Content and Quality of Free-Text Occupation Documentation in the Electronic Health Record

Ranyah Aldekhyyel, MS[1, 5], Elizabeth S. Chen, PhD[2], Sripriya Rajamani, MBBS, PhD, MPH[1, 3], Yan Wang, PhD[1], Genevieve B. Melton, MD, PhD[1, 4]
[1]Institute for Health Informatics, [3]Public Health Informatics Program, and [4]Department of Surgery, University of Minnesota, Minneapolis, MN; [2]Center for Biomedical Informatics, Brown University, Providence, RI; [5]Medical Education Department, College of Medicine, King Saud University, Riyadh, SA

Abstract

Recent recommendations for capturing social and behavioral information in electronic health record (EHR) systems for downstream applications, including research, highlight the need to better represent patient occupation. The objectives of this study were to characterize the content and quality of EHR social history module free-text occupation documentation. After developing categorization schemas, occupation entries with frequencies >5 (n=2,336) and a random sample of those with frequencies ≤5 (n=381) were analyzed. The information contained in the 2,336 entries fell into five groups: occupation (84.7%), occupation details (20.6%), employment status (2.5%), not in labor force (21.6%), and other (2.5%). Quality issues included use of acronyms/abbreviations (9.1%) and misspellings (1.6%). In comparison, quality issues with the 381 entries were: other (29.1%), acronyms/abbreviations (19.0%), and misspellings (9.0%). These findings suggest the need for EHR user training, system enhancements, and content standardization to support use of occupational information for clinical care and research.

Introduction

With the increasing use of Electronic Health Record (EHR) systems driven by various healthcare reform initiatives and the EHR Incentive Program[1], there is an opportunity for enhanced capture of occupation data electronically at the point of care. The importance of documenting social and behavioral factors influencing health status and outcomes has been recognized and supported by recommendations from respected advisory bodies and organizations, such as the National Academy of Medicine (NAM; formerly Institute of Medicine) and the National Institute for Occupational Safety and Health (NIOSH). For instance, in 2011 the NAM published a report entitled "Incorporating Occupational Information in Electronic Health Records,"[2] which highlighted the need for representing occupational information in EHRs through emphasizing the potential benefits to the individual patient as well as the population as a whole, and included recommendations for next steps. The report also illustrated several examples suggesting that the presence of the patient's type of work in the EHR could enable more accurate diagnosis and treatment of specific medical problems, which could lead to improved quality and efficiency of care. NIOSH supported incorporating occupation in the EHR through publishing demonstration projects, which focused on the representation of occupation related information. These include the Occupational Data for Health (ODH) data model[3] (work in progress, currently not available in the public domain), HL7 Clinical Document Architecture (CDA) standard template for "Occupation Data for Health"[4], and various pilot projects to understand and promote the capture of occupation data[3]. All these resources point to occupation as a complex concept with inter-related elements (e.g., occupation, industry, employer, and employment status).

In 2014, follow-up recommendations for capturing social and behavioral information in EHRs were issued in two reports published by the NAM[5, 6]. Occupation/Employment was considered under the socio-demographics domain in the first report ("Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1") and employment was identified as a candidate domain in the second report ("Capturing Social and Behavioral Domains and Measures in Electronic Health Records: Phase 2"). The report also indicated an instrument to standardize the collection of employment information, which was the Multi-Ethnic Study of Atherosclerosis (MESA) question and categories[6] for employment status.

Prior work has focused on reviewing the capture of social and behavioral factors in the EHR, involved characterizing social history information in clinical notes and identifying the eight most common statement types, one of which was occupation which represented 10-15% of statements across three different sources of clinical notes[7]. Subsequently, social and behavioral information was also examined in three public health surveys: Behavioral Risk Factor Surveillance System (BRFSS), National Health and Nutrition Examination Survey (NHANES), and National Health Interview Survey (NHIS)[8]. Occupation was a common survey item with 22 questions and corresponding responses across the three surveys.

Building upon findings from prior work and national recommendations for capture of occupational information in the EHR, the main objective of this research is to understand the practice of documentation of occupation in an institutional EHR by analyzing contents of free-text occupation entries assessing the type of information captured and evaluating the quality of data stored. This study uses similar methodologies for analysis as developed and applied in previous studies that focused on representation of various aspects of social history documentation in the EHR, including tobacco use, alcohol use, drug use and living conditions[9-12]. Specifically, the aim of this study is to determine issues associated with current representation of occupation in an EHR system with potential implications for enhancing system design for discrete data collection, user training, and standardization of data as well as informing the development of natural language processing (NLP) tools to enhance access to and use of structured data for research, clinical care, and population health purposes.

Methods

Study Design

This study involved a retrospective analysis of data collected in the free-text occupation field within the Epic EHR implemented at University of Minnesota (UMN)-affiliated Fairview Health Services accessible through the UMN Clinical Data Repository (CDR). Occupation and related information is collected in the socio-economic section in the social history module of the EHR and is comprised of one structured field (*Employer*) and two free-text fields (*Occupation* and *Comments*) (Table 1). The current study focused on the "occupation" field and was conducted in two phases for: (1) data extraction, preparation and transformation (Figure 1A) and (2) categorization schema development and application (Figure 1B).

Table 1. Example Occupation Entries.

Field	Example 1	Example 2	Example 3
Occupation	Truck Driver	asphalt / roofing, snow plowing in winter	10th grader - PLSHS
Employer	Wal-Mart	Other	NONE
Comments	Night Shift	Increased physical activity in summers	-

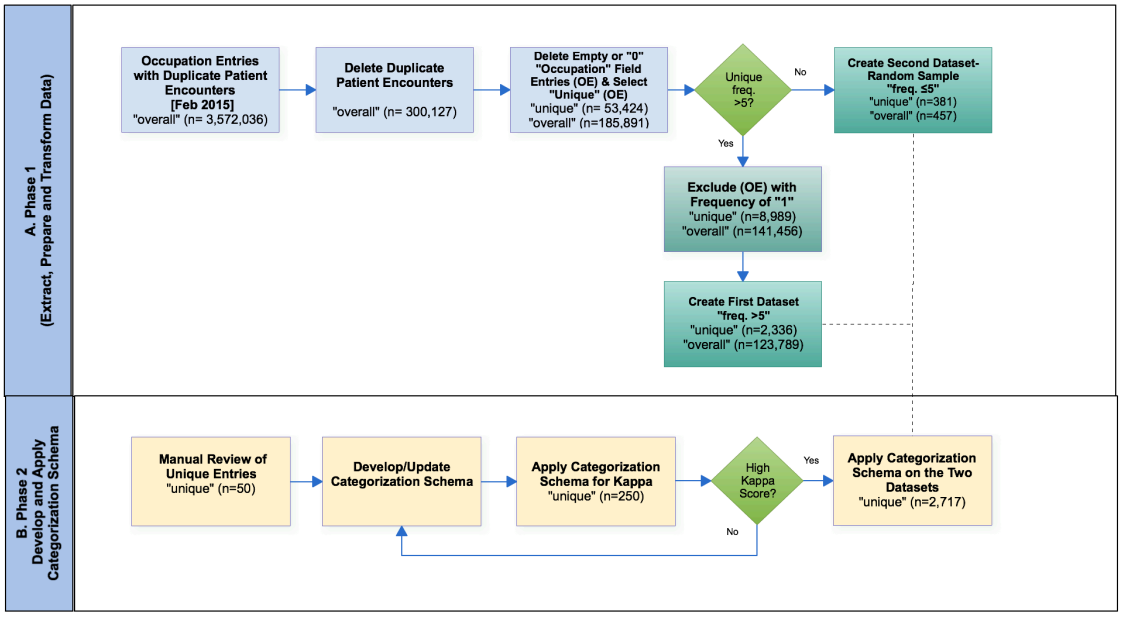

Figure 1: Study Methodology with Two Phases for Extracting, Preparing and Transforming Data (A) and Developing and Applying Categorization Schema (B).

Phase 1: Data Extraction, Preparation and Transformation (Figure 1A)

Occupation data from the UMN CDR collected from May 2000 through February 2015 were extracted resulting in 3,572,036 entries and imported into a relational database (MySQL). Removal of duplicate patient encounters (n=3,271,909) resulted overall in 300,127 entries. Many empty entries were found in the occupation field, which were then removed along with "0" values, as these entries were considered missing data. All entries were then converted to lowercase in order to select the "unique" entries stored in the free-text occupation field. A "unique" entry is an entry with no exact match within the dataset, taking into account space between words and punctuation marks (e.g., "homemaker" is considered a different entry than "home-maker"). The resulting dataset included 53,424 unique occupation entries representing 185,891 overall entries. The data were then divided based on the frequencies of unique occupation entries. A frequency of "5" was chosen as the cutoff for the two datasets representing about 90% of overall entries (25% of unique entries) excluding those with a frequency of "1". The first dataset "freq. > 5" was comprised of unique entries with frequencies of more than 5 (n=2,336) representing about 87% of the overall dataset (n=123,789). A random sample of the second dataset was extracted from the remaining dataset (5 or less) of 51,088 total unique occupation entries. The second dataset "freq. ≤ 5" included 381 unique occupation entries and was based on the total number of unique occupation entries with frequencies of 5 or less to provide a precision of 5% at the 95% confidence level.

Phase 2: Categorization Schema Development and Application (Figure 1B)

Schema Development

Categorization schemas were iteratively developed to manually categorize the "contents" of the unique occupation entries and identify any "data quality" issues with the entries. The "contents" schema was developed for identifying the unique contents of the free-text occupation entries. The "data quality" schema was developed to determine any data quality issues with the "contents" in the free-text occupation entries. Both of the developed schemas allowed for entries to be placed under several categories. The process of developing these schemas was based on an approach used in a prior study focused on analyzing free-text tobacco use documentation in the EHR[9]. The general method for developing these guidelines consisted of two phases. The first phase focused on developing initial categorization schemas based on analysis of 50 unique occupation entries and enhancing the schemas through weekly meetings and discussions involving four subject matter experts in the field of informatics with experience and expertise in clinical care, public health, and standards (GMM, ESC, RA, and SR). The second phase involved calculating inter-rater reliability using the kappa statistic to ensure consistency in categorizing entries between two reviewers (RA and SR) using the final versions of each categorization schema for 250 unique occupation entries from the "freq. > 5" dataset (n=2,336).

Earlier versions of the "contents" schema consisted of three main groups with twelve different categories. A notable topic of discussion among the group was focused on identifying the most applicable category for student-related entries. After review of six different sources of information (Standard Occupational Classification [SOC] System from the Bureau of Labor Statistics in the United States (U.S.) Department of Labor[13], Systematized Nomenclature of Medicine-Clinical Terms [SNOMED-CT][14], North American Industry Classification System (NAICS) from the U.S. Census Bureau[15], MESA[6], NIOSH[3] and MetaMap[16]) and then categorizing 50 entries from the "freq. > 5" dataset, the main groups "not in labor force" and "employment status" were created with associated categories and subcategories. The final version included five main groups, twelve categories, and four subcategories (Table 2). Categories and subcategories were created to provide further detail about the content being documented in the occupation entries. Subcategories are directly linked with a specific category (e.g., the subcategory "Type of Occupation" is directly linked to the "Name of Occupation" category and the subcategories "Student Related –Status", "Student Related –Type", and "Student Related –Other" are directly linked to the "Not in Labor Force –Student" category). This means that no entries can be placed under a specific subcategory without having part of the entry categorized under the related category. An example that explains the relationship between categories and subcategories is "Travel Consultant" where the entry was divided into two words: "Consultant" was categorized as "Name of Occupation" and "Travel" was categorized as "Type of Occupation". Another example is "Law Student" where "Student" was categorized as "Not in Labor Force-Student" and "Law" was categorized as "Student Related-Type".

The "data quality" schema underwent less revision since the categories were more distinct and captured common issues that are found in free-text data. The final version consisted of five main issues: (1) Misspelling, (2) Acronym/Abbreviation, (3) Ambiguous, (4) Multiple Terms, and (5) Other (Table 3). Two reviewers analyzed a set of 250 entries using the final versions of the categorization schemas. Inter-rater reliability was calculated using Cohen's Kappa, achieving κ of 0.94 for contents and 0.86 for data quality issues (percentage agreement of 0.99 and 0.98 respectively). All differences in categorization, between the two reviewers, were revised and resolved prior to applying the schema on the datasets.

Table 2: Categorization Schema for "Contents"

#	Category	Brief Description	Examples
Group (1): Occupation			
1	Name of Occupation	Describes what kind of work the patient does	• Teacher • Technician
1.1	Type of Occupation	Describes the type of a specific occupation. This code is directly linked to name of occupation.	• Electrical Technician - code as (1 and 1.1)
Group (2): Occupation Details			
2	Industry	Describes the type of work the patient's employer or business does. The large perspective of the work sector.	• Human Resources • Customer Service
3	Workplace	Describes the place or location where the patient works.	• Daycare • Warehouse
4	Job Duties	Describes the activity that the patient is performing as part of an occupation/job. Detailed specific task.	• Office work • Data Entry
5	Employer Name	The name of the patient's employer	• Target • Fairview
6	Equipment	Describes the necessary equipment for a particular occupation	• Computer • Heavy equipment
Group (3): Employment Status			
7	Employment Status	Describes that patient's current employment status	• Volunteer • Part time
8	Unemployment Status	Describes that patient's unemployment status	• Unemployed • Currently unemployed
Group (4): Not in Labor Force			
9	Not in Labor Force –Student	"Students" who are neither employed nor unemployed	• Student
9.1	Student Related –Status	Describes the student's current enrollment status. This code is directly linked with "student"	• Full time Student - code as (9 and 9.1)
9.2	Student Related –Type	Describes the student's major at school. This code is directly linked with "student".	• Nursing Student - code as (9 and 9.2)
9.3	Student Related –Other	Describes the grade or level of student. Does not have to be associated with student.	• 1st grade • Sophomore
10	Not in Labor Force –Other	Persons who are neither employed nor unemployed and are not students	• Retired • Home maker
Group (5): Other			
11	NA/None	Includes "NA" and "None"	• NA, None
12	Miscellaneous	Includes numeric values, type or status of patient, "other". Provide details as part of comment/notes.	• Child, Kid • Single

Schema Application

Coders manually categorized unique occupation entries using the two developed categorization schemas. The remaining 2,036 set of occupation entries, from the "freq. > 5" dataset (n=2,336), in addition to the "freq. ≤ 5" dataset (n=381) were divided in half between the same two reviewers who previously categorized the 250 unique occupations, to apply the developed schemas.

"PHY ED TEACHER" is an example of an entry from the "freq. > 5" dataset. This entry was categorized as "name of occupation" and "type of occupation" for "contents" and the "data quality" issues identified were "acronym/abbreviation" and "other-uppercase letters". Other examples of entries include "unemp" that was categorized as "unemployment status" with one identified quality issue of "acronym/abbreviation" and "Student-4th Grade" that was categorized as "not in labor force-student" and "student related-other".

Examples from the "freq. ≤ 5" dataset include "snowplow driver, road maintenence" and "legal assistant, retired 2012". Both entries were placed under several categories. The first entry was categorized as "name of

occupation", "type of occupation" and "industry" with quality issues of "misspelling" and "other-punctuation mark". The second entry was categorized as "name of occupation", "type of occupation", "not in labor force-other" and "miscellaneous-date" with quality issue of "other-punctuation mark".

After completing the manual categorization of unique occupation entries in both datasets (n=2,717), counts and percentages for both the "unique" and "overall" entries were calculated for the frequencies of "unique" occupation entries occurring within the dataset. "Overall" entries within a specific frequency group represent the overall number of times the corresponding "unique" entry occurred in the dataset.

Table 3: Categorization Schema for "Data Quality" Issues

#	Issue	Brief Description	Examples
1	Misspelling	Entry includes misspelling	• Message therapist • softwear engineer
2	Acronym/ Abbreviation	Entry includes acronym or abbreviation	• CAN • Administrative asst.
3	Ambiguous	Entry includes ambiguous information	• Account • Domestic Goddess
4	Multiple Entries	Entry includes two or more distinct terms that represent more than one category or two occupations divided by "/"	• Student/waitress • owner/operator
5	Other	Entry is informal use of word, all in uppercase letters, includes punctuation marks such as ',' or '.', etc. Provide details as part of comment/notes (e.g., "uppercase" or "punctuation mark").	• MANAGER • Society, quit

Results

Table 4 depicts the frequency distribution of unique and overall entries, and illustrates the selection of a frequency of "5" as the cutoff for the two datasets for performing data analysis. The table provides a summary of the "unique" occupation entries (n=53,424), which includes the overall and unique counts and percentages of occupation entries grouped into the frequency of the occurrence of a specific "unique" entry within the dataset. "Frequency" represents the number of times a specific "unique" entry occurred in the dataset. "Unique" occupation entries that occurred more than 10 times within the dataset have been grouped into the " >10" frequency".

Table 4: Frequency Distributions for Unique and Overall Occupation Entries

Frequency	# Unique Entries	% Unique	# Overall Entries	% Overall
>10	1,319	2.5%	116,190	62.5%
10	121	0.2%	1,210	0.7%
9	141	0.3%	1,269	0.7%
8	176	0.3%	1,408	0.8%
7	238	0.5%	1,666	0.9%
6	341	0.6%	2,046	1.1%
5	484	0.9%	2,420	1.3%
4	756	1.4%	3,024	1.6%
3	1,397	2.6%	4,191	2.3%
2	4,016	7.5%	8,032	4.3%
1	44,435	83.2%	44,435	24.0%
Total	**53,424**	**100.0%**	**185,891**	**100.0%**

Figure 2 compares the "overall" percentages of the "contents" of the unique occupation field distributed among the five main content groups between the two datasets. Noticeably, around 22% of the overall data being analyzed (in both datasets) were indicated as "not in labor force". The top five "overall" entries categorized as "not in labor force", regardless of identified data quality issues, were student (n=10,555), homemaker (n=4,693), "housewife" (n=958), retired (n=679) and stay at home (n=541). Results of the "overall" counts of the three

subcategories associated with the category "not in labor force-student" were: "student related – status" (n=31), "student related – type" (n=590) and "student related – other" (n=716).

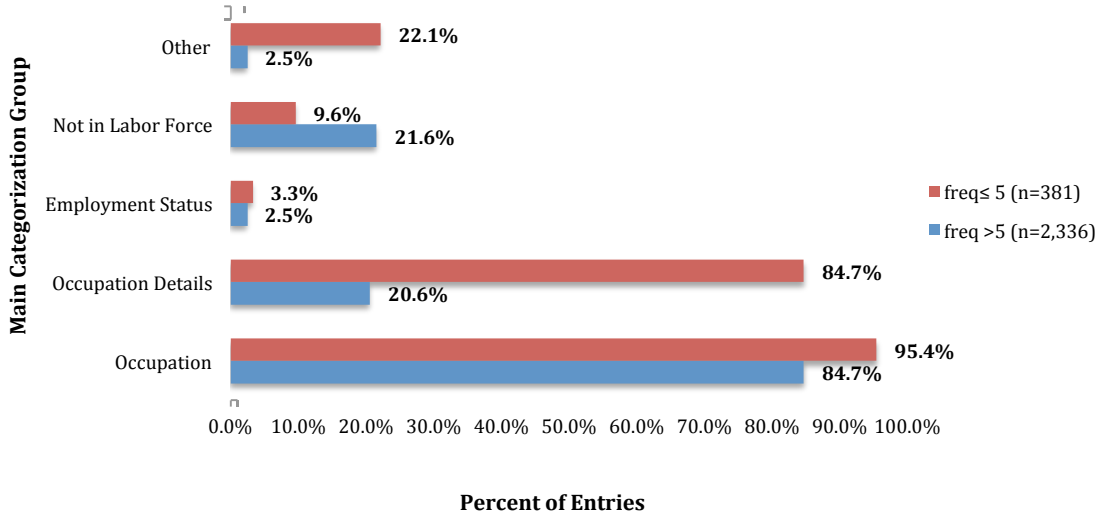

Figure 2: Distribution of the "Overall" Percent of Entries across Main Categorization Groups for "Contents"

Applying the categorization schemas to identify the contents and data quality issues of the two datasets being analyzed are summarized in Tables 5 and 6. The results are presented to show the count and percentages of both the "unique" and "overall" occupation entries between the two datasets. Percent of "unique" entries categorized as "name of occupation" (64.9%) was higher in the "freq. > 5" dataset. As anticipated, the percent of "unique" entries categorized as "miscellaneous" was much higher in the "freq. ≤ 5" dataset (23.9%). Examples of "miscellaneous" entries included "odd jobs", "infant", "data" and "grocery". The "freq. ≤ 5" dataset was also identified as having a higher percentage of quality issues indicated as "other" such as punctuation marks in the entry or the entire entry in uppercase letters.

Table 5. Categorization of Contents for Occupation Entries

#	Category	"freq. > 5" Dataset (n=2,336)			"freq. ≤ 5" Dataset (n=381)		
		Count	% *(unique)*	% *(overall)*	Count	% *(unique)*	% *(overall)*
Occupation							
1	Name of Occupation	1,515	64.9%	59.2%	215	56.4%	57.5%
1.1	Type of Occupation	1,072	45.9%	25.5%	142	37.3%	37.9%
Occupation Details							
2	Industry	234	10.0%	12.1%	27	7.1%	7.4%
3	Workplace	116	5.0%	2.9%	66	17.3%	16.0%
4	Job Duties	164	7.0%	4.1%	43	11.3%	9.4%
5	Employer Name	88	3.8%	0.9%	57	15.0%	14.0%
6	Equipment	22	0.9%	0.5%	2	0.5%	0.4%
Employment Status							
7	Employment Status	13	0.6%	0.2%	12	3.1%	2.6%
8	Unemployment Status	16	0.7%	2.2%	3	0.8%	0.7%
Not in Labor Force							
9	Not in Labor Force -Student	36	1.5%	9.0%	9	2.4%	2.4%
9.1	Student Related -Status	2	0.1%	0.0%	0	0.0%	0.0%
9.2	Student Related -Type	22	0.9%	0.5%	5	1.3%	1.1%
9.3	Student Related -Other	41	1.8%	0.6%	7	1.8%	1.5%
10	Not in Labor Force -Others	73	3.1%	11.5%	21	5.5%	4.6%
Other							
11	NA/None	3	0.1%	0.2%	0	0.0%	0.0%
12	Miscellaneous	76	3.2%	2.2%	91	23.9%	22.1%

Table 6. Categorization of Data Quality Issues for Occupation Entries

#	Issue	"freq. > 5" Dataset (n=2,336)			"freq. ≤ 5" Dataset (n=381)		
		Count	% *(unique)*	% *(overall)*	Count	% *(unique)*	% *(overall)*
1	Misspelling	130	5.6%	1.6%	35	9.2%	9.0%
2	Acronym/Abbreviation	869	37.2%	9.1%	74	19.4%	19.0%
3	Ambiguous	59	2.5%	0.8%	15	3.9%	3.3%
4	Multiple Entries	25	1.1%	0.1%	36	9.4%	7.9%
5	Other	36	1.54	0.9%	118	31.0%	29.1%

Significant quality issues were found in both of the study datasets. When looking at the number of quality issues identified within a single "unique" entry, it was found that in the "freq. ≤ 5" dataset, there were about 21% "unique" occupation entries that had one quality issue identified, 18% with two quality issues, 11% with three quality issues and 1% with four identified quality issues. Comparing these findings with the quality issues identified in the "freq. > 5" dataset, it was found that 9% had one quality issue, 12% with two quality issues, 1% with three quality issues and 1% with four identified quality issues.

Discussion

The overall goal of this study was to understand current documentation practices for occupation in the EHR for informing efforts to structure and standardize this information for subsequent use. Although, when developing the categorization schema, federal laws related to occupation were not considered (e.g., Fair Labor Standards Act), our analysis demonstrates that data being documented in the free-text occupation field under the socio-economic section of the EHR being studied partially reflect current standards and recommended formats from reports that have been published. As per the proposed NIOSH[3] representation as well as other national reports and recommendations, occupation and occupation details can be captured in the following categories: (1) occupation, (2) industry, (3) employment status, (4) employer, (5) work schedule, (6) occupational injury, (7) occupational exposures and (8) work relatedness. Preliminary findings from a related study focused on evaluating the adequacy of the ODH Model developed by NIOSH[3] supported the robustness of the ODH Model for representation of occupational information[18]. Designing the EHR to capture occupational information for a patient in a free-text field resulted primarily in concepts related to occupation, status, industry, and employer. Having the data stored in a free-text field also creates a chance for errors during the data entry process, which could, in turn, result in lower quality of occupation data for secondary applications like research and population health interventions.

Four main observations resulted from our analysis. Within the study dataset, 38% (n=114,236) of the entries had the number "0" entered in the occupation field or had no entry at all. These were considered missing values and thus were not included in the primary analysis. Another observation was that 14% (n=44,435) had an entry that only occurred once in the entire dataset. This was mostly due to user-created terms, errors or descriptive statements being entered in the field. Examples of such descriptive sentences are "homemaker, former rn, 13 kids, 2 biologic" and "fishing guide in Ontario during summers, retired science teacher". Reasons for this issue could be due to the fact that the occupation field is a free-text field, lack of awareness or training on the part of users on entering this information, or the absence of appropriate fields designed for structured documentation of associated occupational information. The third observation was related to the frequencies of the unique occupation entries. As the frequencies of the entries decreased, the relative number of identified quality issues increased. Therefore, the decision to split the data into two separate datasets based on the frequencies of the unique occupations was made for better overall management and analysis. The fourth observation was related to entries that described individuals not in the workforce or student-related entries such as "retired", "on disability", "stay at home mom", "housewife" and "4th grade student". Subject matter experts were faced with finding the most appropriate category to represent this group within the coding schema, as there were not any available standards that clearly addressed this group in a consistent manner. A recommendation from this is that there should be a separate field or set of fields in the EHR that represent these types of individuals and associated information. Also essential is the need to develop standards and measures that would address this specific issue and design the EHR to capture information for these individuals whom are not in the labor force.

The results of this study also highlighted significant data quality issues associated with this data field and the relative low utility of this data for secondary purposes such as research, policy and population initiatives. Due to the fact that the study only analyzed one free-text field (*Occupation*) and represented data from a single EHR in one healthcare system (Epic EHR at Fairview Health Services), these findings may not be generalizable or represent the quality of occupation documentation in other EHR systems or other clinical settings. Since Fairview Health Services includes six different hospitals (four are metropolitan based and two are rural), an

academic practice with quaternary care, and a community practice with a specialized children's hospital, the initial findings potentially represent a breadth of documentation practices for occupational information. In addition, the methods for characterizing the contents and quality of this information could be adapted and applied at other healthcare systems. Future studies will incorporate federal laws related to occupation, normalize the "occupation" field entries, group and aggregate synonym entries, and analyze the occupation field in relation to other associated fields in the social history module (i.e., *Employer* and *Comments* as shown in Table 1), as well as other parts of the EHR that may include occupational information (e.g., clinical notes). More formal data quality assessments, such as those described by Weiskopf and Weng[17], will also be needed.

Examining a single EHR that reflects a standard input mechanism by a particular vendor may result in innate limitations in how occupational data are currently being captured. Broader implications of this work include informing improved EHR interfaces for capturing occupational data and how to codify free-text occupation information stored in the EHR. This study also helps set additional foundation for efforts in the area of NLP to analyze free-text stored in the EHR related to the occupational history of patients. NLP techniques can be used to extract, structure, and encode relevant information from free-text data for subsequent use. Knowing what the target is, in this case occupational information, and the type of model used to extract from text will set the foundation for mapping to standardized terminologies such as SOC[13], NAICS[15], and SNOMED-CT[14] that include codes for occupation and industry. For example, the social context hierarchy in SNOMED-CT[14] is designed to cover social conditions and circumstances significant to healthcare that includes a sub-hierarchy for occupation, which could be used to standardize different values found from this study. Future work could involve developing and evaluating NLP techniques for the different categories of content in the occupation field as identified in this study.

Conclusion

With the increased adoption of EHR systems and the growing recognition of social factors in impacting health outcomes, there is a need to understand the current status of the information being stored and captured in EHRs to increase the value of information that can be obtained. This study involved performing a content analysis of data from a free-text occupation field of an EHR over a selected time period by categorizing the contents of information being captured and identifying associated data quality issues, using developed categorization schemas. The findings of this study have implications in terms of system design, user training, and implementation of relevant standards, including vocabulary related to occupation and industry.

Acknowledgements

This study was supported by the grant from the National Library of Medicine #R01LM011364 and the University of Minnesota Clinical Translational Science Institute #8UL1TR000114.

References

1. Office of the National Coordinator for Health IT (ONC). Health IT Dashboard [Internet]. [updated 2015; cited 2016 June 22]. Available from: http://dashboard.healthit.gov/index.php
2. National Academy of Medicine (NAM). Incorporating Occupational Information in Electronic Health Records: Letter Report. Washington, DC [Internet]. 2011 [cited 2016 June 22]. Available from: http://www.nap.edu/catalog.php?record_id=13207
3. National Institute for Occupational Safety and Health (NIOSH). Structuring Patient Work Information in EHRs to Improve Patient Care and Public Health; Occupational Data for Health (ODH) Model. Public Health Informatics Conference; Atlanta, GA. 2014.
4. Health Level Seven (HL7) - IHE PCC Technical Committee. Integrating the Healthcare Enterprise (IHE) Patient Care Coordination (PCC) Technical Framework Supplement - CDA Content Modules; Trial Implementation Guide [Internet]. 2014 Dec [cited 2016 June 22]. Available from: http://www.ihe.net/uploadedFiles/Documents/PCC/IHE_PCC_Suppl_CDA_Content_Modules.pdf
5. Institute of Medicine (IOM). Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1. Washington, D.C [Internet]. 2014 April [cited 2016 June 22]. Available from: http://www.nationalacademies.org/hmd/Reports/2014/Capturing-Social-and-Behavioral-Domains-in-Electronic-Health-Records-Phase-1.aspx
6. Institute of Medicine (IOM). Capturing Social and Behavioral Domains and Measures in Electronic Health Records, Phase 2. Washington, DC [Internet]. 2014 [cited 2016 June 22]. Available from: http://www.nap.edu/catalog/18951/capturing-social-and-behavioral-domains-and-measures-in-electronic-health-records

7. Chen ES, Manaktala S, Sarkar IN, Melton GB. A Multi-Site Content Analysis of Social History Information in Clinical Notes. AMIA Annual Symposium Proceedings. 2011;2011:227-236.
8. Melton GB, Manaktala S, Sarkar IN, Chen ES. Social and behavioral history information in public health datasets. AMIA Annual Symposium Proceedings. 2012 : 625-34 .
9. Chen ES, Carter EW, Sarkar IN, Winden TJ, Melton GB. Examining the use, contents, and quality of free-text tobacco use documentation in the Electronic Health Record. AMIA Annual Symposium Proceedings. 2014:366-74.
10. Chen E, Garcia-Webb M. An analysis of free-text alcohol use documentation in the electronic health record: early findings and implications. Applied clinical informatics. 2014;5(2):402-15. Epub 20.16/07/14
11. Winden TJ, Chen ES, Lindemann E, Wang Y, Carter EW, Melton GB. Evaluating living situation, occupation, and hobby/activity information in the Electronic Health Record. Proceedings of the 2014 AMIA Annual Symposium; 2014 Nov 15-19; Washington, DC; 2014. p. 139.
12. Carter EW, Sarkar IN, Melton GB, Chen ES. Representation of Drug Use in Biomedical Standards, Clinical Text, and Research Measures. AMIA Annual Symposium Proceedings. 2015;2015:376-385.
13. United States Department of Labor. Bureau of Labor Statistics. The 2010 Standard Occupational Classification System [Internet]. [updated 2010 March; cited 2016 June 22]. Available from: http://www.bls.gov/soc/major_groups.htm
14. U.S. National Library of Medicine. Unified Medical Language System [Internet]. [updated 2016 May; cited 2016 June 22]. Available from: https://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
15. United States Census Bureau. North American Industry Classification System (NAICS) [Internet]. [updated 2012; cited 2016 June 22]. Available from: http://www.census.gov/eos/www/naics/
16. National Library of Medicine (NLM). MetaMap - A Tool For Recognizing UMLS Concepts in Text. [updated 2016 March; cited 2016 June 22]. Available from: http://metamap.nlm.nih.gov/
17. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan 1;20(1):144-51.
18. Rajamani S, Chen E, Aldekhyyel R, Wang Y, Melton GB. Validating the Occupational Data for Health Model: An Analysis of Occupational Information in Reports, Standards, Surveys, and Measures. AMIA Annual Symposium Proceedings. Forthcoming 2016.