Representation of Social History Factors Across Age Groups: A Topic Analysis of Free-Text Social Documentation

Elizabeth A. Lindemann, BS[1], Elizabeth S. Chen, PhD[2], Yan Wang, PhD[3], Steven J. Skube, MD[1], Genevieve B. Melton, MD, PhD[1, 3]
[1]Department of Surgery and [3]Institute for Health Informatics University of Minnesota, Minneapolis, MN; [2]Center For Biomedical Informatics, Brown University, Providence, RI

Abstract

As individuals age, there is potential for dramatic changes in the social and behavioral determinants that affect health status and outcomes. The importance of these determinants has been increasingly recognized in clinical decision-making. We sought to characterize how social and behavioral health determinants vary in different demographic groups using a previously established schema of 28 social history types through both manual analysis and automated topic analysis of social documentation in the electronic health record across the population of an entire integrated healthcare system. Our manual analysis generated 8,335 annotations over 1,400 documents, representing 24 (86%) social history types. In contrast, automated topic analysis generated 22 (79%) social history types. A comparative evaluation demonstrated both similarities and differences in coverage between the manual and topic analyses. Our findings validate the widespread nature of social and behavioral determinants that affect health status over populations of individuals over their lifespan.

Introduction

Increasingly, the importance of social and behavioral factors on an individual's health status are being recognized in clinical decision-making and effective population health management. As use of electronic health record (EHR) systems increases, there is a concomitant increase in EHR documentation that can be used to understand both how social history is documented and how these factors change over an individual's lifespan. Previous studies[1-9] have examined how individual social, behavioral, and environmental factors can be represented with controlled data representations as well as how these factors differ across documentation sources. For example, Rajamani *et al.*[1], Aldekhyyel *et al.*[2], and Lindemann *et al.*[3] examined how occupation is documented in standards and reports, free-text comment fields in an EHR, and free-text documentation from multiple clinical note sources, respectively. Chen *et al.*[4, 5], Wang *et al.*[6, 7], Winden *et al.*[8], and Carter *et al.*[9] examined documentation of tobacco, alcohol, and drug use within free-text fields of the EHR. Finally, Winden *et al.*[10] analyzed living situation, living conditions, and residence in standards. Since social history is often documented as free text and not structured data within clinical notes or in free-text documentation fields of the social history section of the EHR, this documentation and its associated topics, particularly with respect to an individual's lifespan, has had limited characterization. Ultimately, gaining a better understanding of the social history information documented provides an opportunity to understand these factors across populations and across the lives of patients.

As a patient ages, there is some evidence that relevant social history topics change dramatically. For infants and toddlers, many of the social determinants of health (SDOH) of parents and caretakers are particularly important including their mental health status and social support[11-14]. Within the literature, many of the direct references to the SDOH of infants and toddlers (ages zero to 24 months)[15] may refer to breastfeeding habits[16, 17], residence[18], living conditions[19], and palliative care for neonatal infants who have acute or potentially chronic care needs[20-22]. Residence is defined as the physical dwelling in which an individual lives or physical location, while the term living conditions refers to qualitative factors within these environments, such as secondhand smoke exposure, rodent infestations, water quality, and other relevant factors. As children enter early childhood, defined as ages two through five years by the National Institute of Child Health and Human Development (NICHD)[15], topics such as living conditions and residence remain important[23, 24], and other pediatric-specific patient factors are of increasing importance like sleep habits[25] and factors that may cause potential increase in obesity rates[26]. As children reach middle childhood (6-11 years)[15] and enter school systems, exposures related to this new environment are increasingly discussed in literature, such as bullying[27] and school performance. During this time, relational dynamics between both peers and family members and psychosocial effects are also increasingly important[28], and factors associated with diet and obesity remain popular topics[29].

When individuals reach adolescence (12-18 years)[15], SDOH topics broaden considerably and may more closely match social history topics for adult age groups. Documentation of alcohol use, drug use, and tobacco use now reflect individual use rather than exposure from parents or caretakers[30]. Also, sexual history topics may be mentioned during adolescence without the context of abuse[30]. In recent years, mentions of social media habits increase dramatically within literature as a factor of Internet presence[31]. Specifically of interest is the impact of social media usage on psychological state and its association with bullying and social dynamics[32].

The importance of standardizing how SDOH topics are entered into the EHR for clinical decision-making has been recognized by the National Academy of Medicine[33]. Towards this goal, understanding the current state of SDOH entry is important for determining what information is most valued by providers. In this study, we sought to identify how EHR social history topics change within clinical documentation as individuals age, and how social history topics vary by demographics, including gender, race, and ethnicity through automated topic analysis and manual analysis of social history topics in social history documentation for a large integrated healthcare system that serves both metropolitan and rural communities with primary and tertiary care.

Methods

At a high level, this study sought to provide a: (a) detailed analysis of social history topic variation by age and other demographics and (b) validation of the automated topic analysis through a separate manual analysis of social history documentation. An overview of our approach is provided in Figure 1 and was composed of four high level steps: (1) data collection, (2) topic analysis of pediatric and adult social history documentation, (3) manual validation of pediatric and adult social history documentation, and (4) comparison of social history topics from both analyses.

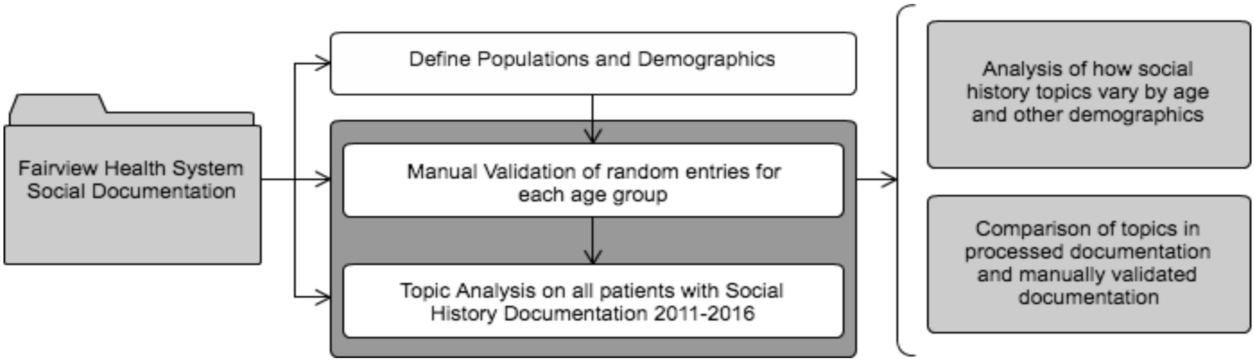
Fairview Health System Social Documentation
Define Populations and Demographics
Manual Validation of random entries for each age group
Topic Analysis on all patients with Social History Documentation 2011-2016
Analysis of how social history topics vary by age and other demographics
Comparison of topics in processed documentation and manually validated documentation

Figure 1. Overview of assessment of social history topics by age and other demographics. Populations and demographics of the patient set were defined first, while manual analysis and topic analysis were performed simultaneously.

Data Collection

Data sources for this study were the free-text Social History Documentation section within the enterprise EHR of Fairview Health Services (FHS). FHS is an integrated healthcare delivery system associated with the University of Minnesota that services both a large metropolitan area and rural parts of greater Minnesota. Social History Documentation from 2011 to 2016 was used for this analysis through the University of Minnesota research Clinical Data Repository (CDR). Each represented patient was used once in this five-year period to avoid oversampling from a single patient. Age groups were defined for pediatric patients according to the NICHD[15]. The age groups for Neonatal Stage (0-27 days after birth), Infants (28 days after birth to 13 months of age), and Toddlers (13 months of age to 24 months of age)[15] were combined, due to lower representation in FHS, compared with other age groups. Pediatric age groups for the analysis therefore consisted of: Infant and Toddler (birth to 24 months of age), Early Childhood (2-5 years), Middle Childhood (6-11 years), and Early Adolescence (12-18 years)[15]. Adult age groups were defined according to the United States Census guidelines[33] with age groups 19-24 years and 25-44 years were combined into a single Young Adulthood category. Adult age groups for the analysis therefore consisted of: Young Adulthood (19-44 years), Middle Adulthood (45-64 years), and Older Adulthood (65 years and older)[34]. A representative number of patients within the FHS was used for each age group. Other demographics including race, gender, and ethnicity were also obtained on patients for analysis in the study.

Manual Analysis

To compare the most common social history topics by age, 200 patient entries were randomly selected for each age group from the years 2011 to 2016. An annotation schema based on previous studies with used[35] consisting of twenty-eight major social history topics (Table 2). Annotators were instructed to annotate the text of the entire sentence for as many social history topics as the text contained. For example, the sentence “She lives with Mom and Dad in a single family home” would be annotated as Living Situation, Residence, and Family. An overlapping set of 10% of entries was annotated to perform inter-rater agreement by study investigators (EL, SS). Manual annotation for major social history topics in each age group resulted in a Cohen’s kappa of 0.67 and percentage agreement 0.92.

Automated Topic Analysis

Social history documents extracted from the CDR were pre-processed with an open source biomedical Natural Language Processing (NLP) pipeline to extract sections and split sections into statements[36]. The preprocessed statements were then parsed by Stanford Probabilistic Context-Free Grammars (PCFGs) parser[37]. We observed that multiple social history topics may be contained in a single statement. For example, the social history statement “Mom lives in an apartment and exercises twice a week” includes both information about the living condition and the hobby of the patient. For this reason, we collected all verb phrases of each social history statement separately as inputs for topic modeling. The subject of the verb phrase (e.g., “Mom” or “Siblings”), the verb (e.g., “exercise” or “live”), the verb prepositional conjunction word (e.g. “in”, “at”, or “with”) and sometimes the object of the verb was also collected. Verb frequencies of each group were computed, and those top occurring verbs were examined to decide if the object of the verb should be collected as part of the input for topic modeling. For instance, if a verb denotes no actions (e.g., “feel” or “report”), then the object of the verb (e.g., “safe” or “immunization”) was collected as part of the input for topic modeling. We also normalized language names, country names, relative names (e.g., “grandmother” or “nephew”), occupation name, and month names into single forms (e.g., “January” or “February”). Each word in the input was then normalized based on the SPECIALIST[38] Lexicon.

To decide the optimum topic numbers for each age group, we computed harmonic means of topic models with a sequence of topic numbers from 5 to 70, step by 3. Figure 2 shows the harmonic means change with different topic numbers chosen in topic modeling. In the following step, a topic model was built for each age group with the computed optimum topic number for each group. The R package ‘topicmodels’[39] was used in this study to compute the optimum topic number and topic model fitting.

Comparison of Major Topics between Corpora

Following annotation, major social history topic occurrences were extracted. The highest frequency topics in the manually validated corpus were compared against topics that arose through topic analysis. A comparative evaluation was performed in order to validate similarities between corpora. Topics generated from topic analysis were grouped according to the annotation schema and coverage was evaluated.

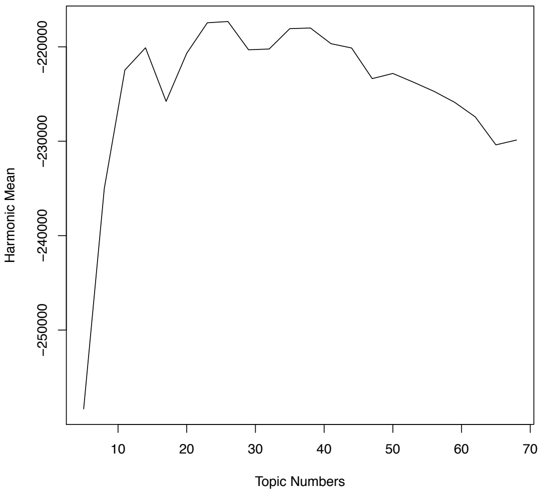

Figure 2. Harmonic means of topic models of infant toddler group with topic number from 5 to 70, step by 3.

Results

Manual Analysis

For each of the seven age groups, a total of 200 patients were selected at random (Table 1). This provided a total of 1,400 documents that were annotated for twenty-eight social history topics. Table 2 summarizes the representation of social history topics across age groups. A total of 8,335 annotations were made, with Young Adulthood containing the highest number of annotations (1,929) followed by Middle Childhood (1,170), and Early Adolescence (1,158). Sentences were annotated at the sentence level for all present social history topics, so in many cases, there were multiple annotations made for each sentence. There were slight discrepancies between study investigators in annotations, surrounding a question referring to guns and occupations of guardians. These differences were resolved after the validation stage was completed.

Table 1 summarizes the demographics seen across the manual validation corpus. While a diverse population is represented, patients were selected at random, and therefore, may not be completely representative of populations across the greater FHS system.

Table 1. Demographic information for patients included in manual validation corpus.

	Infant and Toddler (0-24 months)	**Early Childhood (2-5 years)**	**Middle Childhood (6-11 years)**	**Early Adolescence (12-18 years)**	**Young Adulthood (19-44 years)**	**Middle Adulthood (45-64 years)**	**Older Adulthood (65+ years)**
Number of Patients	200 (14.3%)	200 (14.3%)	200 (14.3%)	200 (14.3%)	200 (14.3%)	200 (14.3%)	200 (14.3%)
Gender – Male	111 (55.5%)	110 (55.0%)	116 (58.0%)	92 (46.0%)	46 (23.0%)	58 (29.0%)	80 (40.0%)
Gender – Female	89 (44.5%)	90 (45.0%)	84 (42.0%)	108 (54.0%)	154 (77.0%)	142 (71.0%)	120 (60.0%)
Number of Races Represented	10	10	9	10	10	6	7
African	14 (7.0%)	25 (12.5%)	24 (12.0%)	14 (7.0%)	9 (4.5%)	3 (1.5%)	7 (3.5%)
African American	9 (4.5%)	20 (10.0%)	19 (9.5%)	21 (10.5%)	13 (6.5%)	15 (7.5%)	2 (1.0%)
American Indian or Alaska Native	2 (1.0%)	1 (0.5%)	4 (2.0%)	3 (1.5%)	3 (1.5%)	3 (1.5%)	2 (1.0%)
Asian	17 (8.5%)	13 (6.5%)	5 (2.5%)	8 (4.0%)	9 (4.5%)	2 (1.0%)	4 (2.0%)
Hispanic or Latino	3 (1.5%)	2 (1.0%)	6 (1.0%)	1 (0.5%)	3 (1.5%)	-	-
Native Hawaiian or Other Pacific Islander	-	-	1 (0.5%)	-	1 (0.5%)	-	-
White	138 (69.0%)	122 (61%)	126 (63.0%)	141 (70.5%)	157 (78.5%)	169 (84.5%)	185 (92.5%)
Some other race	2 (1.0%)	3 (1.5%)	-	1 (0.5%)	-	-	-
Two or more races	3 (1.5%)	4 (2.0%)	6 (3.0%)	2 (1.0%)	1 (0.5%)	-	1 (0.5%)
Unknown	2 (1.0%)	2 (1.0%)	-	1 (0.5%)	1 (0.5%)	-	-
Choose not to answer	13 (6.5%)	12 (6.0%)	13 (6.5%)	10 (5.0%)	5 (2.5%)	8 (4.0%)	1 (0.5%)
Number of Ethnicities Represented	16	14	18	19	16	15	12

For all pediatric age groups, Family had the highest frequency of annotations, followed by Occupation and Living Situation. These topics are often seen together in the same text; for example, "She lives at home with Mom, Dad, and 2 siblings." Occupation can refer to the occupation of an individual, "He works in IT," or that of a parent or guardian, "Mom works from home, while Dad works in an office." While annotations for these categories remained high for all age groups, there is a decrease in the frequency of these annotations, and a subsequent increase in the annotations of the Social History (SH) "Other" category. There is a large portion of the Older Adulthood annotations that account for Marital Status; for example, "He is a widower" or "Lives at home with Husband." Although overall annotations for this Marital Status decrease by Older Adulthood, this type accounts for the third highest annotations for the age group.

Table 2. Distribution of manual annotations across major social history topics (200 notes per group) 2011-2016.

Topic	**Infant and Toddler (0-24 months)**	**Early Childhood (2-5 years)**	**Middle Childhood (6-11 years)**	**Early Adolescence (12-18 years)**	**Young Adulthood (19-44 years)**	**Middle Adulthood (45-64 years)**	**Older Adulthood (65+ years)**
Alcohol Use	-	1 (0.1%)	1 (0.1%)	5 (0.4%)	17 (0.9%)	18 (1.7%)	19 (2.4%)
Animals	22 (2.0%)	15 (1.3%)	27 (2.3%)	45 (3.9%)	60 (3.1%)	11 (1.0%)	12 (1.5%)
Caffeine Use	-	-	1 (0.1%)	3 (0.3%)	118 (6.1%)	58 (5.5%)	31 (3.9%)
Diet	10 (0.9%)	8 (0.7%)	-	1 (0.1%)	178 (9.2%)	133 (12.5%)	81 (10.2%)
Drug Use	5 (0.5%)	1 (0.1%)	1 (0.1%)	4 (0.3%)	11 (0.6%)	6 (0.6%)	9 (1.1%)
Exposure Other	11 (1.0%)	11 (1.0%)	14 (1.2%)	37 (3.2%)	1 (0.1%)	1 (0.1%)	4 (0.5%)
Family	286 (26.0%)	319 (28.3%)	273 (23.3%)	217 (18.7%)	161 (8.3%)	10 (10.4%)	94 (11.9%)
Family History	20 (1.8%)	1 (0.1%)	1 (0.1%)	7 (0.6%)	7 (0.4%)	11 (1.0%)	18 (2.3%)
Hobby	-	3 (0.3%)	12 (1.0%)	29 (2.5%)	5 (0.3%)	1 (0.1%)	5 (0.6%)
Hobby Exposure	-	-	-	-	-	-	-
Hobby Other	-	-	-	-	-	-	-
Living Condition	7 (0.6%)	6 (0.5%)	16 (1.4%)	41 (3.5%)	2 (0.1%)	1 (0.1%)	-
Living Situation	155 (14.1%)	178 (15.8%)	181 (15.5%)	159 (13.7%)	143 (7.4%)	4 (4.3%)	46 (5.8%)
Living Situation Exposure	11 (1.0%)	6 (0.5%)	20 (1.7%)	38 (3.3%)	16 (0.8%)	3 (0.3%)	2 (0.3%)
Living Situation Other	43 (3.9%)	47 (4.2%)	56 (4.8%)	27 (2.3%)	3 (0.2%)	3 (0.3%)	1 (0.1%)
Marital Status	86 (7.8%)	103 (9.1%)	98 (8.4%)	56 (4.8%)	134 (6.9%)	106 (10.0%)	88 (11.1%)
Occupation	189 (17.2%)	184 (16.3%)	223 (19.1%)	201 (17.4%)	141 (7.3%)	103 (9.7%)	72 (9.1%)

Table 2 Continued. Distribution of manual annotations across major social history topics for 200 notes per age group, from the years 2011-2016.

Topic	Infant and Toddler (0-24 months)	Early Childhood (2-5 years)	Middle Childhood (6-11 years)	Early Adolescence (12-18 years)	Young Adulthood (19-44 years)	Middle Adulthood (45-64 years)	Older Adulthood (65+ years)
Occupation Exposure	2 (0.2%)	-	2 (0.2%)	1 (0.1%)	-	1 (0.1%)	1 (0.1%)
Occupation Other	3 (0.3%)	-	-	1 (0.1%)	-	3 (0.3%)	2 (0.3%)
Physical Activity	3 (0.3%)	13 (1.2%)	16 (1.4%)	26 (2.2%)	141 (7.3%)	88 (8.3%)	60 (7.6%)
Physical Activity Exposure	-	-	-	-	-	-	-
Physical Activity Other	-	-	-	-	-	-	-
Residence	88 (8.0%)	79 (7.0%)	60 (5.1%)	72 (6.2%)	104 (5.4%)	38 (3.6%)	54 (6.8%)
Residence Exposure	-	4 (0.4%)	1 (0.1%)	-	-	-	-
Residence Other	-	4 (0.4%)	4 (0.3%)	6 (0.5%)	24 (1.2%)	12 (1.1%)	8 (1.0%)
Social Support	15 (1.4%)	3 (0.3%)	3 (0.3%)	6 (0.5%)	11 (0.6%)	13 (1.2%)	11 (1.4%)
Tobacco Use	63 (5.7%)	21 (1.9%)	22 (1.9%)	24 (2.1%)	76 (3.9%)	21 (2.0%)	22 (2.8%)
Travel	1 (0.1%)	2 (0.2%)	2 (0.2%)	4 (0.3%)	5 (0.3%)	6 (0.6%)	6 (0.8%)
Social History Other	81 (7.4%)	117 (10.4%)	136 (11.6%)	148 (12.8%)	567 (29.4%)	267 (25.2%)	145 (18.3%)
Total Annotations	1101	1126	1170	1158	1929	1060	791

Topic Analysis

A total of 187,920 patients were included in the automated topic analysis. Table 3 summarizes the demographics of the FHS population served from the years 2011 to 2016 with the Social History Documentation part of the analysis. This population is representative of the individuals serviced in the years specified. The age group for Young Adulthood has the most patients represented. Since adult age groups together contain more years compared to pediatric age groups, more patients are represented in adult age groups.

Table 3. Demographic information for patients across FHS Social History Documentation for the years 2011-2016.

	Infant and Toddler (0-24 months)	Early Childhood (2-5 years)	Middle Childhood (6-11 years)	Early Adolescence (12-18 years)	Young Adulthood (19-44 years)	Middle Adulthood (45-64 years)	Older Adulthood (65+ years)
Number of Patients	5,383 (2.9%)	10,837 (5.8%)	13,103 (7.0%)	14,120 (7.5%)	74,530 (39.7%)	44,355 (23.6%)	25,592 (13.6%)
Gender – Male	2,781 (51.7%)	5,683 (52.4%)	6,673 (50.9%)	6,599 (46.7%)	18,934 (25.4%)	15,220 (34.3%)	8,979 (35.1%)
Gender – Female	2,602 (48.3%)	5,153 (47.6%)	6,429 (49.1%)	7,520 (53.5%)	53,504 (71.8%)	29,135 (65.7%)	16,612 (64.9%)
Gender – Unknown	-	-	1 (0.0%)	1 (0.0%)	1 (0.0%)	-	-
Gender – NA	-	1 (0.0%)	-	-	1 (0.0%)	-	-
Number of Races Represented	11	11	11	11	11	11	11
African	293 (5.4%)	784 (7.2%)	1,121 (8.6%)	704 (5.0%)	3,015 (4.0%)	965 (2.2%)	359 (1.4%)
African American	394 (7.3%)	932 (8.6%)	1,158 (8.8%)	1,190 (8.4%)	3,846 (5.2%)	1,849 (4.2%)	537 (2.1%)
American Indian or Alaska Native	73 (1.4%)	152 (1.4%)	125 (1.3%)	184 (1.3%)	710 (1.0%)	395 (0.9%)	120 (0.5%)
Asian	385 (7.2%)	773 (7.1%)	803 (6.1%)	665 (4.7%)	4,439 (6.0%)	1,541 (3.5%)	682 (2.7%)
Hispanic or Latino	61 (1.1%)	199 (1.8%)	389 (3.0%)	340 (2.4%)	1,097 (1.5%)	441 (1.0%)	90 (0.4%)
Native Hawaiian or Other Pacific Islander	16 (0.3%)	31 (0.3%)	28 (0.2%)	34 (0.2%)	140 (0.2%)	54 (0.1%)	17 (0.1%)
White	2,442 (45.4%)	6,171 (56.9%)	7,892 (60.2%)	9,115 (64.6%)	46,302 (62.1%)	33,761 (76.1%)	21,287 (83.2%)
Some other race	30 (0.5%)	61 (0.6%)	131 (1.0%)	129 (0.9%)	464 (0.6%)	294 (0.7%)	150 (0.6%)
Two or more races	56 (1.0%)	218 (2.0%)	410 (3.1%)	224 (1.6%)	535 (0.7%)	122 (0.3%)	26 (0.1%)
Unknown	1,306 (24.3%)	937 (8.6%)	7,892 (3.7%)	690 (4.9%)	7,987 (10.7%)	2,906 (6.6%)	1,499 (5.9%)
Choose not to answer	527 (9.8%)	966 (8.9%)	929 (7.1%)	1,138 (8.1%)	4,985 (6.7%)	2,322 (5.2%)	941 (3.7%)
Number of Ethnicities Represented	54	59	60	63	66	65	63

Table 4 summarizes the ten most frequent topics for each age group identified using the topic analysis methods. These topics were grouped according to the schema of social history topics created for the manual analysis corpus. Many topics are slightly overlapping, causing there to be multiple entries for topics with similar content. For example, Marital Status is mentioned as the second, third, and fourth most occurring topic for Infant and Toddler, but may contain granularity that is not captured by the social history type classification.

Table 4. Representation of ten most frequent topic analysis topics, in order of frequency, with keywords and examples.

Topic # Assigned Name	Keywords	Examples
1 SH Other	Seatbelt Language Abuse Helmets	Primary Language Spoken: English Do you/your family use safety helmets? Abuse: Current or Past (Physical, Sexual, or Emotional) Seatbelts used.
2 Residence	Location Names House Apartment	NAME lives with her parents and sister in PLACE. Environmental History: The family lives in a 6 year old home in a rural setting. Lives with parents and half sister in the upstairs of a house while grandmother and a few cousins live on the first floor of a house.
3 Living Situation Exposure	Smoke exposure Safe Guns	NO: Lead, smokers at home, radon, pool/spa, known TB exposure. Do you feel safe in your home: Yes/No No guns at home.
4 Family	Parents Brother Sister	Lives at home with mother, grandmother, aunt and 2 older half-siblings. Lives with biological mother and maternal half sister.
5 Marital Status	Relationship Married Single	Parent Relationship: Married Divorced-almost. Lives with husband and 2 dogs.
6 Occupation	Work Part-time/Full-time Company Names	Patient did some part time work as a cashier for COMPANY in the past. Mom is a homemaker. Father is an engineer at COMPANY.
7 Living Situation	Lives with Boyfriend/Significant Other (S.O.) Siblings/Parents	Lives in PLACE with S.O. NAME and son, NAME. NAME lives at home with his mother. Lives with her husband and 3 healthy children.
8 Diet	Calcium Diet Food	Calcium intake: eats cheese and drinks a lot of milk. Age solids introduced – 4 months, table food. Balanced diet: Yes
9 Alcohol Use	Alcohol Socially Drinks	Alcohol use is < 1 alcoholic drinks per week Describes intermittent problems with alcohol in terms of excessive drinking in the past. Alcohol use is rare.
10 Tobacco Use	Cigarettes Smoking Exposure	They live in a house with no smoke exposure. Grandmother smokes outside. The patient has 20 yr hx of intermittent pipe and cigar use.

Comparative Evaluation

Table 5 summarizes the comparative evaluation performed to analyze similarities between SH topics that arose through manual analysis and topic analysis. Several topics were consistent through all age groups: Family, Living Condition, Living Situation, Living Situation Exposure, Marital Status, Occupation, Residence, and SH Other. Early Adolescence has the widest variety of social history topics included. Several topics become more prevalent with age, including Physical Activity, Diet, and Caffeine Use.

Table 5. Distribution of topics generated from manual and topic analyses, grouped by manual annotation schema.

Topic	Infant and Toddler (0-24 months)		Early Childhood (2-5 years)		Middle Childhood (6-11 years)		Early Adolescence (12-18 years)		Young Adulthood (19-44 years)		Middle Adulthood (45-64 years)		Older Adulthood (65+ years)	
Alcohol Use		✓	◆		◆	✓	◆	✓	◆	✓	◆	✓	◆	✓
Animals	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆		◆	
Caffeine Use					◆		◆		◆	✓	◆	✓	◆	✓
Diet	◆		◆	✓			◆	✓	◆	✓	◆	✓	◆	✓
Drug Use	◆	✓	◆		◆	✓	◆	✓	◆	✓	◆	✓	◆	✓
Exposure Other	◆	✓	◆	✓	◆	✓	◆	✓	◆		◆		◆	
Family	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓
Family History	◆		◆		◆	✓	◆	✓	◆	✓	◆	✓	◆	✓
Hobby			◆		◆		◆	✓	◆		◆		◆	
Hobby Exposure														
Hobby Other														
Living Condition	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓		✓

Table 5 Continued. Distribution of topics generated from manual and topic analyses, grouped by manual annotation schema.

Topic	Infant and Toddler (0-24 months)		Early Childhood (2-5 years)		Middle Childhood (6-11 years)		Early Adolescence (12-18 years)		Young Adulthood (19-44 years)		Middle Adulthood (45-64 years)		Older Adulthood (65+ years)	
Living Situation	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓
Living Situation Exposure	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓
Living Situation Other	◆	✓	◆	✓	◆	✓	◆		◆		◆		◆	
Marital Status	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓
Occupation	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓
Occupation Exposure	◆						◆						◆	
Occupation Other	◆	✓					◆						◆	
Physical Activity	◆		◆		◆	✓	◆	✓	◆	✓	◆	✓	◆	✓
Physical Activity Exposure														
Physical Activity Other														
Residence	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓
Residence Exposure			◆		◆									
Residence Other		✓	◆	✓	◆		◆		◆		◆		◆	
Social Support	◆	✓	◆		◆	✓	◆	✓	◆	✓	◆		◆	✓
Tobacco Use	◆	✓	◆	✓	◆		◆	✓	◆	✓	◆	✓	◆	✓
Travel	◆		◆		◆		◆		◆		◆		◆	
Social History Other	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓	◆	✓
# of SH Topics Represented	20	17	22	14	22	16	24	18	22	17	22	15	23	16

◆ denotes presence of topic in manual analysis; ✓ denotes presence of topic in topic analysis.

Discussion

Variations Between Topic Analysis and Manual Analysis

The findings of this study demonstrate and validate the hypothesis that social history topics change over the course of an individual's lifespan. The manual analysis conducted demonstrates the breadth of social history information that is contained within social documentation, while the topic analysis provided further granularity and depth to this information with a larger number of patients. Some of the richer pieces of information from the topic analysis includes references to languages individuals speak both primarily and in the home specifically, abuse factors, access to guns, seatbelt use, and helmet usage.

From twenty-eight potential social history types we have previously described, twenty-four were used at least once. There were no entries for Hobby Exposure, Hobby Other, Physical Activity Exposure, and Physical Activity Other. However, all other topics were observed in multiple age groups. There were minor variations in annotations between study investigators, contributing to the Cohen's kappa of 0.67. These discrepancies were related to the occupations of parents and guardians and how to annotate the presence of firearms in the home. These discrepancies have also since been resolved at the close of the annotation process. In comparison, the topic analysis contained twenty-two of these topics. No topics were generated for the four types that were not present in the manual analysis. The number of social history topics represented in the topic analysis was fairly consistent across age groups, although Early Adolescence and Young Adulthood contained the best coverage for topics.

The comparative evaluation showed large areas of overlap between the manual analysis corpus and the topic analysis corpus, despite variations in size. This overlap demonstrates the complexity of information characterized within FHS Social History Documentation, and that this analysis is representative of the large clinical population in the Fairview Health Services system, since each patient with Social History Documentation was represented. The presence of SH Other topics in both the manual analysis corpus and the topic analysis corpus demonstrates how broad SDOH for individuals can be, and how difficult it might be to create structured documentation for SDOH that adequately encompasses an individual's factors.

This work will serve as a basis for further natural language processing efforts, providing more robust tools for examining how social determinants affect individuals as they age. Providing a larger understanding of how SDOH are currently entered into the EHR will support understanding of what providers value and help to further standardization of entry, ultimately aiding in clinical decision-making.

Social Determinant Changes through Lifespan

With respect to social history documentation for pediatric individuals, there is a heavier focus on Living Situation observed prior to adulthood. This is also true for Occupation, possibly due to the focus on occupations of parents and guardians as it relates to a child's care needs, although occupation continues to be well documented through adulthood and potentially in the structured Occupation fields in the EHR elsewhere for adults. Many social history types may refer to usage of parents, guardians, or exposure. For example, Alcohol Use, Drug Use, and Tobacco Use mentions largely refer to habits of individuals a child lives with in these corpora. Hobbies show the highest occurrence for pediatric individuals in the Middle Childhood and Early Adolescence age groups. Family maintains highest frequency, followed by Occupation, and Living Situation for all age groups throughout childhood.

Family remains prevalent as a subject of documentation through adulthood, but accounts for a smaller portion of annotations than in pediatric age groups. Family is the only social history type that is seen consistently in highest frequency groups. As individuals enter Older Adulthood, mentions of Marital Status account for a larger portion of annotations. The prevalence of Family mentions may be due to the relationship family history and with whom an individual associates with has on health status and outcome.

Following SH Other, the topics that arose with highest frequency from the topic analysis concerned Residence, Living Situation Exposure, Family, and Marital Status. These topics are largely related to one another and very difficult to separate by nature of their juxtaposition within. These findings demonstrate the interdependent nature of these factors within an individual's SDOH. While Occupation topics were high in frequency for most age groups, the content of this topic was likely to change with age, and referred more directly to work experience rather than education as individuals enter adulthood.

Both corpora show a change in documentation as individuals reach adolescence and adulthood. Specifically, Caffeine, Diet, and Physical Activity show marked increases in documentation as individuals reach Young Adulthood. The largest change documented in these corpora that occurs as individuals enter adulthood is the increase in annotations for SH Other. The prevalence of SH Other topics is reflected through the topic analysis as summarized in Table 4. This category serves as a place to catch any social, behavioral, and environmental factors that are not related to the other types in Table 2. For all adult age groups, SH Other accounts for the most frequent annotations. This could possibly indicate that as we age, standard language around social history becomes harder to capture consistently and potentially more complex. Documenting the SDOH of parents and guardians might correlate to this, prove to be complex, and directly affect the SDOH of children. The complexity and variety of information included in this category will inform extensions to annotation schema for future work.

As patients within FHS reach adulthood, there is a shift in gender prevalence. This was evidenced in the demographics of the randomly selected manual analysis corpus as well as the topic analysis corpus that accessed all patients with Social History Documentation from 2011 to 2016. In pediatric sets, male patients represent a majority in both corpora. At Early Adolescence, genders are fairly comparable, leading up to a major shift at Young Adulthood. As patients enter adulthood, women represent more than 70% of patients in both corpora. This is likely due to a number of factors, but demonstrates that this analysis may be more accurate to the SDOH of adult women and pediatric males. The ratio between men and women becomes slightly more balanced as individuals enter Middle Adulthood and Older Adulthood. The demographics of this patient set also points to a lack of diversity in race. In all age groups, individuals who indicate their race as "White" predominate documentation. This is also subject to a number of factors, but could also point to this analysis more heavily representing the SDOH of those individuals.

Conclusion

Social history documentation in EHR systems will become increasingly valuable to understand for clinical care and other downstream consistent uses of this information. The content of this documentation over different age groups likely in part reflects changes of social history factors affecting individuals throughout their lifespan. The findings of this study point to the changing nature of SDOH as individuals age, and demonstrate the breadth and depth of SDOH that can affect a patient's health status. Further work to standardize how social, behavioral and environmental factors are documented within the EHR is particularly needed to ensure robust documentation of diverse SDOH topics for different age groups.

Acknowledgements

Research reported in this manuscript was supported in part by the National Library of Medicine of the National Institutes of Health grant R01LM011364 and the University of Minnesota Clinical and Translational Science

Institute of the National Institute of Health grant 8UL1TR000114. The content and opinions expressed are solely the responsibility of the authors and may not represent the official views of the National Institute of Health.

References

1. Rajamani S, Chen ES, Aldekhyyel Y, 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.
2. 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.
3. Lindemann EA, Chen ES, Rajamani S, Manohar N, Wang Y, Melton GB. Representation of Occupation Information in Clinical Texts: An Analysis of Free-Text Clinical Documentation in Multiple Sources. AMIA Jt Summits on Transl Sci Proc. 2017. *In Press.*
4. 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 Annu Symp Proc. 2014 Nov 14;2014:366-74.
5. Chen E, Garcia-Webb M. An analysis of free-text alcohol use documentation in the electronic health record: early findings and implications. Appl Clin Inform. 2014 Apr 16;5(2):402-15.
6. Wang Y, Chen ES, Pakhomov S, Arsoniadis E, Carter EW, Lindemann E, et al. Automated Extraction of Substance Use Information from Clinical Texts. AMIA Annu Symp Proc. 2015;2015:2121-30.
7. Wang Y, Chen ES, Pakhomov S, Lindemann E, Melton GB. Investigating Longitudinal Tobacco Use Information from Social History and Clinical Notes in the Electronic Health Record. AMIA Annu Symp Proc. 2016:1209-1218.
8. Winden TJ, Chen ES, Wang Y, Sarkar IN, Carter EW, Melton GB. Towards the Standardized Documentation of E-Cigarette Use in the Electronic Health Record for Population Health Surveillance and Research. AMIA Jt Summits Transl Sci Proc. 2015 Mar 25;2015:199-203.
9. Carter EW, Sarkar IN, Melton GB, Chen ES. Representation of Drug Use in Biomedical Standards, Clinical Text, and Research Measures. AMIA Annu Symp Proc. 2015;2015:376-85.
10. Winden TJ, Chen ES, Melton GB. Representing Residence, Living Situation, and Living Conditions: An Evaluation of Terminologies, Standards, Guidelines, and Measures/Surveys. AMIA Annu Symp Proc. 2016:20172-2081.
11. Youngblut JM, Brooten D, Glaze J, Promise T, Yoo C. Parent Grief 1-13 Months After Death in Neonatal and Pediatric Intensive Care Units. J Loss Trauma. 2017;22(1):77-96.
12. D'Agata AL, Sanders MR, Grasso DJ, Young EE, Cong X, McGrath JM. UNPACKING THE BURDEN OF CARE FOR INFANTS IN THE NICU. Infant Ment Health J. 2017.
13. Discenza D. Supporting Parents with Mental Health Support in the NICU. Neonatal Netw. 2016;35(1):42-4.
14. Purdy IB, Melwak MA, Smith JR, Kenner C, Chuffo-Siewert R, Ryan DJ, et al. Neonatal Nurses NICU Quality Improvement: Embracing EBP Recommendations to Provide Parent Psychosocial Support. Adv Neonatal Care. 2017;17(1):33-44.
15. Whetzel PL, Noy NF, Shah NH, Alexander PR, Nyulas C, Tudorache T, Musen MA. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011 Jul;39(Web Server issue):W541-5. Epub 2011 Jun 14.
16. Bridges N. The faces of breastfeeding support: Experiences of mothers seeking breastfeeding support online. Breastfeed Rev. 2016;24(1):11-20.
17. Pallotti P. Supporting young mothers who want to breastfeed. Pract Midwife. 2016;19(4):8, 10-2.
18. Rivera M, Sullivan R. Rethinking Child Welfare to Keep Families Safe and Together: Effective Housing-Based Supports to Reduce Child Trauma, Maltreatment Recidivism, and Re-Entry to Foster Care. Child Welfare. 2015;94(4):185-204.
19. Carvalho RA, Santos VS, Melo CM, Gurgel RQ, Oliveira CC. Inequalities in health: living conditions and infant mortality in Northeastern Brazil. Rev Saude Publica. 2015;49:5.
20. Anand KJ, Eriksson M, Boyle EM, Avila-Alvarez A, Andersen RD, Sarafidis K, et al. Assessment of Continuous Pain in Newborns admitted to NICUs in 18 European Countries. Acta Paediatr. 2017.
21. Marchuk A. End-of-life care in the neonatal intensive care unit: applying comfort theory. Int J Palliat Nurs. 2016;22(7):317-23.
22. Falck AJ, Moorthy S, Hussey-Gardner B. Perceptions of Palliative Care in the NICU. Adv Neonatal Care. 2016;16(3):191-200.

23. Molnar BE, Goerge RM, Gilsanz P, Hill A, Subramanian SV, Holton JK, et al. Neighborhood-level social processes and substantiated cases of child maltreatment. Child Abuse Negl. 2016;51:41-53.
24. Rivera M, Sullivan R. Rethinking Child Welfare to Keep Families Safe and Together: Effective Housing-Based Supports to Reduce Child Trauma, Maltreatment Recidivism, and Re-Entry to Foster Care. Child Welfare. 2015;94(4):185-204.
25. Aishworiya R, Chan PF, Kiing JS, Chong SC, Tay SK. Sleep Patterns and Dysfunctions in Children with Learning Problems. Ann Acad Med Singapore. 2016;45(11):507-12.
26. Alvarado SE. Neighborhood disadvantage and obesity across childhood and adolescence: Evidence from the NLSY children and young adults cohort (1986-2010). Soc Sci Res. 2016;57:80-98.
27. Kollerova L, Smolik F. Victimization and its associations with peer rejection and fear of victimization: Moderating effects of individual-level and classroom-level characteristics. Br J Educ Psychol. 2016;86(4):640-56.
28. Quitmann J, Rohenkohl A, Sommer R, Bullinger M, Silva N. Explaining parent-child (dis)agreement in generic and short stature-specific health-related quality of life reports: do family and social relationships matter? Health Qual Life Outcomes. 2016;14(1):150.
29. Tiwari A, Aggarwal A, Tang W, Drewnowski A. Cooking at Home: A Strategy to Comply With U.S. Dietary Guidelines at No Extra Cost. Am J Prev Med. 2017.
30. Van Amstel LL, Lafleur DL, Blake K. Raising our HEADSS: adolescent psychosocial documentation in the emergency department. Acad Emerg Med. 2004;11(6):648-55.
31. Dinleyici M, Carman KB, Ozturk E, Sahin-Dagli F. Media Use by Children, and Parents' Views on Children's Media Usage. Interact J Med Res. 2016;5(2):e18.
32. Wohn DY, Carr CT, Hayes RA. How Affective Is a "Like"?: The Effect of Paralinguistic Digital Affordances on Perceived Social Support. Cyberpsychol Behav Soc Netw. 2016;19(9):562-6.
33. Recommended Social and Behavioral Domains and Measures for Electronic Health Records, The National Academy of Medicine [cited 2017 7/3/2017]. Available from http://nationalacademies.org/HMD/Activities/PublicHealth/SocialDeterminantsEHR.aspx.
34. United States Census Briefs, US Census Bureau, Department of Commerce [cited 2017 3/7/2017]. Available from https://www.census.gov/prod/cen2010/briefs/c2010br-03.pdf.
35. 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.
36. BioMedical Information Collection and Understanding System (BioMedICUS) [Internet]. [cited 8 March 2017]. Available from: https://github.com/nlpie/biomedicus.
37. Klein D, Manning CD. Accurate Unlexicalized Parsing. Proc of the 41st Meeting of the Assoc for Comp Ling. 2003: 423-430.
38. UMLS® Reference Manual [Internet]. Bethesda (MD): National Library of Medicine (US); 2009 Sep-. 6, SPECIALIST Lexicon and Lexical Tools. [cited 7 March 2017]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9680/.
39. Topicmodels: Topic Models [Internet]. 2017 Feb 28 [cited 7 March 2017]. Available from: https://cran.r-project.org/web/packages/topicmodels/.