Comparison of EHR-based diagnosis documentation locations to a gold standard for risk stratification in patients with multiple chronic conditions

Shelby Martin[1]; Jesse Wagner[1]; Nicoleta Lupulescu-Mann[2]; Katrina Ramsey[3]; Aaron A. Cohen[1]; Peter Graven[2]; Nicole G. Weiskopf[1]; David A. Dorr[1]

[1]Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, Portland, OR, USA;
[2]Center for Health Systems Effectiveness, Oregon Health & Science University, Portland, OR, USA;
[3]School of Public Health, Division of Biostatistics, Oregon Health & Science University, Portland, OR, USA

Keywords
Multiple Chronic Conditions; Health Information Systems; Risk Stratification; Forecasting; Data Quality

Summary

Objective: To measure variation among four different Electronic Health Record (EHR) system documentation locations versus 'gold standard' manual chart review for risk stratification in patients with multiple chronic illnesses.

Methods: Adults seen in primary care with EHR evidence of at least one of 13 conditions were included. EHRs were manually reviewed to determine presence of active diagnoses, and risk scores were calculated using three different methodologies and five EHR documentation locations. Claims data were used to assess cost and utilization for the following year. Descriptive and diagnostic statistics were calculated for each EHR location. Criterion validity testing compared the gold standard verified diagnoses versus other EHR locations and risk scores in predicting future cost and utilization.

Results: Nine hundred patients had 2,179 probable diagnoses. About 70% of the diagnoses from the EHR were verified by gold standard. For a subset of patients having baseline and prediction year data (n=750), modeling showed that the gold standard was the best predictor of outcomes on average for a subset of patients that had these data. However, combining all data sources together had nearly equivalent performance for prediction as the gold standard.

Conclusions: EHR data locations were inaccurate 30% of the time, leading to improvement in overall modeling from a gold standard from chart review for individual diagnoses. However, the impact on identification of the highest risk patients was minor, and combining data from different EHR locations was equivalent to gold standard performance.

The reviewer's ability to identify a diagnosis as correct was influenced by a variety of factors, including completeness, temporality, and perceived accuracy of chart data.

Correspondence to:
David A. Dorr
Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University
3181 S.W. Sam Jackson Park Rd., MDYMICE
Portland, OR 97239–3098, USA
Email: dorrd@ohsu.edu

Appl Clin Inform 2017; 8: 794–809
received: December 21, 2016
accepted in revised form: May 31, 2017
published: August 2, 2017
Martin S, Wagner J, Lupulescu-Mann N et al. Comparison of EHR-based diagnosis documentation locations to a gold standard for risk stratification in patients with multiple chronic conditions. Appl Clin Inform 2017; 8: 794–809
https://doi.org/10.4338/ACI-2016-12-RA-0210
Funding
The project described was supported by AHRQ grant number 1R21HS023091–01. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the Agency for Healthcare Research and Quality.

1. Background and Significance

Risk stratification is a process wherein clinical practices assign their patients into different tiers based on factors that may contribute to adverse health outcomes. In ambulatory or longitudinal care, most risk scores are calculated based on patient demographics and the presence of chronic conditions, with patients with more conditions having an increased risk for hospitalizations, emergency department (ED[1]) visits, and higher costs of care. Risk stratification can help better identify and mitigate patients' risk and allocate resources for healthcare more effectively, but risk prediction by current scores is moderately successful at best [1–3]. This may be partially due to the temporal lag and limited scope of traditional data sources. However, as Electronic Health Records (EHRs) are adopted, increased availability of timely and accurate clinical data may improve scores' predictive ability. For example, EHRs may store data and documentation about diagnoses in different locations such as the problem list, in encounters, in medical history, or in the labs, studies, and notes used to confirm diagnoses. EHR data have mixed quality related to completeness, accuracy, and meaning, which may affect risk scores [4], but little work has been done to explore how these qualities of computer-based patient records impact patient risk prediction.

Since EHR data is rich with narrative text and other information, validating patient data via complete review of all patient data (e.g., a gold standard chart review) should improve a risk score's predictive ability by reducing inaccuracies and inconsistencies from any individual EHR documentation location [5]. Diagnoses, for example, may be entered in different locations within the EHR with different intentions: in the problem list (a patient-level description of active problems related to health and well-being) during a hospitalization but never removed if resolved; in the medical history either because it is resolved or present but not pertinent to a particular visit; copied from previous encounters without verification, or added later by coders for billing. However, some EHR documentation locations may be more relevant, accurate, and actionable than others. For instance, other studies have shown that increased problem list completeness can improve data validity and impact outcomes, such as more effective care, through clinical decision support [6, 7]. Unfortunately, while problem lists may be more accurate than other locations, they also tend to be less complete [8], as well as highly variable across sites [9]. For instance, Meystre and Haug found problem list sensitivity for condition presence was initially 9.8% compared to a gold standard, only rising to 41% post-intervention using a Natural Language Processing system to extract medical problems from free-text documents in the electronic chart [10]. Other documentation locations in the EHR may have the diagnosis when the problem list is empty, such as lab results, studies, and medications, but these are even less complete overall [11] and may need to be adjusted for local practice patterns to maximize accuracy. When phenotyping patients to determine genetic influences on disease, this degree of precision is important, but risk stratification may depend less on precision and more on completeness or concepts like salience and temporality. In sum, diagnoses stored in different EHR locations and entered through different software modules likely contain different information about a given condition for a particular patient, and it is important to understand how this variability affects different risk scores' ability to accurately predict future patient outcomes especially as compared to a gold standard summary of the information available.

2. Objectives

Our aim was to understand the degree to which different patient diagnosis documentation locations in the EHR accurately reflect whether a patient has a particular chronic condition, and how this affects risk prediction for future utilization costs, hospitalizations, and ED visits. We hypothesized that

1 Abbreviation definitions: ED – Emergency Department; EHR – Electronic Health Record; HCC – Hierarchical Condition Categories; MCCI – Modified Charlson Comorbidity Index; ACA – Affordable Care Act; CHF – Congestive Heart Failure; PheKB – Phenotype Knowledge Base; ICCIS – Integrated Care Coordination Information System; PD – Patient-Diagnosis; COPD – Chronic Obstructive Pulmonary Disease; CKD – Chronic Kidney Disease; CAD – Coronary Artery Disease

characteristics of EHR data by documentation location (i.e., problem list, medical history, ambulatory encounter diagnoses, and a non-diagnosis code 'rule') affect the accuracy of risk prediction for future patient healthcare utilization.

3. Methods

To achieve our aim, we identified and defined common risk scores and diagnoses that convey risk of future health exacerbations. We then queried an EHR for all potential sources of these diagnoses from adult patients seen in primary care in a large, urban academic medical center in Oregon; data that was collected as part of normal care anywhere in the system was queried. Concurrently, we developed a 'gold standard', formal annotation process that defined a comprehensive chart review for these same diagnoses, and performed the annotation on charts. We then calculated the same risk scores from each individual EHR documentation location with those verified by the gold standard, and compared the predictive validity of the various locations. ▶Figure 1 shows the flow chart of data extraction and patient inclusion and exclusion from the entire study; each component will be discussed below.

3.1 Diagnosis selection and definition and risk scores

We chose diagnoses based on their use in a set of common risk scores. The diagnoses:
- are present in one or more existing risk scores and/or predictive algorithms;
- can be fully defined using codes and rules, and
- can be recognized readily by a human reviewer from documentation in the medical chart.

The thirteen diagnoses chosen were Alcohol/Drug Abuse, Asthma, Attention Deficit Hyperactivity Disorder (ADHD), Breast Cancer, Congestive Heart Failure (CHF), Chronic Kidney Disease (CKD), Chronic Obstructive Pulmonary Disease (COPD), Coronary Artery Disease, Depression, Diabetes (with and without complications), Human Immunodeficiency Virus (HIV), and Ischemic Stroke. These were chosen from 3 risk scores. First, Hierarchical Condition Categories (HCC), was chosen due to wide use and validation in cost prediction; next, a modified Charlson Comorbidity Index (MCCI), chosen for moderate simplicity and validation in utilization [12]; and, third, a count of chronic conditions defined in the Affordable Care Act (ACA) for Medicaid Health Homes, which was chosen for its simplicity and its inclusion of mental health conditions [13].

3.2 EHR documentation location diagnosis definition

Once the diagnoses and risk scores were chosen, we defined the locations of these diagnoses in an EHR (Epic), and then defined the codes or rules that make up these diagnoses. We chose four different EHR documentation locations to extract patient diagnoses based on which ones were likely to have evidence of the diagnoses: Problem Lists, a patient level source of diagnoses used to denote active problems across encounters; Encounter diagnoses, where up to 3 are added for justification of billing codes; Medical History, often used to record all historical diagnoses, whether active or not; and a Phenotypic Rule for other data sources. For the first three, we used standard value sets containing ICD-9 and ICD-10 codes for each of the thirteen diagnoses, identified from the definitions of the risk scores themselves (each risk score contains a set of diagnoses grouped to these conditions); or, if the groupings were too broad (multiple diagnoses in the same category, as in the HCC In these cases, we used specific subsets of these codes groups from quality measure value sets found in the Value Set Authority Center, a service of the National Library of Medicine [14]. The purpose of the phenotypic rule was to use non-diagnosis code information from an EHR to identify the diagnosis; in most instances, these were labs or studies that defined the diagnosis (e.g., HIV PCR). Phenotype repositories, such as PheKB (Phenotype Knowledgebase; www.phekb.org), were reviewed, and specific diagnostic laboratory tests, medications prescribed primarily for a given condition, and other surveys and questionnaires were used to identify these diagnoses. Unlike usual phenotypes, however, the rule did not include the same diagnosis codes from the other sources.

3.3 Eligibility and Extraction

From these codes, a set of patient diagnoses for manual review was extracted from the data from a subset of 3 clinics sharing the same EHR from a sample of patients selected from the Integrated Care Coordination Information System (ICCIS) study data set. The overall dataset has more than 500,000 patients from more than 8 EHRs [17]; our subset of clinics were chosen because the three Oregon primary care clinics had both continuous EHR data from 2008–2014 and linked claims data on costs, ED visits and hospitalizations. From 16,120 patients ever seen these clinics, adult patients were eligible for inclusion if there was EHR evidence of ≥1 eligible condition in any of the EHR documentation locations, and they were seen from 2008 to 2012. From the total set, 900 patients were selected; of these, 750 had complete outcomes data from claims. To establish false positive detection, we randomly selected additional patients and added false diagnoses, representing ~5% of our sample (overall, 4.2% of diagnoses).

3.4 Gold standard creation

A formal annotation process was defined to create a gold standard for patient diagnoses based on chart review. The process, encoded in a workbook (▶Appendix 1), contained instructions on how to determine the presence or absence of each of the thirteen diagnoses for each patient [15, 16]. The project team, consisting of 2 research assistants and 2 faculty with expertise in medicine and informatics, completed a series of iterative review cycles to further refine validation rules, identify data errors, and resolve questions related to diagnosis accuracy when these were unclear. Errors found in the data were noted and the EHR query was revised to exclude these in subsequent sets. From this, a template for patient data review was created to capture key information, including whether the diagnosis suggested was correct, was a related or inactive (e.g., diabetes post-bariatric surgery or childhood asthma) diagnosis, or was incorrect according to the judgment of the reviewer.

A patient diagnosis (PD) was defined as a single specified condition for a particular patient listed as active within a one-year period; each patient may have multiple conditions in the dataset, but codes for the same diagnosis from the same documentation location were combined. The patient identifiers and diagnoses from the EHR query were extracted into the structured template for manual review. A research assistant completed the PD reviews. Besides presence of diagnoses, certainty of a PD (i.e., the degree to which the reviewer was certain their judgment was correct) was assessed on a scale of 1–10, based on volume and clarity of data. The annotated gold standard set with diagnoses and reviewer assessments is provided in de-identified form as a supplemental component. An independent rater reviewed 10% of PD using the same scale for reliability testing.

3.5 Analysis

Analysis goals were to determine the diagnostic and predictive validity and relative value of each EHR documentation location compared to the gold standard; then, to calculate the relative change in predictive validity for the the three outcomes of ED visits, hospitalizations and healthcare costs based on use of each documentation location versus the ‚true' set of diagnoses found in the gold standard. First, we examined how often the reviewer identified each PD as

- correct,
- incorrect, or
- belonging

to a related disease, and we compared the reviewer's average certainty by PD; diagnoses marked correct were coded as true positives, false positives, or true negatives (if a false diagnosis). We evaluated inter-rater reliability using Bias adjusted kappa. To measure the accuracy of EHR diagnoses, we compared each location to the gold standard with respect to sensitivity, specificity, positive predictive value, negative predictive value, and binary correlation as represented by the phi coefficient. Finally, to document common reasons for inconsistencies (e.g., false positives and negatives) between coding and documentation in the chart, we reviewed reasons why diagnoses were labeled as incorrect.

For the predictive validity and relative value of each location, we used the diagnoses identified from each documentation location to calculate a set of the standard risk scores described above. Descriptive statistics and patient outcomes were calculated for each risk score and documentation location. To establish the predictive validity of the locationsfor risk scores, independent multivariable regression models were created to estimate the impact of the score, for each EHR documentation location, on the outcome. Two-part models were created using first zero-weighted analysis, then a generalized linear model with link and variance family specifications for each outcome. The risk scores were calculated on the presence of each condition in each EHR documentation location or combination of different locations. We also calculated risk scores for a combined EHR location, where any of the locations had evidence of the condition. The marginal effect of the score (increase in outcome per 1 unit increase in score, e.g., increase in rate of hospitalization per increase in score) was calculated and tested for significance. Additionally, the difference in marginal effects between locations was tested to determine if the effect of one location was different from that of another by z-statistic[2]. To create the models, we identified the model link function using a model selection routine described in Glick [18] using the best predicting location and risk score combination (Problem List-Encounter-Medical History and HCC) [18, 19]. The routine identifies the link function that performs best using the Pearson Correlation test of bias on the raw scale.

4. Results

4.1 Description of Sample

A total of 2,179 diagnoses across 750 eligible patients were used for analyses. Descriptive statistics for these patients are shown in ▶Table 1 and 2. Due to the selection of high-risk conditions, risk scores were slightly higher than general adult population averages and ranges for the clinic as a whole. Similarly, average costs, hospitalizations, and ED visits were higher in our sample.

Of the 93 (4.2%) false diagnoses, 83 (89.2%) were correctly identified as such. Ten (10.8%) were incorrectly identified as true diagnoses; in four of these cases the false PD was closely related to another diagnosis found in the chart.

Inter-rater reliability was tested by a second, independent reviewer on a random subset of 106 (4.9%) diagnoses. The second reviewer did not know how the primary reviewer had labeled the diagnoses, nor the certainty ranking for these. Results showed 84.5% agreement for diagnosis identification between reviewers translating to 'substantial' agreement (Bias-adjusted kappa coefficient: 0.76).

Excluding false diagnoses, 70.7% of diagnoses were verified as correct, 20.6% as incorrect, and 8.7% were related but different diagnoses. Table II shows the frequency of each diagnosis across all patients and frequencies of diagnoses verified as correct, incorrect, and related or inactive. The frequency of correct diagnosis was highest for depression (87.8%), while the frequency of incorrect diagnosis was highest for attention deficit hyperactivity disorder (ADHD) (44.9%). The highest frequency for related or inactive diagnosis was for breast cancer (28.9%) due to the high number of patients in remission.

Common reasons for determination of incorrect diagnoses included lack of evidence, weak evidence, data entry errors on the part of the reviewer, diagnosis resolved by the start of prediction year, and medication prescribed before or after the prediction year, among others specific to the diagnosis (e.g., patient had kidney transplant but previous diagnosis of chronic kidney disease, or CKD). ▶Appendix 2 ▶Table 3 shows the frequencies for reasons why the gold standard identified the initial diagnosis as incorrect when a specific reason accounted for more than five percent of the total number of diagnoses.

2 This was performed with a T-test of difference in means using the standard errors of the marginal effects. Identical results can be obtained from an interactive model where the documentation location is interacted in the model with the risk score.

Across all conditions and results, the reviewer's average certainty was 8.2 (▶Appendix 2 ▶Table 4). The lowest average certainty for correct diagnosis was for CHF (6.9) while the lowest certainty for incorrect diagnosis was for depression (7.2) and attention deficit hyperactivity disorder (7.2). For related diagnoses, the highest average certainty was for diabetes without complications (8.8), while the lowest was for chronic obstructive pulmonary disease (COPD) (7.4).

4.2 Diagnostic accuracy of documentation location compared to gold standard

Diagnostic accuracy by EHR documentation location is presented in ▶Figure 2 and in ▶Table 4. The outer grey circle represents all PD, as determined by presence of the diagnosis in any one of the four locations or the gold standard. The yellow circle represents all diagnoses from the gold standard. The blue circles denote the diagnoses from each of the four locations (or, in the case of the last panel, the combination of all four locations). The green overlap between the yellow and blue circles represents true positives (TP): diagnoses present in the locations that were confirmed by presence in the gold standard. By extension, the yellow crescent shows false negatives (FN), which were present in the gold standard, but not in the relevant location, the blue crescent shows false positives (FP), which were present in the location, but not the gold standard, and the remaining grey indicates true negatives (TN). The sensitivity (TP/(TP+FN)) was highest for Encounters data and Medical History data (0.54–0.55), and specificity (TN/(TN+FP)) was highest for Problem List data (0.82). The rule-based 'location' without diagnostic codes was neither sensitive nor specific (0.48 and 0.49). Using any location as a positive generated the highest sensitivity (0.95) and lowest specificity (0.19). Binary correlation between the gold standard and documentation locations, as measured with the phi coefficient, was highest for Problem List (0.85) and lowest for Rule Check (0.41).

4.3 Predictive performance

▶Table 5 lists the validation modelling results. Variations in risk scores are presented across the four locations and the gold standard as a fifth documentation location; the risk scores ranged from 0–9 in the samples. The modelling determined the marginal effect of a one point increase in risk score on the outcome in question and the z-score the standard deviations of the marginal effect from the mean, with 1.96 z score significant at $p<.05$. The gold standard performed better than the individual locations for all three risk scores, but performed slightly lower than the combined documentation locations. The marginal effects varied by the risk score ranges; for example, using the problem list, a one-point increase in the HCC risk score resulted in a $39,172 increase in costs, whereas a one point increase in ACA only increased costs by $10,403 due to the ACA risk score's larger range. For all three outcomes, HCC outperformed the other scores. For cost, the HCC z-statistics ranged from 1.69–3.27 ($p=0.09$–0.001), compared to MCCI (0.74–1.48, all $p>0.10$) and ACA (1.49–2.53, $p=0.14$–0.01). Comparing only HCC, the z-statistic for the encounter location was greater than the other individual locations and the combined location, with a difference ranging from 0.23–1.58. Detailed comparisons of relative performance are shown in ▶Table Appendix 4.

Notably, the combined scores outperformed the gold standard on two of three risk scores for cost, but not HCC. A similar pattern was evident for ED visits, with z-statistics between 0.45–3.2, and the gold standard generally outperformed the other individual locations (average difference 0.92), but the combined location was nearly equivalent to the gold standard. However, the gold standard was better than the worst EHR location (Encounter-HCC, with a substantial z-statistic difference of 1.62, with $p=0.11$). Finally, for hospitalizations, the HCC performed best, and the gold standard outperformed Problem List and Medical History (average difference 1.1) but not Encounters or Combined locations.

5. Discussion

We have demonstrated that substantial variation can occur in diagnostic accuracy and validity by EHR locations compared to gold standard chart review. For individual locations, problem list remains the most specific and encounter the most sensitive for our data. Attempting to reproduce the diagnoses through a non-diagnosis code based rule was least sensitive and specific across a number of diagnoses. While relative predictive validity varied substantially when using EHR documentation locations to calculate risk compared to the gold standard, a common task of identifying 'high' risk patients by cut point revealed few differences. Surprisingly, the combined set of data from all locations outperformed the gold standard in a few cases; our hypothesis for this finding was that data volume – the amount of data entered in different places in the EHR – is an independent predictor of future outcomes. This makes sense, logically, since more health care utilization requires more data entry.

Our sensitivity and specificity results are similar to those reported by a 2003 comprehensive study on the misclassification of claims data diagnoses using manual chart review [20]. The study results showed that specificity of claims diagnoses was substantially higher than the sensitivity due to the greater likelihood that codes omitted for patients having a condition are expected to be more common than a mistake in coding when a patient does not have a particular condition. PPV and NPV were high in our sample; however, the algorithm that selected the initial patient population did select 70% of patients and diagnoses accurately, so our prevalence was higher than would be found in a large population.

EHR data are frequently incomplete due to health care fragmentation but their accuracy is often thought to be better than claims alone. Moreover, the temporality of EHR data, which are available nearly as soon as care is delivered, is a major potential benefit over claims data. Still, the percentage of diagnoses identified as correct was lower than anticipated at 70.7%. Reviewing reasons for incorrect diagnoses was helpful in revealing patterns occurring in these diagnoses, which frequently corresponded with the IOM's high quality data attributes. For example, 'lack of evidence' was the most frequent reason for an incorrect diagnosis, corresponding to the completeness attribute. Some diagnoses were found to be resolved; for instance, CKD may be 'treated' by a kidney transplant). Actual errors in labeling diagnoses as incorrect, found through false diagnoses or through inter-rater review, were infrequent (2.9–13.5%, depending on the diagnosis).

We identified several factors that influenced the reviewer's confidence about the diagnosis. These included the volume of supporting data in the EHR, whether the diagnosis could be confirmed with a single test, whether the diagnosis was in remission or resolved by the start of the prediction year (e.g., breast cancer), and when diagnoses were labeled as another condition in the same class but were not identical (e.g., asthma). This reaffirms the importance of issues such as completeness, temporality, and meaning or accuracy in the EHR. Future interventions may target diagnoses with potentially high impact on future risk prediction, and ask for clarification or encourage a single documentation location such as the problem list.

Despite these errors found in the data, the overall impact on risk prediction was minimal. The model we created for outcome prediction showed that the gold standard was slightly better at predicting ED visits, hospitalization, and costs than other individual EHR locations , but not better than combinations of them. Our results add to the evidence that manual review of chart data may improve predictive validity of the risk score, with diagnoses and EHR documentation locations having a wide range of baseline characteristics. However, manual chart review is a time- and labor-intensive process, and the relative gain was not substantial.

Limitations of this study include algorithmic triggers that selected diagnoses as positive that were truly negative (e.g., medications for asthma which may not require an actual asthma diagnosis or HIV tests with equivocal results read as positive by the rule), inability to access certain patient records or sparse patient data in the EHR, presence of codes within a quality measure concept code set that did not match exactly the diagnosis "rule" (e.g., for CHF the concept set includes diastolic heart failure while the set definition should apply to systolic or reduced ejection fraction), and reliance on provider-entered notes for evidence rather than patient self-report or other diagnosis verification. Moreover, a single EHR was studied, and local practice patterns may dictate the relative diagnostic accuracy of EHR locations; the generalizability of the study depends on whether these same docu-

mentation locations are available (likely, based on previous literature) and how different EHRs' modules and the single health systems' polices might shape their use, which can cause variance. Our rules to diagnose the conditions were imprecise partly because formal tests were often not done at the site, leading to lower than expected performance on rules alone; this issue is known for those developing phenotypes, as rigorous criteria will often exclude the vast majority of patients [11]. The low performance, however, is more due to the study design than the potential of these rules.

Future work may compare automated EHR phenotypes from different EHRs and compare similar metrics to understand how practice variation would affect prediction. Additionally, a prospective trial comparing different documentation locations in population-based decision making for allocation of resources would be valuable to understand if these locations may have other, more qualitative differences that assist in decision making. For instance, a problem list item may have more pertinence than individually billed diagnoses when considering the addition of longitudinal care management, whereas encounters may more appropriately drive targeted transitional care programs.

6. Conclusion

We successfully compared diagnoses within four EHR-based diagnosis documentation locations to a gold standard for accuracy and predictive validity, demonstrating that individual EHR locations were 70.7% accurate but that combining diagnosis data across locations matched the predictive validity of the gold standard.

Multiple choice questions

1. When choosing potential data locations in EHRs for diagnoses to use in predictive algorithms, which choice matches the findings?
A. Problem List would provide the best single source for overall accuracy;
B. Combination of sources would provide nearly equal prediction as a gold standard;
C. Encounter diagnoses would be the most precise;
D. Diagnoses from rules, rather than codes, would provide the best inputs.

Answer: B. The findings from the study indicate that for prediction of future events, combining different data locations provides similar prediction calibration and discrimination to a gold standard annotation set of diagnoses despite the loss of precision from the data mashing.
2. Data quality can be thought of as whether data is fit for particular use; in this case, the accuracy of diagnoses for stratification of risk for future outcomes. When considering accuracy of diagnoses from the EHR, what can manual annotation provide?
A. Limited benefit since the diagnoses are coded;
B. Clarification of existing diagnoses only;
C. Deeper understanding of the nuances of diagnoses;
D. Potential to deem up to a third of diagnoses as inaccurate.

Clinical Relevance Statement

This study is relevant to organizations and clinical teams engaged in examining their population of patients and to determine the best way to identify ongoing risks of adverse health outcomes and the resultant hospital stays and costs. It finds that using data from clinical information systems to summarize risks from common clinical diagnoses related to these outcomes leads to variable results in the presence of diagnoses but limited impact on risk prediction. This impact could also be addressed by combining all the potential locations of diagnoses, but the final estimations of risks are moderate and clinicians should be wary of their use in care.

Human Subject Research Approval

The study was performed in compliance with the World Medical Association Declaration of Helsinki on Ethical Principles for Medical Research Involving Human Subjects, and was reviewed by Oregon Health and Science University's Institutional Review Board.

Acknowledgements

Thanks to Doug Rhoton for technical assistance with data extraction and aggregation during this project, and to Steven Roncaioli for his assistance with the literature search.

Conflicts of Interests

There are no conflicts of interests to report.

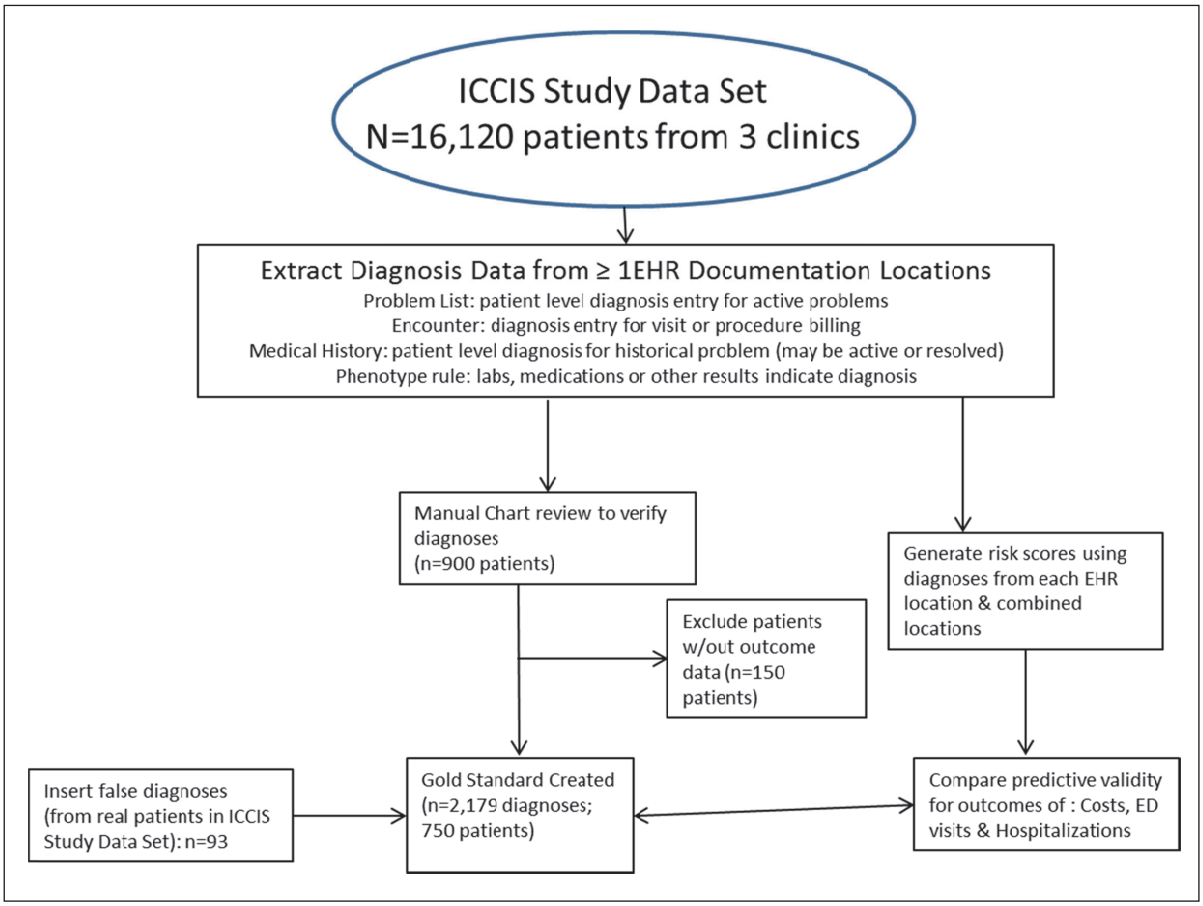

Fig. 1 Flow chart for methods and inclusion and exclusion criteria

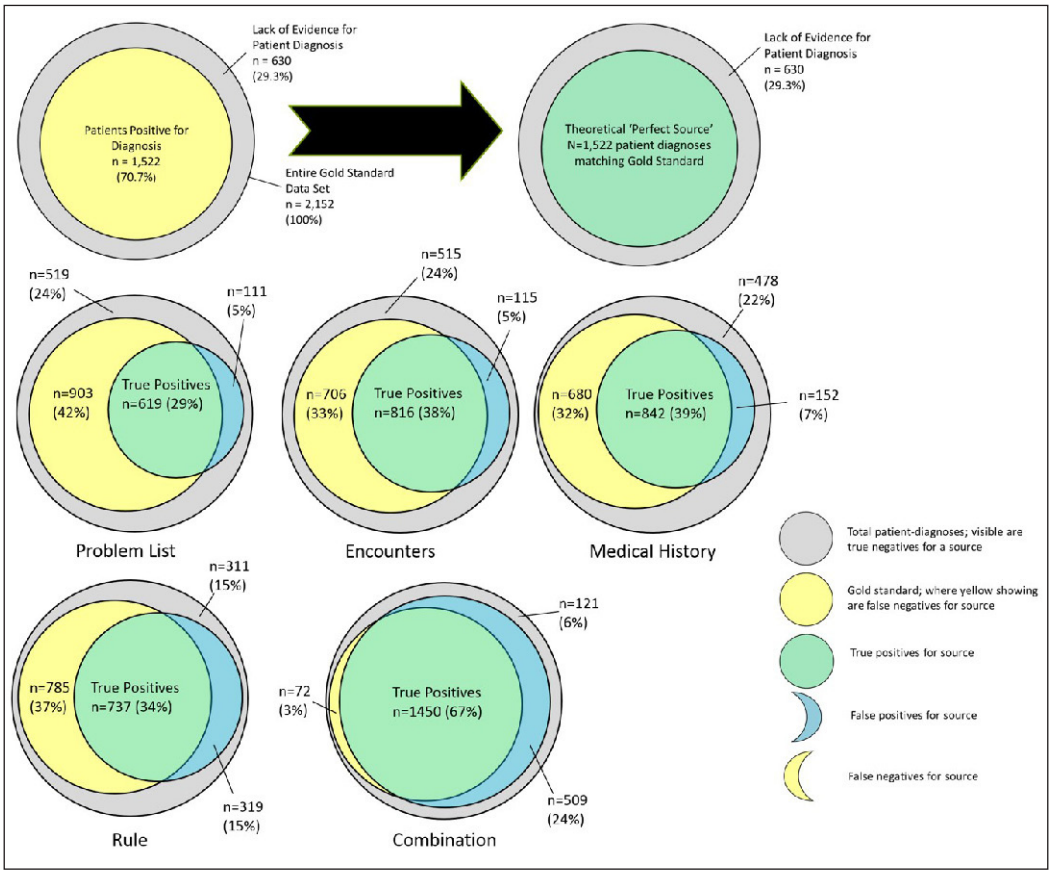

Fig. 2 Diagnostic accuracy by documentation location compared to overall sample and gold standard

Variable	N	%
Total	750	100
Female	423	56.4
Race/ethnicity		
White non-Hispanic	576	76.8
Black non-Hispanic	65	8.7
Asian non-Hispanic	63	8.4
American Indian non-Hispanic	13	1.7
Hispanic	30	4.0
Unknown	3	0.4
	Mean	**SD**
Age mean and spread	58.7	16.3
Age categories (years) 18–35 35–44 45–54 55–64 65–74 75–84 85+	N 69 82 156 161 148 94 40	% 9.2% 10.9% 20.8% 21.5% 19.7% 12.5% 5.3%

Table 1 Descriptive statistics of gold standard patient sample

Table 2 Risk scores and outcomes baseline characteristics for all patients (N=750); ACA = Accountable Care Act eligible conditions for both mental and physical health; MCCI = Modified Charlson Comorbidity Score; HCC = Hierarchical Condition Categories; ED = Emergency department

Risk Scores	Mean	SD	Median	Range
ACA	1.07	1.04	1	0–7.00
MCCI	1.18	1.62	1	0–9.00
HCC	0.93	0.61	0.78	0.12–4.32
Outcomes				
Cost (in USD)	$57,438	$173,445	10,520	$0– 2,042,931
Hospitalizations (number)	0.32	0.93	0	0–12.0
Bed-days (number of inpatient days)	1.79	6.2	0	0–67.0
ED Visits (number)	1.21	3.49	0	0–47.0

Table 3 Frequency of each diagnosis and gold standard review coding for presence; *Correct = gold standard confirms; Related/Inactive = gold standard revealed a related diagnosis or that the diagnosis was in remission or resolved; Incorrect = no evidence of diagnosis seen on gold standard

Diagnosis	Total		Correct		Related/Inactive		Incorrect	
	N	%	N	%	N	%	N	%
Alcohol/Drug Abuse	163	7.5	82	50.3	39	23.9	42	25.8
Asthma	364	16.7	231	63.5	38	10.4	95	26.1
Attention Deficit Hyperactivity Disorder (ADHD)	49	2.2	27	55.1	0	0.0	22	44.9
Breast Cancer	38	1.7	17	44.7	11	28.9	10	26.3
Congestive Heart Failure (CHF)	117	5.4	77	65.8	1	0.9	39	33.3
Chronic Kidney Disease (CKD)	162	7.4	114	70.4	0	0.0	48	29.6
Chronic Obstructive Pulmonary Disease (COPD)	119	5.5	85	71.4	14	11.8	20	16.8
Coronary Artery Disease	153	7.0	113	73.9	0	0.0	40	26.1
Depression	443	20.3	389	87.8	10	2.3	44	9.9
Diabetes without complications	264	12.1	183	69.3	48	18.2	33	12.5
Diabetes with complications	102	4.7	62	60.8	29	28.4	11	10.8
Human Immunodeficiency Virus (HIV)	133	6.1	113	85.0	0	0.0	20	15.0
Ischemic Stroke	72	3.3	48	66.7	0	0.0	24	33.3
Grand Total	2,179	100	1,541	70.7	190	8.7	448	20.6

Table 4 Diagnostic accuracy by location

Source	Sensitivity	Specificity	PPV	NPV
Problem List	0.41	0.82	0.96	0.4
Encounters	0.54	0.75	0.96	0.57
Medical History	0.55	0.76	0.91	0.44
Rule	0.48	0.49	0.87	0.56
Combined (All)	0.95	0.18	0.74	0.38

Table 5 Validation modeling: Risk score and documentation location variation by outcome (N=750); ACA = Accountable Care Act eligible conditions for both mental health and physical health; MCC = Modified Charlson Comorbidity score; HCC = Hierarchical Condition Categories; Combined = Using diagnoses from ANY of Encounter, Medical History, and Problem List. *p ≤ 0.05; Significant results at 95% confidence occur when the z-score is ≥1.96.

	EHR Documentation Locations				
Risk Score (Range)	**Encounter**	**Medical History**	**Problem List**	**Combined**	**Gold standard**
	Mean (Standard Deviation)				
ACA (0–7)	1.6 (1.3)	1.7 (1.3)	1.3 (1.1)	2.4 (1.5)	2.2 (1.4)
MCC (0–9)	1.5 (1.7)	1.7 (1.8)	1.2 (1.5)	2.3 (2.2)	2.2 (2.1)
HCC (0.1–4.3)	0.9 (0.6)	0.9 (0.6)	0.8 (0.6)	1.3 (0.8)	1.0 (0.7)
Outcomes	**Marginal Effects, estimated (z-score)**				
Cost					
ACA	$13,969 (2.53*)	$12,194 (1.68)	$10,403 (1.49)	$11,194 (2.18*)	$11,080 (2.08*)
MCCI	$6,102 (1.26)	$11,169 (0.74)	$8,107 (1.48)	$9,062 (1.17)	$8,841 (1.10)
HCC	$44,661 (3.27*)	$33,705 (1.69)	$39,172 (3.04*)	$35,795 (2.86*)	$45,690 (2.98*)
ED Visit					
ACA	0.169 (1.87)	0.150 (1.58)	0.094 (0.88)	0.170 (2.23*)	0.168 (2.54*)
MCCI	0.031 (0.44)	0.048 (1.71)	0.061 (0.76)	0.063 (2.01*)	0.061 (1.96*)
HCC	0.242 (1.23)	0.321 (1.28)	0.442 (2.30*)	0.393 (3.14*)	0.465 (3.10*)
Hospitalizations					
ACA	0.051 (2.28*)	0.036 (0.35)	0.018 (0.66)	0.044 (0.86)	0.052 (1.07)
MCCI	0.038 (2.31*)	0.027 (0.83)	0.023 (1.21)	0.026 (1.06)	0.026 (1.06)
HCC	0.124 (2.74*)	0.059 (0.24)	0.051 (1.09)	0.107 (1.90)	0.124 (2.03*)

References

1. Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, Kripalani S. Risk prediction models for hospital readmission: a systematic review. JAMA 2011; 306(15): 1688–98.
2. Mehmud S, editor A comparative analysis of claims-based tools for health risk assessment. Society of Actuaries Predictive Model Symposium; 2009 October 8–9, 2009; Chicago, IL.
3. Levine S AJ, Attaway K, Dorr DA, Leung M, Popescu B, Rich J. Predicting the financial risks of seriously ill patients. 2011: CHCF White paper. 2011 [September 21, 2015]; Available from: http://hcpinstitute.org/HCP%20Institute%20Predictive%20Modeling%20High%20Risk%20Patients%20Article_4.pdf.
4. Committee on Improving the Patient Record IoM. The computer-based patient record: meeting health care needs. In: Dick RS SE, Detmer DE, editor. The computer-based patient record: an essential technology for health care, revised edition. Washington, D.C. National Academy Press 1997; 74–99.
5. 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.
6. Jolly SE, Navaneethan SD, Schold JD, Arrigain S, Sharp JW, Jain AK, Schreiber MJ, Simon JF, Nally JV. Chronic kidney disease in an electronic health record problem list: quality of care, ESRD, and mortality. Am J Nephrol 2014; 39(4): 288–96.
7. Bell GC, Crews KR, Wilkinson MR, Haidar CE, Hicks JK, Baker DK, Kornegay NM, Yang W, Cross SJ, Howard SC, Freimuth RR, Evans WE, Broeckel U, Relling MV, Hoffman JM. Development and use of active clinical decision support for preemptive pharmacogenomics. J Am Med Inform Assoc 2014; 21(e1): e93–9.
8. Wright A. Improving Quality by Maintaining Accurate Problem Lists in the EHR (IQ-MAPLE). [September 21, 2015]; Available from: http://grantome.com/grant/NIH/R01-HL122225–01.
9. Wright A, McCoy AB, Hickman TT, Hilaire DS, Borbolla D, Bowes WA, 3rd, Dixon WG, Dorr DA, Krall M, Malholtra S, Bates DW, Sittig DF. Problem list completeness in electronic health records: A multi-site study and assessment of success factors. International journal of medical informatics 2015; 84(10): 784–90.
10. Meystre S, Haug P. Improving the sensitivity of the problem list in an intensive care unit by using natural language processing. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2006: 554–8.
11. Fort D, Weng C, Bakken S, Wilcox AB. Considerations for using research data to verify clinical data accuracy. AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science 2014; 2014: 211–7.
12. Dorr DA, Jones SS, Burns L, Donnelly SM, Brunker CP, Wilcox A, Clayton PD. Use of health-related, quality-of-life metrics to predict mortality and hospitalizations in community-dwelling seniors. J Am Geriatr Soc 2006; 54(4): 667–73.
13. Oregon Health Authority. Patient-Centered Primary Care Home Program. Payment Incentives. [updated July 16, 2014. May 18, 2016]; Available from: http://www.oregon.gov/oha/pcpch/Pages/payment-incentives.aspx.
14. U.S. National Library of Medicine. Value Set Authority Center. [updated 4/20/2016. 5/10/2016]; Available from: https://vsac.nlm.nih.gov/.
15. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011; 18(5): 552–6.
16. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc 2008; 15(1): 14–24.
17. Dale JA, Behkami NA, Olsen GS, Dorr DA. A multi-perspective analysis of lessons learned from building an Integrated Care Coordination Information System (ICCIS). AMIA Annual Symposium proceedings / AMIA Symposium 2012. 2012: 129–35.
18. Glick H. Methods for Cost Estimation in CEA: the GLM Approach. Academy Health, Issues in Cost-Effectiveness Analysis; Washington, D.C. 2008.
19. Jones AM, Lomas J, Moore PT, Rice N. A quasi-Monte-Carlo comparison of parametric and semiparametric regression methods for heavy-tailed and non-normal data: an application to healthcare costs. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2015:n/a-n/a.
20. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol 2004; 57(2): 131–41.