Predicting Causes of Data Quality Issues in a Clinical Data Research Network

Ritu Khare, PhD[1], Byron J. Ruth, MS[1], Matthew Miller, MS[1], Joshua Tucker[1], BS, Levon H. Utidjian, MD, MBI[1,2], Hanieh Razzaghi, MPH[1], Nandan Patibandla, MS[3], Evanette K. Burrows, BS[1], L. Charles Bailey, MD, PhD[1,2]

[1]Departments of Pediatrics and Biomedical & Health Informatics, The Children's Hospital of Philadelphia, Philadelphia, PA, 19104; [2]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, 19104; Information Services Department, [3]Children’s Hospital Boston, Boston, MA, 02115

Abstract

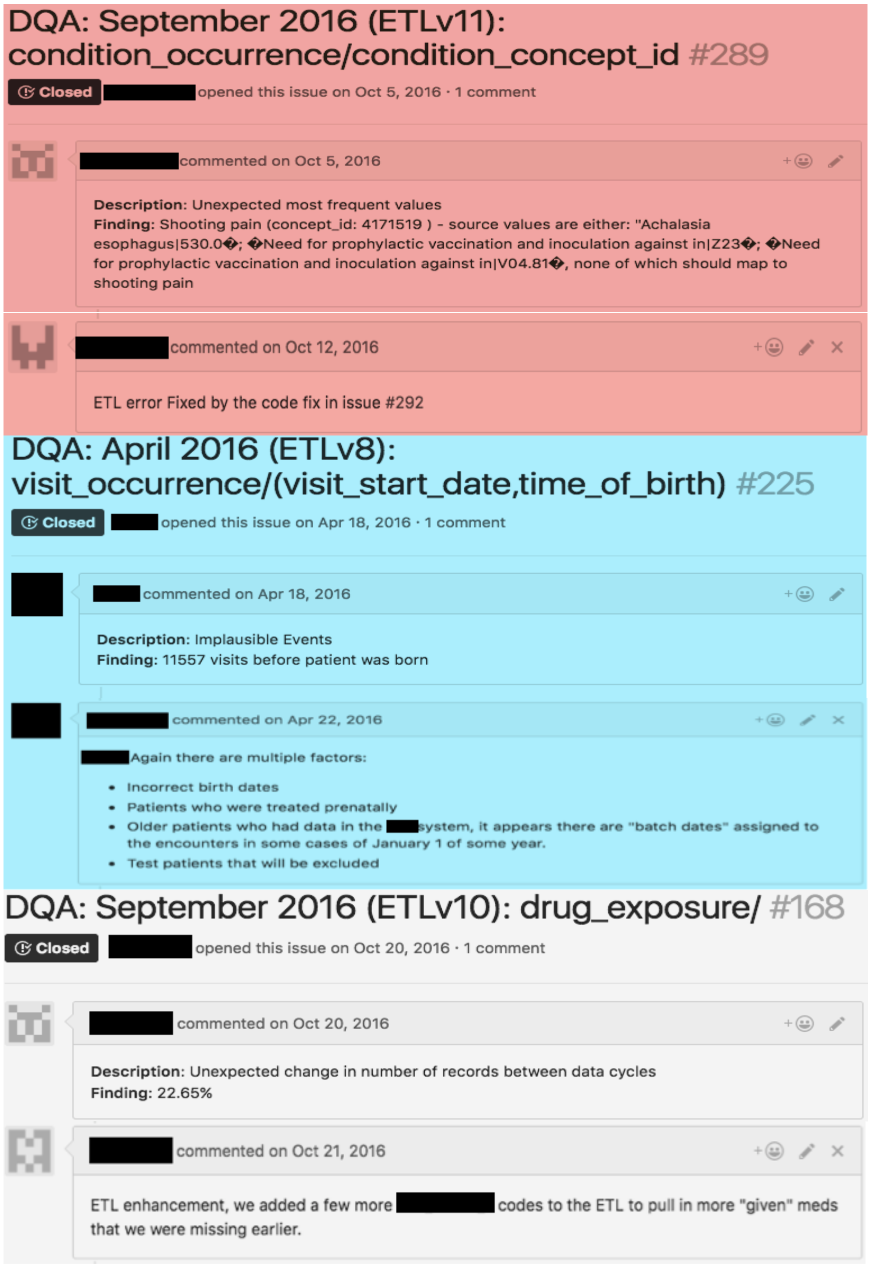
Clinical data research networks (CDRNs) invest substantially in identifying and investigating data quality problems. While identification is largely automated, the investigation and resolution are carried out manually at individual institutions. In the PEDSnet CDRN, we found that only approximately 35% of the identified data quality issues are resolvable as they are caused by errors in the extract-transform-load (ETL) code. Nonetheless, with no prior knowledge of issue causes, partner institutions end up spending significant time investigating issues that represent either inherent data characteristics or false alarms. This work investigates whether the causes (ETL, Characteristic, or False alarm) can be predicted before spending time investigating issues. We trained a classifier on the metadata from 10,281 real-world data quality issues, and achieved a cause prediction F1-measure of up to 90%. While initially tested on PEDSnet, the proposed methodology is applicable to other CDRNs facing similar bottlenecks in handling data quality results.

Introduction and Background

Clinical data research networks (CDRNs) transform electronic health record (EHR) data from multiple institutions into common data models, and make that data available, either in a centralized or a distributed fashion, to conduct a wide range of scientific studies.[1-3] Given that EHRs are designed for clinical operations rather than research use, one of the most critical aspects in building a CDRN is to ensure that the aggregated clinical data are “high-quality” or “ready for research use.”[2-7] CDRN datasets are typically built in an iterative fashion. The data coordinating center executes certain data characterization or validation modules on the dataset to identify any data quality problems; these problems are communicated to the contributing institutions that investigate and resolve the problems, and generate the improved datasets. The investigation of data quality problems is a complex process that involves local replication of the problems, reviews of the relevant extract-transform-load (ETL) code, verification of data assumptions, and discussions about local data characteristics with the interdisciplinary team of clinicians, analysts, researchers, and administrative staff.

In the past, several studies have recommended techniques for conducting data quality assessments on EHR-derived datasets, such as using expert judgment, heuristics, knowledge of impossibilities, gold standard benchmarking, code reviews, conformance to value set domains, and computation of derived values,[2,5,8,9] and more recently Kahn et al. designed a comprehensive ontology to classify data quality checks.[9] However, handling, analysis, or classification of real-world data quality problems or issues, are largely undocumented.[10] Here, we present an automated approach that given a data quality issue, classifies the issue cause as “ETL” vs. “characteristic” vs. “false alarm,” to assist in prioritization and resolution of issues. The main contribution of this work is the use of supervised machine learning to predict the causes of data quality issues and achieve a promising performance.

In this study, we focus on a pediatric CDRN, PEDSnet, that aggregates EHR data from eight of the nation’s largest children’s hospitals[11,12] using the Observational Medical Outcomes Partnership (OMOP) common data model (CDM).[13] PEDSnet has invested substantial efforts in designing and implementing data quality “checks” to evaluate the validity of EHR-derived datasets and identify any data quality “issues” that indicate that the data could be inaccurate or difficult to use for some research purposes.[14] PEDSnet uses “GitHub issues” to report the data quality issues to individual sites[15] as shown in Figure 1. Once the issues are reported, the originating site’s first task is to determine whether the issue is an error in the ETL programming pipeline (Figure 1a), or represents a characteristic or inherent property of data such as EHR data entry error, administrative issues, source data incompleteness, institutional data anomalies, etc. (Figure 1b), or is a false alarm caused due to a programming bug in the PEDSnet data quality program or limitation of the data quality check (Figure 1c). The next task is to resolve any ETL-related issues for the next data submission. Based on a detailed analysis of the six most recent data cycles or iterations in PEDSnet, a partner site on an average investigates 26 new issues in each cycle, out of which only 35% represent resolvable problems (i.e. ETL category), as shown in Figure 2. In addition, based on the analysis of issue timelines in GitHub, we found that a data quality issue is open for 31 days on an average suggesting the potential duration of the issue investigation process; see Figure 3 for a distribution of GitHub issue duration across different types of causes. In sum, issue handling is a major bottleneck toward the iterative development of PEDSnet, as the partner sites need to resort to time-consuming and expensive processes for manual prioritization and investigation of individual issues. In this study, we examine whether the cause of a data quality issue can be predicted before delving into investigation, to help minimize issue fatigue and avoid spending time on issues that cannot be resolved, e.g. characteristic issues, or that should not have been reported at all, e.g. false alarms.



DQA: September 2016 (ETLv11): condition_occurrence/condition_concept_id #289

Closed [redacted]opened this issue on Oct 5, 2016 · 1 comment

[redacted]commented on Oct 5, 2016

Description: Unexpected most frequent values
Finding: Shooting pain (concept_id: 4171519) - source values are either: "Achalasia esophagus|530.0�; �Need for prophylactic vaccination and inoculation against in|Z23�; �Need for prophylactic vaccination and inoculation against in|V04.81�, none of which should map to shooting pain

[redacted]commented on Oct 12, 2016

ETL error Fixed by the code fix in issue #292

DQA: April 2016 (ETLv8): visit_occurrence/(visit_start_date,time_of_birth) #225

Closed [redacted]opened this issue on Apr 18, 2016 · 1 comment

[redacted] [redacted]commented on Apr 18, 2016

Description: Implausible Events
Finding: 11557 visits before patient was born

[redacted] [redacted]commented on Apr 22, 2016

[redacted]Again there are multiple factors:

- Incorrect birth dates
- Patients who were treated prenatally
- Older patients who had data in the [redacted]system, it appears there are "batch dates" assigned to the encounters in some cases of January 1 of some year.
- Test patients that will be excluded

DQA: September 2016 (ETLv10): drug_exposure/ #168

Closed [redacted]opened this issue on Oct 20, 2016 · 1 comment

[redacted]commented on Oct 20, 2016

Description: Unexpected change in number of records between data cycles
Finding: 22.65%

[redacted]commented on Oct 21, 2016

ETL enhancement, we added a few more [redacted]codes to the ETL to pull in more "given" meds that we were missing earlier.

Figure 1. GitHub screenshots of PEDSnet data quality issues illustrating different causes - top to bottom (a) ETL issue (in red), (b) Characteristic issue (in blue), (c) False alarm (in gray).

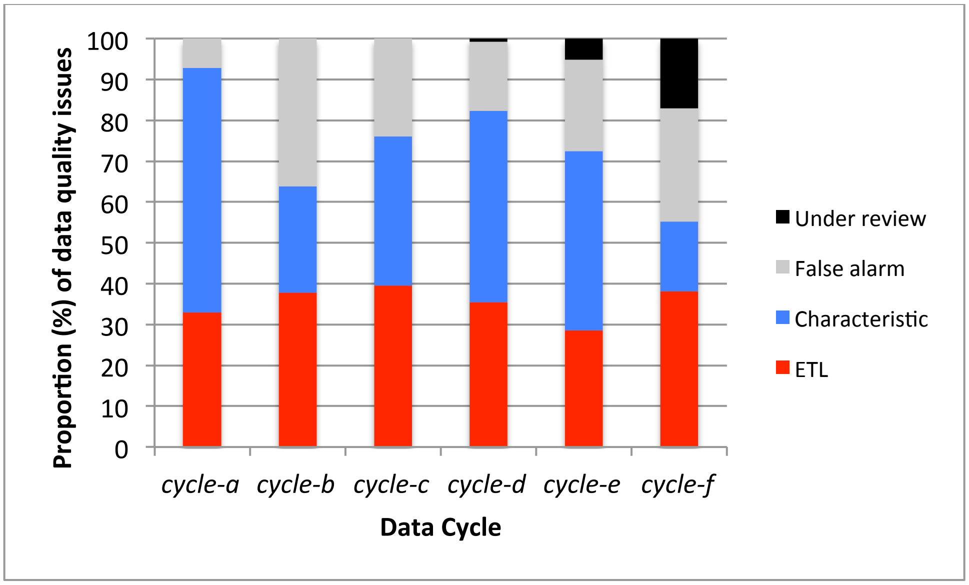


Figure 2. Longitudinal distribution of causes of data quality issues reported to PEDSnet sites

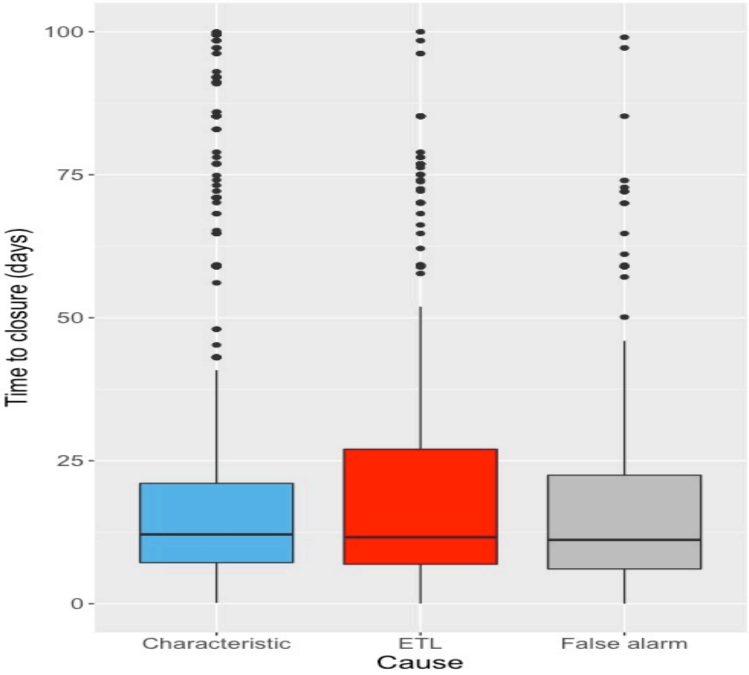


Figure 3. The Github (open – > close) duration of different classes of issues

Methods

As our data source, we use the PEDSnet data quality issue warehouse.[17] The warehouse contains metadata about data quality issues, and manually identified causes of those issues. The metadata includes the affected domain(s) and field(s) in the CDM, the tailored description of the issue, site information, the check type (Table 1) generating the issue, and the version of the CDM adopted for the given data cycle. It should be noted that the "characteristic" issues, once determined, get documented in the subsequent data cycles, but are not reported to the sites to avoid duplication of efforts.

Table 1. Some Examples of Data Quality Check Types and Issues in PEDSnet

Check Type (*alias*: Name)	**Example Data Quality Issues**
InconSource: Inconsistency with source	Distribution of NULL values in `race_source_value` does not match with the distribution of "No Information" concept in `race_concept_id` in `Person` table
InvalidValue: Value set violations	A non-standard concept used for populating the `condition_concept_id` field
UnexFact: Unexpected facts	A medication name entered into the `location.zip` field
ImplEvent: Implausible events	Found encounters with `visit_start_date` occurring after `visit_end_date`
CatOutlier: Categorical outlier	A patient with over 30,000 procedures
UnexTop: Unexpected most frequent values	"injection for contraceptive" as the most frequent procedure at a site
UnexDiff: Unexpected difference from the previous data cycle	Decrease in the number of deaths, or large increase (e.g. 2X) in the number of conditions
MissData: Missing data	Gestational age is not available for 70% of patients
MissFact: Missing expected facts	No "creatinine" lab record found in `measurement` table

We hypothesize that training a machine learning classifier, using meta-data about known issues, can help determine the cause of a data quality issue, and that the classifier can deliver performance sufficient to drive issue prioritization. As input, we selected a variety of features from the PEDSnet issue warehouse, intended to capture several aspects of an issue. Overall, 83 binary features were selected. The features types and instances are described below, and illustrated in Table 2.

- Domain: The CDM table where the issue was observed, e.g. `Person`, `Care_Site`, `Location`, `Death`, `Condition_occurrence`, `Visit_payer`, `Visit_occurrence`, `Procedure_occurrence`, `Measurement`, `Drug_exposure`, `Measurement_organism`, etc.
- Field Type: The type of field where the issue was observed, e.g, numerical fields, foreign keys, concept identifiers, source values, combination of fields, or others.
- Check Type: The type of data quality assessments conducted to identify the issue; some examples are shown in Table 1.
- Prevalence: The number of records affected by the issue, categorized as full (100%), high (30%-100%), medium (1%-30%), low (0%-1%), or unknown.
- Site: The site where the issue is observed, including one of the eight PEDSnet sites.
- CDM version upgrade: A boolean feature denoting whether the PEDSnet CDM version was upgraded since the previous data cycle.

Table 2. Features types and positive features for the example issues shown in Figure 1

Feature Types	**ETL issue (Fig. 1a)**	**Characteristic issue (Fig. 1b)**	**False alarm (Fig. 1c)**
Domain	`Condition_occurrence`	`Visit_occurrence`	`Drug_exposure`
Field Type	Concept identifier	Multiple	-
Check Type	*UnexTop*	*ImplEvent*	*UnexDiff*
Prevalence	Medium	Low	Medium
CDM version upgrade	No	Yes	No

We targeted two classification problems, binary (ETL vs. Non-ETL), and three-way (ETL vs. Characteristic vs. False alarm). We evaluated several classification methods including Naïve Bayes (NB), Decision tree (DT), Decision tree with boosting (DTB), k-Nearest neighbor (KNN), and support vector machine (SVM). We used Python implementation[16] of the classifiers, and used the datasets extracted from the PEDSnet issue warehouse for training. Prior to choosing specific configurations for these learners, a "model grid search" was performed using the

GridSearchCV algorithm to find the optimal set of parameters for each of the target learners as shown in Table 3. The search was performed on 80% of the data using a five-fold stratified training set. Each combination was evaluated against a hold-out set to score the model.

Table 3. The learned parameters for various classifiers using a grid search

Learner	**Parameters**
Decision tree + pruning (DT)	Max depth = 10
Decision tree + pruning + boosting (DT+B)	Max depth = 4 Estimators = 300
k-nearest neighbor (KNN)	K = 5 (binary), 3 (three-way)
Naïve Bayes (NB)	Class Priors = None

Support vector machine (SVM)	Kernel = linear
	Error term = 0.1
	Tolerance = 0.001

Results

We used the July 2017 version of the PEDSnet issue warehouse[17], which contains metadata on 11,434 data quality issues identified over a span of 30 months. We drew two datasets for experimentation, *all-issues-dataset* which includes 10,281 issues after filtering out the issues with unknown causes, and *unique-issues-dataset*, with 4,388 issues, that is a subset of *all-issues-dataset* prepared after excluding duplicate characteristic issues. The class label distributions across both datasets are: 14.37% (ETL), 81.78% (Characteristic), and 3.84% (False alarm); and 33.68% (ETL), 57.31% (Characteristic), and 9% (False alarm); respectively.

Figures 4 and 5 show the receiver operating characteristics (ROC) curves and performance measures for binary classification (ETL vs. Non-ETL), respectively. The results indicate that the k-nearest neighbor and decision tree with boosting algorithms could be promising choices for this problem. The classifiers trained on the *unique-issues-dataset* deliver higher F1 measure for the ETL issues, as compared to the *all-issues-dataset*, given the higher balancedness. Figure 6 shows the performance of classifiers on three-way classification problem. The performance for each class is higher than that of the binary classifiers. The classification performance on characteristic issues is higher than that on ETL or false alarms. This is most likely due to the availability of significantly higher training examples for characteristic issues in both the datasets.

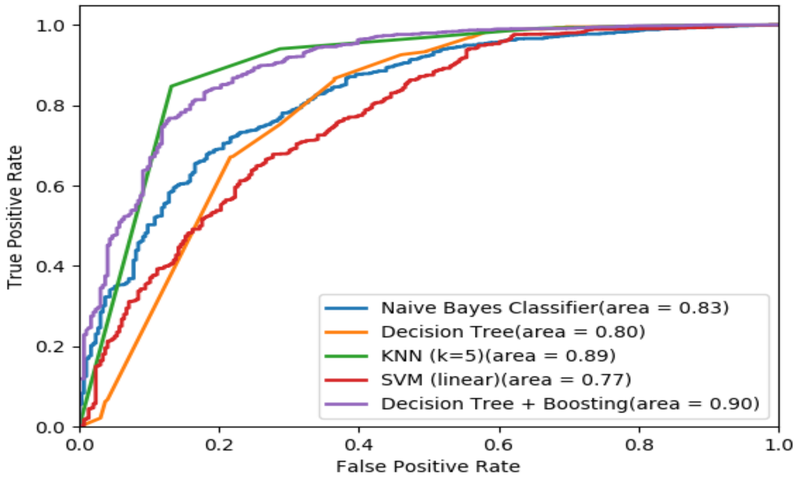


Figure 4. ROC curve for the binary (ETL vs. Non-ETL) classifier trained on *all-issues-dataset*

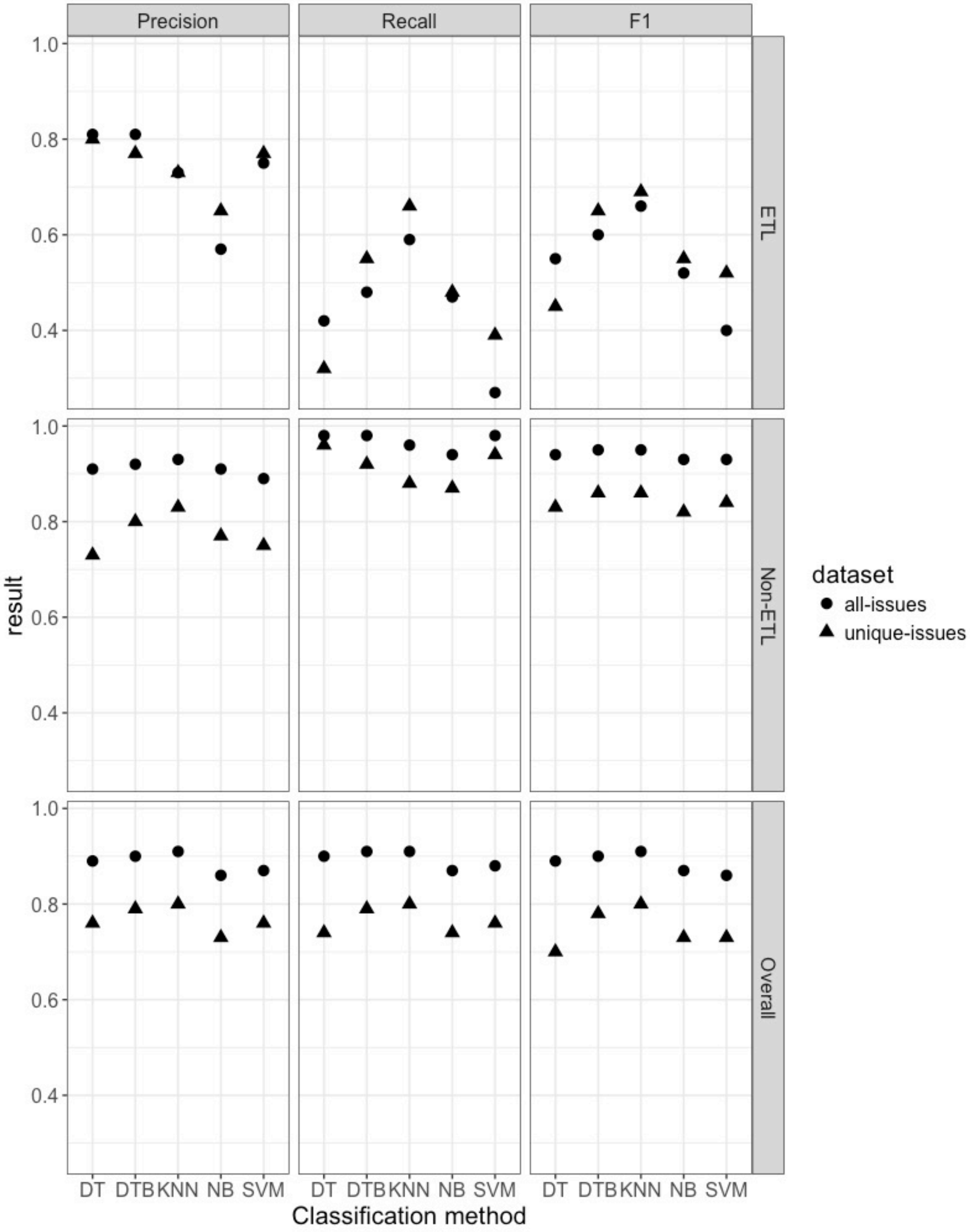


Figure 5. Performance measures for binary (ETL vs. Non-ETL) classification of issues

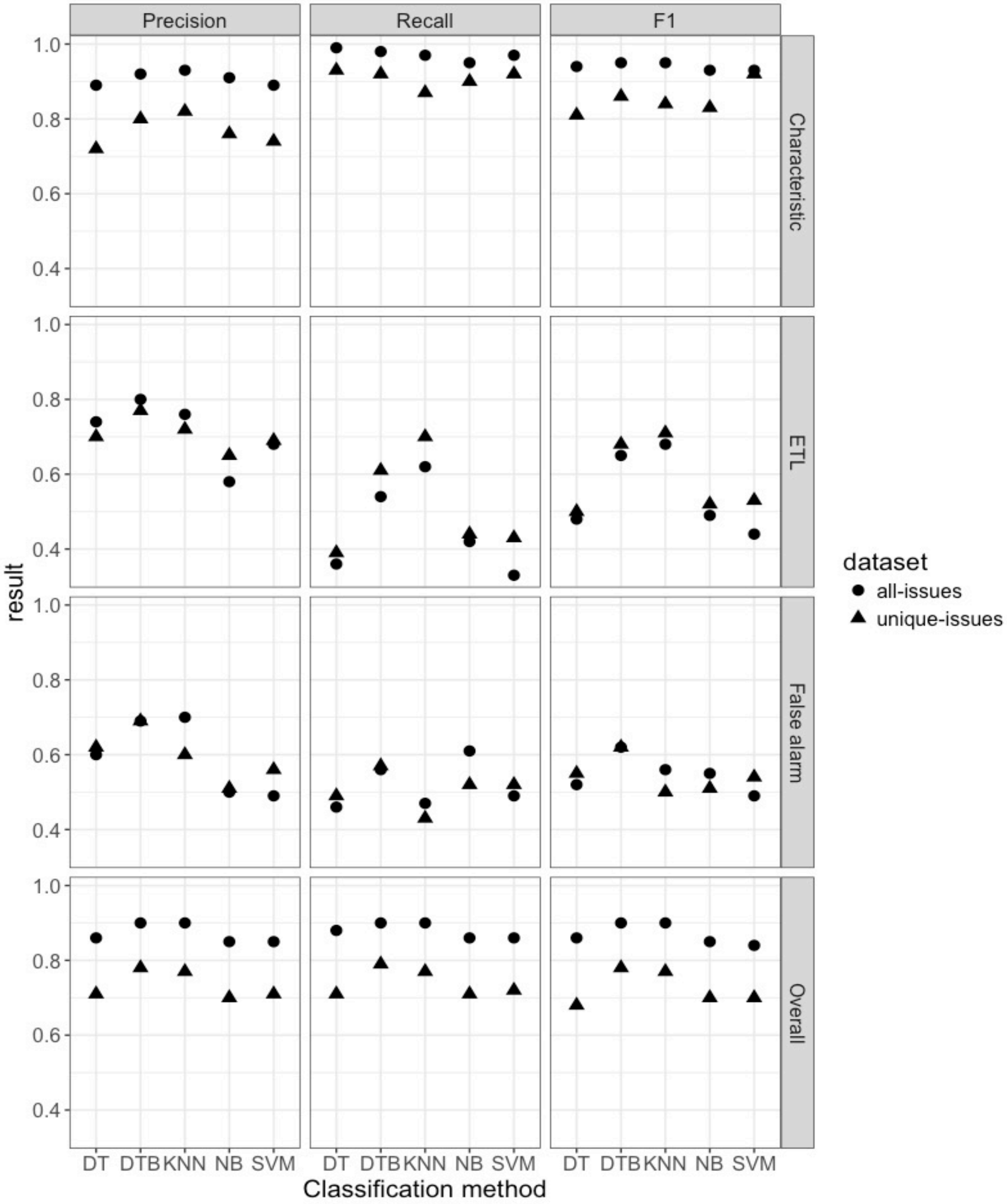


Figure 6. Performance measures for three-way (Characteristic, ETL, False Alarm) classification of issues

To further understand the results, we examined the types of issues that constitute the most frequent error cases in the *all-issues-dataset* using the Decision tree with boosting classifier (Table 4). In general, the issues that are frequently difficult to classify tend to be limited to five major check types representing candidates for further study. In the majority of error cases, the classifier could not determine whether *MissData* (missing data) for `drug_exposure` and `measurement` was due to inherent characteristics or ETL error. Both these domains represent some of the most evolving domains in PEDSnet wherein the sites are gradually populating various fields, and hence the fluctuations in causes of missingness in the past several cycles. Another difficult check type was *UnexDiff* (unexpected difference in the number of records between two data cycles) wherein it is difficult to determine whether the issue is due to an ETL error or due to a natural enlargement of site's dataset, i.e. false alarm.

Table 4. Most frequent error cases (Check type, domain), FP=false positive, FN= false negative

Cause Class	***ETL***		***Characteristic***		***Non-issue***	
	FP	***FN***	***FP***	***FN***	***FP***	***FN***
1	*UnexDiff*, `Measurement`	*MissData*, `drug_exposure`	*MissData*, `drug exposure`	*MissData*, `Measurement`	*UnexDiff*, `Procedure_occurrence`	*UnexDiff,* `Visit_occurrence`
2	*UnexDiff*, `Visit_occurrence`	*MissData*, `Measurement`	*MissData*, `Measurement`	*InvalidConID*, `Provider`	*UnexDiff*, `Drug_exposure`	*UnexDiff,* `Measurement`
3	*InvalidConID*, `Provider`	*MissConID*, `Drug_exposure`	*MissConID*, `Drug_exposure`	*MissData*, `Observation`	*MissData*, `Location`	*MissFact,* `care_site`

Discussion

Based on our experience of conducting iterative data quality assessments on a pediatric CDRN, we find that a majority (>60%) of the data quality issues should receive lower priority for investigation, as they are either false alarms or an inherent characteristic of data that cannot be altered or resolved. In this study, we have studied the cause prediction problem using machine learning classifier that, given a data quality issue, predicts the cause of the issue. The best performing classifier achieved a promising F1-measure of 0.9, and indicates the potential to save significant effort by the data generation teams. While this study was primarily driven by the efficiency challenges faced in PEDSnet and the proposed method was tested on the pediatric dataset, the methodology can be applied to benefit other CDRNs.

By conducting the experiments using several classifiers with different class configurations, we were able to identify strong candidates for real-world implementation and execution. While our interest primarily lies in accurately predicting ETL issues, the performance on ETL issues (F1-measure, 0.71) of all classifiers left substantial scope for improvement, e.g. further analysis of the frequent use cases identified through error analysis. In the future, we plan to extend this work using systematic feature selection, development of more granular causal classes, development of balanced datasets, and assessment of the impact of automatic predictions on user experience.

Acknowledgements

The authors would like to thank all the members of the PEDSnet data committee team, in particular, Daniel Eckrich, Janet Zahner, Kevin Matthews, Melody Kitzmiller, Richard Hoyt, Rob Brown, and Victoria Soucek, for participating in investigations of data quality issues and conducting discussions on GitHub. The authors are indebted to the patients and families contributing to the PEDSnet dataset. Research reported in this work was funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (CDRN-1306-01556).

References

1. Bailey LC, Milov DE, Kelleher K, Kahn MG, Del Beccaro M, Yu F, et al. Multi-Institutional Sharing of Electronic Health Record Data to Assess Childhood Obesity. PLoS One. 2013;8(6):e66192.

2. Brown JS, Kahn M, Toh S. Data quality assessment for comparative effectiveness research in distributed data networks. Med Care. 2013;51(8 Suppl 3):S22-9.

3. Kahn MG, Raebel MA, Glanz JM, Riedlinger K, Steiner JF. A pragmatic framework for single-site and multisite data quality assessment in electronic health record-based clinical research. Med Care. 2012;50 Suppl:S21-9

4. 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;20(1):144-51.

5. Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc. 2002;9(6):600-11.

6. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30-7

7. Terry AL, Chevendra V, Thind A, Stewart M, Marshall JN, Cejic S. Using your electronic medical record for research: a primer for avoiding pitfalls. Fam Pract. 2010;27(1):121-6.

8. Bayley KB, Belnap T, Savitz L, Masica AL, Shah N, Fleming NS. Challenges in using electronic health record data for CER: experience of 4 learning organizations and solutions applied. Med Care. 2013;51(8 Suppl 3):S80-6.

9. 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 (Generating Evidence & Methods to improve patient outcomes). 2016;4(1).

10. Kahn MG, Brown JS, Chun AT, Davidson BN, N. B, Meeker D, et al. Transparent Reporting of Data Quality in Distributed Data Networks. eGEMs (Generating Evidence & Methods to improve patient outcomes). 2015;3(1).

11. Forrest CB, Margolis P, Seid M, Colletti RB. PEDSnet: how a prototype pediatric learning health system is being expanded into a national network. Health Aff (Millwood). 2014;33(7):1171-7.

12. 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;21(4):602-6.

13. OMOP Common Data Model [Internet]. Available from: http://www.ohdsi.org/data-standardization/the-common-data-model/.

14. Khare R, Utidjian L, Schulte G, Marsolo K, Bailey LC. Identifying and Understanding Data Quality Issues in a Pediatric Distributed Research Network. AMIA Annual Symposium; November 16 2015; San Francisco, CA.

15. Browne A, Pennington J, Bailey LC. Promoting Data Quality in a Clinical Data Research Network Using GitHub. AMIA Joint Summit on Clinical Research Informatics. 2015.

16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825−30.

17. Khare R, Utidjian L, Ruth BJ, Kahn MG, Burrows E, Marsolo K, et al. A longitudinal analysis of data quality in a large pediatric data research network. J Am Med Inform Assoc. 2017.