BMC Medical Informatics
and Decision Making

RESEARCH Open Access

Characterizing the limitations of using diagnosis codes in the context of machine learning for healthcare

Lin Lawrence Guo[1], Keith E. Morse[2], Catherine Aftandilian[3], Ethan Steinberg[4], Jason Fries[4], Jose Posada[5], Scott Lanyon Fleming[4], Joshua Lemmon[1], Karim Jessa[6], Nigam Shah[4] and Lillian Sung[1,7*]

Abstract

Background Diagnostic codes are commonly used as inputs for clinical prediction models, to create labels for prediction tasks, and to identify cohorts for multicenter network studies. However, the coverage rates of diagnostic codes and their variability across institutions are underexplored. The primary objective was to describe lab- and diagnosis-based labels for 7 selected outcomes at three institutions. Secondary objectives were to describe agreement, sensitivity, and specificity of diagnosis-based labels against lab-based labels.

Methods This study included three cohorts: SickKids from The Hospital for Sick Children, and StanfordPeds and StanfordAdults from Stanford Medicine. We included seven clinical outcomes with lab-based definitions: acute kidney injury, hyperkalemia, hypoglycemia, hyponatremia, anemia, neutropenia and thrombocytopenia. For each outcome, we created four lab-based labels (abnormal, mild, moderate and severe) based on test result and one diagnosis-based label. Proportion of admissions with a positive label were presented for each outcome stratified by cohort. Using lab-based labels as the gold standard, agreement using Cohen's Kappa, sensitivity and specificity were calculated for each lab-based severity level.

Results The number of admissions included were: SickKids ($n = 59{,}298$), StanfordPeds ($n = 24{,}639$) and StanfordAdults ($n = 159{,}985$). The proportion of admissions with a positive diagnosis-based label was significantly higher for StanfordPeds compared to SickKids across all outcomes, with odds ratio (99.9% confidence interval) for abnormal diagnosis-based label ranging from 2.2 (1.7–2.7) for neutropenia to 18.4 (10.1–33.4) for hyperkalemia. Lab-based labels were more similar by institution. When using lab-based labels as the gold standard, Cohen's Kappa and sensitivity were lower at SickKids for all severity levels compared to StanfordPeds.

Conclusions Across multiple outcomes, diagnosis codes were consistently different between the two pediatric institutions. This difference was not explained by differences in test results. These results may have implications for machine learning model development and deployment.

Keywords Electronic health records, Diagnostic coding practice, Cohort identification, Outcome identification, Machine learning for health

*Correspondence:
Lillian Sung
Lillian.sung@sickkids.ca

Full list of author information is available at the end of the article

Background

Machine learning models based on electronic health records (EHRs) are increasingly being developed and implemented into routine care. They have improved outcomes related to reducing acute care visits among ambulatory cancer patients [1], decreasing in-hospital clinical deterioration [2], increasing serious illness conversations [3], improving platelet utilization [4] and refining antibiotic choice [5] as examples.

To develop models, inputs or features are extracted from EHRs; these reflect different aspects of care such as diagnostic codes, laboratory tests, microbiology results, medication administrations, blood product administration, and procedures. Diagnostic codes are also frequently used to define the outcome of interest or label. How well each institution generates accurate diagnostic codes may vary depending on the coding process specific to the institution [6] and clinical diagnostic practice specific to the hospital unit or physician [6–8]. This variability might influence the performance and generalizability of machine learning models developed at institutions with different diagnostic coverage rates. In pediatric populations, the coverage rates of diagnostic codes and their variability across institutions are underexplored [9, 10].

A challenge to studying the question of diagnostic code coverage is the creation of gold standard labels as the diagnostic codes themselves are often used to develop these labels. One type of clinical data in which the label is inherent within the result itself is laboratory-based outcomes. Abnormal lab tests can be defined using institution-specific reference ranges. In addition, levels of severity (mild, moderate, and severe) for each abnormal lab test can be defined based upon widely accepted thresholds. Thus, evaluating diagnostic code coverage against lab-based definitions provides a pragmatic setting in which to evaluate this question. Consequently, the primary objective was to describe lab- and diagnosis-based labels for selected outcomes at three institutions. Secondary objectives were to describe agreement, sensitivity, and specificity of diagnosis-based labels against lab-based labels.

Methods

Design

This study used data derived from EHRs at three institutions, namely The Hospital for Sick Children (SickKids) in Toronto, Ontario; Lucile Packard Children's Hospital (primarily pediatric-directed care) in Palo Alto, California and Stanford Health Care (primarily adult-directed care) in Palo Alto, California. The overall goal was to compare lab- and diagnosis-based labels for pediatric patients at SickKids vs. Stanford. We included a Stanford adult cohort for descriptive purposes.

Data sources

SEDAR The data source at SickKids was the SickKids Enterprise-wide Data in Azure Repository (SEDAR) [11, 12]. SEDAR contains a curated version of Epic Clarity data that is being used for operational, quality improvement and research purposes. This study was approved as a quality improvement project at The Hospital for Sick Children and consequently, the requirement for Research Ethics Board approval and informed consent were waived by The Hospital for Sick Children.

STARR The Stanford medicine research data repository (STARR) [13] is the clinical data warehouse that contains records routinely collected in the EHR of Stanford Medicine, which is comprised of Lucile Packard Children's Hospital and Stanford Health Care. The data have been mapped to the standard concept identifiers and structure of the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) [14], resulting in a dataset named STARR-OMOP. This study used a de-identified version of STARR-OMOP [13] in which protected health information has been redacted. Because of de-identification, the requirement for Institutional Review Board approval and informed consent were waived by Stanford Medicine.

Cohorts

We defined three cohorts. SickKids was obtained using SEDAR while StanfordPeds and StanfordAdults were obtained using STARR-OMOP and applying age-specific restrictions. Table 1 summarizes the inclusion criteria for each cohort. Across all three cohorts, inpatient admissions were included if they occurred between 2018-06-02 to 2022-08-01. The pediatric cohorts (SickKids and StanfordPeds) included patients who were 28 days or older and younger than 18 years on the day of admission. We excluded neonates 1 to 27 days of age because Lucile Packard Children's Hospital has an obstetrical unit and consequently includes healthy newborns while SickKids does not have an obstetrical unit and does not routinely see healthy newborns. StanfordAdults included adult patients aged 18 or above on the day of admission. Multiple admissions per patient were permitted as long as eligibility criteria were met.

Outcome definitions

We included seven clinical outcomes that have lab-based definitions, namely acute kidney injury (AKI), hyperkalemia, hypoglycemia, hyponatremia, anemia, neutropenia, and thrombocytopenia. We appreciate there are a large number of potential lab-based outcomes; these seven were chosen based on our current research interests and because they are clinically meaningful. The outcomes were chosen *a priori*, before conducting any of the

Table 1 Inclusion criteria and cohort characteristics

	SickKids	StanfordPeds	StanfordAdults
Inclusion Criteria			
Age at Admission	≥ 28 days and < 18 years	≥ 28 days and < 18 years	≥ 18 years
Admission Date	2018-06-02 to 2022-08-01	2018-06-02 to 2022-08-01	2018-06-02 to 2022-08-01
Cohort Characteristics			
Number Admissions	59,298	24,639	159,985
Number Patients	36,585	14,518	103,170
Median Age at Admission [IQR]	6 [2–12]	6 [2–12]	57 [36–71]
Pediatric Age Group, n (%)			
Infant (28 days– 12 months)	8980 (15.1%)	3869 (15.7%)	
Toddler (13 months– 2 years)	5661 (9.5%)	2269 (9.2%)	
Early childhood (2–5 years)	10,263 (17.3%)	4307 (17.5%)	
Middle childhood (6–11 years)	14,837 (25.0%)	6056 (24.6%)	
Early adolescence (12–17 years)	19,557 (33.0%)	8138 (33.0%)	
Sex, n (%)			
Females	27,264 (46.0%)	11,800 (47.9%)	91,770 (57.4%)
Males	32,030 (54.0%)	12,837 (52.1%)	68,207 (42.6%)
Unknown	4 (< 0.1%)	2 (< 0.1%)	15 (< 0.1%)
Patient Outcomes			
In-hospital mortality, n (%)	297 (0.5%)	203 (0.8%)	3088 (1.9%)
Median length of stay [IQR]	2 [1–5]	3 [1–6]	3 [2–6]

Abbreviations: SickKids– The Hospital for Sick Children; Peds– pediatrics; IQR– interquartile range

analyses. We purposely did not include abnormal high and low for the same lab test (for example, hyperglycemia and hypoglycemia) as they may be correlated. For each outcome, we created four lab-based labels based on the test result and one diagnosis-based label; these five labels were evaluated in each patient admission. Appendix 1 in Additional File 1 shows the thresholds for each severity level (mild, moderate, and severe levels) of the lab-based labels; these thresholds were based upon research studies or guidelines [15–21]. We also labeled the result as abnormal if the result was above or below (not both) of the institution-specific reference range. For lab-based labels, units for lab results were normalized, and severity level was nested. For example, a patient admission with severe hypoglycemia would also be included in the analyses for mild and moderate hypoglycemia. For the diagnosis-based label, we considered an outcome to be present if at least one outcome-related diagnosis code was assigned to the admission. Appendices 2 and 3 in Additional File 1 list the diagnosis codes used to define diagnosis-based labels.

Concept Selection for Lab-based and diagnosis-based labels

We adopted different search strategies for concepts in STARR-OMOP and SEDAR due to differences in structure and vocabularies for clinical codes. Diagnosis codes were derived from the "condition_occurrence" table for STARR-OMOP and from the "diagnosis" table for SEDAR. Lab test results were obtained from the "measurement" table for STARR-OMOP and from the "lab" table for SEDAR. For face validation, diagnosis codes and lab result distributions obtained from STARR-OMOP were reviewed by three clinicians (KEM, CA and LS) to identify errors related to normalization or concept selection. At SickKids, this same review was only conducted by one clinician (LS) due to access restrictions.

Baseline characteristics by Cohort

To explore whether there were differences in the cohorts with respect to patients, we described the demographic characteristics and raw lab results of patients between centers. Demographic characteristics included age, sex, length of stay, and the prevalence of in-hospital mortality. For the evaluation of raw lab results, we determined the minimum or maximum result for each lab test per admission and stratified by cohort.

To gain insight into whether there were differences between pediatric institutions with respect to laboratory procedures or clinical practice, we described the institution- and age group-specific reference ranges for abnormal lab results by SickKids and StanfordPeds. The pediatric age groups were defined by the National Institute of Child Health and Human Development [22] as infancy (28 days– 12 months), toddler (13 months– 2 years), early childhood (2–5 years), middle childhood (6–11 years) and early adolescence (12–17 years). In addition, we evaluated lab testing frequency calculated as the number of tests per inpatient day for each admission.

Statistical analysis

The primary objective was to describe lab- and diagnosis-based labels at the three institutions. These were presented as the proportion of admissions with at least one positive label. To describe the odds of a lab- or diagnosis-based label by whether the pediatric admission occurred at StanfordPeds vs. SickKids, analysis was complicated by the large number of admissions and multiple testing (35 separate evaluations for this analysis alone). In addition, there were multiple admissions per patient, resulting in correlation within individuals. To address these concerns, we took several steps. First, we focused on describing the odds ratio (OR) and 99.9% confidence interval (CI) for a lab- or diagnosis-based label by pediatric institution. Second, we described the 99.9% confidence interval rather than the 95% confidence interval to help address multiple testing. Third, we did not calculate *P* values but rather, focused on describing CIs with the exception of comparing lab testing frequency by institution. Finally, to address multiple admissions per patient, OR and 99.9% CI were calculated using mixed-effects logistic regression. Models included each binary label as the outcome, institution and pediatric age group as fixed effects and subject as random intercept. Analysis was performed using the glmer function from lme4 package in R.

The secondary objectives were to describe agreement, sensitivity, and specificity of diagnosis-based labels against lab-based labels. Agreement in each cohort was described using Cohen's Kappa coefficient. Sensitivity and specificity of the diagnosis-based labels were determined using each of the lab-based labels as the gold standard. For each metric, we presented the median and ranges stratified by cohort and lab-based severity (abnormal, mild, moderate and severe).

As an exploratory analysis, we separately evaluated each visited unit during admissions at each pediatric institution. We examined the weighted proportion of positive lab-based labels and positive diagnosis-based labels for each hospital unit and calculated Spearman's rho (r) based on the average across lab-based severity.

Table 2 Distribution of minimum or maximum results for each lab test per admission stratified by cohort

Lab Test	Units	SickKids Median (IQR)	StanfordPeds Median (IQR)	StanfordAdults Median (IQR)
Maximum Creatinine	umol/L	42.0 (29.0–60.0)	34.5 (23.0–53.0)	78.7 (61.0–109.6)
Maximum Potassium	mmol/L	4.5 (4.1–5.0)	4.4 (4.1–4.9)	4.4 (4.1–4.8)
Minimum Glucose	mmol/L	4.9 (4.3–5.5)	4.9 (4.3–5.6)	5.2 (4.6–5.9)
Minimum Sodium	mmol/L	138.0 (136.0–140.0)	137.0 (134.0–139.0)	135.0 (132.0–138.0)
Minimum Absolute Neutrophil Count	10^9/L	2.2 (1.1–3.9)	3.4 (1.9–5.9)	5.5 (3.6–7.7)
Minimum Hemoglobin	g/L	102.0 (85.0–119.0)	99.0 (81.0–117.0)	103.0 (85.0–119.0)
Minimum Platelet Count	10^9/L	231.0 (147.0–321.0)	219.0 (140.0–303.0)	188.0 (139.0–243.0)

Abbreviation: SickKids: The Hospital for Sick Children; Peds: pediatrics; IQR: interquartile range

To describe lab-based reference ranges for pediatric patients, we described the threshold for an abnormal lab test by pediatric age group stratified by institution. Where the threshold varied within an age group, the range was visually depicted using a bar rather than a line. To compare testing frequency between pediatric institutions, mixed-effects linear regression was performed with number of lab tests per admission as the outcome, institution and pediatric age group as fixed effects and subject as random intercept. Analysis was performed using the lmer function from the lme4 package in R.

All analyses were conducted using Python (version 3.7) and R (version 4.1.2).

Results

Baseline characteristics

The number of admissions included were: SickKids (n=59,298), StanfordPeds (n=24,639) and StanfordAdults (n=159,985). Characteristics of the three cohorts are listed in Table 1. The distributions of age, sex, in-hospital mortality, and median length of stay were similar between SickKids and StanfordPeds while the distribution of sex and in-hospital mortality differed at Stanford$_{\text{Adult}}$. Table 2 shows the distribution of minimum or maximum results for each lab test per admission by cohort. Distributions appeared similar between StanfordPeds and SickKids with the exception of minimum absolute neutrophil count, which was lower at SickKids vs. StanfordPeds. Appendix 2 in Additional File 1 shows that the reference ranges varied between SickKids and StanfordPeds. Reference ranges for glucose and sodium were the same for all age groups except infants. Reference ranges for potassium and platelets were notably different by institution across age groups. Appendix 3 in Additional File 1 shows the average number of lab tests performed per inpatient day across all admissions stratified by institution. SickKids performed significantly fewer tests compared to StanfordPeds for all tests.

Prevalence of lab-based and diagnosis-based labels

Table 3 provides the percentage of admissions with a positive lab- and diagnosis-based label. Table 3; Fig. 1 show OR and 99.9% CI. The proportion of admissions with a positive diagnosis-based label was significantly higher for StanfordPeds compared to SickKids across all outcomes, with OR (99.9% CI) for abnormal diagnosis-based label ranging from 2.2 (1.7–2.7) for neutropenia to 18.4 (10.1–33.4) for hyperkalemia. Lab-based labels were more similar by institution although several were significantly different as demonstrated by CIs that did not cross 1.

Agreement between outcome definitions

Figure 2 shows the evaluations of diagnosis-based labels against each of the lab-based labels using Cohen's Kappa

Table 3 Proportion of admissions with positive lab- and diagnosis-based labels by cohort

Severity*	Outcome**	SickKids	StanfordPeds	StanfordAdults	Odds Ratio (99.9% CI)***
Number Admissions		59,298	24,639	159,985	
Lab (Abnormal)	AKI	4,553 (7.7%)	2,266 (9.2%)	41,761 (26.1%)	1.3 (1.1,1.4)
	Hyperkalemia	5,475 (9.2%)	1,596 (6.5%)	9,790 (6.1%)	0.7 (0.6,0.8)
	Hypoglycemia	3,006 (5.1%)	1,613 (6.5%)	12,161 (7.6%)	1.4 (1.2,1.5)
	Hyponatremia	3,888 (6.6%)	4,141 (16.8%)	53,512 (33.4%)	2.9 (2.6,3.2)
	Neutropenia	4,263 (7.2%)	1,804 (7.3%)	5,085 (3.2%)	1.0 (0.9,1.2)
	Anemia	14,839 (25.0%)	10,496 (42.6%)	119,796 (74.9%)	2.4 (2.2,2.6)
	Thrombocytopenia	10,667 (18.0%)	3,844 (15.6%)	44,636 (27.9%)	0.8 (0.7,0.9)
Lab (Mild)	AKI	4,478 (7.6%)	3,461 (14.0%)	31,930 (20.0%)	1.8 (1.6,2.1)
	Hyperkalemia	3,408 (5.7%)	1,848 (7.5%)	9,776 (6.1%)	1.3 (1.2,1.5)
	Hypoglycemia	3,844 (6.5%)	2,197 (8.9%)	13,115 (8.2%)	1.5 (1.3,1.6)
	Hyponatremia	6,169 (10.4%)	5,648 (22.9%)	66,559 (41.6%)	2.6 (2.4,2.8)
	Neutropenia	4,868 (8.2%)	1,921 (7.8%)	5,099 (3.2%)	0.9 (0.8,1.1)
	Anemia	21,232 (35.8%)	11,436 (46.4%)	114,488 (71.6%)	1.6 (1.5,1.7)
	Thrombocytopenia	7,061 (11.9%)	3,844 (15.6%)	44,636 (27.9%)	1.4 (1.2,1.5)
Lab (Moderate)	AKI	1,650 (2.8%)	1,550 (6.3%)	11,321 (7.1%)	2.1 (1.8,2.4)
	Hyperkalemia	1,939 (3.3%)	1,137 (4.6%)	4,790 (3.0%)	1.4 (1.3,1.6)
	Hypoglycemia	2,084 (3.5%)	1,178 (4.8%)	7,366 (4.6%)	1.4 (1.2,1.6)
	Hyponatremia	885 (1.5%)	917 (3.7%)	15,027 (9.4%)	2.5 (2.1,3.0)
	Neutropenia	3,346 (5.6%)	1,172 (4.8%)	2,850 (1.8%)	0.9 (0.7,1.0)
	Anemia	17,039 (28.7%)	9,429 (38.3%)	91,276 (57.1%)	1.6 (1.4,1.7)
	Thrombocytopenia	4,175 (7.0%)	2,349 (9.5%)	19,327 (12.1%)	1.4 (1.2,1.7)
Lab (Severe)	AKI	616 (1.0%)	612 (2.5%)	5,804 (3.6%)	2.4 (1.9,2.9)
	Hyperkalemia	810 (1.4%)	588 (2.4%)	1,331 (0.8%)	1.8 (1.5,2.1)
	Hypoglycemia	1,088 (1.8%)	585 (2.4%)	4,042 (2.5%)	1.3 (1.1,1.6)
	Hyponatremia	301 (0.5%)	269 (1.1%)	3,790 (2.4%)	2.1 (1.6,2.8)
	Neutropenia	2,290 (3.9%)	635 (2.6%)	1,316 (0.8%)	0.7 (0.6,1.0)
	Anemia	2,675 (4.5%)	1,912 (7.8%)	14,239 (8.9%)	1.8 (1.6,2.0)
	Thrombocytopenia	2,328 (3.9%)	1,306 (5.3%)	7,189 (4.5%)	1.4 (1.2,1.7)
Diagnosis	AKI	176 (0.3%)	1,139 (4.6%)	9,440 (5.9%)	15.1 (11.3,20.0)
	Hyperkalemia	38 (0.1%)	353 (1.4%)	2,453 (1.5%)	18.4 (10.1,33.4)
	Hypoglycemia	209 (0.4%)	412 (1.7%)	1,385 (0.9%)	4.3 (3.1,5.8)
	Hyponatremia	388 (0.7%)	708 (2.9%)	5,219 (3.3%)	4.2 (3.4,5.2)
	Neutropenia	776 (1.3%)	790 (3.2%)	1,572 (1.0%)	2.2 (1.7,2.7)
	Anemia	974 (1.6%)	4,238 (17.2%)	30,935 (19.3%)	9.7 (8.3,11.3)
	Thrombocytopenia	192 (0.3%)	2,132 (8.7%)	10,334 (6.5%)	14.8 (10.7,20.6)

* Abnormal, mild, moderate and severe according to Appendix 1 in Additional File 1. Abnormal means either above or below (not both) reference range

** Lab-based measure of acute kidney injury was hypercreatinemia

*** Odds ratio for SickKids vs. StanfordPeds obtained using mixed-effects logistic regression with each binary label as outcome, institution and pediatric age group as fixed effects, and subject as random intercept

Abbreviation: AKI: acute kidney injury; SickKids: The Hospital for Sick Children; Peds: pediatrics

coefficient, sensitivity, and specificity. Overall, diagnosis codes had high specificity (mean = 0.984, standard deviation (SD) = 0.026) but low sensitivity (mean = 0.203, SD = 0.158) and low Kappa (mean = 0.213, SD = 0.132) with lab-based labels. Compared to StanfordPeds, SickKids diagnosis-based labels had lower Kappa statistic and sensitivity, but higher specificity. Notably, the specificity results at SickKids exhibited less variation across outcomes compared to StanfordPeds and StanfordAdults, particularly in the severe category.

Figure 3 plots the weighted proportions of positive diagnosis-based labels against the weighted proportions of positive lab-based labels for each hospital unit at SickKids and StanfordPeds. At StanfordPeds, units with a higher incidence of patients with lab-based labels also had higher incidence of patients with a positive diagnosis-based label for a clinical outcome, evidenced by Spearman r ranging from 0.513 (hyponatremia) to 0.871 (neutropenia). In contrast, the Spearman r's were generally lower at SickKids across all outcomes, ranging from 0.010 (hypoglycemia) to 0.356 (anemia).

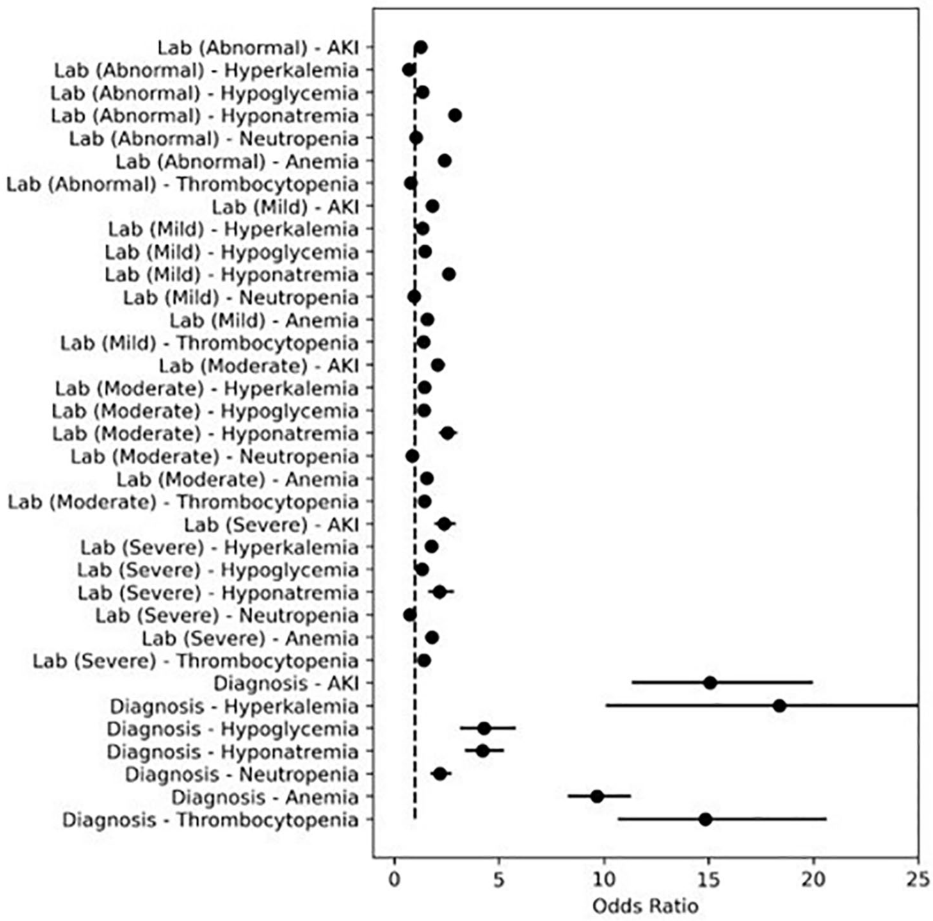

Fig. 1 Odds of each label by whether the pediatric admission occurred at Stanford vs. SickKids
Figure shows odds ratio and 99.9% confidence interval showing odds of an abnormal label by institution. Dashed line indicates an odds ratio of 1. An odds ratio of > 1 corresponds to higher odds of assigning a positive label for StanfordPeds compared to SickKids. Odds ratios were obtained using mixed-effects logistic regression with each binary label as outcome, institution and pediatric age group as fixed effects and subject as random intercept

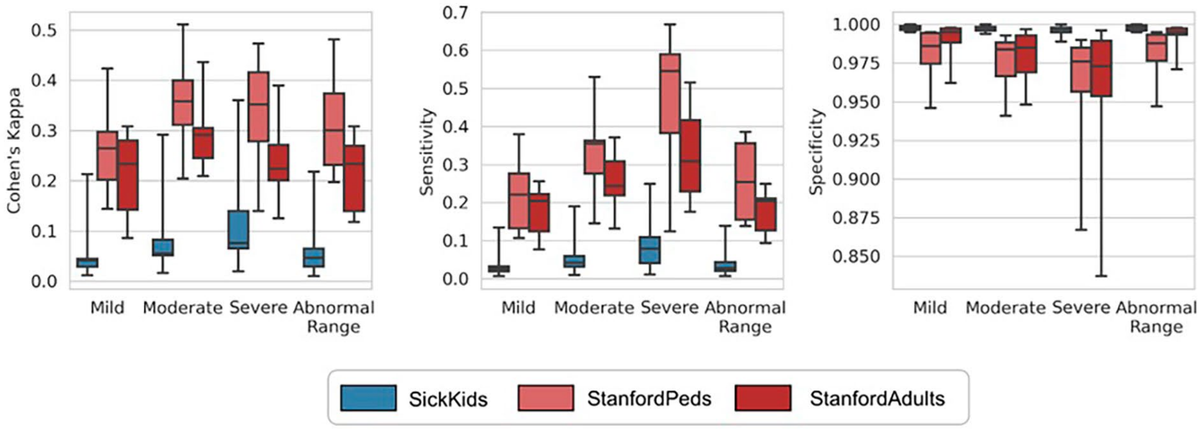

Fig. 2 Cohen's Kappa, sensitivity, and specificity for diagnosis-based labels against lab-based labels
The figure shows median, interquartile range (shaded box) and range (whiskers)

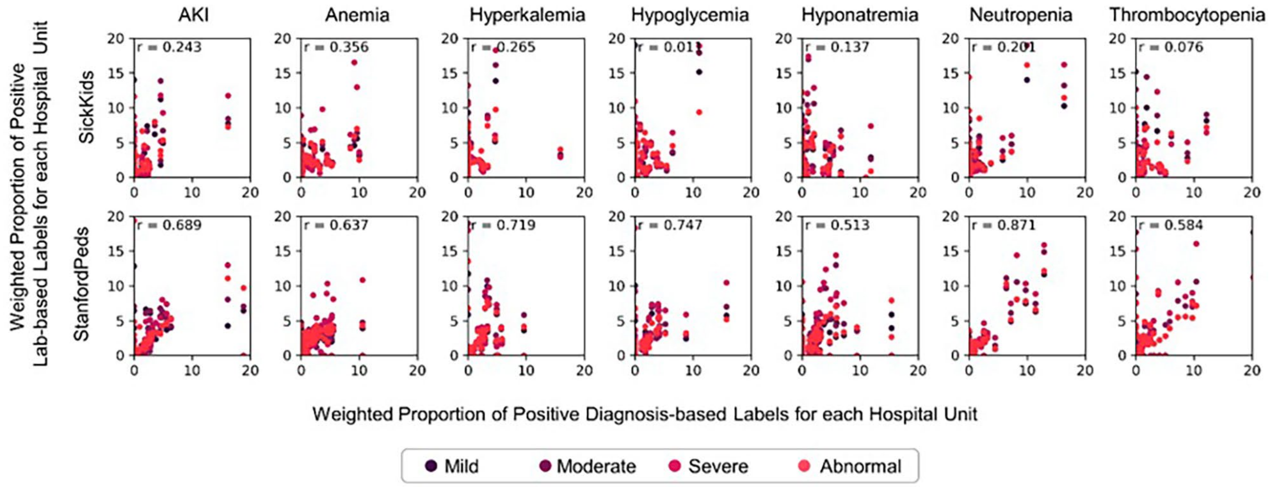

Fig. 3 Agreement between diagnosis-based labels and lab-based labels across hospital units
The numbers on the x- and y- axis represent the weighted proportion of positive lab-based labels and positive diagnosis-based labels for each hospital unit visited during the admission. Spearman rho (r) was calculated based on the average across lab-based severity

Discussion

Our results showed that despite similar demographic characteristics, there were large differences between the two pediatric institutions in the proportion of admissions with diagnosis codes for the evaluated clinical outcomes. In addition, diagnosis-based labels generally had low agreement with lab-based labels and displayed low sensitivity but high specificity when considering lab-based labels as the gold standard, with differences observed between the two institutions. In addition, we found differences between the two institutions in terms of test ordering frequency and even laboratory test references ranges.

These results suggest that if machine learning models are intended for deployment at multiple institutions, reliance on diagnostic codes, either as feature or labels, could be problematic if institutions have different coding practices. Differences in diagnostic coding practices between countries, such as between Canada and the U.S., might also contribute to the observed variations. Second, they suggest that using institutional reference ranges to categorize laboratory test results may contribute to geographic dataset shift. This study contributes to the body of evidence that demonstrates the limitations of using diagnosis codes for outcome identification. Studies have reported low sensitivity rate when using diagnosis codes to identify, for example, acute kidney injury [23], obesity [24], and symptoms of coronavirus disease 2019 [25]. In addition, this study showed differences between and within institutions in diagnostic practice that may have contributed to the differences in the performance of diagnosis codes for outcome identification.

Diagnosis codes from the EHR are commonly queried during feature extraction [26–30], label creation [31], and cohort identification [32]. Heterogeneity in diagnostic practice across hospital units within the same institution (e.g., SickKids) can impact a model's performance within sub-populations or spuriously associate certain units with the outcome of interest during model development. In addition, the cross-institution difference in diagnostic coding practice has implications for network studies as it violates the assumption that coding practice is comparable across institutions and creates heterogeneity in outcome prevalence as an artifact of code availability.

While we found that the proportions of positive lab-based labels were more similar between pediatric institutions, there were significant differences although smaller than that observed for diagnosis-based labels. Possible contributions were the observed differences in lab testing frequency between the two pediatric institutions. In addition, the reference ranges themselves were different for tests with the same absolute interpretation regardless of where the test was conducted. For example, two hypothetical children with the same platelet count could be considered to have a normal test at one institution and an abnormal test at the second institution. Some SickKids reference ranges were based upon those established by the Canadian Laboratory Initiative on Pediatric Reference Intervals (CALIPER) initiative [33], which contributed to the disparity. Nonetheless, this has implications for machine learning models. First, it is common during feature processing to categorize lab test results as normal, high, and low based upon the reference range [26, 34]. Having different reference ranges would thus produce different features despite having the same numerical value. Second, different reference ranges may impact downstream clinical decision making and variability of resultant clinical actions, for example procedures and medication administrations. Since these actions will be recorded in the EHR, impact on clinical decision making can further worsen geographic dataset shift.

The strengths of this study include the ability to evaluate multiple institutions in different countries and the involvement of clinician co-investigators who contributed to the identification of concepts to include in the various label definitions. However, this study is limited for several reasons. First, we only evaluated seven outcomes. In addition, the outcomes were restricted to those that have lab-based definitions in order to use lab tests to develop gold standard labels. Outcomes that are more complex might require chart review to establish gold standards and more sophisticated electronic phenotyping approaches to reach reasonable performance [35, 36]. Next, our analyses were restricted to admissions within a relatively narrow time period (2018–2022). It might be useful to characterize practice differences over time as temporal distribution shift can negatively impact model performance over time [28, 37]. Finally, it is possible that differences in EHR implementation [38] and data transformation processes between the institutions may have contributed to the observed variations.

Conclusion

In conclusion, across multiple outcomes, diagnosis codes were consistently different between the two pediatric institutions. This difference was not explained by differences in test results. These results may have implications for machine learning model development and deployment.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02449-8.

Supplementary Material 1

Acknowledgements

Not applicable.

Author contributions

LLG and LS designed the project with input from all authors. LLG performed all experiments. LLG, KEM, CA, and LS analyzed and interpreted results, with some input from all other authors. LLG and LS wrote the manuscript. All authors revised and commented on the manuscript. All authors read and approved the final manuscript.

Funding

No external funding was received for the study.

Data availability

Access to the de-identified STARR-OMOP requires affiliation with a Stanford Principal Investigator and a Stanford identity. Access to SEDAR is restricted due to security and privacy considerations. However, data supporting the findings of this study are available upon reasonable request. The code to reproduce the Stanford cohorts using STARR-OMOP are available at https://github.com/som-shahlab/starr-omop-bq-datasets.

Declarations

Ethics approval and consent to participate

The use of data from The Hospital for Sick Children was approved as a quality improvement project at SickKids and thus, requirement for Research Ethics Board approval and informed consent were waived by The Hospital for Sick Children. The data from Stanford Medicine was de-identified in which protected health information has been redacted. Because of de-identification, the requirement for Institutional Review Board approval and informed consent were waived by Stanford Medicine. All methods were carried out in accordance with relevant guidelines and regulations in the Declaration of Helsinki.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Author details

[1]Program in Child Health Evaluative Sciences, The Hospital for Sick Children, Toronto, ON, Canada
[2]Division of Pediatric Hospital Medicine, Department of Pediatrics, Stanford University, Palo Alto, CA, USA
[3]Division of Hematology/Oncology, Department of Pediatrics, Stanford University, Palo Alto, CA, USA
[4]Stanford Center for Biomedical Informatics Research, Stanford University, Palo Alto, CA, USA
[5]Universidad del Norte, Barranquilla, Colombia
[6]Information Services, The Hospital for Sick Children, Toronto, ON, Canada
[7]Division of Haematology/Oncology, The Hospital for Sick Children, 555 University Avenue, M5G1X8 Toronto, ON, Canada

Received: 23 March 2023 / Accepted: 30 January 2024
Published online: 14 February 2024

References

1. Hong JC, Eclov NCW, Dalal NH, Thomas SM, Stephens SJ, Malicki M, et al. System for high-intensity evaluation during Radiation Therapy (SHIELD-RT): a prospective Randomized Study of Machine Learning–Directed clinical evaluations during Radiation and Chemoradiation. J Clin Oncol. 2020;38(31):3652–61.
2. Escobar GJ, Liu VX, Schuler A, Lawson B, Greene JD, Kipnis P. Automated identification of adults at risk for In-Hospital clinical deterioration. N Engl J Med. 2020;383(20):1951–60.
3. Manz CR, Parikh RB, Small DS, Evans CN, Chivers C, Regli SH, et al. Effect of integrating machine learning mortality estimates with behavioral nudges to clinicians on Serious Illness conversations among patients with Cancer: a stepped-Wedge Cluster Randomized Clinical Trial. JAMA Oncol. 2020;6(12):e204759–e.
4. Guan L, Tian X, Gombar S, Zemek AJ, Krishnan G, Scott R, et al. Big data modeling to predict platelet usage and minimize wastage in a tertiary care system. Proc Natl Acad Sci U S A. 2017;114(43):11368–73.
5. Yelin I, Snitser O, Novich G, Katz R, Tal O, Parizade M, et al. Personal clinical history predicts antibiotic resistance of urinary tract infections. Nat Med. 2019;25(7):1143–52.
6. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. Health Serv Res. 2005;40(5 Pt 2):1620–39.
7. Burles K, Innes G, Senior K, Lang E, McRae A. Limitations of pulmonary embolism ICD-10 codes in emergency department administrative data: let the buyer beware. BMC Med Res Methodol. 2017;17(1):89.
8. Tang KL, Lucyk K, Quan H. Coder perspectives on physician-related barriers to producing high-quality administrative data: a qualitative study. CMAJ Open. 2017;5(3):E617.
9. Liu B, Hadzi-Tosev M, Liu Y, Lucier KJ, Garg A, Li S et al. Accuracy of International classification of diseases, 10th Revision codes for identifying Sepsis: a systematic review and Meta-analysis. Crit Care Explorations. 2022;4(11).
10. Golomb MR, Garg BP, Saha C, Williams LS. Accuracy and yield of ICD-9 codes for identifying children with ischemic stroke. Neurology. 2006;67(11):2053.

11. Guo LL, Calligan M, Vettese E, Cook S, Gagnidze G, Han O, et al. Development and validation of the SickKids Enterprise-wide data in Azure Repository (SEDAR). Heliyon. 2023;9(11):e21586.
12. Guo LL, Calligan M, Vettese E, Cook S, Gagnidze G, Han O et al. Development and validation of the SickKids Enterprise-wide Data in Azure Repository (SEDAR). In Press.
13. Datta S, Posada J, Olson G, Li W, O'Reilly C, Balraj D et al. A new paradigm for accelerating clinical data science at Stanford Medicine. arXiv preprint arXiv:200310534. 2020.
14. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Studies in health technology and informatics. 2015;216:574.
15. Khwaja A. KDIGO clinical practice guidelines for acute kidney injury. Nephron Clin Pract. 2012;120(4):c179–84.
16. Daly K, Farrington E. Hypokalemia and hyperkalemia in infants and children: pathophysiology and treatment. J Pediatr Health Care. 2013;27(6):486–96. quiz 97– 8.
17. Abraham MB, Jones TW, Naranjo D, Karges B, Oduwole A, Tauschmann M et al. ISPAD Clinical Practice Consensus guidelines 2018: Assessment and management of hypoglycemia in children and adolescents with diabetes. Pediatr Diabetes. 2018;19 Suppl 27:178– 92.
18. Spasovski G, Vanholder R, Allolio B, Annane D, Ball S, Bichet D, et al. Clinical practice guideline on diagnosis and treatment of hyponatraemia. Eur J Endocrinol. 2014;170(3):G1–47.
19. Allali S, Brousse V, Sacri AS, Chalumeau M, de Montalembert M. Anemia in children: prevalence, causes, diagnostic work-up, and long-term consequences. Expert Rev Hematol. 2017;10(11):1023–8.
20. Lustberg MB. Management of neutropenia in cancer patients. Clin Adv Hematol Oncol. 2012;10(12):825–6.
21. Chernecky C, Barbara B. Platelet (thrombocyte) count - blood. Laboratory Tests and Diagnostic Procedures. 6th edition ed. St Louis, MO: Elsevier Saunders; 2013. p. 886-7.
22. Williams K, Thomson D, Seto I, Contopoulos-Ioannidis DG, Ioannidis JPA, Curtis S, et al. Standard 6: Age groups for Pediatric trials. Pediatrics. 2012;129(Supplement3):153–S60.
23. Tomlinson LA, Riding AM, Payne RA, Abel GA, Tomson CR, Wilkinson IB et al. The accuracy of diagnostic coding for acute kidney injury in England– a single centre study.
24. Grams ME, Waikar SS, MacMahon B, Whelton S, Ballew SH, Coresh J. Performance and Limitations of Administrative Data in the identification of AKI. Clin J Am Soc Nephrol. 2014;9(4):682–9.
25. Crabb BT, Lyons A, Bale M, Martin V, Berger B, Mann S, et al. Comparison of International Classification of Diseases and related health problems, Tenth Revision codes with Electronic Medical records among patients with symptoms of Coronavirus Disease 2019. JAMA Netw Open. 2020;3(8):e2017703–e.
26. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969–75.
27. Rasmy L, Xiang Y, Xie Z, Tao C, Zhi D. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digit Med. 2021;4(1):1–13.
28. Guo LL, Pfohl SR, Fries J, Johnson AEW, Posada J, Aftandilian C, et al. Evaluation of domain generalization and adaptation on improving model robustness to temporal dataset shift in clinical medicine. Sci Rep. 2022;12(1):2726.
29. Steinberg E, Jung K, Fries JA, Corbin CK, Pfohl SR, Shah NH. Language models are an effective representation learning technique for electronic health record data. J Biomed Inform. 2021;113:103637.
30. Tang S, Davarmanesh P, Song Y, Koutra D, Sjoding MW, Wiens J. Democratizing EHR analyses with FIDDLE: a flexible data-driven preprocessing pipeline for structured clinical data. J Am Med Inform Assoc. 2020;27(12):1921–34.
31. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Sci Data. 2019;6(1):96.
32. Khera R, Schuemie MJ, Lu Y, Ostropolets A, Chen R, Hripcsak G, et al. Large-scale evidence generation and evaluation across a network of databases for type 2 diabetes mellitus (LEGEND-T2DM): a protocol for a series of multinational, real-world comparative cardiovascular effectiveness and safety studies. BMJ Open. 2022;12(6):e057977.
33. Adeli K, Higgins V, Trajcevski K, White-Al Habeeb N. The Canadian laboratory initiative on pediatric reference intervals: a CALIPER white paper. Crit Rev Clin Lab Sci. 2017;54(6):358–413.
34. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. J Biomed Inform. 2021;113:103621.
35. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. J Am Med Inf Assoc. 2013;20(1):117–21.
36. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inf Assoc. 2016;23(e1):e20–7.
37. Guo LL, Pfohl SR, Fries J, Posada J, Fleming SL, Aftandilian C, et al. Systematic review of approaches to preserve machine learning performance in the Presence of temporal dataset shift in Clinical Medicine. Appl Clin Inf. 2021;12(04):808–15.
38. Glynn EF, Hoffman MA. Heterogeneity introduced by EHR system implementation in a de-identified data resource from 100 non-affiliated organizations. JAMIA Open. 2019;2(4):554–61.

Publisher's Note