Longitudinal analysis of social and behavioral determinants of health in the EHR: exploring the impact of patient trajectories and documentation practices

Daniel J. Feller, MA[1], Jason Zucker, MD[2], Oliver Bear Don't Walk, MS[1], Michael T Yin[2], MD, Peter Gordon, MD[2], Noémie Elhadad, PhD[1]

[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA; [2]Division of Infectious Diseases, Department of Medicine, Columbia University, New York, NY USA

Abstract

Social and behavioral determinants of health (SBDH) are environmental and behavioral factors that impede disease self-management and can exacerbate clinical conditions. While recent research in the informatics community has focused on building systems that can automatically infer SBDH from the patient record, it is unclear how such determinants change overtime. This study analyzes the longitudinal characteristics of 4 common SBDH as expressed in the patient record and compares the rates of change among distinct SBDH. In addition, manual review of patient notes was undertaken to establish whether changes in patient SBDH status reflected legitimate changes in patient status or rather potential data quality issues. Our findings suggest that a patient's SBDH status is liable to change over time and that some changes reflect poor social history taking by clinicians.

Introduction

Social and behavioral determinants of health (SBDH) are environmental and behavioral factors such as unstable housing and substance use disorders that often impede disease self-management and can lead to or exacerbate existing comorbid conditions.[1–4] Due to the established impact of SBDH on health outcomes for persons living with chronic disease, health systems are increasingly attuned to these determinants and the clinically meaningful information they provide.[5–9] However, evidence suggests that providers often struggle to retrieve information related to SBDH, and that those providers who are unaware of SBDH provide worse quality care.[3,5]

Integrating SBDH data into electronic health records (EHRs) has generated widespread interest among health systems and healthcare providers. The informatics community has focused on three approaches to integrating SBDH into EHRs: expanding data collection to include structured data elements representing SBDH[3], associating community-level SBDH information with individual patients[10,11], and using information extraction and Natural Language Processing (NLP) methods to identify patient-level SBDH information contained in clinical documentation.[6,12,13] SBDH such as smoking status[14–16], substance abuse[17–19], and homelessness[20,21] have been the focus of recent NLP approaches, looking for identifying these at the encounter level, but have to date achieved modest-yet promising performance. However, there has been little work so far on characterizing a patient's SBDH status beyond a single encounter.

It is unlikely that a patient's documented SBDH status is invariably consistent with the patient's true state. Patients are often hesitant to disclose sensitive information such as sexual orientation and substance use to healthcare providers[19,20], and may be less likely to share sensitive information with non-physician providers.[24] In addition, the quality of social history taking by clinicians is variable, and providers are liable to make incorrect assumptions about their patient's health.[25,26] As a result, a patient's SBDH status recorded in the patient record may reflect inaccuracies attributable to phenomena inherent to clinical documentation of sensitive information.

There is little knowledge of how social and behavioral determinants of health as expressed in patient records change through time. While it has been established that an individual's SBDH may change across the life course, it is unclear whether the information in EHRs accurately represent such changes.[27] To our knowledge, the only relevant study that examined changes in SBDH documentation was conducted by Bejan et. al. in 2017, which observed cyclic transitions between the at-risk and homeless categories among homeless patients, and less frequent transitions among individuals with stable housing.[20] In this paper, we conduct a longitudinal analysis of multiple SBDH as expressed in the patient record and attempt to answer several important questions. First, we assess whether SBDH change through time and

estimate their respective rates of change. Second, we conduct a survival analysis to examine the timescale of these changes. Third, we employ both quantitative and qualitative methods to examine potential data qualities issues related to SBDH.

Dataset

In this section, we describe the methods used to curate and analyze a longitudinal gold-standard corpus of clinical notes. A protocol for manual annotation of clinical notes was developed and followed by multiple rounds of annotation. The study described herein was approved by the Institutional Review Board at CUIMC and patient informed consent was waived due to the retrospective nature of the study.

Development of Annotation Guidelines for SBDH

Two physicians experienced in the prevention and treatment of HIV reviewed the biomedical literature to identify social and behavioral determinants that increased an individual's risk for acquiring HIV. 29 individuals risk factors were selected and included sexual orientation, housing status, alcohol use and drug use which are the focus of the study described herein.[28]

We obtained document-level annotations for the aforementioned SBDH domains and risk factors and thus considered all SBDH as binary variables. Annotators were instructed to review the entire length of each clinical document and label the presence of both the high-level domain and granular SBDH label. For example, the phrase *"patient denies alcohol use"* would be result in a positive label for the high-level SBDH domain 'alcohol use documented' but a negative label for 'active alcohol use'. The annotation guide can be downloaded at https://github.com/danieljfeller/SBDSH.

Creation of Gold-Standard Corpus for Longitudinal Analysis

A corpus of clinical notes was obtained from the clinical data warehouse at Columbia University Irving Medical Center (CUIMC), a large academic medical center in New York City. For this study, we obtained all individual notes associated with 32 randomly sampled individuals with HIV who received care at CUIMC. The study cohort was characterized by analyzing structured demographic data extracted from the EHR system at CUIMC. Additional details on this cohort are described elsewhere.[9]

Annotation was conducted by 1 physician and 2 medical students who were instructed to manually review every clinical document included in the patient record. An initial set of 3 longitudinal patient records were collaboratively coded by all annotators for training purposes; 76 notes were annotated by each of the 3 reviews and a Kappa of 0.736 was observed across all SBDH labels. Throughout the annotation process, annotators relied on the guidelines described above.

Creation of Gold-Standard Corpus for Data Quality Analysis

We collected a larger gold-standard corpus for analyzing data quality of SBDH documentation. In contrast to the longitudinal corpus described above, we included in this cohort HIV+ individuals associated with only a portion of their clinical notes annotated and required only that individuals had 2 or more notes with SBDH documentation. These notes were annotated for the purposes of a previous study.[28]

Methods

The analysis described herein focused on 4 SBDH; sexual orientation, housing status, alcohol use and drug use. Encounters with confirmed documentation of SBDH were isolated and analyzed to examine changes in a patient's SBDH status. and potential data quality issues. We then manually reviewed pairs of notes authored on the same-day with conflicting documentation to identify possible sources of data quality issues related to SBDH.

SBDH Preprocessing

A distinct dataset was created for each SBDH of interest and included only notes where the respective high-level SBDH was documented. For example, to be included in the analysis of 'drug use' status, it was necessary that a note discuss drug use (even if drug use was a negative label, as in *"no history of substance use"*). This way, all notes in that SBDH dataset had explicit discussion of that SBDH and either positive or negative findings for that SBDH. Notes that did not discuss the SBDH were not included; absence of a certain SBDH in a clinical note most often reflects the fact that this domain was not discussed by patient and provider, rather than evidence that the patient is a negative case.[29]

Survival Analysis

To assess the rate of change in SBDH status, we simply parameterized each annotated document using sequence time (e.g., visit_time$_1$, visit_time$_2$, etc. where visit_time$_i$ represents the time between the first and i^{th} visit in the longitudinal record of a patient) and estimated the likelihood that a patient at visit$_t$ would transition to a different state (e.g., 'never used alcohol' to 'actively using alcohol'), as documented within the documentation at the next visit. This analysis was conducted using the corpus generated by a comprehensive annotation of the entire longitudinal record of the 32 HIV+ individuals in our cohort.

We used survival analysis to analyze the expected duration of time associated with changes in patient SBDH status. Observation periods were established between adjacent notes in a patient's longitudinal history and time was measured in days. An event was defined as any change in SBDH status observed in the subsequent note. Observations were (right) censored when the subsequent note was observed with the same SBDH status as observed in the preceding (index) note. A survival function was estimated for each SBDH using the Kaplan-Meier estimate and can be interpreted as the fraction of clinical notes observed at time T with an observed change in SBDH status documented in the subsequent note:

$$\widehat{S}(t) = \prod_{i:\ t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

where d_i is the number of notes with subsequent SBDH changes at time t_i and n_i is the number notes not associated with any subsequent SBDH change (and who have not yet been censored) at time t_i.

Quantitative Analysis of Data Quality

In order to assess data quality in SBDH documentation, we used a larger cohort of patients who were required only to have multiple notes manually annotated. Using the parameterization of sequence time described above, we considered *illegitimate transitions* to be those that were chronologically impossible; for example, a patient could be documented as having 'never used alcohol' *subsequent* to being documented as 'actively using alcohol'.

In addition, we identified same-day conflicts in SBDH documentation by isolating clinical notes that were written on the same day by distinct providers. Similar to previous analyses, we required that all notes under consideration have confirmed documentation of the relevant high-level SBDH domain. Same-day conflicts were defined as observed discrepancies in SBDH status (e.g., documentation of *"no active alcohol use"* and *"active alcohol use"*).

Qualitative Analysis of Data Quality

We manually reviewed 20 note pairs that exhibited same-day conflicts in patient SBDH status with the goal of developing an understanding of the sources of data quality problems. 5 note pairs were gathered from each of the four SBDH analyzed herein. The sources of data quality issues were identified using a set of annotation guidelines we created for this analysis. Data quality issues reflected 1) *inappropriate use of copy & paste* if content was duplicated across notes of the same type, 2) an *inaccurate problem list* if outdated information contained in structured clinical data was automatically inserted into narrative free-text, 3) *variable history taking* wherein the note with positive documentation contained significantly more information regarding SBDH compared to the note with negative documentation (e.g. provider documents a minimal social history and does not inquire about specific substances, for example") , 4) a *patient hesitant to disclose sensitive information* when it was clear that both notes contained a detailed social history but that the patient gave conflicting answers across the notes, and 5) the use of a *standard note template* which automatically inserted negative and formulaic documentation of SBDH status.

Results

Cohort Characteristics

3273 clinical documents associated with 32 HIV+ individuals were manually annotated and included in the longitudinal corpus. All available clinical notes associated with these patients was annotated with a range of 11 to 473 notes per patients (median 50, mean 102). The longitudinal histories of patients in this analysis were of variable lengths and ranged from 196 days to 3146 days; the mean longitudinal history was 997 days and the median was 772 days.

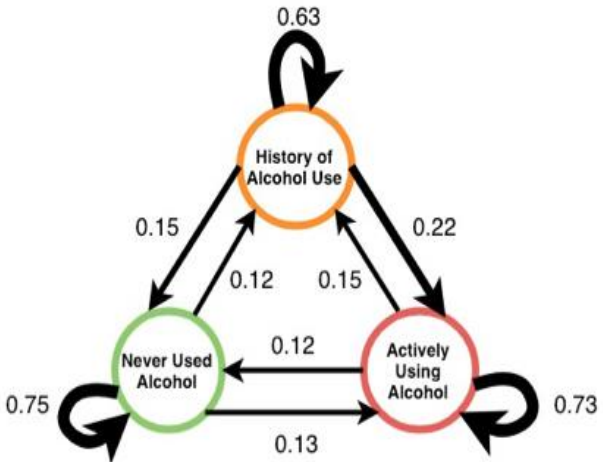

Figure 1. State diagram of changes in documentation of alcohol use across all patients in the cohort.

75% of individuals were male (24) and the average age was 46 years with a standard deviation of 13.5 years. Race and ethnicity information was missing for a majority of the cohort, but 8 patients were documented as African American, 7 as Caucasian Hispanic, 2 as Caucasian non-Hispanic, and 1 as Asian.

The larger cohort used in the data quality analysis included 366 individuals with multiple annotated notes (4294 notes total). All available clinical notes associated with these patients was annotated with a range of 2 to 473 notes per patients (median 2, mean 11). The longitudinal histories of patients in this analysis were of variable lengths and ranged from 0 days to 3146 days; the mean longitudinal history was 469 days and the median was 174 days. 60.3% of individuals were male (221) and the average age was 53 years with a standard deviation of 12.2 years. Race and ethnicity information was missing for a majority of the cohort, but 71 patients were documented as African American, 58 as Caucasian Hispanic, 30 as Caucasian non-Hispanic, and 1 as Asian.

Longitudinal Changes in SBDH

A state diagram illustrating changes of documentation status for alcohol use computed across the cohort of 366 individuals is presented in Figure 1, across 1077 pairs of consecutive notes.

Many patients in the longitudinal cohort of 32 individuals were observed to have temporal changes among all SBDH analyzed. Alcohol status was most likely to change across subsequent notes (23.3%), followed by drug use (10.4%), drug use (8.7%), and sexual orientation (1.1%). A chi-square goodness of fit test found a significant difference in these proportions (χ^2: 29.2, $p < 0.001$).

A Kaplan-Meier plot that presents the 365-day survival function of the 4 SBDH analyzed in this study in presented in Figure 2. The y-axis represents the proportion of notes without changes in SBDH documentation, and the x-axis represents the number of days between each subsequent note. 365-day transition rates for housing status, drug use, and alcohol use were 39.6%, 30.6%, and 36.3%, respectively. The 365-day transition rate for sexual orientation was 6.9%.

Data Quality in SBDH Documentation

While most of the transitions in SBDH status looked sensible, we observed illegitimate transitions in SBDH status as shown in Figure 1. Among 353 notes documenting patient SBDH status as 'active alcohol use, 43 were followed by documentation of 'never used alcohol' (12.2%). Moreover, among 165 notes documented that the patient had 'historical alcohol use', 26 were followed by documentation of 'never used alcohol' (15.7%).

We also observed same-day conflicts in patient SBDH status. For instance, 23.2% of 56 same-day note pairs with alcohol status documented had conflicting indications of alcohol use, 21.2% of 52 same-day note pairs had conflicting indications of substance use, 6.8% of 44 same-day note pairs had conflicting indications of patient housing status, and 8.0% of 25 same-day note pairs had conflicting of sexual orientation. We also observed conflicts in SBDH documentation within a 7-day period. 12.6% of annotated notes associated with the same patient had conflicting

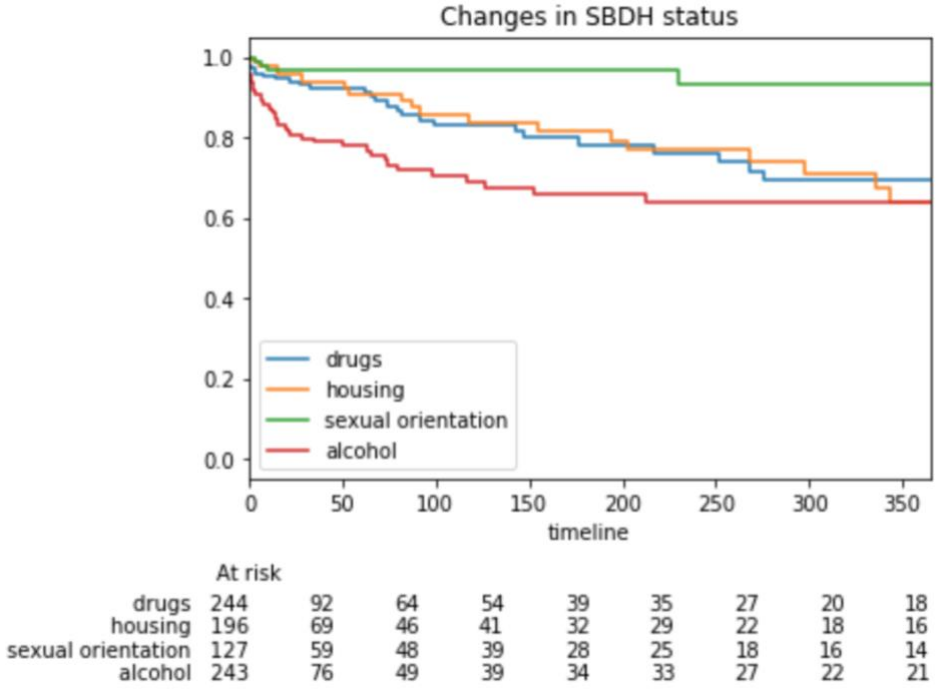

Figure 1. Survival function of SBDH changes for drug use, housing, sexual orientation, and alcohol use.

indications of alcohol use (N = 230). In addition, 14.3% of notes describing drug use exhibited the same changes (N = 265), as did 7.6% of notes describing housing status (N=198). 1.6% of notes with sexual orientation documentation exhibited same-week conflicts (N = 127).

Manual review of 20 same-day conflicts in SBDH revealed multiple sources of poor data quality with incomplete social history taking by clinicians being most common. Among the 20 manually reviewed note pairs with same-day conflicts in SBDH, 14 (70%) conflicts reflected variable quality of social history taking by clinicians, wherein the note with positive documentation contained significantly more detailed information regarding SBDH compared to the note with negative documentation. The majority of such instances entailed particularly brief social histories that may have reflected inaccurate assumptions and/or limited social history taking by healthcare providers. 2 note pairs exhibited evidence of an inaccurate problem list, wherein note content automatically generated from the patient's EHR problem list conflicted with information contained in narrative free-text. 2 note pairs exhibited evidence that a patient was hesitant to disclose SBDH, as it was clear that both notes contained a detailed social history but that the patient gave conflicting answers across the notes. 1 note pair conflict reflected the use of a standard note template, which automatically declared negative SBDH status.

Discussion

The findings of this study suggest that social and behavioral determinant of health as expressed in the patient record exhibit changes over time. Our longitudinal analysis of 4 distinct risk-factors suggests that a patient's SBDH status should be treated as a shifting, mutable variable in electronic systems. We also provide evidence that some changes in SBDH documentation reflect data quality issues and not actual changes in the patient state.

We present the first comprehensive longitudinal analysis of multiple SBDH as expressed in patient records. Four distinct SBDH were examined throughout the course of 32 patient records and exhibited varying rates of change. A patient's recorded alcohol status was most likely to change, as 23.3% of adjacent encounters with documentation of alcohol use contained conflicting information. This may reflect the high prevalence of heavy episodic drinking among persons living with HIV, which has resulted in calls for repeated assessments of alcohol consumption in this population.[30,31] While documentation of substance abuse and housing status were less likely to change compared to alcohol use, these SBDH exhibited changes across as much as 10% of adjacent encounters. Epidemiological studies suggest that these SBDH are likely to change; most individuals who are considered with unstable housing experience only transient periods of homelessness[27], and many HIV+ persons with substance abuse disorder engage in episodic

rather than sustained use.[32,33] These findings suggest that automated approaches to inferring SBDH should not treat these variables as fixed patient characteristics and thus should reevaluate a patient's SBDH status on a regular basis.

In contrast to alcohol use, substance abuse, and housing status, a patient's recorded sexual orientation was unlikely to change, as 1.1% of adjacent encounters with documentation of sexual orientation contained conflicting information. This rate reflects the infrequent changes in sexual orientation observed among persons experiencing stigma and discrimination.[34] This findings suggest that distinct SBDH are likely to change at different rates and thus may be reevaluated on different time scales.

Multiple findings indicated that some changes in patient SBDH status reflected data quality issues and not legitimate changes in the patient state. We observed a high frequency of implausible longitudinal changes in patient SBDH status, wherein a patient transitioned from an active status (e.g. active alcohol use) to having no history of active status (e.g. no history of alcohol use). In addition, we observed same-day conflicts in patient's documented SBDH status. Our manual review of these discrepancies observed that most conflicts reflected the variable nature of social history taking. It has been established that some healthcare providers are reluctant to discuss sensitive issues with their patients, thereby limiting their ability to take a comprehensive social history.[35] The implication of these data quality issues is that any attempt to characterize a patient's SBDH status should not merely reflect the most recent documentation. Decision support systems that aggregate multiple instances of SBDH documentation may provide a more faithful representation of a patient's SBDH status compared to data collected during a single encounter. Information retrieval and classification methods should utilize observation windows that leverage only recent EHR data[36,37], or weight decay techniques that model the decreasing relevance of data elements to computational phenotypes.[38,39]

Future research should conduct a more comprehensive analysis by annotating all notes associated with a large corpus and use techniques such as mutual information to assess how the predictability of future SBDH status relative to existing documentation changes with time.[40] Such an analysis would open the possibility for techniques that could accurately model the relevance of social and behavioral determinants of health documentation.[41,42] In addition, more research is needed to better estimate the likelihood that changes in SBDH documentation reflect true changes in the patient state versus documentation errors. This could be accomplished by obtaining additional sources of data such as validated psychometric instruments or laboratory tests indicative of substance abuse.

Limitations

First, our methods did not enable us to quantify the proportion of SBDH changes that reflected true changes in the patient state and the proportion that reflected documentation errors. A key limitation of this study is our inability to estimate whether a patient's SBDH status is accurate and note that obtaining this information is likely not possible using clinical notes alone. Second, while our annotators achieved a relatively high inter-rater reliability, there were likely some erroneous annotations and thus some temporal changes in SBDH status may reflect annotation errors and rather than changes in documented SBDH status. Third, our findings were generated by analyzing data from a specific patient cohort treated at a single institution. The high prevalence of SBDH within the study cohort may have resulted in a higher frequency of SBDH changes.

Conclusion

Social and behavioral determinants of health as documented in the patient record are subject to change through time. Future approaches to automated inference of SBDH from EHRs should consider the fluid nature of these variables and the impact of data quality on their methods and results.

References

1. Adler, N. E. & Stead, W. W. Patients in Context — EHR Capture of Social and Behavioral Determinants of Health. *N. Engl. J. Med.* **372**, 698–701 (2015).
2. Gottlieb, L. M., Wing, H. & Adler, N. E. A Systematic Review of Interventions on Patients' Social and Economic Needs. *Am. J. Prev. Med.* **53**, 719–729 (2017).
3. Gottlieb, L. M., Tirozzi, K. J., Manchanda, R., Burns, A. R. & Sandel, M. T. Moving electronic medical records upstream: incorporating social determinants of health. *Am. J. Prev. Med.* **48**, 215–218 (2015).

4. Chung, E. K. *et al.* Screening for Social Determinants of Health Among Children and Families Living in Poverty: A Guide for Clinicians. *Curr. Probl. Pediatr. Adolesc. Health Care* **46**, 135–153 (2016).
5. Weir, C. R. *et al.* A qualitative evaluation of the crucial attributes of contextual Information necessary in EHR design to support patient-centered medical home care. *BMC Med. Inform. Decis. Mak.* **15**, (2015).
6. Navathe, A. S. *et al.* Hospital Readmission and Social Risk Factors Identified from Physician Notes. *Health Serv. Res.* (2017). doi:10.1111/1475-6773.12670
7. Feller, D. J. & Agins, B. D. Understanding Determinants of Racial and Ethnic Disparities in Viral Load Suppression. *J. Int. Assoc. Provid. AIDS Care* **16**, 23–29 (2017).
8. Feller, D. J., Akiyama, M. J., Gordon, P. & Agins, B. D. Readmissions in HIV-Infected Inpatients: A Large Cohort Analysis. *JAIDS J. Acquir. Immune Defic. Syndr.* **71**, 407–407 (2016).
9. Feller, D. J., Zucker, J., Yin, M. T., Gordon, P. & Elhadad, N. Using Clinical Notes and Natural Language Processing for Automated HIV Risk Assessment. *J. Acquir. Immune Defic. Syndr. 1999* (2017). doi:10.1097/QAI.0000000000001580
10. Cantor, M. N., Chandras, R. & Pulgarin, C. FACETS: using open data to measure community social determinants of health. *J. Am. Med. Inform. Assoc.* **25**, 419–422 (2018).
11. Cantor, M. N. & Thorpe, L. Integrating Data On Social Determinants Of Health Into Electronic Health Records. *Health Aff. (Millwood)* **37**, 585–590 (2018).
12. Chen, E. S., Manaktala, S., Sarkar, I. N. & Melton, G. B. A Multi-Site Content Analysis of Social History Information in Clinical Notes. *AMIA. Annu. Symp. Proc.* **2011**, 227–236 (2011).
13. Walsh, C. & Elhadad, N. Modeling Clinical Context: Rediscovering the Social History and Evaluating Language from the Clinic to the Wards. *AMIA Summits Transl. Sci. Proc.* **2014**, 224–231 (2014).
14. McCormick, P. J., Elhadad, N. & Stetson, P. D. Use of semantic features to classify patient smoking status. *AMIA Annu. Symp. Proc. AMIA Symp.* 450–454 (2008).
15. Uzuner, Ö., Goldstein, I., Luo, Y. & Kohane, I. Identifying Patient Smoking Status from Medical Discharge Records. *J. Am. Med. Inform. Assoc. JAMIA* **15**, 14–24 (2008).
16. Savova, G. K., Ogren, P. V., Duffy, P. H., Buntrock, J. D. & Chute, C. G. Mayo clinic NLP system for patient smoking status identification. *J. Am. Med. Inform. Assoc. JAMIA* **15**, 25–28 (2008).

17. Yetisgen, M. & Vanderwende, L. Automatic Identification of Substance Abuse from Social History in Clinical Text. in *SpringerLink* 171–181 (Springer, Cham, 2017). doi:10.1007/978-3-319-59758-4_18

18. Carter, E. W., Sarkar, I. N., Melton, G. B. & Chen, E. S. Representation of Drug Use in Biomedical Standards, Clinical Text, and Research Measures. *AMIA Annu. Symp. Proc. AMIA Symp.* **2015**, 376–385 (2015).

19. Carrell, D. S. *et al.* Using natural language processing to identify problem usage of prescription opioids. *Int. J. Med. Inf.* **84**, 1057–1064 (2015).

20. Bejan, C. A. *et al.* Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records. *J. Am. Med. Inform. Assoc. JAMIA* (2017). doi:10.1093/jamia/ocx059

21. Gundlapalli, A. V. *et al.* Extracting Concepts Related to Homelessness from the Free Text of VA Electronic Medical Records. *AMIA. Annu. Symp. Proc.* **2014**, 589–589 (2014).

22. Brooks, H. *et al.* Sexual orientation disclosure in health care: a systematic review. *Br. J. Gen. Pract. J. R. Coll. Gen. Pract.* **68**, e187–e196 (2018).

23. Gerbert, B. *et al.* When asked, patients tell: disclosure of sensitive health-risk behaviors. *Med. Care* **37**, 104–111 (1999).

24. Teixeira, P. A., Gordon, P., Camhi, E. & Bakken, S. HIV Patients' Willingness to Share Personal Health Information Electronically. *Patient Educ. Couns.* **84**, e9–e12 (2011).

25. Haidet, P. & Paterniti, D. A. 'Building' a history rather than 'taking' one: a perspective on information sharing during the medical interview. *Arch. Intern. Med.* **163**, 1134–1140 (2003).

26. Behforouz, H. L., Drain, P. K. & Rhatigan, J. J. Rethinking the Social History. *N. Engl. J. Med.* **371**, 1277–1279 (2014).

27. Kuhn, R. & Culhane, D. P. Applying cluster analysis to test a typology of homelessness by pattern of shelter utilization: results from the analysis of administrative data. *Am. J. Community Psychol.* **26**, 207–232 (1998).

28. Feller, D. J. *et al.* Towards the Inference of Social and Behavioral Determinants of Sexual Health: Development of a Gold-Standard Corpus with Semi-Supervised Learning. *AMIA Annu. Symp. Proc. AMIA Symp.* **2018**, 422–429 (2018).

29. López Pineda, A. *et al.* Comparison of machine learning classifiers for influenza detection from emergency department free-text reports. *J. Biomed. Inform.* **58**, 60–69 (2015).

30. Bertholet, N., Cheng, D. M., Samet, J. H., Quinn, E. & Saitz, R. ALCOHOL CONSUMPTION PATTERNS IN HIV-INFECTED ADULTS WITH ALCOHOL PROBLEMS. *Drug Alcohol Depend.* **112**, 160–163 (2010).

31. Bensley, K. M. *et al.* Patterns of Alcohol Use Among Patients Living With HIV in Urban, Large Rural, and Small Rural Areas. *J. Rural Health Off. J. Am. Rural Health Assoc. Natl. Rural Health Care Assoc.* (2018). doi:10.1111/jrh.12326

32. Brunet, L., Napravnik, S., Heine, A. D., Leone, P. A. & Eron, J. J. Longitudinal opioid use among HIV-infected patients, 2000 to 2014. *J. Acquir. Immune Defic. Syndr. 1999* **75**, 77–80 (2017).

33. Lauby, J. *et al.* Get Real: Evaluation of a Community-Level HIV Prevention Intervention for Young MSM Who Engage in Episodic Substance Use. *AIDS Educ. Prev.* **29**, 191–204 (2017).

34. Everett, B. Sexual Orientation Identity Change and Depressive Symptoms: A Longitudinal Analysis. *J. Health Soc. Behav.* (2015). doi:10.1177/0022146514568349

35. Wimberly, Y. H., Hogben, M., Moore-Ruffin, J., Moore, S. E. & Fry-Johnson, Y. Sexual history-taking among primary care physicians. *J. Natl. Med. Assoc.* **98**, 1924–1929 (2006).

36. Mo, H. *et al.* Desiderata for computable representations of electronic health records-driven phenotype algorithms. *J. Am. Med. Inform. Assoc. JAMIA* **22**, 1220–1230 (2015).

37. Rasmussen, L. V. *et al.* Design patterns for the development of electronic health record-driven phenotype extraction algorithms. *J. Biomed. Inform.* **51**, 280–286 (2014).

38. Zhou, J., Wang, F., Hu, J. & Ye, J. From Micro to Macro: Data Driven Phenotyping by Densification of Longitudinal Electronic Medical Records. in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 135–144 (ACM, 2014). doi:10.1145/2623330.2623711

39. Lipton, Z. C., Kale, D. C. & Wetzel, R. C. Phenotyping of Clinical Time Series with LSTM Recurrent Neural Networks. *ArXiv151007641 Cs* (2015).

40. Hripcsak, G., Albers, D. J. & Perotte, A. Parameterizing time in electronic health record studies. *J. Am. Med. Inform. Assoc. JAMIA* **22**, 794–804 (2015).

41. Chen, J. H., Alagappan, M., Goldstein, M. K., Asch, S. M. & Altman, R. B. Decaying relevance of clinical data towards future decisions in data-driven inpatient clinical order sets. *Int. J. Med. Inf.* **102**, 71–79 (2017).

42. Dasu, T. & Weiss, G. Mining Data Streams. in *Encyclopedia of Data Warehousing and Mining* (2009).