Representation of Drug Use in Biomedical Standards, Clinical Text, and Research Measures

Elizabeth W. Carter, MS[1], Indra Neil Sarkar, PhD, MLIS[1,3], Genevieve B. Melton, MD, PhD[4,5], Elizabeth S. Chen, PhD[1,2]

[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, University of Minnesota, Minneapolis, MN

Abstract

Drug misuse is a prominent cause of morbidity and mortality in the United States. Recent focus on behavioral and social domains in the electronic health record (EHR) has highlighted the need for comprehensive examination of social history information, such as drug use. In this study, representation of drug use was examined in three types of sources: (1) standards from HL7 and openEHR, (2) clinical text from publicly accessible clinical notes and a local EHR, and (3) research measures from the PhenX Toolkit and CDE Browser. In total, 27 elements were identified across the examined sources, revealing a diverse set of values that were found to be associated with drug use type, frequency, method, time frame, and amount. The findings of this study provide insight into the representation of drug use information that may contribute to efforts for standardizing collection and use of these data to support clinical care and research.

Introduction

Morbidity and mortality from pharmaceutical, over-the-counter, and illicit drug misuse in the United States (U.S.) has increased in recent years, while drug poisoning has become a leading cause of injury deaths[1, 2]. Opioid analgesic poisoning deaths tripled between 2000 and 2010[3]. Opioids and benzodiazapenes were the predominant pharmaceutical drugs leading to over 43,000 overdose deaths in 2013[4]. In 2013, 8.8% of adolescents and 9.4% of adults in the U.S. were current (within the past month) users of illicit drugs[5]. Further, drug misuse has been well documented among youths[6] and adults[7, 8] to be associated with comorbid medical problems and mental health disorders. The consequences of such comorbidities for youths can be substantial – including higher rates of treatment, social and academic problems, and suicide attempts[6, 7]. In the general population, co-occurrence of mental health disorders with drug misuse has been documented for anxiety disorders (with marijuana)[9, 10] and suicide deaths (with misuse of prescription drugs) in U.S. veterans[11]. These findings reflect part of the impact of drug use in the U.S. and illuminate the need for standardized representation of drug use in behavioral data that may be used to inform patient care, clinical research, and public health policy.

In recent years, focus on standardized collection of patient information using the electronic health record (EHR) has intensified. The 2009 Health Information Technology for Economic and Clinical Health Act (HITECH) outlined goals for the adoption of EHRs for health providers, in parallel with the objective of "meaningful use" to improve patient care[12]. The Patient Affordable Care Act of 2010 further emphasized the detection of substance use problems and the integration of care with primary care providers[13]. The 2013 set of Meaningful Use Objectives from the Centers for Medicare and Medicaid Services contains one core objective for tobacco use (smoking status); however, there were no detailed objectives for other types of drug use[14]. A collaboration of the National Institute on Drug Abuse Center for Clinical Trials Network (NIDA-CCTN), National Cancer Institute (NCI), and Substance Abuse and Mental Health Services Administration (SAMSHA) supports the development of a standardized clinical quality measure (CQM) that may be considered for inclusion as a Meaningful Use core objective[15]. In 2014, the Institute of Medicine proposed a new set of standard measures for social and behavioral domains that included evaluation of the 'Abuse of Other Substances' domain[16, 17]. However, due to the sensitivity and complexity of data collection for drug misuse, this domain was not included for further consideration for Meaningful Use Stage 3. Existing standards for collection of social history information (e.g., the HL7 implementation guides for Clinical Document Architecture (CDA) Release 2[18] and the 'Substance Use Summary' and 'Substance Use' archetypes in openEHR[19]) that may provide a possible way to represent drug use in EHRs have been assessed in prior work for capturing social history information in clinical notes[20].

The EHR has the potential to serve as a powerful tool to support drug use screening, diagnosis, intervention, and treatment for primary care providers[21]. Wu, *et al.* examined the prevalence of substance use disorders (SUDs) and

patterns of comorbidities in adults by extracting patient information from the 'Habit/Substance Use' domain in the EHR at Duke University Medical Center. Prevalence of SUDs differed by sex, race and ethnicity, whereas comorbidities differed by race and sex, and were more prevalent among patients with SUDs than those without. This study provides supporting evidence for the use of EHRs for research to inform health care. Yet, further standardization of patient data is required to support information sharing among providers. Motivated by the 2006 National Institutes of Health (NIH) National Electronics Clinical Trials and Research projects, NIDA-CCTN led an effort to develop sets of consensus-based common data elements (CDEs) for substance use to be used within the EHR and other data sources[15, 22, 23]. The NIH Common Data Element (CDE) Resource Portal[24] promotes the use of CDEs through numerous initiatives, tools, and resources. Among these are the Cancer Data Standards Registry and Repository (caDSR)[25] that incorporates cancer-specific CDEs from numerous sources; Grid-enabled Measures (GEM)[26] that facilitates the development and sharing of common measures; and, Consensus Measures for Phenotypes and eXposures (PhenX)[27] that provides consensus-based standard measures to be incorporated in genome-wide association and epidemiologic research studies. These sources contain CDEs (a single question denoting a fixed representation of a variable)[28] or measures (a standardized instrument containing a protocol with one or more CDEs)[29], which will henceforth be referred to as "research measures."

With the increased attention to the importance of monitoring and understanding drug use and its impact, there is a need for research focused on comprehensive collection and subsequent use of this information. To address this need, the goal of this study was to examine the representation of drug use in multiple sources for informing EHR and standards development for guiding evidence-based patient care and ultimately improving patient outcomes.

Methods

Study design

Three types of sources were analyzed to identify elements associated with drug use and their corresponding values: (1) *standards* – HL7 CDA-based models[30] and openEHR archetypes[19]; (2) *clinical text* – drug use sentences in clinical notes from the publicly accessible MTSamples.com (MTS)[31] resource and drug use comments from the social history module of the Epic EHR[32] at the University of Vermont Medical Center (UVMMC)[33]; and (3) *research measures* – measures including one or more CDEs from the PhenX Toolkit and CDEs from caDSR[34] using the CDE Browser.

Analysis of standards

The first phase involved exploring current standards for collection of social history information related to drug use. The HL7 Implementation Guide for CDA Release 2: IHE Health Story Consolidation, DSTU Release 1.1 was examined for elements within the Social History Observation template of the Social History section[30]. Archetypes for 'Substance Use Summary' and 'Substance Use' were also identified using the openEHR Clinical Knowledge Manager[35]. Elements described within these standards were merged with a set of elements identified in previous work involving the analysis of social history information in clinical notes, social and behavioral information in public health surveys, and free-text comments associated with tobacco and alcohol use in the EHR[20, 36, 37]. This combined set of 25 elements was used to create an initial set of annotation guidelines for analyzing the clinical text and research measures.

Analysis of clinical text

The second phase consisted of an analysis of clinical texts from MTS and UVMMC. For MTS, sentences describing drug use were identified within 491 clinical notes categorized as 'Consult – History and Physical', which have been used in prior studies[38, 39], using the General Architecture for Text Engineering (GATE)[40]. This provided a set of 130 drug use sentences from 124 (25.3%) notes. At UVMMC, the Epic EHR social history module collects information on drug use using a set of structured fields for Status ('Not Asked,' 'Yes,' and 'No'), Use/week, and Type (e.g., 'benzodiazapines,' 'cocaine,' 'marijuana,' 'heroin,' and 'opioids') as well as a free-text comment field, which was the focus of this study. A random set of 50 drug use comments from January 2014 was first analyzed to determine if and how to enhance the annotation guidelines. For example, the original element *Duration* was found to be too broad and led to the creation of more specific elements *Duration of Use* and *Duration Since Time Point*. Two annotators (EWC and ESC) annotated a subset of 10% of the 130 MTS sentences and 450 UVMMC comments from March 2014 (representing the most recent comments for a random set of 450 patients), achieving an inter-rater reliability using Cohen's kappa of 0.92 and 0.94 respectively.

The brat rapid annotation tool (BRAT)[41, 42] was then used to annotate drug use information in the 130 sentences and 450 comments (Figure 1) using the revised annotation guidelines. These guidelines define the elements (e.g., *Status*, *Type*, or *Time Frame*) as well as relationships between elements such as *Type* and *Time Frame* (e.g., '40 years ago' => 'heroin'). Further analysis involved extracting the annotations (representing element-value pairs) and generating statistics for each element and their corresponding values for each source of clinical text as well as the combined set of annotations for MTS and UVMMC. Using the combined annotations, those with similar meaning were grouped by a common word, phrase, or pattern. For example, the grouping 'Cocaine' represented the annotations 'cocaine,' 'cocaine drug,' 'crack,' and 'crack cocaine,' whereas patterned groupings such as '[#] [time unit] ago' represented 'days ago' and 'over a year ago.'

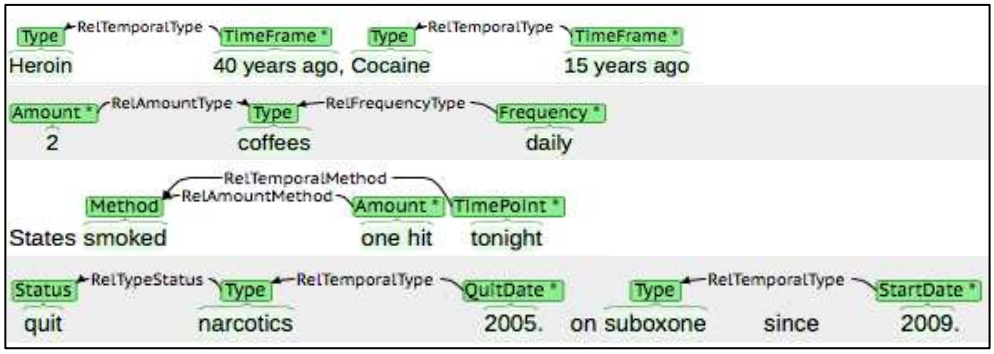

Figure 1. Annotation of clinical text using the brat rapid annotation tool.

Analysis of research measures

The third phase of the study involved analysis of the PhenX Toolkit (V5.7; accessed July 31, 2014) containing 21 domains with over 330 measures and the caDSR[43] using the CDE Browser (V4.0.4; accessed July 31, 2014) providing access to 35 resources containing information about thousands of CDEs. A search for measures using the PhenX Toolkit was done by browsing the domains with a focus on the 'Alcohol, Tobacco, and Other Substances' domain, in addition to a search of all measures by using the following search terms: 'substance,' 'illicit,' 'illegal,' and 'drug.' The same set of search terms was used to identify measures in the CDE Browser. Cumulatively, an initial set of 111 measures was identified and further restricted by excluding measures that explicitly targeted substance abuse and cessation; emotional, physical or social repercussions of drug use or abuse; or, opinions and perspectives about use or abuse. This resulted in a final set of 40 drug use-specific measures containing 40 questions from CDE Browser and 10 drug use-specific measures containing 191 questions from PhenX Toolkit. For some measures, instructions and supplemental information were also included in the analysis. For example, the measure 'Patterns of substance use – adults' in the PhenX Toolkit included a supplemental drug card detailing over 150 drug names and each was designated as a value within the corresponding measure.

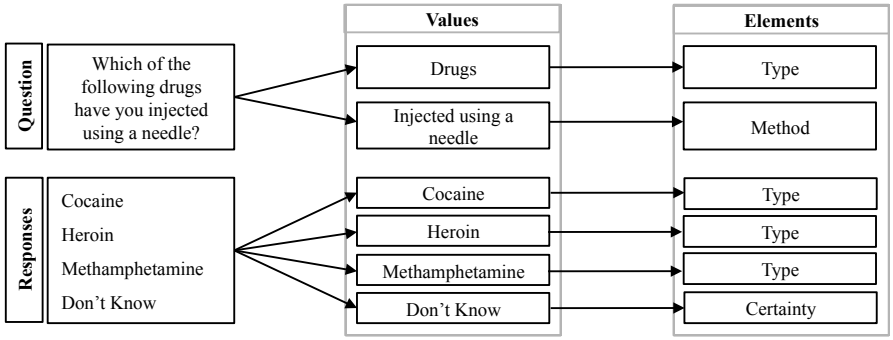

Figure 2. Analysis of questions and responses within research measures.

Annotation of the measures (Figure 2) was performed using the same guidelines for annotating the clinical text by designating words and phrases within questions as well as the responses to questions as values (each equivalent to one annotation), and then assigning a corresponding element to each value. A new element for *Time Point* (reference to a particular period of time) was observed within the questions of the PhenX measures and subsequently added to the guidelines. Prior to annotation of the two sets of identified measures, two annotators (EWC and ESC) evaluated

a subset of 10% of the measures from CDE Browser and PhenX Toolkit and achieved an inter-rater reliability using Cohen's kappa of 0.94 and 0.95 respectively. Similar to the clinical text, further analysis involved extracting the elements and corresponding values and generating statistics for each source as well as the combined set of measures. Using the combined annotations, those with similar meaning were grouped by a common word, phrase, or pattern.

Results

The HL7 and openEHR standards, 50 measures from PhenX Toolkit and CDE Browser containing a total of 131 questions, and 130 MTS sentences and 450 UVMMC comments represented a set of 27 drug use elements. Table 1 summarizes the distribution of these elements across the standards, clinical text, and research measures. In addition, this table shows the number and proportion of annotations associated with each element for MTS sentences (465 annotations), UVMMC comments (1,205 annotations), PhenX Toolkit (2,019 annotations), and CDE Browser (343 annotations). Of the 25 elements from prior studies[20, 36, 37] (indicated by '*' in Table 1), 24 retained their original meaning, *Duration* was further demarcated as *Duration of Use* and *Duration Since Time Point*, and a new element *Time Point* was added, which resulted in a total of 27 elements.

Table 1: Distribution of drug use elements across sources.

	Standards		**Clinical Text**		**Research Measures**	
Element	***HL7***	***openEHR***	***MTS (n=465)❖***	***UVMMC (n=1,205)❖***	***PhenX Toolkit (n=2,019)❖***	***CDE Browser (n=343)❖***
Status*	X	X	35 (7.5%)	103 (8.5%)	–	1 (0.4%)
Certainty*			5 (1.1%)	1 (0.1%)	140 (6.90%)	20 (8.5%)
Negation*		X	110 (23.7%)	54 (4.5%)	61 (3.0%)	17 (7.3%)
Temporal*			–	2 (0.2%)	9 (0.4%)	2 (0.9%)
Start Date*	X	X	–	3 (0.2%)	–	–
Start Age*		X	–	5 (0.4%)	9 (0.4%)	2 (0.9%)
Quit Date*	X	X	–	31 (2.6%)	–	2 (0.9%)
Quit Age*		X	–	–	–	–
Duration of Use			–	11 (0.9%)	7 (0.3%)	2 (0.9%)
Duration Since Quit*			–	24 (2.0%)	40 (2.0%)	–
Duration Since Time Point			1 (0.2%)	3 (0.2%)	7 (0.3%)	–
Time Frame*			6 (1.3%)	84 (7.0%)	174 (8.6%)	7 (3.0%)
Time Point			7 (1.5%)	48 (4.0%)	7 (0.3%)	3 (1.3%)
Method*	X	X	99 (21.3%)	147 (12.2%)	138 (6.8%)	40 (17.1%)
Type*	X	X	145 (31.2%)	327 (27.1%)	673 (33.3%)	152 (65.0%)
Subtype*			6 (1.3%)	2 (0.2%)	–	–
Amount*	X	X	16 (3.44%)	69 (5.7%)	80 (4.0%)	13 (5.6%)
Frequency*	X	X	5 (1.1%)	218 (18.1%)	591 (29.3%)	81 (34.6%)
Context*			1 (0.2%)	19 (1.6%)	5 (0.2%)	1 (0.4%)
Situation*			2 (0.4%)	5 (0.4%)	–	–
Location*			–	2 (0.2%)	–	–
Subject*			–	–	37 (1.8%)	–
Change*		X	–	–	–	–
Triggers*		X	–	–	–	–
Evidence of Dependence*		X	21 (4.5%)	21 (1.7%)	28 (1.4%)	1 (0.4%)
Cessation*		X	1 (0.2%)	6 (0.5%)	9 (0.4%)	–
Other*		X	5 (1.1%)	20 (1.7%)	3 (0.1%)	–
# of Elements	7	15	16	23	18	15

❖ number of annotations; * element derived from previous work[20, 36, 37].

As reflected in Table 1, seven elements were represented in the HL7 implementation guide while the openEHR archetypes cumulatively included 15 elements. MTS clinical text and measures from PhenX Toolkit and CDE Browser contained information for 16, 18, and 15, elements, respectively. Clinical text from UVMMC contained the most diverse drug use information representing 23 elements. *Quit Age*, *Change*, and *Triggers* were only found in the openEHR archetypes whereas *Time Frame* (e.g., 'last 6 months'), *Time Point* (e.g., 'tonight'), and *Context* (e.g., 'at

night for sleep') were specific to the research measures and clinical text. *Situation* (e.g., 'when going out to eat') and *Subtype* (e.g., 'pills: vicodin') occurred only in clinical text.

Within the standards, values were provided for some elements. For example, in the openEHR archetypes, values for *Status* included 'current user,' 'former regular user,' 'former occasional user,' and 'never user' while *Frequency* values included 'daily use,' 'weekly use,' 'irregular use,' and 'no use.' In addition, data types were provided that may be used to infer values for some elements (e.g., Date/Time for *Start Date* and *Quit Date*). Analysis and grouping of values focused on six elements that were found to be the most frequent across the clinical text and research measures: *Type*, *Frequency*, *Method*, *Negation*, *Time Frame*, and *Amount*. *Negation* was the third most frequently occurring element and included six groups of values: 'no,' 'none,' and 'without' (78 total and 3 unique values in measures) and 'denies,' 'does not,' and 'never' (164 total and 10 unique values in clinical text). Tables 2-6 include the value groupings for the remaining top five elements and shows for each element: (1) total number of values, number of unique values, and number of groups; (2) total number of values per group, frequency of group among the total values for element, and number of unique values within the group; and (3) example values for the highest frequency groups per element.

Table 2: Distribution of values for *Type* element. (total # of values per group; (frequency); [# unique values])

Clinical Text 472 Total Values; 109 Unique Values; 34 Groups	**Research Measures** 824 Total Values; 408 Unique Values; 203 Groups
Marijuana: 83 (17.6%) [11]	**Drugs:** 69 (8.4%) [16]
Cannabis	Any other drugs, drug products
Marijana, marijuana, marijuanna, MJ	Drug non-medical use only
Medical marijuana, prescribed marijuana	Drugs not prescribed by a doctor
Caffeine: 77 (16.3%) [10]	**Marijuana:** 50 (6.1%) [12]
Caffein, caffeine	Cannabinoids, cannabis
Coffee, coffees, ice coffee, green tea, tea	Hash, hash oil, hashish
Cola, pepsi, soda, energy drinks	Marijuana, mary jane
Illegal drugs: 62 (13.1%) [10]	**Cocaine:** 47 (5.7%) [10]
Illegal drug, illicit substance	Angel dust
All illicit, illicit drug, illicits	Cocaine, coke, crack
Multiple illicit drugs	Cocaine in chunk form

Table 2 shows the distribution of values for the element *Type* among clinical text and research measures. Values in the group 'Marijuana' were among the most commonly observed at a frequency of 6.1% of 824 values (research measures) and 17.6% of 472 values (clinical text) with the values 'marijuana' and 'cannabis' observed in both sources. The group 'Marijuana' also illustrates the occurrence of misspellings and abbreviations among the clinical text as observed with 'marijana,' 'marijuanna,' and 'MJ.' The group 'Caffeine' was observed only in clinical text and included values representing coffees, teas, and soft drinks.

Table 3: Distribution of values for *Frequency* element. (# of values; (Frequency); [# unique values])

Clinical Text 223 Total Values; 81 Unique Values; 7 Groups	**Research Measures** 672 Total Values; 52 Unique Values; 7 Groups
Occasionally: 114 (51.1%) [13]	**[#]:** 576 (85.7%) [18]
Occasionally, off and on, on occasion	Exact number: e.g., 0, 1, 2
Every few months	Range: e.g., 1-2, 6-9, 20-39
Every once in a while	>/<: e.g., 100 times or more
Intermittent	Calculation: e.g., average # times
[#] [time unit] per [time unit]: 102 (45.7%) [54]	**[#] [time unit] per [time unit]:** 51 (7.6%) [22]
/day, daily	Daily, weekly, monthly
/week, weekly	1 or 2 days a week, 2x weekly
1-2 times per month, 3x a day	About once a day, / day
Every other day, per day everyday	< once a month, monthly or less often
Other: 5 (2.1%) [4]	**Other:** 12 (1.6%) [5]
Chronic	More often than prescribed
Minimal, not usually	Most frequently, usually
Part-time	Never

Values for the *Frequency* element were categorized into seven groups for the clinical text and research measures with the three most frequent groups represented in Table 3. The group '[#] [time unit] per [time unit]' contained similar values for both sources including 'daily' and 'weekly.' In contrast, the most common *Frequency* group for clinical text was represented by the group 'Occasionally' while the measures reflected more precise values such as exact numbers (e.g., '0,' '1,' and '2') or ranges (e.g., '1-2' and '20-39') in the group '[#]'.

Table 4 shows that the top four groups for the *Method* element for clinical text and research measures are nearly equal with the exception that the value 'intake' appears in the clinical text whereas the value 'taken' is found in research measures. Values in the group 'Use' were represented in over 50.0% of the total values in clinical text and research measures. Within the measures, the group 'Inject' contained more descriptive values in the form of phrases such as 'injected using needle' and 'injection into the skin,' while clinical text contained more abbreviated values such as 'inject,' 'boot,' and 'IVDU.'

Table 4: Distribution of values for *Method* element. (# of values; (Frequency); [# unique values])

Clinical Text 246 Total Values; 33 Unique Values; 12 Groups	**Research Measures** 178 Total Values; 42 Unique Values; 16 Groups
Use: 158 (64.2%) [7]	**Use:** 91 (51.1%) [4]
Us, use, used, uses, usage, using	Use, used, using, usage
Utilize	**Take:** 32 (18.0%) [1]
Intake: 32 (13.0%) [1]	Taken
Intake	**Inject:** 15 (8.4%) [11]
Inhale: 22 (8.9%) [6]	Inject, injected
Smoke, smokes, smiokes, smoking	Injected using needle
Vaporized	Injection into the skin, IV injection
Inject: 14 (5.7%) [6]	**Inhale:** 13 (7.3%) [5]
Inject, injection	Inhaled,
Boot	Freebasing, Huffing
IVDU, needles	Smoked, smoking

For the element *Time Frame* in Table 5, information about a past period of time was represented in research measures by values in the group 'Past [#] [time unit]' including phrases such as 'during the last 30 days' and 'past 30 days up to and including today.' In contrast, a past time period was represented in the group 'Past' containing more generalized values such as 'former,' 'past,' and 'prior,' in addition to the group '[#] [time unit] ago' with more precise values including '10 years ago' and '1 week ago' within clinical text.

Table 5: Distribution of values for *Time Frame* element. (# of values; (Frequency); [# unique values])

Clinical Text 90 Total Values; 48 Unique Values; 9 Groups	**Research Measures** 182 Total Values; 31 Unique Values; 9 Groups
Past: 30 (33.3%) [6]	**Past [#] [time unit]:** 65 (35.7%) [7]
Former	During the last 30 days, past 30 days
Past, prior	During the last 12 months, past year
Prev, previous	Past 30 days up to and including today
Current: 18 (20.0%) [5]	**[#] [time unit]:** 60 (33.0%) [6]
Current, currently	_day(s), _week(s), _month(s), _year(s)
Now, present	A day
This time	Number of days_[range:0-30]
[#] [time unit] ago: 9 (10.0%) [9]	**Lifetime:** 43 (23.6%) [4]
Ten years ago	During your life, in your lifetime
2 weeks ago	Any time in her life
1 month ago	Any time in their lives

In Table 6, the element *Amount* was represented by the same four most frequently observed groups in both clinical text and research measures. The group '[#] [amount unit]' contained the most values for measures at a frequency of 23.4% of 94 total values and 30.6% of 85 total values for clinical text, encompassing 22 and 23 unique values

respectively. For both sources, the group 'Other' contained vague values for *Amount* including 'any,' 'other amounts,' 'too much,' 'some,' and 'little'.

Table 6: Distribution of values for *Amount* element. (# of values; (Frequency); [# unique values])

Clinical Text 85 Total Values; 48 Unique Values; 9 Groups	**Research Measures** 93 Total Values; 32 Unique Values; 9 Groups
[#] [amount unit]: 26 (30.6%) [23]	**[#] [amount unit]:** 22 (23.4%) [22]
1 bowl	_/_pills
10 bags	_/_grams
2 cans	_/_ampoules
None: 17 (20.0%) [2]	**[#]:** 17 (17.2%) [1]
None, nothing	Exact number: e.g., 0, 1, 2
[#]: 16 (18.8%) [11]	**None:** 13 (13.8%) [1]
1, 1-2, 6-7	None
one, x1	**Other:** 21 (23.5%) [4]
Other: 41 (37.4%) [9]	Any
Any, any significant, significant	Estimates
Some	Other amounts
Little	Too much

Discussion

Social and behavioral risk factors such as nicotine or alcohol use are known to impact health and are often documented in clinical settings. However, the collection of drug use information remains complex and often disregarded[17]. Documenting drug use patterns and behavior in the EHR could support research and public health policy; aid physicians and other clinicians in identifying drug misuse, abuse, and dependence; and, highlight risk factors for comorbid conditions that could lead to enhanced patient care through prevention, intervention, and cost-effective, targeted treatment. The findings of this study provide insights into the current scope of the collection of drug use information in clinical and research settings, and reflect the wide variation in the breadth and depth of drug use content within standards, clinical text, and research measures. These findings may be used to inform development of a comprehensive drug use model for guiding improved data collection and use in electronic health sources (e.g., EHRs and surveys), which could be further enhanced by understanding the requirements of different end users and use cases.

Numerous sources of measures related to social and behavioral domains were identified in this study, including caDSR and PhenX Toolkit that were examined. Additional resources were also identified (e.g., GEM) that may be further explored in future work. Of the sources examined, PhenX measures contained the most robust drug use content since some measures contain multiple questions and are supplemented with additional information including lists of drug use methods, amounts, and drug types. By contrast, drug use measures identified in the caDSR using the CDE Browser were represented by a single question with additional information statements for most measures and were from resources including the NCI cancer Biomedical Informatics Grid (caBIG; seven measures), Lombardi Cancer Center (LCC; two measures), National Institute on Drug Abuse (NIDA; 26 measures), National Institute of Dental and Craniofacial Research (NIDCR; one measure), and NCI Programs of Research Excellence (SPOREs; four measures). The challenge of identifying drug use-specific measures across the multitude of resources highlights the need to develop a set of standard measures for this domain and supports the recent focus on adopting standard social and behavioral measures in the EHR[17].

Collectively, the analysis of standards, clinical text, and research measures revealed a wide-ranging set of drug use values that were represented by 27 elements. The *Temporal* element was further categorized into a total of nine elements. *Start Date* and *Quit Age* were predominantly represented in standards, whereas *Start Age* and *Quit Date* were found across the source types. The elements *Duration of Use*, *Duration Since Quit*, *Duration Since Time Point*, *Time Frame*, and *Time Point* appeared in clinical text and research measures, and overall, the values for each element were more precisely represented in clinical text. For example, clinical text contained *Time Point* values such as 'Saturday,' '15 years ago,' '2006,' and 'July 4th' while 'first time,' 'last time,' and 'most recent time' were found in the measures. These value disparities highlight the inherent difference in purpose between the two sources: clinical text reflects a more factual account of patient history, whereas measures are a clinically investigative tool seeking information using a generalized set of questions. Of note, reference to the age a person used a drug (e.g.,

'age of 13-14') appeared only within clinical text and was annotated as a value within *Time Frame*; however, in future work the element *Use Age* could be added to represent these values most accurately.

On occasion, the analysis of values presented an organizational challenge. Complex values containing multiple elements were observed in some measures. For example, within the element *Amount*, the value '#' was further defined by the unit of measure including 'bags,' 'buttons,' 'capsules,' 'hits,' and 'rocks,' among other units which are nonstandard but have meaning in the drug use community. This finding thereby demonstrates the need for additional delineation of *Amount* values to include a numerical value as well as a unit of measure. *Type* values also presented a challenge since they were found in clinical text and research measures in the form of a slang or street name, or common, chemical, or pharmaceutical name. Interestingly, 'caffeine' was the second most common *Type* value in clinical text, motivating the exploration of a separate Caffeine model as a possible next step guided by existing standards such as the 'Caffeine Consumption' archetype in openEHR. Groupings for *Type* were designated either by a single value such as 'party pills' or by creating a group for words or phrases clearly representing one concept such as the group 'Methamphetamine' which contains the values 'crank,' 'crystal meth,' 'meth,' 'desoxyn,' and 'fumes of crystal meth.' Next steps for standardization and categorization of *Type* values may include exploring existing drug terminologies such as RxNorm from the National Library of Medicine[44, 45] that provides normalized names for clinical drugs, information and categorization of drugs from NIDA[46], and the Alcohol and Other Drug (AOD) Thesaurus from the National Institute on Alcohol Abuse and Alcoholism (NIAAA)[47]. The availability of a comprehensive resource that hierarchically organizes drug types, variant names, and groupings could be valuable for providing flexibility to support different use cases (e.g., selecting a specific drug type versus a more general drug category).

The findings from this study represent a preliminary synopsis of drug use information identified within diverse sources with future goals involving the creation of a formal representation of an integrated drug use model. To that end, a next step could involve retrieving existing codes associated with measures identified in this study. For example, the PhenX measure 'Substances – 30 day frequency' correlates to the Logical Observation Identifiers Names and Codes (LOINC) concept with the name 'PhenX - substance – 30 days frequency protocol,' whereas the measure 'Substance Abuse Illicit Substance Cocaine Personal Medical History Frequency' from the CDE Browser contains the concepts 'illicit substance,' 'cocaine,' and 'personal medical history' with corresponding NCI Thesaurus codes. In addition, drug use values could be mapped to standardized terminologies and coding systems such as SNOMED-CT[48 49] and the Unified Medical Language System (UMLS)[50]. Subsequent model development could also include alignment of these model components with national and international modeling initiatives such as the Clinical Information Modeling Initiative (CIMI)[51], with the focus of generating a semantically interoperable drug use model.

Conclusion

Recent focus on promoting the collection of behavioral and social information in the electronic health record has highlighted the needs and challenges in developing standardized measures for these domains, including drug use. This study provides a broad perspective on the current representation of drug use information as reflected by standards, in documentation from clinical settings, and within research measures. The findings further provide a foundation for next steps in the development of a comprehensive drug use model that might be used to support research, clinical, and public health applications.

Acknowledgments

The authors thank Tamara Winden for discussions related to the annotation guidelines; Elizabeth Lindemann for contributing to the annotation of clinical notes; and, Yan Wang for supporting aspects of the annotation process. The clinical notes were obtained with permission from MTSamples (https://www.mtsamples.com). Research reported in this manuscript was supported by the National Library of Medicine of the National Institutes of Health under award number R01LM011364. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Johnson NB, Hayes LD, Brown K, Hoo EC, Ethier KA. CDC National health report: leading causes of morbidity and mortality and associated behavioral risk and protective factors: United States, 2005-2013. 2014. Report No.: 1545-8636.

2. Jones CM, Mack KA, Paulozzi LJ. Pharmaceutical overdose deaths: United States, 2010. JAMA. 2013;309(7):657-9.
3. National Center for Health Statistics. Health, United States, 2013: with a special feature on prescription drugs. Hyattsville, MD. : Department of Health and Human Services, 2014. Report No.: 2014-1232.
4. Centers for Disease Control and Prevention. Prescription drug overdose in the United States: fact sheet: Centers for Disease Control and Prevention; [cited 2015 February 25]. Available from: http://www.cdc.gov/homeandrecreationalsafety/overdose/facts.html.
5. The Substance Abuse and Mental Health Services Administration. Substance abuse and mental health estimates from the 2013 national survey on drug use and health: overview of findings. Center for Behavioral Health Statistics and Quality, 2014. Report No.: 14-4863.
6. Roberts RE, Roberts CR, Xing Y. Comorbidity of substance use disorders and other psychiatric disorders among adolescents: evidence from an epidemiologic survey. Drug Alcohol Depend. 2007;88 Suppl 1:S4-13.
7. National Institute on Drug Abuse. Comorbidity: Addiction and Other Mental Illnesses. National Institutes of Health, September 2010. Report No.: 10-5771.
8. Mertens JR, Lu YW, Parthasarathy S, Moore C, Weisner CM. Medical and psychiatric conditions of alcohol and drug treatment patients in an HMO: comparison with matched controls. Arch Intern Med. 2003;163(20):2511-7.
9. Kedzior KK, Laeber LT. A positive association between anxiety disorders and cannabis use or cannabis use disorders in the general population: a meta-analysis of 31 studies. BMC Psychiatry. 2014;14:136.
10. 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;25(6):515-36.
11. Kim HM, Smith EG, Ganoczy D, Walters H, Stano CM, Ilgen MA, et al. Predictors of suicide in patient charts among patients with depression in the Veterans Health Administration health system: importance of prescription drug and alcohol abuse. J Clin Psychiatry. 2012;73(10):1269-75.
12. Blumenthal D, Tavenner M. The "meaningful use" regulation for electronic health records. N Engl J Med. 2010;363(6):501-4.
13. U.S. Department of Health and Human Services. The affordable care act, section by section [cited 2015 February 25]. Available from: http://www.hhs.gov/healthcare/rights/law/index.html.
14. Centers for Medicare and Medicaid Services. Eligible professional meaningful use table of contents core and menu set objectives. [cited 2015 February 25]. Available from: https://http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/EP-MU-TOC.pdf.
15. Ghitza UE, Tai B. Challenges and opportunities for integrating preventive substance-use-care services in primary care through the Affordable Care Act. J Health Care Poor Underserved. 2014;25(1 Suppl):36-45.
16. Institute of Medicine. Capturing social and behavioral domains and measures in electronic health records: phase 2. Washington, DC: The National Academies Press; 2014.
17. Adler NE, Stead WW. Patients in context: EHR capture of social and behavioral determinants of health. N Engl J Med. 2015;372(8):698-701.
18. Dolin RH, Alschuler L, Boyer S, Beebe C, Behlen FM, Biron PV, et al. HL7 Clinical Document Architecture, Release 2. J Am Med Inform Assoc. 2006;13(1):30-9.
19. openEHR. openEHR. [cited 2014 December]. Available from: http://hwww.openehr.org.
20. Melton GB, Manaktala S, Sarkar IN, Chen ES. Social and behavioral history information in public health datasets. AMIA Annu Symp Proc. 2012;2012:625-34.
21. Wu LT, Gersing KR, Swartz MS, Burchett B, Li TK, Blazer DG. Using electronic health records data to assess comorbidities of substance use and psychiatric diagnoses and treatment settings among adults. J Psychiatr Res. 2013;47(4):555-63.
22. Ghitza UE, Gore-Langton RE, Lindblad R, Shide D, Subramaniam G, Tai B. Common data elements for substance use disorders in electronic health records: the NIDA Clinical Trials Network experience. Addiction. 2013;108(1):3-8.
23. Tai B, McLellan AT. Integrating information on substance use disorders into electronic health record systems. J Subst Abuse Treat. 2012;43(1):12-9.
24. U.S. National Library of Medicine. Common data elements (CDE) resource portal. [cited 2015 February 26]. Available from: http://www.nlm.nih.gov/cde/.
25. National Cancer Institute. Cancer data standards registry and repository (caDSR). [cited 2014 February 26]. Available from: https://cdebrowser.nci.nih.gov/CDEBrowser/.
26. Moser RP, Hesse BW, Shaikh AR, Courtney P, Morgan G, Augustson E, et al. Grid-enabled measures: using Science 2.0 to standardize measures and share data. Am J Prev Med. 2011;40(5 Suppl 2):S134-43.

27. Hamilton CM, Strader LC, Pratt JG, Maiese D, Hendershot T, Kwok RK, et al. The PhenX Toolkit: get the most from your measures. Am J Epidemiol. 2011;174(3):253-60.
28. Genetic Alliance. Common data elements: making the mass of NIH measures more useful. [cited 2014 December]. Available from: http://www.geneticalliance.org/sites/default/files/webinararchive/070914Sheehan.pdf.
29. RTI International. PhenX Toolkit. [cited 2015 February 26]. Available from: https://http://www.phenxtoolkit.org.
30. Health Level Seven International. HL7 implementation guide for CDA release 2: IHE health story consolidation, release 1.1 - US realm. [cited 2014 December]. Available from: http://www.hl7.org/implement/standards/product_brief.cfm?product_id=258.
31. MTSamples. MTSamples.com. [cited 2014 December]. Available from: http://mtsamples.com.
32. Epic Systems Corporation. [cited 2014 December]. Available from: http://www.epic.com.
33. The University of Vermont Health Network. The University of Vermont Medical Center. [cited 2014 December]. Available from: https://http://www.uvmhealth.org/medcenter/Pages/default.aspx.
34. National Cancer Institute. Cancer data standard registry and repository (caDSR). [cited 2014 December]. Available from: http://cbiit.nci.nih.gov/ncip/biomedical-informatics-resources/interoperability-and-semantics/metadata-and-models.
35. openEHR. openEHR clinical knowledge mananger. [cited 2014 December]. Available from: http://www.openehr.org/ckm/.
36. 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;2011:227-36.
37. 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; 2014:366-74.
38. Chen ES, Carter EW, Winden TJ, Sarkar IN, Wang Y, Melton GB. Multi-source development of an integrated model for family health history. J Am Med Inform Assoc [Internet]. 2014 Oct 21. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25336591.
39. 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.
40. GATE: General Architecture for Text Engineering. [cited 2014 December]. Available from: http://gate.ac.uk.
41. Brat rapid annotation tool. [cited 2014 December]. Available from: http://brat.nlplab.org.
42. Stenetorp P, Pyysalo S, Topic G, Ohta T, Annaniadou S, Tsujii J. BRAT: a web-based tool for NLP-assisted text annotation. Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics. 2012:102-7.
43. National Cancer Institute. CDE Browser. [cited 2014 December]. Available from: https://cdebrowser.nci.nih.gov/CDEBrowser/.
44. Nelson SJ, Zeng K, Kilbourne J, Powell T, Moore R. Normalized names for clinical drugs: RxNorm at 6 years. J Am Med Inform Assoc. 2011;18(4):441-8.
45. National Institutes of Health. RxNorm [cited 2014 December]. Available from: http://www.nlm.nih.gov/research/umls/rxnorm/.
46. National Institute on Drug Abuse. The science of drug abuse and addiction. [cited 2015 February 26]. Available from: http://www.drugabuse.gov.
47. National Institute on Alcohol Abuse and Alcoholism. AOD thesaurus: alcohol and alcohol problems science database. [cited 2015 February 26]. Available from: http://etoh.niaaa.nih.gov/AODVol1/Aodthome.htm.
48. International Health Terminology Standards Development Organisation. Systemized nomenclature of medicine clinical terms (SNOMED). [cited 2015 February 26]. Available from: http://www.ihtsdo.org/snomed-ct/.
49. International Health Terminology Standards Development Organisation. IHTSDO [cited 2015 February 26]. Available from: http://www.ihtsdo.org.
50. National Library of Medicine. Unified medical language system (UMLS). [cited 2015 February 26]. Available from: http://www.nlm.nih.gov/research/umls/.
51. Clinical Information Modeling Initiative. CIMI [cited 2015 February 26]. Available from: http://informatics.mayo.edu/CIMI/index.php/Main_Page.