Nurse documentation of sexual orientation and gender identity in home healthcare: A text mining study

Computers, Informatics, Nursing

Nurse documentation of sexual orientation and gender identity in home healthcare: A text mining study

--Manuscript Draft--

Manuscript Number:	CIN-D-17-00098R1
Full Title:	Nurse documentation of sexual orientation and gender identity in home healthcare: A text mining study
Article Type:	Peer Reviewed Article
Keywords:	Text mining; nurse documentation; home health care; LGBT health
Corresponding Author:	Ragnhildur I Bjarnadottir, MPH, Ph.D University of Florida Gainesville, FL UNITED STATES
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University of Florida
Corresponding Author's Secondary Institution:	
First Author:	Ragnhildur I Bjarnadottir, MPH, Ph.D
First Author Secondary Information:	
Order of Authors:	Ragnhildur I Bjarnadottir, MPH, Ph.D
	Walter O. Bockting, PhD
	Sunmoo Yoon, PhD
	Dawn W. Dowding, PhD
Order of Authors Secondary Information:	
Manuscript Region of Origin:	UNITED STATES
Abstract:	Health disparities have been documented in the LGBT population, but more research is needed to better understand how to address them. To that end, this observational study examined what is documented about sexual orientation and gender identity in narrative home care nurses' notes in an electronic health record. Lexical text mining approaches were used to examine a total of 862,715 clinical notes from 20,447 unique patients who received services from a large home care agency in Manhattan, New York, and extracted notes were qualitatively reviewed to build a lexicon of terms for use in future research. Forty-two notes, representing 35 unique patients, were identified as containing documentation of the patient's sexual orientation or gender identity. Documentation of sexual orientation or gender identity was relatively infrequent, compared to the estimated frequency of LGBT people in the US population. Issues related to fragmentary language emerged, and variety in phrasing and word frequency was identified between different types of notes and between providers. This study provides insight into what nurses in home healthcare document about sexual orientation and gender identity, their clinical priorities related to such documentation, and provide a lexicon for use in further research in the home care setting.

Cover Letter

College of Nursing
Department of Family, Community
and Health System Science

1225 Center Drive
PO Box 100197
Gainesville, FL 32610-0197
352-273-6508

Leslie H. Nicoll,
PhD, MBA, RN, FAAN
Editor-in-Chief
Computers Informatics Nursing

Dear Dr. Nicoll

We are pleased to submit a revised manuscript of our original research article entitled "Nurse documentation of sexual orientation and gender identity in home healthcare: A text mining study" by Ragnhildur I. Bjarnadottir, Walter Bockting, Sunmoo Yoon and Dawn Dowding for consideration for publication in Computers Informatics Nursing.

All authors have substantially contributed to this manuscript and its conception and design, analysis, drafting the article and revising it critically for important intellectual content.

This manuscript has not been published and is not under consideration for publication elsewhere. We have no conflicts of interest to disclose. All authors have reviewed the manuscript and approved it for submission.

Please address all correspondence to the undersigned.

Thank you for your consideration,

Sincerely,

Ragnhildur Bjarnadottir

Ragnhildur I. Bjarnadottir

Title Page (including all author information)

Nurse documentation of sexual orientation and gender identity in home healthcare: A text mining study

Ragnhildur I. Bjarnadottir, MPH, PhD, RN[a,b], Walter Bockting, PhD[a, c], Sunmoo Yoon, PhD, RN[a], Dawn W. Dowding, PhD, RN[a, d, e]

[a] Columbia University School of Nursing, 630 West 168th Street, Mail Code 6, New York, NY 10032, USA
[b] University of Florida, College of Nursing, 1225 Center Drive, PO Box 100197, Gainesville, FL, 32608, USA
[c] New York State Psychiatric Institute, 1051 Riverside Drive, Unit 15, New York, NY 10032, USA
[d] Center for Home Care Policy & Research, Visiting Nursing Service of New York, 1250 Broadway, 7th floor, New York, NY, 10001, USA
[e]Division of Nursing, Midwifery and Social Work, School of Health Sciences, University of Manchester, Manchester, M13 9PL, UK

Corresponding author:
Ragnhildur I. Bjarnadottir
University of Florida College of Nursing
1225 Center Drive, PO Box 100197, Gainesville, FL, 32608
Phone: +1 352-273-6508
Fax: +1 352-273-6505
Email: rib@ufl.edu

Conflicts of Interest and Source of Funding:

The authors have no conflicts of interest or sources of funding to disclose for this study.

Acknowledgements

We thank the Visiting Nurses Service of New York, as well as Drs. Suzanne Bakken and Robert Lucero for invaluable help and support in the conception, design and analysis of this study.

We would like to thank the editor and reviewers for their thoughtful review and feedback.

Reviewer #1	Response
1. There are numerous spelling, grammar, and writing errors noted, particularly in the front matter of the manuscript. It reads like someone else wrote the Intro and Background and someone else wrote the methods, results, and discussion.	Thank you for your thorough review. We have reviewed the paper and amended spelling, grammar, and writing errors.
2. The article is of interest from as a methodological approach, but I found it was written more for a clinician audience (especially the abstract/intro/background) and not for informatics professionals. This could be improved by limiting the background on information on the population of interest and more on the methods and tools used. Also, what are some practical applications of these methods for informatics professionals?	Thank you, we have reduced the background to include less details on the population of interest and more on the gap in knowledge related to text mining in nurses' notes. For practical applications of these methods for informatics professionals we include the following implications: *"The resulting list of n-grams can be used as a lexicon for future research. Further research should focus on applying and evaluating the lexicon in different settings and adding to its comprehensiveness. The findings also highlight a need for ways of mapping terms related to sexual orientation and gender identity to standardized terminologies in documentation in a way that is meaningful, comprehensive and culturally competent."*
3. I found the results quite tedious to read through. It was very mechanistic and adding some reference to the significance of some of the results would be helpful and more interesting to read. E.g. "The relative frequency of 'LGBT' in the narrative notes was 9.2*10-4 %, compared to 8*10-6 % in the reference corpus." Although it appears there is a difference, I don't know if it is high or low, significant, etc. I know some of this is discussed later in the paper, but some type reference range text might be helpful.	Thank you for your suggestion, we have added some clarifying language in this section and restructured sentences for a clearer read.
4. You state the design of the study as Bag-of-words and in the abstract state it is a text mining study. This is not a research design, it is a method of data extraction. The design in this study is thus not clear.	Thank you for pointing this out. We have amended our design section to indicate that this is a retrospective observational cross-sectional design where text mining approaches were utilized for text extraction and retrieval.
5. Figure1 is in black and white and it is hard to distinguish the differences in the graphs. You might use stronger and lighter shading to make the differences clear or submit them in color.	Thank you, we have amended Figure 1 to sbe in color.
6. Table 1 be consistent with the spacing of your subheadings (I would center sexual orientation and gender identify), also the text under LOINC to the rights looks off like it is right justified.	We have amended table 1 so that the headings are centered and the items below are all left justified.
7. Table 2 separate the headings sexual orientation and gender identity from the list of items with a bar like you did in table 1.	We have added a border to be consistent with table 1.
8. Table 3 spacing and justification/center of subheadings, separate the n-gram tables from each other. You can write, Table 3 (continued) or something of that nature. They are confusing all lumped together.	We have adjusted the spacing and justification and separated the n-gram sections from each other.
Reviewer #2	

In the abstract, sentence 2 there is an extra word - either aimed or examine but not both.	Thank you for your thoughtful review, we have amended this to read *"To that end, the this study examined what is documented about sexual orientation and gender identity in narrative home care nurses' notes in an electronic health record"*
Page 1 - last paragraph - Sentence 1 was confusing. This needs clarification. This paragraph is unclear regarding unstructured data. The term other sources is used and it is unclear if this refers to nursing notes or other types of data.	Thank you, we have amended this paragraph for clarification. The section now reads: *"While widespread implementation of EHRs in health care provides increasing availability of routinely collected electronic data, a large portion of this data is unstructured and complex to extract and analyze.[13] Despite continuing efforts to "standardize clinical documentation, clinicians continue to greatly rely on unstructured or narrative data[12], meaning that up to 75% of available clinical data is unstructured.[13] Managing and utilizing these largely unstructured data, which include nursing notes, comes with challenges, and innovative solutions are needed."*
Page 4- paragraph 2- Clarify sentence 3. The structure of the sentence makes it difficult to understand.	We have amended this paragraph to read: "*Using the AutoMap software's built-in packages, the first step of data cleaning was fix common typographical errors in the text. Next, all numbers, symbols, stop words and noise words were removed. However, pronouns were retained due to their potential significance in examining gender identity. Consequently all text was converted to upper case to remove the issue of case sensitivity."*
Page 6- paragraph 1 - sentence 3- There seems to be a word missing.	Thank you we have added the missing word 'to'. The sentence now read: *"After manual review 21 notes were excluded; 14 due to the word 'gay' appearing as a proper noun rather than in reference to sexual orientation and seven due to errors in pronoun use."*
Page 10 paragraph 1 - sentence 4 - Check the tenses; they don't agree.	Thank you, we have amended the sentences to ensure the whole paragraph is written in present tense.
Page 11- paragraph 1- sentence 4 Once again, review the tenses, they all need to be the same.	We have amended this sentence so that the entire section is in the present tense.
Page 12 - paragraph 2 - sentence 5- there are extra words. The last sentence in this paragraph is cumbersome and needs to be clarified.	Thank you for noting cumbersome structure of this section. We have revised it to say: *"Comparison to the reference corpus reveals that a majority of the n-grams identified have a higher relative frequency in our clinical corpus compared to the reference corpus, despite the apparent growth in literature and public discourse on LGBT issues. This highlights the uniqueness of the nursing language compared to contemporary literature and public discourse. This may further indicate a perceived clinical relevance of this data among home care nurses."*

Abstract

Health disparities have been documented in the LGBT population, but more research is needed to better understand how to address them. To that end, this observational study examined what is documented about sexual orientation and gender identity in narrative home care nurses' notes in an electronic health record.

Lexical text mining approaches were used to examine a total of 862,715 clinical notes from 20,447 unique patients who received services from a large home care agency in Manhattan, New York, and extracted notes were qualitatively reviewed to build a lexicon of terms for use in future research.

Forty-two notes, representing 35 unique patients, were identified as containing documentation of the patient's sexual orientation or gender identity. Documentation of sexual orientation or gender identity was relatively infrequent, compared to the estimated frequency of LGBT people in the US population. Issues related to fragmentary language emerged, and variety in phrasing and word frequency was identified between different types of notes and between providers.

This study provides insight into what nurses in home healthcare document about sexual orientation and gender identity, their clinical priorities related to such documentation, and provide a lexicon for use in further research in the home care setting.

Keywords: Text mining; nurse documentation; home health care; LGBT health

Background and Significance

Significant health disparities have been documented in the LGBT population, but more research is needed to better understand the mechanism behind them and how they can best be addressed.[1–4] In the United States, home care agencies serve 4.9 million Americans a year,[5] and the majority of elderly people requiring long term care services (80%) receive them in the home.[6] Similarly, the home care sector across Europe is growing in both size and importance, but remains vastly understudied.[7]

Experts have pointed out the role of clinical data and documentation in electronic health records (EHR) in expanding the knowledge of LGBT health issues, as evidenced by the Institute of Medicine's (IOM) call to incorporate sexual orientation and gender identity into routine assessment and data collection in healthcare[1,8] and similar recommendations by the Council of Europe.[9] Collecting clinical data in healthcare using EHR's improves the structure and process of such data collection, and may also improve patient outcomes directly, as well as provide a rich source of data for research and clinical decisions support.[10,11]

While widespread implementation of EHRs in health care provides increasing availability of routinely collected electronic data, a large portion of this data is unstructured and complex to extract and analyze.[12] Despite continuing efforts to standardize clinical documentation, clinicians continue to greatly rely on unstructured or narrative data,[13] meaning that up to 75% of available clinical data is unstructured.[12] Managing and utilizing these largely unstructured data, which include nursing notes, comes with challenges, and innovative solutions are needed.

To address this data challenge, there is increasing interest in automated or semi-automated methods, such as text-mining, to analyze clinical text data. In recent studies these methods have shown promise for medical record reviews and retrospective identification of

various adverse health outcomes.[14–17] However, nurses' notes and other nurse-generated data have been largely overlooked as a data source in these studies.[18] There is a need to develop methods to capture the large body of data that already exists in the form of nurses' notes and explore the potentially valuable information it may contain about LGBT patients and their health and care needs.

Objective

This study aimed to examine what is documented about sexual orientation and gender identity in narrative home care nurses' notes in an electronic health record.

Methods

Design

This was a retrospective observational cross-sectional study using text mining approaches. Text mining is a subset of data mining that utilizes a set of computational techniques for retrieval and analysis of human language, aiming to extract and represent meaning for free or unstructured text[19,20]. Prior research has demonstrated that text mining can be effective in identifying data from narrative clinical notes.[19,21–27] Due to the complexity and level of ambiguity in clinical narratives, text mining is the most commonly utilized method to retrieve text information from clinical records.[20,28]

This study utilized a bag-of-words method with n-gram based text retrieval. The bag-of-words is one of the most commonly used methods for text representation and categorization [29]. With this method, the text documents are represented as a multi-set, or so-called bag, and the grammar and word sequence are disregarded. This allows for counting frequencies of words or concepts in text and representing the text quantitatively as vectors. This was an appropriate design for this study, as the aim was to explore the highly understudied topic of sexual

orientation and gender identity documented in nurse narratives in home care patients' electronic health records.

The text mining procedure was performed in the following steps: 1) Data selection, 2) Preprocessing, 3) Transformation, 4) Application of data mining algorithm and 5) Interpretation.[30]

Data corpus and selection

The data for this study was obtained from a large not-for-profit home healthcare provider in the United States with a diverse patient population across New York. The data corpus comprised of nursing narratives from three types of nurses notes; referral, narrative and coordination of care notes. Referral notes are documented at first referral to the agency, during the intake visit. Narrative notes are documented during each visit, when the nurse obtains information that is perceived as important but not captured in structured data in the EHR. Coordination of care notes are used to document the coordination of care with other healthcare and service providers. All notes in the data corpus were documented by home care nurses in the organization's electronic health records system for all patients receiving care in the latest available full year, (2015) in the borough of Manhattan (N=20,447). The borough of Manhattan was selected based on the high density of members of the LGBT community, compared to other boroughs across the New York metropolitan area.[31] This was considered most feasible under the assumption that this would also result in more density of LGBT patients in the data corpus.

Data cleaning and preprocessing

Clinical texts, such as nurses' notes, are generally considered noisy and irregular data,[32] partly due to common typographical errors, abbreviations and fragmentary language.[33] For this reason, and due to the volume of the data, thorough data cleaning and preprocessing is a key step

in the data mining process. This step serves to format the data to a more computer-readable form for further analysis. The Intellij integrated development environment for Java and the AutoMap software were used for data cleaning and preprocessing.

Using the AutoMap software's built-in packages, the first step of data cleaning was fix common typographical errors in the text. Next, all numbers, symbols, stop words and noise words were removed. However, pronouns were retained due to their potential significance in examining gender identity. Consequently all text was converted to upper case to remove the issue of case sensitivity.

Following this, a stemmer was applied in AutoMap to reduce dimensionality of the data. Stemming aims to reduce any inflectional forms of words to their word stems or base forms.[34] This study used the Krovetz stemmer as the main goal of stemming in this case was time efficient data reduction, given that the stemming processes were not likely to effect the keywords or n-gram of interest in this study.

Transformation

Following preprocessing, the AutoMap software was used to transform all text into n-grams, to be used for text categorization. An n-gram is a sequence of a certain number of words or characters from a larger string.[35] Examples of n-grams related to gender identity would be 'transgender' (unigram), 'transgender male' (bigram) and 'male to female' (trigram). This study utilized a combination of unigrams, bigrams and trigrams, which has been found to yield higher accuracy in text categorization, compared to the use of only one type of n-gram.[36]

Application of data mining algorithm

Following the transformation step, a search algorithm was constructed using Intellij integrated development environment with Java to extract n-grams of potential relevance to the

sexual orientation or gender identity of patients. To construct the search algorithm, keywords and phrases were identified based on previously conducted qualitative interviews [37] as well as an examination of commonly used medical terminologies and lexicons and exploration of how sexual orientation and gender identity are coded in these (Table 1). These keywords and phrases included terms related to sexual orientation, sexual and gender identity and/or expression and sexual behavior to ensure that all relevant documentation would be identified. However, it should be noted that these concepts are in no way synonymous or interchangeable. Search terms included in the search algorithm are displayed in Table 2. The search process was iterative, with certain terms added or removed based on search results. Each retrieved note was manually reviewed to assess the context of the identified keywords and determine whether they accurately identified the documentation of a patients' sexual orientation or gender identity. The traditional mining techniques of frequency counts and visualization [38] were then employed to summarize the findings. Finally, the relative frequency of each n-gram was compared to the relative frequency of that n-gram in a reference database. The database utilized for reference was the Google Books n-gram viewer, which allows for the search of n-grams in Google's text corpora, consisting of sources printed between the years 1500 and 2008.[39] The purpose of this was to examine what the relative frequency of n-grams related to sexual orientation or gender identity in nurses' narrative notes is compared to narratives from other fields, such as history, art and humanities. This reference database was therefore used as a proxy for public discourse.

Interpretation

The interpretation stage comprises of an evaluation of findings to determine if the data mining process can be terminated or if further iterations are needed. [30] In this study, the data

mining process was terminated once the iterative search process no longer yielded additional results.

Results

The data corpus comprised of 20,447 referral notes, 234,788 coordination of care notes and 607,480 narrative notes from 20,477 unique patients. A total of 63 notes were identified that contained documentation related to patients' sexual orientation or gender identity. Forty-two notes remained that contained documentation of patients' sexual orientation or gender identity. These consisted of 11 referral notes, 24 narrative notes and seven coordination of care notes. These notes represented 35 unique patients. Eleven patients were identified from referral notes, 2 patients from coordination of care notes and 23 patients from narrative notes. One of the 35 patients was identified in two different types of notes, narrative and coordination of care notes. Of the 35 patients identified as having documentation in their record related to sexual orientation or gender identity, 22 were lesbian, gay or bisexual, 6 were transgender and seven were heterosexual.

Table 3 displays the unigrams, bigrams and trigrams related to sexual orientation or gender identity that were identified through the text mining process. Nine unique unigrams, seventeen unique bigrams and twelve unique trigrams were identified. Of these, seven unique unigrams, eleven bigrams and eight trigrams, were represented in the narrative notes. Coordination of care notes yielded no unigrams or trigrams related to sexual orientation or gender identity, and only two unique bigrams. Four unique unigrams, seven bigrams and five trigrams were represented in the referral notes. Figure 1 shows a comparison of the frequency of n-grams between notes. The n-grams can be broadly classified into five categories: 1) sexual orientation terms, 2) terms on gender identity or expression, 3) terms related to relationships and

family, 4) terms related to sexual behaviors and 5) terms referring to supportive services (Table 3). These categories will be discussed further in the following sections.

Sexual orientation

Five unigrams related to sexual orientation were identified, as well as one bigram and one trigram. The most commonly occurring n-gram related to sexual orientation was the unigram 'LGBT', which stands for lesbian, gay, bisexual and transgender. This unigram occurred nine times in the corpus, although it was exclusively represented in the narrative notes. The relative frequency of 'LGBT' in the narrative notes was higher than in the reference corpus ($9.2*10^{-4}$ % compared to $8*10^{-6}$ %). A manual review revealed that the acronym frequently occurred in relation to community resources or supportive services tailored to the LGBT population.

The unigram 'heterosexual' followed in frequency, occurring six times in referral notes and once in the narrative notes. The relative frequency of this unigram in the referral notes was higher than in the reference corpus ($8.6*10^{-3}$ %, compared to $4.2*10^{-4}$ %). The remaining unigrams in this category were relatively infrequent, occurring once or twice and exclusively in narrative notes.

Gender identity and expression

Three unigrams, six bigrams and five trigrams were identified related to patients' gender identity or gender expression. The most frequently occurring n-gram in this category was the unigram transgender, occurring four times in the referral notes and three times in the narrative notes. The relative frequency in the notes was not substantially higher than in the reference corpus ($5.7*10^{-4}$ % and $3.1*10^{-4}$ %, respectively, compared to $1.0*10^{-4}$ % in the reference corpus).

Four of the bigrams and two of the trigrams represented different phrasing or denotation of the transgender individual's sex and gender, including 'transgender ftm' or 'female to male', to indicate that a patients had been assigned female sex at birth but identified as male gender.

Relationships and family

Two bigrams were identified that referenced the relationships and family of patients. Both referred to female patients and their spouses, either girlfriend or wife. The bigram 'her wife' occurred three times in the coordination of care notes and seven times in narrative notes, and had a relative frequency of $2.7*10^{-4}$ % and $7.2*10^{-4}$ %, respectively, which was substantially higher than the reference corpus ($3.0*10^{-6}$ %). Similarly, the bigram 'her girlfriend' occurred more frequently in the nurses' notes ($3.6*10^{-4}$ % in coordination of care notes and $1.0*10^{-4}$ % in narrative notes, compared to $1.0*10^{-5}$ % in the reference corpus).

Sexual behaviors

Two bigrams and one trigram were identified related to sexual behaviors. In all instances, the sexual behaviors documented were specifically heterosexual sexual activity. Manual review revealed that in all cases, the patient in question was HIV-infected and the documentation of heterosexual sexual activity referred to how transmission occurred. No n-grams were identified related to sexual behaviors or activity with same-sex partners.

Supportive services

The category of supportive services for members of the LGBT community was only represented in narrative notes. One unigram, six bigrams and five trigrams were identified in the text. Of these n-grams, all but one were in reference to the services offered by Services and Advocacy for GLBT Elders (SAGE). Through manual review, it emerged that this was

documented to note that the patient in question had been referred to these services, or was already connected with them. In addition to SAGE, one note documented the use of services at Callen-Lorde Community Health Center, which specializes in healthcare and services targeted to New York's lesbian, gay, bisexual, and transgender communities.[40]

Discussion

To the author's knowledge, no other study has been conducted using natural language processing to examine the documentation of sexual orientation or gender identity in home care nurses' notes. The findings provide insight into how nurses document information about their patients' sexual orientation or gender identity, and provide a lexicon of n-grams for use in further studies on this topic in the home care setting.

Text mining approach and issues

Findings of this study highlight previously documented issues related to the analysis of unstructured text, such as the issue of fragmentary language.[33] This is perhaps best exemplified in the great variation that emerged when a transgender gender identity was documented. A transgender individual assigned male sex at birth but identifying as female was denoted in the unstructured text as 'transgender m-f', 'transgender mtf' and 'transgender male to female', all referring to the same concept but varying based on the provider conducting the documentation. The variation in terminology creates ambiguity and makes the development of an efficient yet comprehensive lexicon challenging. This highlights the importance of continued efforts to develop and implement standardized terminologies for nurse documentation. While great strides have been made to implement and consolidate standardized nursing terminologies [41], standard terminology related to the documentation of sexual orientation and gender identity is lacking. The Department of Health and Human Services and the Office of the National Coordinator for

Health Information Technology (ONC)[42,43] have called for the addition of standardized terms related to sexual orientation and gender identity to be incorporated into the SNOMED CT nomenclature and HL7 standards, but have not yet been added. Future research should examine how to map commonly utilized terms to standard terminologies that will support comprehensive and culturally competent documentation of patients' sexual orientation and gender identity.

Experts have begun examining how best to ask questions about sexual orientation or gender identity in the clinical setting in order to obtain comprehensive information [8]. However, there is a need for ways of mapping the varied and dynamic terms and phrasing of these questions into a standardized nursing terminologies on the documentation side. Such mapping requires input from the clinicians conducting the documentation, such as nurses. The n-grams extracted in this study may provide some insight into the language nurses are most comfortable with using in their documentation, but further research is needed.

Identification of LGBT patients

Despite a large data set and an extensive, iterative search process, relatively few instances of documentation of sexual orientation and gender identity emerged. We identified 28 LGBT patients where sexual orientation or gender identity were documented, in a dataset of 20,477 patients, or around 0.1%. Contrastingly, around 3.8% of the US population are estimated to identify as LGBT (Gates, 2011). This indicates that sexual orientation and gender identity is likely only documented in a small portion of those patients who identify as LGBT. This is consistent with findings from previous qualitative studies, which found that nurses were reluctant to discuss and document their patients' sexual orientation and gender identity.[37,44] However, the search algorithm constructed iteratively in this study was able to comprehensively identify

patients from the records and the resulting lexicon can be used in future research to identify cohorts of LGBT patients for use in health disparities research.

Emphasis in documentation

In instances where sexual orientation was documented, the focus appeared to be on documenting demographic information, such as gender, behaviors that resulted in risk or infection, patient's relationships with spouses and caregivers and relevant community resources. These focus areas are well aligned with the main goals of home care nursing, to promote health, improve function and assist patients to remain at home.[45] Accurate demographic information and assessment of potentially risky behaviors are key to ensuring optimal outcomes and informing patient education and the understanding and identification of informal caregivers, supportive relationships and community resources can improve the individual's ability to avoid hospitalization and remain in the home.[46] This is also consistent with a previous qualitative study among home care nurses, where nurses express an emphasis on documentation informing care and practice, and mainly see clinical relevance of sexual orientation or gender identity data in relation to risky behaviors or caregiver support.[37] Interestingly, a majority of instances where sexual orientation or sexual behaviors was documented in this study were referring to heterosexual activity. This may further highlight a discomfort or the perceived sensitivity of the information when patients are engaged in same-sex relationships or sexual activity.

Comparison between types of notes

As shown in Figure 1, the frequency of n-grams varied greatly between different types of notes. A majority of the n-grams were represented in the narrative notes, which may indicate that discussions about sexual orientation or gender identity are more likely to come up further into the home care episode, rather than at first referral. This is consistent with findings from

qualitative interviews, where nurses expressed the importance of building trust and rapport with the patient before broaching a sensitive topic such as sexual orientation and gender identity.[37] There were however several n-grams that were not represented in narrative notes and only came up in the referral notes. This highlights the importance of tailoring data mining processes to the specific text being analyzed. Different search algorithms may be more or less effective for different sets of notes, and may therefore need to be specifically tailored, particularly if specificity is a priority.

Comparison to public discourse

Comparison to the reference corpus reveals that a majority of the n-grams identified have a higher relative frequency in our clinical corpus compared to the reference corpus, despite the apparent growth in literature and public discourse on LGBT issues. This highlights the uniqueness of the nursing language compared to contemporary literature and public discourse. This may further indicate a perceived clinical relevance of this data among home care nurses.

Limitations

This study has limitations worth noting. Firstly, only one year of data was used to limit the volume of the data analyzed. Including data from a longer period might strengthen the study by providing more data and potentially the emerging of further relevant n-grams. Secondly, this study was only conducted using data from one home care agency and in one borough of Manhattan. It cannot be assumed that the findings are generalizable across different settings or geographical locations. Finally, the method used was a knowledge-based approach, relying on an a priori list of search terms to use in the data mining process, and the list may therefore not have been exhaustive. Despite these limitations, the lexicon developed based on the findings can serve as a base or foundation for future research, to be further developed and improved upon.

Conclusions

The findings of this study provide insight into what nurses in home healthcare document in patient records about sexual orientation and gender identity and their priorities related to such documentation. The resulting list of n-grams can be used as a lexicon for future research. Further research should focus on applying and evaluating the lexicon in different settings and adding to its comprehensiveness. The findings also highlight a need for ways of mapping terms related to sexual orientation and gender identity to standardized terminologies in documentation in a way that is meaningful, comprehensive and culturally competent.

Human Subjects Protections

This study was performed in compliance with the International Ethical Guidelines for Biomedical Research Involving Human Subjects and was reviewed by Institutional Review Boards at the university and the health care agency included in this study

References

1. Institute of Medicine (US) Committee on Lesbian, Gay, Bisexual, and Transgender Health Issues and Research Gaps and Opportunities. *The Health of Lesbian, Gay, Bisexual, and Transgender People: Building a Foundation for Better Understanding*. Washington (DC): National Academies Press (US); 2011. http://www.ncbi.nlm.nih.gov/books/NBK64806/. Accessed November 13, 2013.

2. McKay B. Lesbian, Gay, Bisexual, and Transgender Health Issues, Disparities, and Information Resources. *Med Ref Serv Q*. 2011;30(4):393-401. doi:10.1080/02763869.2011.608971

3. Molina Y, Lehavot K, Beadnell B, Simoni J. Racial Disparities in Health Behaviors and Conditions Among Lesbian and Bisexual Women: The Role of Internalized Stigma. *LGBT Health*. 2014;1(2):131-139. doi:10.1089/lgbt.2013.0007

4. Fredriksen-Goldsen KI, Kim H-J. Count Me In Response to Sexual Orientation Measures Among Older Adults. *Res Aging*. July 2014:0164027514542109. doi:10.1177/0164027514542109

5. Harris-Kojetin L, Sengupta M, Park-Lee E. *Long-Term Care Providers and Services Users in the United States: Data from the National Study of Long-Term Care Providers, 2013–2014*. Hyattsville, MD: National Center for Health Statistics; 2016:1-118. http://www.cdc.gov/nchs/data/series/sr_03/sr03_038.pdf.

6. Congressional Budget Office. *Rising Demand for Long-Term Services and Supports for Elderly People*. Washington, D.C.: Congress of the United States; 2013. https://www.cbo.gov/sites/default/files/113th-congress-2013-2014/reports/44363-LTC.pdf.

7. Genet N, Boerma W, Kroneman M, Hutchinson A, Saltman RB, eds. *Home Care across Europe. Current Structure and Future Challenges*. United Kingdom: European Observatory on Health Systems and Policies; 2011. http://www.euro.who.int/en/about-us/partners/observatory/publications/studies/home-care-across-europe.-current-structure-and-future-challenges. Accessed August 23, 2016.

8. Institute of Medicine (US) Board on the Health of Select Populations. *Collecting Sexual Orientation and Gender Identity Data in Electronic Health Records: Workshop Summary*. Washington (DC): National Academies Press (US); 2013. http://www.ncbi.nlm.nih.gov/books/NBK132859/. Accessed June 15, 2015.

9. Council of Europe. *Discrimination on Grounds of Sexual Orientation and Gender Identity in Europe*. 2nd ed. Strasbourg: Council of Europe; 2011.

10. Holroyd-Leduc JM, Lorenzetti D, Straus SE, Sykes L, Quan H. The impact of the electronic medical record on structure, process, and outcomes within primary care: a systematic review of the evidence. *J Am Med Inform Assoc JAMIA*. 2011;18(6):732-737. doi:10.1136/amiajnl-2010-000019

11. Kalra D, Fernando B, Morrison Z, Sheikh A. A review of the empirical evidence of the value of structuring and coding of clinical information within electronic health records for direct patient care. *Inform Prim Care*. 2012;20(3):171-180.

12. Capurro D, PhD MY, van Eaton E, Black R, Tarczy-Hornoch P. Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multisite assessment. *EGEMS Wash DC*. 2014;2(1):1079. doi:10.13063/2327-9214.1079

13. Bigeard E, Jouhet V, Mougin F, Thiessard F, Grabar N. Automatic extraction of numerical values from unstructured data in EHRs. *Stud Health Technol Inform*. 2015;210:50-54.

14. Simmons M, Singhal A, Lu Z. Text Mining for Precision Medicine: Bringing Structure to EHRs and Biomedical Literature to Understand Genes and Health. *Adv Exp Med Biol*. 2016;939:139-166. doi:10.1007/978-981-10-1503-8_7

15. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*. 2008:128-144.

16. Gerdes LU, Hardahl C. Text mining electronic health records to identify hospital adverse events. *Stud Health Technol Inform*. 2013;192:1145.

17. Harpaz R, Callahan A, Tamang S, et al. Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art. *Drug Saf Int J Med Toxicol Drug Exp*. 2014;37(10):777-790. doi:10.1007/s40264-014-0218-z

18. Hyun S, Johnson SB, Bakken S. Exploring the Ability of Natural Language Processing to Extract Data from Nursing Narratives. *Comput Inform Nurs CIN*. 2009;27(4):215-225. doi:10.1097/NCN.0b013e3181a91b58

19. Cambria E, White B. Jumping NLP Curves: A Review of Natural Language Processing Research. *IEEE Comput Intell Mag*. 2014;9(2):48-57. doi:10.1109/MCI.2014.2307227

20. Kao A, Poteet SR. *Natural Language Processing and Text Mining*. 1 edition. Springer London; 2007.

21. Abbas A, Khan MU, Ali M, Khan SU, Yang LT. A cloud based framework for identification of influential health experts from Twitter. In: *Proceedings of the 15th International Conference on Scalable Computing and Communications (ScalCom)*. ; 2015.

22. Apté C, Damerau F, Weiss SM. Automated Learning of Decision Rules for Text Categorization. *ACM Trans Inf Syst*. 1994;12(3):233–251. doi:10.1145/183422.183423

23. Baldwin KB. Evaluating healthcare quality using natural language processing. *J Healthc Qual Off Publ Natl Assoc Healthc Qual*. 2008;30:24-29.

24. Ding H, Riloff E. Extracting Information about Medication Use from Veterinary Discussions. In: ; 2015.

25. Hripcsak G, Friedman C, Alderson PO, DuMouchel W, Johnson SB, Clayton PD. Unlocking Clinical Data from Narrative Reports: A Study of Natural Language Processing. *Ann Intern Med*. 1995;122(9):681-688. doi:10.7326/0003-4819-122-9-199505010-00007

26. Temple MW, Lehmann CU, Fabbri D. Natural Language Processing for Cohort Discovery in a Discharge Prediction Model for the Neonatal ICU: *Appl Clin Inform*. 2016;7(1):101-115. doi:10.4338/ACI-2015-09-RA-0114

27. Sevenster M, Buurman J, Liu P, Peters JF, Chang PJ. Natural Language Processing Techniques for Extracting and Categorizing Finding Measurements in Narrative Radiology Reports: *Appl Clin Inform*. 2015;6(3):600-610. doi:10.4338/ACI-2014-11-RA-0110

28. Cohen KB. Chapter 6 - Biomedical Natural Language Processing and Text Mining. In: Sarkar IN, ed. *Methods in Biomedical Informatics*. Oxford: Academic Press; 2014:141-177. http://www.sciencedirect.com/science/article/pii/B9780124016781000063. Accessed June 15, 2015.

29. Zhang Y, Jin R, Zhou Z-H. Understanding bag-of-words model: a statistical framework. *Int J Mach Learn Cybern*. 2010;1(1-4):43-52. doi:10.1007/s13042-010-0001-0

30. Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. *AI Mag*. 1996;17(3):37. doi:10.1609/aimag.v17i3.1230

31. Venugopal A. Census Shows Rising Numbers of Gay Couples and Dominicans in New York. *WNYC News*. http://www.wnyc.org/story/146106-census-shows-rising-number-gay-couples-and-dominicans/?utm_source=sharedUrl&utm_medium=metatag&utm_campaign=sharedUrl. Published July 14, 2011. Accessed August 23, 2016.

32. Lasko TA, Denny JC, Levy MA. Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data. *PLOS ONE*. 2013;8(6):e66341. doi:10.1371/journal.pone.0066341

33. Jefferies D, Johnson M, Nicholls D. Nursing documentation: How meaning is obscured by fragmentary language. *Nurs Outlook*. 2011;59(6):e6-e12. doi:10.1016/j.outlook.2011.04.002

34. Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge University Press; 2008.

35. Cavnar WB, Trenkle JM. N-Gram-Based Text Categorization. In: *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*. ; 1994:161–175.

36. Conway M, Doan S, Kawazoe A, Collier N. Classifying disease outbreak reports using n-grams and semantic features. *Int J Med Inf*. 2009;78(12):e47-e58. doi:10.1016/j.ijmedinf.2009.03.010

37. Bjarnadottir R, Boctking W, Trifilio M, Dowding DW. Nurses' Perceptions of Assessing Sexual Orientation and Gender Identity in Home Health Care. *Manuscr Submitt Publ*. 2016.

38. Mladenić D, Grobelnik M. Mapping Documents onto Web Page Ontology. In: Berendt B, Hotho A, Mladenič D, Someren M van, Spiliopoulou M, Stumme G, eds. *Web Mining: From Web to Semantic Web*. Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2004:77-96. doi:10.1007/978-3-540-30123-3_5

39. Google Ngram Viewer. https://books.google.com/ngrams. Accessed May 21, 2016.

40. About Us. Callen-Lorde. http://callen-lorde.org/about/. Accessed May 21, 2016.

41. Hardiker NR, Hoy D, Casey A. Standards for Nursing Terminology. *J Am Med Inform Assoc JAMIA*. 2000;7(6):523-528.

42. 2015 Edition Final Rule. HealthIT.gov. https://www.healthit.gov/policy-researchers-implementers/2015-edition-final-rule. Accessed October 28, 2015.

43. Cahill SR, Baker K, Deutsch MB, Keatley J, Makadon HJ. Inclusion of Sexual Orientation and Gender Identity in Stage 3 Meaningful Use Guidelines: A Huge Step Forward for LGBT Health. *LGBT Health*. 2016;3(2):100-102. doi:10.1089/lgbt.2015.0136

44. Beagan BL, Fredericks E, Goldberg L. Nurses' work with LGBTQ patients: "they're just like everybody else, so what's the difference"? *Can J Nurs Res Rev Can Rech En Sci Infirm*. 2012;44(3):44-63.

45. Ellenbecker CH, Samia L, Cushman MJ, Alster K. Patient Safety and Quality in Home Health Care. In: Hughes RG, ed. *Patient Safety and Quality: An Evidence-Based Handbook for Nurses*. Advances in Patient Safety. Rockville (MD): Agency for Healthcare Research and Quality (US); 2008. http://www.ncbi.nlm.nih.gov/books/NBK2631/. Accessed May 21, 2016.

46. Peikes D, Chen A, Schore J, Brown R. Effects of care coordination on hospitalization, quality of care, and health care expenditures among medicare beneficiaries: 15 randomized trials. *JAMA*. 2009;301(6):603-618. doi:10.1001/jama.2009.126

Figure legend:

Figure 1 shows a comparison of the frequency of n-grams identified between the three types of included notes: Referral notes, coordination of care notes and narrative notes.

Table legends:

Table 1 shows the potential search words identified through literature review and domain expertise for use in n-gram text extraction, by source.

Table 2 shows the n-grams or search terms included in the search algorithm.

Table 3 shows the unigrams, bigrams and trigrams related to sexual orientation or gender identity that were identified through the text mining process and their relative frequency.

Figure 1 - original docx file

Figure 1. Comparison of n-gram frequency between notes

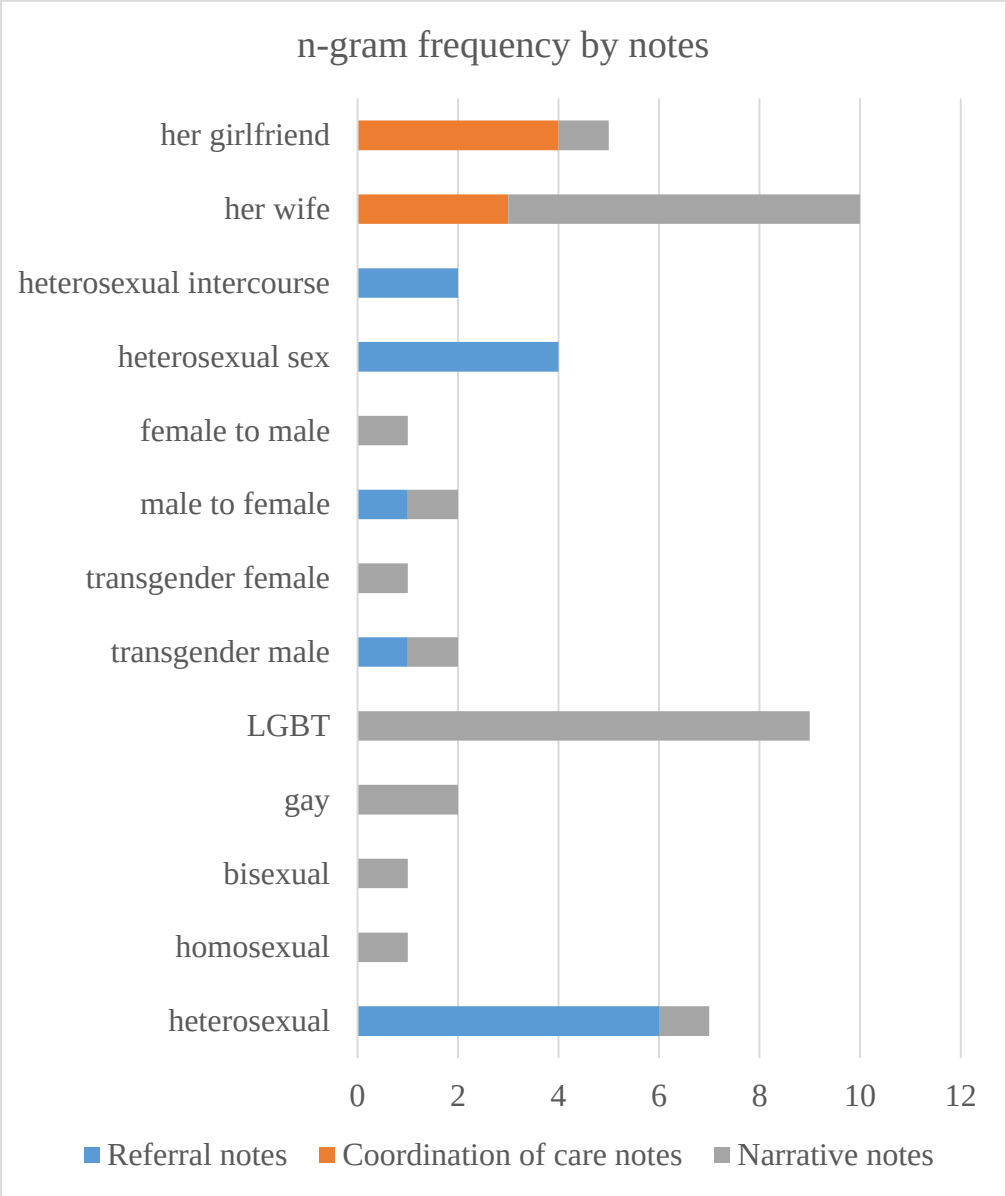

Figure 1. Comparison of n-gram frequency between notes

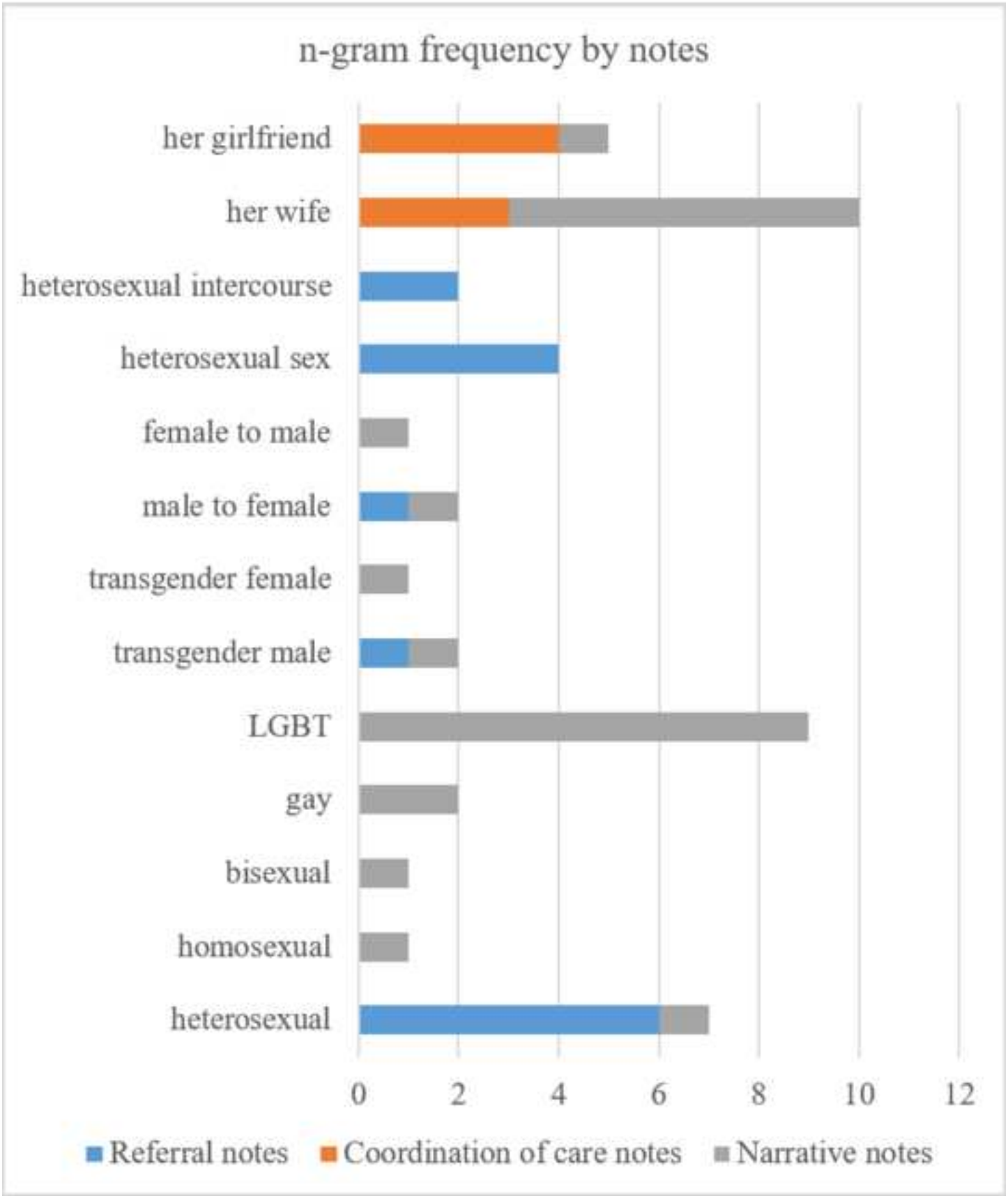

Table 1. Potential search words identified, by source

	From interviews	ICD 9-10	SNOMED	LOINC
Sexual orientation				
	lesbian	High risk heterosexual behavior	Homosexual/homosexuality	Sexual orientation
	gay	High risk homosexual behavior	Homosexual	Bisexual
	bisexual	High risk bisexual behavior	Gay	Heterosexual
	homosexual	Counseling related to patient's sexual behavior and orientation	Lesbianism	Homosexual
	same-sex		Lesbian	
	LGBT		Bisexual state	
	partner		Bisexual	
	his husband'			
	her wife'			
Gender identity				
	Transgender	Gender identity disorder	Transsexual	Gender identity
	Transsexual	Personal history of sex reassignment	Male-to-female transsexual	Identifies as male
	identifies as'		Female-to-male transsexual	Identifies as female
	mtf		Surgically transgendered transsexual	Female-to-male transsexual
	ftm		Surgically transgendered transsexual male-to-female	Male-to-female transsexual
	female to male		Surgically transgendered transsexual female-to-male	Identifies as non-conforming
	male to female			
	preferred pronoun'			

Table 2. Search terms included in search algorithm

Sexual orientation	Gender identity
Heterosexual	Transgender
Homosexual	Transsexual
Lesbian	Gender identity
Gay	Sex reassignment
Bisexual	Identifies as
Sexual orientation	Male to female
LGBT	MtF
Her girlfriend	Female to male
His boyfriend	FtM
Her wife	Preferred pronoun
His husband	

Table 3. N-grams identified in notes

A) Unigrams		Referral notes		Coordination of care notes		Narrative notes		Reference-Google books
Category	n-gram	Frequency	Relative frequency*	Frequency	Relative frequency*	Frequency	Relative frequency*	Relative frequency*
Sexual orientation	heterosexual	6	860.28	0	0.00	1	103.05	426.33
	homosexual	0	0.00	0	0.00	1	103.05	518.04
	bisexual	0	0.00	0	0.00	1	103.05	155.94
	gay	0	0.00	0	0.00	2	206.10	2152.09
	LGBT	0	0.00	0	0.00	9	927.45	82.77
Gender identity or expression	transgender	4	573.52	0	0.00	3	309.15	95.16
	m-f	1	143.38	0	0.00	0	0.00	0.28
	mtf	1	143.38	0	0.00	0	0.00	0.65
Supportive services	SAGE	0	0.00	0	0.00	9	927.45	960.83

** Unit: $\%10^{-6}$*

Table 3. continued

B) Bigrams		Referral notes		Coordination of care notes		Narrative notes		Reference-Google books
Category	n-gram	Frequency	Relative frequency*	Frequency	Relative frequency*	Frequency	Relative frequency*	Relative frequency*
Sexual orientation	homosexual male	0	0.00	0	0.00	1	103.05	2.25
Gender identity or expression	transgender m-f	1	143.38	0	0.00	0	0.00	0.00
	transgender male	1	143.38	0	0.00	1	103.05	0.14
	transgender female	0	0.00	0	0.00	1	103.05	0.07
	transgender mtf	1	143.38	0	0.00	0	0.00	0.00

	sexual reassignment	1	143.38	0	0.00	0	0.00	0.97
	preferred pronoun	1	143.38	0	0.00	0	0.00	0.07
Relationships and family	her wife	0	0.00	3	272.43	7	721.35	2.91
	her girlfriend	0	0.00	4	363.24	1	103.05	9.62
Sexual behaviors	heterosexual sex	4	573.52	0	0.00	0	0.00	540.60
	heterosexual intercourse	2	286.76	0	0.00	0	0.00	5.48
Supportive services	Callen Lorde	0	0.00	0	0.00	1	103.05	0.06
	SAGE LGBT	0	0.00	0	0.00	5	515.25	0.00
	LGBT center	0	0.00	0	0.00	1	103.05	0.25
	LGBT Sv	0	0.00	0	0.00	3	309.15	0.00
	LGBT service	0	0.00	0	0.00	2	206.10	0.00
	gay environment	0	0.00	0	0.00	1	103.05	0.30

** Unit: $\%10^{-6}$*

Table 3. continued

C) Trigrams		Referral notes		Coordination of care notes		Narrative notes		Reference-Google books
Category	n-gram	Frequency	Relative frequency*	Frequency	Relative frequency*	Frequency	Relative frequency*	Relative frequency*
Sexual orientation	Caucasian homosexual male	0	0.00	0	0.00	1	103.05	0.00
Gender identity or expression	male to female	1	143.38	0	0.00	1	103.05	11.73
	female to male	0	0.00	0	0.00	1	103.05	8.48
	sexual reassignment surgery	1	143.38	0	0.00	0	0.00	0.64
	patient is transgender	1	143.38	0	0.00	0	0.00	0.00
	is a transgender	1	143.38	0	0.00	0	0.00	0.09

Sexual behaviors	unprotected heterosexual sex	2	286.76	0	0.00	0	0.00	0.15
Supportive services	LGBT elder support	0	0.00	0	0.00	1	103.05	0.00
	SAGE for LGBT	0	0.00	0	0.00	3	309.15	0.00
	SAGE LGBT SV	0	0.00	0	0.00	4	412.20	0.00
	member of LGBT	0	0.00	0	0.00	1	103.05	0.00
	SAGE LGBT SNR	0	0.00	0	0.00	1	103.05	0.00

** Unit: $\%10^{-6}$*