Assessing the Representation of Occupation Information in Free-Text Clinical Documents Across Multiple Sources

Elizabeth A. Lindemann[a], **Elizabeth S. Chen**[b], **Sripriya Rajamani**[c,d], **Nivedha Manohar**[e], **Yan Wang**[c], and **Genevieve B. Melton**[a,c]

[a]Department of Surgery, University of Minnesota, Minneapolis, MN, USA

[b]Center for Biomedical Informatics, Brown University, Providence, RI, USA

[c]Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA

[d]Public Health Informatics Program, University of Minnesota, Minneapolis, MN, USA

[e]West Side Community Health Services, St. Paul, MN, USA

Abstract

There has been increasing recognition of the key role of social determinants like occupation on health. Given the relatively poor understanding of occupation information in electronic health records (EHRs), we sought to characterize occupation information within free-text clinical document sources. From six distinct clinical sources, 868 total occupation-related sentences were identified for the study corpus. Building off approaches from previous studies, refined annotation guidelines were created using the National Institute for Occupational Safety and Health Occupational Data for Health data model with elements added to increase granularity. Our corpus generated 2,005 total annotations representing 39 of 41 entity types from the enhanced data model. Highest frequency entities were: Occupation Description (17.7%); Employment Status – Not Specified (12.5%); Employer Name (11.0%); Subject (9.8%); Industry Description (6.2%). Our findings support the value for standardizing entry of EHR occupation information to improve data quality for improved patient care and secondary uses of this information.

Keywords

Occupations; Social Determinants of Health; Electronic Health Records

Introduction

The need to accurately capture occupation and other social history information is central for the use of this information in the provision of direct clinical care and for secondary uses like research and risk stratification of patients [1]. As electronic health record (EHR) system use increases within healthcare organizations, there is an opportunity for improved capture of this information at the point of care. Within the United States, the National Academy of

Address for Correspondence: Genevieve B. Melton, MD, PhD, FACMI, Department of Surgery and Institute for Health Informatics, University of Minnesota, MMC 450, 420 Delaware Street SE, Minneapolis, MN, 55455, USA, gmelton@umn.edu.

Medicine (NAM) in their 2011 report "Incorporating Occupational Information in Electronic Health Records" advocates for inclusion of social factors in EHR systems, due to the important impact of social factors on health status and outcomes [2]. Similarly, Phases 1 and 2 of their 2014 reports "Capturing Social and Behavioral Domains and Measures in Electronic Health Records" [3, 4] further emphasize the importance of these factors. The National Institute for Occupational Safety and Health (NIOSH), a leader in efforts to qualify occupation information in the United States, has done extensive work to promote documentation of occupation information in a standard manner by creating the Occupational Data for Health (ODH) data model [5].

Prior work has been done to analyze broadly how social history information is captured in EHR systems [6, 7], public health datasets [8], and how occupation-related information appears within standards and within social history in the EHR [9, 11]. More specifically, using a top-down approach, Rajamani *et al.* examined reports, standards, surveys, and research measures to analyze the ODH model's ability to provide coverage for occupation-related information [9], including the Health Level Seven Clinical Document Architecture content module of the Integrating the Healthcare Enterprise Patient Care Coordination Technical Framework, which incorporates the ODH model [10]. Aldekhyyel *et al.* examined the content and quality of entries within the free-text occupation field of the Fairview Health Services enterprise EHR social history module [11]. The results of these studies provided a comprehensive overview of the current state of representation and standardization of occupation information. Collectively, they highlight that the ODH model is robust in coverage of occupation-related information, and that the content and quality of occupation information within the EHR can be inconsistent and variable.

The main objective of this study was to build upon prior efforts of this group of authors by looking at a range of free-text clinical sources to further inform occupation representations leveraging the NIOSH ODH model. In performing our analysis, we anticipate additional refinements thus enabling future standardization and additional insights into language around occupation, ultimately aiding in standards refinement and future natural language processing (NLP) efforts around occupation information.

Methods

Data Sources

This study utilized six clinical document sources comprising a mix of both publically available and local sources to analyze free-text mentions of occupation and related information within notes. Information was unstructured, yielding a large variety of occupation-related information. The public domain note sources used came from 491 "Consult – History and Physical" notes from MTSamples (MTS) [12] and 200 de-identified "H&P" (History and Physical Examination) notes from the University of Pittsburgh Medical Center (UPMC) [13] obtained via a data use agreement. Four document types were analyzed from the University of Minnesota (UMN)-affiliated Fairview Health Services Epic EHR: (1) Social History Documentation, (2) Social Work notes, (3) Physical Therapy notes, and (4) Occupational Therapy notes available through the University of Minnesota Clinical Data Repository (CDR) from 2013.

Analysis of Clinical Text

Clinical text analysis consisted of three main parts: (1) text de-identification, (2) schema creation, and (3) text annotation. Sentences from Fairview Health Services were anonymized prior to annotation using the Safe Harbor Method [14]. Codes for anonymizing data came from previous work [15]. Further obfuscation was used where necessary to avoid personally identifiable data, and descriptors were created to provide specificity regarding employer type. These included descriptors for level of government employers, hospitality industry employers, healthcare employers, education employers, and Fortune 500 employers. Fortune 500 companies were determined using the Fortune 500 yearly ranking for 2016 [16]. Any remaining companies that did not fit into more specific descriptors were anonymized to "company."

Annotations were made using the brat rapid annotation tool (BRAT) [17]. Annotation schema and guidelines were derived from the NIOSH ODH model categories and elements, providing the parent entities *Occupational History, Usual Occupation, Employment Status, Occupational Injury*, and *Occupational Exposure* and associated child entities. The Systemized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [18] was examined, but it was determined that the ODH model was more relevant for our corpora, given its usage in previous work by this group of authors [9]. An initial set of 25 sentences was annotated with a group of annotators, and 50 overlapping sentences were individually annotated to evaluate the annotation schema and guidelines (Figure 1) by three annotators (EL, SR, NM). Additional entities were created to increase coverage, including *Subject, Negation, Temporal, Occupational Conditions*, and *Occupation Status*. The final schema had 9 parent entities and 32 child entities, for a total of 41 elements (Table 1).

Annotators were instructed to annotate at the most specific level of detail. For example, in the sentence, "She is a registered nurse," the text "registered nurse" is annotated as the child entity *Occupation Description*, rather than the parent entity *Occupational History.* Parent entities were used as general annotation categories when a child entity could not be specified. In another example, "Prior to retirement, pt worked as a civil engineer," the sentence contains annotations relating to both clauses (i.e., "Prior to" is *Temporal – Time Frame*; "retirement" is *Employment Status – Retired*; "worked" is *Employment Status – Not Specified*; "civil engineer" is *Occupational Description).* Relationships were created to describe how entities are connected to each other. An overlapping set of 10% of sentences was annotated to calculate inter-rater reliability between three annotators (achieving a Cohen's kappa of 0.76 and proportion agreement of 0.95). When inter-rater reliability was ascertained, the remaining sentences were annotated using the most expanded version of the schema. Following annotation process, annotations were extracted by element, generating a list of values for each element. These values were grouped by similar meaning for highest frequency elements. For example, for *Occupation Description*, the group 'Legal Occupations' represents annotations 'paralegal,' 'attorney,' and 'tax attorney.'

Results

A total of 868 sentences from the six sources of clinical documents were annotated, yielding 2,005 annotations, which were mapped to 41 entities. The most frequent entities were:

Occupation Description (17.7%), *Employment Status – Not Specified* (12.5%), *Employer Name* (11.0%), *Subject* (9.8%), and *Industry Description* (6.2%). Table 1 summarizes the representation of entities across source type. The Fairview Social History Documentation sentences have the greatest variety of occupation information, containing 39 (95%) of 41 possible entities. The sentences from the Fairview Physical Therapy Notes have the least diversity in occupation-related information, containing annotations for 13 (32%) of the 41 possible entities. No annotations were made for the entities *Volunteer* and *Usual Occupation*.

Tables 2 through 5 provide summaries of value sets for the most commonly represented elements, demonstrating how elements were represented in clinical text. Table 2 summarizes the value sets for the most common entity, *Occupation Description.* Values are grouped according to the Bureau of Labor Statistics 2010 Standard Occupation Classification major groups [19]. Of the major groups in SOC, 22 of 23 are represented across sources. Military Specific Occupations is the only group not represented within these clinical sources, due to its small presence within the corpus, for *Occupation Description* annotations. Healthcare Practitioners and Technical Occupations are most common with 73 overall values (20.6%) and 51 unique values. Farming, Fishing, and Forestry Occupations are the least commonly represented, with 1 entry (0.3%).

The entity *Employer Status – Not Specified* is a child entity of *Occupation Status* that is heavily used through the annotation process, because annotation guidelines specified to annotate all tenses of the verb "work" as *Employment Status – Not Specified.* This was done to capture work behaviors when part-time or full-time employment status is not directly stated (Table 3). Consequently, "Work" is the most frequent group in this data set, with 214 total values (86.3%) and only 6 unique values. The "Other" group included a variety of other terms indicating work status, including "released to regular work" and "has been on light duty."

Table 4 summarizes the value sets for the entity *Employer Name.* All values from Fairview Health Services Epic EHR in this set were anonymized prior to annotation, and are grouped accordingly. Educational Services was the most frequently occurring group with 83 entries (37.6%). Student status is a contributing factor to this result, as schools were annotated as *Employer Name.* Military employment is least commonly seen with 1 entry (0.5%). Five entries (2.3%) did not provide enough information in order to group by employer type.

Table 5 summarizes the value sets for the entity *Subject.* This entity was added to the annotation schema after initial evaluation when a need for granularity was discovered. Parent group entries are most frequent with 137 total entries (74.5%) and 13 unique entries. Names of individuals and references to spouses were also common – de-identified prior to annotation as "Name" – 15.2% and 7.1% respectively.

Discussion

This study's findings demonstrate the wide variety and range of occupation-related information within EHR clinical texts. A total of 2,005 annotations were made for occupation-related information in 868 sentences from six clinical document sources. This

pervasiveness of occupation-related information within free-text notes points to the need for standardized entry within the EHR. In large part, the NIOSH ODH model proves appropriate for representing most of the occupation information, but several additions to the schema were necessary for comprehensive representation in our clinical text corpus. Specifically, the *Subject* entity is one of the most frequently occurring entities (9.8%), and was added during initial schema evaluation when need was assessed. Within this element, parent group entries are highly used, demonstrating a clinical interest in parental occupation and a prevalence of pediatric patients in this dataset. This correlation could also point to the relevance of a parent's occupation on a child's health outcomes. Similarly, the prevalence and interest in spousal occupations within this dataset points to the relevance of a spouse's occupation to an individual's health outcomes.

The Fairview Social History Documentation notes represent the largest portion of the dataset, with 553 sentences and 1,295 annotations. This source also presents the greatest coverage of occupation-related elements, with 39 of 41 possible elements across the dataset. *Volunteer* and *Usual Occupation* (parent entity) are the two elements not represented in the dataset; these elements were also not seen in any other source. The Fairview Physical Therapy Notes present the least coverage of elements with 13 of 41 possible elements. The Fairview Occupational Therapy Notes and Fairview Social Work Notes also present a more focused coverage of elements with 17 and 23 elements, respectively. This could be due to the specialty nature of these notes and more focused reasons for obtaining occupation information. The MTS sentences presented the greatest coverage of elements per sentence annotated with 35 elements represented in 171 sentences.

The addition of the element *Occupational Conditions* was useful in describing daily stressors of an individual's job that have noted long-term effects on health outcomes, but do not necessarily constitute injury or exposure. Among these are items such as "Pt describes standing 12 hour days on concrete for his job," "Patient works at a computer and on the phone all day," and "He is employed in sales, which requires quite a bit of walking, but he is not doing any lifting." While *Occupational Conditions* annotations comprise only 1.1% of total annotations, there were zero duplicate entries, indicating the scope of conditions – both physical and mental demands – individuals face in their occupations.

Occupation Description is the most frequent element seen in this dataset with 354 total values in 868 sentences. Among these values, 267 were unique, and all but one were categorized into 22 of 23 SOC major groups. The ability to classify a large number of unique entries into 22 groups proves that occupation descriptions could benefit from standardized entry of this information within the EHR. While value sets are wide reaching, many occupations fall into a smaller set of groups. *Employer Name* was also pervasive in the dataset with 221 total values. These were categorized into 7 groups based on anonymized codes.

We also observed that ambiguity was common within the *Occupation Status* related entities, such as *Self-Employed* and *Homemaker/Housewife.* For example, some individuals may hold more than one type of employment status (i.e., both a full-time position and a part-time position). Some individuals may have overlapping employment status (i.e., the term "stay at

home mom" implies both a caretaker and homemaker role). Self-employment also presents ambiguity, as this could refer to either full-time or part-time employment status in addition to self-employment.

The fifth most frequent element within the dataset was *Industry Description* (6.2%). This value set was not grouped by any classification set, because a large majority of entries were unique and presented varying degrees of information presented. The issue of discrepancies in information is best represented by *Industry Description*, but was prevalent throughout the dataset. This underscores the lack of standardization currently in place in documenting occupation information within the EHR, and the need for further work to characterize the wealth and variety of occupation information that has potential impacts on health status and outcomes.

This work builds off of previous work [8, 9] to identify how occupation information is represented across the literature and various aspects of the EHR. Work could be done involving comparisons against occupation information in EHR systems from other vendors. Future work will focus on dissemination of this research on the Brown Digital Repository [20], creating granularity within the current model, and lead to better NLP techniques to analyze occupation information. This will contribute towards standardized entry of occupation information within the EHR, promoting data quality and ultimately improving patient care and secondary use of occupation information.

Conclusion

As EHR system use becomes more widespread, it will become imperative to have standardized entry of factors that influence health status and outcomes. Several respected groups have recognized occupation as a factor of health status – with this recognition comes the need to understand and standardize how occupation-related information is being captured within the EHR. This study analyzed free-text clinical notes from a variety of sources in order to characterize the state of occupation-related information within the EHR. The NIOSH ODH data model proved robust in characterizing information content, and additions were made to the annotation schema to provide additional granularity. This work has potential to lead to more detailed and knowledgeable standards, and provides a basis for creating a standardized entry system within the EHR and improved NLP techniques.

Acknowledgments

References

1. Behfourouz HL, Drain PK, Rhatigan JJ. Rethinking the Social History. N Engl J Med. 2014; 371(14):1277–1279. [PubMed: 25271600]

2. National Academy of Medicine (NAM). Washington, DC: 2011. Incorporating Occupational Information in Electronic Health Records: Letter Report [Internet]. Available from: http://www.nap.edu/catalog.php?record_id=13207 [cited 2016 December 9]

3. National Academy of Medicine (NAM). [cited 2016 December 9] Capturing Social and Behavioral Domains in Electronic Health Records: Phase 1. 2014. Available from: http://

www.nationalacademies.org/hmd/Reports/2014/Capturing-Social-and-Behavioral-Domains-in-Electronic-Health-Records-Phase-1.aspx

4. National Academy of Medicine (NAM). [cited 2016 December 9] Capturing Social and Behavioral Domains in Electronic Health Records: Phase 2. 2014. Available from: http://www.nationalacademies.org/hmd/Reports/2014/EHRdomains2.aspx

5. 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.

6. Chen ES, Manaktala S, Sarkar IN, Melton GB. A multi-site content analysis of social history information in clinical notes. AMIA Annu Symp Proc. 2011:227–36. [PubMed: 22195074]

7. 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. AMIA Annu Symp Proc. 2014:139.

8. Melton GB, Manaktala S, Sarkar IN, Chen ES. Social and Behavioral History Information in Public Health Datasets. AMIA Annu Symp Proc. 2012:625–634. [PubMed: 23304335]

9. Rajamani S, Chen ES, 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 Annu Symp Proc. 2016:115–116.

10. Health Level Seven (HL7) – IHE PCC Technical Committee. [cited 2017 April 6] Integrating the Healthcare Enterprise (IHE) Patient Care Coordination (PCC) Technical Framework Supplement – CDA Content Modules: Trial Implementation Guide [Internet]. Available from: http://www.ihe.net/uploadedFiles/Documents/PCC/IHE_PCC_TF_Vol1.pdf

11. Aldekhyyel R, Chen ES, Rajamani S, Wang Y, Melton GB. Content and Quality of Free-Text Occupation Documentation in the Electronic Health Record. AMIA Annu Symp Proc. 2016:1708–1716. [PubMed: 28269929]

12. MTSamples. [cited 2016 December 9] MTSamples.com [Internet]. Available from: http://mtsamples.com

13. University of Pittsburgh NLP Repository. [cited 2016 December 9] Department of Biomedical Informatics, University of Pittsburgh. [Internet]. Available from: http://www.dbmi.pitt.edu/nlpfront

14. U.S. Department of Health and Human Services. [cited 2016 December 9] Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule [Internet]. Available from: http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

15. Moon, S., Pakhomov, S., Melton, G. [cited 2016 December 9] Clinical Abbreviation Sense Inventory [Internet]. Available from: http://conservancy.umn.edu/bitstream/handle/11299/137704/README_DeidentifedSymbolDataSet.txt?sequence=2&isAllowed=y

16. Fortune 500. [cited 2016 December 11] Fortune 500 List [Internet]. Available from: http://beta.fortune.com/fortune500/list

17. [cited 2016 December 9] Brat: Brat rapid annotation tool [Internet]. Available from: http://brat.nlplab.org

18. [cited 2017 April 6] SNOMED International SNOMED CT Browser [Internet]. Available from: http://browser.ihtsdotools.org/

19. Bureau of Labor Statistics. [cited 2016 December 9] The 2010 Standard Occupational Classification System [Internet]. Available from: http://www.bls.gov/soc/major_groups.htm

20. [cited 2017 April 7] Brown University Libraries, SFHERE Project [Internet]. Available from: https://repository.library.brown.edu/studio/collections/id_700/
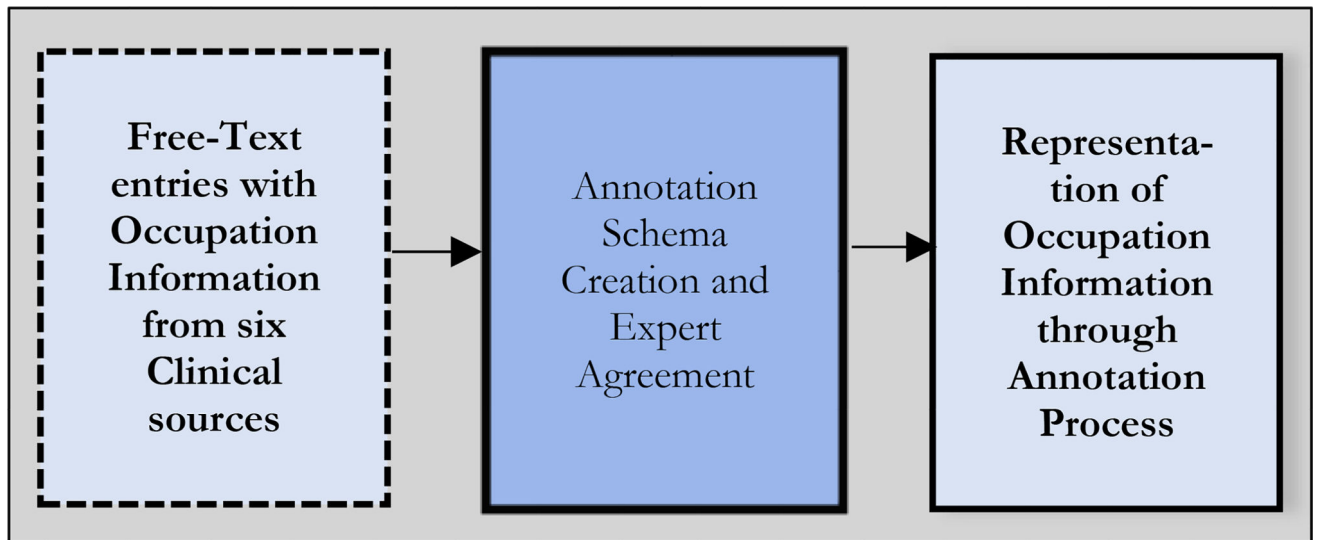
Figure 1.
Overview of Assessment of Occupation Information in Clinical Text

Table 1

Distribution of occupation entities across sources.

Entities	MTS [171] (n=385)*	UPMC [44] (n=101)*	Fairview SH Documentation [553] (n=1295)*	Fairview SW Notes [58] (n=113)*	Fairview OT Notes [16] (n=65)*	Fairview PT Notes [26] (n=46)*	All Sources [868] (n=2005)*
Occupational History	-	-	2 (0.2%)	-	1 (1.5%)	-	3 (0.1%)
Industry Description	22 (5.7%)	10 (9.9%)	80 (6.2%)	6 (5.3%)	5 (7.7%)	2 (4.3%)	125 (6.2%)
Occupation Description	77 (20.0%)	24 (23.8%)	227 (17.5%)	15 (13.3%)	3 (4.6%)	8 (17.4%)	354 (17.7%)
Job Duties	13 (3.4%)	-	55 (4.2%)	3 (2.7%)	3 (4.6%)	2 (4.3%)	76 (3.8%)
Employer Name	36 (9.4%)	7 (6.9%)	151 (11.7%)	18 (16.0%)	1 (1.5%)	8 (17.4%)	221 (11.0%)
Employer Location	7 (1.8%)	4 (4.0%)	48 (3.7%)	3 (2.7%)	-	3 (2.2%)	65 (3.2%)
Usual Occupation	-	-	-	-	-	-	-
Usual Industry Description	1 (0.3%)	-	3 (0.2%)	-	-	-	4 (0.2%)
Usual Occupation Description	5 (1.3%)	1 (1.0%)	4 (0.3%)	-	-	-	10 (0.5%)
Occupation Status	1 (0.3%)	-	3 (0.2%)	-	-	-	4 (0.2%)
Employment Status	4 (1.0%)	-	1 (0.1%)	-	-	-	5 (0.2%)
Full-Time	2 (0.5%)	1 (1.0%)	24 (1.9%)	4 (3.5%)	-	6 (13.0%)	37 (1.8%)
Part-Time	3 (0.8%)	-	9 (0.7%)	4 (3.5%)	-	1 (2.2%)	17 (0.8%)
Not Specified	46 (11.9%)	13 (12.9%)	141 (10.9%)	4 (3.5%)	41 (63.1%)	6 (13.0%)	251 (12.5%)
Self-Employed	4 (1.0%)	1 (1.0%)	2 (0.2%)	2 (1.8%)	1 (1.5%)	1 (2.2%)	11 (0.5%)
Employed but temporarily not working	1 (0.3%)	1 (1.0%)	1 (0.1%)	4 (3.5%)	1 (1.5%)	-	8 (0.4%)
Not Employed	12 (3.1%)	10 (9.9%)	30 (2.3%)	3 (2.7%)	-	-	55 (2.7%)
Retired	29 (7.5%)	-	56 (4.3%)	1 (0.9%)	-	-	86 (4.3%)
Disability (Not employed)	9 (2.3%)	-	4 (0.3%)	1 (0.9%)	-	-	14 (0.7%)
Military Service	2 (0.5%)	1 (1.0%)	1 (0.1%)	-	-	-	4 (0.2%)
Student	-	-	22 (1.7%)	-	-	-	22 (1.1%)
Student Type	1 (0.3%)	-	19 (1.5%)	-	-	-	20 (1.0%)
Student Status	5 (1.3%)	-	9 (0.7%)	1 (0.9%)	-	-	15 (0.7%)
Student Other	3 (0.8%)	-	55 4.2%	1 (0.9%)	1 (1.5%)	2 (4.3%)	62 (3.1%)
Volunteer	-	-	-	-	-	-	-
Not in Paid Workforce (Child)	3 (0.8%)	-	25 (1.9%)	-	-	-	28 (1.4%)
Homemaker/Housewife	8 (2.1%)	-	30 (2.3%)	1 (0.9%)	1 (1.5%)	-	40 (2.0%)

Entities	MTS [171] (n=385)*	UPMC [44] (n=101)*	Fairview SH Documentation [553] (n=1295)*	Fairview SW Notes [58] (n=113)*	Fairview OT Notes [16] (n=65)*	Fairview PT Notes [26] (n=46)*	All Sources [868] (n=2005)*
Caretaker	2 (0.5%)	-	6 (0.5%)	-	-	-	8 (0.4%)
Occupational Injury	2 (0.5%)	1 (1.0%)	3 (0.2%)	-	-	-	6 (0.3%)
Occupational Exposure	1 (0.3%)	2 (2.0%)	2 (0.2%)	-	-	-	5 (0.2%)
Occupational Conditions	8 (2.1%)	-	7 (0.5%)	3 (0.2%)	1 (1.5%)	4 (8.7%)	23 (1.1%)
Temporal	1 (0.3%)	-	3 (0.2%)	-	-	-	4 (0.2%)
Work Schedule	3 (0.8%)	-	5 (0.4%)	2 (1.8%)	1 (1.5%)	2 (4.3%)	13 (0.6%)
Start Date	4 (1.0%)	-	18 (1.4%)	2 (1.8%)	-	-	24 (1.2%)
End Date	2 (0.5%)	-	3 (0.2%)	-	-	-	5 (0.2%)
Days Per Week	-	1 (1.0%)	5 (0.4%)	-	1 (1.5%)	-	7 (0.3%)
Hours Per Week	-	-	3 (0.2%)	-	1 (1.5%)	1 (2.2%)	5 (0.2%)
Duration Years	1 (0.3%)	3 (3.0%)	10 (0.8%)	5 (4.4%)	1 (1.5%)	-	33 (1.6%)
Time Frame	38 (9.9%)	20 (19.8%)	55 (4.2%)	6 (5.3%)	1 (1.5%)	-	120 (6.0%)
Subject	10 (2.6%)	1 (1.0%)	162 (12.5%)	23 (20.4%)	1 (1.5%)	-	197 (9.8%)
Negation	6 (1.6%)	-	11 (0.8%)	1 (0.9%)	-	-	18 (0.9%)
# of Entities	35	16	39	23	17	13	39

([# of sentences per source];

*
n=number of annotations; SH=Social History; SW=Social Work; OT=Occupational Therapy; PT=Physical Therapy

Table 2

Distribution of values for Occupation Description element with grouping based on the 2010 Standard Occupation Classification groups [16]

Occupation Description			
2010 SOC Classification Group (n=23)	**Number of Total Values (n=354)**	**Number of Unique Values (n=267)**	**Frequency (%)**
Healthcare Practitioners and Technical Occupations	73	51	20.6
Business and Financial Operations Occupations	40	39	11.3
Education, Training, and Library Occupations	36	31	10.2
Construction	22	18	6.2
Management Occupations	21	15	5.9
Food Preparation and Serving Related Occupations	20	11	5.6
Office and Administrative Support Occupations	19	16	5.4
Architecture and Engineering Occupations	17	7	4.8
Installation, Maintenance, and Repair Occupations	15	11	4.2
Sales and Related Occupations	11	10	3.1
Transportation and Material Moving Occupations	11	2	3.1
Building and Grounds Cleaning and Maintenance Occupations	9	5	2.5
Personal Care and Service Occupations	8	6	2.3
Community and Social Service Occupations	8	6	2.3
Computer and Mathematical Occupations	7	7	2.0
Arts, Design, Entertainment, Sports, and Media Occupations	7	6	2.0
Legal Occupations	7	3	2.0
Life, Physical, and Social Science Occupations	6	6	1.7
Protective Service Occupations	6	6	1.7
Healthcare Support Occupations	5	5	1.4
Production Occupations	4	4	1.1
Farming, Fishing, and Forestry Occupations	1	1	0.3
Unclassified	1	1	0.3

Table 3

Distribution of values for Employment Status – Not Specified element

Employment Status – Not Specified			
Group (n=2)	Number of Total Values (n=248)	Number of Unique Values (n=13)	Frequency (%)
Work	214	6	86.3
Other	34	7	13.7

Table 4

Distribution of values for Employer Name element

Employer Name		
Group (n=7)	**Number of Total Values (n=221)**	**Frequency (%)**
Educational Services	83	37.6
Private and Publically held companies that do not qualify for Fortune 500 ranking	47	21.3
Health Care and Social Assistance	30	13.6
Fortune 500 Companies	25	11.3
Public Administration	19	8.6
Accommodation and Food Services	11	5.0
Not enough information for classification	5	2.3
Military	1	0.5

Table 5

Distribution of values for Subject element

Subject
184 Total Values; 28 Unique Values; 4 Groups

Group (n=4)	# of Total Values (n=184)	# of Unique Values (n=28)	Frequency (%)
Parent	137	13	74.5
Name	28	3	15.2
Spouse	13	6	7.1
Other Subject	6	6	3.3