit-level. The Visit Clinical Data Agreement module can augment this module to determine whether expected events also occur within the same visit.
Expected Facts Present	The checks in this module evaluate whether a cohort contains specific types of clinical data, such as BMI, office visit procedures, or vital signs.
Clinical Data Values and Ranges	The checks in this module assess whether clinical data outcomes or values align with patient cohort characteristics.
<i>Dataset Fitness</i>	
Clinical Events and Specialty Agreement	The concordance-focused checks in this module assess the quality of specialty data in a study sample. The goal is to evaluate the concordance of specific events (e.g., diagnosis of type I diabetes) and specialist providers or care sites (e.g., endocrinologist or endocrinology clinic) at the visit level.
Visit Clinical Data Agreement	This module tests whether expected clinical events occur within the same clinical encounter. For example, ventilator support should occur in an intensive or critical care unit, or a prescription for an antihypertensive medication should be accompanied by a blood pressure reading. The investigator provides information about the expected clinical events, and the module computes comparisons and distributions, and identifies anomalies.
Date Sequencing	This module provides a set of checks to ensure that the dates within a dataset are plausible. It detects outliers and anomalous values such as dates too far in the past or future, and whether dates assigned to clinical events occur in reasonable order or proximity.
<i>Data Conformance</i>	
Clinical Metadata	The checks in this module assess whether clinical facts are accompanied by appropriate metadata. Examples of metadata-clinical fact pairings include the proportion of prescription drugs with dosing information, lab values with an assigned specimen, or anthropometric measures with an appropriate unit.
Unit & Value Alignment	This module determines whether drug prescriptions/administrations or lab results contain units that are conformant to expected standards.
Duplicate Record Check	The checks in this module identify where there are duplicate rows or values in a given dataset.
<i>Variable Testing</i>	
Expected Variables Present	This module checks for the presence of expected variables and presents a variety of distributions of these variables in the dataset. It allows the investigator to visualize multiple variables for comparison and to show the co-occurrence of variables in the same patient.
Quantitative Variable Distributions	The checks in this module allow the user to evaluate quantitative distributions of clinical values (e.g., lab values, BMI, etc.) or patient characteristics (e.g., number of visits per patient, follow-up time) in a study dataset.
<i>Concept-Set Testing</i>	
Concept-Set Distribution	This module examines distributions of concepts (e.g. codes) that represent a particular variable in a dataset. The purpose is to understand how the concept-sets drive the prevalence or clinical composition of variables for a particular study. For example, to identify asthma in a primary care cohort, wheezing or uncomplicated asthma may be the most common concept represented, but concepts related to asthma exacerbation or severe asthma may be most common in a more specialized respiratory cohort.
Source and Concept	The checks in this module demonstrate how mappings of source system concepts to study vocabularies impact the representation of concepts in the data. It aims to provide researchers with source-to-concept

Vocabularies	mappings and their distributions to identify problems or issues with how standardization may result in information loss. To run these checks, the source code as well as the mapped code must be available in the data.
Unmapped Concepts	The checks in this module provide information about the potential impact of unmapped concepts within a dataset.

Module Implementation

We recognized that the typical analysis would involve examination of the same data in different ways, and that the source data from multiple institutions might not be readily accessible to those making study decisions. Therefore, each module comprises two separable parts: *process* (tabular results) and *output* (visualization). We will use our sickle cell disease cohort as examples of each by applying the **Sensitivity to Selection Criteria** module, where we evaluated the impact of two alternate cohort definitions on the patient population across multiple institutions. The first alternative definition was to require a diagnosis prior to the age of 1 (ASC1), and the second narrowed the cohort to patients with the SCD-SS and SCD-S β_0 subtypes, associated with more severe clinical manifestations (ASC2). We show here differences in encounter-based utilization metrics, specifically, outpatient, ED, and hematologist encounters, as well as clinically-oriented metrics such as prescription drugs, medication administrations, and procedures. Results are normalized to events per patient year to account for variability in follow up, and the median value for each institution is computed. The module allows for any number of utilization metrics, as well as demographic composition. The *process* portion compiles tabular results of the module's DQ checks (Figure 2, Panel A), and can be applied to data locally or at remote sites. The *output* portion of the module takes the tabular results, potentially compiled from multiple *process* runs at different sites, as input and generates a consistent set of visualizations for user interpretation and reporting. These were designed to display results in familiar ways, and present information densely enough that the reader can detect patterns (e.g. stratifications over time or data source) or compare data elements. The **Sensitivity to Selection** module for the sickle-cell cohort produces a graph where the center point represents the value for the original cohort, with ASC1 on the left, and ASC2 on the right (Figure 2, Panel B). The graph displays differences in utilization metrics between the cohorts, with ASC1 showing increased utilization across almost all metrics, particularly ED utilization and specialty encounters. The specialty visits appear bifurcated: in the base cohort, one set of institutions shows about five visits per person year, and the other about 1. Medication administrations are roughly similar between the cohorts, except with a sharp increase seen at one institution in the ASC2 cohort. One institution shows on average more prescription drugs for patients than others, with a sharp increase in the ASC1 cohort. Generally, these results point to heterogeneity in institutional data as well as cohorts, and may require investigation into specialty visits as well as medication administrations and prescription drugs. Patients diagnosed between 0-1 may behave as either a loyalty cohort or a cohort that is more clinically complex, while the latter is clearly true of ASC2.

site	cohort id	cohort characteristic	output	total patients
Site C	ASC1	median emergency visits ppy	0.89	230
Site C	ASC1	median med admin drugs ppy	9.705	230
Site C	ASC1	median outpatient visits ppy	6.39	230
Site C	ASC1	median prescription drugs ppy	4.81	230
Site C	ASC1	median procedures ppy	67.535	230
Site C	ASC1	median specialty visits ppy	3.08	230
Site C	ASC2	median emergency visits ppy	0.62	425
Site C	ASC2	median med admin drugs ppy	11.67	425
Site C	ASC2	median outpatient visits ppy	6.6	425
Site C	ASC2	median prescription drugs ppy	4.94	425
Site C	ASC2	median procedures ppy	67.35	425
Site C	ASC2	median specialty visits ppy	2.04	425
Site C	base cohort	median emergency visits ppy	0.45	717
Site C	base cohort	median med admin drugs ppy	5.17	717
Site C	base cohort	median outpatient visits ppy	4.24	717
Site C	base cohort	median prescription drugs ppy	2.77	717
Site C	base cohort	median procedures ppy	40.8	717
Site C	base cohort	median specialty visits ppy	0.44	717

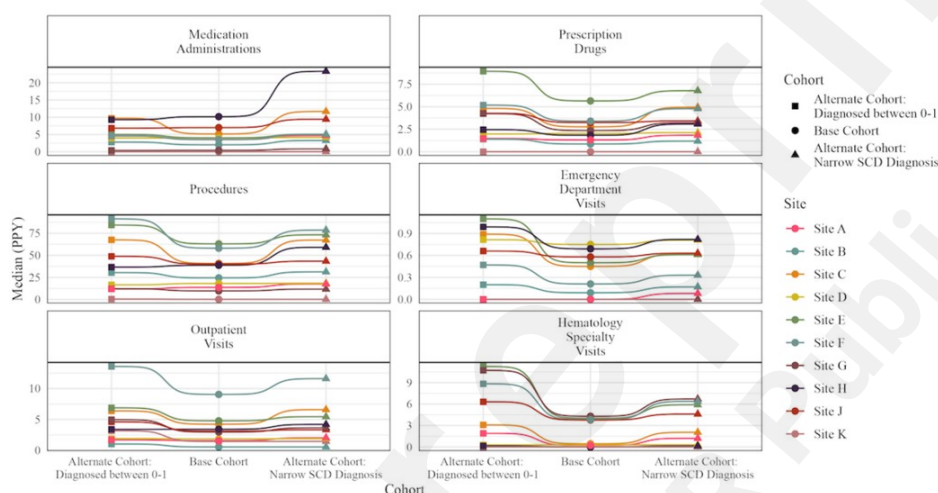


Figure 2. Results generated by the Sensitivity to Selection Criteria SSDQA module for alternative definitions of a SCD cohort. A. Tabular output for selected utilization measures produced by the process component. B. Visualization produced by the output component to allow more effective interpretation of the differences between cohort definitions.

In our design of modules, we accepted that interpretation of dense displays requires some initial orientation, and some summarization. Therefore, the more complex visualizations include additional information via hover displays or aggregated trends to support data exploration. However, the design also considered the suitability of static graphics for inclusion in manuscript supplemental data or dataset descriptors.

As Table 1 indicates, modules were designed to be executed at different steps during a study. We have demonstrated their use in *Cohort Identification*. They are also used to test *Cohort Fitness*, *Dataset Fitness*, *Data Conformance*, *Variable Testing*, and *Concept-Set Testing*. The primary difference between *Cohort Fitness* and *Dataset Fitness* is that the former evaluates fitness of a specific cohort against study requirements while the latter tests for logical cohesion of a dataset. Therefore, anomalies found in the former require investigating study-specific criteria while those in the latter identify broader logical inconsistencies in a dataset. *Data Conformance* modules evaluate a dataset for conformance to a required set of standards, such as ensuring that units for a particular lab

conform to the standard or accepted reporting for that lab. *Variable Testing* focuses on study variables; that is, a set of operational clinical definitions represented in an analysis. To use the sickle cell example, the *Expected Variables Present* module can evaluate whether opioid prescriptions are detected in the dataset. Figure 3 shows multi-institutional output for this module, demonstrating the variation of opioid usage across institutions. In this graph, the mean proportion of patients prescribed the opioid across all institutions is indicated by the size of the circle, as an indicator of overall importance of that medication within the study cohort. For comparison, the color represents the proportion for each specific institution. Anomalously high or low values are indicated by a star. For example, Site A has the highest proportion of patients

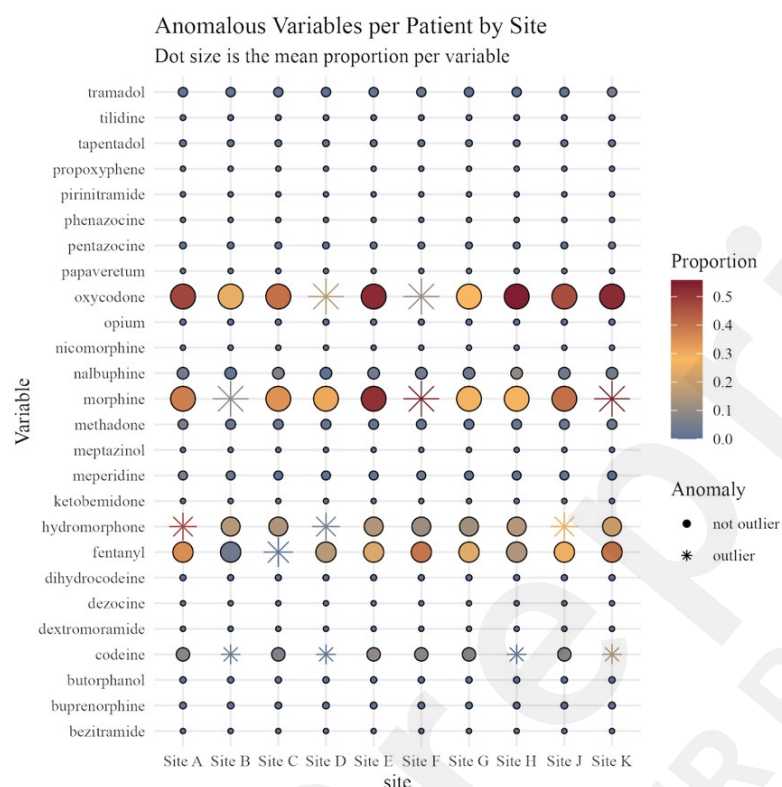


Figure 3. Dot plot showing proportion of patients receiving different opioid prescriptions. Size of dot is the mean proportion across all institutions and color is specific to the institutions. Stars indicate statistical outliers.

across institutions for hydromorphone prescriptions, indicating increased use of second-line opioid analgesics which could mean greater complication rate or significant practice variation. In contrast, Site B has low proportions across all opioid prescriptions generally, which may indicate a less acute population or alternative pain management strategies. The *Concept Set Testing* category focuses on specific vocabularies or ontologies that represent variables. Using the opioid example, the *Concept-Set Distribution* module (*cf.* Figure 5) focuses on the distributions of the individual RxNorm codes representing opioids, rather than drug selection grouped by active ingredients.

One of the unique characteristics of the SSDQA modules is the ability to configure almost all modules to detect anomalies across time. Extending the *Expected Variables Present* module example described above, a researcher can configure the module to observe the proportion of patients prescribed opioids over time. In the multi-site analysis (Figure 3), Site A was anomalously high for hydromorphone prescriptions compared with other institutions. This is confirmed by a consistent trend over time (Figure 4), and the error bars between Site A and other institutions do not overlap. This may prompt the investigator to more deeply investigate characteristics of the patient

population that could possibly explain these differences and whether adjustments need to be made to create a less heterogeneous patient population across institutions. The module also computes the distance of institutional trend lines from the all-site mean, here confirming that Site A is highly anomalous compared with other institutions, while site D is moderately anomalous for low prescription rates. Conversely, despite the transient peak seen for Site B, its overall distance from median prescribing is small.

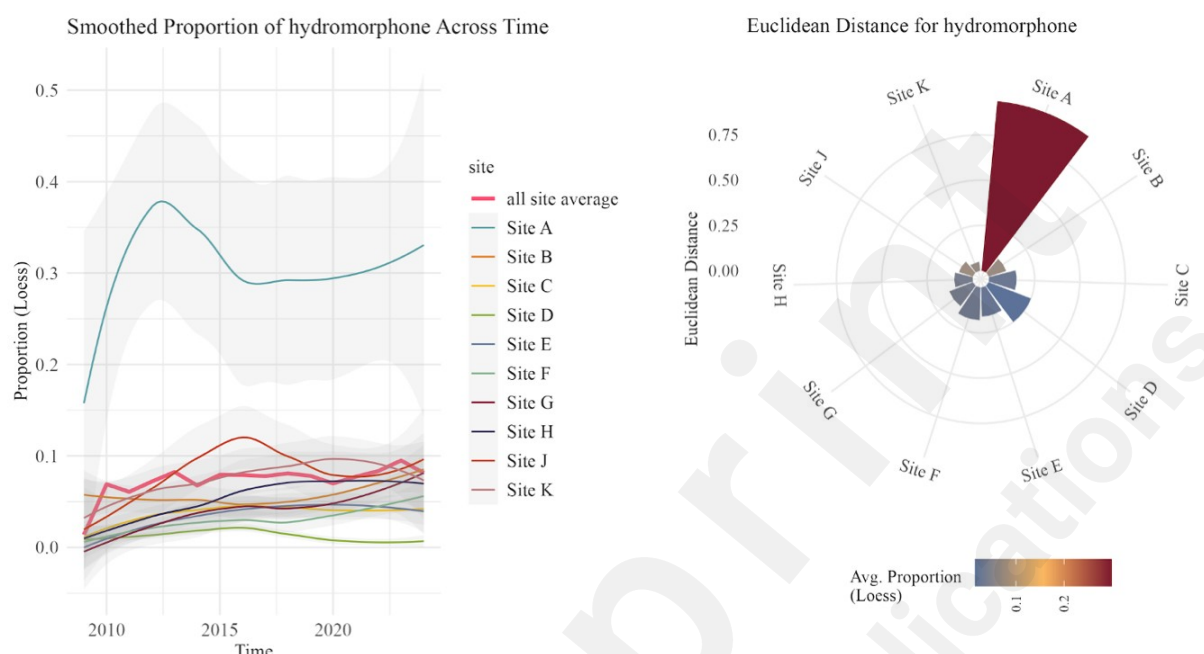


Figure 42. Anom2 method-visualization pair, shown in results of Expected Variables Present longitudinal multi-institutional analysis. A. Smoothed graph shows proportion (y-axis) of patients prescribed hydromorphone over time (x-axis). B. Euclidean Distance is also computed to quantitatively display the distance over time between a single institution and the all-institution mean.

Module Technical Reproducibility

To streamline the modules and enable better reproducibility, we adhered to a set of Output-Visualization dyads (Table 2). For example, several of the exploratory analyses output frequency distributions that can be represented by a bar graph (denoted by Reference Alias: *exp1*). This pairing (frequency distribution and bar graph) uses the same underlying code and configurations across multiple modules. However, distributions can be also visually displayed when appropriate as heatmaps, interactive tables that allow dynamic filtering and searching, and dot plots.

Table 2. Method-Visualization dyads for use in analysis in SSDQA modules

Output-Visualization Alias	Dyad	Data Output	Visualization
anom1		Jaccard Similarity Index	Heatmap
anom2		Euclidean Distance	Line Graph & Radial Bar Graph
anom3		Hotspots Outlier Detection	Dot + Star Plot
anom4		Time Series Anomalies	Control Chart
anom5		Time Series Anomalies	STL Regression
anom6		Distance from Mean	Bar Graph
anom7		Standardized Mean Difference	Heatmap
anom8		Standardized Mean Difference	Dotplot & Summary Table

anom9	Jaccard Similarity Index	Bar Graph
exp1	Frequency Distributions	Bar Graph
exp2	Attrition Counts	Line Graph
exp3	Time Series Exploration	Line Graph
exp4	Frequency Distributions	Interactive Table
exp5	Time Between Events	Histogram & Bar Graph
exp6	Frequency Distributions	Dot Plot
exp7	Frequency Distributions	Heatmap
exp8	Cohort Overlap	UpSet Graph & Bar Graph

As an example, the dyad *exp7* is used in four modules (Table 3). For simplicity, we will demonstrate its use in two modules: *Concept Set Distribution* and *Source and Concept Vocabularies*. In these two modules, *exp7* is used in the single site analyses. We selected Site A to determine whether there may be a connection between the use of a second-line opioid and pain, which would be the indication for their use.

Table 3. SSDQA modules with Method-Visualization pairs

	<i>Single Unit: Cross Sectional</i>		<i>Multi Unit: Cross Sectional</i>	
<i>Method-Visualization Pair Alias</i>	<i>Exploratory</i>	<i>Anomaly</i>	<i>Exploratory</i>	<i>Anomaly</i>
Cohort Attrition	exp2		exp2	anom3
Source and Concept Vocabularies	exp7	anom3	exp4	anom3
Expected Variables Present	exp1	anom1	exp7	anom3
Patient Event Sequencing	exp5		exp1	anom3
Patient Facts	exp1	anom6	exp3	anom3
Concept Set Distribution	exp7	anom1	exp4	anom3
Clinical Events and Specialty Agreement	exp1	anom3	exp6	anom3
Patient Record Consistency	exp1	anom9	exp1	anom3
Sensitivity to Selection Criteria	exp8	anom7	exp1 + exp7	anom8
	<i>Single Unit: Longitudinal</i>		<i>Multi Unit: Longitudinal</i>	
<i>Method-Visualization Pair Alias</i>	<i>Exploratory</i>	<i>Anomaly</i>	<i>Exploratory</i>	<i>Anomaly</i>
Cohort Attrition				
Source and Concept Vocabularies	exp3	anom4 / anom5	exp3	anom2
Expected Variables Present	exp3	anom4 / anom5	exp3	anom2
Patient Event Sequencing	exp3	anom4 / anom5	exp3	anom2
Patient Facts	exp3	anom4 / anom5	exp3	anom2
Concept Set Distribution	exp3	anom4 / anom5	exp3	anom2
Clinical Events and Specialty Agreement	exp3	anom4 / anom5	exp3	anom2
Patient Record Consistency	exp3	anom4 / anom5	exp3	anom2
Sensitivity to Selection Criteria				

The module: *Concept Set Distribution* (Figure 5) shows that the most common diagnoses used for pain in the dataset from Site A map to the SNOMED-CT terms 26942 (*Hemoglobin SS disease with crisis*), 4216915 (*Hemoglobin S sickling disorder with crisis*), and 4329041 (*Pain*), using OMOP standard vocabularies in our database. The latter two concepts, in contrast with the first, are not specific to more severe variants of sickle cell disease.

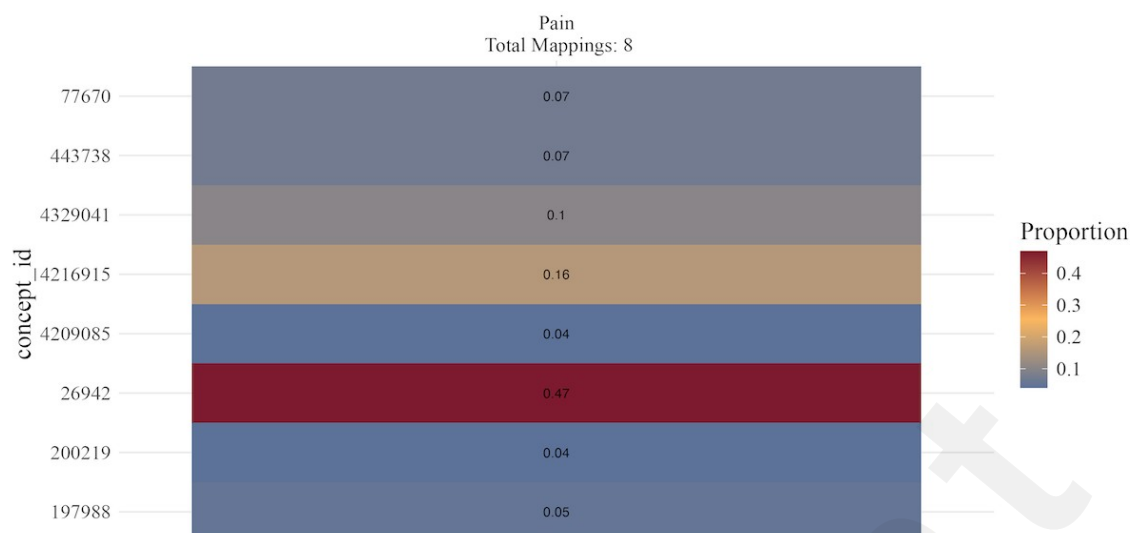
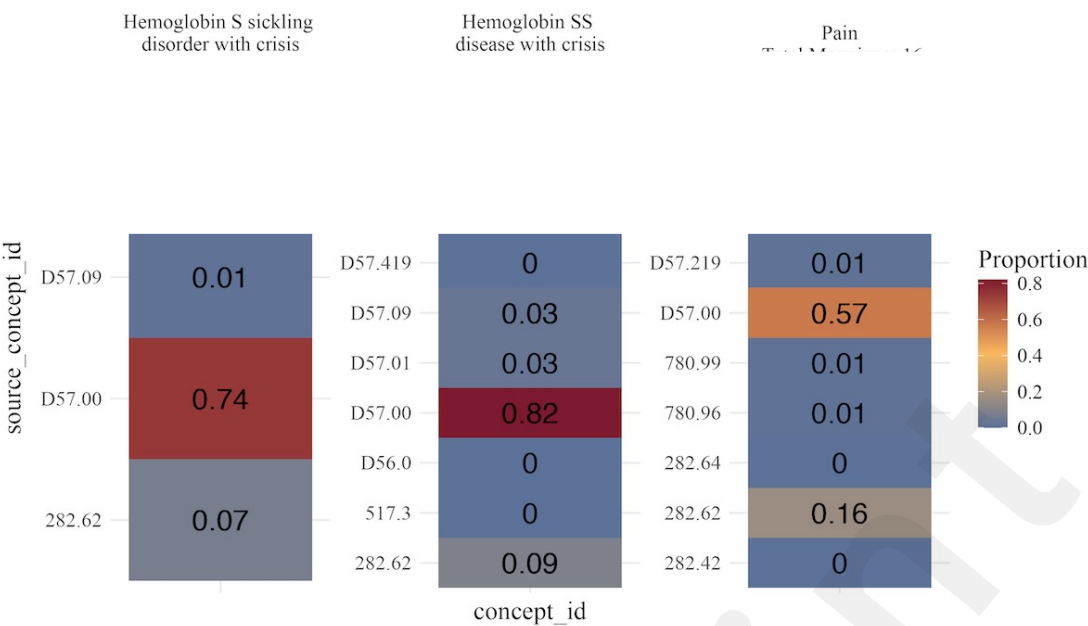


Figure 5. Diagnosis terms in analytic data from a user-supplied codeset representing “Pain”. For readability, the SNOMED-CT terms present in the dataset are shown here as codes rather than text (cf. <https://athena.ohdsi.org>). The color of each cell reflects the relative frequency with which each term occurs in the data.

In addition to evaluating the scope of a codeset, it is often important to know by what path the concepts in the codeset were reached. The **Source and Concept Vocabularies** module (Figure 6) demonstrates how source concepts, that is, data in the dataset designated as reflecting the original source systems, map to the terms used in the analytic models. Figure 6 shows each of six SNOMED-CT terms that may indicate SCD-related pain, with vocabulary mappings from ICD-9-CM and ICD-10-CM, a common constraint on diagnosis entry in EHR systems. This graph shows that the most common mapping to *Hemoglobin S sickling disorder with crisis* and *Pain* are diagnoses specific to the SS variant of sickle cell disease (e.g. ICD-10-CM *D57.00: Hb-SS disease with crisis, unspecified*), but the terms expressed in the analytic variable are inexact and lose specificity. From a data quality perspective, these mappings imply that many patients are experiencing pain typically associated with the SS variant, but that is not being fully expressed in the analytic coding, which should be improved or the analyses should rely on other data to adjust for disease variant. From a clinical perspective, patients with SCD-SS are at risk of the most severe complications, which may be an important consideration for study analyses. For example, it may correlate with the frequent usage of hydromorphone, a more potent opioid (Figures 3 and 4). Taken together, these data indicate that patients at Site A may be more complex and have more severe disease, and this should be accounted for in case definitions and analytic design. It also indicates that Site A may have some erroneous mappings to SNOMED-CT concepts, which may misclassify certain outcomes. For example, the term *Pain* in the analytic vocabulary is non-specific and unrelated to sickle cell disease, but the modules indicate that this may capture pain due to Hb-SS specifically.



Discussion

Principal Results

The tools we have implemented make contributions to better and more systematic research in several ways. First, the checks are driven by a theoretically sound framework that integrates clinical and technical considerations, while retaining opportunity to configure checks to meet the requirements of the user. Second, results are delivered with layered options for interactivity to better interpret output, and modules addressing related goals to support more thorough investigation. Third, we have packaged the modules to reduce the risk of error during per-study re-implementation. Fourth, adoption of common tools enables reproducibility and standardization of SSDQA checks for dissemination in a wider community, in keeping with FAIR principles [59, 60]. Importantly, it further facilitates transparency about data beyond simple descriptions when sharing of datasets is not possible due to patient privacy considerations. Consumers of research results are better able to understand data characteristics that may bear on applicability of the results to their circumstances. Similarly, other researchers can account for differences in data source in order to validate or augment previous work. For example, complex differences in utilization of the health system across cohorts may bias study outcomes, which is only crudely reported in current publications. Finally, by providing of a model that structures user interaction, standardizes approaches to visualizations, and uses consistent calling conventions across packaged modules, we facilitate extension of the current modules by us or others.

Limitations

Nonetheless, several limitations should be noted. First, the universe of potential SSDQA testing is quite large, and while the model is designed for expansion, the current modules cannot cover all possible requirements. However, SSDQA for any single study will exercise only a subset of those possibilities; our current work covers most high frequency cases. Second, use of DQ tools requires an interdisciplinary team comprising technical, methodologic, and clinical domain expertise to effectively design and interpret SSDQA. The modules facilitate the process, but do not replace collaborative interpretation and response. Third, the diversity of modules and configurations, and the data density in results displays, require familiarization in order to make most effective use of the tools. We have attempted to reduce cognitive burden by use of consistent visualizations and configuration methods and good documentation practices, but additional educational materials will be needed to reach a wider audience. We are currently engaged in constructing additional guidance for both DQ design and results interpretation. Fourth, most modules require the data to be in the OMOP or PCORnet[®] CDMs to support module reuse. This may create a barrier to adoption, especially for smaller studies, but institutional availability of CDM-transformed data has risen rapidly, and template code for transformation to and between CDMs is becoming more available.

Comparison with Prior Work

In-depth understanding of data quality is a prerequisite for effective use of clinical data, especially at large scale; the need for better definition in this field has been repeatedly recognized. We present a systematic approach to evaluating study-specific data quality that bridges theory and application to construct modules that are standardized and reproducible. Our work serves as a complement to existing tools such as the OHDSI Data Quality Dashboard[17], PCORnet[®] Data Curation[27], and other network data quality programs [26, 28, 30, 48, 61], which are usually applied to an entire database. While many study teams conduct exploratory analysis addressing problems such as missingness, these are usually re-implemented for each study. Our goal is not to replace the motivation for current practice, but to build on what are often incomplete evaluations that can benefit from better foundation and informative reporting. The flexibility offered in this model allows teams to identify additional aspects of SSDQA that may benefit their study, and tailor the checks to their study requirements. It also fosters collaboration between study teams to identify data quality issues that can inform design decisions.

Conclusions

Effective SSDQA is a critical but complex part of conducting valid research using clinical data. As with other steps of the research process, SSDQA benefits from evolving methodologic practices. The model described here, and the availability of tools that simplify adoption of good practice, provide an opportunity to improve SSDQA execution and reporting. Additional work will undoubtedly be needed as clinical research methods and data sources continue to advance.

Acknowledgments

The authors are grateful for the contributions of data scientists and researchers who shared their experience assessing study-specific data quality. We are also grateful to the patients, families, and clinicians who make possible large-scale learning in networks like PEDSnet about ways to improve health. This work was funded by the Patient-Centered Outcomes Research Institute (PCORI) award

ME-2020C3-21199. The funder played no role in the design of the study; methods development; collection, analysis, or interpretation of data; writing of the manuscript; or the decision to submit for publications. The views expressed here do not represent those of PCORI or participating institutions.

Conflicts of Interest

Jennifer Sciolla is a current member of the Ronald McDonald House Charities of Greater Delaware as well as a Consultant for the Association of Child Life Professionals. Clair Blacketer is an employee of Johnson & Johnson and holds stock and stock options. These conflicts did not interfere with the work performed in this project.

Abbreviations

CDM – common data model; CKD – chronic kidney disease; DQ – data quality; DQA – data quality assessment; EMA – European Medicines Agency; EHR – electronic health record; FAIR – findable, accessible, interoperable, reusable; FDA – Food and Drug Administration; Hb-SS – Hemoglobin SS; ICD-9-CM – International Classification of Diseases, Ninth Revision, Clinical Modification; ICD-10-CM – International Classification of Diseases, Tenth Revision, Clinical Modification; OMOP – Observational Medical Outcomes Partnership; SCD – sickle cell disease; SNOMED-CT – Systematized Nomenclature of Medicine – Clinical Terms; SSDQA – study-specific data quality assessment

Data Availability

The clinical data underlying the results in this manuscript are not available, as they contain patient-level protected health information (PHI). The software used to create the visualizations can be found on GitHub[58].

References

1. Forrest CB, Margolis PA, Bailey LC, Marsolo K, Del Beccaro MA, Finkelstein JA, et al. PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc.* 2014 Jul-Aug;21(4):602-6. PMID: 24821737. doi: 10.1136/amiajnl-2014-002743.
2. Forrest CB, McTigue KM, Hernandez AF, Cohen LW, Cruz H, Haynes K, et al. PCORnet(R) 2020: current state, accomplishments, and future directions. *J Clin Epidemiol.* 2021 Jan;129:60-7. PMID: 33002635. doi: 10.1016/j.jclinepi.2020.09.036.
3. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): Rationale, design, infrastructure, and deployment. *J Am Med Inform Assoc.* 2021 Mar 1;28(3):427-43. PMID: 32805036. doi: 10.1093/jamia/ocaa196.
4. All of Us Research Program I, Denny JC, Rutter JL, Goldstein DB, Philippakis A, Smoller JW, et al. The "All of Us" Research Program. *N Engl J Med.* 2019 Aug 15;381(7):668-76. PMID: 31412182. doi: 10.1056/NEJMSr1809937.
5. Tarabichi Y, Frees A, Honeywell S, Huang C, Naidech AM, Moore JH, et al. The Cosmos Collaborative: A Vendor-Facilitated Electronic Health Record Data Aggregation Platform. *ACI open.*

2021 Jan;5(1):e36-e46. PMID: 35071993. doi: 10.1055/s-0041-1731004.

6. Drozda JP, Jr., Graham J, Muhlestein JB, Tcheng JE, Roach J, Forsyth T, et al. Multi-institutional distributed data networks for real-world evidence about medical devices: building unique device identifiers into longitudinal data (BUILD). *JAMIA Open*. 2022 Jul;5(2):ooac035. PMID: 35663113. doi: 10.1093/jamiaopen/ooac035.

7. Pedrera-Jimenez M, Garcia-Barrio N, Hernandez-Ibarburu G, Baselga B, Blanco A, Calvo-Boyer F, et al. Building an i2b2-Based Population Repository for COVID-19 Research. *Studies in health technology and informatics*. 2022 May 25;294:287-91. PMID: 35612078. doi: 10.3233/SHTI220460.

8. Ohno-Machado L, Agha Z, Bell DS, Dahm L, Day ME, Doctor JN, et al. pSCANNER: patient-centered Scalable National Network for Effectiveness Research. *J Am Med Inform Assoc*. 2014 Jul-Aug;21(4):621-6. PMID: 24780722. doi: 10.1136/amiajnl-2014-002751.

9. Desai RJ, Marsolo K, Smith J, Carrell D, Penfold R, Pillai HS, et al. The FDA Sentinel Real World Evidence Data Enterprise (RWE-DE). *Pharmacoepidemiol Drug Saf*. 2024 Oct;33(10):e70028. PMID: 39385712. doi: 10.1002/pds.70028.

10. Joshua Lin K, Jin Y, Gagne J, Glynn RJ, Murphy SN, Tong A, et al. Longitudinal Data Discontinuity in Electronic Health Records and Consequences for Medication Effectiveness Studies. *Clin Pharmacol Ther*. 2022 Jan;111(1):243-51. PMID: 34424534. doi: 10.1002/cpt.2400.

11. Koffman L, Levis AW, Arterburn D, Coleman KJ, Herrinton LJ, Cooper J, et al. Investigating Bias from Missing Data in an Electronic Health Records-Based Study of Weight Loss After Bariatric Surgery. *Obes Surg*. 2021 May;31(5):2125-35. PMID: 33462670. doi: 10.1007/s11695-021-05226-y.

12. Boyd AD, Gonzalez-Guarda R, Lawrence K, Patil CL, Ezenwa MO, O'Brien EC, et al. Potential bias and lack of generalizability in electronic health record data: reflections on health equity from the National Institutes of Health Pragmatic Trials Collaboratory. *J Am Med Inform Assoc*. 2023 Aug 18;30(9):1561-6. PMID: 37364017. doi: 10.1093/jamia/ocad115.

13. Hubbard RA, Lett E, Ho GYF, Chubak J. Characterizing Bias Due to Differential Exposure Ascertainment in Electronic Health Record Data. *Health Serv Outcomes Res Methodol*. 2021 Sep;21(3):309-23. PMID: 34366704. doi: 10.1007/s10742-020-00235-3.

14. Harton J, Mitra N, Hubbard RA. Informative presence bias in analyses of electronic health records-derived data: a cautionary note. *J Am Med Inform Assoc*. 2022 Jun 14;29(7):1191-9. PMID: 35438796. doi: 10.1093/jamia/ocac050.

15. Haneuse S, Arterburn D, Daniels MJ. Assessing Missing Data Assumptions in EHR-Based Studies: A Complex and Underappreciated Task. *JAMA Netw Open*. 2021 Feb 1;4(2):e210184. PMID: 33635321. doi: 10.1001/jamanetworkopen.2021.0184.

16. Weiskopf NG, Dorr DA, Jackson C, Lehmann HP, Thompson CA. Healthcare utilization is a collider: an introduction to collider bias in EHR data reuse. *J Am Med Inform Assoc*. 2023 Apr 19;30(5):971-7. PMID: 36752649. doi: 10.1093/jamia/ocad013.

17. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. *Journal of the American Medical Informatics Association*. 2021;28(10):2251-7. doi: 10.1093/jamia/ocab132.

18. Castellanos EH, Wittmershaus BK, Chandwani S. Raising the Bar for Real-World Data in Oncology: Approaches to Quality Across Multiple Dimensions. *JCO Clin Cancer Inform*. 2024 Jan;8:e2300046. PMID: 38241599. doi: 10.1200/CCI.23.00046.

19. Feder SL. Data Quality in Electronic Health Records Research: Quality Domains and Assessment Methods. *West J Nurs Res*. 2018 May;40(5):753-66. PMID: 28322657. doi: 10.1177/0193945916689084.

20. Huser V, Li X, Zhang Z, Jung S, Park RW, Banda J, et al. Extending Achilles Heel Data Quality Tool with New Rules Informed by Multi-Site Data Quality Comparison. *Studies in health technology and informatics*. 2019 Aug 21;264:1488-9. PMID: 31438195. doi: 10.3233/SHTI190498.

21. Johnson SG, Speedie S, Simon G, Kumar V, Westra BL. Quantifying the Effect of Data

- Quality on the Validity of an eMeasure. *Appl Clin Inform*. 2017 Oct;8(4):1012-21. PMID: 29241241. doi: 10.4338/ACI-2017-03-RA-0042.
22. Kapsner LA, Kampf MO, Seuchter SA, Kamdje-Wabo G, Grading T, Ganslandt T, et al. Moving Towards an EHR Data Quality Framework: The MIRACUM Approach. *Stud Health Technol Inform*. 2019 Sep 3;267:247-53. PMID: 31483279. doi: 10.3233/SHTI190834.
23. Khare R, Utidjian LH, Razzaghi H, Soucek V, Burrows E, Eckrich D, et al. Design and Refinement of a Data Quality Assessment Workflow for a Large Pediatric Research Network. *EGEMS (Wash DC)*. 2019 Aug 1;7(1):36. PMID: 31531382. doi: 10.5334/egems.294.
24. Lee K, Weiskopf N, Pathak J. A Framework for Data Quality Assessment in Clinical Research Datasets. *AMIA Annu Symp Proc*. 2017;2017:1080-9. PMID: 29854176.
25. Mohamed Y, Song X, McMahon TM, Sahil S, Zozus M, Wang Z, et al. Tailoring Rule-Based Data Quality Assessment to the Patient-Centered Outcomes Research Network (PCORnet) Common Data Model (CDM). *AMIA Annu Symp Proc*. 2022;2022:775-84. PMID: 37128433.
26. Pfaff ER, Girvin AT, Gabriel DL, Kostka K, Morris M, Palchuk MB, et al. Synergies between centralized and federated approaches to data quality: a report from the national COVID cohort collaborative. *J Am Med Inform Assoc*. 2022 Mar 15;29(4):609-18. PMID: 34590684. doi: 10.1093/jamia/ocab217.
27. Qualls LG, Phillips TA, Hammill BG, Topping J, Louzao DM, Brown JS, et al. Evaluating Foundational Data Quality in the National Patient-Centered Clinical Research Network (PCORnet(R)). *EGEMS (Wash DC)*. 2018 Apr 13;6(1):3. PMID: 29881761. doi: 10.5334/egems.199.
28. Sengupta S, Bachman D, Laws R, Saylor G, Staab J, Vaughn D, et al. Data Quality Assessment and Multi-Organizational Reporting: Tools to Enhance Network Knowledge. *EGEMS (Wash DC)*. 2019 Mar 29;7(1):8. PMID: 30972357. doi: 10.5334/egems.280.
29. Sidky H, Young JC, Girvin AT, Lee E, Shao YR, Hotaling N, et al. Data quality considerations for evaluating COVID-19 treatments using real world data: learnings from the National COVID Cohort Collaborative (N3C). *BMC Med Res Methodol*. 2023 Feb 17;23(1):46. PMID: 36800930. doi: 10.1186/s12874-023-01839-2.
30. Taggart J, Liaw ST, Yu H. Structured data quality reports to improve EHR data quality. *Int J Med Inform*. 2015 Dec;84(12):1094-8. PMID: 26480872. doi: 10.1016/j.ijmedinf.2015.09.008.
31. Weiskopf NG, Bakken S, Hripcsak G, Weng C. A Data Quality Assessment Guideline for Electronic Health Record Data Reuse. *EGEMS (Wash DC)*. 2017 Sep 4;5(1):14. PMID: 29881734. doi: 10.5334/egems.218.
32. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc*. 2013 Jan 1;20(1):144-51. PMID: 22733976. doi: 10.1136/amiajnl-2011-000681.
33. Lewis AE, Weiskopf N, Abrams ZB, Foraker R, Lai AM, Payne PRO, et al. Electronic health record data quality assessment and tools: a systematic review. *J Am Med Inform Assoc*. 2023 Sep 25;30(10):1730-40. PMID: 37390812. doi: 10.1093/jamia/ocad120.
34. Ozonze O, Scott PJ, Hopgood AA. Automating Electronic Health Record Data Quality Assessment. *Journal of Medical Systems*. 2023;47(1). doi: 10.1007/s10916-022-01892-2.
35. Girman CJ, Ritchey ME, Lo Re V, 3rd. Real-world data: Assessing electronic health records and medical claims data to support regulatory decision-making for drug and biological products. *Pharmacoepidemiol Drug Saf*. 2022 Jul;31(7):717-20. PMID: 35471704. doi: 10.1002/pds.5444.
36. Raman SR, O'Brien EC, Hammill BG, Nelson AJ, Fish LJ, Curtis LH, et al. Evaluating fitness-for-use of electronic health records in pragmatic clinical trials: reported practices and recommendations. *J Am Med Inform Assoc*. 2022 Apr 13;29(5):798-804. PMID: 35171985. doi: 10.1093/jamia/ocac004.
37. Reynolds MW, Bourke A, Dreyer NA. Considerations when evaluating real-world data quality in the context of fitness for purpose. *Pharmacoepidemiol Drug Saf*. 2020 Oct;29(10):1316-8.

PMID: 32374042. doi: 10.1002/pds.5010.

38. Gatto NM, Reynolds RF, Campbell UB. A Structured Preapproval and Postapproval Comparative Study Design Framework to Generate Valid and Transparent Real-World Evidence for Regulatory Decisions. *Clin Pharmacol Ther.* 2019 Jul;106(1):103-15. PMID: 31025311. doi: 10.1002/cpt.1480.

39. Gatto NM, Vittoe SE, Rubinstein E, Reynolds RF, Campbell UB. A Structured Process to Identify Fit-for-Purpose Study Design and Data to Generate Valid and Transparent Real-World Evidence for Regulatory Uses. *Clin Pharmacol Ther.* 2023 Jun;113(6):1235-9. PMID: 36871138. doi: 10.1002/cpt.2883.

40. Wang SV, Pinheiro S, Hua W, Arlett P, Uyama Y, Berlin JA, et al. STaRT-RWE: structured template for planning and reporting on the implementation of real world evidence studies. *Bmj.* 2021 Jan 12;372:m4856. PMID: 33436424. doi: 10.1136/bmj.m4856.

41. Razzaghi H, Greenberg J, Bailey LC. Developing a systematic approach to assessing data quality in secondary use of clinical data based on intended use. *Learn Health Syst.* 2022 Jan;6(1):e10264. PMID: 35036548. doi: 10.1002/lrh2.10264.

42. Denburg MR, Razzaghi H, Goodwin Davies AJ, Dharnidharka V, Dixon BP, Flynn JT, et al. The Preserving Kidney Function in Children With CKD (PRESERVE) Study: Rationale, Design, and Methods. *Kidney Med.* 2023 Nov;5(11):100722. PMID: 37965485. doi: 10.1016/j.xkme.2023.100722.

43. Razzaghi H, Goodwin Davies A, Boss S, Bunnell HT, Chen Y, Chrischilles EA, et al. Systematic data quality assessment of electronic health record data to evaluate study-specific fitness: Report from the PRESERVE research study. *PLOS Digit Health.* 2024 Jun;3(6):e0000527. PMID: 38935590. doi: 10.1371/journal.pdig.0000527.

44. Diaz-Garelli JF, Bernstam EV, Lee M, Hwang KO, Rahbar MH, Johnson TR. DataGauge: A Practical Process for Systematically Designing and Implementing Quality Assessments of Repurposed Clinical Data. *EGEMS (Wash DC).* 2019 Jul 25;7(1):32. PMID: 31367649. doi: 10.5334/egems.286.

45. Engel N, Wang H, Jiang X, Lau CY, Patterson J, Acharya N, et al. EHR Data Quality Assessment Tools and Issue Reporting Workflows for the 'All of Us' Research Program Clinical Data Research Network. *AMIA Annu Symp Proc.* 2022;2022:186-95. PMID: 35854725.

46. Kamdje Wabo G, Moorthy P, Siegel F, Seuchter SA, Ganslandt T. Evaluating and Enhancing the Fitness-for-Purpose of Electronic Health Record Data: Qualitative Study on Current Practices and Pathway to an Automated Approach Within the Medical Informatics for Research and Care in University Medicine Consortium. *JMIR Med Inform.* 2024 Aug 19;12:e57153. PMID: 39158950. doi: 10.2196/57153.

47. Mang JM, Seuchter SA, Gulden C, Schild S, Kraska D, Prokosch HU, et al. DQAgui: a graphical user interface for the MIRACUM data quality assessment tool. *BMC Med Inform Decis Mak.* 2022 Aug 11;22(1):213. PMID: 35953813. doi: 10.1186/s12911-022-01961-z.

48. Mohamed Y, Song X, McMahon TM, Sahil S, Zozus M, Wang Z, et al. Electronic health record data quality variability across a multistate clinical research network. *J Clin Transl Sci.* 2023;7(1):e130. PMID: 37396818. doi: 10.1017/cts.2023.548.

49. Pedrera-Jimenez M, Garcia-Barrio N, Rubio-Mayo P, Maestro G, Lalueza A, Garcia-Reyne A, et al. Making EHRs Trustable: A Quality Analysis of EHR-Derived Datasets for COVID-19 Research. *Studies in health technology and informatics.* 2022 May 25;294:164-8. PMID: 35612049. doi: 10.3233/shti220430.

50. Ehsani-Moghaddam B, Martin K, Queenan JA. Data quality in healthcare: A report of practical experience with the Canadian Primary Care Sentinel Surveillance Network data. *Health Inf Manag.* 2021 Jan-May;50(1-2):88-92. PMID: 31805788. doi: 10.1177/1833358319887743.

51. Garies S, McBrien K, Quan H, Manca D, Drummond N, Williamson T. A data quality assessment to inform hypertension surveillance using primary care electronic medical record data

- from Alberta, Canada. BMC Public Health. 2021 Feb 2;21(1):264. PMID: 33530975. doi: 10.1186/s12889-021-10295-w.
52. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015 Feb 10. PMID: 25670757. doi: 10.1093/jamia/ocu023.
53. R Development Core Team. R: A Language and Environment for Statistical Computing. 2019.
54. Mitka M. New HIPAA rule aims to improve privacy and security of patient records. Jama. 2013 Mar 6;309(9):861-2. PMID: 23462763. doi: 10.1001/jama.2013.806.
55. Kahn MG, Callahan TJ, Barnard J, Bauck AE, Brown J, Davidson BN, et al. A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data. EGEMS (Wash DC). 2016;4(1):1244. PMID: 27713905. doi: 10.13063/2327-9214.1244.
56. al We. Welcome to the Tidyverse. Journal of Open Source Software. 2019;4(43):1686. doi: <https://doi.org/10.21105/joss.01686>.
57. PCORnet. Common Data Model (CDM) Specification, Version 6.1. [cited 2024]; Available from: https://pcornet.org/wp-content/uploads/2023/04/PCORnet-Common-Data-Model-v61-2023_04_031.pdf.
58. Study Specific Data Quality Analysis. 2024 [cited 2024]; Packages to execute study specific data quality]. Available from: <https://github.com/ssdqa>.
59. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. PMID: 26978244. doi: 10.1038/sdata.2016.18.
60. Haux C, Knaup P. Using FAIR Metadata for Secondary Use of Administrative Claims Data. Studies in health technology and informatics. 2019 Aug 21;264:1472-3. PMID: 31438187. doi: 10.3233/SHTI190490.
61. Estiri H, Stephens K. DQ(e)-v: A Database-Agnostic Framework for Exploring Variability in Electronic Health Record Data Across Time and Site Location. EGEMS (Wash DC). 2017 May 10;5(1):3. PMID: 29930954. doi: 10.13063/2327-9214.1277.

Supplementary Files

Multimedia Appendixes

Supplemental Table 1: Questions asked to data scientists about data quality issues on their studies.

URL: <http://asset.jmir.pub/assets/277363761872def5b38727aef30c6f23.docx>

Supplemental Table 2: List of Participants in Expert Review Panel.

URL: <http://asset.jmir.pub/assets/bfb371902bd8c916783063a485d9dadd.docx>

Supplemental Table 3: Interview guide with members of the Expert Review Panel.

URL: <http://asset.jmir.pub/assets/4823e4375f6ea0b0572515fa7186a6dc.docx>

Supplemental Table 4: List of terms for Check Category, Data Quality Probe, Clinical Goal or Assessment, and Analysis Level.

URL: <http://asset.jmir.pub/assets/ad07b0b342408ec1578ad337f86faa4f.docx>