Social and Behavioral History Information in Public Health Datasets

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

Abstract

Social and behavioral history is increasingly recognized as integral for understanding important determinants of disease and critical for patient care, research, clinical guidelines, and public health policies. Social and behavioral history information in the public health domain, specifically large public health surveys, has not been well described. In this study, a content analysis was performed and information model constructed and contrasted with clinically-based models for each of three widely used public health surveys: BRFSS (Behavioral Risk Factor Surveillance System), NHANES (National Health and Nutrition Examination Survey), and NHIS (National Health Interview Survey). Survey items were predominantly related to alcohol use, drug use, occupation, and tobacco use. Although the clinical social history information model was similar, public health social history demonstrated additional complexity in coding temporality, degree of exposure, and certainty. Our results give insight into ongoing efforts to integrate clinical and public health information resources for improving and measuring health.

Introduction

Social determinants, including behavioral and environmental factors, are increasingly recognized as key modifiable factors for many causes of disease, disability, and mortality in the United States. Some of the common risk factors leading to disease include lack of physical activity, poor diet, tobacco usage, alcohol consumption, exposure to toxic agents particularly occupationally-related, illicit drug use, and high-risk sexual behavior (1). Social factors are also important as an aid and guidepost for proper guidance and counseling for preventative care, for establishment of public health priorities, and for evaluation of interventions at both an individual patient and societal level. A range of studies have demonstrated linkages between social determinants and mortality (1-3), chronic diseases (4), and other co-occurrences such as substance use/abuse and mental health disorders such as depression (5-7).

Simultaneously, there is increasing emphasis on the importance of understanding, utilizing, and managing information related to behavioral and environmental determinant of disease. In 2006, the Institute of Medicine (IOM) report "Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate" was an important call to action for researchers to perform integrated research to study the effects of social, behavioral, and genetic factors on health outcomes (8). To build upon existing efforts and further advance research in this area, the report describes the need for enhancing existing datasets, developing new data sources, and establishing strategies and models for incorporating such factors and their interactions. In addition, increased interest in studying gene-environment interactions upon health also highlights the opportunity to study the influence of social and behavioral variables building upon efforts to understand the causal or interactive relationships between genes and risk factors on disease (9, 10). A follow-up 2011 IOM report, "Incorporating Occupational Information in Electronic Health Records: Letter Report", emphasized the importance of collecting occupational history information from both public health and electronic health record (EHR) systems, highlighting the importance of occupational information in diagnosis and treatment of several diseases, as well as key to establishing policies, interventions, and prevention strategies to improve the health of the working population (11). This report also highlighted three immediate high data priorities: occupation, industry, and work-relatedness.

More broadly, the recent IOM report, "For the Public's Health: The Role of Measurement in Action and Accountability", examined overall population health strategies, processes, and metrics in the context of today's reformed health system (12). The report found that population health information currently lacks a coherent framework or system for understanding the health status of our population. They recommend significant changes to the processes, tools, and approaches to gather information on health outcomes so that decision-makers can obtain sufficient information to make important policies about health in their respective communities. Two of their main recommendations relate to allowing public health agencies and medical care organizations and providers to share

information from clinical and public health sources to inform care and population health priorities. These core recommendations also point to the need to integrate and harmonize information models associated with clinical and public health datasets as complementary resources.

The overarching goal of this study was to understand data collection for social and behavioral factors in widely used public health datasets in order to achieve the following subgoals: (a) enhance information models for social and behavioral factors, (b) aid organization and structure of dataset information to facilitate data integration, and (c) provide a resource to current public health datasets by a particular social factor.

Background

National Health Surveys: BFRSS, NHIS, and NHANES

Several key surveys collect information related to social, behavioral, or disease factors for specific populations and have been used as key resources for a number of studies (8). Currently, there are three main national surveys: Behavioral Risk Factor Surveillance System (BRFSS) (13), National Health Interview Survey (NHIS) (14), and National Health and Nutrition Examination Survey (NHANES) (15). BRFSS and NHIS are annual surveys, whereas NHANES collects data every 2 years. These surveys have been used for a wide range of studies for health services research (16, 17), as adjuncts to other data sources (18, 19), and for multiple other applications (20, 21).

Established in 1984, BRFSS is an on ongoing telephone survey. It is maintained by the Centers for Disease Control and Prevention (CDC) Office of Surveillance, Epidemiology, and Laboratory Services and is a state-based health survey system, which gives information about health risk behaviors, clinical preventive practices, and health care access and use. Adults 18 years or older are asked to take part in the survey. Only one adult is interviewed per household over the telephone via random digit dialing of a cross-section of the population. Each individual state health department conducts the survey. Currently, data are collected in all 50 states, the District of Columbia, Puerto Rico, the U.S. Virgin Islands, and Guam. While BFRSS has a declining and low response rate (national median 34 percent) and little data on children, it is conducted on over 350,000 adults annually, making it the largest telephone health survey in the world.

In contrast, NHIS (established in 1957) aims to monitor U.S. population health by collecting information on a broad range of health topics and is maintained by the CDC National Center for Health Statistics. The survey aims to be representative of U.S. civilian non-institutionalized individuals. Data are collected with face-to-face interviews on all family members by one or more adult respondents that are present at the time of interview. Some additional information is collected about children with an adult respondent for one child per family.

Finally, NHANES began in the early 1960s and was developed as a series of surveys, focusing on different population groups or health foci. The survey is unique in that it combines interviews and physical examinations. NHANES is maintained by the CDC National Center for Health Statistics and was designed specifically to assess the health and nutritional status of U.S. adults and children. NHANES is continuous in nature and currently examines a representative sample of 5,000 persons each year with health interviews being conducted in respondents' homes. In addition to the physical examination portion, respondents are able to enter their own responses on a touch-screen allowing sensitive questions to be answered with privacy (22).

Modeling social history in EHR system clinical texts

Our group recently performed a multi-institutional study on social history information in EHR system clinical notes, comparing the content of statements in this section with current information models, namely HL7 and OpenEHR models (23). The motivation of this study was to gain a better understanding of the content of social history information documented in electronic clinical notes. Social history information from clinical notes from three different sources (MTSamples.com (24), University of Vermont affiliated Fletcher Allen Health Center, and University of Minnesota affiliated Fairview Health Services) was analyzed and the representation of this information with HL7 Clinical Document Architecture (25) and openEHR (26) was evaluated. Overall, 260 clinical notes (consisting of 989 sentences and 1,439 statements relevant to social history) were analyzed and 35 statement types identified (Table 1). Based on review of implementation guides associated with the HL7 Clinical Document Architecture and Clinical Knowledge Manager for openEHR for the top eight statement types across sources, the two standards were found to provide different representations that were only partially capable of capturing the breadth and granularity of information contained within the various statements.

Table 1. Statement Types for Social History in Clinical Notes *(top 12 types across clinical sources in bold)*

• Abuse	• Ethnicity	• **Living Situation**	• Sexual Activity
• **Alcohol Use**	• **Family**	• **Marital Status**	• Sexual Orientation
• **Animals**	• **Functional Status**	• Mental/Emotional Status	• Sick Contact
• Caffeine Use	• Gender Identity	• Military Service	• Social Support
• Daily Activity	• Health Status	• **Occupation**	• **Tobacco Use**
• Diet	• Hobby	• **Physical Activity**	• Travel
• **Drug Use**	• Infectious Disease	• Reproductive Activity	• Weight Management
• **Education**	• Insurance	• **Residence**	• Other
• Environmental/ Occupational Exposure	• Legal	• Safety/Preventative Care	

Materials and Methods

Each public health dataset was analyzed for social and behavioral factor survey items. The representation of the information from survey responses was evaluated and contrasted with the previously described clinically derived social history model (23) (Figure 1). Our overall approach included steps to: (1) analyze social history information collected from three public health data sets, (2) place this information into a public health social history model, and (3) contrast these models with clinical models for social history. We used the 2009 version of BRFSS from the BRFSS Questions Archive; the 2009-10 questionnaire taken from different sub questionnaires (Family Questionnaire, Sample Person Questionnaire, Audio Computer Assisted Personal Self Interview, and Computer Assisted Personal Interview) from NHANES, and the 2010 version of NHIS for this study (excluding supplements such as the NHIS Sample Adult Cancer Questionnaire as this supplement was discontinued following 2010). For BRFSS, only core questions were included which meant that state-specific portions of BRFSS and the disease-related modules not specifically part of the core questions were excluded.
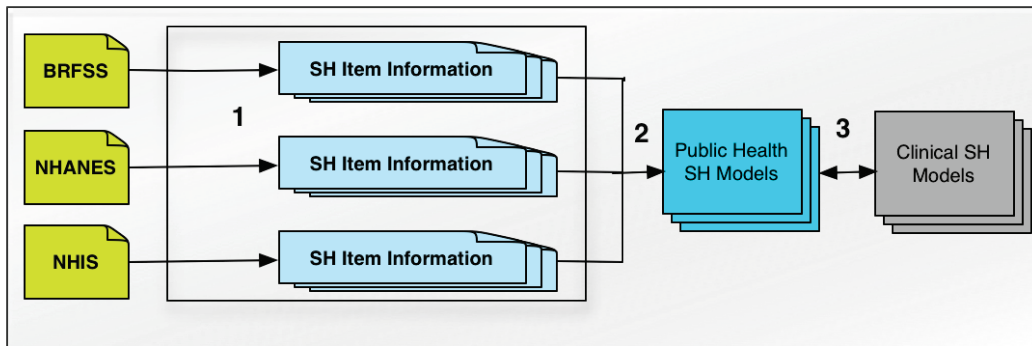

Figure 1: Overview of Materials and Methods.

1. *Coverage of Social and Behavioral History Information in Public Health Dataset Items*
The first phase of the study involved collection of BRFSS, NHIS, and NHANES social and behavioral health items. For the purposes of this analysis, we focused on the twelve most common social history categories observed in clinical notes identified in our previous study (Table 1). All items from the three surveys were reviewed to identify and broadly group survey items into these categories. In a subset of items, inter-rater reliability between two reviewers in the assignment of item categories was assessed with a κ (Kappa) statistic and proportion agreement.

2. *Modeling Social and Behavioral Items from Public Health Dataset Items*
Using an iterative, consensus-based process with a minimum of three individuals in each session with clinical (GM, SM), public health (SM, IS), and biomedical informatics (EC, IS, GM, SM) expertise, the extracted survey items were analyzed in 3 iterations aiming to create a 'Public Health Social History Model' for the broad categories with at least five survey items. For each item, potential information from response stems (answers) to each survey item was encoded (versus the information from the survey item question). While variation in information was found across categories, an attempt was made to define a common set of data elements. Value sets were populated by information found within respective survey question items. With the conclusion of each iteration, the group reviewed modeling onto each element axis and the corresponding enumerated values to resolve any disagreements

and guide the next iteration. Based on the resulting list of items responses, element axes, and value enumerations, items were re-analyzed with respect to structure and detailed content. Similarly, a random subset of 40 items (25%) was assessed individually by two reviewers for proportion agreement in coding the information for each item in all axes for the entire survey item.

3. *Comparison of the Public Health Social History Model with the Clinical Social History Model*
Clinical social history models previously described based off of clinical narrative statements (23) were contrasted with models generated from the three Public Health Datasets, including axes and value enumerations for each axis. The adequacy of the clinical model for representing information within relevant survey items across the three sources was assessed. Each category was explicitly modeled and contrasted in order to understand the overlap and gaps between the models generated with each of the two approaches.

Results

Overall, 161 survey items in the top twelve social history categories were contained in the three surveys. As reflected in Table 2, TOBACCO USE statements were the most common, followed by DRUG USE, ALCOHOL USE, and OCCUPATION question items. Only three social and behavioral history categories were included in all three surveys (i.e., ALCOHOL USE, OCCUPATION, and TOBACCO USE), but these 107 items constituted the majority (66%). Another eight items (5%) in two categories (i.e., MARITAL STATUS and RESIDENCE) were covered in two survey datasets, and the remaining 45 items (28%) in five categories were unique to a single survey. Example items for the six most common categories (at least 5 survey questions) are shown in Table 3. Inter-rater reliability between two reviewers in the assignment of statement types for a subset of 30 items (19% of sample) from BRFSS (n=5), NHIS (n=10), and NHANES (n=15) was excellent with **κ** (1.000) and proportion agreement (100%, 100%, and 100%), respectively.

Table 2: Distribution of Survey Items Across the Three Sources.

Statement Type	Overall	BRFSS	NHANES	NHIS
ALCOHOL USE	23	6	10	7
ANIMALS	-	-	-	-
DRUG USE	31	-	31	-
EDUCATION	1	-	1	-
FAMILY	3	3	-	-
FUNCTIONAL STATUS	2	2	-	-
LIVING SITUATION	-	-	-	-
MARITAL STATUS	2	1	-	1
OCCUPATION	22	5	13	4
PHYSICAL ACTIVITY	8	8	-	-
RESIDENCE	6	3	3	-
TOBACCO USE	62	5	48*	9
Total	161	33	107	21

*3 repeated questions

Models for each of the 6 item categories were created with sets of data element axes and values. Inter-rater reliability between two reviewers for mapping statements fully for all axes in a subset of 40 items (25%) demonstrated proportion agreement 87.5% in coding all axes completely. An attempt was made to define a common set of data elements or axes and to populate the specific values for the survey items. Table 4 depicts the data elements and example values associated with the most frequent categories along with the distribution of coverage over each axis for the 158 unique survey items. In all categories, because the great majority of question items had specific responses about *certainty* (e.g., 'yes', 'no', 'uncertain', 'don't know', 'refused'), which differ from the *status* of an item (e.g., 'past', 'future', 'ever', 'current'), the axis *certainty* was explicitly modeled and separated from *status*. Furthermore, almost all survey items specified *experiencer* in that most survey items were posed to 'you' or 'spouse'. Across the six categories, *type* was specified in 65% of survey items. In the categories ALCOHOL USE, DRUG USE, PHYSICAL ACTIVITY, and TOBACCO USE, 46% of survey items specified information about *frequency*. While some categories, such as ALCOHOL USE and PHYSICAL ACTIVITY, had a high specification for *amount* (96% and 75% respectively), other categories did not. A total of 2 items that mapped to a category did not readily fit into the social history model. For example, with DRUG USE the question: "*<Have you/Have you not> ever been in a drug rehabilitation program?*" and its corresponding answers could not be encoded using the information model.

Table 3: Example Social History Survey Items in Most Common Categories and Responses by Category.

Original Survey Item	Survey Item Responses	Response Information
ALCOHOL USE		
(BRFSS) During the past 30 days, what is the largest number of drinks you had on any occasion?	_ _ =Number of drinks; 77=DK/NS; 99=Refused	Largest <#> drinks on any occasion in past 30 days
(NHANES) During the past 30 days, on how many days did you have at least one drink of alcohol?	1= 0 days; 2= 1 or 2 days; 3= 3 to 5 days; 4= 6 to 9 days; 5= 10 to 19 days; 6= 20 to 29 days; 7= All 30 days; 77 = REFUSED; 99 = DON'T KNOW	<#> days of alcohol use of at least one drink in past 30 days
(NHIS) In the PAST YEAR, on how many DAYS did you have 5 or more drinks of any alcoholic beverage?	000 = Never; 001-365 = 1- 365 days; 997 = Refused; 999 = Don't know	<#> days when drank ≥ 5 drinks in a day in past year
DRUG USE		
(NHANES) Have you ever used cocaine, crack cocaine, heroin, or methamphetamine?	1=Yes; 2=No; 7=Refused; 9=Don't Know	You <have/have not> ever used cocaine, crack cocaine, heroin, or methamphetamine
(NHANES) Have you ever, even once, used a needle to inject a drug not prescribed by a doctor?	1=Yes; 2=No; 7=Refused; 9=Don't Know	You <have/have not> ever used a needle to inject an unprescribed drug
OCCUPATION		
(BRFSS) Thinking about the last time you worked, at your main job or business, how were you generally paid for the work you do? Were you:	1=Paid by salary; 2=Paid by the hour; 3=Paid by the job/task (e.g., commission, piecework); 4=Paid some other way; 7=DK/NS; 9=Refused	<Compensation type > with last time you worked at your main job
(NHANES) Which of the following {were you/was} {NON-SP HEAD/NON-SP SPOUSE} doing last week?	1 =Working at a job or business; 2 = With a job or business but not at work; 3 = Looking for work; 4 = Not working at a job or business; 7 = Refused; 9 = Don't know	<Occupation status> of <you/NON-SP HEAD/NON-SP SPOUSE> last week
(NHIS) What is the main reason you did not work last week/What is the main reason you did not have a job or business last week?	01 = Taking care of house or family; 02 = Going to school; 03 = Retired; 04 = On a planned vacation from work; 05 = On family or maternity leave; 06 = Temporarily unable to work for health reasons; 07 = Have job/contract and off-season; 08 = On layoff; 09 = Disabled; 10 = Other; 97 = Refused; 99 = Don't know	<Reason> for not working last week
PHYSICAL ACTIVITY		
(BRFSS) When you are at work, which of the following best describes what you do? Would you say-- If respondent has multiple jobs, include all jobs.	1=Mostly sitting or standing; 2=Mostly walking; 3=Mostly heavy labor or physically demanding work; 7=DK/NS; 9=Refused	<Predominant activity> at work
(BRFSS) On days when you do moderate activities for at least 10 minutes at a time, how much total time per day do you spend doing these activities?	_:_ _=Hours and minutes per day; 777=DK/NS; 999=Refused	On days of moderate activity, <#:## hr:minute duration> total time per day of moderate activity
RESIDENCE		
(BRFSS) Do you own or rent your home? (Note: "Other arrangement" may include group home or staying with friends or family without paying rent.)	1=Own; 2=Rent; 3=Other arrangement; 7=DK/NS; 9=Refused	<Own/rent> home has <other arrangement>
(NHANES) In what state {were you/was NON-SP HEAD} born?	ENTER 2 LETTER STATE ABBREVIATION TO START THE LOOKUP	<you/NON-SP HEAD/NON-SP SPOUSE> born in <state>
TOBACCO USE		
(BRFSS) Have you smoked at least 100 cigarettes in your entire life?	1=Yes; 2=No;7=DK/NS; 9=Refused	<Have/Have not> smoked ≥100 cigarettes in entire life
(NHANES) {Were/Was} the {BRAND REPORTED IN SMQ.660/brand of} cigarettes menthol or non-menthol?	1=Menthol; 2=Non Menthol; 7=Refused; 9=Don't Know	<mentholated/non-mentholated/unknown> cigarette type
(NHIS) Do you NOW smoke cigarettes every day, some days or not at all?	1 = Everyday; 2= Some days; 3= Not at all; 7 = Refused; 9= Don't know	Current <frequency > smoke cigarettes

DK/NS, don't know/not sure. SP, spouse

Table 4: Public Health Social History Elements, Axes, and Values by Statement Type.

Statement Type	Element	Example Values or *Patterns*	Distribution of element usage
ALCOHOL USE (n=23)	Status	current, ever	17%
	Temporal	*past <#> <timeunit>*, entire life, any one year, *<age in years> first drank*	96%
	Method	oral	96%
	Type	beer, wine, hard liquor, malt liquor	48%
	Amount	≥1 drink, ≥5 drinks (no gender difference), ≥4 drinks(woman)/≥5 drinks(man)	96%
	Frequency	*<#> <days/times> / <week/month/year/past 30 days/past year/entire life>*, almost everyday	52%
	Certainty	yes/present, no/absent, don't know/not sure, refused	100%
	Experiencer	you, spouse	100%
DRUG USE (n=31)	Status	ever	26%
	Temporal	last smoked, *<timeunit> since last use*, last 30 days, age at first use, lifelong	71%
	Method	smoke, injection	35%
	Type	cocaine, crack, heroin, methamphetamine, marijuana, hashish, steroids, other drugs, unprescribed drugs, injection drugs	97%
	Amount	<#> of joints	3%
	Frequency	once per month, <#> times per month, <#> days in last month, <#> times of lifetime use, even once	61%
	Certainty	yes, no, refused, don't know	100%
	Experiencer	you	100%
OCCUPATION (n=22)	Status	current, ever	18%
	Temporal	present, last week, when you were at work	41%
	Method	paid by salary, never worked, paid by the hour, paid by the job/task, paid some other way, looking for work, disabled, retired, employed, on layoff	18%
	Type	*<occupation>*, armed services, industry type, *<activity type>*, taking care of house of family member, going to school, state government employee	45%
	Subtype	another schedule, a rotating shift, a regular night shift, a regular daytime shift	9%
	Amount	<#> hours, <#> days/weeks/months/years	9%
	Frequency	more than 35 hours per week, less than 35 hours per week	18%
	Location	at all jobs and businesses, at main job or business, *<employer name>*	32%
	Certainty	yes, no, refused, don't know	100%
	Experiencer	you, non-SP head, SP	91%
PHYSICAL ACTIVITY (n=8)	Temporal	past month, on days of activity ≥10 minutes	38%
	Type	any physical activity, mostly sitting or standing, mostly walking, mostly heavy labor or strenuous activity, moderate activity, vigorous activity	100%
	Amount	≥10 minutes, *<#:## hours:minutes> total time*	75%
	Frequency	≥1 time/week, <#> times/week	63%
	Certainty	yes/present, no/absent, don't know/not sure, refused	100%
	Experiencer	you	100%
	Location	at work, not at work	38%
RESIDENCE (n=6)	Status	current	50%
	Method	own, rent, other arrangement without paying rent, born	67%
	Type	home, other (group home, stay with friends/family without rent)	17%
	Location	<FIPS country code>, <zip code>, USA, other country, <state abbreviation>	83%
	Certainty	yes/present, no/absent, don't know/not sure, refused	67%
	Experiencer	you, non-SP head	100%
TOBACCO USE (n=62)	Status	current, past (quit), quit attempt	16%
	Temporal	entire life, past <#> <months/days>, *<age in years> first started to smoke*, *<#> <days/weeks/months/years> since quit*, <#> minutes after waking up	16%
	Method	smoke, stop smoke	77%
	Type	cigarettes, chewing tobacco, snuff, snus, cigar, pipe, nicotine patch/gum	71%
	Subtype	*<UPC code>, <brand>, <filtered/unfiltered>, <mentholated/non-mentholated>, <size category>, <brand qualifiers>*	15%
	Amount	*<#> [cigarettes/average cigarettes], ≥100, ≥<#> or <#> puffs*	19%
	Frequency	regularly, some days, daily, not at all, fairly regularly, <#> days/month	35%
	Certainty	yes/present, no/absent, don't know/not sure, refused	85%
	Experiencer	you, spouse, *<name>*, anyone who lives here	90%
	Location	anywhere inside home	3%

TOBACCO USE

Status*	**current** / **quit** / quit attempt	*no/negative** / *past* / *desire quit*
Temporal	**[#] <timeunit> ago/duration** / <age> when last smoked / > <#> hour after waking	*many years* / >= [#] timeunit / entire life
Method	**smoke** / oral	*snort*
Type	**cigarettes** / **cigars** / **snuff**	**chewing tobacco** / **pipe** / nicotine patches/gum
Subtype	<brand> / [filtered/unfiltered] / [mentholated/nonmentholated]	<size category> / <UPC code> / <brand qualifier>
Amount	**<#> cigarettes** / <#> on average / at least 100	<#> packs \ times / significant / at least <#> puffs
Frequency	**daily** / some days / regularly	*heavy* / *socially* / <#> days/month
Certainty*	yes / don't know / refused	no / not sure
Experiencer	anyone who lives here / you	<name> of person in household / spouse
Location	anywhere inside home	

ALCOHOL USE

Status*	**current** / ever / uncertain*	*no/negative** / *denies** / *past (quit)*
Temporal	**[#] <timeunit> ago/duration** / <age> first had a drink / any one year	*in/since [date]* / <timeunit> consumed in past year / entire life
Method	**oral**	
Type	**beer** / **hard liquor** / drinks	**wine** / malt beverage
Amount	*variable* / [#] glasses/pints/bottles/times / maximum <#> drinks	*social* / at least one drink / <#> or more drinks
Frequency	*daily/weekly/monthly/yearly* / <#> days/week\month\year / <#> times in life\past 30 days	*heavy/occasional* / almost everyday
Certainty*	yes / don't know / refused	no / not sure
Experiencer	you	spouse

Key: Dotted line boxes unique to PH Model. **Bolded** text, both models. Underlined text, unique to PH Model. *Italicized* text, unique to Clinical Model. **Status* in clinical model split into *Status* & *Certainty* in PH Model.
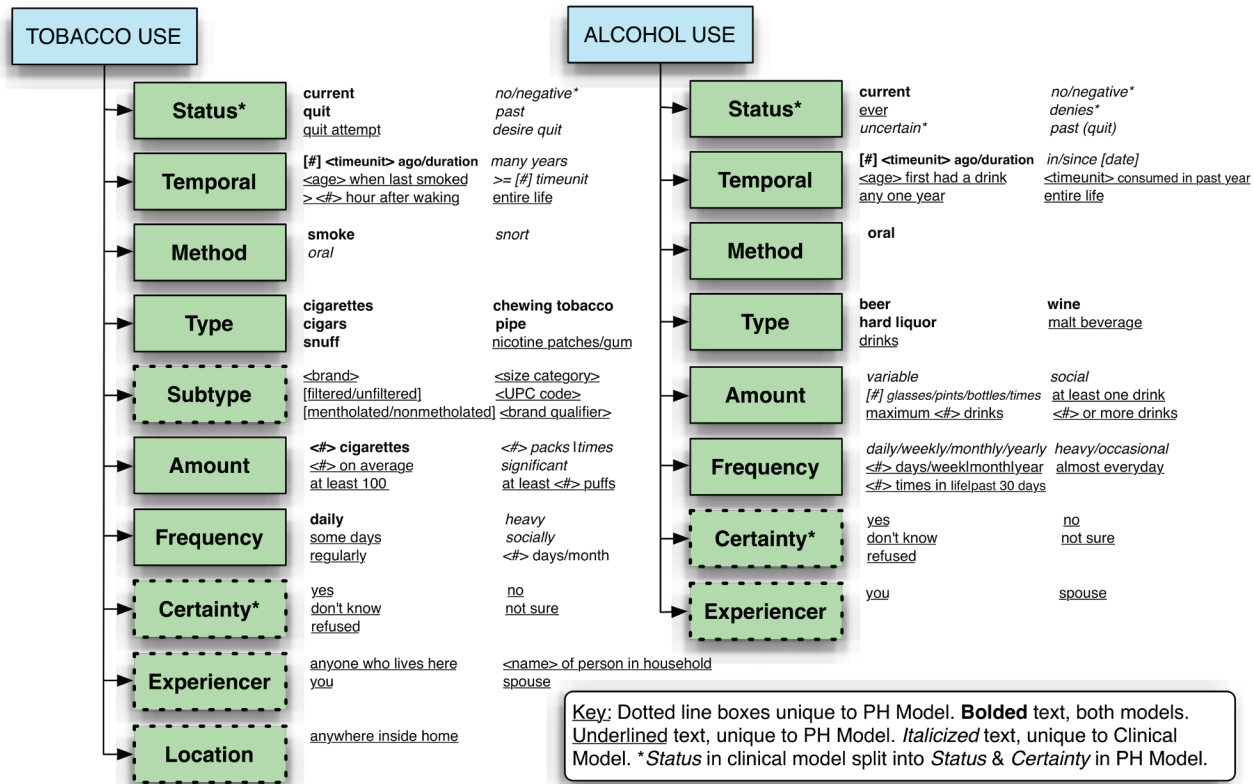
Figure 2. Public Health and Clinical Social History Models for ALCOHOL USE and TOBACCO USE

Formal comparison of the public health model to the clinical model demonstrated significant areas of overlap, as well as areas of distinct axes and value enumerations with the public health model. Figure 2 graphically contrasts the public health model for the two categories with the greatest items, ALCOHOL USE and TOBACCO USE. New axes *experiencer* and *certainty* were added for all categories for the reasons previously described. In addition, several models required the addition of axes such as *subtype* for both TOBACCO USE and OCCUPATION. In the case of TOBACCO USE, *subtype* signified detailed information about cigarette brand and type; for OCCUPATION, *subtype* was used for work shifts or scheduling (e.g., 'day shift' or 'rotating shift'). Overall, while some enumerated values overlapped significantly, such as the axis *type* representing the most general specification of the category (e.g., "beer" and "wine" for ALCOHOL USE), a large number of values for axes were each unique to the clinical or public health model.

Discussion

As the focus of healthcare and medicine increasingly moves from treatment and other therapeutics to prevention for improving the health of our communities, data integration and normalization for surveillance and monitoring of population health will become an even greater priority. This includes harnessing the information from public health datasets, EHR systems, and newly established bi-directional systems. Social and behavioral history information gives improved understanding of health with both an individual patient and population perspective, which will also aid research, patient care, and public health policy. We expand upon work with social history information modeling from clinical notes and biomedical standards to understand public health datasets and their representation of social history information. Overall, this study provides a foundation for examining the complementary information represented in these datasets, and demonstrates some of the challenges in integrating these datasets with clinical sources to leverage these information resources.

Our analysis demonstrated that public health survey items addressing social history had significant depth in detail for certain questions (e.g., TOBACCO USE and OCCUPATION), yet these surveys lacked much or any coverage for some categories of social history items altogether, such as ANIMALS, LIVING SITUATION, or FUNCTIONAL STATUS. Furthermore, because of the detailed nature and need for sufficient validity and reliability

of these survey items, our developed Public Health Social History Models had significant information about *certainty, frequency*, and *temporal* data elements. Furthermore, for the category TOBACCO USE, survey questions included quite specific information about cigarette type, including *subtype* information (an added axis) including information about the size of the cigarette, whether or not it was mentholated, package information <UPC code>, and other detailed information. Similarly, while TOBACCO USE and DRUG USE had elements of "attempt to quit", ALCOHOL USE questions did not cover this important concept in its respective survey items. This issue may be aided through the reuse of values from complementary statement types to build up social history information models further. In our previous study, many of the items were found to be structured similarly and are analogous in other important ways such as TOBACCO USE, DRUG USE, ALCOHOL USE, and CAFFEINE USE. Similarly, ENVIRONMENTAL/OCCUPATIONAL EXPOSURE, TRAVEL, ANIMALS, and SICK CONTACTS can be grouped as they each relate to exposures of different types in different settings.

In addition to the modification and addition of information axes, comparison of the clinical and public health models demonstrated significant complexity in expressions of temporal information and degree in the public health model, the later of which is reflected in the need for the addition of a number of detailed values for *amount* and *frequency*. Furthermore, since *amount, frequency*, and *temporal* information were inter-related in some survey items, these elements had to be explicitly modeled and separated. For example, the following survey item from NHANES is illustrative of this issue:

> 'During the past 30 days, on how many days did you have 5 or more drinks of alcohol in a row…within a couple of hours?'

Here, the *temporal* information is 'past 30 days', the *amount* is '5 or more drinks [in a row]', and the frequency is '<#> days in past 30 days'. Despite the complexity in encoding this information, an understanding of the degree of exposure, such as the concept of 'pack-years' with smoking has importance for disease-risk, particularly in the case of cardiopulmonary disease with smoking and liver disease with alcohol use.

There were also some unique challenges to encoding information with respect to *status* and *certainty* elements. For the purposes of encoding the information from public health survey items, we expanded what was originally *status* alone for clinical social history to two elements, *status* and *certainty*, in this follow-up study. This was done to separate out information about the relative *certainty* of the information of the survey element, which was potentially covered in most survey questions (e.g., *don't know, not sure, no, yes, refused*). In contrast, *status* covered items such as *current, ever, quit,* or *past* including temporal elements related to the information and present status of the encoded information.

While BRFSS, NHANES, and NHIS are the most widely used public health surveys, this study is limited in that our models are not contemporaneous and the surveys continue to be modified on a regular basis. Furthermore, while BRFSS has several state-specific modules, we only analyzed core questions and not state-specific survey items. We also found that while publically available, all three surveys were not easy to browse and search. Finding the exhaustive list of survey items for each category was a non-trivial and time-consuming task. Further work to incorporate other public health and patient-specific data sources is the next step in expanding the social history information model.

As a long-term goal, we aim to improve the capture and standardization of social history information from a range of complementary clinical, public health, and consumer self-report sources. Together, enhanced information models can potentially improve the interoperability and reuse of these data for a wide range of applications to respective stakeholders, improving both research and patient care. This, combined with better methodologies to improve the analysis of this data including associations with disease will provide increased abilities to understand correlations and links between the social determinant of health, disease, and opportunities for prevention.

Conclusion

The goal of this study was to gain a better understanding of the content of social history information in public health datasets in the overall context of existing standards and work in the clinical domain. Through analysis of BRFSS, NHANES, and NHIS social history question items were identified and mapped to a public health social and behavioral history model. Further analysis of these questions revealed a variety of information that differed significantly from information obtained in the clinical domain. These findings provide guidance for enhanced collection and representation of social history from these complementary knowledge resources.

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.   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.
6.   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.
7.   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.
8.   In: Hernandez LM, Blazer DG, editors. Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate. Washington (DC)2006.
9.   Ottman R. An epidemiologic approach to gene-environment interaction. Genet Epidemiol. [Research Support, U.S. Gov't, P.H.S.]. 1990;7(3):177-85.
10.   Ottman R. Gene-environment interaction: definitions and study designs. Prev Med. [Research Support, U.S. Gov't, P.H.S.]. 1996 Nov-Dec;25(6):764-70.
11.   Linda Hawes Clever MEBR, Andrea M. Schultz, and, Catharyn T. Liverman ECotRP, of CfOHNPI, Medicine, editors. Incorporating Occupational Information in Electronic Health Records: Letter Report. Washington, D.C.: The National Academies Press; 2011.
12.   For the Public's Health: The Role of Measurement in Action and Accountability. Committee on Public Health Strategies to Improve Health; Institute of Medicine. Washington, DC: The National Academies Press; 2010.
13.   Behavioral Risk Factor Surveillance System. Office of Surveillance, Epidemiology, and Laboratory Services. [November 1, 2011.]; Available from: http://www.cdc.gov/brfss/.
14.   National Health Interview Survey. National Center for Health Statistics.   [November 1, 2011.]; Available from: http://www.cdc.gov/nchs/nhis.htm.
15.   National Health and Nutrition Examination Survey. National Center for Health Statistics.   [November 1, 2011.]; Available from: http://www.cdc.gov/nchs/nhanes.htm.
16.   Kasteridis P, Yen ST. Smoking Cessation and Body Weight: Evidence from the Behavioral Risk Factor Surveillance Survey. Health Serv Res. Feb 22.
17.   Schousboe JT, Gourlay M, Fink HA, Taylor BC, Orwoll ES, Barrett-Connor E, et al. Cost-effectiveness of bone densitometry among Caucasian women and men without a prior fracture according to age and body weight. Osteoporos Int. Feb 17.
18.   Rothman EF, Sullivan M, Keyes S, Boehmer U. Parents' Supportive Reactions to Sexual Orientation Disclosure Associated With Better Health: Results From a Population-Based Survey of LGB Adults in Massachusetts. J Homosex. Feb;59(2):186-200.
19.   Senechal M, Bouchard DR, Dionne IJ, Brochu M. Lifestyle Habits and Physical Capacity in Patients with Moderate or Severe Metabolic Syndrome. Metab Syndr Relat Disord. Feb 21.
20.   Underwood JM, Townsend JS, Stewart SL, Buchannan N, Ekwueme DU, Hawkins NA, et al. Surveillance of demographic characteristics and health behaviors among adult cancer survivors--Behavioral Risk Factor Surveillance System, United States, 2009. MMWR Surveill Summ. Jan 20;61(1):1-23.
21.   Ladak F, Gjelsvik A, Feller E, Rosenthal S, Montague BT. Hepatitis B in the United States: ongoing missed opportunities for hepatitis B vaccination, evidence from the Behavioral Risk Factor Surveillance Survey, 2007. Infection. Jan 12.

22.   National Health and Nutrition Examination Survey . Survey Operations.   [November 1, 2011]; Available from: http://www.cdc.gov/nchs/nhanes/about_nhanes.htm#operations.

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

24.   MTSamples. [March 1, 2011]; Available from: http://www.mtsamples.com/.

25.   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 Jan-Feb;13(1):30-9.

26.   OpenEHR   Specifications.   [March   1,   2011];   Available   from: http://www.openehr.org/svn/specification/TAGS/Release-1.0.2/publishing/architecture/rm/ehr_im.pdf.