Journal of
Health & Medical Informatics

Zeng et al., J Health Med Informat 2011, S3
DOI: 10.4172/2157-7420.S3-001

Case Report Open Access

Characterizing Clinical Text and Sublanguage: A Case Study of the VA Clinical Notes

Qing T. Zeng[1,2], Doug Redd[1,2], Guy Divita[1,2*], SamahJarad[3,4], Cynthia Brandt[3,4] and Jonathan R. Nebeker[1,2]

[1]*Biomedical Informatics Department at the University of Utah, Salt Lake City*
[2]*Division of Geriatrics, Department of Internal Medicine at the University of Utah, Salt Lake City*
[3]*Division of Epidemiology, Department of Internal Medicine at the University of Utah, Salt Lake City*
[4]*VA West Haven, CT, USA*
[5]*Yale Center for Medical Informatics, Yale University, New Haven, CT, USA*

Abstract

Objective: To characterize text and sublanguage in medical records to better address challenges within Natural Language Processing (NLP) tasks such as information extraction, word sense disambiguation, information retrieval, and text summarization. The text and sublanguage analysis is needed to scale up the NLP development for large and diverse free-text clinical data sets.

Design: This is a quantitative descriptive study which analyzes the text and sublanguage characteristics of a very large Veteran Affairs (VA) clinical note corpus (569 million notes) to guide the customization of natural language processing (NLP) of VA notes.

Methods: We randomly sampled 100,000 notes from the top 100 most frequently appearing document types. We examined surface features and used those features to identify sublanguage groups using unsupervised clustering.

Results: Using the text features, we are able to characterize each of the 100 document types and identify 16 distinct sublanguage groups. The identified sublanguages reflect different clinical domains and types of encounters within the sample corpus. We also found much variance within each of the document types. Such characteristics will facilitate the tuning and crafting of NLP tools.

Conclusion: Using a diverse and large sample of clinical text, we were able to show that there are a relatively large number of sublanguages and variance both within and between document types. These findings will guide NLP development to create more customizable and generalizable solutions across medical domains and sublanguages.

Introduction

Successful NLP generally requires a good understanding of the characteristics of the target text corpus. Medical NLP has benefited from thorough analyses and effective parsing of sublanguage syntax, vocabulary, and sentence types [8-14] Past efforts in sublanguage analysis have typically focused on a small number of document types such as X-ray reports or discharge summaries. This effort defines a broader perspective and seeks to identify sublanguages within a very large and diverse clinical text corpus.

The Veterans Health Administration (VHA) is the largest single medical system in the United States, providing care to millions of veterans. It operates 163 hospitals, 804clinics, and 135 nursing homes. The Veterans Health Information Systems and Technology Architecture (VistA) is the electronic health record (EHR) used by the VHA . It is one of the most widely used EHRs in the world.

Much of the information in VistA, such as progress notes, discharge summaries, radiology reports, microbiology results, pathology reports, and family histories, are in the form of unstructured and semi-structured text. Until recently, this information has not been accessible for research nor has it been usable for performance measurement, decision support, and surveillance.

A number of studies have applied Natural Language Processing (NLP) techniques to VistA free text data [1-7], with promising results. These studies have, however, only explored a very small fraction of the vast amount of VistA notes in terms of domain and facility coverage. Text and sublanguage analysis is necessary to scale up the NLP development for the VA as well as non-VA data.

In this paper we describe our analysis of a representative sample (n=100,000) of the large VistA text note corpus. The sample covers 100 different document types and the corpus data came from VA Regions 1 and 4, covering 10 different Veterans Integrated Service Networks (VISNs), ranging from New England to Rocky Mountain West.

Background

CHIR and VINCI

Currently, the Consortium for Healthcare Informatics Research (CHIR), a multi-project informatics research initiative funded by the VHA, is focused on mining free text notes. It tackles NLP methodological issues including de-identification, information extraction, and clinical inference and modeling. CHIR also includes several applied projects. VA Informatics and Computing Infrastructure (VINCI), a large informatics initiative, is responsible for creating a secure, high-performance environment for analysis, improving researchers' appropriate access to data, and providing advanced analytical tools to researchers. In regard to free-text data, VINCI currently provides

***Corresponding author:** Guy Divita, Research computer scientist, Division of Epidemiology, Department of Internal Medicine, University of Utah, Salt Lake City, USA, E-mail: Guy.Divita@hsc.utah.edu

Received November 22, 2011; **Accepted** December 22, 2011; **Published** December 26, 2011

Citation: Zeng QT, Redd D, Divita G, Jarad S, Brandt C, et al. (2011) Characterizing Clinical Text and Sublanguage: A Case Study of the VA Clinical Notes. J Health Med Informat S3. doi:10.4172/2157-7420.S3-001

access to over a billion notes. The effort reported in this paper is part of CHIR and VINCI research.

Sublanguage

According to Harris, sublanguage is "a subset of the sentences of a language forms a sublanguage of that language if it is closed under some operations of the language: e.g., if when two members of a subset are operated on, as *by and* or *because*, the resultant is also a member of that subset [15]." The lexical, syntactic, semantic, discourse, and document structure properties of a sublanguage may differ from those of the general language. The sublanguages in the medical domain have been subjects of a number of empirical and theoretical studies [10-13,16,17]. Medical NLP systems have been heavily tailored to parse specific sublanguages. Indeed, NLP systems tend to have the most success in narrow domains where the sublanguages are well defined and understood. This poses a challenge in adapting medical NLP systems to new domains, since researchers have suggested there are a number of sublanguages in the clinical notes.

V3NLP

Part of the analysis we describe here utilizes a NLP system called V3NLP. V3NLP has been released within the VA, but not yet to the general public. The intension is to release this product as open-source software when the VA decides upon on an appropriate Open Source policy. V3NLP incorporates and adapts processing modules from HITEx, cTAKES and MetaMap. It also contains new modules not previously available in HITEx [18], cTAKES [19] and MetaMap [20]. The concept-mapping pipeline, for instance, contains a sectionizer, tokenizer, sentence splitter, POS tagger, phrase chunker, concept mapper, local filters and local terminology identification. Figure 1 shows the two backend pipeline platforms that are utilized within V3NLP, SLAP, and FLAP. FLAP, The Framework Launching Application, provides the capability to dynamically launch and configure NLP pipelines in the Unstructured Information Management Architecture (UIMA) Asynchronous Scale out feature. SLAP, so named to be comparable with FLAP, is a pipeline designed around services and modules that marshal into and out of a Common Model interface.

Materials and Methods

Material

In this study, we examined the 569 million TIU notes currently available in the *VINCI text note repository*. These data came from VA Regions 1 and 4, covering 10 different Veterans Integrated Service Networks (VISNs) in the northeastern and western United States. It contains data from roughly 5.6 million patients, from October 1996 through December 2009.

These notes were labeled with one of 2481 Enterprise Document Types. 2481 document types is thought to be too granular a categorization of these notes. The top 100 document types cover 70.66% of the VINCI text note repository. Untitled notes accounted for 6.8 % of the VINCI text note repository. We calculated the document-type distribution on the entire data set. We extracted a representative sample of 100,000 notes for further analysis. The 100,000 notes were selected through random sampling of 1000 per document type from the top 100 most frequent document types.

Methods

Distribution of document type: In VistA, clinicians and administrators could create their own document titles. However, most document titles are assigned to one of the 2481 Enterprise Document Types. Example of documents types are *NURSING NOTE*, *TELEPHONE ENCOUNTER NOTE*, *MENTAL HEALTH NOTE*. Some of these notes do not have standard document types assigned to them. In this study, we refer to them as the *UNTITLED* type. We examined the frequency of unique document types. We also examined the prevalence of the all upper and lower case documents. Notes that are in only one case introduce additional ambiguity, making it more challenging to distinguish acronyms with words such as *OR* (Operating Room) versus *or* the conjunction. It is useful to know what sublanguages these all upper or all lower case notes appear in, and if they are prevalent

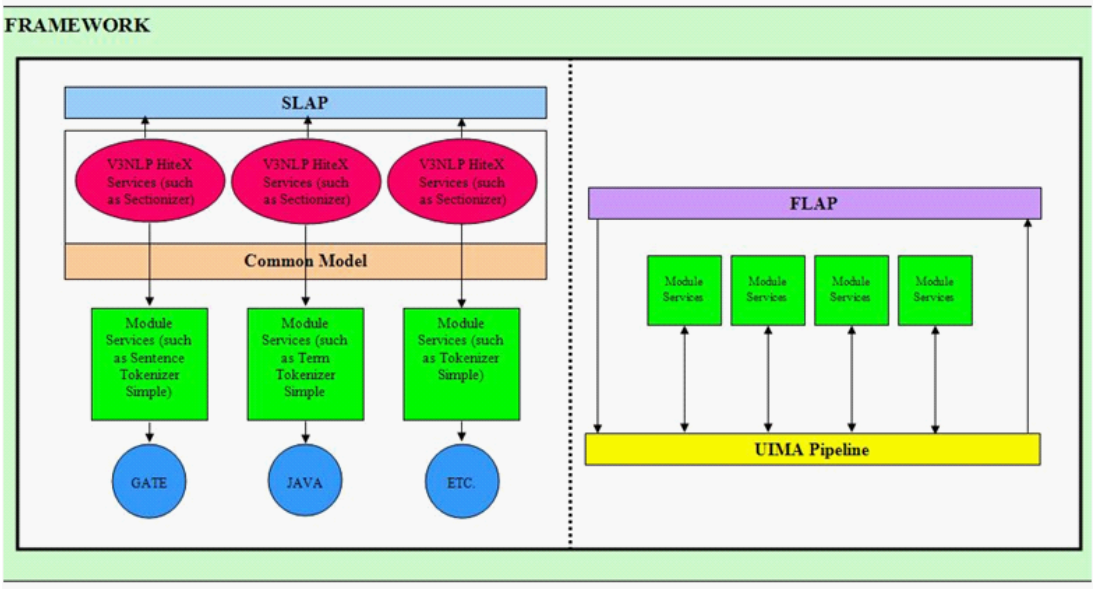

Figure 1:

enough to warrant a need to address this challenge. Existing NLP systems handle ASCII-7 or more recently UTF-8 character sets. Special consideration is needed to handle richer character sets, first to insure programs do not fail when hitting a strange character outside of expected ranges, and second to semantically correctly handle such non ASCII-7 characters, such as ©.

Document and sentence length: We measured the document and sentence length by characters and by tokens. We define tokens as sequences of characters bounded by whitespace or punctuation marks, with exceptions for tokens that are (a) numeric terms with specific patterns of digits, decimal marks, hyphens, and whitespace; (b) abbreviations that include punctuation marks; (c) dates with specific patterns of digits, punctuation marks, and whitespace; (d) words containing punctuation marks, such as apostrophes indicating possessiveness; or (e) punctuation marks external to other tokens, such as those separating sentences, phrases, or sections. In our analysis, a sentence boundary is defined as (a) the beginning or end of a document; (b) a blank line; (c) a period that is external to other tokens (e.g. not part of an abbreviation); (d) a question mark or exclamation point; or (e) a section boundary. Document length greatly affects information retrieval metrics such as TF/IDF. Some domains are less suited to use of these metrics due to very short average document lengths. Parsing tools are likewise sensitive to sentence length, with performance issues when very short or very long sentences are encountered.

Document structure analysis (section): We identified sections and section headers in the notes. What section a term appears in provides relevant context for information extraction, word sense disambiguation, categorization and information retrieval tasks. The section headers were defined by a list of previously known section header patterns (n=907) in V3NLP and a small set of regular expressions intended to capture unseen headers in the sample of this study. Examples of known headers include *Diagnosis:* or *PLAN*.

The set of regular expressions looked for patterns where the first word of a line is in initial caps, and where there is a delimiter such as a colon, followed by a line break. There was no attempt to conflate found section headings in this study; thus variants *problems*, *PROBLEM*, *Problemlist*, and *Patient Problems* were treated as different section headers in the statistics. We calculated the number of sections and report the mean and standard deviation in the sample overall and in each document type. We also describe the most frequently used section headers.

Ngram analysis (tokens): Frequent adjacent tokens find a use in NLP whether it is for vocabulary discovery, information retrieval indexes, spelling and term suggestion aids, or predictive language models. One through three grams (adjacent tokens) were created from a window of *n* tokens running across each document. (Only one and two grams are discussed in this paper). We considered lexical variants as distinct grams. Grams were put into hashes by document type to accumulate frequency counts. The challenge was to do this on such a large sample. Optimized hashes were used [21]. Grams with frequencies of 10 or more were kept. Only useful grams were kept for the 2 grams and beyond. Grams were dropped if they began or ended with function words (prepositions, determiners, pronouns, and the like.) It was noted whether each gram could have been an acronym or abbreviation or contained an acronym or abbreviation, through the use of the acronym lookup module within NLM's LVG API [22]. It was noted whether or not each gram was a string in the UMLS [23] via normalizing each gram with NLM's *norm* API, followed by a lookup into the UMLS's normalized string index. The indication of UMLS coverage for unigrams should be taken with some skepticism. The mapping was done ignoring context in the document. The resulting tables include gram, if the gram was an acronym, how ambiguous in the UMLS it is, the total frequency in the data and the frequency for the individual document type. These tables are the basis for the data in the results section. The tables are available for use upon approval.

Semantic analysis (concept,semantic group, co-occurrence): To better understand the semantic content in the notes, we use the V3NLP system described above to extract UMLS concepts from the text. Each UMLS concept is associated with a semantic type. There are 134 semantic types, which can be further grouped into 15 semantic groups for higher-level analysis [24]. These semantic groups are: Activities & Behaviors (ACTI), Anatomy (ANAT), Chemicals & Drugs (CHEM), Concepts & Ideas (CONC), Devices (DEVI), Disorders (DISO), Genes & Molecular (GENE), Geographic Areas (GEOG), Living Beings (LIVB), Objects (OBJC), Occupations (OCCU), Organizations (ORGA), Phenomena (PHEN), Physiology (PHYS), and Procedures (PROC). We analyzed and compared the distribution of the 15 semantic groups in each document type.

In addition, we analyzed the co-occurrence of the concepts and semantic groups in each document type. The discovery of co-occurrence patterns is a relatively common procedure in sublanguage analysis. The co-occurrence of syntactic or semantic categories may be examined. In this study, we focused on semantic co-occurrence.

Clustering analysis: The characteristics of the different document types are different to varying extents. To explore how they relate to each other, we performed clustering analysis. To represent the data, we used the following features set: 250 top frequent section headers, 1000 top frequent concepts, 1000 top frequent tokens, and 15 semantic groups. We used the occurrence frequencies of the features as computed from the notes. TF/IDF was applied to the notes for normalization followed by dimensionality reduction using Singular Value Decomposition (SVD). The clusters of document types were obtained by first applying k-means to the notes in the new embedded space followed by hierarchical clustering. Hierarchical clustering, namely complete linkage clustering, was performed using MATLAB on the similarity matrix constructed between the centroids of the k-means clusters based on cosine measure. The output is a dendrogram that shows the clusters of the document types.

Results

Character sets and newlines

The notes that have come out of the database were observed to be encoded in Code Page 1252. The notes do include salient non ASCII-7 characters such as © *2000*. It has been observed that there is some non ASCII-7 noise caused by OCR errors, or inadvertent insertion of binary data. The prevalence of non ASCII-7 is 0.06%. The string <*CRLF*> serves as the newline delimiter rather than ASCII 10 and 13 characters. This is likely to change to a traditional newline delimiter when the next snapshot occurs. Discussions with the data managers indicated that this is an English only corpus.

Distribution of document type

The distribution of document types by their prevalence is highly

skewed to a handful of types. Figure 2 shows the distribution of the top 10 document types. These 10 document types cover 40% of the sample. This has implications for document sample selection when studies are being formulated. There are 98 document types that include excessive shouting (all uppercase) documents. Such document types include *GROUP COUNSELING NOTE*s. Note that such excessive shouting only occurred in 15.3% of this *GROUP COUNCELING NOT*E category. Overall, such excessive shouting occurred in 5.7% of the whole sample. Slightly less than .15% of the sample is exclusively in lowercase, and only occurs in 38 document types. *OPERATVE REPORT*s had the highest prevalence of being exclusively lowercased, with an occurrence of 2.5% within this document type. It was not studied whether or not these mono-cased documents were less clinically relevant, but it is known that these mono-cased documents came from the top 100 document types, and that there was an attempt to filter out non-medical document types such as computer down time prior to final document type selection.

Document and sentence length

The average (mean) document has 299 tokens and 1083 characters

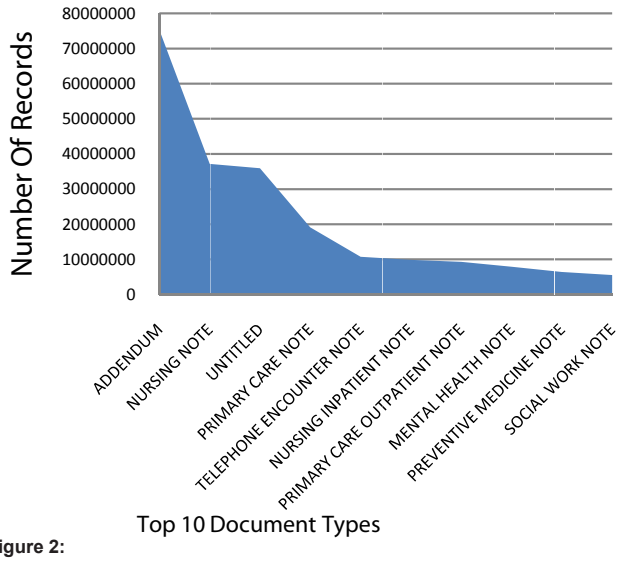

Figure 2:

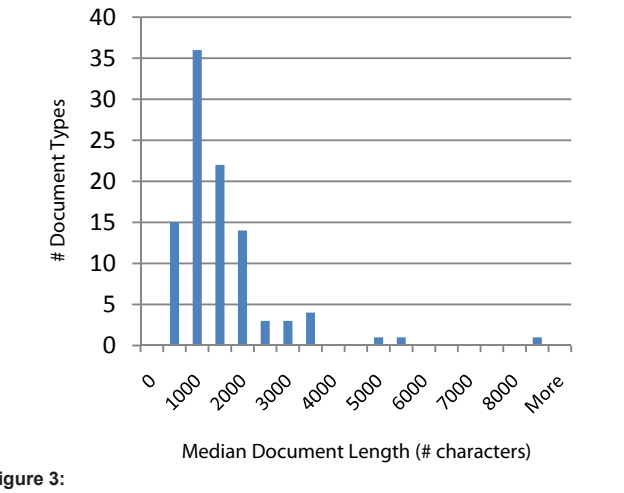

Figure 3:

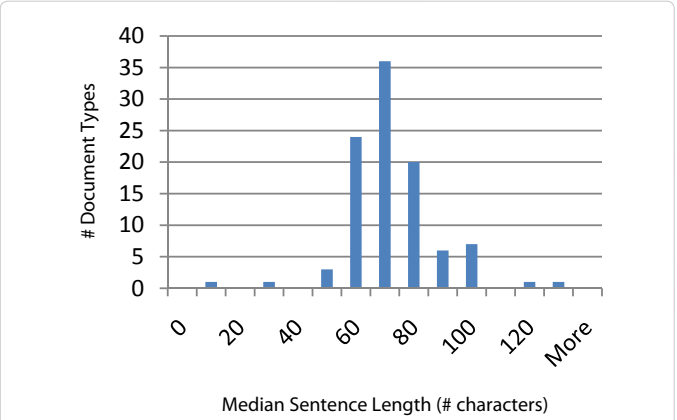

Figure 4:

in it. The medians are 166 tokens and 970 characters (interquartile ranges are 325 and 1905, respectively). Figure 3 shows the distribution of the average document type by average document character length. A portion of this corpus (15.3%) had documents that were less than10 characters in length, and they occurred in 40 document types. The most prevalent document type with less than 10 characters in it is the *OPERATIVE REPORT* document type, where 2.5% were under 10 characters in length. This may not be a coincidence.

Figure 4 shows the average sentence lengths by document type within this sample. *NURSING SKIN ASSESSMENT NOTE*s have the longest sentences with an average length of 275 characters. *ADDENDUM, DISCHARGE SUMMARY, PODIATRY NOTE, SOCIAL WORK NOTE and TREATMENT PLAN NOTE* are document types that have the median sentence length of 94 characters. *DIALYSIS NOTE*s have the shortest sentences with an average of 44 characters per sentence. The mean sentence length is 113 characters (median of 64). Interestingly, the length of sentences and documents are not well correlated – documents with the shortest sentences are not the shortest documents.

Document structure analysis (section)

The average document has 8 sections, but there is a large variance of 8.33 standard deviations to that statistic. The document types with the highest average number of sections (27) were *NURSING ADMISSION EVALUATION NOTE*s but such documents had a large standard deviation of 16.45. *H & P NOTE*s had a similar number of average sections, and an equally large standard deviation of 13.88. The document types with the lowest average number of sections (2) were *NO SHOW NOTE* (stdev = 2.81), *GROUP COUNSELING NOTE* (stdev = 1.38), *IMMUNIZATION NOTE* (stdev = 1.26), and *SCANNED NOTE* (stdev = 0.66).Figure 5 shows the distribution of documents by the number of sections contained. The document types with 2 sections per document (min) are: *GROUP COUNSELING NOTE*, *IMMUNIZATION NOTE*, *NO SHOW NOTE*, and *SCANNED NOTE*. The 10 most frequent section headers are: *ENT, PLAN, ASSESSMENT, DATE, PAIN, ALLERGIES, PULSE, OTHER, WEIGHT*, and *MEDICATIONS.*

Ngram analysis (tokens)

In this sample, there were 52,075 classes of tokens with a frequency of 10 or more, accounting for 14,187,955 token instances. The sample contains 28,138,596 instances of tokens with a frequency of 2 or more.

Figure 6 shows the top 20 useful word tokens from this sample. *Active, no, not, patient, pt,* and *yes* appear on this list and are worthy of note. *Active* shows up often within template medication lists, where each medication is labeled with *active* or *inactive* in a table format. *No* showing up so often points out that there are many negated entities in this sample. *Yes* is on this list, and illustrates the prevalent use of yes/no questions of the form question *[ ] yes [ ] no*. *Patient* and *pt* are very prevalent. *Pt* is the top acronym in the sample. Although *pt* is a widely ambiguous acronym, all evidence suggests that the majority use in this sample is for patient.

Figure 7 shows the relative proportion of the frequency of *no*, *yes*, and *pt* for each of the 100 document types. The graph shows spikes indicating that some document type's use this word much more frequently relative to all the words used in that document type. One hypothesis is that there is something unique about those document types that show more relative use of a word. In the cases of *yes, no*, and *pt*, there is prevalent use of the same template, rather than a concentration of a unique sense. For instance, document types with a high number of *yes* and *no* clearly contain a high number of boiler plate templates. The text surrounding the *yes* and *no* needs to be processed differently. The term *yes* or *no* modifies the semantics and logical

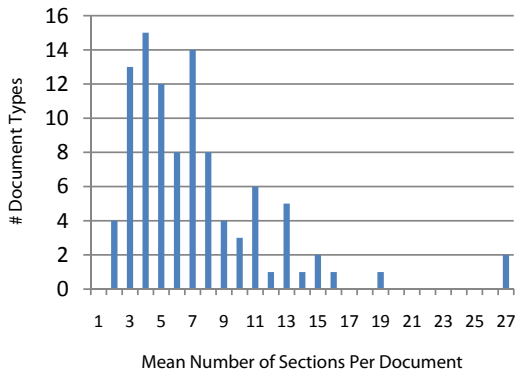

Figure 5:

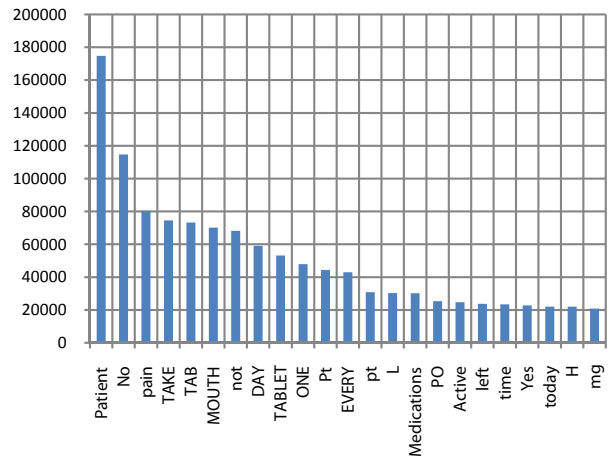

Figure 6:

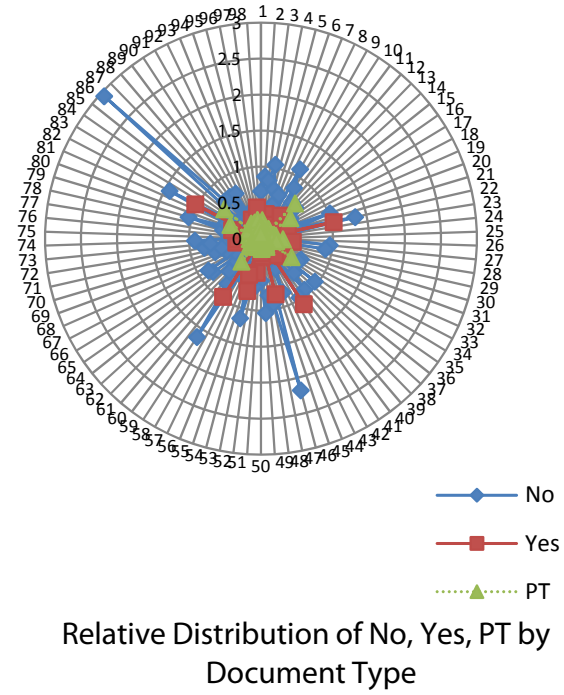

Figure 7:

assertion of the phrase or sentence preceding it. One real application in NLP is that some documents contain the question "Is the patient a *smoker*?" A typical NLP tool would extract the concept smoker from the sentence. The answer may be "No," which would be recognized as a negation. Since this negation belongs to a different sentence, the common clinical NLP tool would not negate the finding of "smoker."

There were 34,457 tokens that are potentially covered by the UMLS, and 17,618 tokens that were not. The majority of those that were not covered included function words, modifiers, and units of measure. Some words such as *Non-VA* might hold significance outside the UMLS as a clinically significant class, and might be considered local terminology. Only one non covered token showed up at the top of the list: *non tender*, which turned out to be a missed spelling variant of a covered concept. There were 5,558 token classes labeled as acronyms or abbreviations in this sample. Of these, 4,182 had coverage in the UMLS.

Bi-grams were run on the same sample for the purpose of cataloging multi-word term frequencies. Because of this, bi-grams that started with function words or ended with function words were thrown out. The filtered result was 9,794,704 bigrams. The top frequency two word terms included *outpatient medications*, *active outpatient*, *blood pressure*, *medications status* and *chest pain*. Figure 8 lists the top such terms.

Semantic analysis (concept, semantic group, co-occurrence)

We extracted 52,470 distinct concepts from the whole sample. The most frequent concepts are: *patients*, *tablet dosage form*, *active*, *day*, and *Tablet Dosing unit*. The semantic group distributions of the concepts of the top 10 most frequent document types is shown in figure 9. The distributions are not dramatically different from each other, though there are clearly distinctions. In the two types of primary care notes,

for instance, the relative frequency of chemical and drugs and anatomy concepts are about 4 times of that in the *SOCIAL WORK NOTE*. On the other hand, the semantic group distributions of the two types of primary care notes are almost identical.

The semantic group co-occurrence patterns (Figure 10) presents yet another picture. In the top 10 document types, the *PREVENTIVE MEDICINE NOTE* stood out with the most frequent co-occurrence being that between *concepts & ideas* and *living beings* instead of *concepts & ideas* and *disorders*. The two types of primary care notes in this case had a slightly different co-occurrence patterns: *chemical and drugs* is more connected to other groups in one type than the other.

Clustering analysis

The hierarchical tree provides an estimate of the similarity among document types (Figure 11). Note that we were able to identify 5 cohesive clusters of the document types that we labeled based on the theme of that cluster. The mental health cluster, for example, comprises of 12 document types all of which address the mental health condition of patients except for a few outliers such as ADDENDUM and ADMINISTRATIVE NOTES which further implies that not all closely clustered document types belong to the same clinical sub domain. Similarly, the pharmacy and medicine cluster contained document types mostly related to medicine. That we were able to identify cohesive clusters was promising however, it remains obvious that there exists overlap in sublanguages across document types as appears from the unlabeled clusters in the dendrogram.

Discussion

To adapt and develop NLP tools for diverse clinical data sets, we analyzed a large VistA text corpus. While the analyzed corpus (n=569 million) contains over 2000 document types, this study focused on a representative sample (n=100,000) from the top 100 most frequent document types. There have been a number of corpus and sublanguage analysis studies in the clinical domain; however, few were conducted on a corpus of this scale.

Our analysis suggests that the large VistA clinical text note corpus is very diverse in terms of text features. Among the 100 document types, we found large variance in every characteristic that we measured. For instance, the shortest document type (*SCANNED NOTE*) contains only 244 characters on average while the longest one (*NURSING ADMISSION EVALUATION NOTE*) contains 8412 on average. The average sentence length also differed drastically ranging from 44 characters per sentence (*DIALYSIS NOTE*) to 275 (*NURSING SKIN ASSESSMENT NOTE*). The word "hearing" is the most frequent word in *AUDIOLOGY NOTE* and "active" is the most frequent word in *E & M OF ANTICOAGULATION NOTE*. The distribution and co-occurrence of semantic groups of the concepts in *SOCIAL WORK NOTES* is clearly different from those of the *PRIMARY CARE NOTE*. There are also large variances within the document types, though they still tend to be smaller in scale when compared with between document type differences. For instance, the *IMMUNIZATION NOTE* with a mean of 501 characters has a standard deviation of 393 characters,

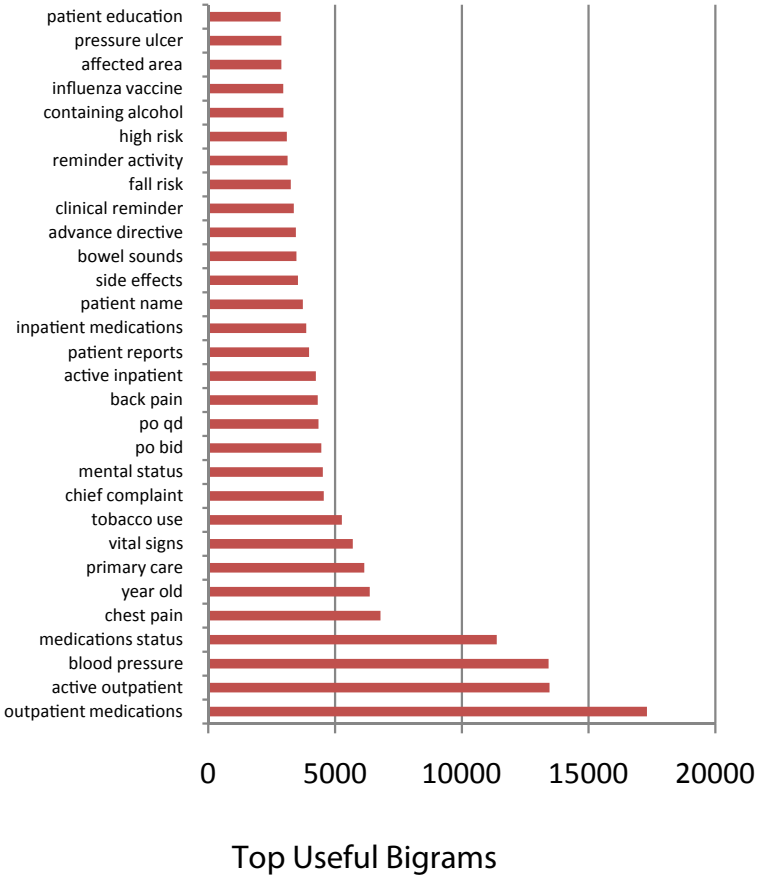

Figure 8:

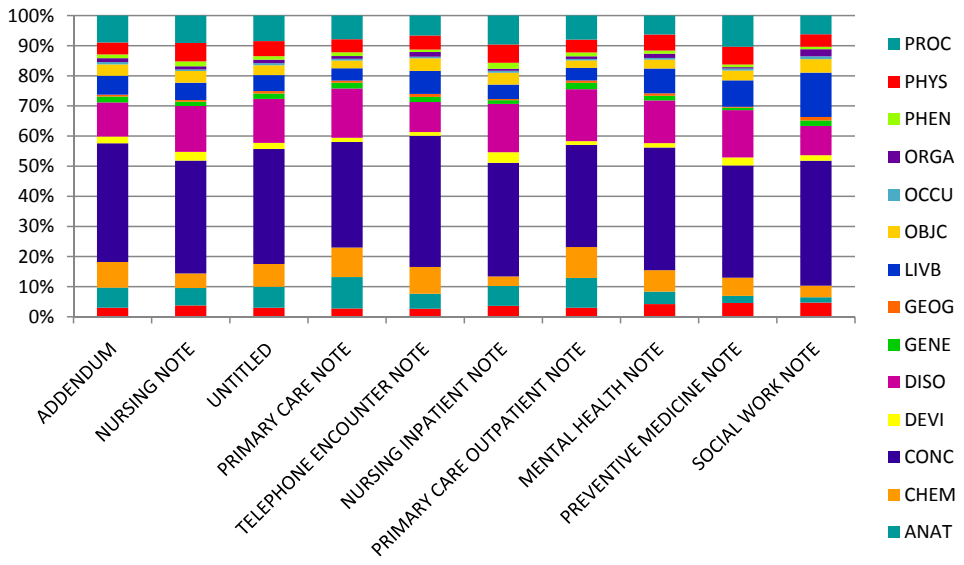

Figure 9:

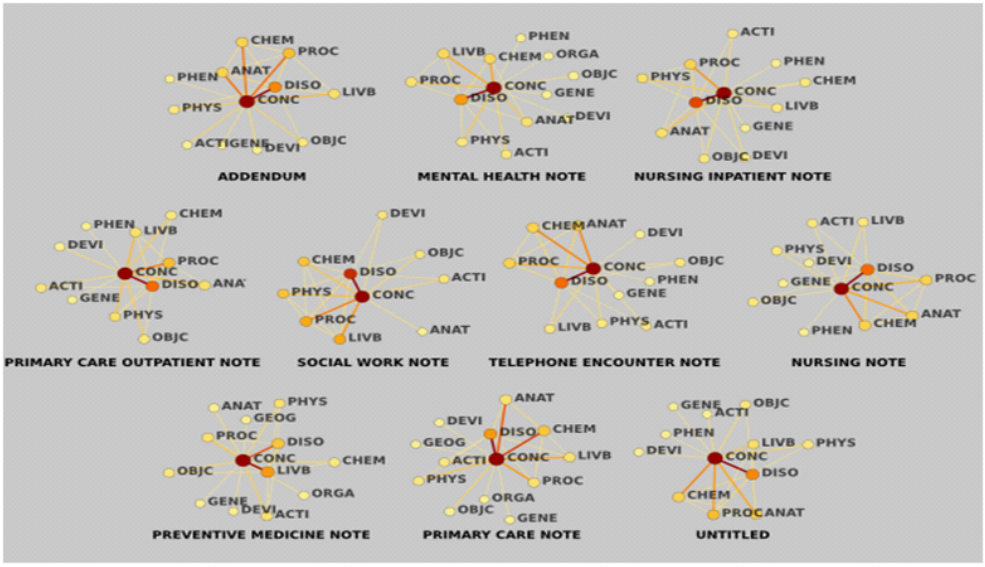

Figure 10: Semantic Group Co-occurrence Pattern. The semantic groups are represented by nodes and the frequency of their co-occurrences by edges. Within each document type, the color and diameter of the nodes indicate the percentage of the concepts in a semantic group (i.e. larger and darker nodes contain more concepts). Shorter edges and darker color indicate more frequent co-occurrences.

while the *DISCHARGE SUMMARY* with a mean of 5919 characters has a standard deviation of 4476 characters. Each document type also has a range of section headers, suggesting none followed a single template.

The findings of our study point to the existence of 16 sublanguages in the sample. As the hierarchical clustering tree shows, the document types have varying degrees of similarity. A total of 16 sublanguages were identified using the cut-off we chose. The sublanguages reflect the clinical domain and type of service provided to patients. Grouping document types into sublanguages facilitates the application and customization of NLP applications to large and diverse clinical text data sets such as the VA text corpus. Prior studies usually focused a on few document types when analyzing clinical texts and clinical sublanguage. Our results argue the importance of taking a broader perspective.

While it may be natural to use document type as a proxy for sublanguage, we need to consolidate the number of document types. For example, there are 88 different types of discharge notes and summaries within the 2481 document types in this corpus. Within the 2481 document types, it could be argued that some types should be dropped. For instance *COMPUTER DOWNTIME* is not likely to be a document type that has clinical relevancy. That being said, this should be done with care. In one pre-sample, *NO SHOW* notes were dropped because a cursory review showed that these were empty notes. A length analysis of document types showed that these notes do contain content and a source familiar with these notes had indicated that these are also clinically relevant. The language in *NURSING ADMISSION EVALUATION NOTES* can certainly be considered

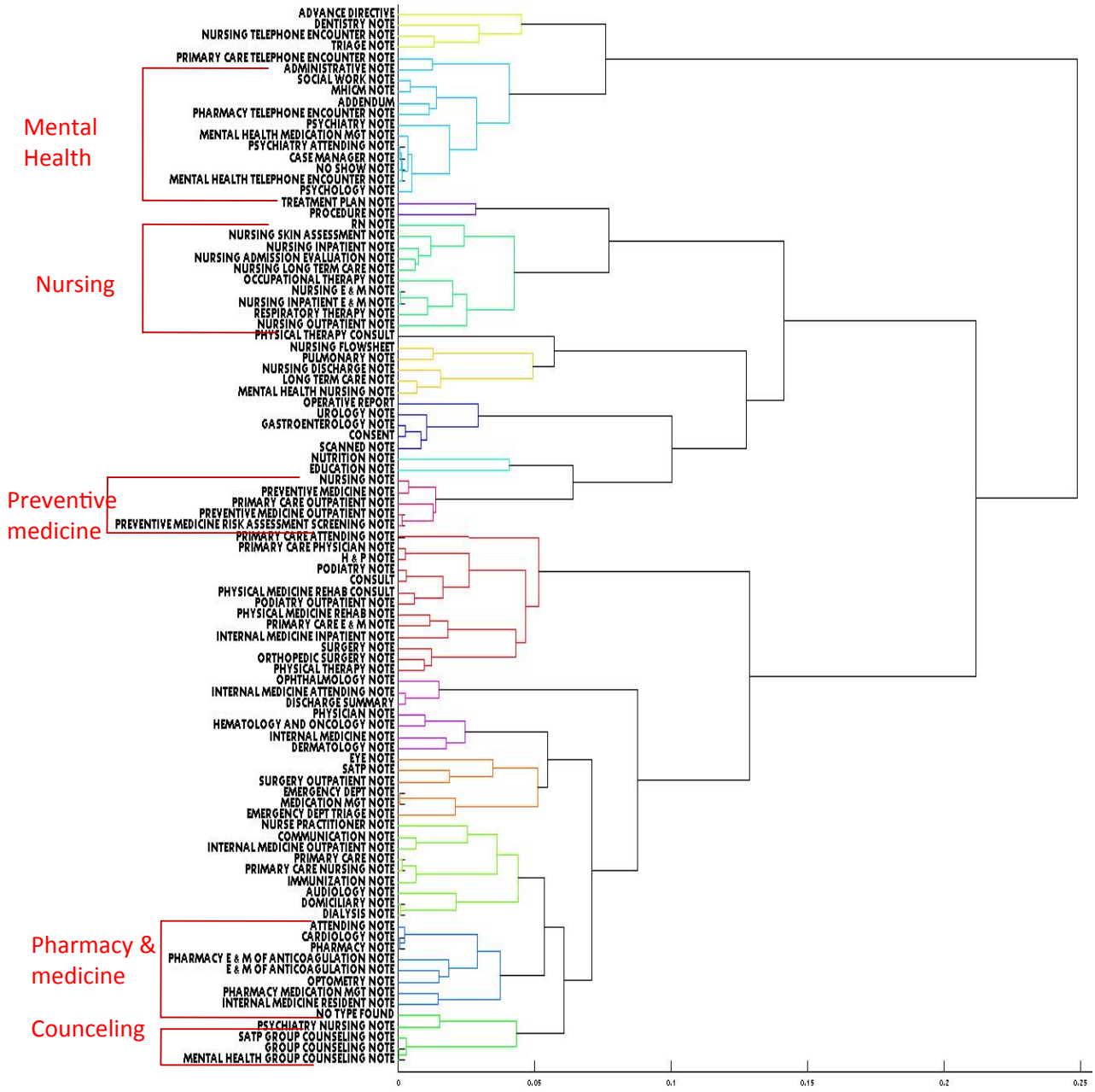

Figure 11: Hierarchical clusters of the document types.

a distinct sublanguage-, since it differs the most from the rest of the document types. On the other hand, the *GROUP COUNSELING NOTE, SATP GROUP COUNSELING NOTE*, and *MENTAL HEALTH GROUP COUNSELING NOTE* clearly fall into the same sublanguage group. Whether *UROLOGY NOTE* and *SURGERY NOTE* should be considered to have the same sublanguages will be more application-dependent.

The characteristics of the sublanguages will guide our NLP research and development. For sublanguages that are relatively "new" to the existing NLP tools, development is required. For instance, the high usage of "yes," "no" and "active" indicates certain document types contain high prevalence of semi-structured data. Current methods for processing semi-structured data are largely regular expression-based which requires human coding. More automated methods will be of great interest. For other sublanguages, adaption is in order. For example, to process document types with all upper or lower case documents, case sensitive regular expressions or dictionary lookup Should be modified.

As previously mentioned, 6.8% of the sample is untitled, and another 14% has been titled with *ADDENDUM*. While it is possible to track down what the addendum is to, the records are not thus marked. The addendum often serves a different purpose than the parent document anyway, such as to update findings or as a type of chat space.

This points out the further need for an effort to empirically classify document types beyond the given Enterprise document type. It is hoped

that this paper raises the awareness of topics regarding document type and document type recognition, the diversity of content, evidence of duplicative, templated text within the corpus and evidence of the many sublanguages within this corpus. Each of these topics should be explored prior to pulling samples from the corpus to make sure the samples are not skewed inadvertently.

This study is not a comprehensive sublanguage analysis of the VistA corpus. There are a number of vocabularies, syntax and document structure features that remain to be examined. There are also many less frequent document types to be included.

The sublanguage analysis was only conducted on the document level; while there are reasons to believe the languages in different sections and possibly across document types sometimes have very distinct features. More significantly, in-depth analysis of each of the sublanguages will be needed in order to customize the NLP functions for them.

Future Work

The n-gram corpus statistics will be used for vocabulary discovery, as features for acronym word sense disambiguation, and as a language model for spelling suggestion. More work should be done to refine document types for this corpus to something that is both salient and manageable. The sublanguages that are emerging will guide word sense disambiguation tasks. It would be interesting to compare elements from this corpus with other medical record corpora. More importantly, we plan to systematically adapt the V3NLP tool for the VistA sublanguages and benchmark the performance before and after the adaptation.

Acknowledgement

This work is funded by VA grants CHIR HIR 08-374 and VINCI HIR-08-204. The authors thank the members of the VINCI NLP team for their support and comments.

References

1. Rubin D, Wang D, Chambers DA, Chambers JG, South BR, et al. (2010) Natural language processing for lines and devices in portable chest x-rays. AMIA Annu Symp Proc 2010: 692-696.
2. South BR, Shen S, Jones M, Garvin J, Samore MH, et al. (2009) Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinformatics 9:S12.
3. D'Avolio LW, Nguyen TM, Farwell WR, Chen Y, Fitzmeyer F, et al. (2010) Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). J Am Med Inform Assoc 17:375-382.
4. South BR, Phansalkar S, Swaminathan AD, Delisle S, Perl T, et al. (2007) Adaptation of the NegEx algorithm to Veterans Affairs electronic text notes for detection of influenza-like illness (ILI). AMIA Annu Symp Proc 11:1118.
5. DeShazo JP, Turner AM (2010) An interactive and user-centered computer system to predict physician's disease judgments in discharge summaries. J Biomed Inform 43: 218-223.
6. Gundlapalli AV, South BR, Phansalkar S, Kinney AY, Shen S, et al. (2008) Application of Natural Language Processing to VA Electronic Health Records to Identify Phenotypic Characteristics for Clinical and Research Purposes. Summit on Translat Bioinforma 2008: 36-40.
7. Brown SH, Speroff T, Fielstein EM, Bauer BA, Wahner-Roedler DL, et al. (2006) eQuality: electronic quality assessment from narrative clinical reports. Mayo Clin Proc 81: 1472-1481.
8. Laippala V, Ginter F, Pyysalo S, Salakoski T (2009) Towards automated processing of clinical Finnish: sublanguage analysis and a rule-based parser. Int J Med Inform 78: e7-12.
9. Pyysalo S, Salakoski T, Aubin S, Nazarenko A (2006) Lexical adaptation of link grammar to the biomedical sublanguage: a comparative evaluation of three approaches. BMC Bioinformatics 3: S2.
10. Friedman C, Kra P, Rzhetsky A (2002) Two biomedical sublanguages: a description based on the theories of Zellig Harris. J Biomed Inform 35: 222-235.
11. Stetson PD, Johnson SB, Scotch M, Hripcsak G (2002) The sublanguage of cross-coverage. Proc AMIA Symp 2002: 742-746.
12. Johnson SB (1998) Conceptual graph grammar--a simple formalism for sublanguage. Methods Inf Med 37: 345-352.
13. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ (1994) Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1:142-160.
14. Patterson O, Igo S, Hurdle JF (2010) Automatic acquisition of sublanguage semantic schema: towards the word sense disambiguation of clinical narratives. AMIA Annu Symp Proc 2010: 612-616.
15. Special Issue: Sublanguage. Dedicated to the memory of Zellig Harris (2002). J Biomed Inform 35: 213-277.
16. Morrison FP, Kukafka R, Johnson SB (2005) Analyzing the structure and content of public health messages. AMIA Annu Symp Proc 2005: 540-544.
17. Rossi Mori A, Galeazzi E, Gangemi A, Pisanelli DM, Thornton AM (1991) Semantic standards for the representation of medical records. Med Decis Making 11: 76-80.
18. Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, et al. (2006) Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. BMC Med Inform Decis Mak 6:30.
19. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, et al. (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 17: 507-513.
20. Aronson AR (2001) Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp 2001: 17-21.
21. Friedman E. GNU Trove 2001-2009.
22. Browne A, C. L. Lvg <www.specialist.nlm.nih.gov> Accessed 2011.
23. Unified Medical Language System (UMLS) < Accessed 2010.
24. McCray AT, Burgun A, Bodenreider O (2001) Aggregating UMLS semantic types for reducing conceptual complexity. Stud Health Technol Inform 84: 216-220.

This article was originally published in a special issue, **Biomedical Informatics** handled by Editor(s). Dr. Emekalam Antony, Elizabeth City State University, USA