Evaluation of Data Quality of Multisite Electronic Health Record Data for Secondary Analysis

Alicia L. Nobles, Ketki Vilankar, Hao Wu, Laura E. Barnes
Department of Systems and Information Engineering
University of Virginia
Charlottesville, VA, USA
Email: {aln2dh, kkv2ad, hw4tm, lbarnes}@ virginia.edu

Abstract— **Currently, a large amount of data is amassed in electronic health records (EHRs). However, EHR systems are largely information silos, that is, uses of these systems are often confined to management of patient information and analytics specific to a clinician's practice. A growing trend in healthcare is combining multiple databases to support epidemiological research. The College Health Surveillance Network is the first national data warehouse containing EHR data from 31 different student health centers. Each member university contributes to the data warehouse by uploading select EHR data including patient demographics, diagnoses, and procedures to a common server on a monthly basis. In this paper, we focus on the data quality dimensions from a subsample of the data comprised of over 5.7 million patient visits for approximately 980,000 patients with 4,465 unique diagnoses from 23 of those universities. We examine the data for measures of completeness, consistency, and availability for secondary use for epidemiological research. Additionally, clinical documentation practices and EHR vendor were evaluated as potential contributors to data quality. We found that overall about 70% of the data in the data warehouse is available for secondary use, and identified clinical documentation practices that are correlated to a reduction in data quality. This suggests that automated quality control and proactive clinical documentation support could reduce ad-hoc data cleaning needs resulting in greater data availability for secondary use.**

Keywords-electronic health records; data quality; big data; multiple data vendors; metrics

I. INTRODUCTION

Electronic health records (EHRs) are electronic versions of a patient's medical history, maintained by a provider, that contain information relevant to a patient's care including demographics, diagnoses, medical procedures, medications, vital signs, immunizations, laboratory results and radiology images [1]. EHRs are the "big data" of healthcare; not only do EHRs allow more efficient management of a patient's clinical information, but the data collected in EHRs provides valuable opportunities for secondary use. Employing big data analytics, such as mining, on the immense, longitudinal data contained in EHR systems holds potential for valuable clinical, operational, and research applications [2]–[6]. Not only does EHR data provide a unique opportunity to examine longitudinal data across a diverse population, it is also advantageous because it eliminates or reduces patient recruitment and data collection, thereby reducing research costs and accelerating the time required for discoveries to reach application [7].

With the advent of federal initiatives supporting the adoption of EHR systems for "meaningful use" of health information technology and penalizing those who do not adopt these systems [8], the utilization of EHRs will continue to rise and the volume of amassed data will continue to grow. In 2011, the National Hospital Ambulatory Medical Care Survey reported 73% of hospital outpatient departments and 84% of emergency rooms used an EHR system, a dramatic increase compared to 29% and 46% in 2006 [9]. The U.S. Congressional Budget Office estimates that by 2019, 90% of physicians and 70% of hospitals will utilize comprehensive EHRs [3], [10].

Collectively, the data contained in EHR systems surpasses the size of many existing registries and data repositories [7]. Although EHR systems collect similar data, they were not built with the intent of interoperability [5], and, instead, are largely "information silos [4]". With the wide variety of implementations and permutations of EHR systems (e.g., there are currently more than 100 different EHR vendors on the market), the reality of merging and analyzing data from disparate systems poses significant challenges for research scientists due to variable data quality [4], [7]. In addition to typical single-source data cleaning needs (e.g., correction for data entry errors, missing data, or invalid data), cleaning of multisite EHR data has unique challenges [11] including consolidating data from heterogeneous systems [4] and standardizing varying clinical documentation practices [12], [13]. While it is generally accepted that data quality is dependent on the intended usage of the data (i.e., fitness for use) [6], [7], [14]–[16], there is no standardized methodology for assessing EHR data quality and, often, ambiguity in the terminology of data quality dimensions [7], [15], [17]–[19]. With all of the potential promises that secondary use of EHR data hold, as the popular expression "garbage in, garbage out" implies, one must start with a good database to produce a meaningful analysis [20].

In this paper, we examine data quality issues associated with secondary use of a structured multisite EHR database created for research purposes and identify potential factors contributing to the variance of data quality across participating institutes. The majority of previous studies have

978-1-4799-9926-2/15/$31.00 ©2015 IEEE

focused on conceptual frameworks of data quality [7], [15]–[19], [21], [22], a prototype platform to merge EHR data [23], or a population with a health condition resulting in a smaller study design [13], [24]. This paper expands the body of literature by examining data quality of a subsample of the College Health Surveillance Network (CHSN), a national data warehouse containing EHR records of patient visits to 31 different student health centers (SHCs) for medical care over a 4.5-year period. Each SHC participating in CHSN is unique with varying offerings of medical services, payment models for the services, and clinical documentation practices. We anticipate that the variation in administration and clinical documentation practices for the SHCs will contribute to the variance in data quality. Our research questions are:

- Is the EHR data complete, consistent, and available for secondary use?
- What clinical documentation practices potentially contribute to the variance of data quality amongst the single source sites?

The remainder of this paper describes the CHSN data warehouse, identifies and defines data quality dimensions, examines the data quality issues and potential contributing factors, discusses the challenges of using the EHR data for research, and suggests potential approaches to mitigate data quality issues.

II. Data Source

CHSN is a national data warehouse with aggregated de-identified longitudinal EHR data for SHCs at 31 universities spanning all geographic regions of the United States. Each month, the participating universities upload data derived from EHRs used for clinical and billing purposes for all patient encounters for students aged 15 to 50 that utilized the SHC in the previous month to a central data warehouse as shown in Fig. 1 [25]. Additionally, ten of the universities also upload data from their counseling centers [25]. Uploaded data include a confidential unique patient identifier, date of visit, diagnosis and procedure codes associated with the visit, and demographic information (i.e., all universities provide age, sex, and graduate/undergraduate standing of the patient and 13 universities provide ethnicity of the patient). Monthly data uploads are ongoing with most of the universities' initial uploads occurring in January 2011. Two of the universities began uploading in mid-2011 and one university began uploading in mid-2012 [25].

Total student enrollment in the CHSN universities represents 26% of the student enrollment in the 108 universities that are classified as universities with very high research activity [13] providing a unique opportunity for longitudinal epidemiological research studies for college student health.

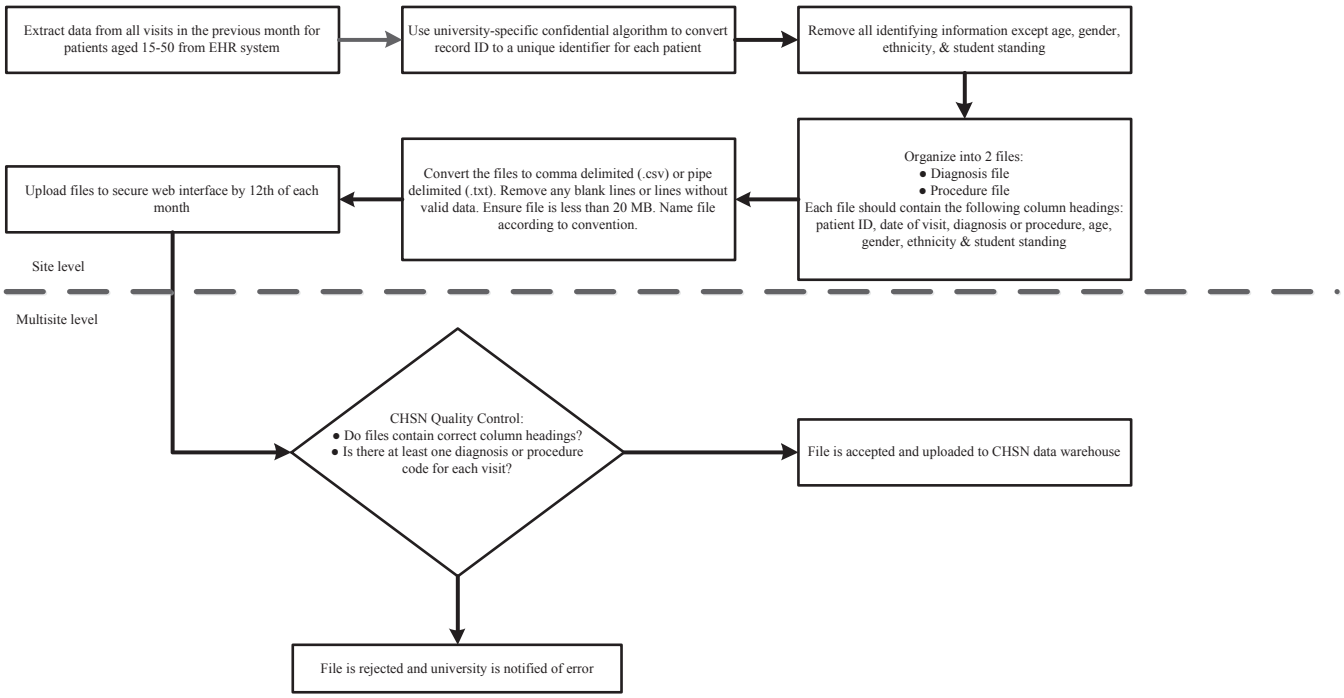

Figure 1. Monthly upload process.

The data used in this study is comprised of approximately 980,000 patients documenting over 5.7 million visits to 23 of the 31 SHCs for medical care between January 2011 and July 2015. The data contains 4,465 unique diagnoses for the patients. The EHR data examined include patient demographics (excluding ethnicity), diagnoses, and procedures. Ethnicity was excluded from the study because approximately half of the SHCs provide this data, and, of those SHCs, the data's trustworthiness is questionable (e.g., one SHC recorded all patients' ethnicity as "white"). This study focuses on 23 SHCs because of their participation in a survey conducted in Fall 2014 regarding clinical documentation practices. Since the survey, eight additional universities have joined CHSN. To maintain patient confidentiality, the universities as well as the EHR vendor are de-identified in this paper.

III. METHODS

A. Data Quality Dimensions

We conducted a literature search to identify studies and review articles about data quality in general and specific to EHRs. Data quality dimensions were developed in accordance with the reviewed literature and with the end-use of epidemiological research in mind. For the purposes of this study, we defined data quality in terms of the optimal scenario for epidemiological research (i.e., the intersection of patient demographics [excluding ethnicity], diagnosis, and procedure) as presented in Fig. 2. The data quality dimensions and measurements of the dimensions are presented in Table I. Measurements are based on the number of visits that qualify under the definition of each data quality dimension. The number of patients is based on the visits identified for each data quality dimension. We defined the data quality measurements as follows:

- **Completeness:** percentage of visits with an entry in the diagnosis and procedure fields,
- **Consistency:** percentage of visits that are complete and use the conventional coding schema for diagnoses (i.e., the International Classification of Diseases [ICD-9] established by the World Health Organization [26]) and procedures (i.e., Current Procedural Terminology [CPT] established by the American Medical Association [27]), and
- **Availability for use:** percentage of visits that are complete and consistent (i.e., the data available for research after cleaning).

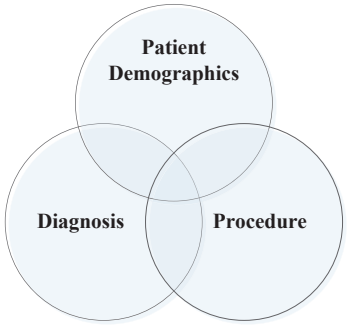

Figure 2. Venn diagram of EHR data contained in the CHSN data warehouse. The optimal scenario for epidemiological research is data comprised of patient demographics, diagnosis, and procedure.

Using the database queries presented in Fig. 3, counts of complete, consistent, and available data were extracted from the data warehouse for each university for all patient visits occurring in a 53-month time period from January 2011 to May 2015. Since clinical documentation practices are variable across student health centers, the rules presented in Fig. 4 were used to pre-process the data contained in the diagnosis and procedure fields to data that is consistent with the ICD-9 and CPT codes.

B. Clinical Documentation Practices

Each of the 23 CHSN member universities completed a web-based survey in November 2014 to gather information on the clinical documentation practices at the SHC. The survey assessed the following questions:

- Does the university's monthly data upload include patient visits to the counseling center?
- Are the providers at the SHC required to enter a diagnosis and procedure code for every patient visit?
- Does the the SHC use any university-specific codes for diagnoses or procedures (i.e., does the university have any special codes that they input into the diagnosis or procedure field that does not comply with the ICD-9 or CPT codes)?
- Does the SHC conduct periodic audits to evaluate the EHR data quality?
- Does the SHC bill all medical services to a third-party payer (i.e., private insurance company)?

The survey also included free-form text fields where the university could supply additional explanation about their clinical documentation practices.

TABLE I. DATA QUALITY DIMENSION AND MEASUREMENTS

Dimension	Conceptual Definition	Measure (Percentage)
Completeness	Presence of data	$\frac{N\ visits\ with\ entry\ in\ diagnostic\ field\ AND\ procedure\ field}{N\ visits}$
Consistency	Conformance of data values to other values in dataset	$\frac{N\ visits\ with\ ICD-9\ in\ diagnostic\ field\ AND\ CPT\ in\ procedure\ field}{N\ complete\ visits}$
Availability for Use	Data accessible to end user	$\frac{N\ visits\ with\ ICD-9\ in\ diagnostic\ field\ AND\ CPT\ in\ procedure\ field}{N\ visits}$

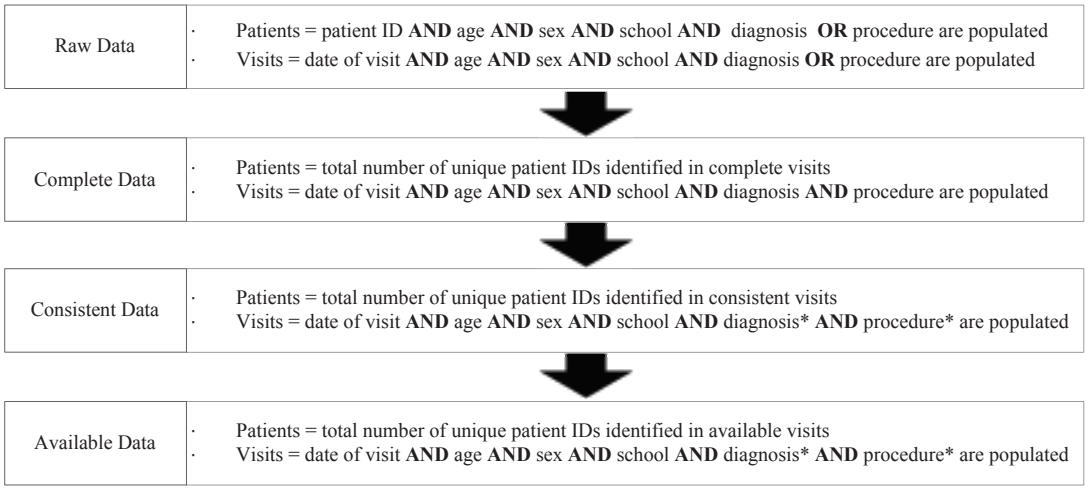

Figure 3. Database queries to extract counts of patient visits that are complete, consistent, and available.

C. *Exploratory Analysis of Variance in Data Quality*

In this study, the data quality measure of availability is dependent on completeness and consistency, therefore, only the variance in completeness and consistency were explored. We investigated factors that contributed to the variance in the data quality measures of completeness and consistency, both qualitatively and quantitatively.

Inferential statistics on completeness and consistency were performed to determine if statistical differences exist based on clinical documentation practices. The effects (i.e., documentation practices) evaluated included:

- **CC_Upload:** Binary variable (yes or no) indicating if the university uploads counseling center data to CHSN data warehouse.
- **Stnd_Codes:** Binary variable (yes or no) indicating if the university adheres to the ICD-9 and CPT standards when documenting diagnosis and procedures, respectively.
- **Bill_Out:** Binary variable (yes or no) indicating if all health services are billed to a third-party payer.
- **Audits:** Binary variable (yes or no) indicating if the university conducted periodic audits of EHR data quality.
- **EHR_Vendor:** Binary variable (most common EHR vendor versus other vendor) indicating if the university uses the most common vendor (*N = 10*) (EHR vendor E1) or a different form of vendor.

Statistical differences in data quality measures were identified using the t-test. The statistical significance level was set at $p = 0.05$.

IV. RESULTS

A. *Data Quality Dimensions*

Table II presents the data completeness, consistency, and availability for use for the 23 universities.

1) Completeness: Overall, 73.9% of the patient visits were complete (i.e., had entries in all fields) retaining 78.4% of the original number of patients in the CHSN data warehouse. At the university-level, data completeness ranged from 39.4% for University X to 99.8% for University A (*M = 75.5%, SD = 14.1%*).

2) Consistency: Overall, 95.3% of the complete patient visits were coded consistently (i.e., diagnosis and procedure codes adhered to the ICD-9 and CPT standards) retaining 98.3% of the number of patients from the complete data. At the university-level, data consistency ranged from 62.3% for

ICD-9	CPT
· Example: 487.0 indicates influenza with pneumonia · Numeric code with the exception of V code and E code. · V code or E code will have a leading V or E respectively. · Code may have a decimal in it . · Before decimal, there is a maximum of 3 places (Range: 1-3). · After decimal, there may be a maximum of 2 digits (Range: 0-2). · Remove any codes with all zeros.	· Example: 90658 indicates a flu shot. · Max length of 6 digits. · Truncate after first five characters. · If first digit is alphabetical, then convert to 99 and retain any numbers (e.g., some universities indicate provider using a letter, therefore J4210 is converted to 994210). · Value must be greater than 0,but less than 99999 or greater than 990000. · Remove any special characters (e.g., $,-,').

Figure 4. Rules for pre-processing diagnosis and procedure codes to standard ICD-9 and CPT formats.

TABLE II. DATA COMPLETENESS, CONSISTENCY, AND AVAILABILITY

University	Raw Data		Complete Data				Consistent Data				Available Data			
	Patients	*Visits*	*Patients*		*Visits*		*Patients*		*Visits*		*Patients*		*Visits*	
	No.	*No.*	*No.*	*Percent*	*No.*	*Percent*	*No.*	*Percent*	*No.*	*Percent*	*No.*	*Percent*	*No.*	*Percent*
A	52,633	295,638	52,545	99.8%	294,903	99.8%	52,505	99.9%	293,996	99.7%	52,505	99.8%	293,996	99.4%
B	39,034	291,654	37,633	96.4%	277,599	95.2%	37,448	99.5%	276,739	99.7%	37,448	95.9%	276,739	94.9%
C	25,876	134,399	23,234	89.8%	122,638	91.2%	23,243	100.0%	122,482	99.9%	23,243	89.8%	122,482	91.1%
D	28,093	184,240	25,218	89.8%	170,569	92.6%	24,540	97.3%	165,730	97.2%	24,540	87.4%	165,730	90.0%
E	50,537	232,719	47,252	93.5%	209,360	90.0%	47,249	100.0%	209,297	100.0%	47,249	93.5%	209,297	89.9%
F	18,623	68,335	16,593	89.1%	61,462	89.9%	16,581	99.9%	61,276	99.7%	16,581	89.0%	61,276	89.7%
G	60,959	265,532	52,889	86.8%	218,014	82.1%	53,128	100.5%	212,115	97.3%	53,128	87.2%	212,115	79.9%
H	57,386	238,008	38,989	67.9%	189,594	79.7%	38,971	100.0%	189,006	99.7%	38,971	67.9%	189,006	79.4%
I	28,173	119,493	24,044	85.3%	89,188	74.6%	24,019	99.9%	89,086	99.9%	24,019	85.3%	89,086	74.6%
J	18,960	119,537	15,085	79.6%	88,545	74.1%	15,081	100.0%	88,445	99.9%	15,081	79.5%	88,445	74.0%
K	31,353	137,362	22,373	71.4%	101,369	73.8%	22,373	100.0%	101,369	100.0%	22,373	71.4%	101,369	73.8%
L	40,943	265,449	34,001	83.0%	192,332	72.5%	33,946	99.8%	191,544	99.6%	33,946	82.9%	191,544	72.2%
M	43,261	306,141	25,311	58.5%	216,723	70.8%	25,305	100.0%	216,230	99.8%	25,305	58.5%	216,230	70.6%
N	41,266	238,417	30,842	74.7%	165,277	69.3%	30,932	100.3%	164,841	99.7%	30,932	75.0%	164,841	69.1%
O	35,586	237,873	23,791	66.9%	162,053	68.1%	23,779	99.9%	161,784	99.8%	23,779	66.8%	161,784	68.0%
P	64,362	510,286	54,769	85.1%	335,156	65.7%	54,731	99.9%	334,630	99.8%	54,731	85.0%	334,630	65.6%
Q	56,942	360,550	36,066	63.3%	241,429	67.0%	36,017	99.9%	234,624	97.2%	36,017	63.3%	234,624	65.1%
R	45,814	341,597	36,510	79.7%	200,647	58.7%	36,298	99.4%	195,737	97.6%	36,298	79.2%	195,737	57.3%
S	87,698	344,116	77,140	88.0%	254,093	73.8%	70,033	90.8%	197,179	77.6%	70,033	79.9%	197,179	57.3%
T	34,902	258,045	25,023	71.7%	144,687	56.1%	24,993	99.9%	144,308	99.7%	24,993	71.6%	144,308	55.9%
U	34,676	211,849	22,985	66.3%	179,048	84.5%	19,959	86.8%	111,622	62.3%	19,959	57.6%	111,622	52.7%
V	49,292	279,634	17,916	36.3%	192,173	68.7%	15,874	88.6%	145,577	75.8%	15,874	32.2%	145,577	52.1%
X	33,720	252,939	27,990	83.0%	99,552	39.4%	27,794	99.3%	99,056	99.5%	27,794	82.4%	99,056	39.2%
Total	980,089	5,693,813	768,199	78.4%	4,206,411	73.9%	754,799	98.3%	4,006,673	95.3%	754,799	77.0%	4,006,673	70.4%

University U to 100.0% for Universities E and K (*M = 95.7%, SD = 9.8%)*.

1) Availability for Use: Overall, 70.4% of the patient visits are available for secondary use (i.e., available for research purposes as defined in this study) retaining 77.0% of the original number of patients in the CHSN data warehouse. At the university-level, data completeness ranged from 39.2% for University X to 99.4% for University A (*M = 72.3%, SD = 15.6%)*.

B. Clinical Documentation Practices

Table III presents the results of the clinical documentation practices survey.

Ten of the 23 universities include data from counseling center visits in their monthly data uploads. The remaining 13 universities do not; however, mental health visits to primary care services are included.

Out of the 23 universities, only seven universities indicated that there are scenarios where a diagnosis and procedure code may not be coupled. Four of which indicated that some nurse only visits do not qualify for a procedure code, or some procedures, such as vaccinations, may not require a diagnosis code. Three universities indicated that they use a university-specific code for some internal purposes such as no shows, cancellations, sale of supplies (e.g., over-the-counter medication dispensed onsite), or a charge of a health service fee.

Twelve of the 23 universities indicated that they only use diagnosis and procedure codes that comply with the ICD-9 and CPT standards. The remaining 11 universities either indicated usage of university-specific codes or not recording a diagnosis or procedure code.

Eighteen of the 23 universities indicated that they conduct periodic audits for quality and usage of EHR data. However, there is no uniformity on the frequency of the audits or agreement on the type of audit conducted. For example, the frequency of audits may vary from daily to yearly, and audits can consist of physicians reviewing diagnosis codes to ensure a correct diagnosis or certified coders reviewing ICD-9 and CPT codes to ensure visits are correctly documented for third-party payers. Because of the lack of uniformity in frequency and agreement on type of audit, audits were not evaluated as an effect contributing to data quality.

Thirteen of the 23 universities bill all medical services to a third-party payer, seven universities bill some services, but not all (e.g., vaccinations are billed to a third party payer, but primary care services are not), and three universities do not bill any of their medical services to a third-party payer.

Eight different EHR vendor systems are utilized by the 23 universities. The most common vendors are E1 (*N = 10*), E2 (*N = 5*), and E3 (*N = 3*). The remaining five universities each use a different EHR vendor.

C. Variance of Data Quality

Table IV presents the variance in data quality for the significant effects.

TABLE III. RESULTS OF CLINICAL DOCUMENTATION SURVEY

Univ	Effects				
	CC_Upload	*Stnd_Codes*	*Audits*	*Bill_Out*	*EHR_Vendor*
A	No	No	Yes	Yes	E3
B	No	Yes	Yes	Yes	E2
C	No	Yes	Yes	Yes	E4
D	Yes	No	Yes	No	E2
E	No	Yes	Yes	Yes	E3
F	Yes	Yes	No	Yes	E2
G	No	Yes	Yes	No	E5
H	No	No	No	Yes	E1
I	No	No	Yes	Yes	E1
J	No	Yes	Yes	Yes	E1
K	No	No	Yes	Yes	E6
L	No	Yes	No	No	E1
M	Yes	No	Yes	No	E1
N	Yes	Yes	Yes	Yes	E1
O	No	No	Yes	No	E3
P	Yes	Yes	Yes	No	E1
Q	Yes	Yes	Yes	No	E1
R	Yes	No	Yes	Yes	E7
S	Yes	No	Yes	No	E1
T	Yes	Yes	No	No	E1
U	No	No	No	Yes	E2
V	No	Yes	Yes	No	E2
X	Yes	No	Yes	Yes	E8

TABLE IV. VARIANCE IN DATA QUALITY FOR SIGNIFICANT EFFECTS

Effect	Level	Mean	SD
CC_Upload	No	81.1%	10.4%
	Yes	68.3%	15.6%
EHR_Vendor	E1	68.9%	7.2%
	Other	80.7%	16.1%
Bill_Out	No	70.2%	10.6%
	Yes	79.6%	15.5%

1) Completeness: Lack of completeness was the largest contributor to reduced availability of data for secondary use resulting in 26.2% of the raw data being excluded.

Universities uploading counseling center data had statistically lower data completeness (*M = 68.3%, SD =* 15.6%) than did universities who do not upload counseling center data (*M = 81.1%, SD = 10.4%*), *t(21) = 2.36, p = 0.03*. Additionally, for the universities uploading counseling center data, mental health visits (in this study, defined by a procedure indicating the visit was for mental health services) had lower rates of completeness (i.e., a diagnosis

code is coupled with the procedure code for each visit) (*M = 53.5%, SD = 40.5%*) compared to primary care visits (*M = 91.5%, SD = 7.9%*). Two of the eight universities that upload counseling center data indicated that they do not record a dignosis code when patients see a provider in the counseling center. Four of the eight universities indicated that they do not use ICD-9 and/or CPT codes for recording diagnosis or procedure; most likely these universities use the Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV) codes for diagnosis which serve as a surrogate for ICD-9 codes, but do not correspond to CPT codes.

Universities using EHR vendor E1 had statistically lower data completeness (*M = 68.9%, SD = 7.2%*) than did universities using other types of EHR vendor (*M = 80.7%, SD = 16.1%*), *t(21) = 2.15, p = 0.04*. However, six of the 10 univerisies that use this vendor also upload counseling center data which may result in lower data completeness measures.

Universities that do not bill out all medical services to a third-party payer trended towards, but not statistically signifcant, lower data completeness (*M = 70.2%, SD = 10.6%*) than did universities that do bill out all medical services (*M = 79.6%, SD = 15.5%*), *t(21) = 1.64, p = 0.12*.

2) Consistency: Lack of consistency contributed considerably less to the reduction in the availability of data for secondary use resulting in only 3.6% of the raw data being excluded. No statistically significant differences were identified for data consistency based on clinical documentation practices. In genreal, most universities had high measures of data consistency (> 97.2%), however, three universities, Universities S, U, and V, have lower data consistency (77.6%, 75.8%, and 62.3%, respecitvely). The qualitiative responses provided in the survey did not provide insight into why these universities have lower data consistency.

3) Availability for Use: As mentioned previously, the measure of availability is dependent on the measures of completeness and consistency. The majority of the variance in availability is due to the variance in completeness, since most universities perform very well in regards to consistency.

D. Limitations of the Study

There are several limitations to the present study. First, these data quality measures examine the most optimal scenario where all data elements (excluding ethnicity) are present. Depending on the objective of the research, the data quality measurements presented in this study may be greater. For example, a research objective focused on the surveillance of influenza during flu season would only require diagnosis codes; therefore, the data available for this research purpose may be greater than that available presented in this study.

Second, the clinical documentation practices survey was not developed in parallel with this study, but rather prior to the data analysis to investigate potential predictors of quality. The survey was an online questionnaire developed with brevity in mind, and excludes a high level of detail on clinical documentation practices. Future work will focus on analysis at targeted universities identified across the spectrum from low to high performers in each quality metric.

Third, the response for the question regarding periodic audits was an unstructured free-text form. The variability in responses indicates the frequency and definition of audits may have been misinterpreted reducing the trustworthiness of the response.

Finally, our sample size was relatively small (N = 23) which may result in only very large effects being statistically significant.

V. CONCLUSIONS

Given that previous studies on EHR data quality focused on standardization of terminology as one of the major issues, we anticipated lower measures of data consistency and were surprised to discover the largest impediment to availability of data for secondary use is data completeness. We found that local administration and clincal documentation practices may considerably impact data quality, especially data completeness. This study revealed several factors correlated to data completeness issues including uploading counseling center data, type of EHR vendor, and billing of medical services to a third-party payer.

Counseling center data suffered from a lack of documentation of both a diagnosis and procedure code. Mostly, this is due to local data documentation procedures including using either proprietary coding schemas or the DSM-IV coding schema to document patient visits. Variations in user-friendliness of EHR vendor (e.g., graphical user interfaces and required navigation for data entry) can affect the user's interaction with the vendor increasing differences in data quality [24]. Additionally, each vendor varies in the data quality checks that are embedded in the system [24]. Typically, third-party payers only accept medical claims that document both a diagnosis and procedure. Due to the complex funding structures for student health centers, some centers do not bill all medical services to a third-party payer, thereby reducing incentive to thorough documentation of visits for the purpose of reimbursement for services provided. Other factors not examined in this study, such as clinical workflow and usage of medical coders, may also contribute to differences in data quality.

Overall, a substantial quantity of data in the CHSN data warehouse is available for secondary use. However, the quantity of data available could be increased by automating quality control and performing proactive clinical documentation support that improves the completeness and consistency of the data upfront rather than focusing on ad-hoc data cleaning [13]. Ad-hoc data cleaning is usually a time-intensive process, so preventing "dirty" data is an important step in reusing data for secondary purposes [11].

There are several initiatives that could support and incentivize more proactive quality control of clinical documentation. First, real-time advanced or automated data

validation and feedback embedded within EHR systems should be developed [13]. Validation may consist of ensuring demographics are correctly recorded (e.g., to avoid the aforementioned scenario where all patients are identified as "white") and counting diagnosis and procedure codes over time to identify overuse of certain codes which may indicate potential misuse. Feedback may consist of identifying co-occurrence of diagnosis codes indicating potential co-morbidities and other diagnoses consistent with the patient's demographics (e.g., diagnoses consistent with a patient's gender). Second, participants of data warehouses should agree to a "standard content" for uploaded data [13]. Third, participants of data warehouses should be provided feedback regarding data quality (e.g., feedback specific to the participant's data quality, other participant's data quality, and how each participant's data quality impacts the data warehouse holistically). Finally, research conducted using the data warehouse should be shared with the data warehouse community to reinforce the importance of the data for secondary use.

As we move forward in an era of emphasis on using EHR data for secondary use, our focus should shift to how we can combine disparate data sources to increase and improve data reuse for meaningful research to enable safer, faster, more cost-effective, and efficient healthcare [2]. To the best of our knowledge, this study is the first large-scale comparison of data quality of multisite EHR data for secondary use and attempt to examine variance in data quality by looking at clinical documentation practices. We hope that this study will contribute to the ongoing discussion of approaches to examine EHR data quality [15] and encourage initiatives to promote proactive clinical documentation practices to improve the data quality of research data warehouses.

ACKNOWLEDGMENT

The authors wish to thank the leadership of the participating schools for their commitment of time and resources to the College Health Surveillance Network Project and to their vision for furthering a better understanding of the epidemiology and health care needs of America's college students.

REFERENCES

[1] U. C. for M. and M. Services, "Electronic Health Records," 26-Mar-2012. [Online]. Available: https://www.cms.gov/Medicare/E-Health/EHealthRecords/index.html?redirect=/EHealthRecords/. [Accessed: 17-Aug-2015].

[2] R. Nambiar, R. Bhardwaj, A. Sethi, and R. Vargheese, "A look at challenges and opportunities of Big Data analytics in healthcare," *Proc. - 2013 IEEE Int. Conf. Big Data, Big Data 2013*, pp. 17–22, 2013.

[3] A. Bernard, "Healthcare Industry Sees Big Data As More Than a Bandage." [Online]. Available: http://www.cio.com/article/2383577/data-management/healthcare-industry-sees-big-data-as-more-than-a-bandage.html. [Accessed: 20-Jul-2015].

[4] C. Castaneda, K. Nalley, C. Mannion, P. Bhattacharyya, P. Blake, A. Pecora, A. Goy, and K. S. Suh, "Clinical decision support systems for improving diagnostic accuracy and achieving precision medicine," *J. Clin. Bioinforma.*, vol. 5, no. 1, p. 4, Mar. 2015.

[5] Murdoch T.B. and Detsky A.S., "The inevitable application of big data to health care," *Jama*, vol. 309, no. 13, pp. 1351–1352, 2013.

[6] A. R. Tate, N. Beloff, B. Al-Radwan, J. Wickson, S. Puri, T. Williams, T. Van Staa, and A. Bleach, "Exploiting the potential of large databases of electronic health records for research using rapid search algorithms and an intuitive query interface.," *J Am Med Inf. Assoc*, vol. 21, no. 2, pp. 292–8, 2014.

[7] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research.," *J. Am. Med. Inform. Assoc.*, vol. 20, no. 1, pp. 144–51, Jan. 2013.

[8] R. Steinbrook, "Health Care and the American Recovery and Reinvestment Act," *N Engl J Med*, vol. 360, no. 11, pp. 1057–1060, 2009.

[9] E. Jamoom and E. Hing, "Progress With Electronic Health Record Adoption Among Emergency and Outpatient Departments: United States, 2006–2011," *NCHS Data Brief, No 187. Hyattsville, MD: National Center for Health Statistics.*, 2015. [Online]. Available: http://www.cdc.gov/nchs/data/databriefs/db187.htm#summary. [Accessed: 23-Jun-2015].

[10] R. A. Sunshine, "Letter to Honorable Charles B. Rangel, Chairman, Committee on Ways and Means, U.S. House of Representatives. Washington, DC: Congressional Budget Office, January 21, 2009." [Online]. Available: http://www.cbo.gov/sites/default/files/hitechrangelltr.pdf. [Accessed: 23-Jun-2015].

[11] E. Rahm and H. H. Do, "Data Cleaning: Problems and Current Approaches," *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering*, 2000. [Online]. Available: http://dc-pubs.dbs.uni-leipzig.de/files/Rahm2000DataCleaningProblemsand.pdf. [Accessed: 23-Jun-2015].

[12] AHIMA, "Assessing and Improving EHR Data Quality (Updated)," *Journal of AHIMA*, 2013. [Online]. Available: http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_050085.hcsp?dDocName=bok1_050085. [Accessed: 18-Aug-2015].

[13] T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, "Secondary Use of EHR : Data Quality Issues and Informatics Opportunities," *Summit on Translat. Bioinforma.*, pp. 1–5, 2010.

[14] D. M. Strong, Y. W. Lee, and R. Y. Wang, "Data Quality in Context," *Commun. ACM*, vol. 40, no. 5, pp. 103–110, 1997.

[15] S. Dungey, N. Beloff, S. Puri, R. Boggon, T. Williams, and A. R. Tate, "A pragmatic approach for measuring data quality in primary care databases .," *Biomed. Heal. Informatics (BHI), 2014 IEEE-EMBS Int. Conf.*, pp. 797–800, 2014.

[16] S. Grannis, "Data Quality in the Era of 'Big Medical Data': Can We Really Believe That What the Data Says is Really Real?," 2012. [Online].Available: https://www.indianactsi.org/site/2012am/slides/ictsi2012_grannis_20120831.pdf. [Accessed: 02-Aug-2015].

[17] K. Thiru, A. Hassey, and F. Sullivan, "Systematic review of scope and quality of electronic patient record data in primary care.," *BMJ*, vol. 326, no. 7398, p. 1070, 2003.

[18] M. G. Kahn, M. A. Raebel, J. M. Glanz, K. Riedlinger, and J. F. Steiner, "A Pragmatic Framework for Single-site and Multisite Data Quality Assessment in Electronic Health Record-based Clinical Research," *Med. Care*, vol. 50, no. 0, pp. S21–S29, 2012.

[19] D. G. T. Arts, N. F. De Keizer, and G.-J. Scheffer, "Defining and improving data quality in medical registries: a literature review, case study, and generic framework.," *J. Am. Med. Inform. Assoc.*, vol. 9, no. 6, pp. 600–611, 2002.

[20] E. Grant, "The promise of big data | News | Harvard T.H. Chan School of Public Health." [Online]. Available: http://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/. [Accessed: 17-Jul-2015].

[21] K. Häyrinen, K. Saranto, and P. Nykänen, "Definition, structure, content, use and impacts of electronic health records: a review of the research literature.," *Int. J. Med. Inform.*, vol. 77, no. 5, pp. 291–304, May 2008.

[22] M. N. Zozus, W. E. Hammond, B. B. Green, M. G. Kahn, R. L. Richesson, S. A. Rusincovitch, G. E. Simon, and M. M. Smerek, "Assessing Data Quality for Healthcare Systems Data Used in Clinical Research."

[23] S. Rea, J. Pathak, G. Savova, T. A. Oniki, L. Westberg, C. E. Beebe, C. Tao, C. G. Parker, P. J. Haug, S. M. Huff, and C. G. Chute, "Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project.," *J. Biomed. Inform.*, vol. 45, no. 4, pp. 763–71, Aug. 2012.

[24] K. S. Chan, J. B. Fowles, and J. P. Weiner, "Electronic health records and the reliability and validity of quality measures: a review of the literature.," *Med. Care Res. Rev.*, vol. 67, no. 5, pp. 503–27, Oct. 2010.

[25] A. Keller and J. C. Turner, "College Health Surveillance Network: Epidemiology and Health Care Utilization of College Students at U.S. Four Year Universities," *J. Am. Coll. Heal.*, 2015.

[26] Centers for Disease Control and Prevention, "International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM). National Center for Health Statistics. Hyattsville, MD." [Online]. Available: http://www.cdc.gov/nchs/icd/icd9cm.htm. [Accessed: 30-Aug-2015].

[27] A. M. Association., *CPT 2011 Standard Edition.* Chicago, IL: American Medical Association Press, 2010.