Modeling Clinical Context: Rediscovering the Social History and Evaluating Language from the Clinic to the Wards

Colin Walsh, MD, Noémie Elhadad, PhD
Department of Biomedical Informatics, Columbia University, New York, NY

Abstract

Social, behavioral, and cultural factors are clearly linked to health and disease outcomes. The medical social history is a critical evaluation of these factors performed by healthcare providers with patients in both inpatient and outpatient care settings. Physicians learn the topics covered in the social history through education and practice, but the topics discussed and documented in real-world clinical narrative have not been described at scale. This study applies large-scale automated topic modeling techniques to discover common topics discussed in social histories, to compare those topics to the medical textbook representation of those histories, and to compare topics between clinical settings to illustrate differences of clinical context on narrative content. Language modeling techniques are used to consider the extent to which inpatient and outpatient social histories share in their language use. Our findings highlight the fact that clinical context and setting are distinguishing factors for social history documentation, as the language of the hospital wards is not the same as that of the ambulatory clinic. Moreover, providers receive little feedback on the quality of their documentation beyond that needed for billing processes. The findings in this study demonstrate a number of topics described in textbooks – schooling, religion, alternative health practices, stressors, for example - do not appear in social histories in either clinical setting.

Introduction

Increasing attention has been paid in the last decade to the importance of social and cultural factors with respect to health and disease. The Institute of Medicine in 2006 released its report, "Genes, Behavior, and the Social Environment", which called for "transdisciplinary, collaborative research" regarding "key social variables" affecting health outcomes such as education, income, social networks, and work conditions.[1] Behavioral, social and environmental factors have been linked to a panoply of issues including overall mortality, chronic diseases such as heart failure, and mental disorders such as rates of suicide.[2–5]

The routine history and physical examination in both inpatient and ambulatory settings is intended to query patients about social, behavioral, and environmental factors. Physicians in training learn the requisite components of a complete social history from medical textbooks on history-taking and the physical exam (Figure 1).[6] But the quality and comprehensiveness of these histories vary in clinical practice. Social histories can be collected in a heterogeneous manner depending on clinical setting and provider-patient rapport. Actual documentation of those histories can range from structured, coded data to electronic free-text to pen and paper charting.

Personal and Social History

- **Occupation**
- **Last year of schooling**
- **Home situation and significant others**
- **Sources of stress, both recent and long term**
- **Important life experiences, such as military service**
- **Leisure activities**
- **Religious affiliation and spiritual beliefs**
- **Activities of daily living (ADLs)**
- **Lifestyle habits**
  - **Exercise and diet**
  - **Safety measures**
  - **Alternative health care practices**

Figure 1. Textbook components of personal and social history.

Assessing quality of documentation remains a challenge to conduct at scale. While physicians are taught to conduct comprehensive histories to capture a holistic view of patient wellness, the gold standard is an individual-level chart

review with either the patient or physician as unit of analysis. Methods to permit large-scale semi-automated assessment of clinical narrative could have implications in the domains of quality of care and of practice improvement. Both clinical housestaff trainees and licensed practitioners who have moved on from the supervisory environment of residency programs might benefit from such tools. Documentation of social history, both because it relies on narrative and it has no established best-practices outside of the textbook guidelines, is in need of robust, scalable content-analysis methods.

The underlying research questions driving this study stem from the need to understand the common topics conveyed in a large collection of real-world social histories as well as the parity of those histories to "textbook teaching" in medical education. A related question considers what – if any – qualitative and quantitative differences exist between the language of social histories taken in inpatient and outpatient clinical settings

The internal medicine service at Columbia University Medical Center incorporates both inpatient hospital wards and ambulatory clinic settings and as such provides an opportunity to answer these research questions at scale. Both clinical contexts comprise hundreds of physicians with thousands of patient encounters per year. Techniques of topic modeling and of language modeling are used to quantify differences between those settings. Topic modeling is an attractive approach to discover aspects of social histories in a large collection of notes; it is unsupervised and as such remains robust to the idiosyncrasies of clinical language; it does not need any gold-standard annotation. Language modeling permits quantification of the extent to which inpatient and outpatient social histories differ in their lexical choice, again, in an unsupervised and robust fashion.

Background

Prior studies have considered the content of social histories in clinical text in public datasets and in small samples of clinical notes.[7,8] Melton et al performed content analysis of social and behavioral history datasets in the public health domain and showed an emphasis on smoking, alcohol use, drug use, and employment data.[7] Chen et al performed manual evaluation of the content of a small number of social histories contained in clinical notes and mapped them to HL7 and openEHR standards.[8] Our study is situated in the framework of unsupervised learning from a large corpus of clinical notes through both topic and language modeling.

Topic modeling is an established method, which, given a collection of documents and a pre-defined number of topics (K), identifies ranked lists of words (topics) according to which the documents can be described.[9,10] It operates as a type of dimensionality reduction from the entire vocabulary of documents to the K topics. Topic modeling has been widely used in non-clinical fields of natural language processing (NLP) and is gaining purchase in clinical natural language processing.[11–15] While these methods have been applied to clinical text, they have not been applied to the task of topic modeling of social histories in free-text clinical notes.

Language modeling is an older technique applied in multiple areas of language technology.[16] Given a training set, it assesses the likelihood of a new string of text (e.g., a sentence or a document) in a testing set. By comparing the likelihood of one text – one test set – against language models derived from different training sets, it is possible to determine the combination of training set and test set that are closest in lexical pattern.

Methods

Dataset

After obtaining approval of the Institutional Research Board, a dataset was collected of electronic clinical text taken from the electronic health record at Columbia University Medical Center from 2005-2009 for inpatient documentation and from 2008-2009 for outpatient documentation. The disparity between years relates to the increased use of the "Primary Provider Clinic Note" Type following 2008. No other clinic note type clearly identified the author as primary care provider, and this identification was necessary for the intended comparison between admitting inpatient physician and outpatient primary care provider. Two note types were extracted: "Medicine Admission Notes", the preferred electronic note completed by residents and attending physicians on the internal medicine service during the study period; "Primary Provider Clinic Note", the preferred note used by ambulatory care physicians who identified themselves as primary providers for that encounter. Both note types were written in a single blank text box without section headings or coded data entry. Any section headings in the notes were written by providers including abbreviations and shorthand (e.g. "Social History", "SocHx", "SH").

To avoid the bias introduced by including multiple admission notes from a single patient, only the first admission note for each unique medical record number was included in the test corpus. This bias has previously been described.[11,15] Similarly, the phenomenon of "copy-paste" or "copy-forward" in which providers copy text from previously written notes was avoided by the selection of the initial admission note for each unique patient record.

Data Preprocessing
The clinical text was imported into the R statistical environment and relevant admission and primary provider notes were identified. Free-text entries (entire admission and primary provider notes) were parsed into XML with section headings corresponding to content headings in notes, e.g. "History of Present Illness", "Medications", "Physical Exam". The social history sections were extracted from the XML output into a single file and then parsed further at the level of sentences.[17,18] All text was normalized to lowercase, and punctuation was removed. Overall, the dataset consisted of individual social history sections with sentence boundary information.

Topic Modeling
Topic modeling assumes that a document is a mixture of topics, and a topic is a mixture of words. The study goal was to capture the aspects of social histories that are common across a large population of patient records. To coax the topic modeling towards discovering these aspects, sentences in social history sections were considered to be individual documents. The hypothesis is that a sentence is a mixture of topics to be discovered.
Stop words and artifactual characters ("h" as part of "h o" or "history of") were removed in formatting text for latest Direchlet Allocation (LDA). Two sub-corpora were created – all inpatient social histories 2005-2009 and all outpatient social histories 2008-2009 – and LDA was applied to each independently.

A practicing internal medicine physician reviewed each set of topics manually. Topics were mapped whenever possible to one or several textbook topics (as described in Figure 1).

Perplexity Analysis of Inpatient- and Outpatient-Derived Language Models
Topic modeling allowed identification of aspects of inpatient and outpatient social histories. Language modeling and perplexity analysis enabled quantification of whether the language of the two settings differ (even if they cover similar topics). Perplexity invokes the concept of cross entropy to compare two different language models using a test set because the true underlying probability distribution in a corpus is unknown. Cross entropy is a step to quantifying the uncertainty in modeling language in a corpus imperfectly; the lower the cross entropy (and the perplexity which relies on it), the better the fit of the model to the data.

Two years of clinical text data were isolated for perplexity analysis; one training set of inpatient text was constructed from 2008 inpatient social histories, and one training set of outpatient text was constructed from 2008 outpatient social histories. The inpatient and outpatient social histories from 2009 comprised the two testing sets. Two N-gram language models were trained, one inpatient and one outpatient. Perplexity was calculated within and between clinical settings, e.g. the inpatient language model was tested against both inpatient and outpatient testing sets. The lower the perplexity, the better the lexical fit between unseen text and the training set. Thus, when comparing the perplexity of one corpus against two different models, the set with the lowest perplexity represented the training set that matched the unseen text more closely.

Experimental Setup
For experiments with topic modeling, variational inference implementation of LDA was used.[19] The default parameters were used including random topic initialization, variational inference set to converge algorithmically, and a maximum of one hundred iterations of expectation maximization to estimate hyperparameters of $\propto$ and β. The number of topics was set manually and various numbers of topics were tested: ten, twenty, thirty. Results are reported for ten topics only.

For experiments with language modeling, the Stanford Research Institute Language Modeling (SRILM) toolkit was used with a trigram model with Kneser-Ney smoothing.[20,21]

Results

From 2005-2009, the clinical data repository contained 64,610 notes of type "Medicine Admission Note". After removal of duplicate medical record numbers (implying subsequent admissions) and text parsing to identify sections, the Inpatient Social History Corpus was composed of 48,944 documents. From 2008-2009, the repository contained 15,154 notes of type "Primary Provider Clinic Note". After removal of duplicates and parsing, the Outpatient Social History Corpus contained 7,796 documents.

Topic Modeling

Table 1. Inpatient Social History Topic Models

Topic	Word Clusters (rank order of words, ten words shown)	Topic Label (manually assigned)
1	etoh, drugs, denies, not, yrs, but, home, tob, never, worked	Lifestyle Habits, Employment
2	live, ny, ppd, nl, boyfriend, stds, remote, still, new, Washington	Home Situation, Sexual History
3	works, wife, last, us, marijuana, heavy, only, grandchildren, weekends, prev	Family, Lifestyle Habits
4	lives, quit, her, husband, illicits, since, son, pt, alcohol, mother	Family, Lifestyle Habits
5	use, now, work, independent, former, factory, none, age, history, old	Employment
6	dr, ago, sexually, working, adls, one, here, illicit, occasional, drinks	Background, Employment, Lifestyle Habits
7	tobacco, years, daughter, active, currently, alone, drug, smoked, past, she	Smoking (Lifestyle Habit)
8	no, children, came, worked, but, beer, this, iadls, clear, vices	Family, Lifestyle Habits, Activities of Daily Living
9	day, gt, retired, moved, disability, does, ivdu, sexual, school, months, two, uses	Employment, Support, Lifestyle Habits, Education
10	smoking, used, social, nyc, who, cocaine, stopped, family, separated, alone	Lifestyle Habits (smoking, drugs), Social Support

Table 2. Outpatient Social History Topic Models

Topic	Word Clusters (rank order of words, ten words shown)	Topic Label (manually assigned)
1	now, work, home, pt, adls, independent, kids, all, care, iadls	Family, Activities of Daily Living
2	not, active, currently, sexually, working, last, hx, does, hiv, sexual	Sexual History
3	denies, use, drug, alcohol, illicit, other, habits, up, wt, bp	Lifestyle Habits
4	children, dr, her, married, here, born, moved, nyc, sister, living	Family, Support
5	years, ago, quit, yrs, smoking, social, smoked, occasional, ppd, cocaine	Lifestyle Habits (smoking, illicit drugs), Preventive Testing
6	wife, worked, since, us, hha, retired, previously, his, factory, came	Employment
7	but, she, one, he, who, time, separated, old, well, father	Family
8	lives, daughter, works, husband, alone, son, mother, unemployed, two, Bronx	Family, Employment

9	day, used, never, past, gt, former, year, week, per, disability	Modifiers
10	no, etoh, drugs, tobacco, illicits, tob, ted, cigs, rare, rrr	Lifestyle Habits

Results of topic modeling are presented Tables 1 and 2. The inpatient social history topics included multiple representations of lifestyle habits – specifically smoking, alcohol use, illicit drug use. The outpatient social history topics also included lifestyle habits as well as topics more clearly composed of family, social support, sexual history, and employment.

In comparison to the textbook version of the social history presented in Table 1, it is clear that lifestyle habits – tobacco, alcohol, illicit drug use, specifically – are the most commonly reflected aspects of social history within inpatient clinical text. Inpatient social histories allude to topics of family, social support, and education, for example, but words in these categories are mixed with other aspects of the social history. The outpatient social histories demonstrate aspects in common with inpatient histories, but the aspects are more clearly defined and more likely to be topics composed entirely of a single aspect. Outpatient Topic #2 for example includes multiple terms related to sexual history and HIV testing while Inpatient Topic #2 mingles words associated with sexual history ("boyfriend", "stds" for sexually transmitted diseases) with other aspects including preventive testing ("ppd") and living environment ("lives", "ny" for New York). Textbook topics that were not apparent in either inpatient or outpatient social histories include: schooling, religious affliation, significant stressors, major life experiences, alternative health practices.

Perplexity Analysis of Inpatient and Outpatient Social History Language Models

Table 3 summarizes the results of perplexity calculation using training sets of inpatient 2008 and outpatient 2008 social history text on testing sets of inpatient 2009 and outpatient 2009 social history text. Out-of-vocabulary words (OOVs) are included as are the numbers of sentences and words in each training corpus.

OOVs from inpatient to outpatient settings are much larger across settings than within the same setting (60K unknown words vs. 26K) and, similarly, inpatient social histories are better modeled by the inpatient language model than the outpatient one (202.9 vs. 331.9 perplexity). Language of outpatient social histories is also better modeled according to an outpatient language model (123.3 vs. 385.8 perplexity). This fact remains even though the number of OOVs in the outpatient test set outnumbers the OOVs in the inpatient set when tested by the outpatient language model (18K vs. 16K).

Table 3. Testing across clinical contexts (inpatient to outpatient and vice versa) reveals higher perplexity.

Training Set	Testing Set	# Sentences in Training Set	# Words in Training Set	OOVs in Testing Set	Perplexity of Testing Set
Inpatient 2008	Inpatient 2009	51,978	373,390	26,075	202.9
Inpatient 2008	Outpatient 2009	51,978	373,390	60,767	385.8
Outpatient 2008	Inpatient 2009	29,764	219,513	16,805	331.9
Outpatient 2008	Outpatient 2009	29,764	219,513	18,175	123.3

Discussion

The principal findings of this study demonstrate that large-scale topic modeling can reveal topics of interest within a chosen section of clinical text. Inpatient social histories correspond to topics of interest to physicians in an intermediate acuity clinical setting. Lifestyle habits – particularly tobacco, alcohol, and illicit drug use – as well as basics of employment and family support are frequently queried areas within social histories. However, these topics in inpatient text are frequently mixed within sentences and are therefore not clearly demarcated. Outpatient social history text, on the other hand, reveals more clearly delineated topic areas as well as more complete social histories being taken by primary care providers. This result corresponds intuitively with clinical teaching and with an

encounter in a lower acuity care setting in which continuity of care is more obviously anticipated. Both topic models failed to reveal some topics emphasized as important to a holistic view of a patient's wellness. The social history topics did not include education or schooling necessary to approximate health literacy, religious affiliation that might inform patients' treatment decisions or end-of-life choices, or alternative health practices that may affect patients' health overall (e.g. alternative weight loss therapies or alternative therapeutics that might interact with prescribed medications). It remains an open question as to whether clinical histories should reflect the textbook closely or whether clinical workflows in practice are appropriately different than traditional teaching.

The perplexity analysis reveals another important finding. The language of the wards is not the same as that of the clinic. Language models derived from a particular clinical context are prone to more entropy and higher computation requirements when they are applied across care settings. This result has important implications for language modeling as clinical context should be considered in the development and evaluation of novel tools in this space. Applications of these findings could fall into the domains of documentation quality, medical education, or billing. Utilizing content modeling through LDA and language models such as those described here might enable physicians to identify and correct deficiencies in clinical notes for quality purposes or to maximize the level of billing for the exams that they perform. Medical students could be prompted to ask patients about as-yet undocumented topics in notes as they learn to take comprehensive histories. Clinical context should be taken into account to optimize the performance of requisite language models in all of these areas.

Strengths of this study include a large amount of real-world clinical text as well as the use of entirely free-text documentation as the original dataset. It demonstrates the strength of automated methods to assign and extract sections from otherwise unformatted blocks of text in which a variety of abbreviations and shorthand are common (e.g. "Social hx", "SocHx", "SH"). The use of unsupervised methods, which do not assume any semantic information about the words in the text, enabled identification of highly meaningful words with respect to social history that would otherwise be difficult to extract from text automatically. For instance, "dr" in this analysis referred not to doctor, but to Dominican Republic, a common country of origin for many of the patients in this study population. While there was no gold standard available in this corpus, textbook information was leveraged to validate and compare the results of topic modeling.

Limitations of this study include a larger dataset for inpatient data than outpatient data as a result of the note types selected for study. Similarly, notes were selected from 2005-2009 and not more recent years because of changes in documentation practice in both care settings since 2009. Free-text documentation without templates or other coded data is now much less common, but such semi-structured data demonstrates less clearly the flexibility and robustness of the natural language processing methods used here. The noise of clinical data is another limitation common to all methods incorporating clinical text. The perplexity analysis was limited by the asymmetry of sizes of corpora for training – a result of the difference in number of visits between inpatient and outpatient settings. It is possible that some elements of social history were discussed outside of the sections labeled "Social History" by providers (such as History of Present Illness), but these were filtered out at the corpus processing stage. And because this analysis took place at corpus-level, there may be low-frequency relevant topics in individual notes outside of social histories sections; these were not the subject of this investigation.

This study extends the work of others by demonstrating that topics of clinical text can be rediscovered through unsupervised machine learning. It also demonstrates the computational impact of modeling language without concern for clinical context. The lower perplexity associated with testing language models within opposed to across clinical settings has important implications for subsequent endeavors in this discipline. Finally, it suggests a system of practice and documentation improvement could discover areas of deficiency in documentation in real-world clinical text. Providers do not currently have dedicated mechanisms of documentation quality review outside of that necessary to accomplish billing tasks. This method could be one step to the creation of such a system to help providers improve at a task they perform multiple times per day without feedback – documentation of the clinical encounter.

Future research can extend this technique to discover topics in other sections of the clinical note or in other note types entirely. Other provider types (surgeons, nurse practitioners, emergency medicine physicians) also incorporate different lexicons in their clinical narratives; these methods could be applied to those contexts to reveal aspects of that language. Language models can be employed for specific levels of care and care settings. Clinical experience

dictates that the language of the intensive care unit varies from the medical wards, for example, and this technique could be one step to verify and subsequently overcome that obstacle. The results of this study could also inform further data modeling including predictive analytics and the extraction of free-text data for incorporation into coded datasets.

Conclusion

Factors of social support, cultural facets, and behaviors are important components of the medical history collected in both inpatient and outpatient encounters. The high quantity of clinical text – often unstructured free-text – requires automated methods that can elucidate underlying aspects of those histories and pave the way for secondary use, quality assessment, clinical training, and decision support incorporating such data in a structured way. This study demonstrated the aspects of social history from both inpatient and outpatient clinical encounters. Inpatient encounters are typified by shorter social histories with an emphasis on immediately cogent behaviors such as smoking, alcohol use, and illicit drugs. Outpatient social histories include greater breadth and depth of topics covered, but neither type of social history touched on significant aspects of patient wellness such as schooling, religious affiliation, or life stressors. Perplexity analysis suggests strongly that clinical context and setting must be incorporated in the design of computational methods to process and benefit from the clinical narrative.

Acknowledgements

This work is supported by T15 LM007079 (CW) and National Library of Medicine award R01 LM010027 (NE).

References

1. Institute of Medicine. Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate. 2006.

2. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. JAMA : the journal of the American Medical Association. 2004 Mar 10;291(10):1238–45.

3. Kleiman EM, Liu RT. Social support as a protective factor in suicide: Findings from two nationally representative samples. Journal of affective disorders. Elsevier; 2013 Mar 2

4. De la Cámara AG, Guerravales JM, Tapia PM, Esteban EA, Del Pozo SVF, Sandubete EC, et al. Role of biological and non biological factors in congestive heart failure mortality: PREDICE-SCORE: a clinical prediction rule. Cardiology journal. 2012 Jan;19(6):578–85.

5. Rasic DT, Belik S-L, Elias B, Katz LY, Enns M, Sareen J. Spirituality, religion and suicidal behavior in a nationally representative sample. Journal of affective disorders. 2009.

6. Bickley LS, Szilagyi PG, Bates B. Bates' guide to physical examination and history-taking / Lynn S. Bickley, Peter G. Szilagyi. 11th ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2013. p. xxv, p. 994

7. Melton GB, Manaktala S, Sarkar IN, Chen ES, Health C. Social and Behavioral History Information in Public Health Datasets. Proceedings of the AMIA Annual Symposium. 2012. pp. 625–34.

8. Chen ES, Manaktala S, Sarkar IN, Melton GB. A Multi-Site Content Analysis of Social History Information in Clinical Notes. Proceedings of the AMIA Annual Symposium. 2011. pp. 227–36.

9. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003;3:993–1022.

10. Blei D. Probabilistic Topic Models. Communications of the ACM. 2012 Nov;77–84.

11. Arnold C, Speier W. A Topic Model of Clinical Reports. 2012;90024:1031–2.

12. Salleb-aouissi A, Radeva A, Passonneau RJ, Xie B, Khattak FK, Tomar A, et al. Diving into a Large Corpus of Pediatric Notes. Proceedings of the ICML Workshop on . 2011;

13. Perotte A, Bartlett N, Elhadad N, Wood F. Hierarchically Supervised Latent Dirichlet Allocation. Proceedings of the Neural Information Processing Systems Conference (NIPS). 2011;1–9.

14. Halpern Y, Horng S, Nathanson LA, Shapiro NI. A Comparison of Dimensionality Reduction Techniques for Unstructured Clinical Text. Proceedings of the ICML Workshop on Clinical Data Analysis. 2012;

15. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC bioinformatics. 2013 Jan 16. 14(1):10.

16. Manning CD, Schütze H. Foundations of Statistical Natural Language Processing. May 1999.

17. Li Y, Lipsky Gorman S, Elhadad N. Section classification in clinical notes using supervised hidden markov model. Proceedings of the ACM international conference on Health informatics (IHI). 2010. pp. 744.

18. Tsuruoka Y, Tateishi Y, Kim J, Ohta T. Developing a Robust Part-of-Speech Tagger for Biomedical Text. Lecture Notes in Computer Science. 2005. pp. 382–92.

19. Blei DM. Latent Dirichlet Allocation in C. Available from: http://www.cs.princeton.edu/~blei/lda-c/

20. Chen SF, Goodman J. An empirical study of smoothing techniques for language modeling. Computer Speech & Language. 1999 Oct;13(4):359–93.

21. Wang W. SRILM at Sixteen : Update and Outlook. Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop. 2011.