AUTOMATED METHODS TO EXTRACT PATIENT NEW INFORMATION FROM CLINICAL NOTES IN ELECTRONIC HEALTH RECORD SYSTEMS

A DISSERTATION SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL OF UNIVERSITY OF MINNESOTA
BY

RUI ZHANG

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

ADVISER: SERGUEI V. PAKHOMOV, PH.D.

NOVEMBER 2013

Acknowledgements

My earnest gratitude to both Dr. Serguei Pakhomov, my adviser, for his insightful and invaluable contributions to my academic progress, and Dr. Genevieve Melton-Meaux, my research mentor, for providing exceptional guidance, inspiration and motivation to successfully complete my dissertation. You two are the best mentors ever.

Special thanks to my committee members: Dr. Hongfang Liu, the director of Mayo Clinic's clinical natural language processing (NLP) program, for her research expertise to consistently support my work; Dr. Stuart Speedie for his contributions to my academic progress; Dr. Bonnie Westra for offering her clinical and research expertise to support my work. Many thanks to my colleges in the NIP/IE group: Dr. Bridget McInnes, Dr. Ying Liu, Yan Wang, Dr. Riea Moon, Robert Bill, James Ryan and Ben Knoll for their support.

I would like to thank the Semantic Knowledge Representation (SKR) team at the Lister Hill National Center for Biomedical Communications, an intramural research division of the National Library of Medicine, National Institutes of Health: Dr. Thomas Rindflesch, Dr. Marcelo Fiszman, Dr. Mike Cairelli, Dr. Halil Kilicoglu, Dr. Graciela Rosemblat, Dr. Guocai Chen, Dr. Dongwook Shin, and Dr. Liz Workman for their guidance during my research internship.

Also, I would like to thank Dr. Connie Delaney (IHI Acting Director), Dr. Lael Gatewood (IHI Associate Director), Dr. David Pieckiewicz (IHI Director of Graduate Studies), and the outstanding faculty, staff, and student fellows: Dr. Terry Adam, Dr. Gyorgy Simon, Dr. Saif Khairat, Dr. Venkatesh Rudrapatna; Jessica Whitcomb-Trance,

Faith Goenner, Elizabeth Madson, Wenjun Kang, Lindsay Bork, Eva Gibney-Jones, Megan Boehm; Dr. Oladimeji Farri, Dr. Piper Svensson-Ranallo, Dr. Mike Grove, Yue Zhou, Yi Zhang, Yanrong Zhu, Zhen Hu, Merdi Rafiei, and David Marc.

Finally, I would like to express my appreciation to the University of Minnesota Graduate School for awarding Doctoral Dissertation Fellowship (DDF) and Health Informatics Fellowship to support my study and dissertation; the Healthcare Information and Management Systems Society (HIMSS) for awarding scholarship to support my study; the University of Minnesota affiliated Fairview Health Services for providing access to relevant patient records, and the invaluable time and efforts of medical interns, residents at the University of Minnesota who participated in my studies.

Dedication

To:

Maojun Zhu & Juzhong Zhang

Xiaopei Wang

Abstract

The widespread adoption of Electronic Health Record (EHR) has resulted in rapid text proliferation within clinical care. Clinicians' use of copying and pasting functions in EHR systems further compounds this by creating a large amount of redundant clinical information in clinical documents. A mixture of redundant information (especially outdated and incorrect information) and new information in a single clinical note increases clinicians' cognitive burden and results in decision-making difficulties. Moreover, replicated erroneous information can potentially cause risks to patient safety. However, automated methods to identify redundant or relevant new information in clinical texts have not been extensively investigated.

The overarching goal of this research is to develop and evaluate automated methods to identify new and clinically relevant information in clinical notes using expert-derived reference standards. Modified global alignment methods were adapted to investigate the pattern of redundancy in individual longitudinal clinical notes as well as a larger group of patient clinical notes. Statistical language models were also developed to identify new and clinically relevant information in clinical notes. Relevant new information identified by automated methods will be highlighted in clinical notes to provide visualization cues to clinicians. New information proportion (NIP) was used to indicate the quantity of new information in each note and also navigate clinician notes with more new information. Classifying semantic types of new information further provides clinicians with specific types of new information that they are interested in finding. The techniques developed in this research can be incorporated into production

EHR systems and could potentially aid clinicians in finding and synthesizing new information in a note more purposely, and could finally improve the efficiency of healthcare delivery.

Table of Contents

List of Tables

List of Figures

CHAPTER 1 INTRODUCTION

In this chapter, I will introduce the current state of healthcare documentation with respect to redundant information, such as the existing problems and previous studies on the redundancy. I will also describe the significance and specific aims of the research presented in this dissertation.

1.1 Current State of Healthcare Documentation with Redundancy

1.1.1 Irrelevant Redundant Information in Clinical Texts

Implementation of electronic health record (EHR) systems has resulted not only in fundamental changes in clinical workflow but also rapid proliferation of electronic clinical texts. While EHR adoption provides the opportunity for health care organizations to promote better quality, decrease costs, and increase efficiency in healthcare, there are some side-effects of EHR use, which may not always be desirable (1, 2). One of the functionalities of many EHR systems is the ability to reuse information previously documented in notes by copying from previous clinical documents and pasting into the current clinical note of a patient. Within the time-constrained clinical settings, clinicians use this function to shorten the time spent on the process of documenting multiple encounters with the same patient. While having information readily available in EHR systems is helpful, excessive redundant information can lead to an increased cognitive burden, information overload, and difficulties in effectively distilling relevant information for effective decision-making at the point of care (3).

This “copy-and-paste” issue is ubiquitous in many healthcare organizations’ EHR systems. Weir et al., found that approximately 20% of 1,891 notes at the Salt Lake City Veterans Affairs (VA) health care system contained copied text (4). Hammond et al. have also reported their observations of copying and pasting in VA Computerized Patient Record System (CPRS) as early as one decade ago (5). They found that 9% of progress notes in the VA CPRS contained duplicated information. The ratio of copy events to total documents was over 50% and the occurrence of copy events increased greatly over time (5).

Rapid proliferation of clinical texts with redundant, duplicated information can potentially increase the cognitive burden of clinicians (4-8). Issues caused by redundant information in clinical texts have recently become a recognized problem in EHR systems, with more and more health care organizations looking for solutions for this problem (9-12). As a main type of documentation in EHR systems, clinical note contains longitudinal textual information of patient medical history. Viewing clinical notes to understand a patient’s medical history and synthesize current clinical condition is the fundamental clinical task for healthcare providers in time-constrained clinical settings. Undoubtedly, a very long and complex clinical note with redundant information propagated from older notes can increase the difficulties of following a patient’s treatment process and changes in medical condition up through a specific visit. Irrelevant and redundant information also deemphasizes the importance of relevant new information, especially when the important information is smaller in size in comparison to redundant information. In some cases, outdated medical information was also kept in the newer notes. For example,

previously prescribed medications are often repeated in the current note even if the patient has had no change in regimen or, even worse, in many cases clinical notes include medications that the patient is no longer taking (10). In a recent report, a physician note included information that the patient was on day 5 of antibiotics, which was copied 6 days in a row (13).

Redundant information in clinical documents can also potentially increase the clinical decision-making difficulty of clinicians. Clinical decision-making is a complex process based on synthesizing evidence from clinical notes, diagnostic studies, and other available clinical information. Having a biased search of clinical evidence by physicians can lead to failures in considering adequate alternative diagnostic possibilities (7). A mixture of redundant (especially outdated or incorrect information) and new information in a single clinical note could potentially confuse the clinician when the clinical history has been documented over a long period of time. For example, in one study, the fact that one outdated recommendation was copied from a hospitalization note dictated 7 years prior produced discord with other parts of the medical chart and confused a clinician (9). A similar situation was also reported by a nurse when she saw documentation of an event that happened 4 years prior was repeated in subsequent recent notes (13).

Moreover, various types of copied information can lead to different degrees of potential risks. Inappropriate reuse of replicated information may introduce errors into the EHR patient records, and erroneous information can be propagated over notes leading to a potentially unsafe situation (5). Hammond et al. classified copied texts into six categories based on the severity of risks and found that 36.6% of all copied events could

lead to higher risks of patient harm, fraud or tort claim exposure (5). The top three categories that contained high-risk information include physical and mental examination, history of present illness, and past medical history. Furthermore, another recent study showed that the existence of redundant information in notes results in decreased use of clinical notes by clinicians (6).

Therefore, irrelevant redundant information prevalent in the EHR systems decreases healthcare efficiency and is a potential risk for patient safety. As its counterpart, new information and clinically relevant information can potentially provide direct information needed for understanding a patient's medical history but is often obscured by the volumes of irrelevant information. If systems could be developed to help physicians identify and navigate relevant new information, it could potentially decrease the cognitive burden and decision-making difficulties for clinicians, ultimately improving the efficiency of health care workflow.

1.1.2 Previous Studies to Detect Redundancy in Clinical Texts

Few groups have previously tried to quantify the redundancy in clinical documents by using different methods. Weir et al. manually chart reviewed 1,891 notes in the Salt Lake City VA health care system notes and found that approximately 20% of at the contained copied text (4). Although the chart review is an accurate method, it is a human intensive process and inefficient for analyzing a large amount of notes. Hammond et al. performed a pair-wise comparison of all patient documents to identify matches to at least 40 consecutive word sequences in all document paris (167,076 progress notes for 1,479 patients) in the VA CPRS (5). However, their methods could only find redundancy with

long duplicated word sequences (*i.e.*, 40 words), but could miss the smaller changes of the medical condition, such as the dosage change of a medicine. Wrenn et al. quantified the percentage of information in a collection of 1,670 inpatient notes with four types (sign-out note, progress note, admission note and discharge note) from 100 patients at New York-Presbyterian Hospital by using global alignment (14). They found an average of 78% and 54% redundant information in sign-out and progress notes, respectively.

More recently, Cohen et al. used the Smith-Waterman text alignment algorithm to quantify the amount of redundancy both in terms of word and semantic concept repetition (15). They used a similar corpora of patient notes as Wrenn et al. did (14) and observed word sequence redundancy levels (*i.e.,* the percentage of alignment of two documents) of 29% and non-standard distribution of concept level redundancy. They also investigated the impact of redundancy in the corpus on the text mining applications: collocation identification and topic modeling. For example, collocations extracted from a redundant EHR corpus were significantly different from those from a non-redundant EHR corpus; while, the results were similar when the topic modeling Latent Dirichlet Allocation (LDA) was applied to these two EHR corpora (15). Therefore, they suggested examining the redundancy of the corpus before applying any text mining techniques. All these studies did not utilize a reference standard in evaluating the performance of these methods, thus it was difficult to compare their performance with each other.

1.1.3 Automated Clinical Text Summarization to Solve Redundancy Issue

To address the issue of redundant information in clinical texts, one solution is to summarize clinical texts (16, 17). Several reports have focused upon automated

summarization of patient narrative (16, 18) which may be viewed as one way of reducing the amount of work needed to process patient records; however, these techniques typically provide summarized narrative of a patient separate from the clinical documents. They do not help clinicians focus on potential critical types of information "in situ" within the original document. The purpose of text summarization is to reduce the size of text while retaining important pieces of information. Instead of spending much time on processing information from a large number of clinical notes, clinicians would only need to review a short automated summarization abstracted from notes. A structured overview of patient information, *Patient Worksheet*, has demonstrated the feasibility and benefits of automated generation of patient summaries that may improve patient health outcomes (17).

While automated text summarization methods provide time-savings in capturing relevant information, they may not be the best approach to solve issues with redundant information. Summarization is the process of retrieving relevant information based on the need of clinicians, followed by text generation of information either in unstructured text format or in structured format. The design of automated summarization systems requires a comprehensive understanding of cognitive reasoning used by clinicians when they review and summarize clinical records (8). Without the comprehensive understanding, a representative model of text summarization is hard to generalize since it depends on who will review the summarization and which clinical tasks the physicians will focus on. Text summarization of an isolated patient clinical record abstracts and separates information from clinical notes (19), thus lacking the connection to the source records. Clinicians may

actually spend more time on identifying or validating summarization due to the potential loss of information with summarization rather than save time. Moreover, summarization loses contextual information from original clinical notes that may help clinicians to understand clinical changes. Thus, the investigation of alternative solutions to redundancy problems is needed.

1.1.4 Visualization of Information in Clinical Narratives

User interface (UI) design problems related to usability and human-computer interactions have been reported as a barrier to effective use of EHR. Re-design of EHR systems UI to aid clinicians, including visualization of relevant new information, is a potentially effective solution. Reported methods and applications for visualization of clinical documents have not focused on either usability (20) or optimal presentation (21) of these texts. Visualization techniques have been demonstrated to be an effective way to represent information from time-stamped, longitudinal clinical records to clinicians and medical researchers (22-24). Optimally organized views of these data allow for comprehensive understanding and comparisons of clinical parameters over time. Researchers have focused primarily on the visualization of structured data, such as laboratory data or radiology imaging (22). While one group has developed a knowledge-based system to display information related to a specific concept of interest based on the semantic relationships (25), visualization techniques for unaltered original clinical texts remain largely unexplored. Currently, clinicians have few tools or cues to help with reviewing clinical notes, thus visualization cues or tools for distilling relevant information in clinical notes are needed.

1.2 Significance

This dissertation addresses the important and ubiquitous issue that has to do with the impact of redundant information present in modern EHRs on quality and efficiency of healthcare delivery. Clinical texts of EHR systems are widely used for patient care documentation and communication between clinicians in the health care system. As one of the primary ways that clinicians aggregate clinical data from various sources, including laboratories, medications and diagnostic studies, EHR systems contain both structured data and unstructured texts. While structured data can be easily analyzed and aggregated, it can be hard for clinicians to interpret this information due to the loss of the clinical context. Unstructured narratives contain complete information that helps clinicians communicate and retrieve relevant information for analyzing complex medical histories and evaluating the current clinical condition of patients. Existence of abundant irrelevant redundant information in EHR systems may result in information overload for clinicians, thus increasing difficulties in information analysis and decreasing clinical workflow efficiency.

The task of reviewing multiple patient notes and synthesizing relevant information in a limited time is challenging. Computational tools to identify the relevant new information can help focus a clinician's attention on making clinical decisions rather than struggling to distill information. To develop such automated tools, researchers have used text summarization techniques to capture key information in the clinical notes. These tools can provide clinicians a brief overview of patient information but also can potentially remove information that might be important for clinicians to synthesize and

understand complex patients. Previous studies have demonstrated that clinical narrative is critical for decision-making. Therefore, the development of methods to keep clinical narratives intact but still highlight relevant new information could be very helpful in focusing attention on the new information within the original clinical text. Successful implementation of the solution to this problem will have a major impact on healthcare and the filed of health informatics, and promise to make healthcare delivery process safer and more effective for patients.

1.3 Specific Aims

This dissertation hypothesizes that development of advanced clinical natural language processing methods can better assist clinicians to identify and navigate new and clinically relevant information from the EHR systems. The overarching goal of this study is to develop automated methods to identify relevant new information for assisting clinicians to navigate and review the notes with new information more efficiently and quickly. I propose to address this goal through the following four specific aims:

1) To modify global alignment to investigate the redundancy patterns in outpatient clinical notes both for individual patients and whole corpus
2) To develop statistical language models to identify and visualize new and clinically relevant information in the clinical notes document
3) To enable navigation of relevant new information in clinical notes by quantifying the proportion of new information in each note

4) To provide clinicians specific types of new information by classifying semantic types of new information and extracting biomedical concepts

The studies focused on these four aims were designed to systematically investigate the redundancy problem from understanding hidden redundancy patterns, to developing methods for identifying and visualizing new information, to navigating notes with new information, and to extracting specific type of new information. This dissertation provides potential ways to lessen the clinicians' cognitive burden caused by the redundant information in EHR systems.

CHAPTER 2 BACKGROUND

In this chapter, I will provide background knowledge for developing methods in this dissertation. I will introduce the sequence alignment algorithms, statistical *n*-gram language models and semantic similarity measures as well as Unified Medical Language System (UMLS) knowledge base and some existed NLP tools developed by U.S. National Library of Medicine (NLM).

2.1 Sequence Alignment

In this section, I will firstly introduce basic background on sequence alignment algorithms widely used in bioinformatics, and explain the Needleman-Wunsch algorithm that I modified to investigate redundant patterns (Aim 1).

2.1.1 Global Alignment and Local Alignment

In bioinformatics, sequence alignment was initially developed to identify similar regions of sequences of DNA, RNA and proteins to discover or speculate relationships of function, structure and evolution between sequences. Alignment techniques can be generally separated into two categories: global alignment and local alignment.

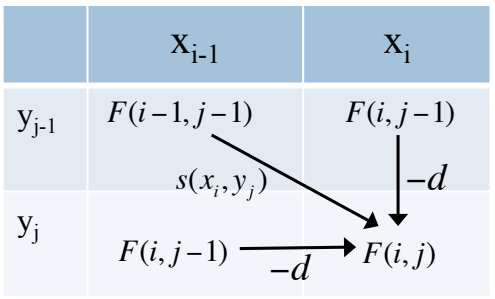
Global alignment identifies the overall similarity of the entire length of a sequence compared to another sequence, and thus is most suitable when the two sequences have a significant degree of similarity throughout and are of similar length. The classic global alignment method is Needleman-Wunsch algorithm (26) that was the first application of dynamic programming in biological sequence alignment.

In contrast, local alignment detects similarity of smaller regions within long sequences. This type of alignment is most suitable when comparing substantially different sequences, which possibly differ significantly in length, and contain only short regions of similarity. The classic local alignment method is Smith-Waterman algorithm (27), which compares segments with all possible lengths.

Typically, there is no difference between global and local alignment when sequences are sufficiently similar. While local alignment allows for the measurement of overlap or similarity over short sequences, it does not provide an aggregate measure of local similarities throughout one sequence compared to another. In contrast, global alignment assumes a single full alignment of two sequences of interest (28).

2.1.2 Needleman-Wunsch Algorithm

The Needleman-Wunsch algorithm calculates a score for two sequences by assigning the penalty values to matches, mismatches, insertions and deletions. As shown in Figure 2-1, one column represents an element in sequence X, and one row indicates each element of sequence Y. The alignment score between x_i and y_j is the maximum value of scores calculated from adjacent northwestern cell, up cell, and the left cell by either adding a substitution score or subtracting a gap penalty. A traceback matrix is also created to record the paths to generate those scores for each cell. To obtain the best alignment, the traceback process starts from the lower right corner, traces back by following the path that recorded in the traceback matrix, and finally stops at the upper left corner of the score matrix.



$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(x_i, y_j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$

where $s(x_i, y_j)$ *is the substitution score,*
d *is the gap penality.*

Figure 2-1. Illustration of score matrix calculation for aligning sequences X and Y by using the Needlman-Wunsch algorithm.

2.1.3 Global Alignment for Clinical Texts

Alignment methods are an option for identifying redundancy in clinical texts. Although alignment methods were originally designed to compare two DNA base pairs in sequences, it has been used within computational linguistics and natural language processing (NLP) to compare the similarity of two texts (14). Instead of aligning letter than represent the bases of two genes, alignment methods align words in two texts. Different penalty scores can be assigned for insertions, deletions, and mismatches. With the best possible alignment, the minimum possible distance between two texts can be calculated by using dynamic programming.

A limitation of the global alignment approach in the context of clinical reports becomes apparent in situations when note sections may appear out of "normal" sequence. For example, if the same two sections in several clinical notes are in a different order but are otherwise highly redundant, the global alignment approach would be unsuitable and would grossly underestimate the degree of redundancy between the notes. In contrast, local alignment techniques alone would not be suitable as these measures would provide

a measure of similarity over a short sequence but no aggregate measure over the entire note of information similarity.

2.2 Statistical Language Models

In this section, I will present a brief overview of the foundations of statistical language modeling pertinent to the work (Aim 2&3) in this dissertation. More detailed and in-depth information on SLM can be found in Jurafsky and Martin (29) and Manning and Schutze (30). An *n*-gram model is a method widely used in the field of computational linguistics and NLP. Important concepts with this include cross entropy and perplexity, which can be used to identify new or redundant information in clinical notes.

2.2.1 Markov Assumption

An *n*-gram model is a type of statistical language model (SLM), which predicts the probability of a word based upon all previous words (29, 30). Assuming that each word is independent, the probability of a complete string of words is:

$$P(w_1, w_2 \ldots, w_{n-1}, w_n) \tag{1}$$

The probability of the word can be decomposed by using chain rule of probability:

$$\begin{aligned} P(w_1^n) &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1 w_2)\ldots P(w_n \mid w_1 w_2 \ldots w_{n-1}) \\ &= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1^2)\ldots P(w_n \mid w_1^{n-1}) \\ &= \prod_{k=1}^{n} P(w_k \mid w_1^{k-1}) \end{aligned} \tag{2}$$

Due to the complexity of calculating the probability of a word given a long sequence of preceding words (*i.e.*, the calculation of $P(w_n \mid w_1^{n-1})$ is computationally intractable), several simplifications have been introduced. An approximation called the *Markov assumption* states that the probability of one word can be based on the prior few words instead of all previous words. The general equation for the *n*-gram approximation with *Markov assumption* is therefore:

$$P(w_k \mid w_1^{k-1}) \approx P(w_k \mid w_{k-n+1}^{k-1}) \tag{3}$$

As shown in equation (3), the probability of a word given all previous words is approximated as the probability of the word given previous *n* words using *Markov assumption*. When substituting (3) into (2), the equation becomes:

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k \mid w_{k-1}) \tag{4}$$

2.2.2 *N*-grams Models

Based on the *Markov assumption,* bigram models simplify the probability of one word given all previous words by the probability of one word given only one previous word. For example, the probability of

P (congestion | a female presenting with a chief complaint of nasal) (5)

is substituted by the probability of

P (congestion | nasal) (6)

A *n*-gram model (which check *n-1* previous words) is a $(n\text{-}1)^{th}$ order Markov model. For example, trigram and four-gram models in this example are respectively:

$$P \text{ (congestion | of nasal)} \tag{7}$$

$$P \text{ (congestion | complaint of nasal)} \tag{8}$$

2.2.3 Statistical Estimators for *N*-gram Models

Probability can be estimated by relative frequency. One estimator of probability is called Maximum Likelihood Estimate (MLE). For example,

$$P_{MLE}(w_1 \cdots w_n) = \frac{C(w_1 \cdots w_n)}{N} \tag{9}$$

$$P_{MLE}(w_n \mid w_1 \cdots w_{n-1}) = \frac{C(w_1 \cdots w_n)}{C(w_1 \cdots w_{n-1})} \tag{10}$$

where N in equation (9) is the number of training instances. MLE is unsuitable for statistical inference in NLP due to the sparseness of the data. MLE assigns zero to unseen events, and the zeros will propagate since the probability of a long string is computed by multiplying probabilities of subparts. Therefore, smoothing techniques such as the Good-Turing discounting are used to compensate for data sparseness.

Good-Turing (GT) estimator attributes another better method for determining the probability or frequency of items:

$$If\ C(w_1 ... w_n) = r > 0,\ P_{GT}(w_1 \cdots w_n) = \frac{r^*}{N},\ where\ r^* = \frac{(r+1)S(r+1)}{S(r)} \tag{11}$$

$$If\ C(w_1...w_n) = 0,\ P_{GT}(w_1 \cdots w_n) = \frac{1 - \sum_{r=1}^{\infty} N_r \frac{r^*}{N}}{N_0} \approx \frac{N_1}{N_0 N} \tag{12}$$

In equation (11), S is the function that fits the observed values of (r, N_r) and $S(r)$ is the expectation of the frequency. This smoothing method substitutes low frequency n-grams and is quite accurate. This method is suitable for large numbers of observations of data and assumes that the distribution is binomial. GT estimator works well for n-grams, although words and n-grams do not have a binomial distribution.

Ney and Essen proposed two additional discounting models for estimating frequencies of n-grams with sparse data. One is absolute discounting: If $C(w_1 \cdots w_n) = r$,

$$P_{abs}(w_1...w_n) = \begin{cases} (r - \delta)/N & if\ r > 0 \\ (B - N_0)\delta / N_0 N & otherwise \end{cases} \tag{13}$$

where δ is a small constant number for all non-zero MLE frequencies and B is the number of target feature values.

Another is linear discounting: If $C(w_1 \cdots w_n) = r$,

$$P(w_1...w_n) = \begin{cases} (1 - \alpha)r / N & if\ r > 0 \\ \alpha / N_0 & otherwise \end{cases} \tag{14}$$

where α is a constant slightly less than one.

These estimates make the probability of unseen events a small number instead of zero and rescale the other probabilities. The absolute discounting approach is very

successful, while the linear discounting is hard to justify. The linear discounting method does not even approximate higher frequencies.

2.2.4 Cross Entropy and Perplexity

Entropy, *H*, is the average uncertainty of a single variable. It is used to measure the amount of information in a random variable. It is expressed as:

$$H(p) = H(X) = -\sum_{x \in X} p(x) \log_2 p(x) \tag{15}$$

Cross entropy measures the closeness between a random variable X with true probability distribution $p(x)$ and a model. Cross entropy is inversely related to the average probability of words that a model assigns in the test data. Cross entropy of a language $L = (X_i) \sim p(x)$ containing a sequence of n words by using a language model m is defined as:

$$H(L,m) = -\lim_{n \to \infty} \frac{1}{n} \sum_{x_{1n}} p(x_{1n}) \log m(x_{1n}) = -\lim_{n \to \infty} \frac{1}{n} \log m(x_{1n}) \approx -\frac{1}{n} \log m(x_{1n}) \tag{16}$$

Another measure that related to the cross entropy is the perplexity, defined as $2^{H(L,m)}$. Both cross entropy and perplexity can be used to detect how similar between the language for training the model and the test language. For example, lower cross entropy or perplexity indicates better generalization of the model to the test language, on the other words, the test language is more likely similar with the language used for training the model. While higher cross entropy or perplexity implies new information in the test language that was not used to estimate the parameters of the model. Thus, the perplexity

allows one to estimate the degree to which the texts in a given clinical report are similar to the texts in the prior clinical notes.

2.3 Biomedical Knowledge Bases and Existing NLP Tools

Biomedical knowledge bases provide resources for researchers to implement NLP applications within biomedical and health domain. One of widely used knowledge bases in the United States is the UMLS, which brings many health and biomedical vocabularies and standards together to support biomedical research (31). UMLS consists of the Metathesaurus, Semantic Network, and SPECIALIST lexicon.

Metathesaurus includes over 100 biomedical vocabularies, code sets and thesauri, covering comprehensive vocabularies (*e.g.*, Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT)), Medical Subject Headings (MeSH), laboratory and observational data (*e.g.*, Logical Observation Identifier Names and Codes (LOINC)), diseases (*e.g.*, International Classification of Diseases and Related Health Problem (ICD)), and procedures and supplies (*e.g.*, Current Procedural Terminology (CPT)). UMLS assigns each biomedical concept with a Concept Unique Identifier (CUI). Multiple terms from different vocabularies with the same concept or meaning share a unique CUI.

The Semantic Network provides a hierarchically structured ontology of biomedical knowledge. It currently consists of 54 relationships between 135 semantic types. Each concept in the Metathesaurus is assigned at least one sematic type. There are five major relationships including physical (*e.g.*, PART_OF), spatial (*e.g.*,

LOCATION_OF), temporal (*e.g.*, PRECEDES), functional (*e.g.*, TREATS), and conceptual (*e.g.*, DIAGNOSES).

The SPECIALIST NLP tools are computer programs designed specifically for various biomedical NLP applications. These tools include the lexical variant generator (LVG), normalized string generator (Norm), word index generator (Wordind), dTagger POS tagger and others. For example, LVG is an application to perform lexical transformation of words. MetaMap is another program to automatically map biomedical concepts in the text to the UMLS Metathesaurus (32). MetaMap provides various options, such as acronyms, abbreviations, negation detection, and word sense disambiguation, to meet the needs of NLP researchers. Researchers can also integrate MetaMap APIs into their applications to develop advance NLP tools.

2.4 Semantic Similarity

Similarity is a fundamental concept that is essential to automated information integration, case-based similarity, inference, and information retrieval tasks (33, 34). With respect to assessing similarity at a patient or case level, effective metrics quantify how similar two patients are to one another, based upon the question at hand (such as overall similarity or similarity from a diagnostic standpoint) (35, 36). These comparisons have been conceptualized as measures of similarity based upon complex *sets* of concepts representing each case and has been classically described as a commutative or symmetrical measure (*i.e.*, Similarity of (A,B) = Similarity of (B,A)). However, in the context of measuring information (semantic) redundancy in EHR documents that are

created sequentially over time, asymmetrical measures may be required as information is continually added and/or repeated in more recent notes compared to older documents.

In contrast to case-based similarity, concept-level semantic similarity metrics quantify the closeness in meaning between two concepts (versus two groups of concepts such as in the example of case-based similarity) by determining the closeness of concepts using different measures such as closeness in a hierarchy (37, 38). Semantic similarity has been studied extensively both in general language and in biomedicine. Automated measures of semantic similarity can be generally classified into knowledge-based approaches and knowledge-free approaches. Knowledge-free approaches rely upon statistical measures such as term frequency and co-occurrence data. Because of the complexity of the medical domain, rich use of synonymy and related concepts, the performance of knowledge-free approaches may not be optimal. Knowledge-based approaches utilize additional information, such as ontological information, definitional data, or domain information to enhance these methods.

In the context of automated measurement of information redundancy, measures of semantic similarity may be useful to perform semantic normalization between pieces of text that are being compared to determine the degree of redundancy. For example, theoretically, it may be useful to treat orthographically different but semantically synonymous or highly similar terms as equivalent (*e.g.*, *heart* vs. *cardiac*) when comparing two texts to identify new information in this research, thus potentially increasing accuracy of methods. Several groups have developed methods to measure semantic similarity based on various types of relationships between concepts, such as

broader/narrower (RB/RN) (39), parent/child (PAR/CHR) (40) and is a relations (41, 42). To provide an open-source framework for measuring semantic similarity and comparing results using various methods, the UMLS::Similarity package was developed based on the UMLS (38).

Semantic similarity measures are generally classified into two categories: path-based measures, and information content (IC)-based measures.

2.4.1 Path-based Similarity Measures

Path-based measures depend on the length of the path between two concepts in a biomedical vocabulary.

1) Rada developed a measure, called conceptual distance (CDist), based on the path length of two concepts in MeSH using RB/RN relations (39). Caviedes and Cimino further implemented this method to examine this measure on MeSH, SNOMED and ICD9 (40). The similarity score is:

$$score_{path}(c_1,c_2) = \frac{1}{length(c_1,c_2)} \tag{17}$$

where $length(c_1,c_2)$ is the shortest path length between the two concepts.

2) Wu and Palmer considered the depth of two concepts in the UMLS and the least common subsumer (LCS) using *is-a* relations (43). The method they developed was:

$$score_{wup}(c_1,c_2) = \frac{2 \times depth(lcs(c_1,c_2))}{depth(c_1) + depth(c_2)} \tag{18}$$

where $depth(lcs(c_1, c_2)$ is the depth of the LCS of the two concepts.

3) Leacock and Chodorow extended the path measure by using the depth of the taxonomy (44). They measured similarity by:

$$score_{lch}(c_1,c_2) = -\log(\frac{\min path(c_1,c_2)}{2 \times D}) \quad (19)$$

where min*path* is the shortest path between two concepts and *D* is the maximum depth of the taxonomy.

4) Nguyen and Al-Mubaid incorporated both the depth and LCS in the measure (45) in their similarity measure:

$$score_{nam}(c_1,c_2) = \log_2([length(c_1,c_2)-1] \times [D - depth(lcs(c_1,c_2)) + 2]) \quad (20)$$

All of the above path-based measures provided computationally simple approaches for determining the degree of semantic similarity of two concepts.

2.4.2 IC-based Similarity Measures

The Information Content (IC) of a concept is defined as negative the log likelihood of $P(c)$, $-logP(c)$, where $P(c)$ is the probability of the concept c[46]:

1) Resnik defined the similarity of two concepts as the IC of their LCS (46):

$$score_{res}(c_1,c_2) = IC(lcs(c_1,c_2)) \quad (21)$$

2) Jiang and Conrath used the IC of each individual concept and their LCS to estimate similarity (47):

$$score_{jcn}(c_1,c_2) = \frac{1}{IC(c_1) + IC(c_2) - 2 \times IC(lcs(c_1,c_2))} \quad (22)$$

3) Lin extended the similarity score as (48):

$$score_{lin}(c_1,c_2) = \frac{2 \times IC(lcs(c_1,c_2))}{IC(c_1) + IC(c_2)} \quad (23)$$

Compared to path-based similarity measures, IC-based measures are relatively computationally expensive, but they have been demonstrated to provide statistically significantly higher accuracy than the path-based measure (49).

2.5 Information Redundancy and Clinically Relevant New Information

In contrast to semantic similarity, information redundancy (at the semantic level) between two items has been studied less. Information redundancy is conceptually a measure of the degree of identical and/or redundant information in an item of interest (subject item) contained within another item (target item). For example, when comparing subject item A to target item B, the redundancy of information within subject item A contained in target item B (*i.e.*, $R(A,B)$) is conceptually the information contain in both A and B, normalized by the information in item A -- |A and B|/|A|. Measures of similarity (whether individual concepts or sets of concepts) are commutative, as the subject and target item (concepts or sets of concepts) can be interchanged. However, redundancy therefore depends upon which item is subject and target and is not commutative. As shown in left panel of Figure 2-1, $R(A,B) \neq R(B,A)$.

Identification of relevant, new (*i.e.*, non-redundant) information has been largely unexplored. New information in a target item is information that is different from previously mentioned information or has not been mentioned previously in the subject item(s). In contrast, relevant information in a target item is information that is relevant to a particular task if it increases the likelihood of accomplishing the goal (in this case, a clinician understanding a given patient). Relevant new information is information in a target item not contained within subject item(s) relevant to a particular task, thus depending on the selection of target and subject items. For example, when comparing subject item(s) *S* and target item *T*, relevant new information of *T* is the information that was contained in *T* but not in *S* relevant to a task. The result is different when switching subject and target items. Therefore a relevant new information measurement is not commutative with *T* and *S*. If one compares a new note with an old note as shown in the Figure 2-2, the red region is the new information that is only contained in the new note. Within the new information, some information may not be useful for clinicians' synthesis of patients' conditions, for example, the visit location. So I defined relevant new information as the information that is new and clinically relevant for clinicians' judgment, as shown in the blue region in Figure 2-2.

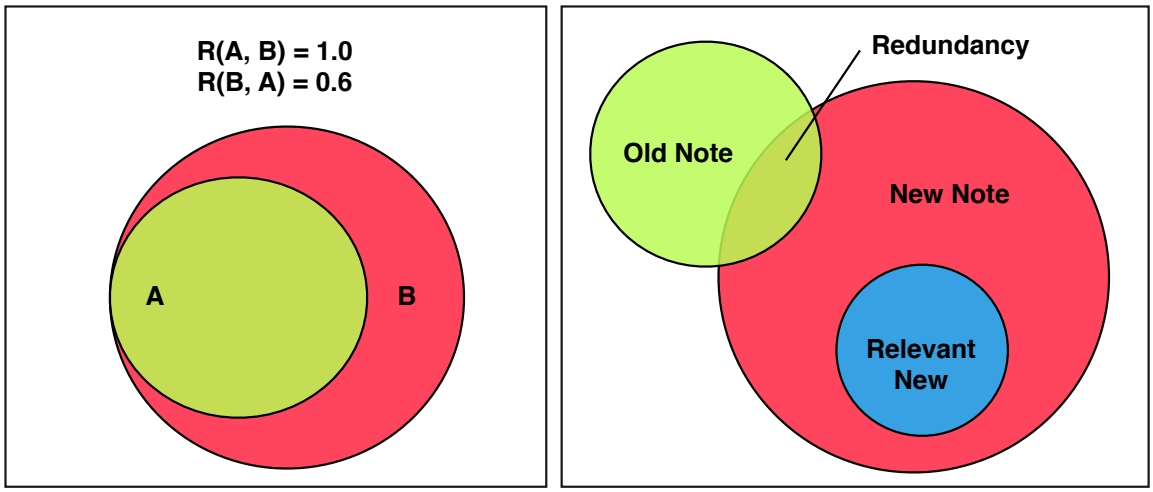


Figure 2-2. Venn diagram of information redundancy and relevant new information in clinical notes.

2.6 Charlson Comorbidity Index (CCI)

Comorbidity is defined as having one or more disorders other than a primary disorder. The Charlson comorbidity classification systems are some of the most common used in health research (50). CCI provides a simple and applied method for predicting the 10-year risk of mortality for patients based off of 22 conditions or disorders. CCI is a score summing the weighted scores (1, 2, 3, or 6) of each disease depending on the comorbidity group as shown in Table 2-1 (51).

Table 2-1. Enhanced ICD-9-CM coding algorithm and assigned points for the Charlson comorbidity index.

Comorbidities	Enhanced ICD-9-CM	Points
Myocardial infarction	410.x, 412.x	1
Congestive heart failure	398.91, 402.01, 402.11, 402.91, 404.01, 404.03, 404.11, 404.13, 404.91, 404.93, 425.4-425.9, 428.x	1
Peripheral vascular disease	093.0, 437.3, 440.x, 441.x, 443.1-443.9, 447.1, 557.1, 557.9, V43.4	1
Cerebrovascular disease	362.34, 430.x-438.x	1
Dementia	290.x, 294.1, 331.2	1
Chronic pulmonary disease	416.8, 416.9, 490.x-505.x, 506.4, 508.1, 508.8	1
Rheumatic disease	446.5, 710.0-710.4, 714.0-714.2, 714.8, 725.x	1
Peptic ulcer disease	531.x-534.x	1
Mild liver disease	070.22, 070.23, 070.32, 070.33, 070.44, 070.54, 070.6, 070.9, 570.x, 571.x, 573.3, 573.4, 573.8, 573.9, V42.7	1
Diabetes without chronic complication	250.0-250.3, 250.8, 250.9	1
Diabetes with chronic complication	250.4-250.7	2
Hemiplegia or paraplegia	334.1, 342.x, 343.x, 344.0-344.6, 344.9	2
Renal disease	403.01, 403.11, 403.91, 404.02, 404.03, 404.12, 404.13, 404.92, 404.93, 582.x, 583.0-583.7, 585.x, 586.x, 588.0, V42.0, V45.1, V56.x	2
Any malignancy, including lymphoma and leukemia, except malignant neoplasm of skin	140.x-172.x, 174.x-195.8, 200.x-208.x, 238.6	2
Moderate or severe liver disease	456.0-456.2, 572.2-572.8	3
Metastatic solid tumor	196.x-199.x	6
AIDS/HIV	042.x-044.x	6

CHAPTER 3 METHODS

In this chapter, I will introduce two different systems: 1) modification of global alignment to investigate redundancy characteristics in outpatient clinical notes; 2) applied statistical language models to identify relevant new information in longitudinal clinical notes. For each system, I firstly developed the methods and evaluated them by expert-annotated reference standard, and then applied the best method to investigate the specific topics.

3.1 Modifying Global Alignment to Detect Information Redundancy

The objective of this part is to explore several possible approaches for measuring redundancy in clinical text, particularly between longitudinal notes for the same patient. To this end, I developed an expert-derived reference standard of redundancy, several redundancy metrics with modification of classic dynamic programming global alignment techniques, and enhanced these metrics using both statistical and knowledge-based tools.

3.1.1 System Design

The process of method development includes four parts: data collection, methods implementation, reference standards, and methods evaluation, as shown in Figure 3-1. Methods implementation part also consists of text pre-processing, baseline redundancy measurements and enhancements to baseline redundancy measures. More details of each part will be provided as below.

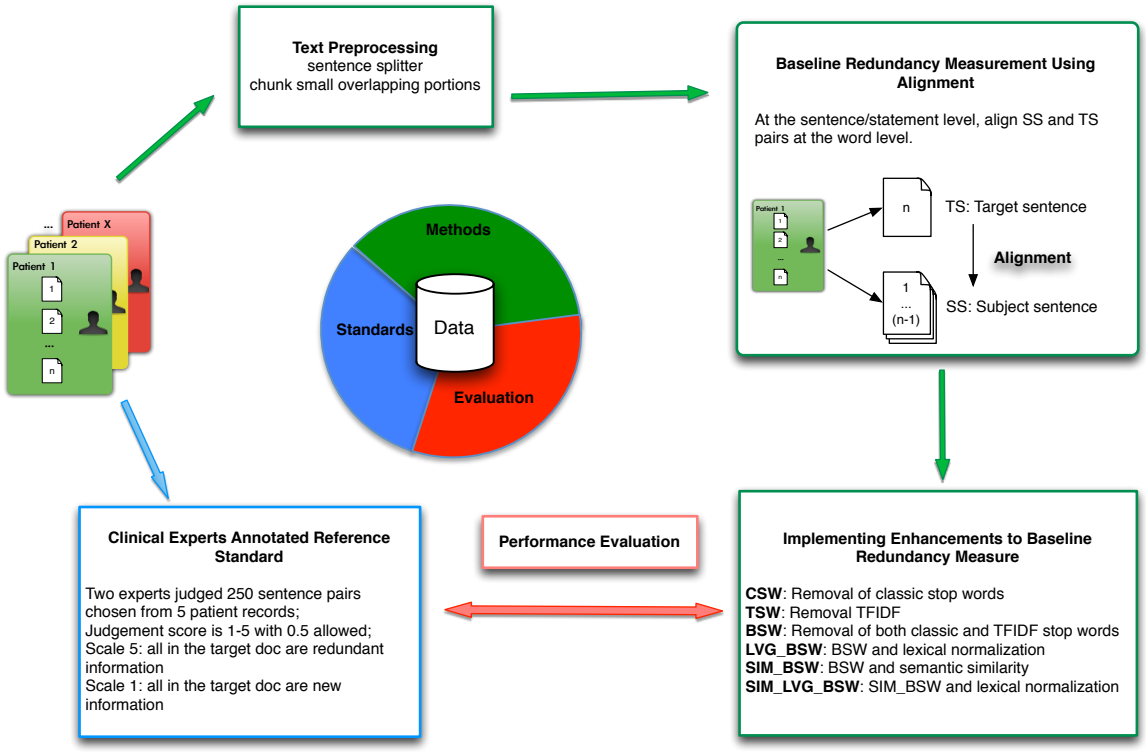


Figure 3-1. System architecture of method development for detecting redundant information.

3.1.2 Study Setting and Data Preparation

One hundred and seventy-eight complete outpatient clinical records from University of Minnesota Medical Center, Fairview Health Services in patients with angina, diabetes, or congestive heart failure followed by the Pharmaceutical Care Department for optimal medication management were used for this study. Each complete outpatient record contained all clinical notes including office visits, allied health nursing notes, telephone encounters, and results during a one-year period from December 2008 to November 2009. These notes were originally created in the Epic EHR system and extracted in text format. Inpatient notes from any of the Fairview Health Services hospitals were excluded

for the purposes of redundancy measurement for this analysis. It is assumed that there was no redundant information in the first document. Intra-document redundancy and semantic alignment were not considered in this study. Outpatient notes were organized chronologically for each patient as detailed in the "Text pre-processing" section of this paper and utilized for this analysis. University of Minnesota institutional review board approval was obtained and informed consent waived for this minimal risk study.

3.1.3 Automated Redundancy Measures

As a baseline metric, alignment between two texts was performed using a modification of the Needleman-Wunsch algorithm, a dynamic programming technique commonly used in the bioinformatics field to align protein or nucleotide sequences. This algorithm was modified to the constraints of clinical notes, as described, with text pre-processing and a sentence/statement alignment process at the word level (as opposed to a character level). I present an overview of this measurement's processing in Figure 3-2.

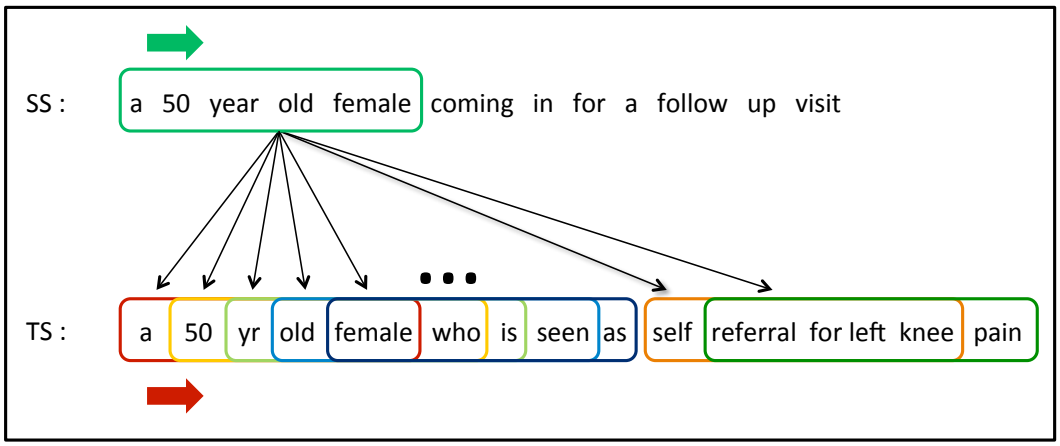
A. Sentence Pair

SS	a 50 year old female coming in for a follow up visit.
TS	a 50 yr old female who is seen as self referral for left knee pain.

B. Prior Alignment

SS	a 50 + year old female + + + + + + coming in for + + a follow up visit.
TS	a 50 yr - old female who is seen as self referral - - for left pain.

C. Window Sliding



D. Progressive Modification of Sentences

	Subject Sentence (SS)
Baseline	a 50 year old female coming in for a follow up visit
CSW	50 year old female coming follow up visit
TSW	50 year old female coming follow up
BSW	50 year old female coming follow up
LVG_BSW	50 year old female come follow up
SIM_BSW	50 year old female come follow up
SIM_LVG_BSW	50 year old female come follow up
	Target Sentence (TS)
Baseline	a 50 yr old female who is seen as self referral for left knee pain
CSW	50 yr old female who seen as self referral left knee pain
TSW	50 yr old female who seen as self referral left knee pain
BSW	50 yr old female who seen as self referral left knee pain
LVG_BSW	50 yr old female who see as self referral left knee pain
SIM_BSW	50 year old female who see as self referral left knee pain
SIM_LVG_BSW	50 year old female who see as self referral left knee pain

Figure 3-2. Schematic of automated redundancy measures between a subject sentence (SS) and a target sentence (TS). A) Sentence pair; B) Global alignment including matches, additions (+) and subtractions (-); C) First frame of SS align with all frames (differed by color, omit few frames in the middle) of TS; D) Sentences are modified by various measure.

3.1.4 Baseline Sentence/Statement Redundancy Measurement Using Alignment

The content of one text of interest, the *subject text*, was compared to another text of interest, the *target text*. At the sentence or statement level, each pair of subject and target sentences (SS and TS respectively) was aligned at the word level. The Needleman-Wunsch algorithm was modified to align a window in the SS to the TS in an iterative fashion. The alignment steps were as follows, 1) each SS was split into overlapping frames of five consecutive words; 2) each frame of the SS was then aligned with all frames of the TS by advancing a sliding window one word at a time (Figure 3-2C); 3) the maximum alignment score for a pair of frames was defined as the number of matched words for each frame of the SS with penalties of word addition or deletion; 4) the window positioned over the SS text was then advanced by one word and aligned as described in the previous step; 5) the final alignment score for the SS text as compared to the TS text was calculated by averaging all maximum TS scores for each SS frame; 6) scores were then normalized by the window size and used to build up a baseline redundancy matrix between pairs of SS vs. TS texts (score range from 0 to 1). Baseline measures were then used to perform stratified random sampling to create the reference standard sentence pair set.

3.1.5 Reference Standard

Two hundred and fifty sentence pairs selected based on the baseline redundancy scores were chosen from five patient records using stratified random sampling. The sampling consisted of splitting the pairs of sentences in each record into quintiles and then selecting five random sentence pairs with scores in each of the quintiles. Two physicians

were asked to judge the redundancy of information on a scale from 1 to 5, with "0.5" scores allowed (*e.g.*, 2.5). The physicians were asked to base their assessments on how much information contained in the SS text they were able to also find within the TS text. They were also asked to compare the information content of the texts as opposed to just comparing the words. The highest score of 5 indicates that all the information in the SS was contained in the TS, while lowest score of 1 indicates that none of the information in the SS was contained in the TS. After calculating agreement, physician scores were averaged to form the reference standard to validate the automated scoring methods.

Inter-rater reliability between the two experts was calculated using Cronbach's Alpha (52). Inter-rater agreement with Cronbach's Alphas was 0.91. The correlation between the expert ratings was also high (0.871 shown in Table 4-1). Expert ratings were averaged to create the reference standard for evaluating all the methods. Both 1) expert evaluations were correlated to one another as a measure of optimal upper-bound performance and 2) automated redundancy measures were correlated to the reference standard using the non-parametric Spearman's rank correlation coefficient (53).

3.1.6 Implementing Enhancements to Baseline Redundancy Measure

Each technique was implemented by using a window size of 5 words. The choice of this window size was motivated by prior work showing that the average length of medical terms found in outpatient clinical notes is between 4 and 5 words (54). In addition to the baseline redundancy metric, which aligned unaltered raw text (Baseline), I experimented with the following modifications of the baseline redundancy measure:

1) Removal of classic stop words (55) (CSW). This method was based on removing stop words (*e.g.,* “the”, ”a”, ”for”, ”it”, ”this”, etc.) that are generally removed by text indexing and retrieval systems.
2) Removal of stop words defined by Term Frequency–Inverse Document Frequency (TFIDF) using optimal thresholds of the TFIDF distribution based upon the entire note corpus (TSW). TFIDF is another method used in standard text indexing and retrieval systems to remove or deemphasize words that occur frequently in many documents and thus are less likely to be useful for ranking documents by their relevance to a query.
3) A combination of CSW and TSW, with removal of both classic stop words and stop words defined by optimal TFIDF thresholds (BSW).
4) Removal of both stop word types (BSW) and lexical normalization to effectively treat lexically different forms of the same term as equivalent when aligning text using Lexical Variant Generation (LVG) (56) (LVG_BSW).
5) Removal of both stop words (BSW) and treating terms with high semantic similarity as equivalent when aligning text using the UMLS and path-based UMLS::Similarity measures (23) (SIM_BSW). For this, a cut-off score of 0.8 was used from the UMLS::Similarity measure to identify synonymous or near-synonymous terms (*e.g.*, “above” – “upper”, “advice” – “guidance”, etc.).
6) Removal of both stop word types (BSW), aligning text using UMLS::Similarity, and lexical normalization using LVG (SIM_LVG_BSW).

As an additional baseline, texts were aligned as described previously by Wrenn et al. (14) using the Levenshtein edit-distance algorithm at a word-level without window movements (Figure 3-2B). The window size was also examined as a factor and was varied to increments of 4, 5, 8, and 10. Figure 3-2 shows two example sentences and illustrates how each approach was implemented.

3.1.7 Investigating Redundancy Patterns

For each patient, all notes were arranged chronologically and each note of interest (target note) was compared with all the previous notes (subject notes). At a document level, windowing was not allowed to cross sentence boundaries so as not to penalize for not preserving information across sentences. The score for each frame was defined as the maximum score with the automated method, comparing the target frame text with the text from all previous notes using LVG_BSW method. Using this technique, a set of frame scores and their distribution were created for each note. A mean score was assigned for each note by averaging all frame scores. Based on these mean scores, the redundancy between documents was derived first as descriptive statistics of scores over all the documents. A physician (GM) examined three patient records and recorded the purpose of each visit and any noteworthy clinical events. These observations were then overlaid graphically with the mean redundancy scores of documents chronologically.

Last, average document redundancy scores for patients over time were calculated to detect temporal redundancy trends. Redundancy scores for each patient document were normalized to account for different numbers of notes in each record. Normalization was performed by pooling redundancy scores for each patient into even quartiles

chronologically over the entire time period, so that the first 25% contained the earliest notes and fourth 25% had the most recent notes. Using the approximately 900 clinical note corpus from 178 patients, each data point with standard error bars illustrates the average redundancy of more than 200 clinical notes. A smoothed curve along with the original data points were included to visualize the trend in redundancy scores across patient notes with time.

3.2 Applying Statistical Language Models to Identify and Visualize Relevant New Information

In this section, I will introduce the method development for identifying relevant new information in longitudinal clinical notes. The goal of this study is to investigate techniques to identify relevant new information and demonstrate visualization of relevant new information within clinical texts.

3.2.1 System Design

Nine patient records (each patient record contained at least 10 longitudinal notes) were selected for this study. The experimental design and system architecture are illustrated in Figure 3-3, with the system being developed with four patient records ("training records") and the system tested on the remaining five records ("test records"). The workflow consisted of: 1) collecting patient documents and document metadata from the clinical document repository; 2) text preprocessing; 3) application of *n*-gram models and various enhancements trained on *n* previous documents to identify relevant new information of the $(n + 1)^{th}$ document for a given patient; 4) creation of an expert-derived

reference standard with expert manual annotation; and 5) evaluation of automated method performance.

3.2.2 Data Preparation

Medical records from University of Minnesota Medical Center, Fairview Health Services were used in this study. These notes were extracted in text format from the *Epic* EHR system (57), which were created during a one-year period (12/2008 to 11/2009). Outpatient notes (*i.e.*, office visits, allied health notes, telephone notes, results) were arranged chronologically. Institutional review board approval was obtained and informed consent waived for this minimal risk study.

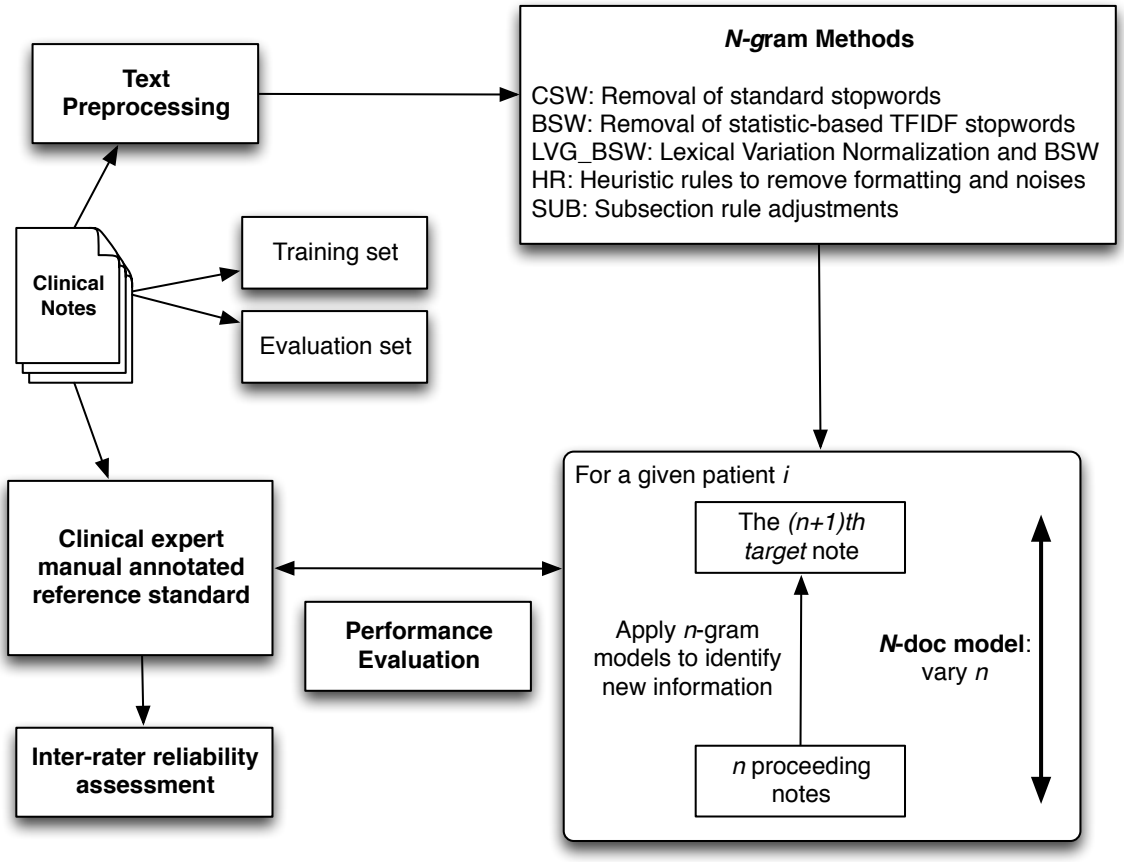


Figure 3-3. Experimental design and system architecture for statistical language models to identify relevant new information.

3.2.3 Text Pre-processing

Since not all sentences in *Epic* clinical notes are well-formed (*e.g.*, "review of system" may appear as the only text on a line or as part of an enumeration), incomplete sentences or "statements" were treated as sentences in this study. Each note was further separated into smaller chunks at a sentence/statement level.

3.2.4 Baseline Relevant New Information Identification Using Bigram Model

As a baseline metric, a single bigram language model was built based on all preceding unaltered documents (*e.g.,* from the 1^{st} to n^{th} document) to identify relevant new information within the target document (*e.g.,* the $(n + 1)^{th}$ document). The bigram counts of each bigram model were used to classify relevant new information in the target document. Initially, the count threshold value was set to zero for each bigram. For example, if $C(w_2 | w_1)$ did not appear in any subject documents, then w_2 following w_1 in the target document was considered as a new word.

3.2.5 Manually Annotated Reference Standard

Nine outpatient clinical records with ten office visits per patient record were selected for this study. Four records were used for training and developing the system (about 6,200 sentences and statements) and five records (including one evaluated by both experts) were used for evaluation (about 9,700 sentences and statements). Two physicians were asked to identify new and clinically relevant information within each document (starting from the second document) based on all the preceding documents chronologically for each patient record using their clinical judgment. Each medical expert annotated five patient records with one record overlap with both. New information in documents was annotated with the General Architecture for Text Engineering (GATE) (58), which allows for the annotation of text and XML files through a graphical user interface (GUI), with a customized annotation schema.

In order to measure agreement between two clinician experts in the task of identifying new information, the overlap between annotations was measured in one of the

nine outpatient clinical records manually annotated. Cohen's Kappa statistic and percent agreement (59) were used to assess inter-rater reliability of the two physicians judgments at a sentence or statement level. If one or more words were marked as relevant new information by experts, the whole sentence or statement was considered as relevant new information for evaluating methods. Performance of automated methods compared to the reference standard was then measured by accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) at a sentence or statement level for the five test records.

3.2.6 System Enhancements to Baseline

In addition to the baseline, which relied bigram counts from unaltered raw text using the bigram language model (Baseline), several modifications were explored (Table 4-1).

1) CSW: removal of classic stop words (55). This method was based on stop words often removed by text indexing and retrieval systems (*e.g.*, "the", "a", "for", "it", "this").
2) BSW: removal of both classic stop words and stop words defined by Term Frequency–Inverse Document Frequency (TFIDF) using optimal thresholds of the TFIDF distribution based on the entire note corpus. TFIDF is used to remove the deemphasized words that occur frequently in the corpus and thus are less likely to be useful for providing new information.
3) LVG_BSW: removal of both stop word types (BSW) and lexical normalization to effectively treat lexically different forms of the same term as equivalent when building up *n*-gram model using Lexical Variant Generation (LVG) (60).

4) HR: heuristic rules to remove clinical note formatting and other noise, as well as both stop word types and lexical normalization.
5) SUB: heuristic rule-based adjustments based on section content, removal of clinical note formatting and noise, as well as both stop word types and lexical normalization. Heuristic rules were created for several subsections. For example, *follow-up* was always treated as new information although the information could be repeated in different documents. (*e.g.*, "…return in 2 weeks." means two different weeks in different documents). Other known statements, such as previous allergies, were eliminated as possible new information (Table 3-1).

Table 3-1. Modification of methods and examples.

Baseline	A 50 year old female coming in for follow up visit
CSW	50 year old female coming follow up visit
BSW	50 year old female coming follow up
LVG_BSW	50 year old female come follow up
HR	
Noise	**Examples**
Visit Information	Date, Time, Provider, Department, Center, Previous Visit, Encounter Number.
Signatures	NAME, MD; NAME DATE Signed; Date Reviewed: DATE; Chart Reviewed by: NAME;
Medication Details	Class, Sig, Route, Special Instructions, Level of Service
Order Information	Priority, Class, Associated Diagnosis, Comments, Order #, Spec. #.
Note Information	Vitals History Recorded; History reviewed and updated in Epic, Progress Notes Scan on.
SUB	
Subsection	**Examples**
Follow-up (Relevant New)	FOLLOW-UP: patient will keep a food record and return in 2 weeks.
Past Medication	Prescription as of MM/DD/YYYY Current Outpatient Medication.
Clinic and Patient Information	FAIRVIEW MAPLE GROVE MEDICAL CENTER, Patient Information, Patient Demographics Address, Phone.
Allergies	Allergies As of Date.

3.2.7 Varying *N*-gram Models

Other *n*-gram models, including trigram ($n = 3$) and four-gram ($n = 4$), with the best performance bigram model system were tested.

Here, counts $C(w_3 | w_1 w_2)$ and $C(w_4 | w_1 w_2 w_3)$ were used for the trigram and four-gram model, respectively. Count threshold values (*e.g.,* 1 and 2) were also varied to assess effect.

3.2.8 *N*-doc Models

The model was applied to the *n* preceding documents and assessed algorithm performance. This model was built based on *n* preceding documents, "*N*-doc model", using the SUB bigram algorithm on the previous 1 to 9 documents. Assuming *n* is chosen as 3, for example, the language model was built only using the 3 previous documents.

3.2.9 Relevant New Information Visualization

As a gestalt result, documents in XML format were generated using the best method. These files were opened in GATE to show highlighted text and compare with the expert standard.

3.3 Quantifying Relevant New Information to Navigate Notes

In this section, I will introduce the methods to navigate clinical notes by using the identified relevant new information. Information navigation of electronic notes is essential for physicians reviewing a complex patient (with a long series of longitudinal clinical notes with historical medical information). The ability to highlight new and

relevant information in clinical notes provides clinicians with the ability to navigate notes more purposefully. Moreover, there is limited investigation into the sources of redundant information within a specific clinical note, which can be important in understanding the behaviors of clinicians in generating new clinical notes, as well as inform the development of future tools for new information identification.

The main aim of this study is to describe an automated method to quantify new information and navigate to notes with new information, and to investigate possible new information (or its inverse - redundant information) patterns for individual patient records. As a secondary aim, I also sought to understand "copy and paste" behaviors and to provide a potential method to navigate notes.

3.3.1 Data Collection

EHR notes were retrieved from University of Minnesota Medical Center affiliated Fairview Health Services. Similar to previous studies, patients with multiple co-morbidities were randomly selected, allowing for relatively larger numbers of longitudinal records in the outpatient clinic setting. These notes were extracted in text format from the Epic™ EHR system (57) during a six-year period (06/2005 to 06/2011). To simplify the study, the notes were only limited to office visit notes (Figure 3-4A). Each note was indexed based on chronological order (*e.g.*, note A1 indicates the 1st note of patient A). Institutional review board approval was obtained and informed consent waived for this minimal risk study.

3.3.2 Manually Reviewed Annotation as Gold Standard

Two medical interns (physicians aged 26 and 30) were asked to identify new information within each document (starting from the second document) based on all the preceding documents chronologically for each patient record using their clinical judgment. Each medical expert annotated five patient records with one record overlapping with both. Annotation of new information in clinical notes was implemented by using the General Architecture for Text Engineering (GATE) (58). GATE allows for the annotation of text and XML outputs through a graphical user interface, with a customized annotation schema.

To achieve a high-quality gold standard, the physicians were first asked to annotate one sample note (based on historical notes) and then compared and discuss the annotations with each other to reach a consensus on annotation standards for new information. Each physician later manually annotated another 10 notes based on the same historical notes to measure agreement. Cohen's Kappa statistic and percent agreement (59) were used to assess inter-rater reliability at a sentence or statement level.

Overall, longitudinal outpatient clinical notes from 15 patients were selected for annotation. To better evaluate the method, raters annotated the same last 3 notes as the target notes compared to historical notes of each patient's note set, but used different numbers of previous notes as the reference clinical history (*e.g.*, one used the previous 5 notes, the other used the previous 10 notes). Overall, each medical intern annotated 45 notes. Twenty of them were used for training and developing the system and another twenty-five for evaluation. Performance of automated methods was then compared to the

reference standard and measured for accuracy, precision, recall, and F-measure at a sentence or statement level.

Also, a 5th year resident physician manually reviewed two randomly selected patient records. Without seeing the results, the physician first reviewed the most current history (11-20th notes of a given patient) and then the target notes (21-38th). Any new information found in the target notes not in the previous 10 notes was noted. The physician then also reviewed another 10 historical notes prior (1-10th), and marked if there was any additional new information within notes 21-38th not recognized with review of the earlier notes. This annotation was then compared with the automatically computed redundancy results.

3.3.3 New Information Pattern Analysis

The automated method was first used to identify new information and quantify the new information proportion (NIP) in each note. The principle method used for this study is based on the updated *n*-gram statistical language models reported previously (61) and described in Section 3.2. In short, it was a bigram language model with classic stopword removal, term frequency-inverse document frequency (TF-IDF) stopword removal, application of lexical variation generation (LVG), and the adjustment of the model through several heuristic rules.

The developed computational model was used to identify new information in all notes (starting from the 21st note) based on the previous n (=1, 2 … 20) longitudinal clinical notes. The matrix of NIP (number of sentences with new information/number of all sentences per note) with the dimension of 2,918 × 20 was then obtained, where 2,918

is the number of all notes having at least 20 historical notes for the whole 100 patient corpus. For example, the orange cell in column 20 and row Nn_j (Figure 3-4B) represents the NIP contained in the note Nn_j (Figure 3-4A) calculated based on the previous 20 clinical notes. Thus, each row represents each note, and columns are the corresponding numbers of previous notes used in the language model to predict new information in that target note. This matrix was used to investigate the impact of the number of previous clinical notes in the model on the NIP scores. Twenty arithmetic means were obtained by averaging NIP scores in each array, and correlated with the previous note numbers to find the relationship.

Notes of the same patient were clustered as a group (*e.g.*, longitudinal notes of the patient A: A21, A22 … A38) and averaged notes sharing the same note index from different patients (*e.g.*, 21st notes from all patients: A21, B21 … N21) to get representative NIP scores based on all patient notes. NIP scores were then plotted to investigate the overall patterns of how new information changed over time.

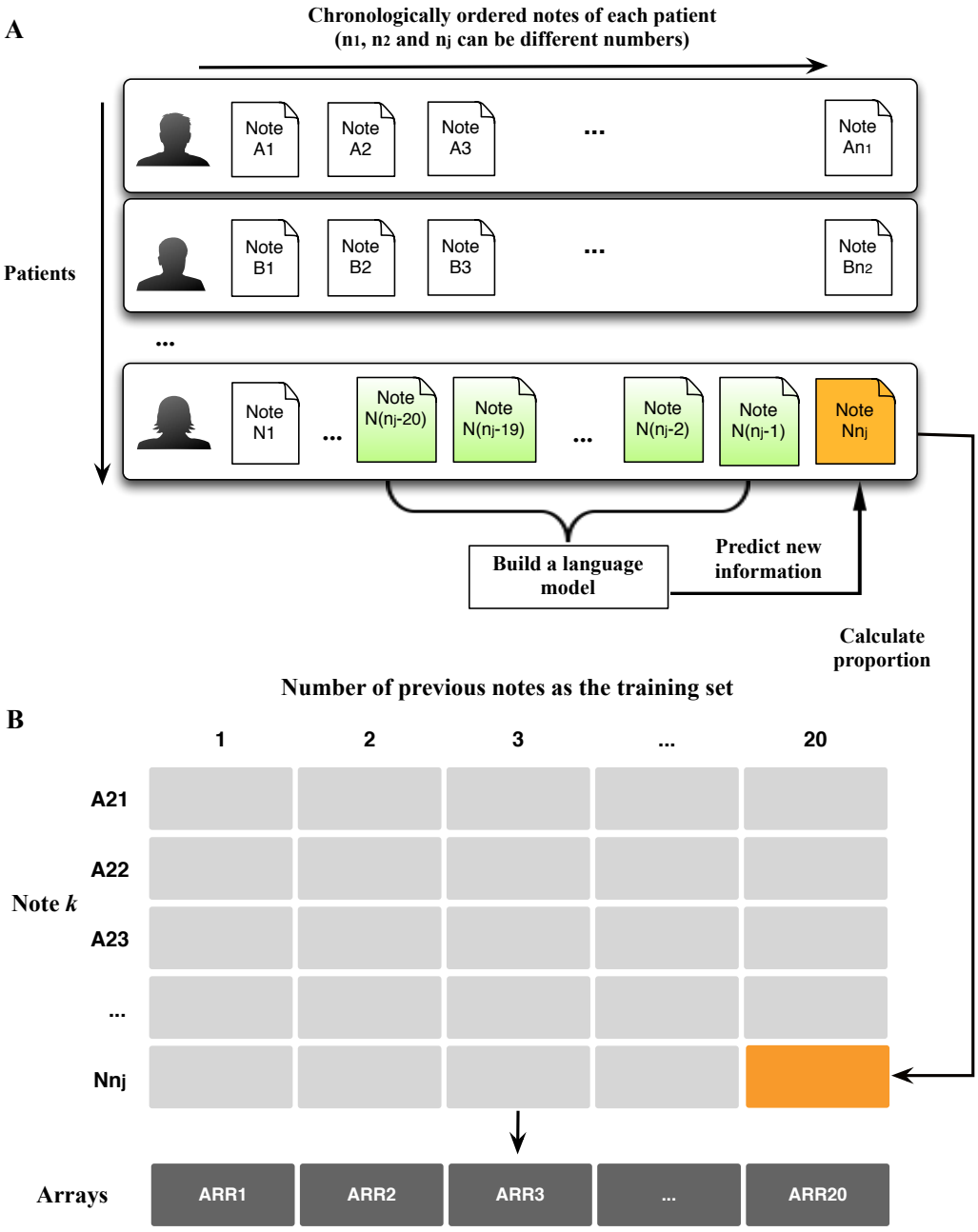


Figure 3-4. (A) longitudinal data set; (B) score matrix of new information proportion (NIP). Build a language model to calculate the NIP of note *k* (A) and generate the corresponding cell in the matrix (B).

3.4 Classifying Semantic Types of Relevant New Information

In this section, I will introduce the methods to classify the types of new information. Prior studies have focused upon methods to identify relevant new information and utilization of the identified new information on note navigation. One of the recognized gaps in these approaches is that these methods do not intrinsically provide more details about the types of new information (*e.g.,* medication, disorders, symptoms). Categorization of new information might aid clinicians in finding specific types of new information more easily and purposely within notes. The objective of this study was to extract specific types of relevant new information, specifically problem/disease (or comorbidities), medication, and laboratory and evaluate this approach by relating the appearance of new disease information with the Charlson comorbidity index calculated at the same time point.

3.4.1 System Design

The methodological approach for this study included: 1) modifying the reference standard to include the information type; 2) identification of new information using an *n*-gram modeling technique; 3) extraction of semantic types of identified new information; 4) calculation of the comorbidity index; and 5) correlation of CCI with new disease information.

In the study, CCI was used as a way to determine the time points at which the burden of disease for each given patient was increasing and subsequently to relate these time points to the appearance of new information in clinical notes. It is assumed that a

new diagnosis adding to the CCI score should also result in introduction of identifiable new information into the clinical notes.

3.4.2 Data Collection

Outpatient EHR notes were retrieved from the University of Minnesota Medical Center affiliated Fairview Health Services. For this study, 100 geriatric patients with multiple co-morbidities were randomly selected, allowing for relatively large numbers of longitudinal notes in the outpatient clinic setting. To simplify the study, only office visit notes were limited and arranged chronologically. These notes were extracted in text format from the Epic™ EHR system (57) between 06/2005 and 06/2011. Institutional review board approval was obtained and informed consent waived for this minimal risk study.

3.4.3 Automated Methods to Classify New Information Types

The *n*-gram models described previously (Section 3.2) were used to identify new information (61). In brief, after text pre-processing, *n*-gram models with classic and TF-IDF stopwords removal, lexical normalization, and heuristic rules to remove note formatting and adjustments by section were performed.

After obtaining new information within each note, this text was mapped to the UMLS (31) using MetaMap (32) with the options to allow acronym/abbreviation variants (-a) and NegEx results (--negex). From this, semantic types were extracted using scores of 600 and over as the cutoff. To simplify the analysis, the detailed analysis was

restricted to the specific types to identify information about problem/disease, medication, and lab results (Table 3-2).

Table 3-2. Sections and semantic types for identifying category of new information.

Category	Semantic Types
Problem/Disease	[Disease or Syndrome], [Finding], [Sign or Symptom]
Medication	[Clinical Drug], [Organic Chemical, Pharmacologic Substance], [Biomedical or Dental Material]
Laboratory	[Laboratory Procedure], [Therapeutic or Preventive Procedure], [Diagnostic Procedure], [Amino Acid, Peptide, or Protein], [Biologically Active Substance]

3.4.4 Manually Reviewed Annotation as Gold Standard

A resident (3rd year) manually reviewed chronologically ordered office visit notes from five individual patients to identify new information. This annotation was then compared with the automatically computed new information proportions (NIP) and extracted biomedical terms of various categories.

3.4.5 Calculating Various Types of New Information Proportion of Patient Notes

To calculate the NIP of each note, the method was trained on previous n (*e.g.*, 1, 2, …) notes to predict the new information of $(n+1)^{th}$ note for the whole corpus (100 patients). NIP was defined as the number of sentence (at least contain one piece of new information) divided by the total number of sentences of each note. NIP on the number (at a sentence or statement level) of various types of NIP (*e.g.*, NDIP for new diseases, NMIP for new medications, NLIP for new lab results) for each note were further

quantified. New information proportion of various types for each patient over time was then plotted. For the purposes of graphical display of notes temporally, the dates of patient notes by a random offset of +/- 1 to 364 days were adjusted.

3.4.6 Calculating Patients' Temporal CCI

All the International Classification of Disease-9th clinical modification revision (ICD-9-CM) codes were extracted from EHR records for each patient at each visit. Charlson comorbidities based on the enhanced ICD-9-CM coding algorithm (Table 2-1) were used to assign points to each note. Temporal CCIs for each patient were collected for further correlate CCI with new information identification for problems and diseases.

3.4.7 Correlation Between CCI and New Diseases Information Proportion (NDIP)

To correlate NDIP with CCI, new disease information was first extracted and calculated the proportion of this type of new information in each note. Whether trends of CCI scores with NDIP (*e.g.*, the time point of score increase) was checked to verify the correctness of the method and also examined if there was a potential use in clinical settings. To correlate the direction of new information score change in each note, both CCI and NDIP were further translated to the trend scores. To implement this, all scores for each patient were firstly ordered chronologically and each score was reassigned as -1, 0, or 1 when the score was lower than, the same as, or higher than the score at the previous time point. Mathematically, assuming that there are n clinical notes for a given patient j, both CCI and NDIP scores can be translated to *Trans_score* (*e.g.*, TCCI and TNDIP, respectively) based on the formula:

$$Trans\_score_{j,k} = \begin{cases} -1 & if\ score_{j,k} < score_{j,k-1} \\ 0 & if\ score_{j,k} = score_{j,k-1} \\ 1 & if\ score_{j,k} > score_{j,k-1} \end{cases}, \ where\ k \in [1,n]$$

To investigate if NDIP will change at the same point as CCI change, I checked only if NDIP had the same direction (increase or decrease) when CCI changed. In other words, whether TNDIP had the same value at points when TCCI changed was verified. The new generated trend scores TCCI and TNDIP at the turn points were then correlated to check if both had a similar trend with disease changes.

CHAPTER 4 RESULTS

In this chapter, I will present the obtained results by using the methods previously described and keep the same sequence as Chapter 3.

4.1 Modifying Global Alignment to Investigate Redundancy Patterns

I will show the evaluation results of automated methods and some findings of the redundancy patterns in the outpatient clinical notes.

4.1.1 Evaluation of Automated Redundancy Measures

All measures and their correlation with the reference standard are listed in the Table 4-1. TF-IDF scoring to experiment with different thresholds for stop-word removal was performed. As illustrated in Figure 4-2, there were several potential stop-word cutoffs. The first 3 cutoffs of 2E-6, 4E-6, and 6E-6 were tested, and it was found that a cutoff of 2E-6 provided the highest correlation (0.780, 0.777, and 0.778 respectively).

Table 4-1 summarizes the correlations with the reference standard for each of the methods, including using LVG_BSW with different window sizes. Comparison between various methods for calculating redundancy scores on the reference standard showed that removing stop words with both the classic stop word list and the optimized TFIDF scoring yields higher correlations with human redundancy judgments than using global alignment or the baseline local alignment. Adding lexical normalization further improves correlation albeit by a small amount; in contrast, semantic normalization with

UMLS::Similarity path-based measure using a threshold of 0.8 does not either improve the correlation with the reference standard, nor does it make the correlation worse.

Table 4-1. Correlation of methods compared to reference standard.

Method	**Spearman Coefficient**
Experts*	0.871
Prior (global alignment)	0.759
Methods of redundancy	
Baseline (window 5)	0.781
CSW (window 5)	0.785
TSW (window 5)	0.780
BSW (window 5)	0.814
SIM_BSW (window 5)	0.816
LVG_BSW (window 5)	0.824
SIM_LVG_BSW (window 5)	0.823
Varying window size	
LVG_BSW (window 4)	0.834
LVG_BSW (window 5)	0.824
LVG_BSW (window 8)	0.803
LVG_BSW (window 10)	0.801

*Experts = correlation of ratings between two raters. Prior = Prior method (14); Baseline = unaltered raw text.

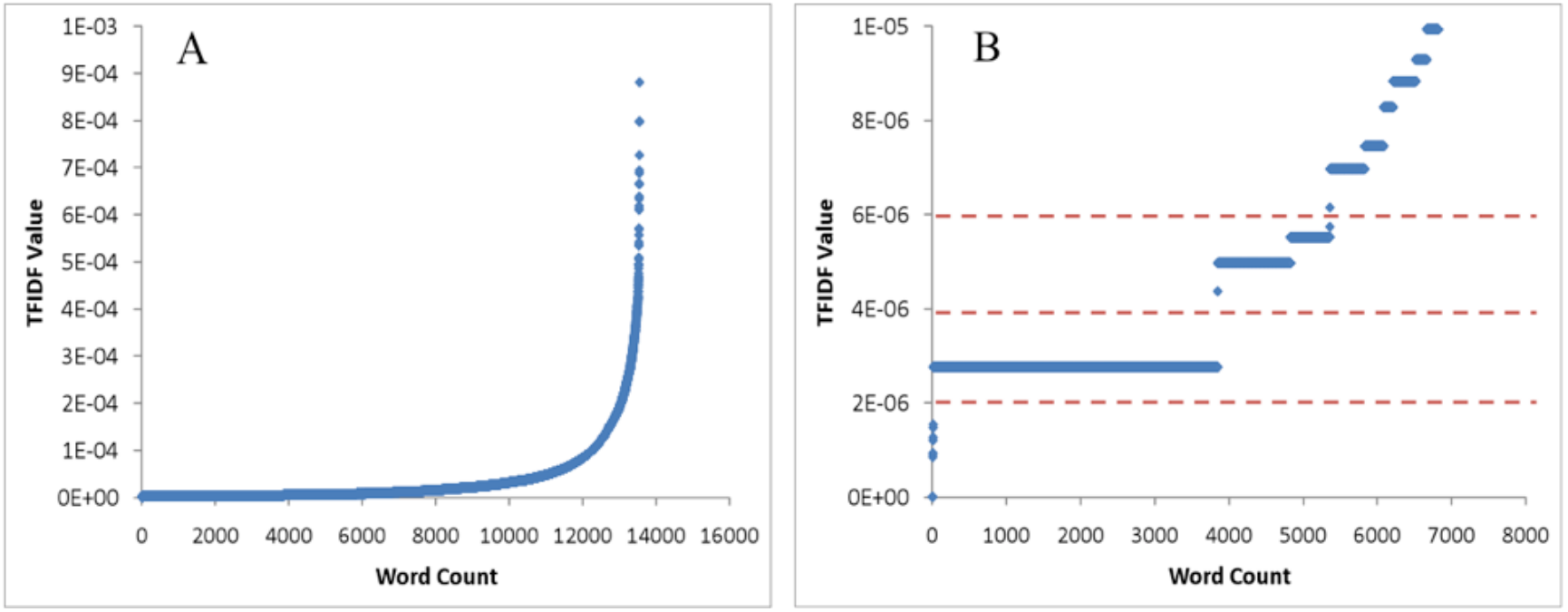


Figure 4-1. A) TFIDF value distribution of the whole corpus; B) Magnified view of TFIDF distribution showing three TFIDF cutoff values, which are marked as red dashed lines.

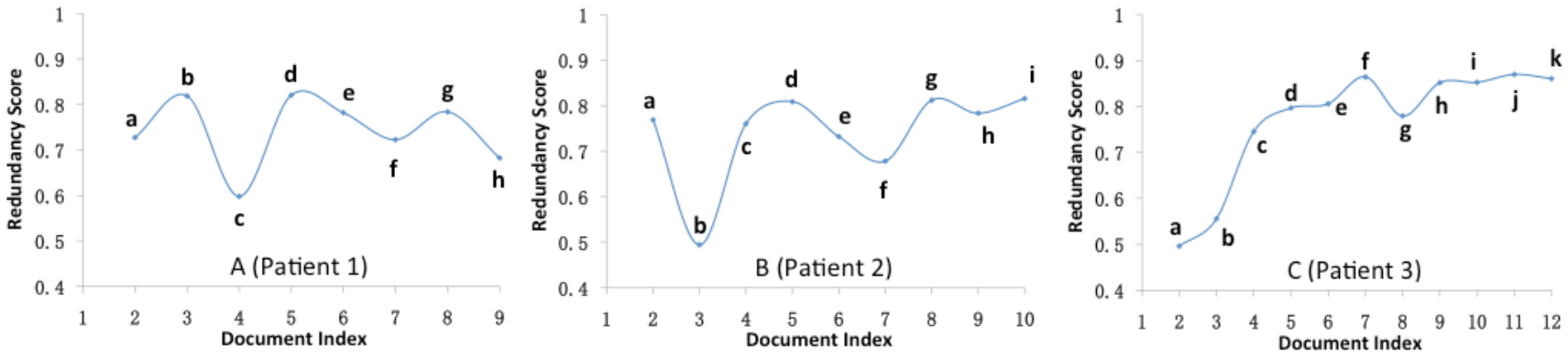


Figure 4-2. Patterns of redundancy scores in outpatient documents with documents indexed in a chronological order.

A) Patient 1: a) health maintenance visit; b) health maintenance visit, minimal upper respiratory tract symptoms; c) motor vehicle accident (MVA) with multiple musculoskeletal complaints, headache; d) follow-up of MVA symptoms; e) pre-operative general assessment for minor surgery; f) care following emergency department for congestive heart failure (CHF) exacerbation; g) health maintenance visit; h) visit for total body itchy rash, diagnosed with scabies.

B) Patient 2: a, c, & d) health maintenance visit; b) change in insurance and change in medication (short note); e) new upper respiratory tract infection (URI); f) urinary tract infection & fever; g) ongoing URI symptoms; h & i) diabetes-focused health maintenance visit.

C) Patient 3: a) right lower extremity (RLE) ankle tender and red (short note); b) recurrent RLE cellulitis and rash; c) follow-up of RLE symptoms; d, e, f, g, h, i, j, & k) health maintenance visit and ongoing RLE symptoms.

4.1.2 Outpatient record redundancy

The mean, standard deviation, maximum and minimum of redundancy scores for all the patient documents in the corpus were 0.74, 0.14, 0.96 and 0 respectively. Several different patterns of redundancy scores were observed when examining individual patient records (Figure 4-2). Three outpatient records with at least 8 notes were examined with visit purpose and clinical events notes. These events and the mean document redundancy scores were plotted (Figure 4-2). The presence of cycles in redundancy scores at the individual patient record level was observed, which appeared to correlate with clinical events in most cases.

Figure 4-3 shows the means and standard errors of the redundancy scores pooled into quartiles (groups 1-4) over all clinical notes in all available patient records with 4 or more notes (because of the split into quartiles). While redundancy at the individual patient record level appears to be cyclical (Figure 4-2), overall redundancy scores across all patient records temporally have a clear upward trend (Figure 4-3). The redundancy scores in the fourth quartile (most recent) were significantly higher than all earlier quartiles. The scores in the 3rd quartile were also higher than in the 2nd quartile but this was not statistically different. There also exists a good linear relationship ($R^2 = 0.89$) between the number of groups and corresponding redundancy scores.

A linear regression line (with function and R^2) between redundancy scores and group number was drawn to visualize a trend. Means with different letters were significantly different ($p < 0.05$).

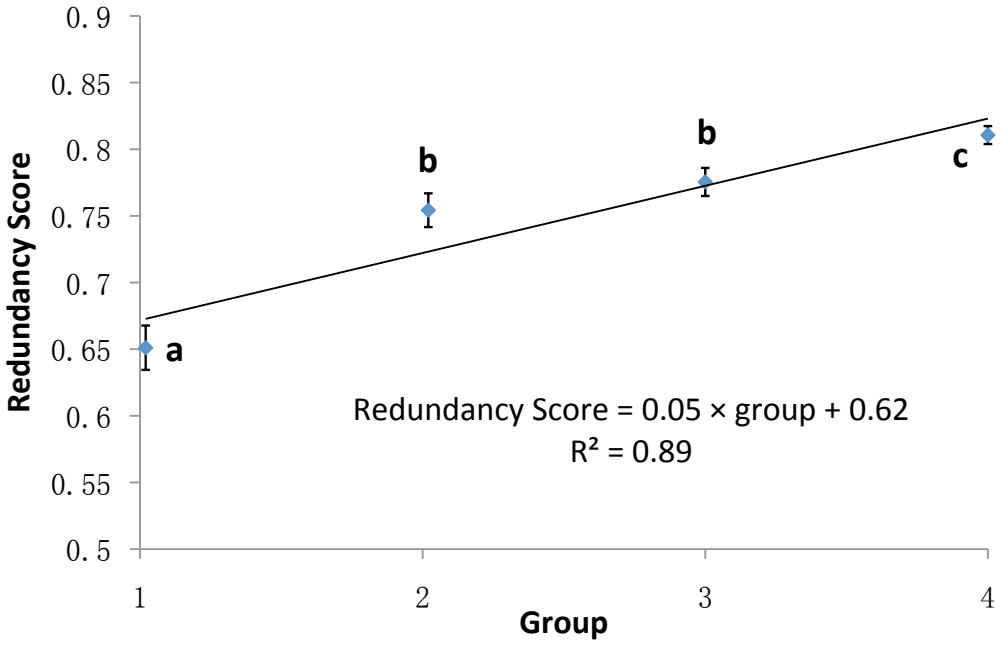


Figure 4-3. Redundancy scores (mean±standard error) of document quartiles.

4.2 Applying Statistical Language Models to Identify and Visualize Relevant New Information

4.2.1 *N*-gram Models Performance Evaluation

Cohen's Kappa coefficient of two annotators for the overlap clinical documents was 0.65 and percent agreement was 0.94. To determine the threshold TFIDF value, the first three cutoff values of 2E-6, 4E-6, 6E-6 were tested and 2E-6 was then chosen due to similar performance. Table 4-2 shows performance for different methods. The use of bi-grams with a count threshold of zero, addition of lexical normalization, removal of both stop word types, as well as heuristic rules including section-specific rules, resulted in best performance.

Table 4-2. Comparison of methods with reference standard. ACC = Accuracy; SEN = Sensitivity; SPE = Specificity; PPV = Positive Prediction Value; NPV = Negative Prediction Value. Relevant new information defined when count ≤ to count threshold value.

Methods (*n*-gram)	**ACC**	**SEN**	**SPE**	**PPV**	**NPV**
Count Threshold Value = 0					
Baseline (bigram)	0.471	0.678	0.442	0.144	0.909
CSW (bigram)	0.507	0.957	0.444	0.193	0.987
BSW (bigram)	0.649	0.942	0.608	0.250	0.987
LVG_BSW (bigram)	0.654	0.961	0.611	0.255	0.991
HR (bigram)	0.829	0.889	0.820	0.456	0.982
SUB (bigram)	0.894	0.757	0.914	0.552	0.964
SUB (trigram)	0.800	0.738	0.808	0.341	0.958
SUB (four-gram)	0.805	0.738	0.814	0.348	0.958
Count Threshold Value = 1					
SUB (bigram)	0.854	0.761	0.866	0.441	0.963
Count Threshold Value = 2					
SUB (bigram)	0.833	0.786	0.848	0.417	0.966

4.2.2 *N*-doc Model Evaluation

Table 4-3 shows the accuracy, sensitivity, specificity, PPV and NPV as *N* increases from 1 to 9. Overall accuracy and PPV increased with increasing document numbers, resulting in decreasing sensitivity with increasing documents in the model.

Table 4-3. Statistical results increasing the number of previous documents in the *N*-doc model. ACC = Accuracy; SEN = Sensitivity; SPE = Specificity; PPV = Positive Prediction Value; NPV = Negative Prediction Value.

# Pre Docs	ACC	SEN	SPE	PPV	NPV
1	0.834	0.756	0.845	0.398	0.962
2	0.852	0.733	0.868	0.434	0.959
3	0.861	0.720	0.880	0.451	0.958
4	0.860	0.715	0.880	0.455	0.957
5	0.871	0.689	0.896	0.477	0.955
6	0.866	0.703	0.890	0.478	0.954
7	0.862	0.695	0.887	0.480	0.951
8	0.885	0.653	0.918	0.527	0.950
9	0.883	0.636	0.920	0.543	0.944

4.2.3 Relevant New Information Visualization

Figure 4-4 shows example screen shots of clinical notes highlighted by these methods in comparison to the expert reference standard. Relevant new information at a word level is highlighted as green in comparison to reference standard relevant new information in purple. In Sec1, formatting and signature were not marked for both (True Negative (TN)), and the first paragraph was marked in both as relevant new information (True Positive (TP)). The automated method wrongly marked the second paragraph, which is a False Positive (FP). In Sec2, relevant new information about MUSCULOSKELETAL was marked in both (TP). But another piece of relevant new information "Negative for temperature intolerance, skin/hair changes" was marked by the automated method (FP). In Sec3, the diagnosis was correctly marked, however, the plan was not marked as relevant new information by automated method (FN).

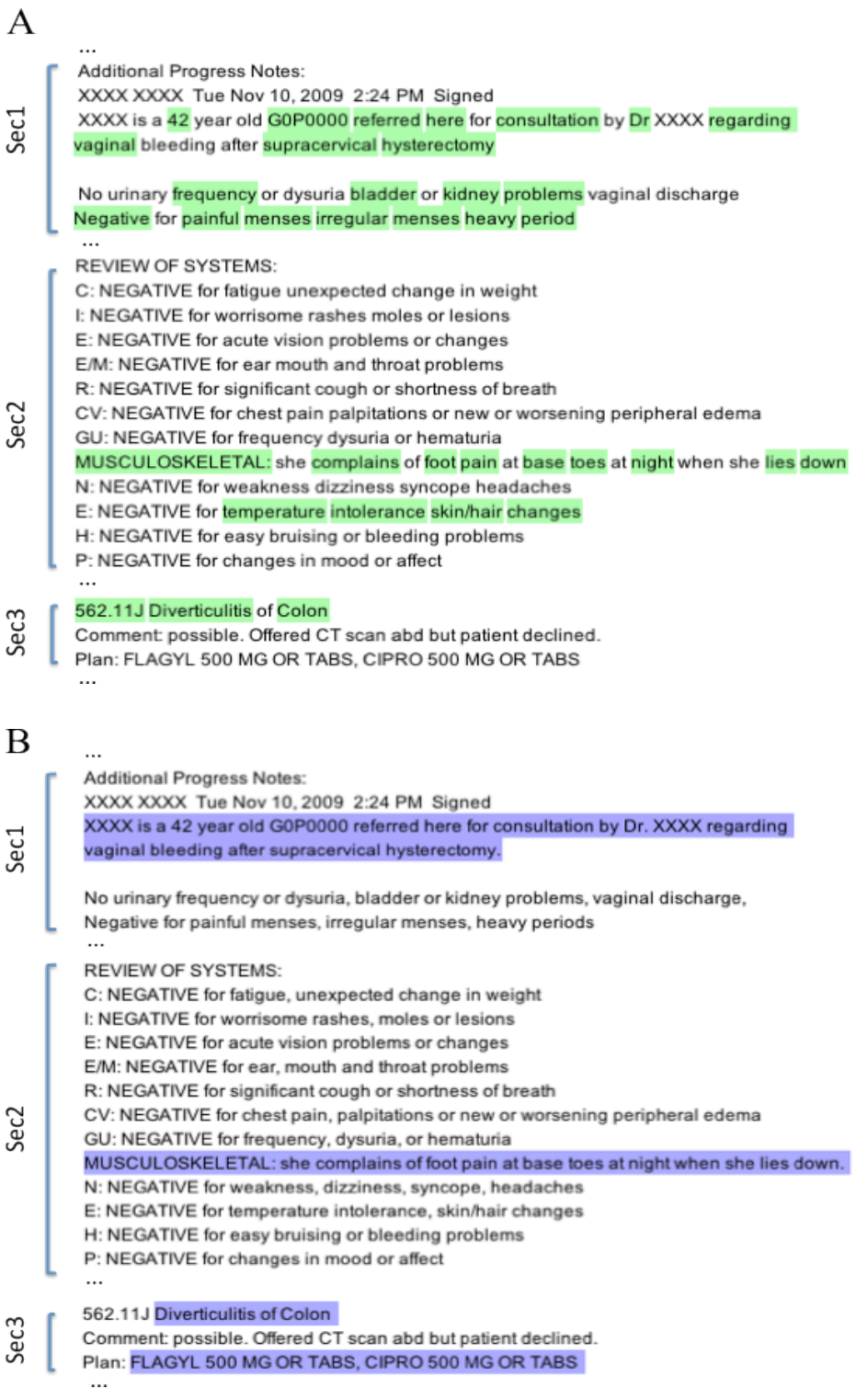


Figure 4-4. Visualization of relevant new information with A) automated method and B) reference standard.

4.3 Quantifying Relevant New Information to Navigate Clinical Notes

4.3.1 Annotation Evaluation and Model Performance

The two raters showed good agreement on the task of identifying new information on the overlapped annotation. Cohen's Kappa coefficient of two annotators for the overlap clinical documents was 0.80 and percent agreement was 97% on new information identification at the sentence/statement level. The results generated by automated results were compared with the refined reference standard. The accuracy, precision, recall, and F-measure are 0.83, 0.72, 0.71, 0.72, respectively.

4.3.2 Changes in the Amount of New Information

After averaging all 2,918 NIP scores for the array (Figure 3-4B), 20 arithmetic mean NIP scores were obtained. The means were then plotted with the number of previous notes and fitted with a logarithm function (Figure 4-5).

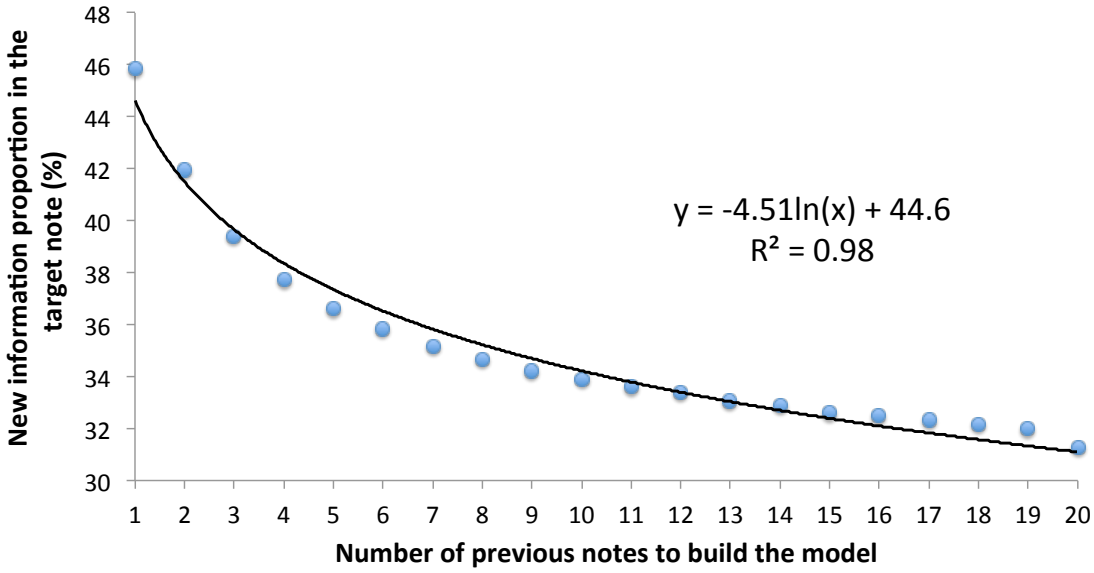


Figure 4-5. Scatter plot and fitted line of new information proportion with the numbers of the previous notes.

4.3.3 New Information Patterns

Figure 4-6A shows patterns of NIP scores in longitudinal clinical notes with the consideration of all patient notes. The four cyclical patterns indicate similar shapes, although the four curves have different NIP values. When the number of previous notes changed from 1 to 5, NIP dropped significantly; whereas the change was more gradual from note 5 to note 10 and even less from note 10 to note 20.

Figure 4-6B & 4-6C show longitudinal clinical notes of two patients with new information manually identified by the resident physician (in boxes). Solid lines show the NIP based on the previous 10 notes, and dotted lines show new information based on the previous 20 notes. Longitudinal clinical notes from both patients appeared to have a cyclical pattern, characterized by alternating periods of peaks (larger NIP) and troughs

(smaller NIP). Based on the NIP scores of two patient notes, larger NIP scores (usually larger than 20%) correlated to notes with more new information content and a smaller NIP scores (usually less than 20%) corresponded to notes without a significant amount of new information. One exception is that the high NIP in note #27, patient 1 (Figure 4-6B) did not contain new patient clinical history but instead had newly information of note template, including "glucose self monitoring: SELF MONITORING:104315::'once daily'". Also, note #25, patient 2 (Figure 4-6C) had a relatively lower NIP score but was judged to have new information of clinical significance (eye twitching).

Comparing solid lines (10 notes) to dotted lines (20 notes), it was found that most NIP scores did not decrease much, with some keeping the same score and a few notes having significantly lower scores. For example, note #21, patient 1 (Figure 4-6B) had a lower score when using longer patient history, which correlated to an old sinus infection found in the note #8. Also in patient 1, note #29 (Figure 4-6B) contained new information based on the previous 10 notes (*i.e.*, note #19-28), including symptoms, surgical history, and social history, which were also found in the note #18, thus the NIP dropping compared with all the pervious 20 notes (*i.e.*, note #9-18). In patient 2 record, Figure 4-6C, it was also found a similar change for note #27, where new information of symptoms was found in note #13.

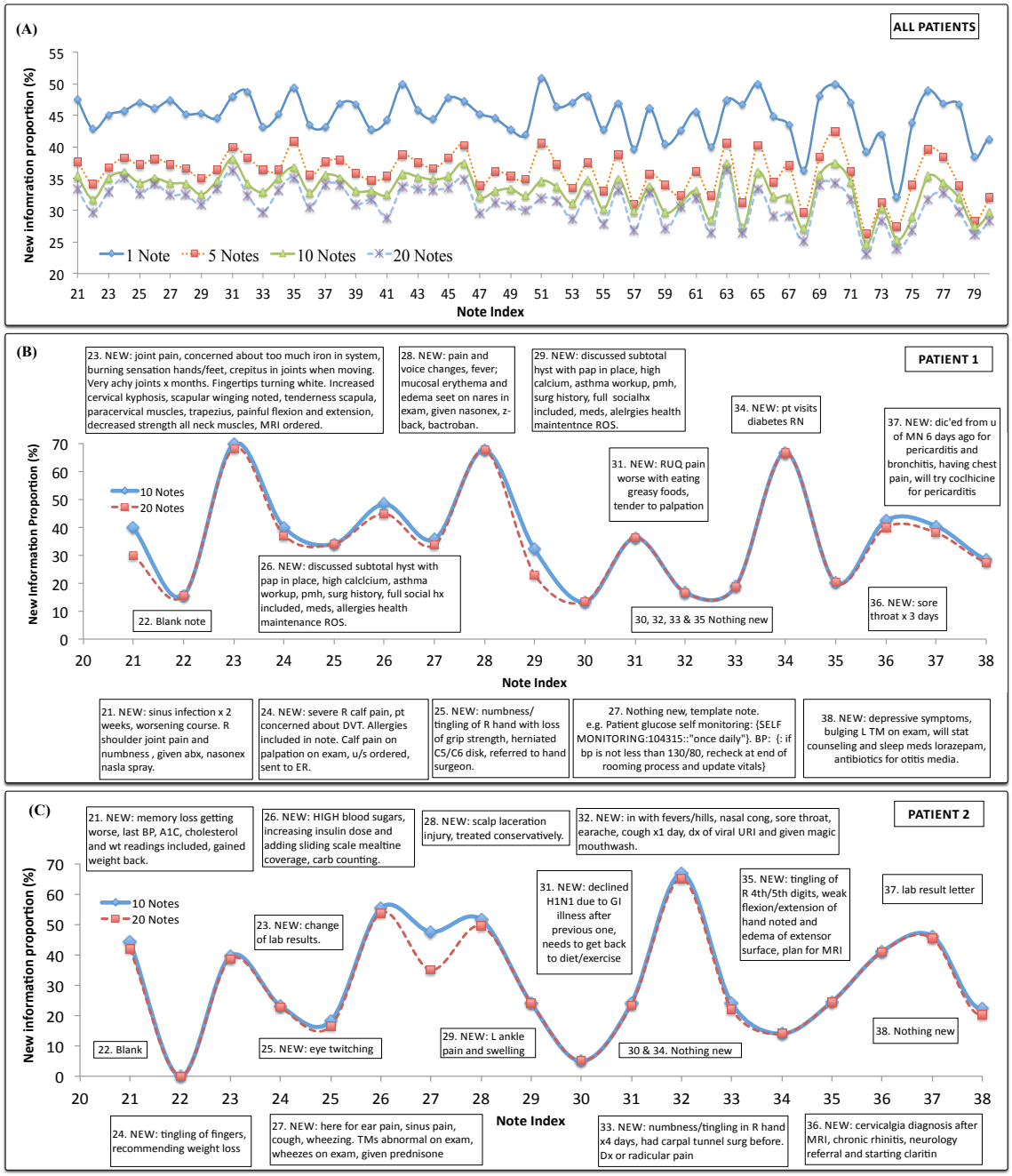


Figure 4-6. Patterns of new information proportions in clinical notes. (A) Overall pattern of new information proportions based on the averaged scores over patients; (B) & (C) New information proportions of longitudinal notes based on the previous 10 and 20 notes from two individual patients. New information contents shown in the boxes were annotated by the expert and compared with the previous 10 notes.

4.4 Classifying Semantic Types of Relevant New Information

4.4.1 Identification of Various Types of Relevant New Information

After calculating new information using the reference standards at the sentence level, the percentage of various categories (*e.g.*, lab, problem, medication etc.) of relevant new information annotated by medical experts were obtained (Figure 4-7). The top three categories were problem (34.1%), medication (31.7%) and laboratory results (17.3%). Other types include procedures of imaging (5.0%), family history (2.8%) social history (2.7%), medical history (2.4%), surgery history (0.4%), and others (3.6%).

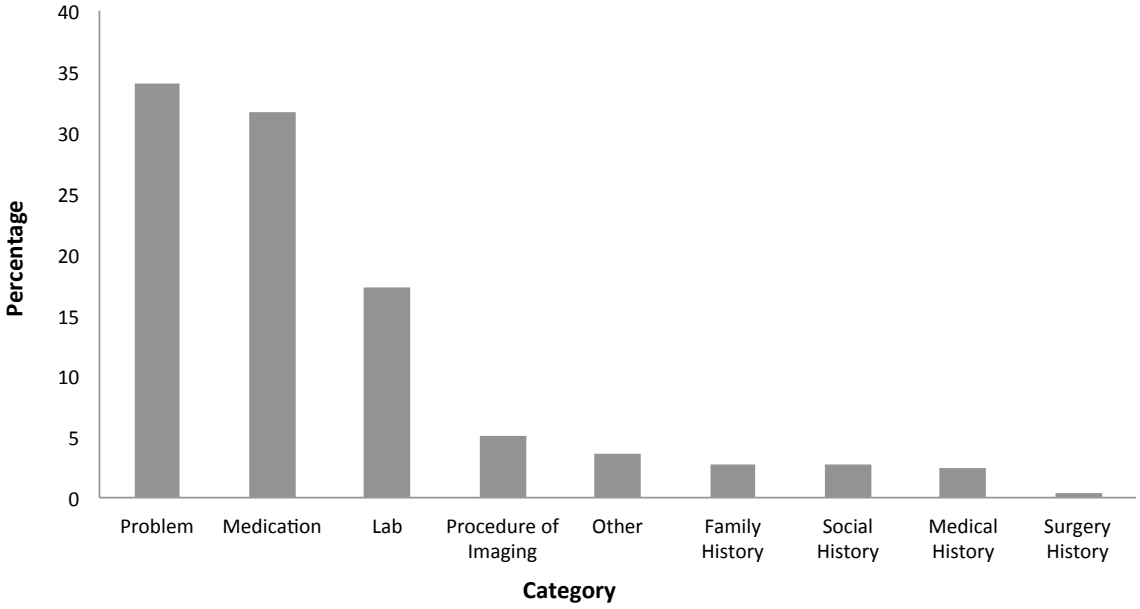


Figure 4-7. Percentages of various types of new information in reference standards.

NIP, NDIP, NMIP, and NLIP were then calculated for each note where NIP = NDIP + NMIP + NLIP + NOIP. Note that NOIP represents other types of new

information proportion (*e.g.*, Mental Process). Individual patients were then selected and the NIP, NDIP, NMIP, and NLIP were plotted as illustrated in one patient in Figure 4-8 and 4-9. Manually reviewed new information for each note is also provided in the boxes associated with each note (Figure 4-8). Overall, notes with higher NIP correlated with more new information, and notes with lower NIP scores tended to not contain significant new information. Key biomedical concepts were extracted for each information category and were marked (using the automated new text extracted) for each note in Figure 4-9.

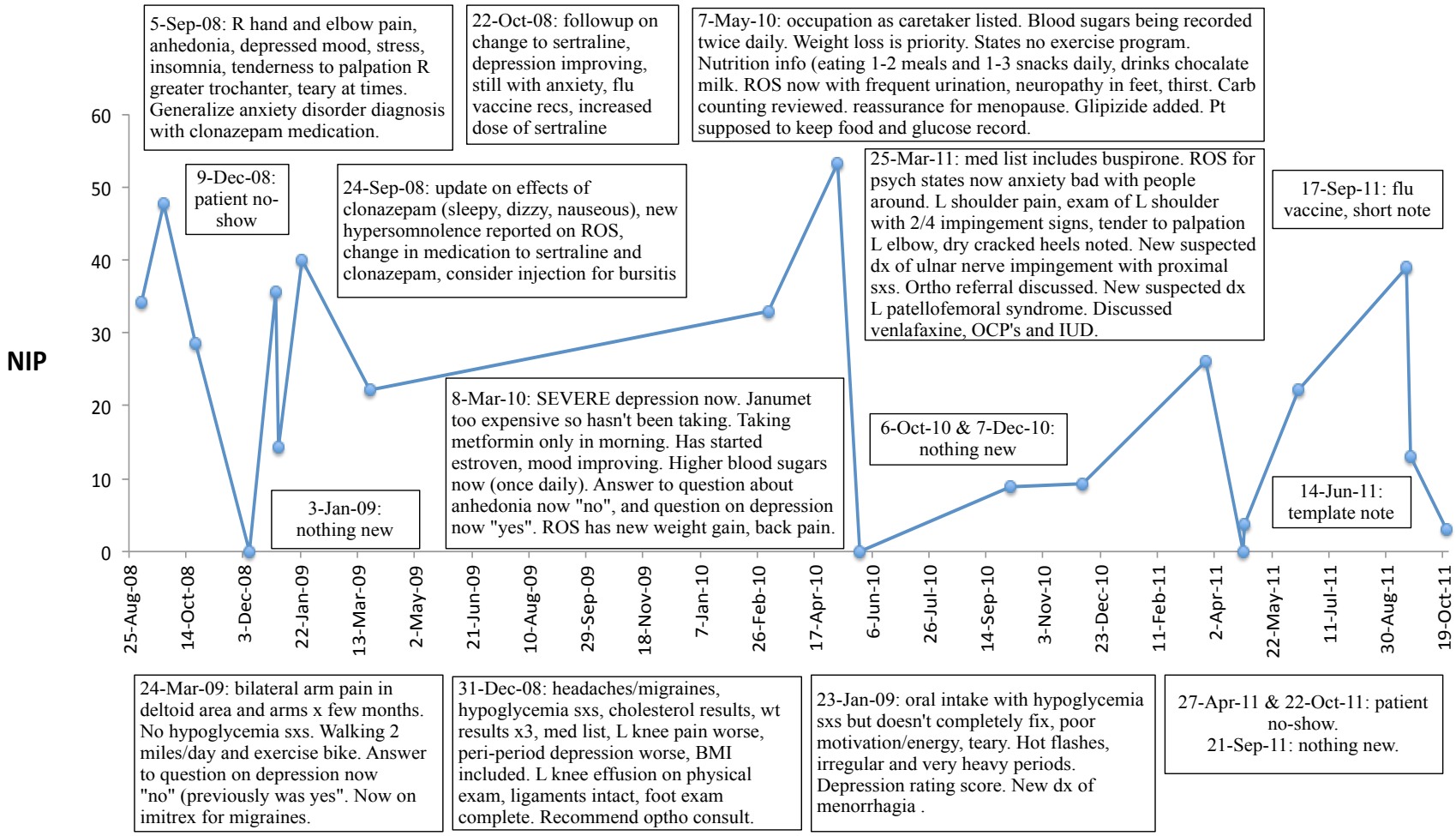


Figure 4-8. New information proportion (NIP) of clinical notes an illustrative patient. Boxes contain summarized new information.

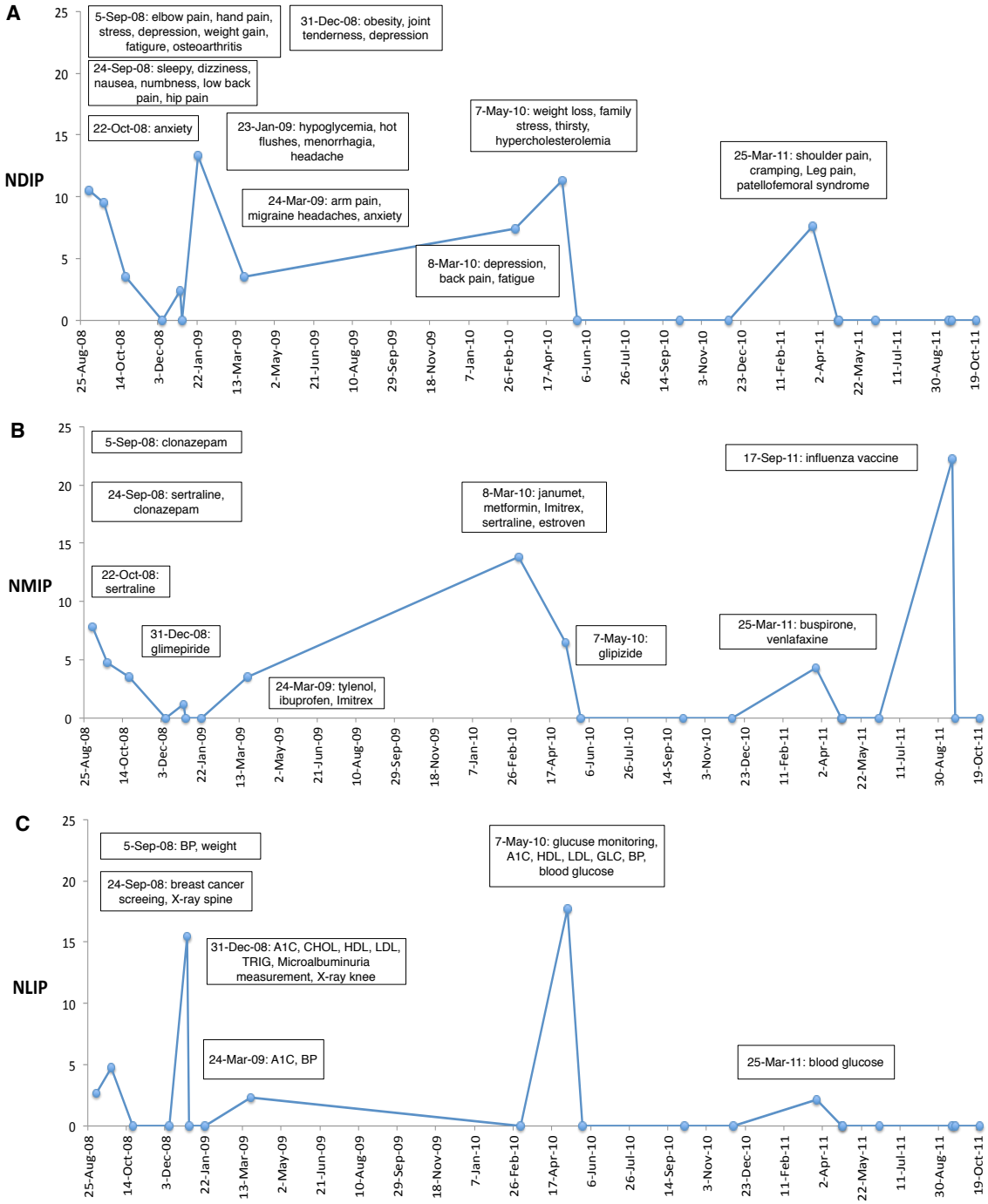


Figure 4-9. Plot of (A) NDIP (disease), (B) NMIP (medication), and (C) NLIP (laboratory) over time for the same patient as Figure 4-8. Biomedical concepts for each note included in boxes. NDIP, new problem/disease information proportion; NMIP, new medication information proportion; NLIP, new laboratory information proportion.

4.4.2 Correlation of New Disease Information with Charlson Comorbidity Index

Figure 4-10 illustrates the relationship between NDIP and CCI scores for another patient. Here, the diagnoses of diabetes without chronic complication, cerebrovascular disease, and renal disease were found on the note #4, #17, and #31, respectively. When correlating these CCI turn points with NDIP, it was found that two out of three NDIPs (#4, #31) increased with CCIs. After correlating NDIPs with CCIs in the whole patient corpus, a correlation score of 0.63 was obtained.

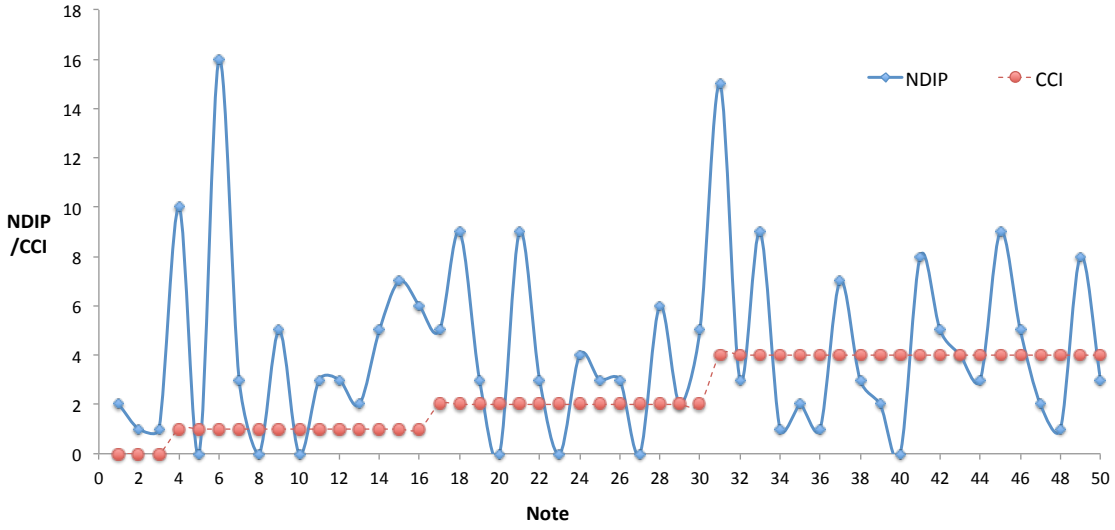


Figure 4-10. Relationship between NDIP and CCI scores for a selected patient. Diagnosis of diabetes without chronic complication, cerebrovascular disease, renal disease found on the note #4, #17, and #31, respectively. NDIP, new problem/disease information proportion and CCI, Charlson comorbidity index.

CHAPTER 5 DISCUSSION

In this chapter, I will discuss the effectiveness and difficulties of method development described in Chapter 3, the meaning and significance of those findings presented in Chapter 4, and how those methods can potentially assist clinicians to review clinical notes.

5.1 Modifying Global Alignment to Investigate Redundancy Patterns

This study focuses on the exploration of new methods to investigate information redundancy in clinical documentation. In this exploratory study, an expert-based reference standard was developed and compared to automated measures, including a previous measure based on global alignment, a baseline measure using alignment over short word sequences and enhancements to this measure using a combination of knowledge-based and statistical corpus-based knowledge-free approaches. With respect to the overall level of redundancy with time in clinical text, the results are consistent with previous reports (14) and confirm the finding that information redundancy in clinical notes (in this case outpatient documents) is significant. The results indicate that content words (as opposed to standard and statistically-based stopwords) are most important to be considered as features for redundancy identification. However, lexical normalization and semantic similarity may also be promising techniques for follow-up studies. Furthermore, the sliding window technique with aggregation performs significantly better than global alignment for assessing clinical text redundancy.

Different automated methods to quantify redundancy were correlated with the expert-derived reference standard. The use of a combination of standard stopword removal and TFIDF threshold stopword removal was optimal over either single type of stopword removal alone. In addition, both lexical normalization and semantic similarity enhanced these measures, although by a small amount. While lexical normalization results were slightly better in this study, it is possible that further enhancements to semantic similarity measures, including more effective mapping to named-entities with text chunking or shallow parsing, would help to identify multi-word concepts (*i.e.*, diabetes mellitus) and improve performance. Other potential enhancements include application of abbreviation and acronym disambiguation as these are common in clinical text. Also, only the path measure from the semantic similarity package with a single cut-off was utilized, and other semantic similarity and relatedness measures that the group has developed, including second-order vector-based measures were not integrated to the system (62).

The investigation of the effect of various window sizes indicated that the performance of the alignment approach was improved with the decrease in window size as seen in Table 4-1. For window size of eight and ten, the performance worsened. This could be due to the average length of selected sentence pairs (13 words) approaching this larger window size. Although a smaller window size of four resulted in slightly higher correlation with human judgments of redundancy, smaller window sizes may also result in generating spurious alignments between portions of medical terms rather than entire terms. This may be an issue when the measure is applied to a large set of clinical

documents rather than individual sentences. In addition, computational efficiency decreases with decreasing window sizes resulting in more text frames to be compared. The window size of five therefore represents a tradeoff between accuracy and efficiency, as well as meaningfulness of generating alignments that capture most of the content of a medical term. While somewhat inefficient to have the sliding window align over the entire text, it is anticipated to be a tractable method, as I would envision applying these metrics to text a single time and storing this information as part of a display feature in a graphical user interface for electronic text.

Several patterns of redundancy scores with different patient documents were observed. Figure 4-2 shows that most patients demonstrated cyclical patterns in the mean redundancy scores for documents of a given patient. To investigate if redundancy scores could detect redundant and new information, three patient records were reviewed and changes in redundancy scores generally correlated with clinical events, such as a motor vehicle accident, loss of insurance and a medication change, or a new visit to the emergency room for a congestive heart failure exacerbation. A document with redundant information had a high redundancy score on the peak of the graph and a document with a significant event had a lower redundancy score resulting in a trough on the curve (*i.e.*, Figure 4-2A(c), 4-2B(b) and 4-2B(f)). These findings indicate another potentially interesting and beneficial use of automated approaches for identifying redundant and new information. These approaches may be used to identify salient or unusual events in the patient's history and thus may aid the clinician in quickly constructing the "background" for the current visit.

There was also an observed trend towards an increase in overall redundancy of information with time in the clinical records, which was reported previously by Wrenn et al. (14). In practical terms, the methods I developed for assessing redundancy of information in clinical notes could potentially be used to automatically identify non-redundant information and present notes to the clinicians at the point of care in an electronic health record system in a more easily digestible manner. Furthermore, the method may be useful for quantifying redundancy during different periods in patient care history and testing for associations with adverse events and other patient outcomes such as hospital admissions, morbidity and mortality.

5.2 Applying Statistical Language Models to Identify and Visualize Relevant New Information

This study focuses on identification and visualization of relevant new information in medical texts. In this study, techniques to detect relevant new information in clinical notes were explored. An expert-based reference standard was developed and used it to evaluate several approaches, including baseline *n*-gram techniques, as well as several enhancements such as rule-based and statistical knowledge-free approaches. The study shows that the content words, knowledge-base rules (*e.g.*, formatting and noise removal, and section rules in the clinical notes) are important features that need to be included when distinguishing between relevant new versus redundant information.

It was observed that heuristic rules helped to improve system performance, including section-specific rules. Informal analysis of the removed content shows that

noise is introduced by structural attributes of clinical notes. For example, most clinical reports in the system include visit information located at the head of the document, such as visit date, time, encounter number, provider, previous visit et al as well as the signatures at the end of some subsections to indicate the document have been reviewed or recorded. Prior to removing these items, the *n*-gram model marked them as relevant new information since some parts of these items typically change from visit to visit. Other examples include the details of medications, such as medication class, route and special instructions. While this medication information may change slightly over time, it was judged by physician annotators as irrelevant. Section headings constitute another cue indicative of redundant or irrelevant information. For example, the content of the "Past Medications" and "Allergies" sections was always marked by annotators as redundant information; in contrast, the "Follow-up" section was always marked as relevant new information. A particularly interesting example consists of two documents showing exactly the same sentence, "patient will keep a food record and return in 2 weeks", marked by experts as relevant new information.

It was observed that the two medical experts sometimes showed slightly different views on annotating relevant new information. Both annotators failed to mark some of the new information. It was also found that one expert annotated more carefully than the other resulting in fewer missing values. For example, one annotator failed to mark "body mass index is …" as relevant new information in contrast to the automated approach. Another example is the statement "Height: 5'10"" stated multiple times in previous notes that was marked as relevant new in the reference but redundant by the methods,

accounted as FN. So it is necessary to refine the reference standard for further study by adjudicating both annotations by a third expert or by refining instructions to annotators in subsequent experiments.

The methods are currently not designed to identify relevant new information at the semantic level. Due to this, the approach is sensitive to such variation as the use of acronyms and word order changes. Another interesting observation is that trigram and four-gram approaches performed worse than the bigram model. This is not surprising because the models were trained on a relatively small corpus (*e.g.*, about 2,000 sentences and statements). Thus these data may have been too sparse for higher than bigram order modeling.

To investigate the effect of the number of previous documents used for modeling, the number of preceding documents was varied from 1 to 9. The PPV increased on average by 15% with the accuracy staying the same. Thus it was concluded that patient records with longer and presumably more complex histories would result in more effective modeling of relevant new information.

Overall, this study of automated visualization of relevant new information via highlighting showed that this was a simple and effective way to present information to physicians when they review complex medical documents. The results of the explorative study are significant in the context of the development of next generation EHRs that should take into account human factors such as information overload in text, which can affect patient safety and quality of care.

5.3 Quantifying Relevant New Information to Navigate Clinical Notes

This study focuses on developing methods to improve note navigation in longitudinal notes. Preliminary studies have shown that improved information navigation with notes may be a promising feature in future EHR document interface design (63). There is a need for both back-end algorithms design to facilitate this, as well as improved front-end user interfaces with these capabilities within EHRs.

To build such an information navigation system, annotation is a vital step but also a challenging task faced in the previous work (61) and in this study. Although inter-rater agreement was high, one rater was found to annotate more carefully than the other. Another issue was that the exact boundary of redundant versus new information was not well defined. To obtain a high-quality gold standard, which can help develop more accurate methods, good baseline communication between annotators was established before performing actual annotation for the method development. For example, one rater only annotated the piece of new information (lab values), but another rater still marked the corresponding information such as the title, date of the lab. Enhanced communications between annotators along with clearer guidelines improved the gold standard's quality.

Changing the number of previous clinical notes in longitudinal patient records changes the size of training data for the language model. For a given note, as the model includes a longer history before the target note, it includes more clinical history about the patient and the model can recognize relatively more redundant information if information

in the target note was included earlier in the patient's history. If new information continues to decrease when the model "sees" more historical notes, it indicates several possibilities about a patient's history including 1) that the history may contain information in the target note copied from earlier notes, 2) that clinicians may express events similarly and that there is a balance between what is old or new and that some events that repeat may actually be new (*i.e.*, a repeat flu infection one year later), or 3) that by adding large amounts of notes to the model, it at some point may contain too much noise to detect new events.

This method was used to estimate NIP and averaged all NIP scores of 2,918 notes from 100 patients. NIP scores decreased as the number of previous notes increased (Figure 4-5), indicating that physicians either copy information from as far as 20 previous notes or use similar forms of expression to describe similar events. Interestingly, the trend almost perfectly ($R^2 = 0.98$) fits a logarithmic function. The decrease of new information logarithmically appears related to the length of the clinical history. Based on the trend generated from all patient notes, approximately 55% of information in the current note was redundant compared to the immediate previous note. In other words, 55% of redundant information may have been copied and pasted from or present in the previous along with the current note for other reasons. Approximately an additional 11% of information in the current note was propagated from previous 2-10 notes; and another additional 4% from of information from the previous 11-20 notes. These numbers can be different for individual patient's records, but this overall trend indicates the boundaries of the source of redundant information.

Patients with chronic diseases come to the physician's office often, thus generating as many as 90 longitudinal notes (Figure 4-6A) over the study period. The cyclical pattern indicated that the collection of longitudinal notes contains both important notes with new information and less important notes with mostly redundant information. This finding was consistent with the previous finding using global alignment methods (64). In that study, a set of 10 notes from three patients were randomly chosen and up to 10 previous notes were used to quantify redundancy scores. It showed similar cyclical patterns and an overall uncharacterized trend of increasing redundancy.

Here, new information (and its counterpart - redundancy) patterns of notes for individual patients were further investigated and the scores were compared with human judgment. There was high correlation between calculated NIP scores and clinically significant events. Higher NIP scores correlated to more new information in the notes and lower NIP scores indicated less new information. It is observed that 20% may be a suitable threshold value only based on the NIP scores of the two patient notes (Figure 4-6B&C). However, further work is necessary to calibrate the measure threshold to distinguish notes with more new information from those with less. It is also observed that unexpectedly high NIP scores can be seen with the introduction of templates absent in historical notes. One prerequisite condition for an accurate statistical language model is that the training set should be representative of the test set. This finding also provides us with a challenge that potentially could be addressed by incorporating EHR document templates into the model.

This analysis also helped to identify the original notes that served as the sources of redundant information found in target notes. A significant drop indicated that the target note contained a lot of replicated information. Some notes were found to have negligible score changes between 11 and 20 notes in the model, due to the existence of few additional pieces of additional significant information in the previous 11-20 notes.

Future research includes development of more robust automated methods for identifying new information as well as the development of a user interface and navigation tools to assist clinicians with identification of new information and efficient utilization of clinical notes. Summarizing or providing key words of new information may be another approach for providing clinicians better tools for improved note navigate in EHRs at the point of care.

5.4 Classifying Semantic Types of Relevant New Information

Automated methods to identify relevant new information represent a potential set of techniques to improve the process of reviewing clinical notes. Some formative studies have demonstrated that visualization of new information within clinical notes based on these techniques may save time in reviewing notes and helped to decrease the likelihood of missing important historical information (63). Subsequent work has demonstrated that NIP measures may be useful in identifying notes with clinically relevant new information (65). While notes with higher NIP scores usually correlate with new findings, clinicians may still confuse the details of what is the cause of a cyclical pattern. More pertinent questions include answering issues such as *why are these notes (with high new*

information scores) important? and *what specific new information does this note contain?*. This study examines types of new information in several important categories by dividing original NIP scores into various types of new information.

In comparing annotations by residents with UMLS concepts using automated methods, there were several key consistent findings. Some types of problem/disease information where automated methods identified information not included in the physician-generated reference standard were found. In the example symptoms of elbow pain, hand pain, and depression were identified in the reference standard but other symptoms such as anhedonia or insomnia were not identified. In contrast, with medications, automated methods incorrectly identified some medications. For example, (Figure 4-9b) new medications of clonazepam (5-Sep-08), sertraline (24-Sep-08), metformin (8-Mar-10), estroven (8-Mar-10), glipizide (7-May-10), and buspirone (25-Mar-11) were found via automated methods and by the expert annotators, but the method incorrectly found "janumet" from the sentence "… janumet was too expensive, so she did not take it." (8-Mar-10). Although NegEx functionality in MetaMap was used to account for negation, the automated method did not effectively deal with the co-reference issue (it refers to janumet). Another example is "vernlafaxine" from the note (25-Mar-10) "… another future option may be to try venlafaxine". Here, the physician only recommended the medicine instead of prescribing it accounting for another false negative example. Finally, with respect to laboratory information, there were few examples where the physician annotator marked laboratory data. One reason for this is that glucose, hemoglobin A1C tests are routine monitoring tests, and clinicians will not focus on that

unless there are significant changes of the results. Mapping issues with respect to acronyms for laboratory procedures were also faced. For example, “A1C” had to be translated to its full name “Hemoglobin A1C” to be recognized by MetaMap. In the future work, more detailed information (*e.g.*, if the value excess the normal range) may be provided other than just listing laboratory name to aid clinicians to pay more attention to the specific lab results with unexpected values.

A relationship between the change of CCI and NDIP was also found. When CCIs increased, some NDIPs also increased. This finding provides some initial evidence that these methods can find certain important new information from notes, such as comorbidity information. In this study, one of the main reasons why correlation scores were not high was that the NDIP measure contains many other diseases/problems in addition to the Charlson comorbidity groups such as hypertension. Further investigations, including extraction of new information only for diseases belonging to one of Charlson comorbidities and correlation this information with CCI scores, may provide significant improvements in these methods.

CHAPTER 6 LIMITATIONS

In this chapter, I will explain the limitations of the research in the aspects of data, annotation, and methods, and also propose several future research directions to avoid these limitations.

6.1 Limitations of Data

Among all four studies, only patients with chronic diseases were selected to allow for a larger number of longitudinal clinical notes for each patient. To simplify the study, I limited the note type to only the office visit note in outpatient clinical settings, most of which were written by physicians. This limits the adaptation of the methods in different clinical settings and types of clinical notes. I will not be confined to more uniform outpatient documents and examine various types of clinical notes written by different service teams, such as physicians or nurses, to understand the effect of document types and clinical sublanguages. Future research will also include larger data sets to develop more robust methods, as well as to evaluate the methods on clinical narratives in inpatient settings. In addition, user studies are under way to visualize the relevant new information in clinical notes and then incorporate with the current EHR systems.

6.2 Limitations of Annotations and Evaluations

All evaluation results were based upon the reference standard that physicians annotated. Manually annotation is a human intensive and expensive analysis process, so only small

sets of reference standard were obtained for training and evaluating the methods. This is one of the barriers most NLP researchers have to encounter. The difference of clinical experience and subjective judgment on the relevant new information between annotators also affects the quality of the reference standard. In addition, in the study to classify the semantic types of new information, although the results were compared with the reference standard, the annotation was not built at the same (biomedical term) level as new information was automated extracted. Future studies will include enlarging the number of annotations and improving the quality of reference standard. Also, the specific annotation guideline needs to be established to alleviate the annotator bias. A larger corpus of high quality reference standard will support developing more effective methods and making evaluations more accurately.

6.3 Limitations of Methods

As mentioned in the hypothesis of these studies, the intra-document redundancy, or the duplicated contents appearing in the same note, was not considered in the dissertation. This would likely be, however, a significantly more computationally difficult problem since the methods would require comparison not only with previous documents but also with all previous text within the current document.

All methods in this dissertation focused only on the lexical level. Semantic level issues were out of the scope of this dissertation, such as co-reference (e.g., “it”, “this”)

and experiencer detection (e.g., "patient", "sister"). For example, "Pt[a] has diabetes" and "His mother has diabetes" shared most of the words, but they are semantically different as the experiencers are changed. Acronym and symbol disambiguation were also not included in the studies. Future research will add more semantic components to make the system more accurate and comprehensive.

Moreover, relevant new information was only limited to the addition of information in the newer notes in all of the studies. The deletion of relevant new information in the more recent clinical notes was not considered in this dissertation. For example, the information that a drug was removed from the patient's medication list is also very important for clinicians' diagnosis and synthesis of patients' change of medical conditions. Due to the asymmetric nature of new information identification process, this type of new information can be obtained by comparing the object notes and target note in reverse, but it still deserves additional investigation to represent this kind of relevant new information to clinicians.

With respect to limitations for the first study to investigate redundancy patterns, the study did not utilize a separate development set of documents, and as such, represents pilot data. I plan to confirm and validate the findings on other document sets and to correlate findings of redundancy to cognitive issues that clinicians experience when consuming clinical texts. I also plan to validate these results at a document level to see if the findings at a statement level generalize with an expert-derived reference standard.

---

[a] Pt is an abbreviation of patient.

The methods to classify semantic types of relevant new information have certain other limitations. Using a mapping technique such as that provided with MetaMap does not give additional types of information such as the change of dosage for a specific drug. Currently, I only looked at three types of new information; other types of information such as Mental Process will be valuable additional semantic types to explore. In future research, I will use existing tools such as those used for plagiarism detection to refine methods for identification of new information. The use of specialized modules such as MedEx (66) can be also considered to extract more details of the change of medication use, other than just providing drug name.

Further work to implement and perform usability test with clinicians of new information visualization within clinical notes and user interfaces is also needed. I believe that having advanced navigation function with visualization functionality (61, 63) could potentially aid clinicians in finding and synthesizing new information both at a note and patient level of granularity.

CHAPTER 7 CONCLUSION

Overall, this dissertation investigated the ubiquitous redundancy problem in clinical documentation systems and proposed potential solutions from different aspects. The first two studies focused on the identification of redundant (and its counterpart new) information in clinical notes by using different methods. The next two studies focused on how to use the new information identified by the methods to aid clinicians' navigation of notes in a series of longitudinal patient clinical records. The third study proposed the new information proportion (NIP) as a useful method to navigate important notes as well as help to understand the clinicians' "copy and pasting" behavior. The fourth study further classified the semantic types and extracted key biomedical terms from the new information to provide clinicians more details for new information. The combination of information identification and visualization on the term level for each note ("micro"), as well as information navigation at the note level ("macro"), will be a potential way to help clinicians navigate directly to notes with specific types of new information they care more about and then review the note with highlighted new information more purposefully.

Redundancy found in clinical narratives due to the inappropriate copying and pasting of parts of the medical record is highly prevalent in EHR documentation systems. In this dissertation, I have described the impact of the redundant information on the efficiency of healthcare and provided potential solutions to solve this issue.

To understand the redundancy issue and their patterns in clinical notes, different automated methods were developed to measure the redundancy and evaluated by comparing with the reference standards built by clinicians. As a result of the studies described in this dissertation, I found that the most optimal method for this task is the modified global alignment method, which operates by sliding a window across clinical text in search of redundant information. Another important finding of this work is that this alignment algorithm can be fine-tuned with techniques such as classic and TF-IDF stopwords removal, and lexical normalization. These techniques can generally increase the accuracy of the methods used to measure redundancy. Using these methods for measuring the amount of redundant information present in clinical notes, it is found that the redundancy increased over time for the whole corpus of outpatient clinical notes. Furthermore, the amount of redundancy appears to have a cyclical pattern for longitudinal clinical notes within an individual patient record. This study investigated the redundancy patterns at the note level both in the whole corpus and for the individual patient.

To identify the new information for each note, I further developed a statistic language model – *n*-gram model to identify new (the counterpart of redundant) and clinically relevant information for each clinical note. The *n*-gram language models were further enhanced by the similar refinement techniques to those reported in Study 1, including classic and TF-IDF stopwords removal, and lexical normalization. The best method, bigram with modification, was used to identify new information on the word level. The heuristic rules classifying relevant or irrelevant information were further used

to distill new information, which is either the note format or non-clinically significant to the physicians. Clinical notes were visualized by highlighting relevant new information identified by the bigram language model. The visualization of relevant new information provides a potential solution for clinicians when they review clinical notes with a large amount of redundant information.

To investigate the use of statistical language models for information navigation with patient longitudinal clinical notes, new information in longitudinal clinical notes for a given patient was quantified as NIP by using the automated methods. The higher NIP scores of clinical notes for a given patient indicate the notes with more new information (or less redundant information) often with clinically significant events. The cyclical pattern of NIP scores for a given patient was also consistent with the redundancy pattern found in the first study. The findings that notes with higher NIPs were highly in accordance with notes manually reviewed by a physician suggested that NIP is a good parameter to navigate notes. New information in longitudinal notes had an overall logarithmic relationship with the length of historical notes used to create the language model. Physicians tended to copy information from the most current notes. The analysis can also help to find the source of redundant information in a given note. Language models may be used as a potential information navigation tool for clinical notes.

The information navigation in the third study is incomplete, since the clinicians only know which notes contains more relevant new information based on the calculated NIP scores; they are unclear what type of information is new before they review the notes. To better synthesize and navigate the notes for clinicians, the methods to classify

the relevant new information based on the semantic types of the contents were further developed. The relevant new information identified by the methods was mapped to UMLS Metathesaurus and the corresponding semantic types and key biomedical terms have been extracted for new information. These types of information, such as problem, medication, and laboratory are important information related to the change of patient's health conditions and the information that clinicians usually pay more attention to analyze the patients. Instead of putting all types of information together in the third study, I split relevant new information into different types for longitudinal clinical notes for a given patient. The key biomedical terms are also provided to clinicians to see which diseases, medications, or laboratory tests/results changed were added. In this study, automated methods provide more navigation functions on information types and key terms for each note before clinicians' review notes, thus providing a potential navigation tool for clinicians to identify information to changes in health status more effectively and quickly.

The studies in this dissertation should provide innovation to the next generation EHR system implementation to enhance the efficiency of reviewing and using clinical documentation, and improve the satisfaction of clinicians with EHR systems.

BIBLIOGRAPHY

1. Harrington L, Kennerly D, Johnson C. Safety issues related to the electronic medical record (EMR): synthesis of the literature from the last decade, 2000-2009. *Journal of Healthcare Management*. 2011;56(1):31-43.

2. Koppel R, Kreda DA. Healthcare IT usability and suitability for clinical needs: challenges of design, workflow, and contractual relations. *Studies in Health Technology and Informatics*. 2010;157:7-14.

3. Stead WW, Lin HS, editors. Computational Technology for Effective Health Care: Immediate Steps and Strategic Directions. *The National Academies Collection: Reports funded by National Institutes of Health*. Washington (DC). 2009.

4. Weir CR, Hurdle JF, Felgar MA, Hoffman JM, Roth B, Nebeker JR. Direct text entry in electronic progress notes. An evaluation of input errors. *Methods of Information in Medicine*. 2003;42(1):61-7.

5. Hammond KW, Helbig ST, Benson CC, Brathwaite-Sketoe BM. Are electronic medical records trustworthy? Observations on copying, pasting and duplication. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2003:269-73.

6. Hripcsak G, Vawdrey DK, Fred MR, Bostwick SB. Use of electronic clinical documentation: time spent and team interactions. *Journal of the American Medical Informatics Association*. 2011;18(2):112-7.

7. Patel VL, Kaufman DR, Arocha JF. Emerging paradigms of cognition in medical decision-making. *Journal of Biomedical Informatics*. 2002;35(1):52-75.

8. Reichert D, Kaufman D, Bloxham B, Chase H, Elhadad N. Cognitive analysis of the summarization of longitudinal patient records. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2010:667-71.

9. Markel A. Copy and paste of electronic health records: a modern medical illness. *The American Journal of Medicine*. 2010;123(5):e9.

10. Siegler EL, Adelman R. Copy and paste: a remediable hazard of electronic health records. *The American Journal of Medicine*. 2009;122(6):495-6.

11. Yackel TR, Embi PJ. Copy-and-paste-and-paste. *Journal of the American Medical Association.* 2006;296(19):2315.

12. Thielke S, Hammond K, Helbig S. Copying and pasting of examinations within the electronic medical record. *International Journal of Medical Informatics*. 2007;76 Suppl 1:S122-8.

13. Embi PJ, Weir C, Efthimiadis EN, Thielke SM, Hedeen AN, Hammond KW. Computerized provider documentation: findings and implications of a multisite study

of clinicians and administrators. *Journal of the American Medical Informatics Association*. 2013;20(4):718-26.

14. Wrenn JO, Stein DM, Bakken S, Stetson PD. Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association*. 2010;17(1):49-53.

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;14:10.

16. Van Vleck TT, Stein DM, Stetson PD, Johnson SB. Assessing data relevance for automated generation of a clinical summary. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2007:761-5.

17. Wilcox AB, Jones SS, Dorr DA, Cannon W, Burns L, Radican K, et al. Use and impact of a computer-generated patient summary worksheet for primary care. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2005:824-8.

18. Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2005:106-10.

19. Weed LL. Medical Records That Guide and Teach. *M.D. Computing: Compters in Medical Practice*. 1993;10(2):100-14.

20. Liu H, Friedman C. CliniViewer: a tool for viewing electronic medical records based on natural language processing and XML. *Studies in Health Technology and Informatics*. 2004;107(Pt 1):639-43.

21. Johnson SB, Bakken S, Dine D, Hyun S, Mendonca E, Morrison F, et al. An electronic health record based on structured narrative. *Journal of the Amercian Medical Informatics Association*. 2008;15(1):54-64.

22. Bui AA, Aberle DR, Kangarloo H. TimeLine: visualizing integrated patient records. *IEEE Transactions on Information Technology in Biomedicine*. 2007;11(4):462-73.

23. Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. *Artificial Intelligence in Medicine*. 2010;49(1):11-31.

24. Wang TD, Plaisant C, Shneiderman B, Spring N, Roseman D, Marchand G, et al. Temporal summaries: supporting temporal categorical searching, aggregation and comparison. *IEEE Transactions on Visualization and Computer Graphics*. 2009;15(6):1049-56.

25. Zeng Q, Cimino JJ, Zou KH. Providing concept-oriented views for clinical data using a knowledge-based system: an evaluation. *Journal of the Amercian Medical Informatics Association*. 2002;9(3):294-305.

26. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*. 1970;48(3):443-53.

27. Smith TF, Waterman MS. Identification of common molecular subsequences. *Journal of Molecular Biology*. 1981;147(1):195-7.

28. Mount D. Bioinformatics sequence and genome analysis. Cold Spring Harbor Laboratory. 2001.

29. Jurafsky D, Martin JH. Speech and Language Processing. Upper Saddle River, NJ: Prentice Hall. 2009.

30. Manning CD, Schütze H. Foundations of Statistical Natural Language Processing. Cambridge, Massachusetts: The MIT Press. 2003.

31. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*. 2004;32(Database issue):D267-70.

32. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. *Journal of the Amercian Medical Informatics Association*. 2010;17(3):229-36.

33. Lin J, Wilbur WJ. PubMed related articles: a probabilistic topic-based model for content similarity. *BMC Bioinformatics*. 2007;8:423.

34. Errami M, Wren JD, Hicks JM, Garner HR. eTBLAST: a web server to identify expert reviewers, appropriate journals and similar publications. *Nucleic Acids Research*. 2007;35(Web Server issue):W12-5.

35. Cao H, Melton GB, Markatou M, Hripcsak G. Use abstracted patient-specific features to assist an information-theoretic measurement to assess similarity between medical cases. *Journal of Biomedical Informatics*. 2008;41(6):882-8.

36. Melton GB, Parsons S, Morrison FP, Rothschild AS, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. *Journal of Biomedical Informatics*. 2006;39(6):697-705.

37. Pakhomov S, McInnes B, Adam T, Liu Y, Pedersen T, Melton GB. Semantic Similarity and Relatedness between Clinical Terms: An Experimental Study. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2010:572-6.

38. McInnes B, Pedersen T, Pakhomov S. UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity. *Proceedings of the American Medical Informatics Association Annual Symposium.* 2009:431-5.

39. Rada R, Mili H, Bicknell E, Blettner M. Development and Application of a Metric on Semantic Nets. *IEEE Transactions on Systems, Man and Cybernetics*. 1989;19(1):17-30.

40. Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric for the UMLS. *Journal of Biomedical Informatics*. 2004;37(2):77-85.

41. Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*. 2003;19(10):1275-83.

42. Pedersen T, Pakhomov SV, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*. 2007;40(3):288-99.

43. Wu Z, Palmer M. Verbs semantics and lexical selection. *Proceedings of the 32nd Meeting of Association of Computational Linguistics*. 1994:133-9.

44. Leacock C, Chodorow M. Combining local xontext and WordNet similarity for word sense identification. WordNet: An electronic lexical database. 1998;49(2):19.

45. Nguyen HA, Al-Mubaid H. New ontology-based semantic similarity measure for the biomedical domain. *Proceedings of the IEEE International Conference on Granular Computing*. 2006:623-8.

46. Resnik P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *International Joint Conference for Artificial Intelligence.* 1995:448-53.

47. Jiang J, Conrath D. Semantic similarity based on corpus statistics and lexical taxonomy. *Proceedings on International Conference on Research on Computational Linguistics*. 1997:19-33.

48. Lin D. An information-theoretic definition of similarity. *Proceedings of the International Conference on Machine Learning*. 1998:296-304.

49. McInnes B, Pedersen T, Liu Y, Melton GB, Pakhomov S. Knowledge-based Method for Determining the Meaning of Ambiguous Biomedical Terms Using Information Content Measures of Similarity. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2011:895-904.

50. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases*. 1987;40(5):373-83.

51. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*. 2005;43(11):1130-9.

52. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16:297-334.

53. Gibbons JD. Nonparametric methods for quantitative analysis. 3rd ed: American Sciences Press. 1997.

54. Savova GK, Harris M, Johnson T, Pakhomov SV, Chute CG. A data-driven approach for extracting "the most specific term" for ontology development. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2003:579-83.

55. Stopword List. Available from: http://www.textfixer.com/resources/common-english-words.txt.

56. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. *Proceedings of Annual Symposium on Computer Application in Medical Care*. 1994:235-9.

57. EPIC. Available from: http://www.epic.com.

58. GATE. Available from: http://gate.ac.uk.

59. Hunt RJ. Percent agreement, Pearson's correlation, and kappa as measures of inter-examiner reliability. *Journal of Dental Research*. 1986;65(2):128-30.

60. McCray AT, Srinivasan S, Browne AC. Lexical methods for managing variation in biomedical terminologies. *Proceedings of Annual Symposium on Computer Application in Medical Care*. 1994:235-9.

61. Zhang R, Pakhomov S, Melton GB. Automated Identification of Relevant New Information in Clinical Narrative. *Proceedings of the ACM SIGHIT International Health Informatics Symposium*. 2012:837-41.

62. McInnes BT, Pedersen T, Pakhomov SV. UMLS-Interface and UMLS-Similarity : open source software for measuring paths and semantic similarity. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2009:431-5.

63. Farri O, Rahman A, Monsen KA, Zhang R, Pakhomov S, Pieczkiewicz DS, et al. Impact of a prototype visualization tool for new information in EHR clinical documents. *Applied Clinical Informatics*. 2012;3(4):404-18.

64. Zhang R, Pakhomov S, MaInnes BT, Melton GB. Evaluating Measures of Redundancy in Clinical Texts. *Proceedings of the American Medical Informatics Association Annual Symposium*. 2011:1612-20.

65. Zhang R, Pakhomov S, Lee JT, Melton GB. Navigating longitudinal clinical notes with an automated method for detecting new information. *Studies in Health Technology and Informatics*. 2013;192:754-8.

66. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. *Journal of the Amercian Medical Informatics Association*. 2010;17(1):19-24.