A Multi-Site Content Analysis of Social History Information in Clinical Notes

Elizabeth S. Chen, PhD[1,2], Sharad Manaktala, MBBS[4,6], Indra Neil Sarkar, PhD, MLIS[1,3], Genevieve B. Melton, MD, MA[4,5]
[1]Center for Clinical & Translational Science, [2]Department of Medicine, [3]Department of Microbiology & Molecular Genetics, University of Vermont, Burlington, VT
[4]Institute for Health Informatics, [5]Department of Surgery, [6]Division of Epidemiology and Community Health, University of Minnesota, Minneapolis, MN

Abstract

Within Electronic Health Records (EHRs), the social history section contains information relevant to social, behavioral, and environmental determinants of health. While social history is playing an increasingly important role in patient care, biomedical research, and public health, little analysis has been done to describe content in the EHR or the adequacy of existing standards for representing this information. In this study, social history sections from 260 clinical notes containing 989 sentences and 1,439 statements were analyzed from three sources. In total, 35 statement types were identified along with categories of information within statements for each type. For the 8 most common types, HL7 CDA and openEHR were found to provide different representations capable of capturing the breadth and granularity of information to some extent. The results of this study provide valuable insights for guiding efforts in the enhanced collection, standardization, and use of social history information in the EHR.

INTRODUCTION

Behavioral and environmental risk factors are increasingly among the leading preventable causes of morbidity and mortality in the United States[1]. These risk factors include cigarette smoking, poor diet and physical inactivity, alcohol consumption, exposure to microbial and toxic agents, high-risk sexual behavior, and illicit drug use. A great deal of evidence has been accumulated into the social, behavioral, and environmental determinants of health, which demonstrate linkages between risk behaviors, morbidity, and mortality[1-3] including the impact of specific risk behaviors on chronic diseases such as heart disease[4] and the high co-occurrence of substance use/abuse and mental health disorders such as depression[5-7]. These findings may be used for guiding and supporting preventive medicine, evaluating interventions and quality improvement initiatives, informing public health policies, and setting governmental and societal priorities at the population-level.

The increased adoption of electronic health record (EHR) systems has the potential for enhanced collection and access to a wide range of information about an individual's lifetime health status and health care[8]. In the process of recording an individual's health history, information related to behavioral and environmental risk factors and social status is traditionally documented within the "social history" section of a clinical history and physical examination and within the "social history" module of an EHR system. EHR systems could thus serve as a rich source for providing knowledge regarding risk factors including their impact, temporal progression and severity, and relationship to other health conditions[9,10]. An anticipated challenge is that social history information may be available predominantly in clinical notes in unstructured form where automated methods will be needed to facilitate the extraction and subsequent use of this information. Previous studies have focused upon information extraction (e.g., applying Natural Language Processing [NLP] technologies) to determine patient smoking status[11], tobacco cessation treatment[12,13], and family history information from clinical notes[14-16].

As reflected by the definitions in Table 1 and by reports emphasizing its importance and need for accurate and complete documentation[17], social history has a clinically significant role. The present study is thus motivated by a need to gain a better understanding of social history information in electronic clinical notes. Findings from this study may provide insights to the current quality of social history documentation, point to ways to improve the collection of comprehensive social history information in the EHR (e.g., through enhancing structured social history modules or guiding the development of detailed templates for notes), and lend guidance to the development of NLP techniques to address social history information. The objective of this study was to provide an in-depth content analysis of clinical notes from multiple sources in an effort to characterize social history information according to broad categories and then according to specific information within each of these categories. Based on this initial analysis, an assessment of existing information models was performed to determine their adequacy for representing the potential breadth and complexity of information captured within the social history section of clinical notes.

Table 1: Definitions and/or Roles of Social History.

Source	Definition
(Anderson and Schiedermayer, 2010)[17]	"The social history can provide vital early clues to the presence of disease, guide physical exam and test-ordering strategies, and facilitate the provision of cost-effective, evidence-based care."
(Clinician's Pocket Reference, 2001)[18]	"Psychosocial (Social) History: Stressors (financial, significant relationships, work or school, health) and support (family, friends, significant other, clergy); life-style risk factors, (alcohol, drugs, tobacco, and caffeine use; diet; and exposure to environmental agents; and sexual practices); patient profile (may include marital status and children; present and past employment; financial support and insurance; education; religion; hobbies; beliefs; living conditions); for veterans, include military service history; Pediatric patients: Include grade in school, sleep, and play habits."
(Clinical Clerkships: The Answer Book, 2005)[19]	"Social history that may be useful in the patient's current management (e.g., social habits such as alcohol, nicotine, or narcotics that may result in withdrawal) or useful in discharge placement."
(Continuity of Care [CCD] Quick Start Guide, 2007)[20]	"This section contains data defining the patient's occupational, personal (e.g., lifestyle), social, and environmental history and health risk factors, as well as administrative data such as marital status, race, ethnicity and religious affiliation. Social history can have significant influence on a patient's physical, psychological and emotional health and wellbeing so should be considered in the development of a complete record."
(openEHR Specifications, 2008)[21]	"*Social history/situation*: current and previous social situation (e.g., in nursing care, details of feeding, sleeping arrangements) are documented as Observations." "*Lifestyle*: there are various Observation archetypes for recording aspects of lifestyle, including exercise, smoking/tobacco, alcohol, drug use and so on."

MATERIALS AND METHODS

Social history information in clinical notes from three different sources was analyzed and the representation of this information in existing standards was evaluated (Figure 1). The overall approach involved three major phases: (1) collect and analyze social history statements from a publicly accessible resource (MTSamples.com [MTS]) to generate an initial list of statement types and models for each type (*training set*); (2) similarly analyze statements from two institutions, Fletcher Allen Health Care (FAHC) and Fairview Health Services (FHS), to evaluate and extend the list of statement types and models (*test sets*); and, (3) assess existing standards, HL7 CDA-based models and openEHR archetypes, with respect to the identified social history statement types and models.

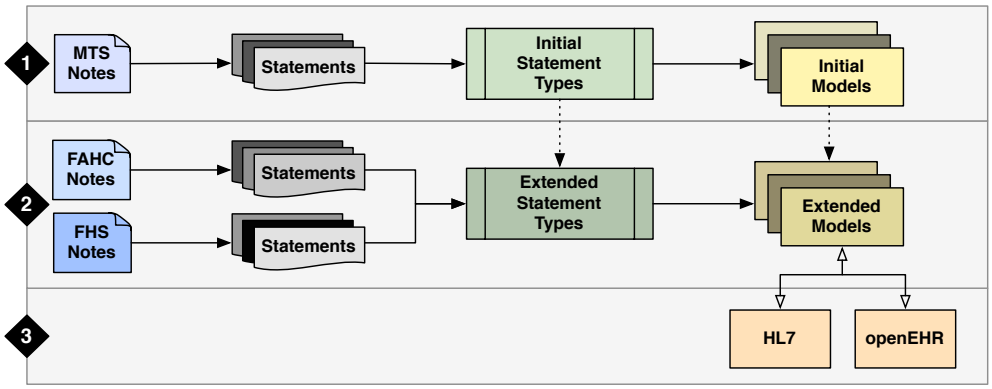

Figure 1: Overview of Materials and Methods.

1. *Analysis and Representation of Social History Information in a Public Dataset of Clinical Notes*

The first phase of the study involved collection and analysis of clinical notes from MTSamples (MTS), a public Web site containing over 4,500 sample transcription reports for different specialties and work types (e.g., Cardiovascular/Pulmonary, Surgery, Discharge Summary, and Office Notes)[22]. From over 450 reports categorized as "Consult – History and Physical", a random sample of 60 notes containing social history sections was identified (e.g., denoted by a section header of "Social History" or "Personal History"). Notes with hybrid sections (e.g., header of "Social/Family History") or containing social history information not in a social history-specific section (e.g., notes without any sections or information in the "History of Present Illness" section) were excluded.

Using an iterative, consensus-based process, the sample of 60 notes was analyzed in 3 iterations with 20 notes in each session towards creating an initial list of "statement types". For each note, the social history section was

extracted and split into sentences (e.g., split by "."), which were further divided into statements conveying discrete items of information from the original sentence. For example, the sentence "Denies alcohol and tobacco use" would be split into two statements, "Denies alcohol use" and "Denies tobacco use". Each statement was then categorized based on its contents and assigned a high-level "statement type" by two reviewers (e.g., the previous two statements could be categorized as "ALCOHOL USE" or "TOBACCO USE"). At the end of each iteration, the entire group reviewed the statements and assigned types to resolve any disagreements and guide the next iteration. Based on the resulting list of statement types, statements associated with each type were re-analyzed with respect to structure and detailed content. From this analysis, an initial set of models consisting of elements and values reflecting the varying types of information across the statements were generated.

2. ***Analysis and Representation of Social History Information in Local Datasets of Clinical Notes***

For the second phase of the study, 100 clinical notes from both Fletcher Allen Health Care (FAHC) and Fairview Health Services (FHS) were obtained and analyzed. FAHC is the tertiary care medical center affiliated with the University of Vermont[23]; transcribed inpatient and outpatient notes categorized as consults or evaluations from a legacy clinical information system were used. FHS is the regional integrated health care network affiliated with the University of Minnesota[24]; dictated and transcribed inpatient admission notes, inpatient consults, and outpatient consults were included.

Similar to the approach used for the MTS set of notes, the FAHC and FHS sets only included notes with explicit social history sections, which were extracted and split into sentences and then individual statements. Using the initial list of statement types and models from the first phase as guidelines, each statement was assigned a high-level statement type and attempts were made to map corresponding information to elements and values in the respective model by one reviewer from each institution (a formally trained informatician [ESC] and an informatics graduate student with medical training [SM]) for the respective set of statements. An additional reviewer at each institution (a formally trained informatician [INS] and a physician with informatics training [GBM]) then performed the same analysis for a subset of 10% of the statements in order to assess inter-rater reliability. To accommodate potentially new statements and information found in the FAHC and FHS sets, the list of statement types and models were extended as needed.

3. ***Assessment of Existing Information Models for Social History***

The last phase of the study was focused on assessing existing standards for representing social history information with an initial focus on HL7 and openEHR. Various implementation guides associated with the HL7 Clinical Document Architecture (CDA)[25], a specification of the syntax and semantics for clinical documents to support exchange, were reviewed with respect to social history[26,27]. These included guides and examples for the Continuity for Care Document (CCD)[20,28], History and Physical Reports[29], Plan-to-Plan Personal Health Record (P2PPHR)[30,31], Public Health Case Reports (PHCR)[31], and Healthcare Associated Infection Reports (HAIRPT)[29]. For openEHR, a review of existing archetypes that aim to provide formal models of domain concepts (e.g., blood pressure or prescriptions) was performed using the Clinical Knowledge Manager (CKM)[32]. For the most frequent statement types (identified in the first two phases of the study), the adequacy of the HL7 CDA-based models and openEHR archetypes for representing information within relevant statements across the three sources was explored.

RESULTS

In total, 260 clinical notes consisting of 989 sentences and 1,439 statements were analyzed. The 60 notes from MTS contributed 183 sentences and 298 statements, 100 notes from FAHC provided 415 sentences and 642 statements, and 100 notes from FHS included 391 sentences and 499 statements. Notes from FAHC and FHS covered a range of specialties including internal and family medicine, surgical specialties, and medical specialties.

With respect to statement types, an initial list of 27 statement types were identified based on the MTS notes and an additional 8 types were added after review of the FAHC and FHS notes (Table 2). These additional types include those created to accommodate statements related to household and other daily activities (DAILY ACTIVITY), criminal history or legal issues (LEGAL), and stress or mood (MENTAL/EMOTIONAL STATUS). As reflected in Table 2, 48.6% of the statement types were found to be common across the three sources (e.g., CAFFEINE USE, EDUCATION, and PHYSICAL ACTIVITY), 34.3% common to two sources (MTS and FAHC such as ANIMALS for statements related to pets, MTS and FHS such as ABUSE for statements related to physical/emotional abuse, or FAHC and FHS such as MILITARY SERVICE), and 17.1% unique to a source (e.g., INSURANCE for FAHC). Example statements for the 8 most frequent statement types (reflecting the combination of top statement types from

each source) are shown in Table 3 (highlighted in bold in Table 2). Inter-rater reliability between two reviewers in the assignment of statement types for a subset of 10% of statements from FAHC (n=65) and FHS (n=50) yielded κ (0.974, 0.953) and proportion agreement (96.9%, 96.0%), respectively.

Table 2: Distribution of Statement Types Across the Three Sources.

Statement Type	MTS	FAHC	FHS	Statement Type	MTS	FAHC	FHS
ABUSE	0.67%	-	1.60%	**LIVING SITUATION**	**6.06%**	**5.58%**	**8.62%**
ALCOHOL USE	**16.50%**	**13.95%**	**11.22%**	**MARITAL STATUS**	**6.40%**	**5.89%**	**11.42%**
ANIMALS	3.70%	0.78%	-	MENTAL/EMOTIONAL STATUS	-	1.40%	0.60%
CAFFEINE USE	1.01%	2.02%	1.00%	MILITARY SERVICE	-	0.31%	0.20%
DAILY ACTIVITY	-	2.02%	1.00%	**OCCUPATION**	**10.77%**	**14.26%**	**12.62%**
DIET	0.67%	1.86%	-	PHYSICAL ACTIVITY	1.68%	5.12%	0.60%
DRUG USE	**10.10%**	2.95%	2.40%	REPRODUCTIVE ACTIVITY	-	0.62%	-
EDUCATION	1.35%	1.55%	2.60%	**RESIDENCE**	3.37%	**5.89%**	**7.41%**
ENVIRONMENTAL/ OCCUPATIONAL EXPOSURE	0.67%	1.40%	-	SAFETY/ PREVENTATIVE CARE	-	0.16%	-
ETHNICITY	0.34%	-	0.40%	SEXUAL ACTIVITY	0.34%	0.78%	0.20%
FAMILY	**8.75%**	**7.13%**	**14.22%**	SEXUAL ORIENTATION	0.34%	-	-
FUNCTIONAL STATUS	2.36%	1.40%	1.60%	SICK CONTACT	1.01%	-	-
GENDER IDENTITY	-	-	0.40%	SOCIAL SUPPORT	1.01%	2.95%	2.20%
HEALTH STATUS	0.34%	0.16%	0.20%	**TOBACCO USE**	**17.51%**	**15.66%**	**12.42%**
HOBBY	0.67%	2.33%	-	TRAVEL	1.35%	0.93%	0.40%
INFECTIOUS DISEASE	0.67%	0.31%	-	WEIGHT MANAGEMENT	0.67%	0.47%	-
INSURANCE	-	0.31%	-	OTHER	1.68%	1.71%	5.21%
LEGAL	-	0.16%	0.60%				

Table 3: Example Statement Types.

ALCOHOL USE	**MARITAL STATUS**
• The patient does not take any drinks • He has not had any alcohol in the last year • She takes one glass of wine per day • She occasionally uses alcohol • He drinks socially about three beers per week • He has history of alcohol dependence	• The patient is single • The patient is currently widowed • She is married to her husband for the last four years • She is in the process of getting divorced • Recently remarried • This is the second marriage for both of them
DRUG USE	**OCCUPATION**
• She has a history of cocaine use five years ago • He has very rare marijuana use • He has a history of heroin addiction • He denies any recent usage of polysubstances • She denies any history of intravenous drug abuse • She has a history of narcotic abuse	• She is a housewife • The patient is currently in school • Part-time farmer • She works as a nurse in a newborn nursery • The patient is a retired sanitation engineer • He is unemployed
FAMILY	**RESIDENCE**
• He has two daughters who live in the area • One son deceased • Wife is living and well • Her husband is disabled • Mother worked part time and is not planning to return to work • Plans to adopt two children	• He lives in a dorm there • He lives in New Jersey • They live in a multilevel home. • He has recently moved from Florida • She currently lives in a co-op apartment • She is a native of Texas
LIVING SITUATION	**TOBACCO USE**
• He lives alone • She lives with her parents. • The patient lives with his wife • He lives with daily nursing aids • She has recently moved in with a family member • She lives with several roommates	• She smokes 5 cigarettes per day, has done so for 10 years • Denied tobacco use • He used to smoke pipe until about 17-18 years ago • The patient quit smoking cigars in 2004 • Former smoker, smoked two packs per week for approximately five to six years • She does not use any significant tobacco

Table 4: Elements and Values for Statement Types and Distribution of Elements for FAHC and FHS.

Statement Type	Element	Example Values or *Patterns*	FAHC	FHS
ALCOHOL USE	Status	current, past, past (quit), denies, never/no history, no/negative	92.2%	82.1%
	Temporal	*<#> <timeunit> [duration/ago], [in/since/until] <date>*	6.7%	5.3%
	Method	oral	53.3%	30.3%
	Type	beer, wine, hard liquor	10.0%	1.7%
	Amount	variable, unknown, *<#> [glasses/drinks/bottles/times], <#>-<#> cans*	18.9%	14.2%
	Frequency	daily, weekly, monthly, yearly, socially, occasionally, rarely	46.7%	26.7%
DRUG USE	Status	current, past, past (quit), denies, never/no history, no/negative	94.7%	83.3%
	Temporal	*<#> <timeunit> [duration/ago], [in/since/until] <date>*	5.3%	-
	Method	smoke, snort/intranasal, oral, intravenous, inject	21.1%	8.3%
	Type	marijuana, cocaine, heroin, benzodiazepines, narcotics	84.2%	16.6%
	Amount	*<#> times*	15.8%	-
	Frequency	occasionally, rarely, very rarely, regularly	10.5%	-
FAMILY	Status	alive and well, alive and ill, deceased, disabled, unknown	15.2%	14.0%
	Temporal	*last <time period>*	2.2%	2.8%
	Type	children, spouse, daughter, son, mother, father, extended family	95.7%	94.3%
	Amount	*<#>, <#> <age>, <#> younger*, grown/adult	56.5%	64.7%
	Other	*<location>, <medical conditions>, <occupation>, <residence>*	32.6%	5.6%
LIVING SITUATION	Status	current, past, uncertain	97.2%	100.0%
	Temporal	*<#> <timeunit> [duration/ago], [in/since/until] <date>*	-	11.6%
	Method	lives alone, lives with	92.2%	97.7%
	Type	children, spouse, husband, wife, mother, father, significant other	75.0%	79.1%
MARITAL STATUS	Status	current, past, uncertain, denies, in process	100.0%	96.4%
	Temporal	*<#> <timeunit> [duration/ago], [in/since/until] <date>*	7.9%	12.2%
	Type	married, divorced, widowed, single, separated, engaged, in relationship	100.0%	98.2%
	Amount	*<#> times*	-	10.5%
OCCUPATION	Status	current, past, future, denies, unknown	93.1%	96.8%
	Temporal	recently, [*in/since/until] <date>, <#>-<#> <timeunit> [duration/ago]*	5.9%	9.5%
	Method	self-employed, employed, unemployed, retired, fired, quit, medical leave	73.9%	44.4%
	Type	homemaker, student *<field>, <job/position>*	67.3%	30.1%
	Level	*<#> [grade/school years]*	-	15.8%
	Location	home, *<school/institution>, <job/employer/organization>, <location>*	34.7%	33.3%
	Extent	part-time, full-time, night shift, day shift, weekends, summer	5.9%	3.1%
RESIDENCE	Status	current, past, future	94.7%	89.1%
	Temporal	*<#> <timeunit> [duration/ago], [in/since/until] <date>*	7.9%	16.2%
	Method	grew up, born in, originally from	10.5%	21.6%
	Type	house, apartment, dorm, assisted living facility, multi-level home	13.2%	35.1%
	Location	local, *<location>*	97.4%	67.5%
TOBACCO USE	Status	current, past, past (quit), denies, never/no history, no/negative	95.0%	88.7%
	Temporal	*since <date/time period> until <date/time period>*	28.7%	20.9%
	Method	smoke, snort, oral	80.2%	69.3%
	Type	cigarettes, cigar, pipe, chewing tobacco, snuff	12.8%	8.0%
	Amount	minimal, significant, *[<|>] <#> [cigarettes/packs/pack-years/times]*	28.7%	17.7%
	Frequency	daily, weekly, monthly, yearly, socially, occasionally, rarely	17.8%	6.4%

Based on statements from the MTS notes, initial models for each of the 27 statement types were created that included a set of data elements and values. While variation in information was found across the statements types, an attempt was made to determine if a common set of data elements could be defined. In testing the models with statements from the FAHC and FHS sets, the data elements were found to be sufficient and only extensions to the value sets were needed. For the 8 additional statement types, models could be created using the same data elements and the value sets were populated by information found within the respective statements. Inter-rater reliability between two reviewers in mapping statements to the respective models for 10% of the statements showed proportion agreement of 93.8% for FAHC and 88.0% for FHS. Table 4 depicts the data elements and example values (or patterns) associated with the most frequent statement types along with the distribution of coverage within the FAHC and FHS statements. For a given statement, not all elements may be represented and rules may be needed to specify valid combinations of elements and values (e.g., for MARITAL STATUS, the statement "The patient is single" could be represented by *status*="current" and *type*="single" and for OCCUPATION, a value for *level* would apply if *status*="current" and *type*="student"). As reflected by the percentages, a value for "status" is frequently explicitly

stated or implied in statements while "temporal" information is less frequently stated across the majority of statement types. Between the two institutions, results could be viewed as comparable for elements across statement types. In the FAHC set, the "type" element consistently has a higher percentage (except for RESIDENCE); in the FHS set, the "temporal" element has a higher percentage for more than half of the statement types. In comparing statements related to substance use, DRUG USE statements appear to include less information related to amount and frequency than ALCOHOL USE and TOBACCO USE for both institutions.

Based on the review of available implementation guides and searching in CKM for the most frequent statement types, Table 5 summarizes relevant HL7 CDA-based models and openEHR archetypes that could be applied or adapted. For most of the types, the "social history observation" is specified (that is based on the HL7 Clinical Statement Model) and is described as covering marital status, ethnicity, smoking, exercise, diet, employment, toxic exposure, alcohol use, drug use, and other social history. Other specifications include the "occupation observation" for OCCUPATION (defined in the PHCR implementation guide), models focused on family history that could potentially be adapted for FAMILY statements, and the use of maritalStatusCode for representing MARITAL STATUS as administrative information in the header of a CDA document. In openEHR, several archetypes related to substance use were found for providing a summary or overview about use (e.g., Alcohol Use Summary archetype) as well as specific use of a substance at a given time (e.g., Alcohol Consumption archetype). Other archetypes related to personal or professional demographics could potentially be used to cover some aspects of information in the FAMILY, MARITAL STATUS, and OCCUPATION statements. In both standards, existing representations for "address" could be further explored for representing RESIDENCE statements.

Table 5: Alignment with HL7 and openEHR.

Statement Type	HL7 CDA-based Models	openEHR Archetype
ALCOHOL USE	Social history observation	Alcohol Use Summary, Alcohol Consumption
DRUG USE	Social history observation	Substance Use Summary, Substance Use
FAMILY	Family history observation, Clinical Genomics Family History Model	Individual's Personal Demographics, Person, Risk of condition based on family history
LIVING SITUATION	Functional status (Problem observation, Result observation)	*No archetype found*
MARITAL STATUS	maritalStatusCode, Social history observation	Extended Personal Demographics
OCCUPATION	Social history observation, Occupation observation	Professional Individual demographics
RESIDENCE	Addr	Address
TOBACCO USE	Social history observation	Tobacco Use Summary, Tobacco Use

In taking a closer look at representing information within social history statements, Figure 2 depicts the use of specific HL7 CDA-based models and openEHR archetypes for the following TOBACCO USE and OCCUPATION statements:

"She smokes 5 cigarettes per day, has done so for 10 years"

```
Status: current
Temporal: 10 years (duration)
Type: cigarettes
Amount: 5 cigarettes
Frequency: daily
```

"She works as a nurse in a newborn nursery"

```
Status: current
Type: nurse
Location: newborn nursery
```

While offering different representations, both HL7 and openEHR are able to represent a majority of the information provided in these two example statements with some information loss. One potential limitation may be found with respect to temporal information, which can take several forms in statements including specific time points (e.g., "since 1990" or "quit in 2004"), durations (e.g., "30 years ago" or ">=10 years"), and vague descriptions or estimates (e.g., "recently", "many years", or "2-3 months"). As reflected in Figure 2A-B for the TOBACCO USE statement, specific time points are expected (e.g., effectiveTime for HL7 and Date/Age commenced or ceased in openEHR) requiring calculations for cases where duration is provided (e.g., "2001" is specified to address "for 10 years" in the example based on the current year of 2011) or limiting the ability to capture specific temporal information. One notable difference between the models is the combination of several types of information within individual elements in the HL7 model (e.g., amount and frequency in "value") compared with the separation of these information in openEHR; depending on the use case, one representation may be preferred over the other. As demonstrated in Figure 2D, the HL7 occupation statement, that specifies the use of the Standard Occupational Classification (SOC)[33] for occupational categories and North American Industry Classification System (NAICS)[34]

for industry types, is able to represent the example OCCUPATION statement but loses some granularity about working in the "newborn nursery" due to the use of the NAICS code for "General Medical and Surgical Hospitals". This finding highlights potential limitations due to the coverage of existing code sets and terminologies that will be essential to address (which is out of scope for the current study and further described in the discussion).

<table>
<tr><th colspan="3">Statement: "She smokes 5 cigarettes per day, has done so for 10 years"</th></tr>
<tr><th>HL7
"social history observation"</th><th>openEHR
"Tobacco Use Summary"</th><th>openEHR
"Tobacco Use"</th></tr>
<tr>
<td><observation classCode="OBS"
moodCode="EVN">
...
<code code="230056004"
codeSystem="2.16.840.1.113883.6.96"
displayName="Cigarette consumption"/>
<statusCode code="completed"/>
<effectiveTime>
<low value="2001"/>
<high value=""/>
</effectiveTime>
<value xsi:type="ST">5 cigarettes per day</value>
</observation>
(A)</td>
<td>Substance: Tobacco
Usage Status: Current User
Consumption Summary
Form: Cigarettes - manufactured
Method of use:
Date commenced: 2001
Age commenced:
Date ceased:
Age ceased:
Comment:
Cessation attempts:
(B)</td>
<td>Substance: Tobacco
Consumption details
Form: Cigarettes - manufactured
Method of use:
Frequency:
Amount:
Number smoked: 5/d
Grams of tobacco:
Triggers:
Readiness for change:
Evidence of dependence:
(C)</td>
</tr>
<tr><th colspan="3">Statement: "She works as a nurse in a newborn nursery"</th></tr>
<tr><th colspan="2">HL7
"occupation and industry type observation"</th><th>openEHR
"Professional Individual Demographics"</th></tr>
<tr>
<td colspan="2"><!-- Occupation observation -->
<observation classCode="OBS" moodCode="EVN">
...
<code code="11341-5" codeSystem="2.16.840.1.113883.6.1"
displayName="History of occupation" />
<text>She works as a nurse in a newborn nursery.</text>
<statusCode code="completed"/>
<value xsi:type="CD" code="29-1141"
codeSystem="2.16.840.1.113883.6.96" displayName="Nurses, Registered"
/>
…
<entryRelationship typeCode="REFR">
<!-- Industry type observation -->
<observation classCode="OBS" moodCode="EVN">
<code code="21844-6" codeSystem="2.16.840.1.113883.6.1"
displayName="Industry Hx" />
<statusCode code="completed" />
<value xsi:type="CD" code="622110"
codeSystem="2.16.840.1.113883.6.85" displayName="General Medical and Surgical Hospitals" />
</observation>
</entryRelationship>
</observation>
(D)</td>
<td>Name
…
Professional details
Professional Role
Unstructured role: nurse
Structured role
Period of involvement
Grade
Specialty
Team
Professional Identifier
Telecoms
Address
…
Organisation
Name of Organisation: newborn nursery
…
(E)</td>
</tr>
</table>

Figure 2: Representation of Statements using HL7 and openEHR. Example TOBACCO USE statement as an HL7 "social history observation" (A), using the "Tobacco Use Summary" openEHR archetype (B), and using the "Tobacco Use" archetype (C). Example OCCUPATION statement represented as an "occupation and industry type observation" as defined in the HL7 CDA for PHCR (Public Health Case Reports) implementation guide (D) and using the "Professional Individual Demographics" archetype from openEHR (E).

DISCUSSION

The ability to extract, encode, and structure social history information from clinical notes in the EHR using automated methods could assist in developing a more complete picture of individual health and health at a population level. Using potentially complementary sources such as EHR system content to study and monitor risk factors could enhance existing efforts in supporting optimal patient management, biomedical research, and public health initiatives. This study represents a first step towards achieving these goals through in-depth examination of existing documentation of social history information in clinical notes and current standards for representing this information for a range of subsequent uses. Collectively, the results of this study provide insights into the current state of collection and representation of social history in the EHR that may help guide future efforts for enhanced use of this information.

Overall, analyzing clinical notes from three different sources provided broad and complementary coverage of statement types and information. Next steps include expanding the analysis to additional notes, examining social history information in the entire note rather than in a specific social history section (e.g., in a hybrid "Social History/Family History" section or in notes without section headers), and comparing content and structure of information across institutions (e.g., between FAHC and FHS) and specialties (e.g., internal medicine, pediatrics, and psychiatry).

Based on the review of over 250 notes, 35 different statement types were identified through an iterative consensus-based process. With each iteration, numerous questions arose resulting in discussions regarding the coverage of specific statement types, whether certain types should be added, removed, or combined, and how to represent embedded information within statements. One point of discussion was with respect to students and categorizing past education compared with current education where a decision was made to assign EDUCATION and OCCUPATION respectively. In some cases, similar information is found in different statement types and should be represented and accessible in a way that can accommodate a range of use cases (e.g., through linkages or relationships). For example, while the focus is on describing family members in FAMILY statements (e.g., "He has no children"), this information may be included and considered supplementary in LIVING SITUATION and SOCIAL SUPPORT statements (e.g., "He lives with parents" and "accompanied by mother" respectively). Further work is needed to formalize definitions of the various statement types and create more comprehensive models that can capture the wealth of contextual information found within statements.

For each statement type, a set of common models was created based on information contained within almost 1,500 statements from existing clinical notes from MTS, FAHC, and FHS. A subset of these models representing the combined top statement types across the three sources (ALCOHOL USE, DRUG USE, FAMILY, LIVING SITUATION, MARITAL STATUS, OCCUPATION, RESIDENCE, and TOBACCO USE) were used to perform an initial assessment of existing standards (HL7 CDA-based models and openEHR archetypes) for representing social history information. Overall, these standards were found to provide different representations that are capable of capturing much of the information contained within the various statements to some extent. Aside from those 8 statement types, HL7 also provides specifications for representing FUNCTIONAL STATUS in CCD and openEHR includes archetypes for "Caffeine Consumption" and "Cessation attempts". Additional work is needed to further evaluate and harmonize the models developed as part of this study with HL7, openEHR, and other information modeling standards (e.g., following a process described in other studies[35-37]). Another important next step will be alignment with biomedical terminologies, which will involve assessing the adequacy of existing code sets and terminologies for social history information and addressing issues such as pre-coordination and post-coordination. For example, LOINC, SNOMED CT, and MEDCIN include a range of concepts related to different aspects of social history from substance use to living situation.

The long-term goals of this effort are to work towards improving the capture and standardization of detailed social history information in the EHR as well as adapting and developing automated NLP methods for extracting this information from clinical notes both retrospectively and prospectively. Recent studies have described the creation of formal annotation schema for guiding the development and evaluation of NLP systems (e.g., for clinical conditions in emergency department reports[38] and information for Inflammatory Bowel Disease in notes[39]). The insights gained from the present study regarding the content and structure of various types of social history information and relationship to existing standards will be valuable for guiding efforts to create a comprehensive annotation schema for social history information. In addition, potential implications for EHR system development and enhancement include expanding structured social history modules or designing clinical note templates to capture the breadth and granularity of information as documented within existing clinical narratives.

CONCLUSION

The goal of this study was to gain a better understanding of the content of social history information in clinical notes and representation of this information in existing standards. Through analysis of social history sections in notes from three different sources, numerous types of statements were identified. Further analysis of these statements revealed a variety of information that can be captured by current standards to some extent. The findings provide guidance for enhanced collection and representation of social history information in the electronic health record.

Acknowledgments

Medical transcription reports were obtained with permission from MTSamples (http://www.mtsamples.com). The authors thank Fletcher Allen Health Care and Fairview Health Services for their support of this study. In particular, the authors thank Rhonda Kost and Charles Roger for their valuable assistance in obtaining clinical notes for FAHC and FHS respectively.

References

1. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. JAMA. 2004 Mar 10;291(10):1238-45.
2. Danaei G, Ding EL, Mozaffarian D, Taylor B, Rehm J, Murray CJ, et al. The preventable causes of death in the United States: comparative risk assessment of dietary, lifestyle, and metabolic risk factors. PLoS Med. 2009 Apr 28;6(4):e1000058.
3. Babor TF, Sciamanna CN, Pronk NP. Assessing multiple risk behaviors in primary care. Screening issues and related concepts. Am J Prev Med. 2004 Aug;27(2 Suppl):42-53.
4. Zaret BL, Cohen LS, Moser M, Yale University. School of Medicine. Yale University School of Medicine heart book. New York: William Morrow and Co.; 1992.
5. Jane-Llopis E, Matytsina I. Mental health and alcohol, drugs and tobacco: a review of the comorbidity between mental disorders and the use of alcohol, tobacco and illicit drugs. Drug Alcohol Rev. 2006 Nov;25(6):515-36.
6. Brook JS, Cohen P, Brook DW. Longitudinal study of co-occurring psychiatric disorders and substance use. J Am Acad Child Adolesc Psychiatry. 1998 Mar;37(3):322-30.
7. Huang FY, Ziedonis DM, Hu HM, Kline A. Using information technology to evaluate the detection of co-occurring substance use disorders amongst patients in a state mental health system: implications for co-occurring disorder state initiatives. Community Ment Health J. 2008 Feb;44(1):11-27.
8. Institute of Medicine (U.S.). Committee on Improving the Patient Record., Dick RS, Steen EB, Detmer DE. The computer-based patient record : an essential technology for health care. Rev. ed. Washington, D.C.: National Academy Press; 1997.
9. Kukafka R, Ancker JS, Chan C, Chelico J, Khan S, Mortoti S, et al. Redesigning electronic health record systems to support public health. J Biomed Inform. 2007 Aug;40(4):398-409.
10. Stewart WF, Shah NR, Selna MJ, Paulus RA, Walker JM. Bridging the inferential gap: the electronic health record and clinical evidence. Health Aff (Millwood). 2007 Mar-Apr;26(2):w181-91.
11. Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):14-24.
12. Hazlehurst B, Frost HR, Sittig DF, Stevens VJ. MediClass: A system for detecting and classifying encounter-based clinical events in any electronic medical record. J Am Med Inform Assoc. 2005 Sep-Oct;12(5):517-29.
13. Hazlehurst B, Sittig DF, Stevens VJ, Smith KS, Hollis JF, Vogt TM, et al. Natural language processing in the electronic medical record: assessing clinician adherence to tobacco treatment guidelines. Am J Prev Med. 2005 Dec;29(5):434-9.
14. Goryachev S, Kim H, Zeng-Treitler Q. Identification and extraction of family history information from clinical reports. AMIA Annu Symp Proc. 2008:247-51.
15. Friedlin J, McDonald CJ. Using a natural language processing system to extract and code family history data from admission reports. AMIA Annu Symp Proc. 2006:925.
16. Melton GB, Raman N, Chen ES, Sarkar IN, Pakhomov S, Madoff RD. Evaluation of family history information within clinical documents and adequacy of HL7 clinical statement and clinical genomics family history models for its representation: a case report. J Am Med Inform Assoc. 2010 May 1;17(3):337-40.
17. Anderson RA, Schiedermayer D. The social history matters! Acad Med. 2010 Jul;85(7):1103.
18. Gomella LG, Haist SA. Clinician's Pocket Reference. 9th ed: McGraw-Hill Professional Publishing; 2001.

19. Weiss JG. The Answer Book: Saint-Frances Guide to the Clinical Clerkships. 1st ed: Lippincott Williams & Wilkins; 2005.
20. http://www.himssehra.org/ASP/tools.asp.
21. http://www.openehr.org/svn/specification/TAGS/Release-1.0.2/publishing/architecture/rm/ehr_im.pdf.
22. http://www.mtsamples.com/.
23. http://www.fletcherallen.org/.
24. http://www.fairview.org/.
25. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, et al. HL7 Clinical Document Architecture, Release 2. Journal of the American Medical Informatics Association : JAMIA. 2006 Jan-Feb;13(1):30-9.
26. http://wiki.hl7.org/index.php?title=Structured_Documents.
27. http://wiki.hl7.org/index.php?title=Product_CDA_R2_IG.
28. ftp://ftp.ihe.net/TF_Implementation_Material/PCC/schemas/ccd/CCD-final.pdf.
29. http://www.hl7.org/dstucomments/index.cfm.
30. http://wiki.hl7.org/index.php?title=Plan-to-Plan_Personal_Health_Record.
31. http://www.hl7.org/Special/committees/structure/index.cfm.
32. http://www.openehr.org/knowledge/.
33. http://www.bls.gov/soc/.
34. http://www.census.gov/eos/www/naics/.
35. Bakken S, Warren JJ, Casey A, Konicek D, Lundberg C, Pooke M. Information model and terminology model issues related to goals. Proc AMIA Symp. 2002:17-21.
36. van der Kooij J, Goossen WT, Goossen-Baremans AT, Plaisier N. Evaluation of documents that integrate knowledge, terminology and information models. Stud Health Technol Inform. 2006;122:519-22.
37. Chen ES, Zhou L, Kashyap V, Schaeffer M, Dykes PC, Goldberg HS. Early experiences in evolving an enterprise-wide information model for laboratory and clinical observations. AMIA Annu Symp Proc. 2008:106-10.
38. Chapman WW, Dowling JN. Inductive creation of an annotation schema for manually indexing clinical conditions from emergency department reports. J Biomed Inform. 2006 Apr;39(2):196-208.
39. South BR, Shen S, Jones M, Garvin J, Samore MH, Chapman WW, et al. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinformatics. 2009;10 Suppl 9:S12.