Measuring the Quality of Unstructured Text in Routinely Collected Electronic Health Data: A Review and Application

by

Marcello Nesca

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Department of Community Health Sciences

University of Manitoba

Winnipeg

ABSTRACT

Introduction: Routinely collected electronic health data (RCEHD), which are generated during the process of managing and monitoring the health care system, can be comprised of structured, semi-structured, or unstructured information. Electronic medical records (EMRs), one type of RCEHD, often contain unstructured text data (UTD), which are typically prepared for analysis (i.e., preprocessed) and analyzed using natural language processing (NLP) techniques. At present, there are few studies about the specific types of NLP methods used to preprocess UTD to address data quality issues prior to analysis or modelling, or information about the impact of preprocessing on EMR data quality.

Purpose & Objectives: The purpose was to examine preprocessing methods for UTD and evaluate the quality of UTD in EMRs. The objectives were to: 1) systematically document current research and practices for preprocessing UTD to describe or improve its quality, and 2) apply data quality indicators identified from current research and practices to UTD in EMRs from the Manitoba Primary Care Research Network and describe the quality of these data.

Methods: Objective 1 involved a scoping review. Scopus, Web of Science, ProQuest, and EBSCOhost were searched for literature on current research and practices to prepare UTD for analysis, up to and including 2021. For objective 2, a case study was undertaken where data quality indicators and preprocessing methods identified in the scoping review were applied to UTD from EMRs. Notes were first converted into a data frame, then an exploratory analysis of the EMR notes using preprocessing methods was conducted. Analyses were stratified by clinic site and year of note creation. The words or characters affected by the preprocessing methods were also described.

Results: A total of 41 articles were included in the scoping review for objective 1; over 50% were published between 2016 and 2021 and over 90% were empirical research articles. Almost one-fifth of the articles were published in health sciences journals. Common preprocessing methods used to restructure and reorganize UTD included removal of stop words, punctuation, and numbers, word tokenization, and parts of speech tagging. Data quality indicator topics for UTD in EMRs included misspelled words, security, word variability, sources of noise, quality of annotations, ambiguous abbreviations, and manual annotations. For objective 2, we selected 193,206 clinical encounter notes from EMRs for 6891 patients between 1985 and 2020 in 12 clinic sites. Descriptive statistics, including means, standard deviations (SDs), frequencies, and percentages, were used to characterize data quality. Overall, the clinical encounter notes contained an average (standard deviation [SD]) of 27.3 (27.0) stop words, 25.7 (27.8) punctuation symbols, 12.1 (11.1) spelling errors, and 2.9 (2.6) special characters. The average (SD) length of a clinical encounter note was 555.8 (551.1) characters, and 71.5 (59.7) words. Lexical diversity, a measure of the range of vocabulary used, had a mean (SD) of 86.2 (11.9). There was little variation in these characteristics of UTD in EMRs over time, but substantial variability across clinic sites.

Conclusion: This study identified multiple data quality indicators that have been used to preprocess UTD in published literature and demonstrated their application to real-world data. Future research in four areas would be beneficial: 1) creation of operational definitions for data quality dimensions specific to text data properties, 2) inspection of other text data quality indicator topics, 3) exploration of quality indicators for short text documents, and 4) mapping operational definitions for text data quality dimensions to data quality indicators for RCEHD.

ACKNOWLEDGEMENTS

I would like to thank and acknowledge my supervisor Dr Lix for the knowledge, support, and feedback during this thesis project, and for the opportunity to work in the field of data quality. I would also like to thank my committee members for their key perspectives in this research, in particular Dr Katz for his expertise in the field of Family Medicine and administrative data, and Dr Leung for his expertise in the field of Computer Science in database and data mining. A big thank you to MaPCReN for providing EMR data for analyses. I also wish to acknowledge the Visual and Automated Disease Analytics program for further advancing my overall data science skills, especially developing an interest in natural language processing. I wish to acknowledge a colleague, and a good friend, Rob Balshaw for bringing my statistics and R coding skills to a new level. Lastly and most importantly, I wish to acknowledge my wife Annie, along with family and friends Thank you for your support.

Marcello Nesca
Sept 2018 – Dec 2021

TABLE OF CONTENTS

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Definition
RCEHD	Routinely collected electronic health data
EMR	Electronic medical records
EHR	Electronic health records
UTD	Unstructured text data
CIHI	Canadian Institute for Health Information
MCHP	Manitoba Centre for Health Policy
ICES	Institute for Clinical Evaluative Services
PHAC	Public Health Agency of Canada
NLP	Natural language processing
NER	Named entity recognition
DTM	Document term matrix
MaPCReN	Manitoba Primary Care Research Network
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SD	Standard deviation
IQR	Interquartile range
CI	Confidence interval
CPCSSN	Canadian Primary Care Sentinel Surveillance Network
EDA	Exploratory data analysis
CNN	Convolutional neural network
URL	Uniform resource locator
HTML	Hypertext markup language
SOAP	Subjective, Objective, Assessment, Plan

Chapter 1 - Introduction

1.1 Background

Routinely-collected electronic health data (RCEHD) are data generated during the process of managing and monitoring the healthcare system(1,2). While RCEHD were never intended to be used for research, they are now widely used for studies about health and healthcare use. There are many benefits associated with using RCEHD for research, including: 1) Population reach – RCEHD often cover entire populations and therefore can be used to study rare outcomes and population sub-groups; 2) Longitudinal information – RCEHD are often collected over long periods of time, and therefore enable long-term follow-up; 3) Reduced biases – attrition and non-response bias are reduced when using RCEHD for research; 4) Reduced cost – RCEHD are not as costly to collect as primary data; 5) Real world setting – RCEHD present opportunity for studies about real-world changes in healthcare policy or practice(3).

RCEHD can be comprised of structured, semi-structured, and unstructured data. However, most research studies that use RCEHD have primarily relied on structured data. Diagnosis and medication codes, which are examples of structured data, are typically stored in relational databases and are identifiable by humans or machines(4,5) because the data are comprised of standardized fields or parameters. For example, an electronic medical record (EMR), which is used to record the information gathered during a patient's visit to a physician or other healthcare provider, can contain diagnosis codes, dates of patient visits, and prescription medication codes. Semi-structured data typically follow some standard conventions for their creation and may contain some identifiable medical terms or codes that can facilitate human or machine readability(6–8). In contrast, unstructured data do not have an identifiable format or structure(5). A lot of unstructured data is in text format, but may also have other formats, such as images, voice recordings, and video data.

Unstructured text data (UTD) are particularly common in EMRs, one type of RCEHD. It is estimated that nearly 80% of the information contained in EMRs is unstructured(9). EMRs, which are routinely generated as part of the patient care process in both outpatient and inpatient settings, are increasingly being used for research because they can provide important clinical context about the patient’s health and healthcare use(4). This includes information about behavioral risk factors such as smoking and alcohol use, clinical measures such as blood pressure, and personal and family health histories of chronic diseases or health conditions.

The quality of data has been defined as their fitness for use(10–12), which means that data meets the needs of the user for a specific task or purpose, such as identifying individuals with a specific chronic disease. More specifically, data quality has been defined as “the totality of features and characteristics of a data set, that bear on its ability to satisfy the needs that result from the intended use of the data,”(13). Given that RCEHD are increasingly being used for research, it is important to consider the fitness of the data for research projects, including epidemiologic studies or health services utilization studies. However, data is also often assessed without having a specific purpose in mind; it may be described with respect to a broad range of purposes, such as conducting population health studies. Thus, questions such as “How complete and how consistent are the data?” or “How precise is the information?” may be addressed in order to provide a general perspective about the quality of the data.

There are a number of consequences of poor data quality. For example, Kiefer notes that poor data quality can “slow down innovation processes”(14). Poor data quality may also increase the time required to prepare a dataset for use, which can impact the timeliness of research outputs. The potential for inaccurate research results is often of paramount importance when data are of poor quality (i.e., are inaccurate or incomplete) (15).

Data quality is a multidimensional construct, which means that there is no single summary measure of quality that exists. Quality includes such data attributes (i.e., dimensions) as relevance, consistency, accuracy, comparability, timeliness, accessibility and usability(11,12,15,16). Frameworks that depict these various dimensions of data quality have been developed by organizations that create or use electronic data. For example, data quality frameworks have been developed by national organizations, such as Statistics Canada and the Canadian Institute for Health Information (CIHI), and the Public Health Agency of Canada (PHAC), as well as provincial/territorial data centres, such as the Manitoba Centre for Health Policy (MCHP), Ontario’s ICES, and Population Data BC. Many of these frameworks also capture indicators (i.e., measures) of the dimensions of data quality. Examples of indicators include the percent of missing values, the percent of invalid values, and the availability of data documentation. The MCHP and ICES frameworks also make a distinction between data quality for a specific use (i.e., research project) and general measures of data quality that can be assessed when data are first acquired into the organization, to provide an overall picture of their accuracy, completeness, and comprehensiveness.

Most data quality frameworks have been developed with structured data in mind. For example, an indicator such as the percent of missing values is not relevant for UTD. However, Kiefer (14) has argued that most, if not all, data quality dimensions developed for structured data are also relevant for UTD, although she emphasized the importance of relevance, interpretability, and accuracy when assessing UTD fitness for use. Kiefer suggested that indicators of these dimensions of UTD quality could be developed by comparing UTD to “three classes of ideal data sets”, which are: (a) the data expected by the data consumer, to define interpretability, (b) the data optimal for a specific task, to define relevance, and (c) the data that represent the real world, to define accuracy. While Kiefer provides an important overview of the issue of quality for UTD, she and others have noted that there has been little research about data quality dimensions and indicators of these dimensions for UTD.

The usability of UTD for research generally requires the application of natural language processing (NLP) techniques. NLP is a form of artificial intelligence used to understand and interpret natural language and, in some applications, to generate natural language (i.e., to generate text from speech). NLP techniques to understand and interpret UTD include topic modeling, sentiment analysis, aspect mining (e.g., identifying different parts of speech), text summarization, and named entity recognition (NER) (e.g., identifying people, places, and other entities in unstructured data). Topic modelling is a method for determining the content of a document by clustering words(17,18) (e.g., determining the topics from abstracts of published articles(19)). Sentiment analysis is used to determine the emotional context of a document or a group of documents (e.g., determining negative or positive reviews of a product sold online)(20).

To prepare UTD for one or more of these NLP techniques, preprocessing of the data is an essential step. One definition of preprocessing data is the set of operations taken on a data set before modelling and analysis(21). Furthermore, the role of preprocessing is focused on processes that make the data more meaningful for analysis, and solving any data issues that would prevent analysis to extract any meaningful knowledge(21). Simply put, preprocessing is essential to create data that are fit for use. Preprocessing of UTD includes such actions as removing stop words (i.e., common words in a language), removing punctuation, tagging (i.e., identifying or labelling) parts of speech, and transforming abbreviations into words or phrases so that they can be easily interpreted. Accordingly, some researchers have suggested that indicators that measure the outputs of data preprocessing steps, such as the number or percent of abbreviations and the number or percent of spelling errors, could be used to characterize UTD quality. Some types of NLP, such as NER, involve training classification models to learn to identify data entities, such as parts of speech, diseases, or geographic locations(22). This requires the use of annotated databases as a "gold standard" (i.e., reference standard), which have been tagged (e.g., parts of speech have been documented or labelled) using manual or automated methods(22). The quality of these annotated databases and the accuracy of classification models based on NLP applications involving annotated databases have also been proposed as indicators of UTD quality. Additionally, several unsupervised (or supervised) methods, which are used for internal (or external) validation have been used for preprocessing and may be used to develop indicators for data quality. For example in a study by Zennaki et al, a recurrent neural network was used for unsupervised and semi-supervised parts of speech tagging in languages for which there are no labeled training data(23).

In summary, there are potentially many indicators that could be used to assess UTD quality. However, there is little relevant literature and few, if any, guidelines on the data quality indicators that might be recommended for inclusion in data quality frameworks for UTD, or that might be used to guide data preprocessing in studies that apply NLP methods to UTD. In addition, there have been few studies that have investigated the impact of UTD quality assessment on the performance of text analyses using NLP. In one example, Alnajran et al. (24), compared a heuristic (i.e., problem-based) NLP UTD preprocessing method to a conventional method of UTD preprocessing, to assess the impact of preprocessing on the performance of similarity measures for short texts from Twitter. The authors hypothesized that a methodical and systematic approach to data preprocessing could lead to improvements in the measurement of similarity between texts and reduce error rates; they found some evidence to support their hypothesis. Overall, they recommended a comprehensive and systematic approach to preprocessing of UTD in order to improve its quality for research.

1.2 Purpose and objectives

Given this background, the purpose of this research was to examine preprocessing methods for UTD and evaluate the quality of UTD in RCEHD with a focus on EMRs. The objectives were to:

1) Systematically document current research and practices to preprocess UTD to describe or improve its quality, and

2) Apply and describe data quality indicators identified from current research and practices to UTD in EMRs.

1.3 Thesis organization

The thesis is organized as follows: Chapter 2 provides a review of relevant literature about methods to assess the quality of in EMRs and other RCEHDs, UTD quality indicators, and NLP techniques to preprocess UTD. Chapter 3 describes the methods used to conduct a systematic review about preprocessing of UTD and conduct a case study about preprocessing for UTD in EMRs. Chapter 4 presents the results of the scoping review and the case study. The final chapter of this thesis explores the key findings and their interpretation, the research strengths and limitations, and the significance of the research.

Chapter 2 - Literature Review

This chapter provides an overview of methods and approaches to assess the quality of EMRs and other RCEHDs, UTD quality indicators and characteristics of UTD for consideration when describing data quality, NLP techniques to assess data quality. The chapter concludes with a summary of the literature review that identifies gaps in existing research about UTD quality and the expected contributions of the proposed research.

2.1 Methods to measure the quality of EMRs and other RCEHD

This section provides an overview of the history of data quality frameworks, the dimensions of data quality in RCEHD, and methods for measuring data quality.

2.1.1 History of data quality dimensions and indicators in frameworks

Data quality indicators identified in frameworks were initially based on ISO9000 standards, where products and services were associated with quality standards for processes and production. These frameworks originally included 1) management responsibilities, 2) operation and assurance costs, 3) research and development, 4) production, 5) distribution, 6) personnel management, and 7) legal functions(25). However, the extension of ISO9000 standards to data quality research proved to be difficult; these standards were not intuitive for characterizing data quality for research and operationalizing the indicators(25).

2.1.2 Dimensions of data quality in RCEHD

Kerr proposed a comprehensive approach to develop a data quality strategy and framework for the New Zealand Ministry of Health's RCEHD. This approach included: 1) defining dimensions of data quality, 2) establishing a balance between scientific rigour for researchers and ease of understanding for data users, 3) ensuring standardization but with sufficient flexibility that a framework can be applied to different types of data, 4) transparency, and 5) assessment tools for data quality (26). The combination of government policy needing high quality RCEHD along with a rapid increase in research using RCEHD highlighted a need for a data quality framework for use with health data(10,27,28).

Dimensions of quality that have been used to characterize structured data in RCEHD, include accessibility, usability, security and confidentiality, accuracy, timeliness, completeness, consistency, integrity, provenance, internal and external validity, interpretability, value, and relevance(11,12,15,16,29). A literature review(30) about the properties that are relevant to measuring data quality in EMRs revealed a large diversity in dimensions and indicators. The main dimensions of data quality that were identified in the literature review included accuracy, completeness, comparability, and validity. Timeliness of data in EMRs was not assessed as often as the other data quality features(30). This is a concern because timeliness depends on if the data are up to date to conduct the task at hand(15,31). The article concluded that there is a lack of standardized methods to measure data quality in EMRs. Since the range of measurement methods is diverse, this makes it difficult for methodological comparisons across EMR data sources or frameworks(28).

Weiskopf and Weng (32) conducted a review of published, peer-reviewed articles to examine methods for data quality assessment and dimensions of data quality in electronic health records (EHRs). Key terms used to define the literature search strategy included terms for data quality and its synonyms (e.g., data accuracy, data reliability), as well as terms for EHR and its synonyms (e.g., computerized medical record). A total of 95 articles were included in the review. Almost three-quarters of these articles focused on structured EHR data, while the remaining one-quarter of these articles focused on both structured and unstructured EHR. The authors identified 27 unique terms used to describe dimensions of data quality. For example, recency and timeliness were used to describe currency, while agreement, consistency, reliability, and variation were used to describe concordance. Five dimensions of data quality were extracted from the articles: completeness, correctness, concordance, plausibility, and currency(32). Information about methods of data quality assessment were also extracted from the articles and summarized into seven categories: 1) Gold standard - where an external dataset with or without information from the EHR was used as a comparison standard. 2) Data element agreement – where two or more EHR data points are compared to determine their consistency. 3) Element presence – a decision is made whether or not desired or expected data points are present in the data. 4) Data source agreement – data from EHRs is compared with data from another source to assess agreement. 5) Distribution comparison – summary statistics for EHR data are compared to their expected values. 6) Validity check – methods that measure the credibility of EHR data are applied. 7) Log review – this involves an examination of the documents that contain information about data entry practices. Weiskopf and Weng then mapped data quality dimensions to the methods used to measure data quality. Completeness was the most frequently measured data quality dimension; it was mentioned in 64% of the articles. Methods used to measure completeness included utilizing a gold standard, element presence, data element agreement, data source agreement, and distribution comparison. Correctness was measured in 60% of the articles; methods such as utilizing a gold standard, data element agreement, validity checks, data source agreements, and log reviews were used to measure correctness. Weiskopf and Weng noted some gaps to be addressed in future research, including adopting a framework or taxonomy for data quality which is consistent with previous research(28), raising awareness about the increasing need to utilize EHRs for research, utilizing methods from other disciplines to measure data quality, and adopting statistical methods for data quality assessment(32).

The objective of the scoping review conducted by Bian et al was to synthesize data quality dimensions and assessment methods and to evaluate a data quality framework currently in use – the Patient Centred Clinical Research Network's data quality framework (PCORnet)(33). The scoping review included previously-reviewed articles from systematic reviews conducted by other authors (include articles from the Weiskopf and Weng review), data quality frameworks, and relevant articles published up to February 2020(33). The aim of the scoping review was to cross-tabulate data quality dimensions and data quality assessment methods. The authors then focused on identifying gaps in PCORnet by mapping the data characterisation checks from PCORnet to the scoping review extracted results. A total of three literature reviews, 20 data quality frameworks, and 226 data quality studies were included in the review. The authors extracted 14 data quality dimensions from the articles: 1) currency, 2) correctness/accuracy, 3) plausibility, 4) completeness, 5) concordance, 6) comparability, 7) conformance, 8) flexibility, 9) relevance, 10) usability/ease-of-use, 11) security, 12) information loss and degradation, 13) consistency, and 14) understandability/interpretability(33). Furthermore, the authors extracted three new data quality assessment methods on top of those identified in Wieskopf and Weng's literature review: 1) conformance checking, which includes a) checking the uniqueness of objects that should not be duplicated, b) dataset agreement with prespecified or additional constraints, and c) agreement of object concepts and format granularity between two or more datasets(33). 2) Qualitative assessment – descriptive qualitative measures with group interviews and interpreted with grounded theory(33). 3) Security analysis – analysis of reports about security issues(33). The scoping review results were then compared to the data quality dimensions and assessment methods found in PCORnet. Bian et al found that the data quality assessment methods missing from the PCORnet data quality frameworks were data element agreement, data source agreement, comparison to a gold standard, and qualitative assessment(33). Furthermore, the data quality dimensions missing from PCORnet were comparability, flexibility, relevance, usability/ease-of-use, consistency, and understandability(33).

Hinds et al conducted a scoping review to describe characteristics of published Canadian studies about the quality of administrative health data(27). Data quality dimensions were selected from the data quality frameworks presented by MCHP, ICES, Statistics Canada, PHAC, and CIHI. The authors found that almost 97% of articles identified in the scoping review were validation studies. Hospital discharge abstracts and physician billing claims were the most frequently validated data sources. Furthermore, one or more elements of administrative data were compared to a reference standard, and more than half of the studies conducted were validating a disease algorithm or case definition. Among the validation studies the data quality dimensions most often addressed were reliability and correctness. Furthermore, the most common measures of validity were sensitivity and positive predictive value. The authors also discussed several gaps in the literature; most of the studies about data quality were conducted in a few provinces which makes it difficult for cross-provincial comparisons. Few studies reported on the quality of EMRs. Lastly, is the authors identified a need for more studies to examine different data quality dimensions and noted there is no commonly used methods or reference standards to measure the quality of administrative health data(27,28).

The scoping reviews described in this chapter reveal that knowledge synthesis methodologies can be used to investigate data quality dimensions and the methods to measure them (27,28,32,33). Furthermore, these reviews also provide guidance about how to use data quality dimensions and methods for informing and building data quality frameworks(27,33). As a result, these scoping reviews are the first step in creating a structure for capturing data quality dimensions and common methods to measure the quality of data within RCEHD, including EMRs.

2.1.3 Methods for measuring data quality in RCEHD

Through the application of a data quality framework to a set of data, quality can be measured using individual indicators that correspond to one or more data quality dimensions. For example, in the MCHP data quality framework, the dimension of accuracy is associated with indicators of completeness and correctness. Completeness is measured using the rate (i.e., percent) of missing values, as well as with measures of geographic coverage. The MCHP data quality framework involves an assessment of data quality in two phases: during data acquisition and during the research phase. The framework provides a formalized process for evaluating quality at the acquisition level; it is accompanied by a description of the techniques and tools used to assess data quality. The MCHP framework contains six data quality dimensions: accuracy, internal validity, external validity, timeliness, interpretability, and value.

ICES created the DataFit Toolkit, which is a data quality user guide similar to the MCHP data quality framework(34). In the ICES framework, a distinction is made between quality measured at the database level and quality measured at the research level. When new data are acquired or existing data are refreshed, the Datafit Toolkit is applied. The data quality dimensions in the Datafit Toolkit are accuracy, internal validity, external validity, timeliness, and interpretability. Accuracy is assessed using indicators of completeness and correctness; completeness is measured by the rate of missing values. The Datafit Toolkit includes the dimension of relevance; this dimension is not included in the MCHP data quality framework.

PHAC's data quality framework has four stages: 1) planning – data quality processes are prepared and prioritized for a data holding, 2) implementation – data quality processes are applied, 3) evaluation – the data holding is evaluated for its quality, and 4) feedback – feedback from the evaluation is cyclical and used in the planning stage(11). The dimensions of quality represented in this framework are: accuracy, timeliness, usability, serviceability and relevance. The PHAC quality framework differs from the MCHP and ICES frameworks in the quality dimension of serviceability. Indicators of serviceability include 1) consistency of the database over time, 2) using standard classification systems, 3) the ability to link the data and re-identify it, 4) using appropriate methodology according to best practices, 5) data element integrity where data is reasonably collected at its finest detail, along with the maintaining the original data elements, 6) data dictionary standards where a data dictionary exists, and data elements correspond to it. PHAC lists some indicators of data quality dimensions but only do so conceptually. For example, a criteria for accuracy is providing an operational definition for the relevant subject matter along with explicitly documenting the population reference(11).

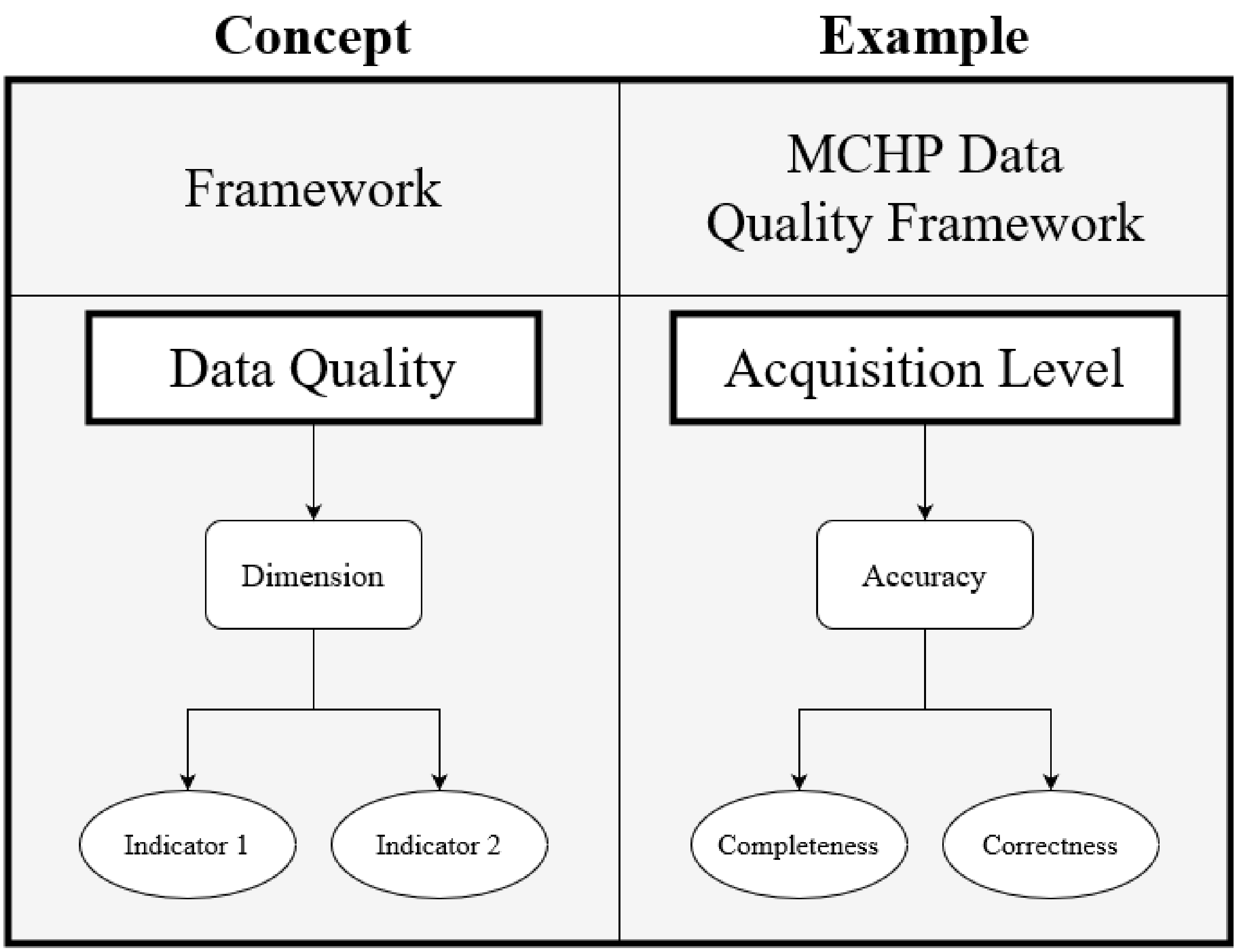
CIHI is a not-for-profit government funded organization that has also created a data quality framework. Methods to measure data quality involve five phases of the information lifecycle: 1) The creation and capture of data from research organizations, health care providers, or institutions which are facilitated by CIHI to produce standards and support data providers. 2)

Research organizations set up processes to prepare data though submission specifications, standards and tools set by CIHI. 3) Data is then prepared, which means that it is verified, cleaned and transformed for statistical analysis. 4) Processed data is then used for statistical analysis to generate tables, reports, interactive tools and infographics. 5) Data is distributed and disseminated by CIHI stakeholders(12). The dimensions of quality in the CIHI framework are 1) relevance – information that caters to the needs of stakeholders. 2) accuracy and reliability – information that correctly and consistently describes the phenomena in question. 3) comparability and coherence –information is comparable over time and across jurisdictions using common standards and methods. 4) timeliness and punctuality – this dimension refers to how quickly information is made available and if information is delivered on time. 5) accessibility and clarity – the degree to which supplemental information is easily obtainable and clearly presented(12). No indicators are included in the CIHI data quality framework.

Statistics Canada's data quality framework is guided by several principles: 1) quality is a multidimensional construct; 2) quality is relative; 3) all employees are involved in assuring quality; 4) quality involves each phase of a business process model; 5) a team approach is best to balance the dimensions of data quality; 6) quality assurance adapts to specific programs or use cases; 7) users of the data need to be informed about the quality dimensions to measure statistical information that is appropriate for use; and 8) data quality is a continuous process(16). The framework developed by Statistics Canada has six data quality dimensions: relevance, accuracy, timeliness, accessibility, coherence, and interpretability(16). Conceptual definitions of indicators of data quality are provided, but operational definitions are not. For example, relevance can be measured by: 1) capturing emerging trends and changes to sociological and economical frameworks while identifying gaps in statistical programs, 2) communicating information to data users and stakeholders, 3) ensuring production of high-quality and timely information through planning and priority setting, 4) supporting current statistical programs and new initiatives by redirecting funds to such initiatives and procuring additional funding(16).

The frameworks reviewed in this chapter contain both guiding principles and methods for evaluating the quality of data in organizations. While there is a certain structure used to develop frameworks for data quality (e.g., the quality frameworks contain guiding principles and quality dimensions; see Figure 1 for an example of a data quality framework), the quality dimensions are not standardized across research and government organizations. However, a commonal aspect of the data quality frameworks reviewed in this section is that they are intended for data that are structured. No framework has been proposed that represents quality for UTD.

Figure 1 - Concept and example of a data quality framework



2.2 UTD quality indicators

Quality indicators for UTD, similar to quality indicators for structured data, can be used to assess the fitness of the data for specific purposes, such as the application of machine-learning algorithms or statistical models(35). However, some quality indicators are also intended to provide a general overview of the key attributes or characteristics of UTD. While there is limited literature on UTD quality, there are some key papers that provide an overview of UTD quality indicators.

Indicators to characterize data quality include measures of noise(14,36). Noisy text data – as Kiefer defines it – result in difficulties of both humans and machines to read text data due to variability in words or sentences (i.e., words that are spelled wrong, abbreviations, special characters that are not alphanumeric or punctuation based or unnecessary punctuation symbols)(14). Berndt et al discusses how sources of noise are introduced as more data become available(35). These sources of noise include misspellings, abbreviations, and slang. Sonntag refers to noise indicators as defining representational quality and notes that noise indicators in UTD include misrepresentations of text that are not consistent or concise(36). Sonntag gives an example where representation is indicated through consistency of data structures or data values, where a reduction of redundancy in data is ranked as having high consistency. Representational deficiencies include incorrect data values (e.g., typographical errors), spelling errors, duplicates of documents, and homonym conflicts. Homonyms are words that have the same spelling but different meaning; they can lead to ambiguity of words or abbreviations(36). Kiefer focused on data quality at the text level (i.e., at the word or sentence level) while Sonntag focuses on quality at the document level. It is important to note that UTD is often characterized as being high-dimensional, where the number of features (variables) exceed the number of observations(37). The reduction of noisy text - and its high dimensionality - is related to the bias-variance error trade off. Bias refers to an error that occurs by having statistical or machine learning models trained on too few features (i.e., an overly simplistic model that misses information, which is known as under fitting). Variance is an error that refers to having an overly complex model trained on too many features and is not generalizable to future data (i.e., overfitting)(38,39). Since UTD is characterised as high-dimensional, building statistical or machine learning models without caution could lead to having an imbalance between the bias-variance trade off where models could have high variance and low bias. There are many ways to reduce complexity (high variance – too many features) in UTD and reducing sources of noise is a common and needed preprocessing step for data quality purposes.

Another quality indicator is the fit of the model in a training dataset; Kiefer describes how well a machine reads data as input and how similar it is to the data output expected by the end user(14). However, Berndt et al note that while text classification algorithms applied to training data may not be susceptible to noise as the amount of text data increases, they can be susceptible to noise if training data are comprised of short texts(35). Examples of short texts include clinical notes and social media posts. Spasic et al, in their systematic review, synthesized evidence about the properties of UTD used to train machine learning techniques(40). Two examples of these properties are size and annotation of training data. Size includes the number of documents or words(40). Annotation occurs when text data elements are tagged with specific indicators (e.g., a word is tagged with its part of speech or identified named entity). Regardless of whether annotation is conducted using automatic or manual methods, it can lead to better representation of data for predictive modeling(40). Lastly, when discussing model fit on training data, Taleb et al notes that as data become more unstructured and heterogeneous, some data quality dimensions can be affected(41). For example the ease of reading UTD is assessed by transforming it into structured data, which would allow for accessible measurement and query(41). The authors conclude that as data becomes more structured, the ability to evaluate it and measure its quality increases(41). As a result, the amount of structure within a dataset may serve as an indication of quality.

Specificity or relevance may also be used to describe UTD quality. Both Kiefer and Sonntag describe specificity and relevance as having domain specific language in the context of a specific use case(14,36). However, while Kiefer notes that while specificity is an indicator of relevance, Sonntag considers relevance to be an indicator of contextual quality.

For the data quality dimension of relevance, Kiefer describes a quality indicator of frequent key words, where certain words appear more than others depending on the discipline, industry or trade(14). For example, a weather report is meant to have short and incomplete sentences, whereas a pilot’s manual contains strictly-defined and specific vocabulary(36).

There is a lack of consistency in definitions of accuracy for UTD, although several authors note that an improvement in accuracy is a desirable outcome(14,36,42). Kiefer describes accuracy from an operational perspective, such as the percentage of correct decisions from a classification algorithm(14). Sonntag takes a more contextual perspective, where data that are error-free and precise are considered to be of high quality(36). Assale et al discuss accuracy from the perspective of applying preprocessing steps so that clinical UTD will be more reliable and accurate. Preprocessing steps can include reduction of noise (i.e., variability), reduction of heterogeneity (e.g., abbreviation and word ambiguation) and redundancy, reduction of the amount of data through dimension reduction methods (e.g., unsupervised methods), conversion of data into a form that is suitable for analysis, and privacy and anonymity processing(42). However, there is a lack of studies that demonstrate the impact of preprocessing methods on the outcomes of text mining, especially in short text scenarios such as clinical notes in EMRs(43).

There is no single agreed upon method for measuring UTD quality, although there are some similarities in the indicators described in existing literature, such as indicators of accuracy. However, the methods proposed to measure these aspects of data quality are heterogeneous. The costs of having highly heterogeneous methods of measuring indicators are that they may not be comparable across different datasets, while the benefits can be that data quality indicators can be highly specific to a particular context.

2.3 NLP techniques to measure UTD quality

UTD contain characteristics that are unique to data analysis when compared to structured data. These characteristics include properties such as abbreviations, word length, character length or spelling errors. Therefore, the quality of UTD can be measured using text specific analyses, namely, NLP techniques(44). Two approaches to solving NLP-based problems are rule-based methods and machine/statistical learning methods.

Rule-based NLP methods are pattern matching methods that involve setting rules; these rules can be used for several purposes, such as operationalizing indicators of noisy text data. A rule-based method such as regular expressions is used to describe patterns in string data(45) and to capture special characters, URLs, mail (or email) addresses, emoticons, abbreviations, or pause-filled words. According to Kiefer, noise can be measured using certain indicators captured by regular expressions, such as: percentage of sentences that cannot be parsed (i.e., breaking down a sentence in its component parts), unknown words, punctuation, long and short sentences, incorrect use of upper or lowercase, special characters, URLs, mail or email addresses, emoticons, abbreviations, pause-filling words (e.g., “uhh” or “umm”), rare words (i.e., words that that have a low frequency of occurrence), and spelling errors (14). Kiefer also uses lexical diversity, which is defined as how ‘rich’ a vocabulary in a text is, or the ‘range’ of vocabulary used(44,46). Kiefer calculated lexical diversity using the type-token ratio where the number of different words (the ‘type’) is divided by all the words in the text (i.e., tokens)(44,47).

Other rule-based methods include sentence parsing, synonym extraction, the original Levenshtein distance metric and the evolved Damerau-Levenshtein distance metric. Sentence parsing is splitting out words in a sentence to either its conceptual parts or different parts of speech or grammar development using rules and patterns (i.e., what words are nouns or verbs)(48–50). Synonym extraction is the process of automatically determining one or two or more words that have the same meaning, and therefore are semantically related(51). Rule-based methods can achieve this through building a dictionary corpus containing synonyms of words(51). Lastly, the original Levenshtein distance is a rule-based metric from information theory that is used to determine the "closeness" of two strings (i.e., similar concept to correlation)(42). Levenshtein distance was used to calculate the number of edits required to change one word to match a second word(42) (e.g., a "distance of one" requires one edit on the first word to match the second word). The Damerau-Levenshtein distance metric is an evolved distance metric that also takes into account inverted letters in a word which are common in spelling errors (e.g., words that contain "ie" are commonly reversed as "ei")(42).

Machine/statistical learning uses probabilistic or classification modelling to describe the structure of data or perform predictions based from inputs of a dataset(52). Machine/statistical learning algorithms are classified as supervised or unsupervised(52). Sonntag presented Word Sense Disambiguation (WSD) as an NLP method to identify and resolve homonym conflicts within text; this task determines which sense of the word belongs in a certain context(53). Latent Semantic Indexing – an unsupervised NLP method – is the computational framework for WSD(36). Other machine/statistical learning methods can be applied to assist in annotations (i.e., tagging words or documents) or testing for the quality of annotations(54–56). For example, supervised learning algorithms can automate parts of speech tagging (e.g., tagging a word as either a noun or a pronoun or a verb), or tagging named entities (i.e., words identified as persons, places, or addresses) by either training data on manual annotations, or using an external dictionary(54–56).

2.3.1 Other techniques for measuring UTD quality

Frequencies and percentages can also be useful to describe UTD quality. Rare words are identified by counting all the words less than some threshold in the corpus through NLP measures such as tokenization or regular expressions(14). For example, if the frequency of a word is less than the average frequency of all words, it is considered rare. Kiefer further explained that rare words are also found by counting words that are not found in either a standard dictionary or a generated dictionary(14).

Measures of linguistic quality using NLP techniques have also been proposed, including readability, comprehension, sentence cohesion, and reader coherence(57–59). The readability measures have been used on health care data and their accuracy has been validated(58). Sentence cohesion and reader coherence were measured as well. However, beyond Batini et al's work, the only measures that are available to measure local and global sentence cohesion are the Tool for the Automatic Analysis of text COhesion (TAACO), which is used to measure how well a text is understood(60), and the Coh-Metrix – the analysis of text cohesion, which, – similar to the TAACO, measures how well a text is understood and its readability(61,62). Unfortunately, the cohesion measures come in standalone applications, with no current integration into programming languages such as R (statistical computing and graphics environment) or Python (general purpose, object oriented, and high-level programming environment).

2.4 Summary

This chapter provided a review of current literature about data quality indicators for UTD, the methods to measure data quality in RCEHD and EMRs, and NLP methods to measure UTD quality. Overall, there is some literature that discusses indicators of UTD, such as measures that characterize noisy data, fit of a NLP model to training data, specificity, relevance (context), and accuracy. Methods to measure the quality of data were mostly portrayed from the perspective of data quality frameworks. The frameworks proposed by five organizations, which represent both research and government perspectives, were reviewed. Common features of these frameworks include the use of data quality dimensions as a guide to measure data quality, and the specification of indicators to measure the data quality dimensions. However, these frameworks were created for the purpose of measuring or characterizing the quality of structured data, not unstructured data. Lastly, NLP methods for measuring text quality were reviewed; this section highlighted techniques and ways to measure data quality dimensions through indicators, such as measures of noisy text data. Other aspects of data quality measurement are unique to text data, such as readability or lexical diversity.

2.5 Gaps in existing literature and novel contributions of the proposed research

Gaps exist within research about techniques to measure the quality of UTD. No data quality framework has been created to guide the analysis of UTD, especially for text data found in RCEHD. Furthermore, while data quality frameworks exist, there is no formalized structure to capture and measure the indicators of data quality that is standardized among different organizations. Since existing data quality frameworks are primarily defined for structured data, it is difficult to formalize a structured quality framework for UTD. One persistent challenge is the lack of formal or consistent processes to report on methods for data quality assessment(28,63).

Chapter 3 - Methods

To achieve objective 1, a scoping review of published research literature about practices and methods to prepare UTD for analysis was conducted. To address objective 2, a case study was undertaken in which preprocessing methods were applied to UTD from clinical encounter notes in EMRs. Data quality indicators and their methods were generated in the analysis.

Ethics approval for the case study was provided by the University of Manitoba Health Research Ethics Board (HREB). Data access approval was given by the Manitoba Primary Care Research Network (MaPCReN) Director.

3.1 Scoping review

The purpose of a scoping review is to map (i.e., describe) the literature on a new topic or research area and identify key concepts, gaps in the research area, and types and sources of evidence to inform future research (64). Key steps include defining the search strategy and study inclusion and exclusion criteria, screening and extracting data from eligible articles, and synthesizing/summarizing the extracted information(65).

The scoping review framework developed by Arksey and O'Malley was used(64). The steps in this framework are: 1) define the research question, 2) identify relevant studies, 3) select studies, 4) chart the data, and 5) collate, summarize, and report results(64).

3.1.1 Search strategy

The scoping review was conducted using the approach outlined in the *Cochrane Handbook for Systematic Reviewers* and reported according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) criteria(66). Specifically, methods to define inclusion and exclusion criteria, develop the data extraction methods, and create the PRISMA flowchart were adopted. A librarian was consulted to assist with the development of the search strategy, search terms, citation indices to review, and inclusion and exclusion criteria.

The search strategy included the concepts of 1) data quality, 2) NLP, and 3) data preprocessing. These concepts were chosen because: 1) preliminary searches on literature demonstrated little outcomes on specific topics such as a) data quality indicators for text and b) measuring the effects of preprocessing methods on the outcomes of NLP methods. 2) preprocessing methods are inherently linked to data quality indicators because preprocessing methods are used to prepare UTD for analysis. The individual search terms used were:

- NLP: “natural language*”, NLP, “text mining”, “computational linguistics”, “text analy*”, “free text”
- Data Preprocessing: “pre process*”, “preproces*”, “Text processing”, “Preprocessing phase”, “Text preprocessing”, “Data processing methods”, “Pre-processing step”, “Pre-processing method”, “Preprocessing techniques”, “Preprocessing phase”
- Data Quality: “data qual*”, “data accur*”, “quality control”, “usability”, “timeliness”

Conditional statements were used to combine the search terms (e.g., "natural language*" OR NLP OR "text mining" OR "computational linguistics" OR "text analy*" OR "free text"). The search strategy process with the associated search terms is outlined in Figure 2. Ford et al’s systematic review on extracting text from EMRs to improve case detection informed the search terms used for NLP(63). Hinds et al’s review informed the search terms used for data quality(27). Lastly, the preprocessing terms used in the search strategy were acquired from keyword tags or words related to “preprocessing” in abstracts from associated literature, then a librarian assisted in identifying different synonyms(43,67,68).

The review included empirical research articles and review articles and was conducted over two time periods. An initial search was executed with an unrestricted minimum date criterion, with an end date of April 15 2020. An updated search was conducted that had an end date of May 15 2021. The databases searched were Scopus, Web of Science, EBSCOhost and ProQuest. Furthermore, we searched the reference sections of the selected articles to identify additional articles.

3.1.2 Inclusion and exclusion criteria

An article was selected for inclusion if it met one or both of the following criteria:

1) It described research about preprocessing methods for UTD, or it described UTD quality measures or methods, or it was a review article that discussed preprocessing methods for UTD. Examples of preprocessing methods include stop word removal, word stemming, lemmatization, removal of punctuation, text segmentation, lowercase conversion, correction of spelling errors, expansion of abbreviations, expansion of contractions, and word normalization. UTD quality methods include NLP methods and text mining methods, or

2) It was about methods or processes to create a gold standard (or reference) dataset to validate UTD. Examples include text annotation, part of speech tagging, and word sense disambiguation.

An article was excluded if it met one or more of the following criteria:

1) It was about methods for sentiment analysis, ontologies, semantic models, geo-spatial analysis, or qualitative research (e.g., methods to analyze interview or focus group data), or
2) It was about methods to construct lexicon databases, dictionaries, or language databases, or
3) It was about the creation of software programs or proprietary solutions for text analysis. Examples include software programs for cleaning or preprocessing UTD, conducting analyses of UTD, and implementing supervised or unsupervised machine learning methods.
4) It was not printed in the English language.
5) It was not a scholarly article from the ProQuest database only.

3.1.3 Title and abstract screening and full text screening

Title and abstract screening were conducted for all articles identified through the implementation of the search strategy, after duplicates were removed. Initial training was undertaken for the title and abstract screening. The initial training was completed by the author and supervisor on a 10% sample of all articles identified from the application of the initial search strategy (n = 1082), to ensure consistency when applying the inclusion and exclusion criteria. Another subset of training was completed by the author and supervisor on another 10% of the articles selected for title and abstract screening; this was used to calculate percent agree. Percent agreement was calculated by counting all articles the author and supervisor agreed upon and dividing by the total number of articles reviewed then multiplying by 100. A 95% confidence interval (CI) based on the binomial distribution was subsequently calculated for percent agreement. Rayyan, a web application for systematic and scoping reviews, was used to manage and organize articles through the process of title and abstract screening(69). The minimum acceptable agreement for the title and abstract screening process was 80%. Differences of opinion on the title and abstract screening were resolved by consensus. Full text screening of all articles selected after title and abstract screening was conducted next, to identify the articles to retain in the scoping review.

3.1.4 Data extraction and analysis

Articles that met the inclusion criteria proceeded to data extraction. The data extraction form was created using Google Forms (see Appendix A). The following types of information were extracted from each article: 1) characteristics of the article, 2) characteristics of the data, 3) characteristics of preprocessing methods to restructure or reorganize the UTD, 4) characteristics of supervised and unsupervised techniques to validate UTD, 5) characteristics of software used for UTD processing. The systematic review conducted by Hinds et al(27) was used to inform the types of information extracted from the articles, such as the characteristics of articles and validation methods. The reviews conducted by Spasic et al and Kiefer provided guidance on the characteristics of text data and the preprocessing methods that were included in the data extraction form(14,40,44). Lastly, an inspection of the topics for data quality indicators for UTD were explored. Articles that used EMR data were focused on. The indicator topic, methods used to define the indicator, strengths and limitations of the indicator, and the use case were described.

Data extraction training was completed by the author and advisor on a 10% sample of the articles selected for full data extraction. A data dictionary was created to ensure the consistency of the data extraction methods; it is included in Appendix B. Percent agreement was used to calculate the overall agreement between the reviewers. A 95% CI based on the binomial distribution was calculated for percent agreement. The number of variables on which there was agreement was divided by the total number of variables and multiplied by 100. Variables with open-ended responses were excluded from the calculation of percent agreement; any differences in agreement were settled with a consensus process. The minimum acceptable agreement was 80% for the entire set of extracted elements. Elements for which there was a lack of agreement went through a process of consensus to discuss the rationale for decision making. Data from Google Forms were exported to a *.csv format and were described using frequencies and percentages. R software was used to analyze the scoping review data.

3.2 Case study of preprocessing methods applied to clinical notes in EMR data

The purpose of the case study was to apply and describe indicators for measuring the preprocessing methods that were identified from the scoping review and previous research to UTD from clinical encounter notes in EMRs. Previous research has informed that applying preprocessing methods address quality issues (due to reducing dimensionality of UTD(37,39)) before applying text-based modelling. These preprocessing methods were found in previous literature (e.g., removal of stop words, removal of punctuation) and were also informed by the scoping review. Therefore, by measuring preprocessing methods, this can act as several indicators that inform us of the extent of variability of both words and the clinical note itself within UTD in EMRs.

The preprocessing methods that were applied in the case study included checking for stop words, punctuation symbols, special characters, spelling errors, and counting character (string) length, and word length(14,35). Lexical diversity was also measured(44). The steps in the case study were: (a) convert the clinical encounter notes into a data frame, (b) conduct an exploratory data analysis (EDA), and (c) apply preprocessing methods to the clinical encounter notes.

3.2.1 Data source

Data for objective 2 were clinical encounter notes found in EMRs from MaPCReN, which is part of the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). CPCSSN is a pan-Canadian multi-disease surveillance network where patient information is collected from participating primary care clinic EMR systems(70,71). The goals of CPCSSN are to utilize EMR data for disease surveillance, quality improvement, and facilitate primary care research(70,71).

The study data were extracted in 2020. The clinical encounter notes were created between 1985 and 2020. MaPCReN text data has been previously de-identified, which means that identifying or potentially-identifying information has been removed(72). The variables included in the clinical encounter notes were the note ID, the ID of the clinic where the data was extracted, MaPCReN's ID, the patient ID, the title of the note, the date the note was created, and the encounter note text data. A text file was provided by MaPCReN in tab delimited format. The file size was approximately 1 GB.

3.2.2 Preparing the UTD for a structured data format

To ensure computational feasibility of the analysis, a sample of the first 10% (2,127,514) of the lines in the text file was selected for the case study. The data were initially loaded as a vector and all string values and variables were converted to lowercase because upper case characters take precedence over lowercase characters when strings are placed in sequential order(73). Subsequently, the strings were subsetted to remove any spaces as carriage returns (i.e., “”). This reduced the total lines of the text file to 1,818,602 lines. Regular expressions were utilized to detect a new value for the note ID and to identify the date when the note was created. The application of these regular expressions to the text data were used to determine where the text for a single clinical encounter note began and ended. A structured data frame was then created from the cleaned vector. The final dataset used in the study was comprised of 193,206 clinical encounter notes. As mentioned by Taleb et al, by structuring data that are initially unstructured this allows for ease of data quality inquiry(41).

To apply methods to measure characteristics of UTD such as lexical diversity, a document term matrix (DTM) was created. A DTM is text data that are converted to matrix form. The rows correspond to documents (e.g., individual clinical encounter notes) and each column represents one term. The DTM facilitates measurement of term frequency.

3.2.3 Exploratory data analysis on characteristics of EMR notes

EDA was used to describe the characteristics of the EMR clinical encounter notes. EDA is an important preliminary step in understanding the overall dataset which is inherently tied to the overall quality of the data(45,74). While EDA does not have specific steps, EDA is used to generate questions about data (i.e., hypothesis generation), check for data quality (e.g., missing data, violation of statistical assumptions, and distributions of variables), create data visualizations such as figures or tables (if needed), and lastly, identify models that might be applied to the data (45,74).

In the EDA for the EMR clinical encounter notes, frequencies, measures of central tendency (i.e., mean and median), and measures of variability (standard deviation (SD), interquartile range (IQR) and range) were used to describe the unprocessed data. The number of unique clinical encounter notes, patients, and clinic sites were described. The number of clinical encounter notes per decade (based on the date the note was created) was described. The number of clinical encounter notes with missing text was described by counting the number of notes that had no text within the free text field.

3.2.4 Application of preprocessing methods for UTD quality assessment

After the EDA, preprocessing methods informed by the scoping review and other relevant literature were applied to the EMR clinical encounter notes(14,35,44,75). This analysis focused on common preprocessing steps identified from the scoping review. First, counts of stop words, punctuation, special characters, word length, character length, and spelling errors, as well as lexical diversity scores were produced for each note(14,24,35,44,75). Stop words were generated from a stop word dictionary found in the tidytext package in R(17). All the stop words were converted into regular expressions patterns to detect each instance of a stop word in each note. There is no dictionary for punctuation symbols; these were detected using regular expressions. Special characters were detected using regular expressions from characters found in the character map application in Windows 10. Spelling errors were identified used the hunspell package in R. The spell-checking dictionaries that hunspell uses are based on dictionary files typically found on an operating system (e.g., using the file names: "en_US.aff" and "en_US.dic")(76). Lexical diversity was calculated for each note using the type token ratio from the quanteda and quanteda.textstats package in R(77). To measure lexical diversity, the note field was first tokenized (i.e., separation of words or a sequence of letters separated by a space into a matrix's columns or rows), with punctuation and stop words removed. Then the tokenized dataset was converted into a DTM. Counting character (i.e., string) length per note involved utilizing a string length function. To produce word length (i.e., words per note), punctuation was removed and then words were tokenized and counted for each unique record id. Retaining punctuation when counting words would have inflated the estimate of the number of words when tokenized. Measures of central tendency and variability for stop words, punctuation, special characters, spelling errors, character length, word length, and lexical diversity were calculated.

Guided by the MCHP data quality framework(15), temporal stability of a number of measures was assessed using a time series analysis based on the date when the note was created. The time series analysis was applied to the annual mean values for stop words, punctuation symbols, special characters, spelling errors, character length, word length, and lexical diversity. Lastly, 95% CIs were produced for each of the preprocessing methods.

Stop words, punctuation symbols, and special characters were chosen for analysis based on scoping review articles that utilized EMR data(20,24,35,42,78). Spelling errors and lexical diversity are measures based from previous literature(14,44). Character length and word length were chosen based on a scoping review article on short text data that analyzes outlier information(75). Lastly, following MCHP privacy protocol on suppression of small values, any analyses that reported n = 6 or fewer entities per subgroup was filtered from the data. Accordingly, any note that was created prior to 1999 was removed.

The same preprocessing methods utilized for temporal stability were also used for each clinic site. Mean numbers of stop words, punctuation symbols, special characters, word length, character length, and lexical diversity were calculated for each site along with their 95% CIs. Site 25 was removed from the analyses because it had fewer than six clinical encounter notes. Furthermore, since clinic sites other than 2 and 7 had small sample sizes (i.e., the majority of clinic sites had fewer than 50 notes), they were excluded from the main analysis, but included in Appendix D.

A frequency analysis was conducted for stop words, punctuation symbols, and special characters. For spelling errors, an inspection of the top 20 words that were spelled incorrectly was explored to determine what the dictionary deemed as incorrect or correct spelling. Regarding word length, notes containing 32 words (first quartile) and less were inspected. For character length, a similar approach was taken for the first quartile of character length, where notes containing 216 characters or less were inspected. Frequencies and percentages for both were taken for each word length group or character length group (e.g., count of notes with 1 to 10 words or 1 to 20 characters) up to the first quartile. The first quartile of words and characters per note was chosen for inspection because the articles in the scoping review cited potential text quality issues when notes contain few words(35,75,79).

Chapter 4 – Results

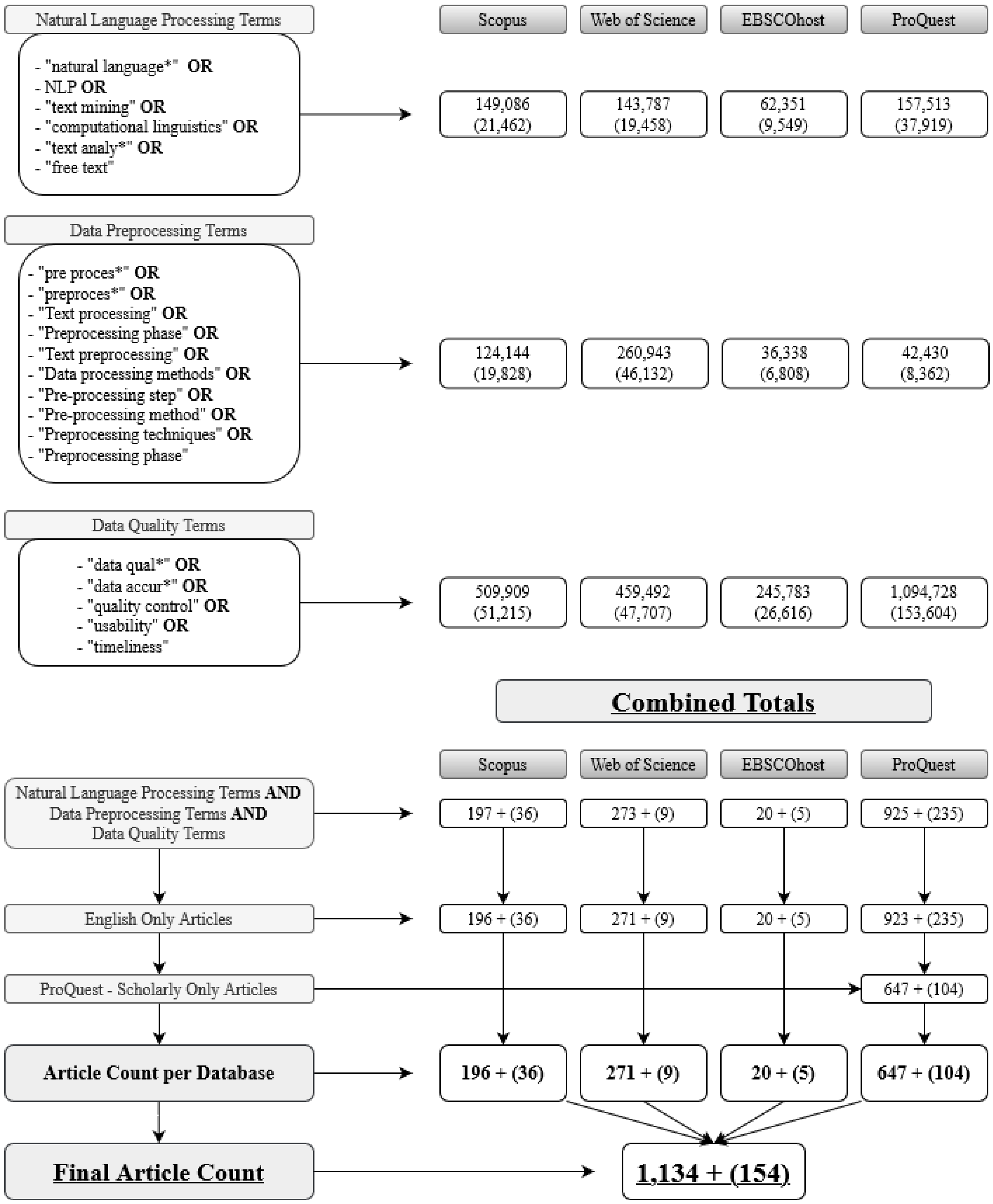
4.1 Scoping review

For objective 1, the scoping review was conducted using the following steps: 1) apply the search strategy; 2) apply the inclusion and exclusion criteria for title and abstract screening and full text screening; 3) extract information from the articles; and 4) describe and visualize the characteristics of the articles.

4.1.1 Search strategy and screening results

The initial search (Figure 2, numbers not in parentheses), which encompassed the period up to April 15, 2020, identified 197 articles from Scopus, 273 articles from Web of Science, 20 articles from EBSCOhost, and 925 articles from ProQuest.

Figure 2 - Scoping review search strategy results



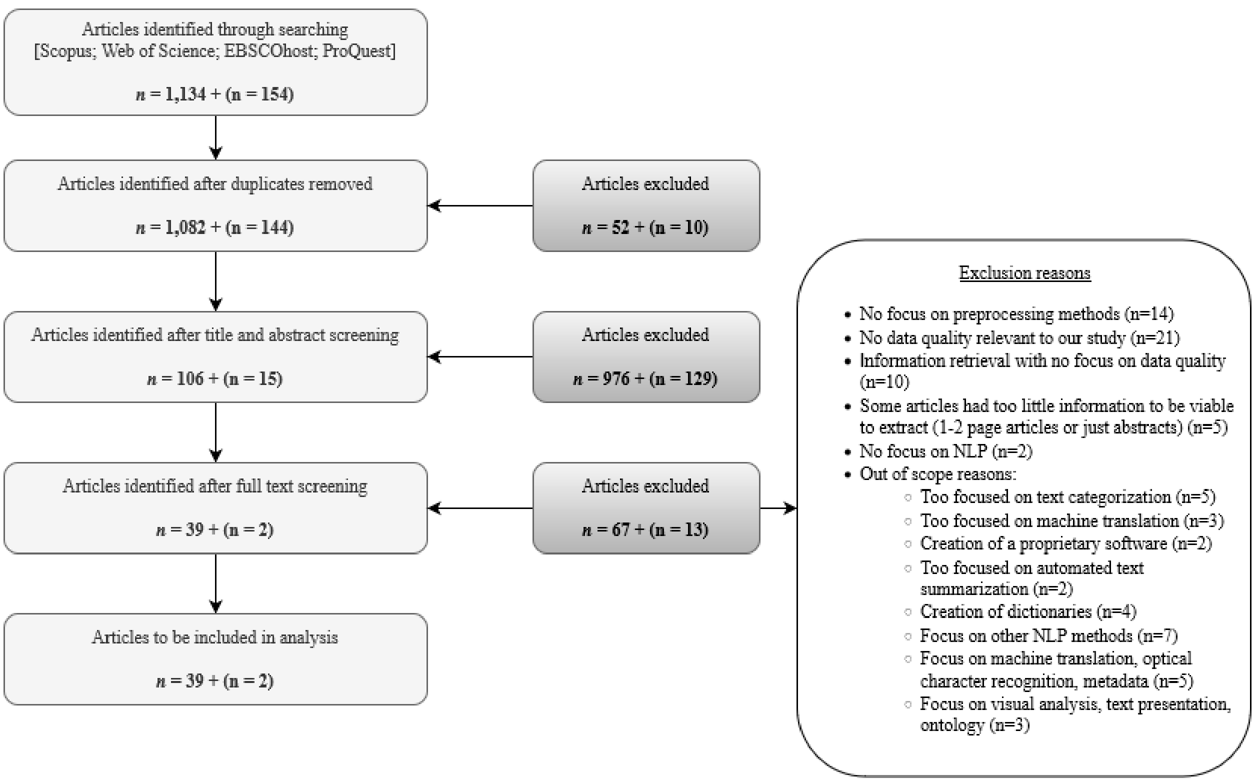
Note: Numbers not in parentheses are for the search completed in April 2021; number in parentheses are for the search completed in May 2021

After removal of duplicates, the search yielded 196 articles from Scopus, 271 articles from Web of Science, 20 articles from EBSCOhost, and 647 articles from ProQuest. Thus, a total of 1,134 articles was retrieved from the initial search. There were 52 duplicate articles that were deleted; this reduced the total to 1,082 articles for title and abstract screening.

The search was updated on May 15, 2021 (Figure 2, numbers in parentheses). An additional 154 articles were retrieved from the search; 10 duplicates were removed. Accordingly, 144 articles were selected for title and abstract screening. In summary, a total of 1,226 articles were retrieved from the search.

Initial training for title and abstract screening yielded 83.3% (95% CI: 75.2%, 89.2%) agreement on the 10% sample of articles reviewed by the author and their supervisor during the initial search. The second training by the author and their supervisor, yielded 86.1% (95% CI: 78.3%, 91.4%) agreement. Figure 3 provides a PRISMA flow chart that illustrates the application of the inclusion and exclusion criteria on both title and abstract screening and full text screening.

Figure 3 - Scoping review PRISMA flowchart



Note: Numbers not in parentheses are for the search completed in April 2021; number in parentheses are for the search completed in May 2021

After full text screening was completed, a total of 41 articles remained. These articles and their titles are reported in Appendix C. Issues that arose in the application of the inclusion and exclusion criteria were resolved through consensus between the author and their supervisor. Reasons for article exclusion were: 1) article did not focus on preprocessing methods, 2) no data quality relevant to our research (e.g., improving library indexing strategies), 3) focus on information retrieval rather than data quality, 4) abstracts only, and 5) no focus on NLP. Articles were out of scope because: 1) focus on text categorization, 2) automated summaries – articles about the use of NLP methods to produce automatic summaries of long text, and 3) creation of lexicons – articles about creation of language dictionaries or large word databases. However, note that we included four articles where a language treebank was created because the article focused on data quality elements such as quality of annotations(56,80–82), 4) the focus was on other NLP methods such as sentiment analysis, 5) optimization of techniques – articles in which the authors developed a proprietary framework or using supervised learning methods to enhance annotations, 6) machine translation – articles about technological enhancements such as optical character recognition or automated transcription of voice data, and 7) the focus was on visual analysis, or text presentation methods, which were unrelated to the focus of the scoping review.

4.1.2 Data extraction results

Four of 39 articles from the initial search were selected for the calculation of agreement for the full text extraction; the overall agreement was 94.6% (95% CI: 88.9%, 97.4%).

Table 1 summarizes the characteristics of the articles selected for full data extraction. Overall, 90.2% of the articles reported the results of empirical research and another 7.3% (n = 3) were review articles. Only a single article was classified as a case study. More than half (51.2%) of the articles were published between 2016 and 2021. No articles were published prior to 2002, even though the search strategy was implemented with unrestricted date criteria. In terms of disciplinary area, 61% of articles were deemed to be from computer science and engineering disciplines, while 39% were from the health sciences, social sciences, humanities, and business. Articles that were in a particular disciplinary area did not necessarily use data from their disciplinary area. For example Gero and Ho’s article used biomedical data(83) but the discipline of the research was computer science.

A variety of types of text data were represented in the 31 empirical articles, 3 review articles, and 1 case study. This included EMRs (i.e., clinical notes, progress notes, patient safety records(35,42,54,78,84–87)), lexical documents (i.e., language treebanks, WordNet database(56,80,88–92)), organizational documents (i.e., maintenance logs/data, accident reports, requirements documentation(93–96)), abstracts and scientific articles (i.e., PubMed and various engineering journals(55,83,97,98)), various corpora (i.e., non-language corpora, nonmedical/medical/biomedical corpora, language corpus (82,98–100)), social media data (i.e., Twitter, meme tracker from various social media websites(24,75,101)), product reviews (i.e., general product, Chinese tourism, Amazon product(20,102,103)), and news articles (i.e., magazines, newswires, consumer reports(81,101,104)).

The case study focused on data quality elements to assess statistical text mining algorithms for clinical progress notes in EMRs(35). Specifically, this article discussed approaches for text analysis and the sources of noise that could affect the outcome of algorithms for identifying fall-related injuries in clinical progress notes. This case study used preprocessing methods, including removal of stop words and lowercase conversion, to prepare the clinical progress notes for analysis. Manual annotation of the clinical progress notes was used to enhance the quality of the data for application of an NLP algorithm.

Regarding the three review articles, one article used preprocessing methods on clinical notes to reduce word variability; the second article was about preprocessing methods to enhance exploratory methods for hypothesis generation(42,101). The third article discussed preprocessing methods to improve sentiment analysis(20). All three review articles were literature reviews that examined a process method framework for UTD preprocessing. For example, the Assale et al article presented a process to measure for error prevention and error correction with an emphasis on disambiguation of abbreviations in medical text data(42). Assale et al demonstrated error correction using rule-based methods such as Damerau-Levenshtein (also known as: Levenshtein-Damerau) to preprocess UTD(42). Allen et al presented a preprocessing method to improve word count quality in exploratory analyses(101). This preprocessing method consisted of steps that utilize tokenization, removal of punctuation and symbols, lowercase conversion, removal of stop words, stemming words and appending titles to words (if applicable)(101). After these methods were applied, sentiment analyses and topic modelling were conducted.

It should be noted that Assale et al's review referenced Berndt et al's article in how to operationally define data quality(35,42). None of the information extracted between these two articles were redundant. Assale et al's review extracted information about word variability (i.e., spelling errors), while Berndt et al's article extracted information about noisy text (i.e., preprocessing methods addressing noisy text)(35,42).

Table 1. Characteristics of the scoping review articles (n=41)

	n	%
Article type		
Empirical research	37	90.2
Review article	3	7.3
Case study	1	2.4
Publication year		
2002 - 2009	7	17.1
2010 - 2015	13	31.7
2016 - 2021	21	51.2
Disciplinary area		
Computer Sciences and Engineering	25	61.0
Health Sciences	8	19.5
Social Sciences, Humanities, Business	8	19.5
Type of data*		
Electronic Medical Records	8	19.5
Lexical (e.g., language treebanks)	8	19.5
Organizational Documents	6	14.6
Product Reviews	4	9.8
News Articles	4	9.8
Corpora (e.g., biomedical text of gene entities)	4	9.8
Abstracts and Articles from Scientific Journals	4	9.8
Social Media	3	7.3
Administrative	1	2.4
Other	5	12.2

* Categories are not mutually exclusive

The last review article, which was by Gharatkar et al, also presented a pipeline for preprocessing of UTD. This pipeline included tokenizing the data and removing special characters, URLs, hashtags, slangs, and stop words. Spell checking and stemming were also implemented. Sentiment analysis was performed after the preprocessing methods were applied to the data(20).

With respect to the empirical articles, almost all (85.4%) of these articles described preprocessing methods to improve NLP algorithm performance. However, one article(24) offered an empirical approach to compare the effect of a proposed preprocessing methodology to a baseline preprocessing method on short text similarity measures; results showed that the proposed method outperformed the baseline method(24). Other empirical articles discussed methods for manual or automated annotation. For example, Westpfahl et al discussed the creation of a gold standard corpus for teaching and research about spoken German(82). In this article, the corpus was manually annotated with part of speech tagging (i.e., identifying and annotating words that are nouns, adverbs, verbs, and other parts of speech) and lemmatization, a form of word stemming, where the morphological base of a word is returned(82).

In terms of data size and terms used to describe data size, the UTD described in the articles were characterized in many ways (Table 2). Documents and words were mentioned most frequently (e.g., citing how many documents were used or how many words were found in a type of UTD). More than half (53.7%) of the articles mentioned the volume of documents, and 46.3% of the articles mentioned the count of words in each document. Elements of size often used to describe the size of structured data such as “how many rows (records)” or “how many columns (features)” were among the least common ways to describe UTD size.

Table 2. Characteristics of text data and quality assessment in the scoping review articles (n = 41)

	n	%
Measures used to describe data size*		
Documents	22	53.7
Words	19	46.3
Phrases (sentences)	8	19.5
Rows (records)	6	14.6
Features	4	9.8
Data annotation*		
Manual	20	48.8
Automated	17	41.5
No annotation	15	36.6
Data availability		
Publicly available	14	34.1
Not publicly available	27	68.9
Preprocessing methods*		
Reorganizing or restructuring methods	35	85.4
Internal validation	23	56.1
External validation	11	26.8
Other	4	9.8
Data quality dimensions*		
Accuracy	28	68.3
Relevance	14	34.1
Comparability	13	31.7
Usability	7	17.1
Completeness	7	17.1
Validity	7	17.1
Readability	6	14.6
Accessibility	3	7.3
Timeliness	3	7.3

*Categories are not mutually exclusive

Overall, over one third (36.6%) of the articles did not employ text annotation, while 48.8% of the articles discussed manual annotation (e.g., an experienced coder assigning codes using ICD-9-CM(105)), and 41.5% of articles discussed automated annotation (e.g., words were automatically annotated for part of speech analyses, grammar tagging, and assistance with manual annotation processes(24,80,82,83,88,100,106,107)).

Almost all (85.4%) of the articles discussed using restructuring and reorganizing methods to prepare UTD for analysis (e.g., removing stop words, punctuation, removing URLs). Figure 4 describes the types of restructuring and reorganizing methods that were used for all articles, while Figure 5 shows the same information for the subset of articles from health science disciplines.

Figure 4 - Restructuring and reorganizing methods used in the scoping review articles

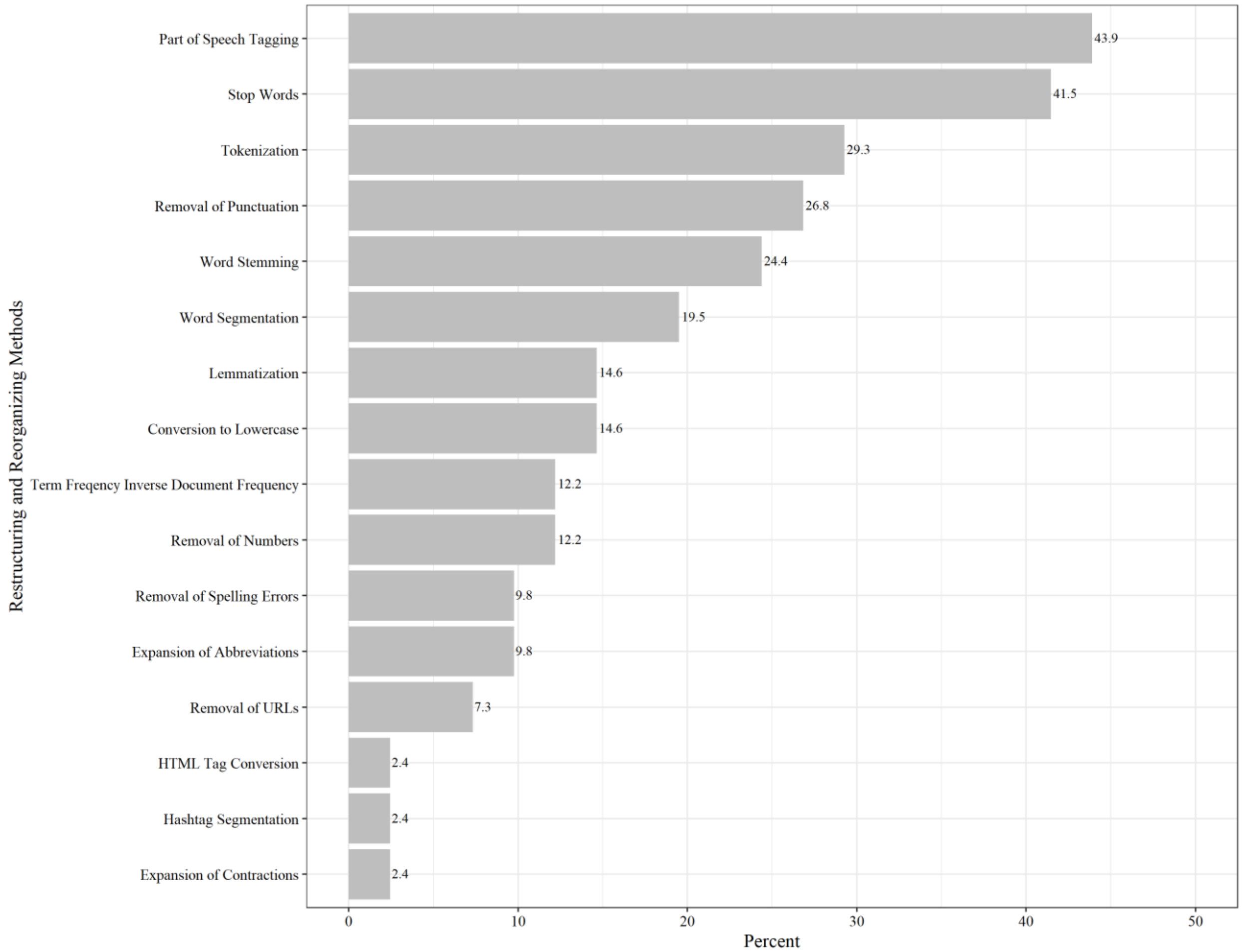
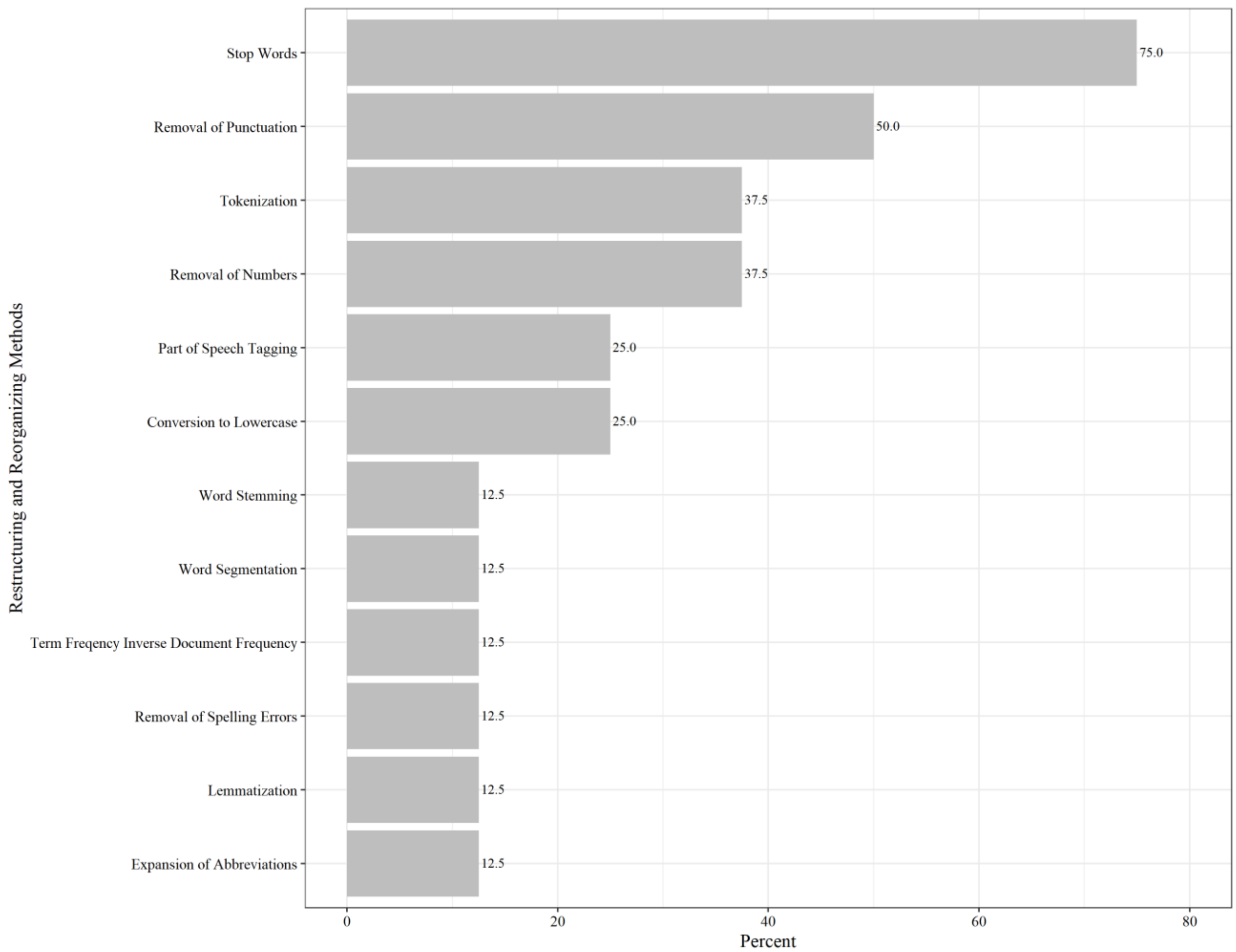


Figure 5 - Restructuring and reorganizing methods used in scoping review articles from the health sciences



The types of restructuring and reorganizing methods for both Figures 4 and 5 were common in 4 of the top 5 methods. These common restructuring and reorganizing methods are (in no order): parts of speech tagging, stop word removal, punctuation removal, and tokenization. The different restructuring and reorganizing methods among the top 5 methods for Figures 4 and 5 are: word stemming occurred more in all articles for the scoping review, whereas the removal of numbers was more common in health science articles.

We identified the articles that explicitly mentioned a data quality dimension; the three data quality dimensions that were most frequently mentioned were: accuracy (68.3%), relevance (34.1%), and comparability (31.7%). Furthermore, "data quality" or "quality" as terms were described or referenced in several ways among the 41 articles. Several articles discussed quality either from the perspective of data quality (or information quality), or using terminology from data or information quality dimensions (e.g., accuracy, correctness, interpretability)(75,86,96,102,103). Other articles discussed enhancing data quality by focusing on either utilizing or improving preprocessing methods(24,42,75,84,85,87–89,91,95,98,100,101,108,109). Several of these articles only mention "data quality" in passing, since the main focus of these articles were the preprocessing methods. Other articles discussed quality from an "annotation" perspective where by tagging documents or words will enhance the quality of algorithms either through the creation of treebanks or manually annotating using expert opinion(35,54–56,78,80–82,96,97,99,103,110,111). Lastly, some articles discussed quality through utilizing preprocessing methods that were primarily focused on achieving a more "accurate or relevant" outcome for algorithmic model performance (20,79,83,90,92–94,102,104,106,112). Overall, while articles mentioned terminology such as "accuracy" or "relevance", data quality as a concept (or its individual dimensions), was referenced or described in multiple ways. Articles that focused on the data quality dimensions (or measuring them) were rare.

The software used to process data was mostly user-defined software (54.0% of articles). One of the user-defined software tools that was used was Stanford CoreNLP, a Java-based program that supports six different languages. Stanford CoreNLP was used to annotate text using different preprocessing methods such as parts of speech tagging, NER, tokenization, determining sentence boundaries (e.g., text is divided into sentences using rule based methods), numeric entities and time values (e.g., numbered addresses and dates are annotated)(113). Stanford POS Tagger(114) was another common software reported in the articles. It is described as a more in-depth part of speech tagger than the Stanford CoreNLP program. Four articles explicitly mentioned the use of Python software, including spaCy, NLTK, and scikit-learn. SpaCy and NLTK are NLP libraries while scikit-learn is a machine learning library(115–117). Two articles explicitly mentioned the R packages tm and caret. The former is used for text mining in R(118), while caret is a predictive modelling package for classification and regression analysis(119).

Overall, many of the articles described an improvement of algorithm outcome performance through a variety of preprocessing techniques to enhance UTD quality. However, there were no articles that reported on indicators of UTD quality before preprocessing. While the data quality dimensions were not mapped to specific preprocessing methods, "accuracy" was used for evaluating outcomes in a confusion matrix and used as a descriptor for utilising preprocessing methods to enable unsupervised and supervised algorithms outcomes more accurate. That is, all preprocessing methods were closely tied to the data quality dimension of accuracy.

4.1.3 UTD quality indicator topics in EMR data

Table 3 provides a detailed assessment of data quality indicator topics for UTD in articles that focused on EMR data. Seven data quality indicator topics were discussed. Two articles reported on the assessment of misspelled words. In the first article, spelling errors were checked in both clinical terms and general terms and were applied to three different types of clinical UTD: 1) clinical notes from primary care clinics, 2) free-text allergy entries, and 3) free-text medication orders. The methods utilized were: 1) dictionary based - where a word was checked against a dictionary (e.g., string matching), 2) NER, which was utilized to detect person names to avoid classifying names as spelling errors, and 3) regular expressions, which were used to detect string patterns of emails, URLs, and commonly misspelled words that were not in the dictionary. The strength of the method is it utilizes multiple approaches to detect spelling errors. However, the authors did not investigate whether spelling errors or their corrections impacted the quality of the data.

In the second article, spelling errors were corrected by utilizing a three-module approach: 1) a string-to-string edit distance known as the Levenshtein-Damerau distance, 2) syntactic correction (i.e., grammatical correction) to correct word order provided by data in the first module, and 3) word sense disambiguation for words with similar parts of speech (e.g., distinguishing between two similar words that are classified as verbs).

The second indicator topic was security, which was reported in one article by Pantazos et al(86). The focus was on addressing four different criteria to assess the quality of de-identified UTD: 1) medical correctness – each health record must provide an accurate picture of a patient (while still being de-identified), 2) anonymity –so that it is not possible to identify a patient, 3) readability – each record must represent reality where patient names and addresses look real, however have de-identified information, and 4) consistency – where existing identifiers are replaced by new de-identified information must be consistently false (e.g., a real patient name "John Doe" has been de-identified as "Jack Tie" and the entire clinical note history must be consistent with the de-identified name "Jack Tie"). For each of these methods, a mapping table (i.e., a table that maps identifiable information with de-identifiable information), was utilized to replace existing identifiers. For example, an address “123 Somewhere Street” is identified by the mapping table and replaced with a de-identified address “456 Elsewhere Avenue”. Ambiguous medical words (e.g., Parkinson’s disease versus the surname Parkinson, or different spellings of the word “diabetes”) were deleted if they occurred less than 200 times in the EMR database. The strength of this method is that it aims to mimic a clinical note that contains identifiable information using de-identified data. However, there were some limitations of this method; spelling errors were not addressed, pharmaceutical names were overlooked, and several clinical abbreviations were mistaken as ambiguous names. Furthermore, because zip codes were scrambled so that specific locations could not be identified, geographical analysis was not possible.

The third indicator topic, which was reported in one article by Assale et al, was about reducing word variability in typographical corrections, where typographical errors in one word can create variability in how the word is spelled (e.g., the word “hello” could be misspelled as “helo”, “hallo”, and “ello” and thus create variants of the word “hello”)(42). The methods used were multi-faceted: 1) preprocessing methods to restructure and reorganize UTD, including tokenization, and removal of stop words, numbers, and non-ASCII symbols, 2) counting word frequency (and percentage) to identify potential typographical errors (i.e., any word that was commonly spelled that occured above 80% of the area under the power law curve was not considered a typographical error), 3) the Levenshtein distance metric was used to search for 80% of the most frequently spelled words and compare them to the other 20% that were considered typological errors, 4) using the Damerau-Levenshtein distance for words that contained letters that were switched around (e.g., consc[ie]nce versus consc[ei]ence), 5) ambiguous associations from the Damerau-Levenshtein distance metric were discarded, and 6) replaced multi-associated words with the most frequent word (e.g., “helo”, “hallo”, and “ello” were replaced with “hello).

The fourth indicator topic was sources of noise. Sources of noise in text primarily involves preprocessing methods that restructure and reorganize UTD. Some of these methods are: 1) Converting uppercase letters to lowercase letters. 2) Tokenization (separation of words (or a sequence of letters separated by a space) unto a matrix’s columns or rows). 3) Removal of tokens with less than three characters or tokens that do not contain alphabetical characters. 4) Normalizing terms (correcting spelling errors or expansion of abbreviations). 5) Removal of stop words. 6) Removal of tokens that have a frequency of one. The strength is that these preprocessing methods are relatively straightforward to apply and are conceptually easy to grasp (e.g., removal of punctuation is easy to implement).

The fifth indicator topic was the quality of annotations; two articles focused on this topic. In the first article, two subject matter experts manually annotated clinical notes about fall-related injuries. Each document was classified as “fall” or “not fall”. Spot checks by a clinical expert were performed on 10 documents in every 1000 annotations. Inter-rater reliability was evaluated using Cohen’s Kappa scoring metric. Cohen’s Kappa score between annotators was 0.90 (on n = 275 annotations), which was considered to represent high reliability. The strength of this method is that it uses clinical expertise to determine fall-related injuries within each clinical note. Supervised methods that involved classification algorithms highly benefit from the quality of annotations from expertise. A limitation of conducting manual annotations involve cost in paying annotators for their expertise. At the same time, manual annotation takes a significant amount of time and commitment to ensure quality of annotations (e.g., in this article it required spot checks and time needed from a third clinician expert). Another limitation is that it may be unrealistic to manually annotation a large sample of documents. In the second article, restructuring and reorganizing methods were used in combination with tagging parts of speech, NER, and lastly tagging relationships among named entities that were previously annotated by NER. These tags were applied by annotators with backgrounds in computational linguistics, while resident physicians were solving any disagreements between annotators. Inter-annotator agreement was calculated to evaluate the quality of annotations. These annotation guidelines were proposed to provide a method of ensuring high annotation quality for Chinese clinical texts. A strength of this method is that the authors have inherently built a concept of data quality within their annotation guidelines to build a corpus by utilizing annotation quality assurance measures. A limitation was that the clinical text used to build the corpus only covered two departments containing a limited amount of medical terminology. However, results showed that the annotation guidelines were of high quality, as measured by inter-annotator agreement.

Ambiguous abbreviations was the sixth indicator topic that was addressed; it was the focus of one article. In this article, the authors utilized deep learning methods (i.e., convolutional neural networks (CNN)); no feature engineering or preprocessing was required. The CNN was trained on word embeddings, which are representations of words in a vector. These word embeddings were extracted from journal articles found in PubMed. The CNN was also trained on parts of speech tags and other covariates such as author name. A major strength of CNNs within deep learning methods, is that they automate feature engineering and preprocessing methods. Thus, CNNs allow for overall reduction in workload. The limitation is that the effects of disambiguating the abbreviations were not tested to assess whether this method impacted the quality of the text.

The last indicator topic was about the reduction of manual workload when annotating documents. The article that focused on this topic used a K-nearest neighbour classifier, which is a supervised learning method, to predict the classification of documents between "diagnostic errors" or "device related complications". Several preprocessing methods were applied to the data; specifically, punctuation, stop words, and white space were removed. All character values were set to lowercase. The document was then converted to a DTM, and documents were annotated prior to classification. The key strength of this method was the use of a supervised learning technique to assist with document annotation, instead of relying on manual annotation. However, the limitation of this research was it used a small sample size during the training phase. Thus, generalizability can be a concern, so there is no way to tell if the trained data can be used on new notes (i.e., training dataset may not be sufficient for a test dataset).

Table 3. UTD restructuring and reorganizing data quality indicators described in scoping review articles using EMR data

Indicator Topic	Source	Methods	Strengths / Limitations	Use Case
Misspelled words	Lai et al	The authors measured spelling errors for clinical and general terms. A clinical term was a word that was not a name or was not present in the Aspells(85) default dictionary. A general term was any other word that was not a clinical term. These methods were used to identify both clinical and general terms: 1) used many dictionaries for misspelling detection 2) used Named Entity Recognition (NER) to avoid misclassification of person names as misspellings 3) used regular expressions to correct emails, URLs, and commonly misspelled words	**Strengths:** A combination of previous methods that worked really well. **Limitations:** 1) Since its a composite of previous methods so its incomparable to previous methods. 2) Authors did not attempt to investigate whether misspellings & misspelling corrections impacted on the quality of the information extracted from medical texts.	Applied to: 1) clinical notes from primary care clinics 2) free-text allergy entries 3) free-text medication orders
	Ruch et al	Methods used to improve spelling correction: module 1) string to string edit distance known as Damerau-Levenshtein module 2) syntactic correction - addresses word order and agreement problems module 3) processing words with the same parts of speech by applying contextual word sense disambiguation	**Strengths:** NLP tools such as NER and lexical disambiguation can reduce spelling correction, this can allow for automated processes. **Limitations:** Authors did not attempt to investigate whether misspellings & misspelling corrections impacted on the quality of the information extracted from medical texts.	Applied to electronic patient records

Security	Pantazos et al	The methods used for de-identification has to meet the criteria of: 1) Medical correctness - each health record must show a true medical picture of a patient 2) Anonymity - its not possible to see who the real patient is 3) Readability - the health record has to represent reality 4) Consistency - the patients identifiers has to be consistent with the entire medical picture Each method utilized a mapping table to replace existing identifiers with new identifiers. To treat ambiguity, when the de-identification program meets an ambiguous word it: 1) gets deleted if its a rare word (occurs less than 200 times); 2) gets left if it occurs more than 200 times permutation tables - that mapped existing identifiers with new ones - ensured readability and consistency distorted identifier tables - mapped existing civil registration numbers to a scrambled one - ensured readability and correctness	**Strengths:** While de-identifying EMR data, the authors tried to preserve essential components of note to mimic the original note. **Limitations:** 1) Did not address spelling errors. 2) Overlooked pharmaceutical names. 3) Missed several clinical abbreviations. 4) Cannot do analysis on geography due to scrambled zip codes. 5) For statistical purposes de-identification should not care about readability or consistency.	De-identified an existing Danish EMR database
Reducing word variability	Assale et al	1) Tokenization, removal of stop words, numbers and non-ASCII symbols 2) Counted frequency of words - words occurring above an 80% percentage were considered correct, less frequent words were checked if they were present in an Italian dictionary or a medical dictionary, whatever is left over was considered typographical. 3) Used distance metric between strings - "Levenshtein distance". Search in the 80% of most frequent words that had a "distance 1" from the typos. Distance 1 signifies that typological words differ from 1 letter insertion/deletion/substitution from the original word. 4) Also take into account "Damerau-Levenshtein distance" metric where it also takes into account inversions between letters - because its common to invert two adjacent letters (distance 2). 5) Manually inspected to verify no association errors. Ambiguous associations were discarded. 6) Multi-associated words (i.e., words with the same meaning but varied in spelling) were replaced with the most frequent one.	**Strengths:** Proposed a method for reducing word variability **Limitations:** 1) Number of false positives is high. 2) Cannot guarantee to correct all errors.	Applied to the anamnestic summaries of Endocrinology and Rheumatology
Sources of noise	Berndt et al	1) converting text to lowercase 2) tokenization 3) removal of tokens with less than three characters or no alphabetical characters 4) normalizing terms 5) removal of stop words 6) removing tokens that only occur once	**Strengths:** Authors addressed text noise by introducing preprocessing methods to reduce noise **Limitations:** The study only used one dataset	Applied to clinical progress notes

Quality of annotations	He et al	Annotation methods included: 1) word segmentation 2) part of speech tagging (with shallow and full parsing of parts of speech tags) 3) named entity tagging 4) relational tagging (i.e., finding relationships among named entities) Annotation quality was evaluated using F1 measure, precision, and recall	**Strengths:** Authors have built a concept of data quality into the creation of a corpus by assessing the quality of annotations **Limitations:** Since there was a limit of annotation resources, the corpus created only covered two departments of a hospital, thus lacking medical terminology in other departments	Applied to: 1) discharge summaries 2) clinical progress notes
	Berndt et al	Quality of Annotations include: 1) "Fall" or "Not fall" was given an operational definition 2) Manual annotation evaluated with Cohen's kappa for interrater agreement	**Strengths:** Generally manual annotations are used as a reference standard to be used for machine learning classification algorithms **Limitations:** Manual annotations are time consuming, costly, and can lead to manual error	Applied to clinical progress notes
Ambiguous abbreviations	Joopudi et al	Utilized a convolutional neural network (CNN) that: 1) Was trained on word embeddings (i.e., a representation of words in a vector) located on journal articles in PubMed 2) Was trained on parts of speech tags 3) Used clinical notes meta information such as author	**Strengths:** Using deep learning allows the user to bypass feature engineering tasks which can be time consuming, furthermore the CNN model worked well on disambiguating abbreviations **Limitations:** While the authors have shown that methods used to disambiguate abbreviations works well, there is no investigation if the corrections have had an impact on the quality of information.	Applied to: 1 & 2) de-identified longitudinal patient records with clinical notes (the difference between these two datasets is that the 2nd dataset was manually annotated) 3) publicly available dataset created by University of Minnesota
Reducing manual annotation	Liang et al	Utilized KNN classifier (supervised method) to predict the type of documents (i.e., a document is either "diagnostic errors" or "device related complications"). The process was as follows: 1) Noise Removal: removal of punctuations, words were set to lowercase, white space was removed, stop words were removed 2) Document was converted to a document term matrix 3) Documents were pre-annotated as either "diagnostic errors" or "device related complications" to create a gold standard comparison 4) Evaluated using F measure and Accuracy	**Strengths:** Demonstrated a process that enhances automatic annotation processes that include data quality elements. **Limitations:** The sample size in the study was small and this method was not demonstrated on other types of data such as EMR clinical notes.	Applied to patient safety documents from WebM&M - publicly available

4.2 Case study

For objective 2, a case study was conducted to apply and describe preprocessing methods identified from the scoping review and previous literature to UTD from clinical encounter notes in EMRs. Recall that by measuring the preprocessing methods allows to understand the extent of variability of both words and the clinical note itself within UTD in EMRs. The following steps were taken: 1) An exploratory data analysis on the characteristics of EMR notes were conducted. 2) Preprocessing methods were applied to the clinical notes in the EMR.

4.2.1 Results of exploratory data analysis

Tables 4 summarizes the results of the EDA on characteristics of 193,206 EMR clinical encounter notes. The results reported in Table 4 reveal that data were available for 6891 patients; the year in which a note was created spanned over 32 years, from 1985 to 2020, with no notes appearing in the years 1990, 1994, 1995 and 1997. Less than 0.1% of the notes had no characters or words in them (i.e., the note field was empty with a string length of 0).

Table 4. Characteristics of EMR clinical encounter notes (N= 193,206)

	n
Patients	6,891
Clinic sites	12
Time period	
1985 - 1990	11
1991 - 2000	5,806
2001 - 2010	89,708
2011 - 2020	97,681
Notes with missing text	28

Figure 6 displays the variability in years; few notes were collected electronically in the late 1980s. However, the frequency of notes dramatically increased from 1999 onward. There was a larger increase in the frequency of notes between 2008 and 2010. Figure 7 displays that the frequency of notes was concentrated in clinic sites 2 (38,141 notes) and 7 (153,208 notes), the other 10 clinic sites displayed a lower sample size per site.

Figure 6 - Frequency of notes by year of note creation

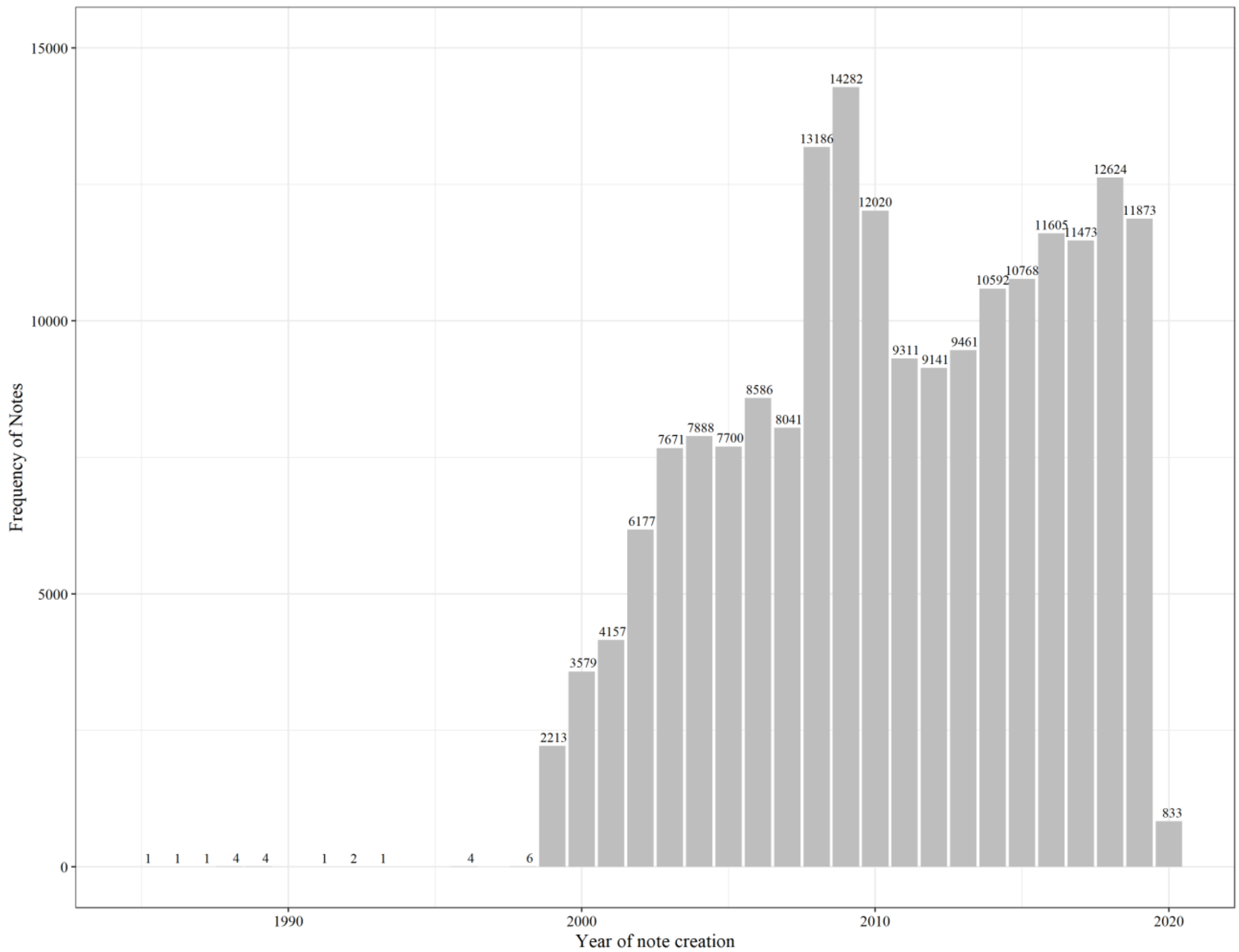
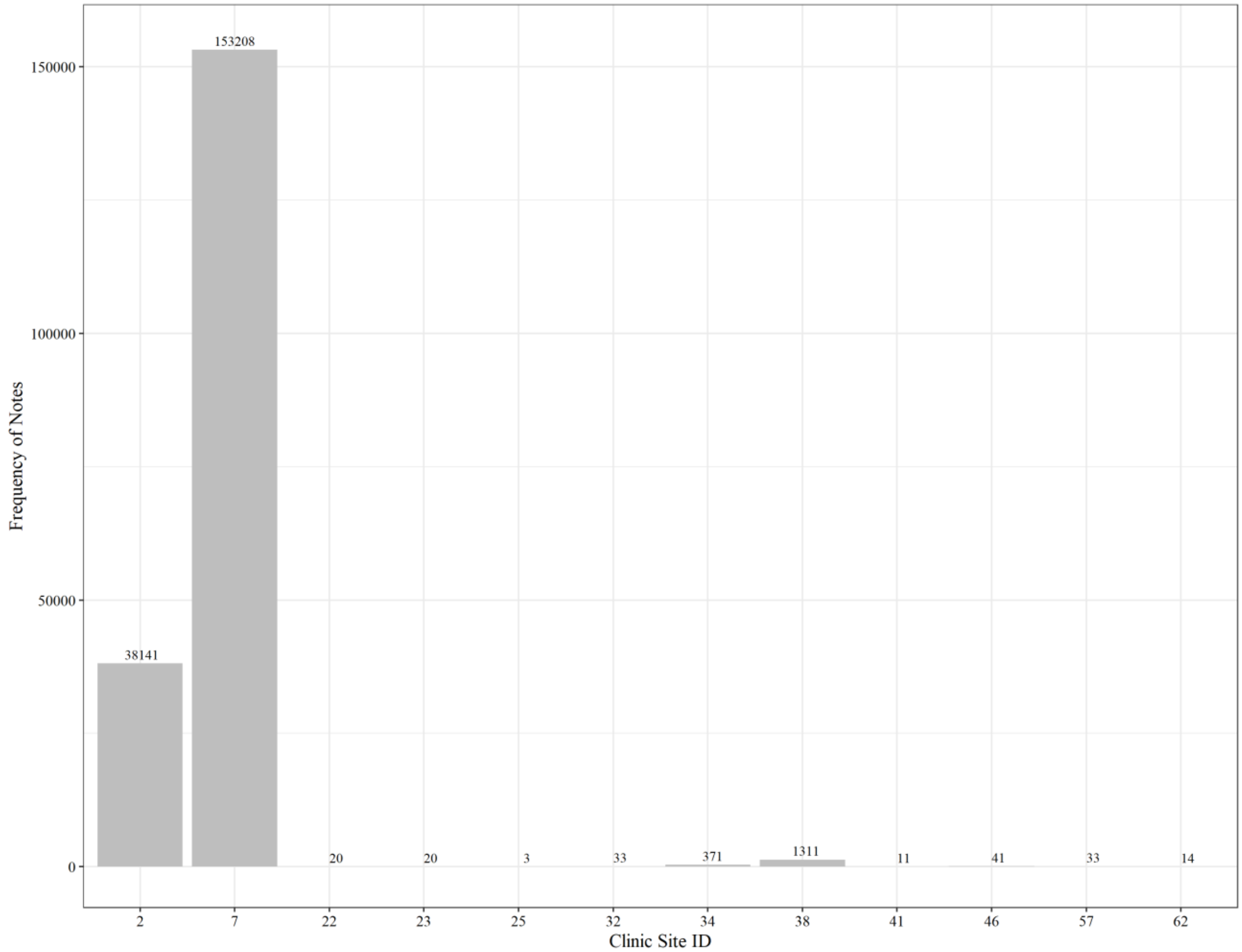


Figure 7 - Frequency of notes per clinic site



4.2.2 Results of the application of preprocessing methods

Table 5 contains information about the preprocessing methods applied to the clinical encounter notes in the EMR data. Each preprocessing method was counted per note (i.e., how many stop words were contained in a note). Overall, stop words and punctuation have higher mean and median elements than spelling errors and special characters. However, focusing briefly on spelling errors per note, the percentage of all notes that contained spelling errors was 97.8%. At the same time, the variability measures among all preprocessing methods (except for lexical diversity) are high which is also evidenced in the range (min and max). For perspective, the mean character length (555.8) is slightly higher than two Twitter posts (a Twitter post is 250 characters). Furthermore, the mean word length (71.5) is between 30 to 60 seconds of spoken speech.

Table 5. Characteristics of preprocessing methods applied to EMR clinical encounter notes (N= 193,206)

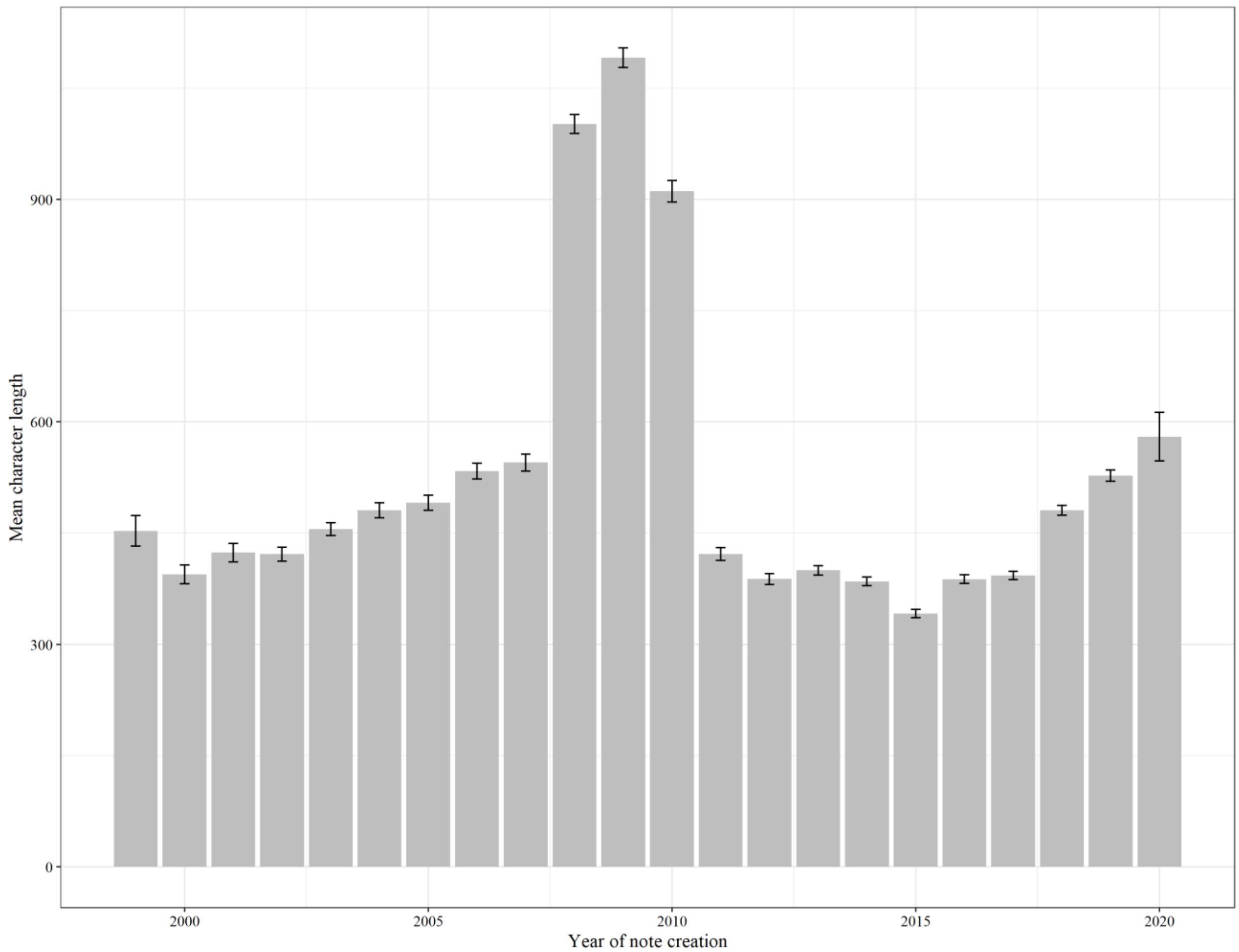
Characteristic	Mean (SD)*	Median (Q1, Q3)*	Min, Max*
Stop word elements	27.3 (27.0)	20 (10, 36)	1, 1075
Punctuation symbols	25.7 (27.8)	18 (10, 29)	1, 833
Spelling errors[1]	12.1 (11.1)	9 (5, 15)	1, 255
Special character elements	2.9 (2.6)	2 (2, 4)	1, 36
Character (string) length	555.8 (551.1)	390 (216, 648)	2, 14056
Word length[1]	71.5 (59.7)	59 (32, 95)	1, 1966
Lexical diversity[1,2,3]	86.2 (11.9)	88.2 (81, 94.9)	22.4, 100

[1] *Punctuation removed;* [2] *Stop words removed;* [3] *Number reported is a percentage*
** Each measure is based from counting each preprocessing method per note*

The time series analyses revealed some consistent trends among most preprocessing methods (Figure 8, and 11 to 14). There is a generally stable trend until 2008. The variation between 2008 and 2010 may indicate changes in the capture of different types of notes which warrants further investigation. Two examples of a low-quality notes from the year 2010 where one contains several questions and their respective responses, similar to that of a survey. The other note appears to contain areas that a clinician would "enter data" such as dates, or checkmarks on a checklist. These EMR notes, contained a low lexical diversity, a high stop word count on the first note, a lower-than-average stop word count on the second note, and both contained a very high punctuation count. Both notes did not contain any meaningful clinical patient information.

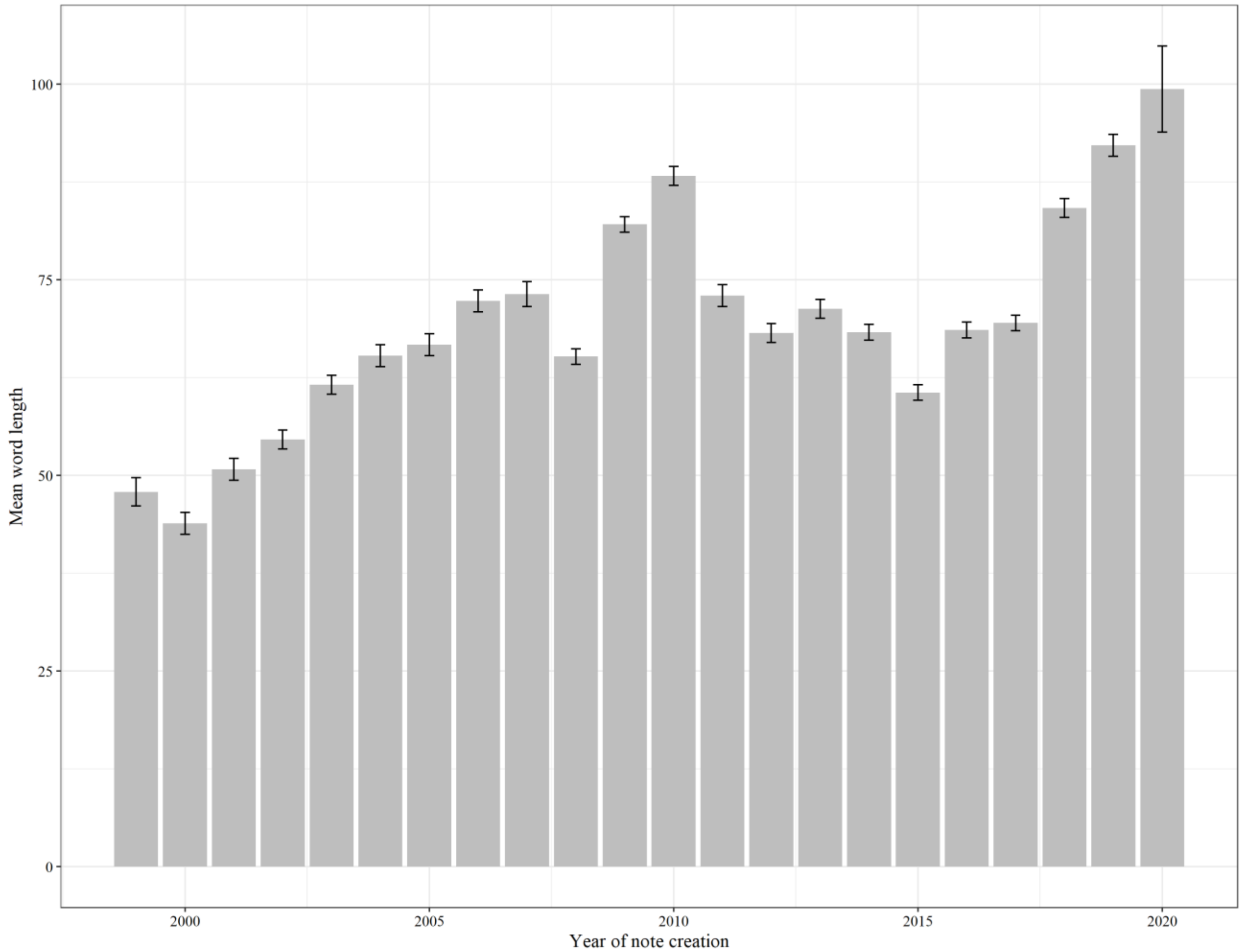
Regarding mean word length (Figure 9), there was an increasing trend during the years when the note was first created. Lastly, while overall sample size was small (i.e., only 2.9% of all notes contain special characters), Figure 10 shows a decreasing trend over time in the mean number of special characters in each clinical encounter note.

Figure 8 – Mean character length per note by year of note creation



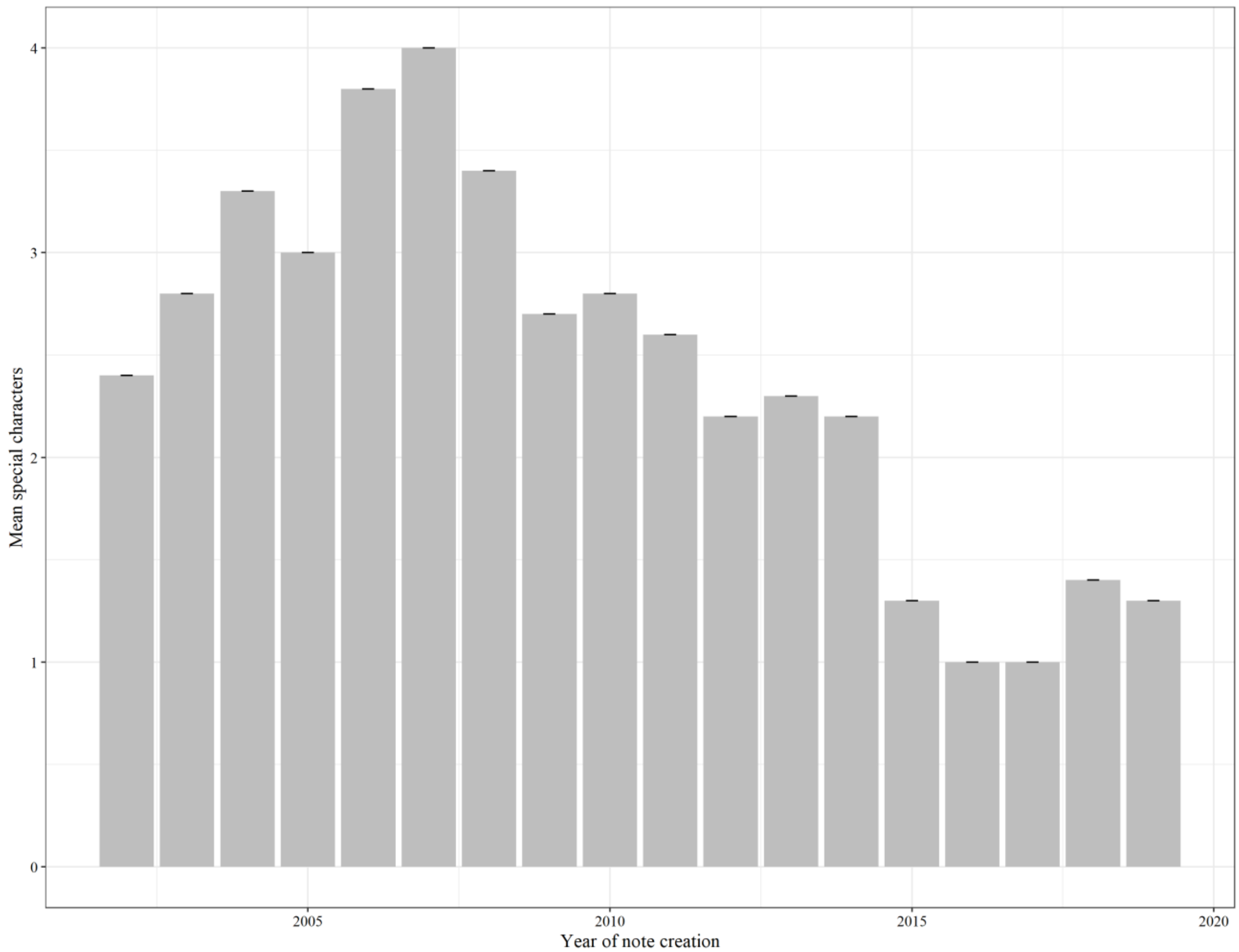
Note: Error bars represent a 95% confidence interval

Figure 9 – Mean word length per note by year of note creation



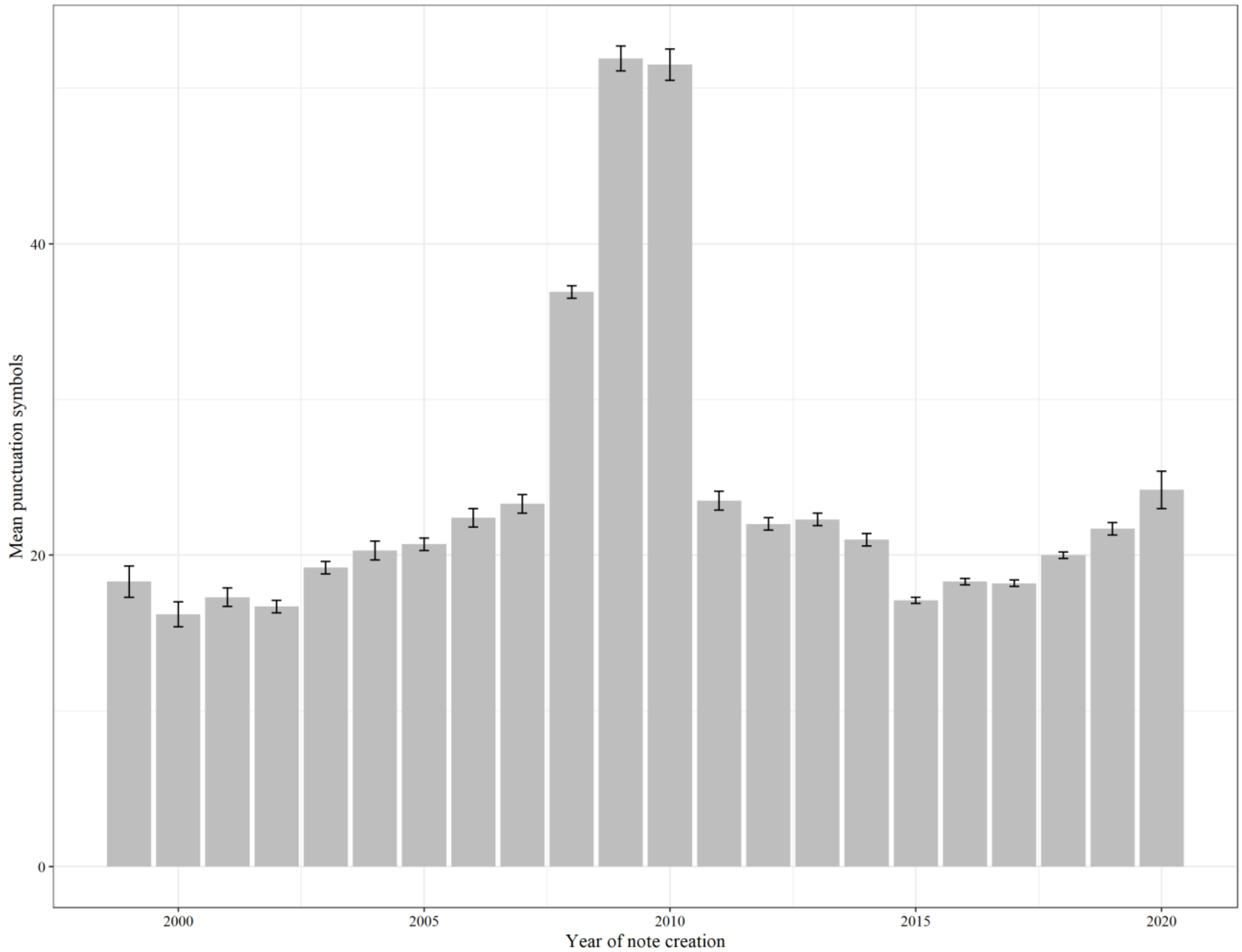
Note: Error bars represent a 95% confidence interval

Figure 10 - Mean number of special characters per note by year of note creation



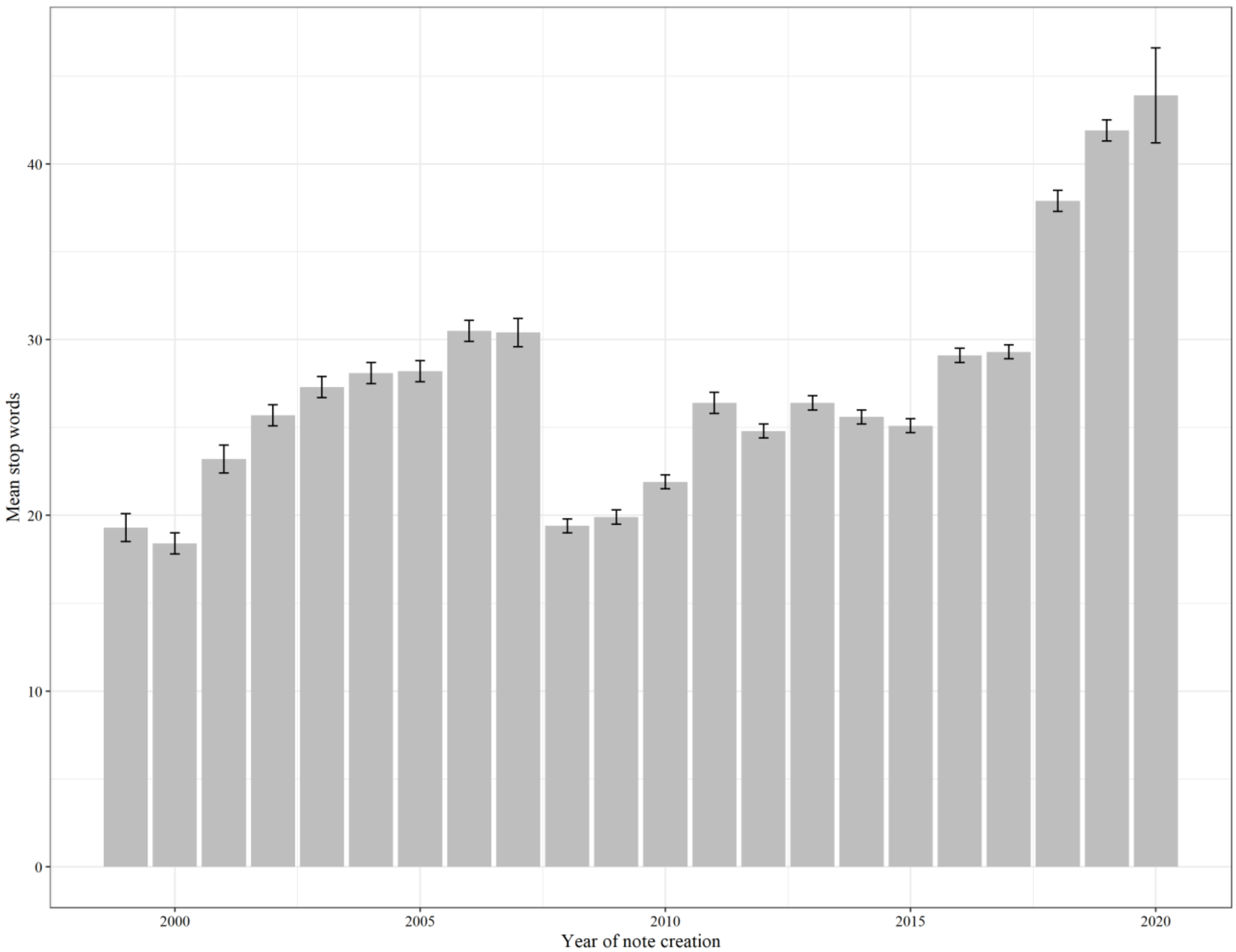
Note: Error bars represent a 95% confidence interval

Figure 11 - Mean number of punctuation symbols per note by year of note creation



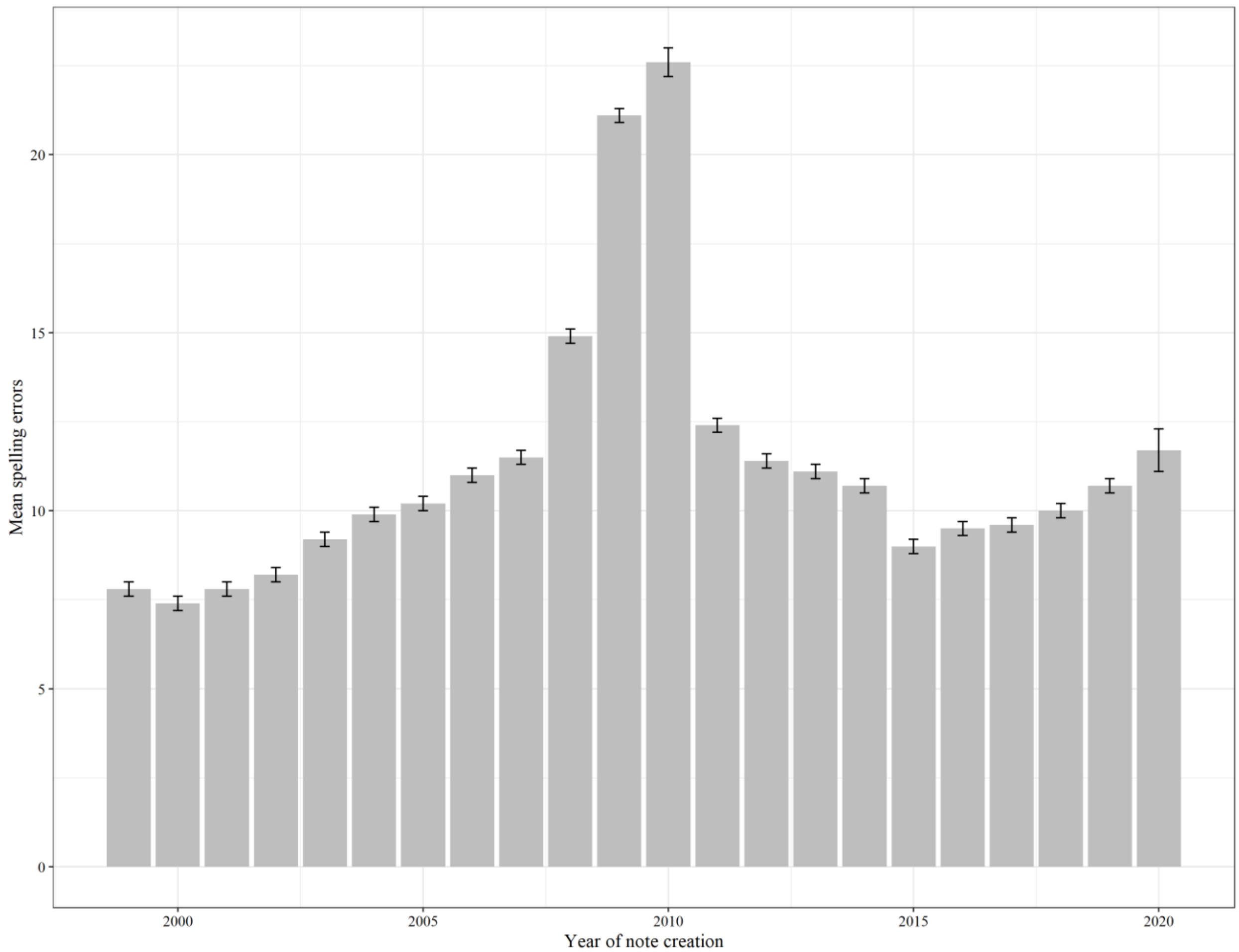
Note: Error bars represent the 95% confidence interval

Figure 12 - Mean number of stop words per note by year of note creation



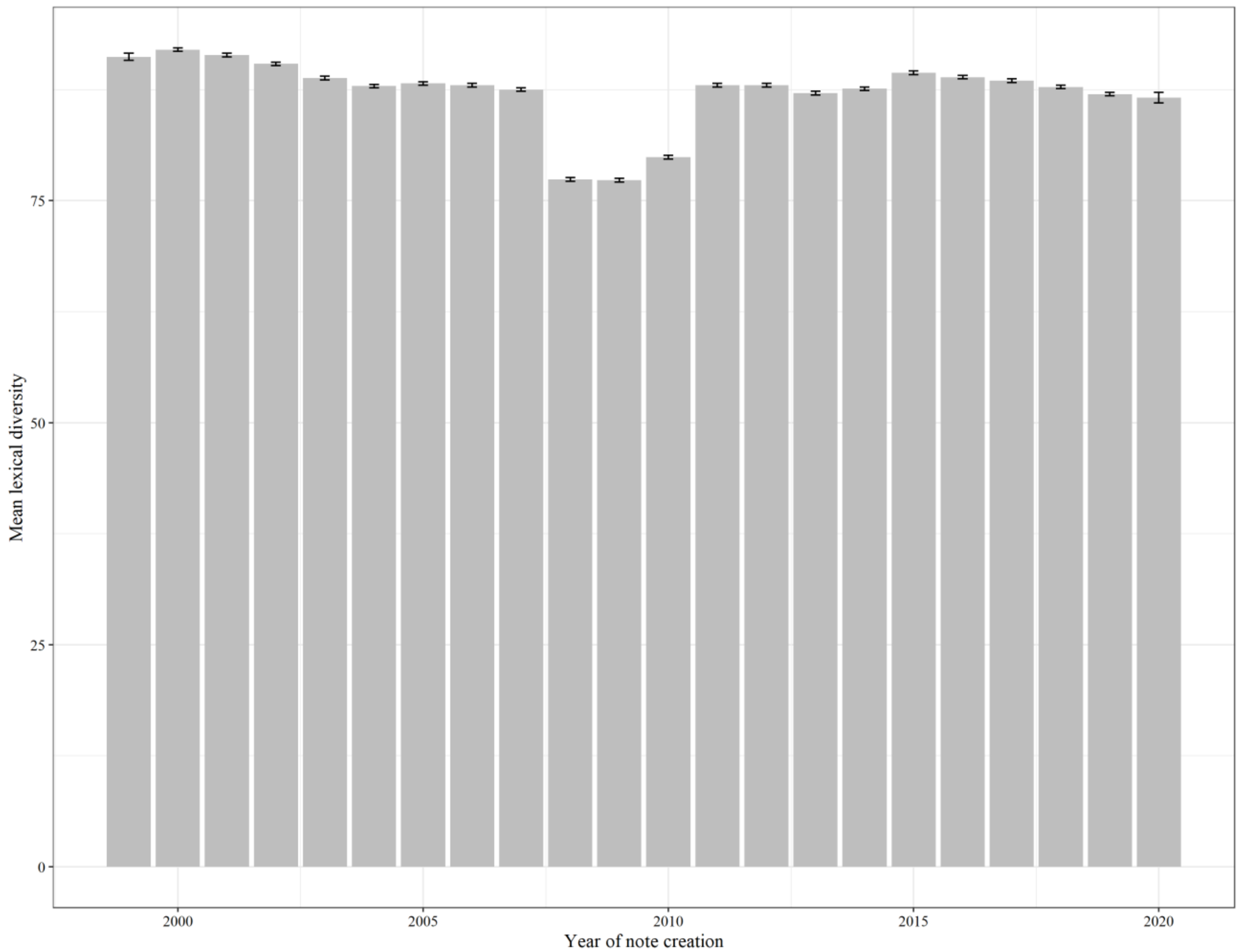
Note: Error bars represent the 95% confidence interval

Figure 13 - Mean number of spelling errors per note by year of note creation



Note: Error bars represent the 95% confidence interval

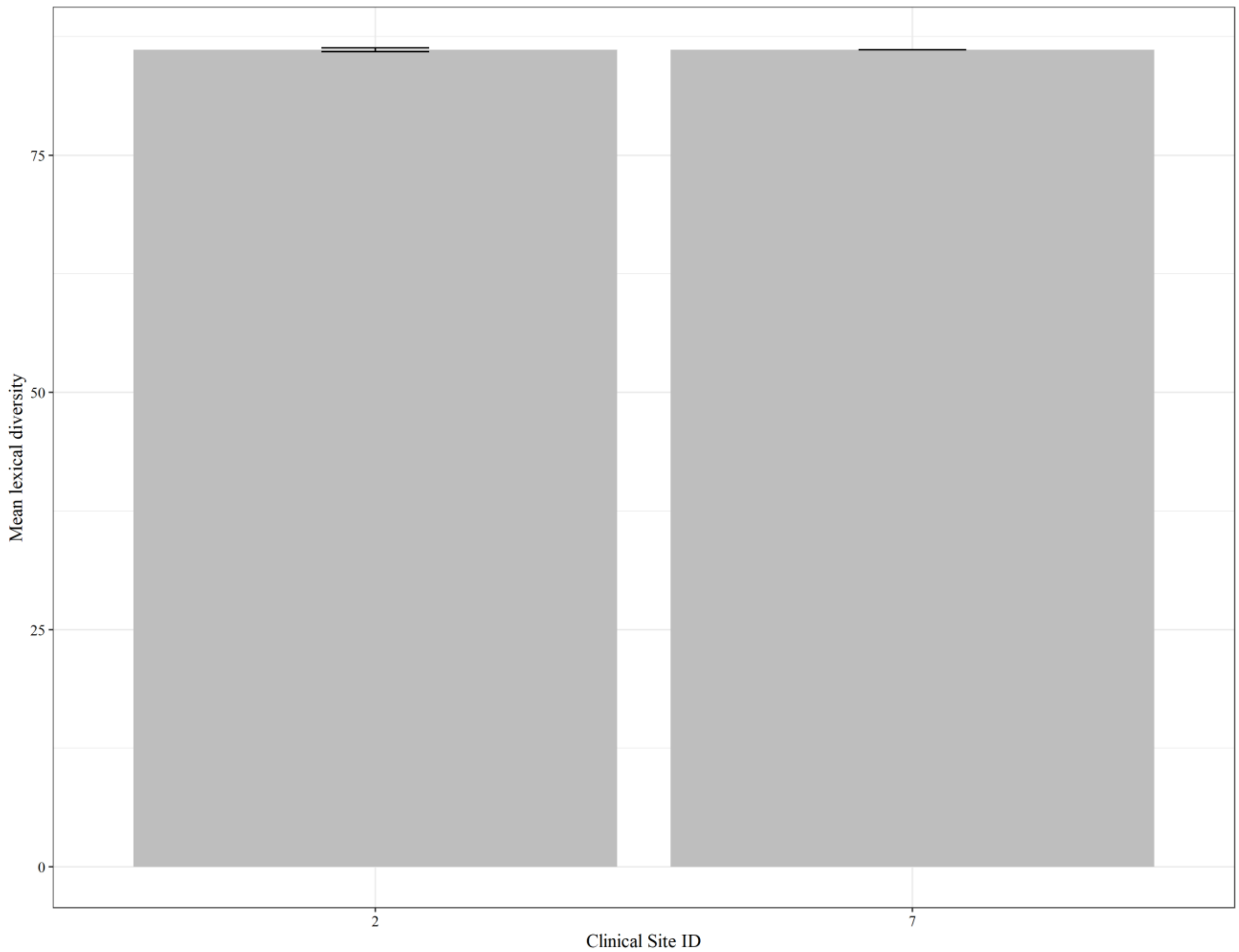
Figure 14 – Mean lexical diversity score per note by year of note creation



Note: Error bars represent the 95% confidence interval

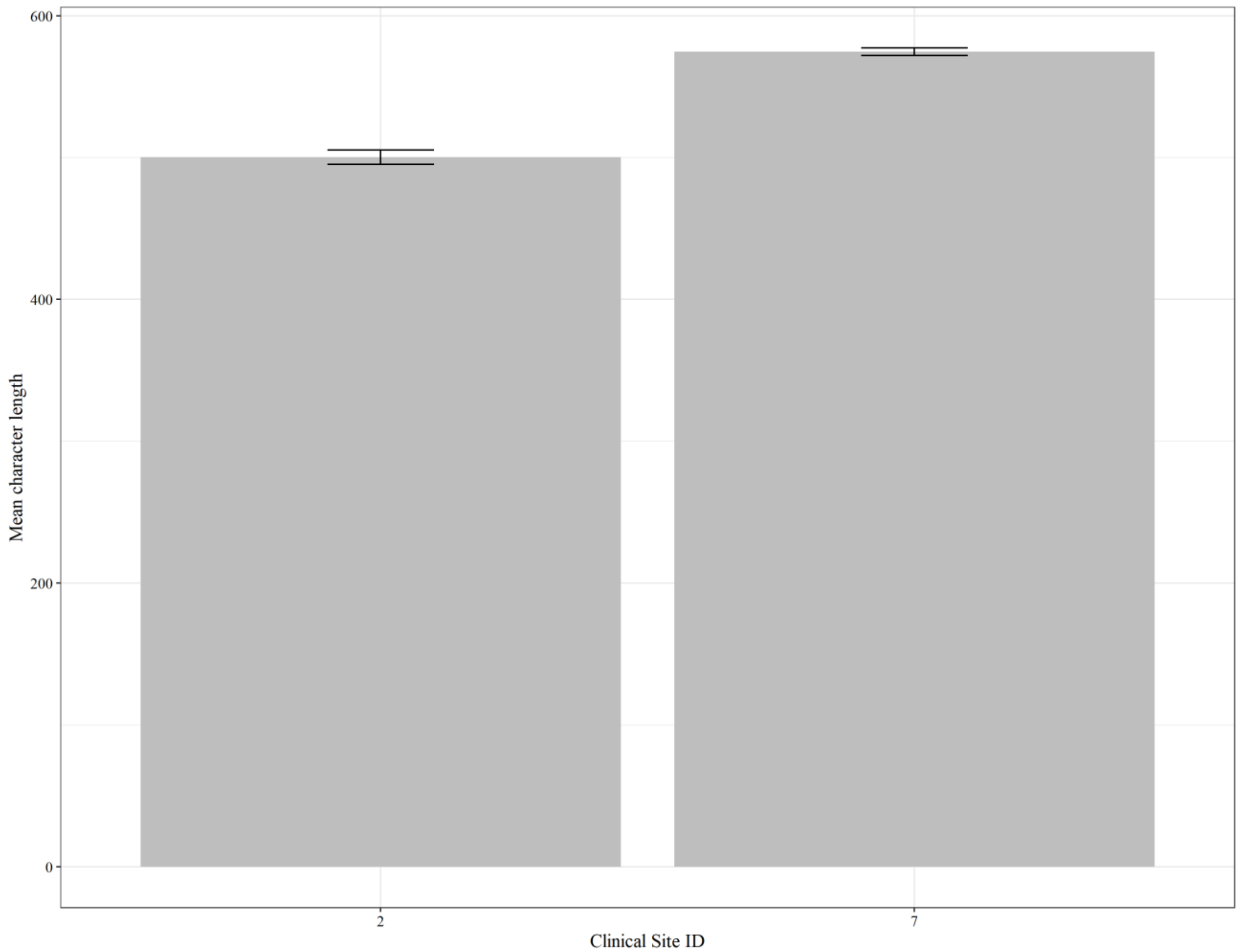
Clinic site consistency depicts the variability among different clinic sites. Figures 15 to 20 reveal limited variability or differences between clinic sites 2 and 7, however there is a visible pattern that clinic site 2 has on average a higher word count, stop word count and spelling error count. Appendix D reveals that sites 32 and 46 have high variation in most of the measures (spelling errors, punctuation, word length and character length). Considering the low sample sizes among the clinic sites other than 2 and 7, there are some sites that contain a narrow CI width, which indicates consistency among notes per clinic site. For example, lexical diversity is fairly stable with narrow CIs for all sites, along with narrow CIs with other preprocessing methods in site 34 and site 57.

Figure 15 - Mean lexical diversity per note by clinic site



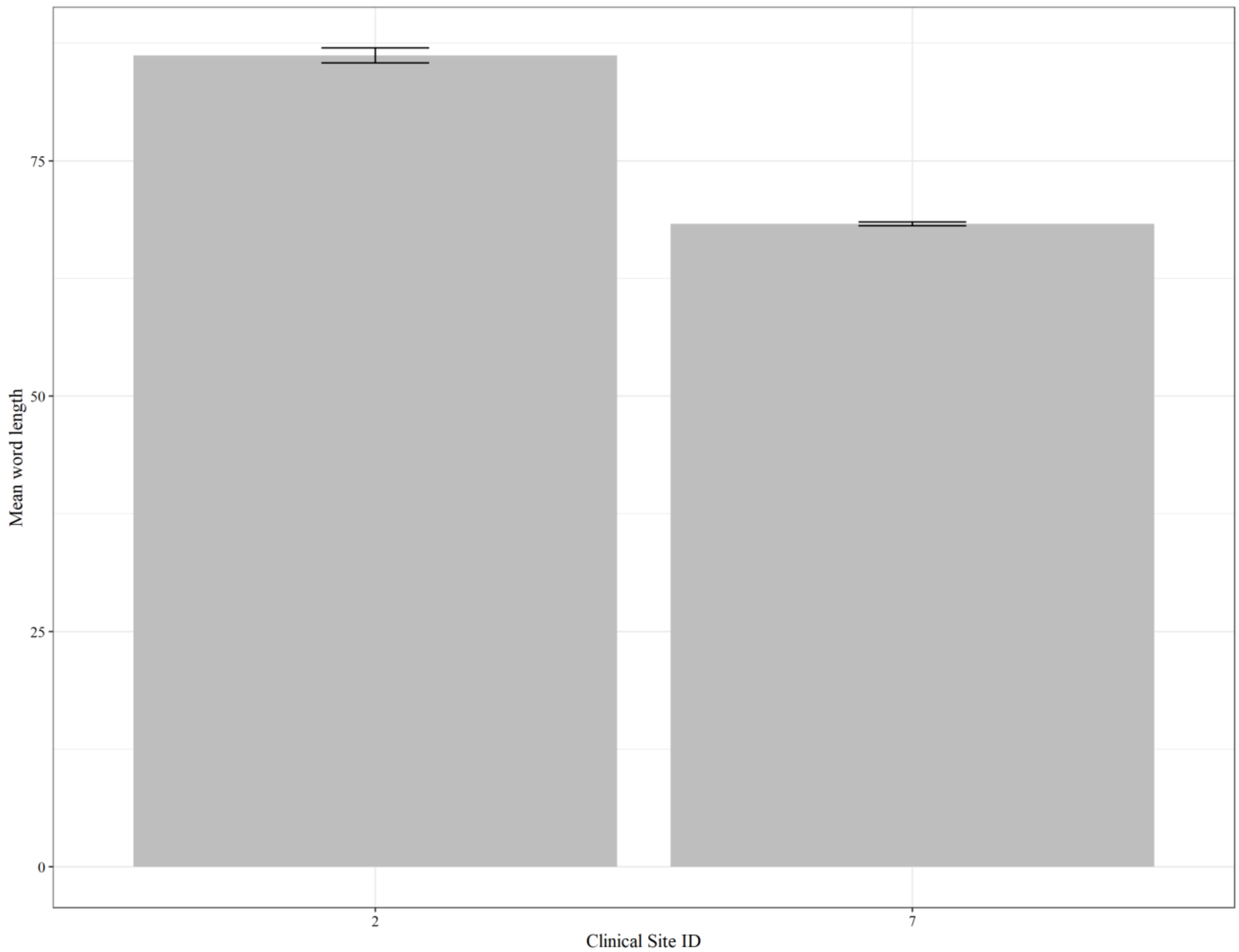
Note: Error bars represent a 95% confidence interval

Figure 16 - Mean character length per note by clinic site



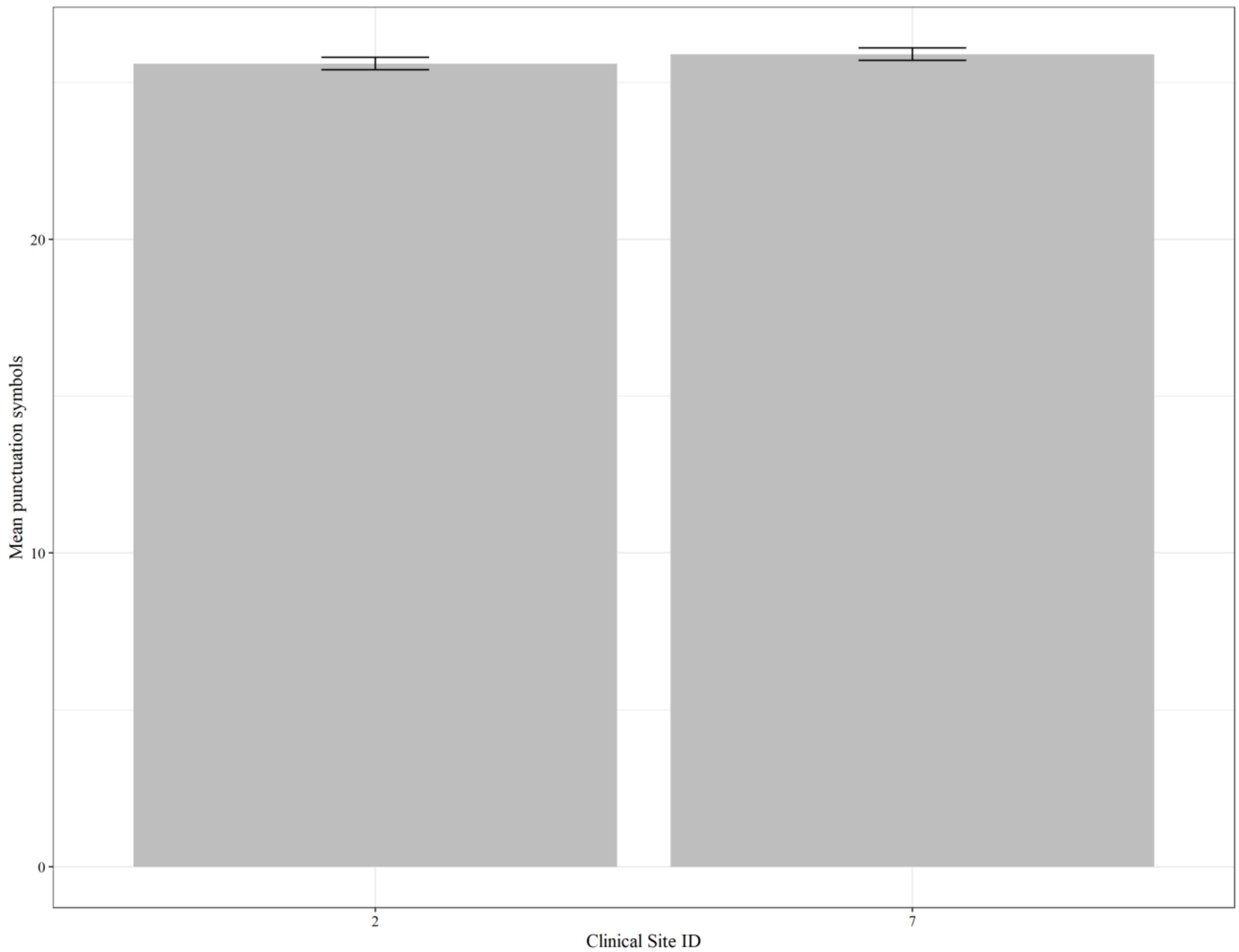
Note: Error bars represent a 95% confidence interval

Figure 17 - Mean word length per note by clinic site



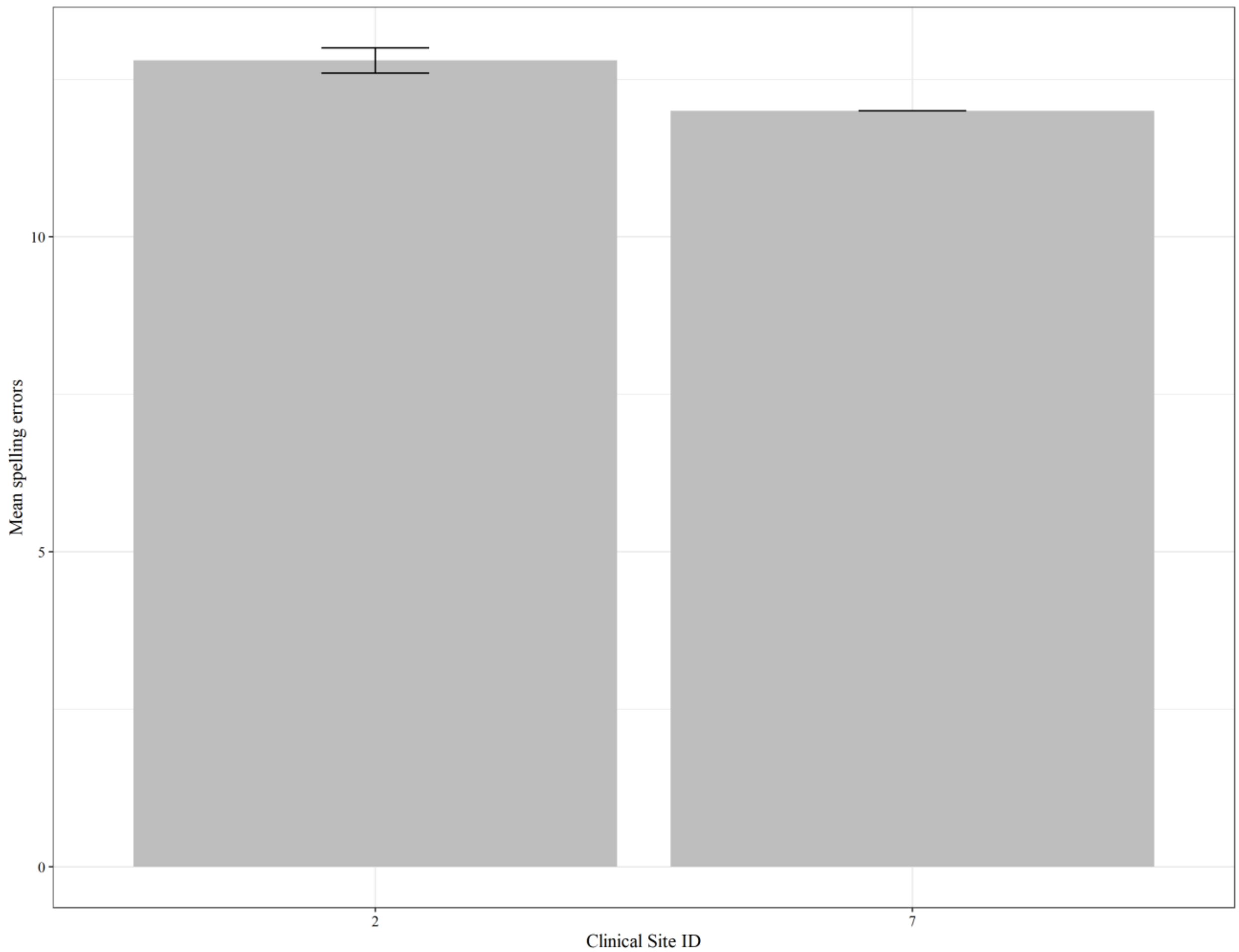
Note: Error bars represent a 95% confidence interval

Figure 18 - Mean punctuation per note by clinic site



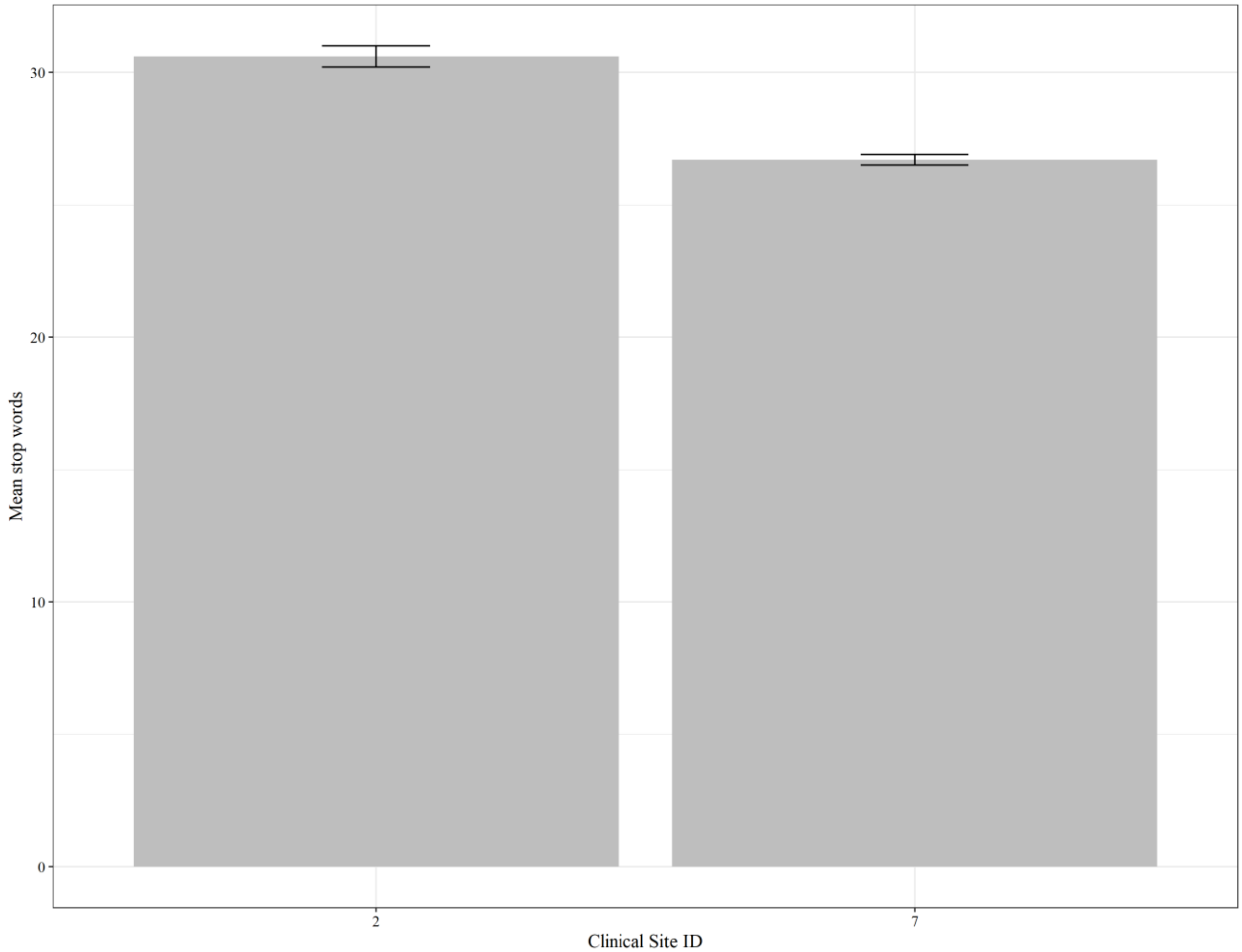
Note: Error bars represent a 95% confidence interval

Figure 19 - Mean spelling errors per note by clinic site



Note: Error bars represent a 95% confidence interval

Figure 20 - Mean stop words per note by clinic site



Note: Error bars represent a 95% confidence interval

4.2.3 Results of the inspection of words after application of preprocessing methods

After applying the preprocessing methods to the EMR notes, the data were inspected to determine the effect on words, punctuation, and characters. Table 6 displays the most frequent stop words, punctuation, and special characters.

Table 6. Frequent preprocessing elements in EMR clinical encounter notes

10 most frequent stop words	**n**	**% of all stop words**
no	266672	5.1
a	236642	4.5
to	225222	4.3
and	177633	3.4
s	149868	2.9
for	148771	2.8
p	145226	2.8
o	141414	2.8
in	126384	2.4
of	118696	2.3
6 most frequent special characters	**n**	**% of all characters**
amp* (i.e., &)	16177	99.0
°	143	0.9
š	16	0.1
±	9	0.1
½	2	0.0
ã	1	0.0
10 most frequent punctuation symbols	**n**	**% of all punctuation**
,	1072809	21.8
:	844613	17.2
.	797994	16.2
[	601548	12.2
]	601447	12.2
/	321264	6.5
-	295812	6.0
*	100368	2.0
)	52036	1.1
'	47488	1.0

**This is the ampersand (&); it results in an error when loading or converting data from one type to another*

The top 10 punctuation symbols listed in Table 6, make up over 96% of all punctuation symbols used in the EMR data, the top 5 symbols (,:.[]) make up close to 80% of all punctuation symbols used in the EMR data. The "amp" special character was the most common of all special characters. The "amp" character is the "&" symbol and represents 99% of all special characters. Lastly, for stop words the top 10 most used stop words account for 33.3% of all stop words used.

Some notable stop words identified are “no” which makes up 5.1% of all stop word usage, and the four letters “s”, “o”, “a”, “p” which make up 13.0% of all stop words.

Table 7. 20 most frequent misspelled words in EMR clinical encounter notes

Misspelled words	Identified meaning	n	% of all misspelled words
Dt	date	168734	7.4
yyyymmdd	date	82646	3.6
bp	blood pressure	73438	3.2
Dr	doctor	50917	2.2
resultval	result value	34235	1.5
fbs	fast blood sugar	29928	1.3
Fu	follow up	29438	1.3
ldl	low density lipo protein	24193	1.1
htn	hypertension	23255	1.0
mo30	Unknown	20377	0.9
abd	Abdominal	19794	0.9
Rx	prescription	19176	0.8
chol	cholesterol	17535	0.8
meds	medication	17344	0.8
Sx	surgery or symptoms	15532	0.7
qd	daily	13685	0.6
hba1c	HBGA1C Hemoglobin A1C	13554	0.6
fup	follow up	13390	0.6
crea	creatinine	13241	0.6
prn	as necessary	13170	0.6

Table 7 displays the 20 most misspelled words, which make up over 30% of all misspelled words. However, 19 out of 20 words were identified as medical or other abbreviations. The two most frequently misspelled words were abbreviations of dates; they accounted for over 11% of all spelling errors. Furthermore, aside from the identified dates listed previously, the other abbreviations were characterized as medical abbreviations except for “mo30”.

Focusing on Table 8, the first quartile of word and character length counts, and percentages were inspected. Notes that contained between 1 and 22 words made up over 65% of the first quartile of notes containing 32 words or less. Notes that contained between two and 101 characters made up 39% of the first quartile of notes containing 216 characters or less. Notes with 102 to 216 characters accounted for almost 62% of the first quartile of notes containing 216 characters or less.

Table 8. 1st quartile of word and character length per note

	n	%
Word length per note		
12 - 22	18012	37.2
23 - 32	16378	34.0
1 - 11	13919	28.6
Character length per note		
102 - 121	6574	13.5
162 - 181	5476	11.5
122 - 141	2630	10.4
82 - 101	4687	10.1
182 - 201	4773	9.9
142 - 161	4223	8.8
42 - 61	3943	8.3
2 - 21	3836	8.1
202 - 216	3702	7.7
62 - 81	3384	6.9
22 - 41	2859	5.6

Table 9 focuses on the 20 most frequent words within the first quartile of word length (i.e., in notes that contained 32 words or less). Certain observations are similar to the misspelled words. For example, the top 20 words that contained medical abbreviations made up almost 12% of all words within notes containing 32 words or less. The remaining words (except for “1”) were stop words; stop words were not removed when counting words per note. Stop words made up almost 10% of the words among the 20 most frequent words within the first quartile of word length.

Table 9. Top 20 words among notes that contain 32 words or less

Top 20 words	Context	n	%
a	medical abbreviation	23313	2.7
s	medical abbreviation	22799	2.7
p	medical abbreviation	21723	2.6
o	medical abbreviation	14879	1.8
to	stop word	11897	1.4
for	stop word	10322	1.2
no	stop word	9881	1.2
dr	medical abbreviation	8451	1.0
and	stop word	7453	0.9
by	stop word	6817	0.8
in	stop word	6516	0.8
of	stop word	4984	0.6
1	unknown	4916	0.6
bp	medical abbreviation	4757	0.6
with	stop word	4741	0.6
on	stop word	4713	0.6
pt	medical abbreviation	4498	0.5
will	stop word	4351	0.5
see	stop word	4314	0.5
well	stop word	4303	0.5

4.3 Summary of results

The scoping review included 41 articles. Most were empirical research articles. The number of articles increased over time. Furthermore, while most articles were from the computer science and engineering disciplines, many articles used EMR data or created corpora using medical data. Most articles used frequency measures to count documents and words that were used to describe the size of UTD. Furthermore, articles that contained annotations were also documented to understand if these were manual processes or automated processes. Most articles utilized manual annotation procedures (e.g., expert opinion). Restructuring and reorganizing methods (e.g., removal of stop words) were primarily used for preparing text data for analysis purposes. Internal and external validation methods were used for various statistical or machine learning algorithms. Lastly, the data quality dimensions were logged based on the number mentions in an article (i.e., an explicit mention of the term “accuracy” or “relevance” and so on). The top three mentions were “accuracy”, “relevance” and “comparability”.

The scoping review further revealed that although there is an abundance of research about increasing accuracy using NLP techniques, very little of this research reports indicators of data quality for UTD. Methods to improve algorithm outcomes are focused on a variety of preprocessing methods such as restructuring or reorganizing techniques or validation methods. The scoping review provided guidance as to which text quality indicator topics have been explored for EMR data. Some of the notable quality indicators focused on characteristics and properties specific to text, such as misspelled words, lexical diversity, security, ambiguity of abbreviations, variability of words (i.e., can be classified as misspelling, or words that are spelled the same but have a different meaning depending on the context), and text data that contain noise such as punctuation, stop words, and special characters. Other indicators that do not fall into categories that are specific to text data are character length and word length per note.

The case study involved a detailed analysis on preprocessing methods. The methods were informed by the scoping review and previous literature and were applied to EMR clinical encounter notes. The case study revealed that the distribution of preprocessing methods such as stop words, punctuation, special characters, word and character length, and lexical diversity, can be assessed using EDA methods. Furthermore, the application of concepts from previous data quality frameworks (i.e., measurement over time and by clinic sites) were used to find any anomalies such as outliers over time or between clinic sites. While attempting to detect misspellings, an abundance of abbreviations utilized in clinical settings were identified. Lexical diversity may be useful to identify notes that have a large amount of word redundancy. Lastly. an inspection of word length from the notes that contained the first quartile of word length revealed that two thirds only have up to 22 words; at the same time the majority of the words were either stop words or medical abbreviations.

Chapter 5 – Discussion and Conclusion

5.1 Summary and discussion

The purpose of this research was to examine preprocessing methods for UTD and evaluate the quality of UTD in RCEHD with a focus on EMRs. For objective 1, a scoping review was undertaken to examine preprocessing methods used to prepare UTD for analysis. The steps in the scoping review were: 1) apply the search strategy; 2) apply the inclusion and exclusion criteria for title and abstract screening and full text screening; 3) extract information from the articles; and 4) describe and visualize the extracted information, including i) characteristics of the articles, ii) characteristics of text data and preprocessing methods, and iii) data quality indicators described in articles about EMR data.

For objective 2, preprocessing methods identified from the scoping review were applied to UTD in clinical encounter notes from MaPCReN, a major source of EMRs in Manitoba. The following steps were taken for the case study: 1) an EDA on the characteristics of the EMR notes, and 2) an application of preprocessing methods. The application included i) describing characteristics of preprocessing methods in EMR notes, ii) plotting preprocessing methods over time of when the note was created and by clinic site to observe if there were variations over time or by site, and lastly iii) an inspection of the individual elements (words, characters, punctuation, spelling errors) that were subjected to the preprocessing methods.

The scoping review captured both preprocessing methods and data quality indicator topics for UTD. Few articles in the scoping review discussed the quality of UTD before analysis or preprocessing was initiated. The main discussion points raised in this published literature were in the challenges of defining data quality, the choice of data quality indicators for UTD depends on the context of the text data, and the challenges associated with EMR UTD are different than other types of text data.

One difficulty with describing the quality of UTD is the lack of standardized terminology. Strong et al differentiate “information quality” from “data quality” in terms of its specific goals; information quality is about assessing the needs of information users while data quality refers to the fitness of data for its intended use(120). However in Chen and Tseng’s article the authors do not make that distinction; their indicators of information quality are similar to those found in data quality frameworks(102). In the scoping review and supporting literature, it was difficult to settle on a standard definition of data quality for text data and its respective dimensions in specific contexts.

Amongst the data quality measures for text data identified from the scoping review, the data quality indicators are tailored specifically to the EMR data being assessed, which suggests that data quality frameworks for UTD may depend on the context or type of data that is being assessed. Language usage must be contextual to the environment (i.e., vernacular used in product reviews differ from vernacular used in EMR notes). For example, several of the data quality indicators for UTD that Chen and Tseng's article refers to, do not necessarily apply to data in EMRs such as the data quality dimension for "objectivity", and its indicators that assesses whether a product review is an opinion rather than factual(102).

Some of the challenges with UTD in EMRs are different than the challenges with UTD found in social media data or organizational reports. One of the biggest challenges of UTD in EMRs is ensuring privacy, anonymity, and confidentiality of patient data. Pantazos et al states that as UTD in EMRs increase in usage, it must achieve two goals 1) that it is de-identified and anonymized and 2) EMRs that are de-identified must contain accurate information about the unidentified patient and be coherent to the reader(86). The scoping review revealed that high-quality UTD within EMR must provide readability, correctness, and consistency of a patient's record (i.e., the de-identified record must maintain a level of accuracy that reflects the patient). Regardless of whether UTD in EMRs are de-identified or not, the data still needs to meet a certain quality standard so that it is usable for research. For example, names, addresses, and dates that are de-identified at the patient level must be consistent throughout an entire patient record.

The case study provided a real-world example of the variability of words in clinical encounter notes for data quality purposes (e.g., spelling errors). The selection of preprocessing methods was informed by the scoping review and previous literature. The key findings were characteristics of EMR data based on measuring preprocessing methods differ versus other types of text data, the EDA of EMR notes and each individual preprocessing method, the analyses of when the note was created along with clinic sites, a discussion on the inspection of individual words and note types, and lastly, a discussion on the overall insights derived from the case study.

The study results highlight that 97.8% of all notes contained spelling errors as identified by the spell checker using a generalized dictionary. However, this high percentage can be attributed, in part, to the large number of abbreviations and other words that are medical terminology, (e.g., “prednisone”), that the spell checker could not identify. Lexical diversity was high for EMR data when compared to other types of text data, possibly because medical professionals use clinical encounter for multiple purposes, such as to describe the purpose of the visit, list chronic health conditions that the patient encounters, detail prescription medications of a patient, describe mental health, and document risk behaviors. Kiefer evaluated five different types of text data: social media, prose, news, chat, and industry data that contained downtime records in a production line. Based on the indicators Kiefer described, social media data, chat and industry data contained high levels of spelling errors, and all five types of text data had low lexical diversity. Kiefer also claimed that a high number of abbreviations for a single word or concept can lead to poor data quality (e.g., diabetes has several different abbreviations such as: “db”, “diab”, “dm”)(44). However, within clinical settings, abbreviations are standard in practice. Thus, there are generally-accepted abbreviation lists that have been created; an example is the clinical practice guidelines for primary care nurses from the Government of Canada website(121).

In the EDA, a frequency analysis was conducted for categorical variables (clinic site, year of note creation). It revealed a large range in notes per clinic site, resulting in a lack of precision (i.e., high variability) in measures of data quality for all but clinic sites 2 and 7. There were a larger number of notes from the years from 2000 to 2019. In particular, between the years of 2008 and 2010 there was a large increase in the frequency of notes.

The average number of stop words, punctuation symbols, and special characters combined make up a large amount of textual variability. For example, on average there were roughly 27 stop words, about 28 punctuation and special characters, and lastly, at least 12 spelling errors per note. There were few special characters per note; the most frequent special characters were “&amp” or “amp” (after removal of punctuation). The former signifies that there is a conversion error that occurs when either exporting data from one database to another or converting from one file type to another. Conversion errors can result in data quality issues because these errors result in special characters being added or removed. Regarding character or word length, the distribution of length is skewed, which indicates a high SD, thus resulting in mean not being a good measure of central tendency, median as a measure of central tendency would be a valid alternative. Lastly, the mean lexical diversity score suggests that most notes have high lexical diversity, meaning that most notes contain a diverse set of vocabulary and tended to not repeat words. However, when looking specifically at notes that contain a low lexical diversity score, many words were repeated. For example, the note with the lowest lexical diversity score (22.4%) repeated words such as: “refills”, “tab”, “cap”, “for” and “mo30”.

An analysis of variations from when the note was created, and clinic site reveals that measuring preprocessing methods can show variations in data that may have low or high quality. For example, the measurement of punctuation counts per note per year revealed notes that were akin to clinical surveys or requisition forms with no contextual information. The literature search conducted by this research project did not reveal any similar analyses in the field of data quality where preprocessing methods were applied to UTD in EMRs over time or by clinic site. However, the MCHP data quality framework was instrumental in guiding these types of analyses, this exemplifies those methods utilized for structured data can similarly be applied to UTD.

A further inspection of words reveals that certain spelling errors contain words that would raise concern if filtered out or have an automated spelling correction correct the “error”. For example, most spelling errors identified legitimate medical abbreviations. Recall that misspellings make up a mean of around 12% per note after punctuation has been removed. In this research there was no medical dictionaries to be found in R libraries therefore we utilized a general open-source dictionary. Since roughly 30% of misspelled words have been identified in medical terminology this can be concerning as general dictionaries may not be the correct source for detecting spelling errors in medical text. Upon further inspection, lowercase dates (e.g., july versus July), medical abbreviations, and several words that are associated with medical terminology that have been fully spelled are detected as misspelled (e.g., creatinine, lymphadenopathy, prednisone). However, it should also be noted that several words associated with medical terminology that have been fully spelled out are correct (e.g., chronic, pain, blood, lipid, symptoms). A silver lining could be that if a general dictionary cannot identify correct spellings of medical abbreviations, the benefit is that it does detect medical abbreviations indirectly through spelling errors. which can aid in the application of other algorithms to identify abbreviations and disambiguate them.

For the stop word dictionary from the tidytext package in R, there are two interesting observations. First, the word “no” is filtered out when removing stop words; this can be problematic because “no” has a negation function (i.e., “no Parkinsons”, or “no symptoms”), and can result in a loss of context of data when analysing UTD. Second, the letters “s, o, a, p” are filtered out as well, which is indicative in medical terminology as an abbreviation of “Subjective”, “Objective”, “Assessment”, and “Plan”. SOAP is intended to form a structure when typing out a clinical note in primary care settings to capture all relevant information during a clinical encounter with a patient. Filtering these letters out could result in a loss of contextual information.

Punctuation symbols made up 26.5% of all data elements. Identifying and counting punctuation symbols has been useful for the identification of a non-standard note. For example, as previously mentioned there have been some notes in EMR records that have had either requisition forms, or “checklists” copied onto the text field in the EMR record. This can lead to a quality issue since some of these “checklists” have a large amount of punctuation which contain questions and options (e.g., history of back pain? [yes/no]), at the same time, the note does not identify what options the primary care clinician chose or wrote. As a result, contextual information is not available, or is not clear, nor any meaning can be derived within these types of notes. The identification of punctuation symbols has been instrumental in identifying the quality of a note from the perspective of context and meaning. A red flag would be if there is a larger than average amount of punctuation characters within a note, the note should be closely inspected.

Lastly, for character and word length, notes with few words (e.g., fewer than 12 words) or having very few characters (e.g., fewer than 22 characters) may explain quality issues since no context or meaning can be derived from sparse notes. Another potential methodological issue for text analysis in EMR notes is that words are tokenized before being counted. This may explain issues since tokenization involves counting single letters (e.g., “S”, “O”, “A”, “P”). As suggested in previous literature, notes with few words or characters may be removed. Removing notes with few words, then removing notes with few characters, is one potential process, since the elimination of notes with few words would also result in the elimination of notes with few characters. Thus, a potential approach would be to count the words, then count the character length within those words, then eliminate words with low character counts.

The preprocessing methods successfully identified a subsection of notes as having low quality due to a lack of contextual information from copying and pasting requisition forms or survey questions onto the EMR note field. Furthermore, preprocessing methods that were applied using the date when the note was first created, captured a period of time (i.e., 2008-2010) where notes were ‘irregular’.More specifically, the measurements of the preprocessing methods revealed variability over time. This suggests that measuring preprocessing methods can capture low quality notes.

The case study revealed some valuable insights. Firstly, in accordance with Taleb et al’s article, converting all the EMR notes to a structured (tabular) data format enabled exploratory analyses of the characteristics of the dataset and characteristics of the preprocessing methods to be streamlined(41). Converting the unstructured elements of the EMR notes and semi-structured elements of the other variables (e.g., clinic site id, patient id) into a single data structure facilitates quality evaluation(41). This should be the first step taken to prepare UTD for analysis or data quality assessment. Second, in Spasic and Nenadic’s(40) systematic review about clinical text data, the properties of text data include size, provenance, collection methods, and annotation. EDA can be used to characterize EMR data. For example, this included average word length, average character length, and frequency of annotation. Third, few articles in the scoping review conducted an EDA of text data. Christen et al presented a table where the authors reported on the amount of outlier data, through looking at string or word length(75). Although the authors did not use EMR or health-related data, inspection of outliers from preprocessing is recommended. Lastly, this research focused on UTD from EMRs gathered in primary care clinics. These methods may not generalize to other types of RCEHD. There may be different software systems that record and collect information, and there are different ways of expressing written language that is dependent on the context (i.e., emergency room notes differ from primary care notes). In addition, the results of this analysis may not generalize to all primary care clinics. Furthermore, there are subsets of the population that experience inequity of access to health care, and thus may not have a family physician or have difficulties receiving care in a primary care clinic(122). For these reasons, individuals captured in EMRs from primary care clinics are unlikely to be representative of individuals in the general population. However, it should be noted that the text data elements examined in this research are linguistically common. For example, stop words are the most frequent words used in the English language, and punctuation symbols are used for grammatical order and organization. Furthermore, any free text field can be characterized by word length, lexical diversity can be measured to assess the repetitiveness of a text field, and spelling errors can be identified. The preprocessing methods used in this research can be applied to other RCEHD, so long as the text data in other clinical settings closely mimic the properties of an EMR note (i.e., each document is a short text note, its context is the medical setting, and there are other variables such as patient id, clinic site id, and when the note was created to structure a row of data).

5.2 Strengths and limitations

This research has some limitations. For objective 1, the scoping review was restricted to English language articles. No grey literature was included. Accordingly, the scoping review may not represent all published articles about UTD quality. Furthermore, articles that discussed preprocessing methods utilized for performance improvement of algorithmic modelling outcomes were not included. It was not feasible to address all different modelling techniques for text data (i.e., collect all articles that conducted a sentiment analysis or other methods). The articles selected were those that focused on quality of text data, or that focused on preprocessing methods with an emphasis on data quality. Choosing key words for the scoping review search strategy was challenging. This was due in large part to the lack of standardized terminology along with a diversity of terminology within the NLP and data quality literatures.

For objective 2, one limitation is that this study looked at a small, restricted set of preprocessing methods to measure data quality in UTD. Other indicator topics were not explored, such as disambiguation of abbreviations. Some measures of data quality, such as relevance and accuracy(14,57), require the use of annotated documents to evaluate the source documents. It was beyond the scope of this research to develop and apply these gold standard annotated documents to measure data quality. There is significant work required to establish gold standard annotated documents and there are currently no clear guidelines for their development. A second limitation pertains to the dictionaries used for identifying stop words and spelling errors. This research used a general dictionary to identify stop words and spelling errors, because R software packages do not use spelling dictionaries created specifically for medical/clinical purposes or for EMR data. Accordingly, several medical terms were identified as spelling errors when the words were spelled correctly. The adoption of a medical dictionary would result in fewer misclassified spelling errors. Similarly, the stop word dictionary used for this research removed certain one letter words that are medical abbreviations. Furthermore, the stop word dictionary removed negating words (e.g., “no” or “not”), which can cause contextual issues. A third limitation pertains to not using the variable “Note Type”. Note types may be used to discover different types of notes, which may indirectly affect note quality. However there was a lack of literature referring to guidelines to utilize note types as an indicator for data quality for text. Lastly, the case study only used 10% of the total lines of the entire EMR datafile which may affect generalizability of the results. This subset was used because of the the computational limitation of converting the text data to structured dataset.

Despite these limitations, the research has many strengths. The major strength of this research is that it used a systematic approach to examine and apply preprocessing methods to describe data quality indicators for UTD in EMRs. The combination of methods, which included a scoping review and a case study, was helpful for exploring indicators of data quality for UTD. For example, through this scoping review privacy and security were identified as unique issues of text data quality. Furthermore, data quality indicators from established data quality frameworks (i.e., MCHP's data quality framework) can be conceptually leveraged for UTD. For example, trends in the text preprocessing measures over time, and variations across clinic sites were inspired by time series analyses and geographic analyses used in the MCHP data quality framework. Lastly, this research contributes to the limited literature about quality indicators for UTD in EMRs. Data quality for UTD is an important area for research within the medical field because written language in EMRs is different from other types of text data found in textbooks or social media.

5.3 Research significance

The information extracted from UTD in EMRs can complement information extracted from structured data about health and health service use. This study has shown that many of the same indicators used in existing data quality frameworks are relevant and may apply to UTD, but new indicators must be considered that describe the properties of text data in RCEHD. Many indicators identified in this research can stand on their own (i.e., no need to utilize external data for validation purposes) similar to several indicators in existing data quality frameworks for structured data. An example is missing data indicators(15). However, at the same time, specialized reference standards are also important for validating UTD(27). Analyses of the quality of clinical encounter notes can provide a basis for future studies about the reasons why people seek health services. Researchers that utilize EMRs can benefit from this research because it can open new avenues for analyses of RCEHD.

Organizations are increasingly using large volumes of data and new methods of analytics(123). It is estimated that 80% of all data in organizations are comprised of text data(123). UTD quality indicators can be used by health care organizations to check the consistency and redundancy of patient information(124). These measures can also facilitate the routine use of UTD in health research. By structuring text data and then applying data quality indicators to UTD in RCEHD, researchers and users of data can reliably expect a certain level of quality in the results of research or reporting. Organizations should not proceed with ad-hoc quality checks since EMRs are more prone to quality issues than text in novels or structured organizational documents. Data quality must be approached with rigour and standardization.

5.4 Recommendations for future projects

A number of opportunities for future research exist. First, –a scoping review could be conducted to identify operational definitions for data quality dimensions specific to text data within RCEHD. Akin to Weiskopf et al's scoping review on discovering methods and dimensions(32), the scoping review could be used to create definitions for dimensions of data quality for UTD in RCEHD. This research has shown that there are unique characteristics of text data that are not present in numerical data (e.g., grammar rules, or punctuation). Thus, operationalizing data quality definitions that specifically address text data properties is an important step towards structuring a quality framework for text data. Second, a case study could be undertaken for the other text data quality indicator topics identified from this research (e.g., disambiguation of abbreviations). While there have been studies to disambiguate abbreviations(84) and correct spelling/typographical errors (42,85), it would also be beneficial to identify the impact of word variability on algorithm outputs in clinical text data. Third, a more focused scoping review about other text quality indicators could be conducted. In particular, quality indicators for short text documents such as social media posts, reviews, and EMRs, could be explored(24,75,101–103,125). Since EMRs are characterized by short texts, it would be interesting to examine other text quality indicators appropriate for these types of texts. Lastly, a documentation project that maps operational definitions for data quality dimensions to the data quality indicators for RCEHD should be explored similar to the work conducted by Weiskopf et al(32).

5.5 Conclusion

Quality of UTD, while important for ensuring the validity and generalizability of research based on information extracted from UTD, appears to be a topic that has received limited attention in research, particularly for EMRs. Data quality is important for understanding the suitability of data for a specific research purpose. Both the scoping review and case study provide useful information about data quality indicators that might be recommended for routine application to UTD. This research investigated some of the preprocessing methods appropriate for clinical encounter notes found in EMRs and reported on quality indicators identified from existing studies about UTD from multiple disciplines.

Measuring data quality is challenging because data quality is a multidimensional construct that is dependent on the context in which the data will be used. However, there are many similarities between the dimensions of data quality for structured and unstructured text. Many indicators require access to specialized gold standard datasets or dictionaries to assess quality. However, there are a few general-purpose indicators of data quality that do not require external data; most of these focus on the measurement of noise in the data. UTD from clinical encounter notes in EMRs are of generally poor quality; preprocessing is an essential step to improve quality of the data to benefit research about health and health services use.

References

1. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med [Internet]. 2015;12(10). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4595218/

2. Nicholls SG, Langan SM, Benchimol EI. Routinely collected data: the importance of high-quality diagnostic coding to research. Can Med Assoc J [Internet]. 2017;189(33):E1054–5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5566604/

3. Jorm L. Routinely collected data as a strategic resource for research: Priorities for methods and workforce. Public Health Res Pract [Internet]. 2015;25(4):e2541540. Available from: https://www.phrp.com.au/issues/september-2015-volume-25-issue-4/routinely-collected-data-as-a-strategic-resource-for-research-priorities-for-methods-and-workforce/

4. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc [Internet]. 2014;21(5):801–7. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4147606/

5. Taylor C. Structured vs. unstructured data [Internet]. 2018. Available from: https://www.datamation.com/big-data/structured-vs-unstructured-data.html

6. Luo L, Li L, Hu J, Wang X, Hou B, Zhang T, et al. A hybrid solution for extracting structured medical information from unstructured data in medical records via a double-reading/entry system. BMC Medical Inform Decis [Internet]. 2016 [cited 2020 Feb 3];16(1):114. Available from: https://bmcmedinformdecismak-biomedcentral-com.uml.idm.oclc.org/articles/10.1186/s12911-016-0357-5

7. Kristianson KJ, Ljunggren H, Gustafsson LL. Data extraction from a semi-structured electronic medical record system for outpatients: A model to facilitate the access and use of data for quality control and research. Health Informatics J [Internet]. 2009 [cited 2020 Feb 3];15(4):305–19. Available from: https://journals-sagepub-com.uml.idm.oclc.org/doi/10.1177/1460458209345889

8. Johnson SB, Bakken S, Dine D, Hyun S, Mendonça E, Morrison F, et al. An electronic health record based on structured narrative. J Am Med Inform Assoc [Internet]. 2008 [cited 2020 Feb 3];15(1):54–64. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2274868/

9. Pak H. Unstructured data in healthcare [Internet]. 2018. Available from: https://artificial-intelligence.healthcaretechoutlook.com/cxoinsights/unstructured-data-in-healthcare-nid-506.html

10. Smith M, Lix LM, Azimaee M, Enns JE, Orr J, Hong S, et al. Assessing the quality of administrative data for research: a framework from the Manitoba Centre for Health Policy. J Am Med Inform Assoc [Internet]. 2018 [cited 2018 Nov 18];25(3):224–9. Available from: http://academic.oup.com/jamia/article/25/3/224/4102320

11. Agboola Y, Camara Y, Brown A, Zhao H, Kendall O. The PHAC Data Quality Framework: A useful tool to serve surveillance programs. In Canada: Public Health Agency of Canada; 2011. Available from: https://www.researchgate.net/publication/305469008_The_PHAC_Data_Quality_Framework_A_Useful_Tool_to_Serve_Surveillance_Programs

12. CIHI. CIHI's information quality framework [Internet]. CIHI; 2017. Available from: https://www.cihi.ca/sites/default/files/document/iqf-summary-july-26-2017-en-web_0.pdf

13. Arts DGT, de Keizer NF, Scheffer G-J. Defining and improving data quality in medical registries: A literature review, case study, and generic framework. J Am Med Inform Assoc [Internet]. 2002;9(6):600–11. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC349377/

14. Kiefer C. Assessing the quality of unstructured data: An initial overview. In: LWDA [Internet]. 2016. p. 62–73. Available from: https://www.semanticscholar.org/paper/Assessing-the-Quality-of-Unstructured-Data%3A-An-Kiefer/a7d31b09498a16201f7044eb77ff14d27c1c559b

15. Azimaee M, Smith M, Lix L, Ostapyk T, Burchill C, Orr J. MCHP data quality framework [Internet]. Winnipeg Manitoba Canada University of Manitoba MCHP; 2018. Available from: http://umanitoba.ca/faculties/health_sciences/medicine/units/chs/departmental_units/mchp/protocol/media/Data_Quality_Framework.pdf

16. Government of Canada SC. Statistics Canada's quality assurance framework, 2017 [Internet]. Statistics Canada; 2017. Available from: https://www150.statcan.gc.ca/n1/pub/12-586-x/12-586-x2017001-eng.pdf

17. Julia Silge, David Robinson, Colin Fay, Emil Hvitfeldt, Kenneth Benoit, Kanishka, et al. juliasilge/tidytext: tidytext 0.2.2 [Internet]. Zenodo; 2019 [cited 2020 Jan 6]. Available from: https://zenodo.org/record/3355453

18. Griffiths TL, Steyvers M. Finding scientific topics. PNAS [Internet]. 2004 [cited 2021 Jan 8];101(suppl 1):5228–35. Available from: http://www.pnas.org/content/101/suppl_1/5228

19. Kee YH, Li C, Kong L, Jieyi Tang C, Chuang K-L. Scoping review of mindfulness research: a topic modelling approach. Mindfulness [Internet]. 2019;10:1474–88. Available from: https://link-springer-com.uml.idm.oclc.org/article/10.1007%2Fs12671-019-01136-4

20. Gharatkar S, Ingle A, Naik T, Save A. Review preprocessing using data cleaning and stemming technique. In: 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). 2017. p. 1–4. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/8276011

21. Famili A, Shen W-M, Weber R, Simoudis E. Data preprocessing and intelligent data analysis. Intell Data Anal [Internet]. 1997 [cited 2021 Feb 21];1(1):3–23. Available from: https://www.sciencedirect.com/science/article/pii/S1088467X98000079

22. Ritter A, Clark S, Etzioni O. Named entity recognition in tweets: an experimental study. In: Proceedings of the 2011 conference on empirical methods in natural language processing [Internet]. Association for Computational Linguistics; 2011. p. 1524–34. Available from: https://aclanthology.org/D11-1141

23. Zennaki O, Semmar N, Besacier L. Unsupervised and lightly supervised part-of-speech tagging using recurrent neural networks. In: 29th Pacific Asia Conference on Language, Information and Computation (PACLIC) [Internet]. Shangai, China; 2015 [cited 2021 Jul 18]. Available from: https://hal.archives-ouvertes.fr/hal-01350113

24. Alnajran N, Crockett K, McLean D, Latham A. A heuristic based pre-processing methodology for short text similarity measures in microblogs. In: Proc - Int Conf High Perform Comput Commun, Int Conf Smart City Int Conf Data Sci Syst, HPCC/SmartCity/DSS [Internet]. Institute of Electrical and Electronics Engineers Inc.; 2019. p. 1627–33. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/8623003

25. Madnick SE, Wang RY, Lee YW, Zhu H. Overview and framework for data and information quality research. ACM J Data Inf [Internet]. 2009 [cited 2019 Nov 17];1(1):22. Available from: http://doi.acm.org/10.1145/1515693.1516680

26. Kerr K. The development of a data quality framework and strategy for the New Zealand Ministry of Health. 2000;14. Available from: http://mitiq.mit.edu/Documents/IQ_Projects/Nov%202003/HINZ%20DQ%20Strategy%20paper.pdf

27. Hinds A, Lix LM, Smith M, Quan H, Sanmartin C. Quality of administrative health databases in Canada: A scoping review. Can J Public Health [Internet]. 2016;107(1):e56-61. Available from: https://link-springer-com.uml.idm.oclc.org/article/10.17269/cjph.107.5244

28. Thiru K, Hassey A, Sullivan F. Systematic review of scope and quality of electronic patient record data in primary care. Br Med J [Internet]. 2003 [cited 2019 Dec 16];326(7398):1070. Available from: http://www.bmj.com/content/326/7398/1070

29. Orfanidis L, Bamidis PD, Eaglestone B. Data quality issues in electronic health records: an adaptation framework for the greek health system. Health Informatics J [Internet]. 2004 [cited 2019 Nov 17];10(1):23–36. Available from: https://doi.org/10.1177/1460458204040665

30. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: A review of the literature. Med Care Res Rev [Internet]. 2010;67(5):503–27. Available from: http://journals.sagepub.com/doi/10.1177/1077558709359007

31. Mashoufi M, Ayatollahi H, Khorasani-Zavareh D. A review of data quality assessment in emergency medical services. Open Med Inform J [Internet]. 2018 [cited 2021 May 2];12:19–32. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5997849/

32. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc [Internet]. 2013;20(1):144–51. Available from: https://doi-org.uml.idm.oclc.org/10.1136/amiajnl-2011-000681

33. Bian J, Lyu T, Loiacono A, Viramontes TM, Lipori G, Guo Y, et al. Assessing the practice of data quality evaluation in a national clinical data research network through a systematic scoping review in the era of real-world data. J Am Med Inform Assoc [Internet]. 2020;27(12):1999–2010. Available from: https://doi-org.uml.idm.oclc.org/10.1093/jamia/ocaa245

34. Azimaee M, Kalappa G, Ji S, Hu M, Qian C. Datafit Toolkit - user guide - a practical approach for population data quality assessment [Internet]. ICES; 2019. Available from: https://raw.githubusercontent.com/icescentral/DataFit-Toolkit/master/DataFit%20Tool%20Kit%20-%20User%20Guide.pdf

35. Berndt DJ, McCart JA, Finch DK, Luther SL. A case study of data quality in text mining clinical progress notes. ACM Trans Manag Inf Syst [Internet]. 2015 [cited 2020 Jun 5];6(1):1:1-1:21. Available from: http://doi.org/10.1145/2669368

36. Sonntag D. Assessing the quality of natural language text data. In Gesellschaft für Informatik e.V.; 2004. p. 259–63. Available from: https://dl.gi.de/handle/20.500.12116/28866

37. Milios EE, Shafiei MM, Wang S, Zhang R, Tang B, Tougas J. A systematic study on document representation and dimensionality reduction for text clustering. Technical Report Technical Report CS-2006-05, Faculty of Computer Science [Internet]. 2006;29. Available from: https://www.cs.usask.ca/~spiteri/DR_Proj_ver04.pdf

38. Simon N. Statistical learning and biomedical data: bias variance tradeoff and generalization error [Internet]. 2020; Toronto. Available from: http://sorataba.org/sora-taba-annual-workshop/

39. Neal B. On the bias-variance tradeoff: textbooks need an update. arXiv:191208286 [cs, stat] [Internet]. 2019 [cited 2021 Sep 2]; Available from: http://arxiv.org/abs/1912.08286

40. Spasic I, Nenadic G. Clinical text data in machine learning: systematic review. JMIR Med Inform [Internet]. 2020;8(3):e17984. Available from: https://medinform.jmir.org/2020/3/e17984/

41. Taleb I, Serhani MA, Dssouli R. Big data quality assessment model for unstructured data. In: 2018 International Conference on Innovations in Information Technology (IIT) [Internet]. 2018. p. 69–74. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/8605945

42. Assale M, Dui LG, Cina A, Seveso A, Cabitza F. The revival of the notes field: leveraging the unstructured content in electronic health records. Front Med [Internet]. 2019;6:66. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2019.00066/full

43. Dařena F, Žižka J. Interdependence of text mining quality and the input data preprocessing. In: Silhavy R, Senkerik R, Oplatkova ZK, Prokopova Z, Silhavy P, editors. Artificial Intelligence Perspectives and Applications [Internet]. Cham: Springer International Publishing; 2015. p. 141–50. (Advances in Intelligent Systems and Computing). Available from: https://link-springer-com.uml.idm.oclc.org/chapter/10.1007/978-3-319-18476-0_15#citeas

44. Kiefer C. Quality indicators for text data. In: BTW 2019–Workshopband [Internet]. Gesellschaft für Informatik, Bonn; 2019. p. 145–54. Available from: https://dl.gi.de/handle/20.500.12116/21801

45. Grolemund G, Wickham H. R for data science [Internet]. [cited 2019 Dec 26]. Available from: https://r4ds.had.co.nz/

46. Yu G. Lexical diversity in writing and speaking task performances. Appl Linguist [Internet]. 2010;31(2):236–59. Available from: https://academic-oup-com.uml.idm.oclc.org/applij/article/31/2/236/178147

47. Koizumi R. Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens? Vocab learn instr [Internet]. 2012 [cited 2021 May 4];01(1):60–9. Available from: http://www.vli-journal.org/issues/01.1/issue01.1.10.html

48. Shaalan K, Khaled. Rule-based approach in Arabic natural language processing. Int J Inf Commun Technol [Internet]. 2010;3:11. Available from: https://www.semanticscholar.org/paper/Rule-based-Approach-in-Arabic-Natural-Language-Shaalan/0f7c3f51adaa6809d4e293b26463e5dec5159fa0

49. Santaholma M. Grammar sharing techniques for rule-based multilingual NLP systems. In: Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA 2007) [Internet]. University of Tartu, Estonia; 2007. p. 253–60. Available from: https://aclanthology.org/W07-2438/

50. Brill E. Transformation-based error-driven learning and natural language processing: a case study in part-of-speech tagging. Comput Linguist Assoc [Internet]. 1995;21(4):543–65. Available from: https://aclanthology.org/J95-4004

51. Wang T, Hirst G. Exploring patterns in dictionary definitions for synonym extraction. Nat Lang Eng [Internet]. 2012 Jul [cited 2021 Aug 31];18(3):313–42. Available from: http://www.cambridge.org/core/journals/natural-language-engineering/article/exploring-patterns-in-dictionary-definitions-for-synonym-extraction/245FA15B0DA0CEC0D0A68B13DC1BBA1A#

52. Nichols JA, Herbert Chan HW, Baker MAB. Machine learning: applications of artificial intelligence to imaging and diagnosis. Biophys Rev [Internet]. 2018 Sep 4 [cited 2021 Sep 1];11(1):111–8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6381354/

53. Jurafsky D, Martin J. Speech and Language Processing - an introduction to natural language processing, computational linguistics, and speech recognition [Internet]. 3rd Edition Draft. 2020. 623 p. Available from: https://web.stanford.edu/~jurafsky/slp3/

54. He B, Dong B, Guan Y, Yang J, Jiang Z, Yu Q, et al. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts. J Biomed Inform [Internet]. 2017 [cited 2020 Jun 5];69:203–17. Available from: http://www.sciencedirect.com/science/article/pii/S153204641730076X

55. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform [Internet]. 2014;47:1–10. Available from: https://www-sciencedirect-com.uml.idm.oclc.org/science/article/pii/S1532046413001974?via%3Dihub

56. Nguyen QT, Miyao Y, Le HTT, Nguyen NTH. Ensuring annotation consistency and accuracy for Vietnamese treebank. Lang Resour Eval [Internet]. 2018 [cited 2020 Jun 6];52(1):269–315. Available from: https://doi-org.uml.idm.oclc.org/10.1007/s10579-017-9398-3

57. Batini C, Scannapieco M. Information quality dimensions for maps and texts. In: Batini C, Scannapieco M, editors. Data and Information Quality: Dimensions, Principles and Techniques [Internet]. Cham: Springer International Publishing; 2016. p. 53–86. (Data-Centric Systems and Applications). Available from: https://doi.org/10.1007/978-3-319-24106-7_3

58. Hedman AS. Using the SMOG formula to revise a health-related document. Am J Health Educ [Internet]. 2008 [cited 2019 Dec 12];39(1):61–4. Available from: https://doi.org/10.1080/19325037.2008.10599016

59. McLaughlin GH. SMOG grading: A new readability formula. J Read [Internet]. 1969;12(8):639–46. Available from: https://ogg.osu.edu/media/documents/health_lit/WRRSMOG_Readability_Formula_G._Harry_McLaughlin__1969_.pdf

60. Crossley SA, Kyle K, McNamara DS. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. Behav Res Methods [Internet]. 2016 [cited 2019 Dec 25];48(4):1227–37. Available from: https://doi.org/10.3758/s13428-015-0651-7

61. Graesser A, McNamara D, Louwerse M, Cai Z. Coh-Metrix: Analysis of text on cohesion and language. Behav res meth instrum comput [Internet]. 2004;36:193–202. Available from: https://link-springer-com.uml.idm.oclc.org/article/10.3758/BF03195564

62. Dowell N. Analyzing language and discourse with Coh-Metrix [Internet]. 2014. Available from: https://www.academia.edu/7522203/Analyzing_Language_and_Discourse_with_Coh-Metrix

63. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. Extracting information from the text of electronic medical records to improve case detection: a systematic review. J Am Med Inform Assoc [Internet]. 2016 [cited 2019 May 21];23(5):1007–15. Available from: http://academic.oup.com/jamia/article/23/5/1007/2379833

64. Arksey H, O'Malley L. Scoping studies: towards a methodological framework. Int J Soc Res Methodol [Internet]. 2005;8(1):19–32. Available from: https://doi.org/10.1080/1364557032000119616

65. Levac D, Colquhoun H, O'Brien KK. Scoping studies: advancing the methodology. Implement Sci [Internet]. 2010;5:69. Available from: https://doi.org/10.1186/1748-5908-5-69

66. Higgins J, Thomas J, Chandler J, Cumpston M, Li T, Page M, et al. Cochrane Handbook for Systematic Reviews of Interventions version 6.2 [Internet]. 6.2. Cochrane, 2021; Available from: https://training.cochrane.org/handbook

67. Vijayarani M. Preprocessing techniques for text mining - an overview. Int J Comput Netw Commun [Internet]. 2015;5(1):7–16. Available from: https://www.researchgate.net/profile/Vijayarani-Mohan/publication/339529230_Preprocessing_Techniques_for_Text_Mining_-_An_Overview/links/5e57a0f7299bf1bdb83e7505/Preprocessing-Techniques-for-Text-Mining-An-Overview.pdf

68. Malak P. Text preprocessing: a tool of information visualization and digital humanities [Internet]. Information Visualization Techniques in the Social Sciences and Humanities. IGI Global; 2018 [cited 2020 Jun 6]. p. 86–104. Available from: www.igi-global.com/chapter/text-preprocessing/201305

69. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. Systematic Reviews [Internet]. 2016;5(1):210. Available from: http://dx.doi.org/10.1186/s13643-016-0384-4

70. Birtwhistle RV. Canadian Primary Care Sentinel Surveillance Network. Can Fam Physician [Internet]. 2011;57(10):1219–20. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3192094/

71. Canadian Primary Care Sentinel Surveillance Network (CPCSSN) [Internet]. [cited 2021 Jun 24]. Available from: https://cpcssn.ca/

72. MCHP. Glossery terms - De-identified information [Internet]. 2016 [cited 2021 Jun 24]. Available from: http://mchp-appserv.cpe.umanitoba.ca/viewDefinition.php?definitionID=104125

73. Severance CR, Blumenberg S, Hauser E. Python for everybody: exploring data in Python 3 [Internet]. North Charleston, SC, USA: CreateSpace Independent Publishing Platform; 2016. 242 p. Available from: https://www.py4e.com/book

74. Seltman HJ. Experimental design and analysis - chapter 4 exploratory data analysis. In: Experimental design and analysis [Internet]. 2018. Available from: https://www.stat.cmu.edu/~hseltman/309/Book/chapter4.pdf

75. Christen P, Gayler RW, Tran K-N, Fisher J, Vatsalan D. Automatic discovery of abnormal values in large textual databases. J Data Inf Qual [Internet]. 2016;7(1–2):1–31. Available from: http://doi.org/10.1145/2889311

76. Ooms J. High-Performance stemmer, tokenizer, and spell checker [Internet]. 2020 [cited 2021 Jun 25]. Available from: https://docs.ropensci.org/hunspell/

77. Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, et al. quanteda: An R package for the quantitative analysis of textual data. Journal of Open Source Software [Internet]. 2018 [cited 2020 Jan 5];3(30):774. Available from: https://joss.theoj.org/papers/10.21105/joss.00774

78. Liang C, Gong Y, Christian N., Kuziemsky C.E., Kushniruk A.W., Borycki E.M. Enhancing patient safety event reporting by K-nearest neighbor classifier. Stud Health Technol Informatics [Internet]. 2015;218:93–9. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84951923066&doi=10.3233%2f978-1-61499-574-6-93&partnerID=40&md5=9dd7ea3d83936147533292dec0f1447b

79. Rianto, Mutiara Achmad Benny, Wibowo Eri Prasetyo, Santosa PI. Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. J Big Data [Internet]. 2021;8(1):26. Available from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00413-1

80. Nguyen P-T, Le A-C, Ho T-B, Nguyen V-H. Vietnamese treebank construction and entropy-based error detection. Lang Resour Eval [Internet]. 2015 [cited 2020 Jun 6];49(3):487–519. Available from: https://doi-org.uml.idm.oclc.org/10.1007/s10579-015-9308-5

81. Xue N, Xia F, Chiou F, Palmer M. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. Nat Lang Eng [Internet]. 2005 Jun 1 [cited 2020 Jun 6];11(2):207–38. Available from: http://doi.org/10.1017/S135132490400364X

82. Westpfahl S, Schmidt T. FOLK-Gold — A Gold Standard for Part-of-Speech-Tagging of Spoken German. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) [Internet]. Portorož, Slovenia: European Language Resources Association (ELRA); 2016 [cited 2020 Jun 6]. p. 1493–9. Available from: https://www.aclweb.org/anthology/L16-1237

83. Gero Z, Ho J. PMCVec: Distributed phrase representation for biomedical text processing. J Biomed Inform [Internet]. 2019 [cited 2020 Jun 5];100:100047. Available from: http://www.sciencedirect.com/science/article/pii/S2590177X19300460

84. Joopudi V, Dandala B, Devarakonda M. A convolutional route to abbreviation disambiguation in clinical text. J Biomed Inform. 2018;86:71–8. Available from: https://www-sciencedirect-com.uml.idm.oclc.org/science/article/pii/S1532046418301552

85. Lai KH, Topaz M, Goss FR, Zhou L. Automated misspelling detection and correction in clinical free-text records. J Biomed Inform [Internet]. 2015 [cited 2019 Aug 3];55:188–95. Available from: http://www.sciencedirect.com/science/article/pii/S1532046415000751

86. Pantazos K, Lauesen S, Lippert S. Preserving medical correctness, readability and consistency in de-identified health records. Health Informatics J [Internet]. 2017 Dec 1 [cited 2020 Jun 6];23(4):291–303. Available from: https://doi.org/10.1177/1460458216647760

87. Ruch P, Baud R, Geissbühler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artificial Intelligence in Medicine [Internet]. 2003 Sep 1 [cited 2020 Jun 6];29(1):169–84. Available from: http://www.sciencedirect.com/science/article/pii/S0933365703000526

88. Hoste V, Hendrickx I, Daelemans W, Bosch AVD. Parameter optimization for machine-learning of word sense disambiguation. Natural Language Engineering [Internet]. 2002 Dec [cited 2020 Jun 6];8(4):311–25. Available from: http://www.cambridge.org/core/journals/natural-language-engineering/article/parameter-optimization-for-machinelearning-of-word-sense-disambiguation/ED5DAA7F32FEFBC5D1C81B22DEA5F43B#

89. Qin kan, Yujiu Yang, Wenhuang Liu, Xiaodong Liu. An integrated approach for detecting approximate duplicate records. In: 2009 Asia-Pacific Conference on Computational Intelligence and Industrial Applications (PACIIA). 2009. p. 381–4. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/5406409

90. Uejima H, Miura T, Shioya I. Improving text categorization by resolving semantic ambiguity. In: 2003 IEEE Pacific Rim Conference on Communications Computers and Signal Processing (PACRIM 2003) (Cat No03CH37490) [Internet]. IEEE; 2003. p. 796–9. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/abstract/document/1235901

91. Zurini M. Word sense disambiguation using aggregated similarity based on WordNet graph representation. Inform econ [Internet]. 2013 [cited 2020 Jun 6];17(3):169–80. Available from: https://ideas.repec.org/a/aes/infoec/v17y2013i3p169-180.html

92. Šnajder J. DerivBase.hr: A high-coverage derivational morphology resource for croatian. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) [Internet]. Reykjavik, Iceland: European Language Resources Association (ELRA); 2014 [cited 2020 Jun 6]. p. 3371–7. Available from: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1090_Paper.pdf

93. Edwards B, Zatorsky M, Nayak R. Clustering and classification of maintenance logs using text data mining. In: Conf Res Pract Inf Technol Ser [Internet]. CRC for Integrated Engineering Asset Management, Faculty of Information Technology, Queensland University of Technology, PO Box 2434, Brisbane 4001, QLD, Australia: AusDM; 2008. p. 193–9. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84870484659&partnerID=40&md5=f34d58eba076a357439884fae90dc336

94. Marzec M, Uhl T, Michalak D. Verification of text mining techniques accuracy when dealing with urban buses maintenance data. Diagnostyka [Internet]. 2017 Dec 20 [cited 2020 Jun 6];15(3):51–7. Available from: http://www.diagnostyka.net.pl/Verification-of-text-mining-techniques-accuracy-when-dealing-with-urban-buses-maintenance,81413,0,2.html

95. Stewart M, Liu W, Cardell-Oliver R, Wang R. Short-text lexical normalisation on industrial log data. In: 2018 IEEE International Conference on Big Knowledge (ICBK). 2018. p. 113–22. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/8588782

96. Subha R, Palaniswami S. Quality factor assessment and text summarization of unambiguous natural language requirements. In: Unnikrishnan S, Surve S, Bhoir D, editors. Advances in Computing, Communication, and Control. Berlin, Heidelberg: Springer; 2013. p. 131–46. Available from: https://doi-org.uml.idm.oclc.org/10.1007/978-3-642-36321-4_12

97. Jung Y. A semantic annotation framework for scientific publications. Qual Quant [Internet]. 2017 [cited 2020 Jun 6];51(3):1009–25. Available from: https://doi.org/10.1007/s11135-016-0369-3

98. Kang N, van Mulligen EM, Kors JA. Comparing and combining chunkers of biomedical text. Journal of Biomedical Informatics [Internet]. 2011 Apr 1 [cited 2020 Jun 6];44(2):354–60. Available from: http://www.sciencedirect.com/science/article/pii/S1532046410001577

99. Tissot H, Del Fabro MD, Derczynski L, Roberts A. Normalisation of imprecise temporal expressions extracted from text. Knowl Inf Syst [Internet]. 2019 Dec 1 [cited 2020 Jun 6];61(3):1361–94. Available from: https://doi.org/10.1007/s10115-019-01338-1

100. Vítovec P, Kléma J. Gene interaction extraction from biomedical texts by sentence skeletonization. CEUR Workshop Proc [Internet]. 2011;802:NA. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84891772107&partnerID=40&md5=4c64e3778af84417d329e274327ffff7

101. Allen TT, Sui Z, Akbari K. Exploratory text data analysis for quality hypothesis generation. Qual Eng [Internet]. 2018 [cited 2020 Jun 5];30(4):701–12. Available from: https://doi.org/10.1080/08982112.2018.1481216

102. Chen CC, Tseng Y-D. Quality evaluation of product reviews using an information quality framework. Decis Support Syst [Internet]. 2011 [cited 2020 Jun 5];50(4):755–68. Available from: http://www.sciencedirect.com/science/article/pii/S0167923610001478

103. Yang H-L, Chao AFY. Sentiment annotations for reviews: an information quality perspective. Online Inf Rev [Internet]. 2018 [cited 2020 Jun 5];42(5):579–94. Available from: https://doi.org/10.1108/OIR-04-2017-0114

104. Scharkow M. Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. Qual Quant [Internet]. 2013 [cited 2020 Jun 6];47(2):761–73. Available from: https://doi.org/10.1007/s11135-011-9545-7

105. Lauría EJM, March AD. Effect of dirty data on free text discharge diagnoses used for automated ICD-9-CM coding. In: AMCIS [Internet]. 2006. Available from: https://aisel.aisnet.org/amcis2006/188

106. Abad ZSH, Karras O, Ghazi P, Glinz M, Ruhe G, Schneider K. What works better? A study of classifying requirements. In: 2017 IEEE 25th International Requirements Engineering Conference [Internet]. 2017. p. 496–501. Available from: https://arxiv.org/abs/1707.02358

107. Briesch D, Hobbs R, Jaja C, Kjersten B, Voss C. Training and evaluating a statistical part of speech tagger for natural language applications using Kepler Workflows. Procedia Comput Sci [Internet]. 2012 [cited 2020 Jun 5];9:1588–94. Available from: http://www.sciencedirect.com/science/article/pii/S1877050912002955

108. HaCohen-Kerner Y, Miller D, Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation. PLOS ONE [Internet]. 2020 May 1 [cited 2021 Feb 21];15(5):e0232525. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232525

109. Song JW, Chung KC. Observational studies: cohort and case-control studies. Plast Reconstr Surg [Internet]. 2010 Dec [cited 2019 Nov 5];126(6):2234–42. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2998589/

110. Moreno I, Boldrini E, Moreda P, Romá-Ferri MT. DrugSemantics: A corpus for named entity recognition in Spanish summaries of product characteristics. J Biomed Inform [Internet]. 2017 [cited 2020 Jun 6];72:8–22. Available from: https://doi.org/10.1016/j.jbi.2017.06.013

111. Song B, Wu P, Zhang Q, Chai B, Gao Y, Yang H. Intelligent assessment of 95598 speech transcription text quality based on topic model. IOP Conf Ser: Mater Sci Eng [Internet]. 2019 [cited 2020 Jun 6];563(4):2001. Available from: https://doi.org/10.1088%2F1757-899x%2F563%2F4%2F042001

112. Lauría EJM, March AD. Combining bayesian text classification and shrinkage to automate healthcare coding: A data quality analysis. ACM J Data Inf Qual [Internet]. 2011 [cited 2020 Jun 6];2(3):13:1-13:22. Available from: http://doi.org/10.1145/2063504.2063506

113. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D. The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations [Internet]. Baltimore, Maryland: Association for Computational Linguistics; 2014 [cited 2021 Mar 18]. p. 55–60. Available from: http://aclweb.org/anthology/P14-5010

114. Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03 [Internet]. Edmonton, Canada: Association for Computational Linguistics; 2003 [cited 2021 Mar 18]. p. 173–80. Available from: http://portal.acm.org/citation.cfm?doid=1073445.1073478

115. Bird S, Klien E, Loper E. Natural language processing with Python [Internet]. 2009 [cited 2021 Mar 19]. Available from: https://www.nltk.org/book/

116. Montani I, Honnibal M, Van Landeghem S, Boyd A. spaCy: industrial-strength natural language processing in python [Internet]. Zenodo; 2021 [cited 2021 Mar 19]. Available from: https://zenodo.org/record/4593273

117. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res [Internet]. 2011 [cited 2021 Mar 19];12(85):2825–30. Available from: http://jmlr.org/papers/v12/pedregosa11a.html

118. Feinerer I, Hornik K, Meyer D. Text mining infrastructure in R. J Stat Softw [Internet]. 2008 [cited 2021 Mar 19];25(1):1–54. Available from: https://www.jstatsoft.org/index.php/jss/article/view/v025i05

119. Kuhn M. Building predictive models in R using the caret package. J Stat Softw [Internet]. 2008 [cited 2020 Jun 11];28(1):1–26. Available from: https://www.jstatsoft.org/index.php/jss/article/view/v028i05

120. Strong DM. Information quality: managing information as a product. In: LIU L, ÖZSU MT, editors. Encyclopedia of Database Systems [Internet]. Boston, MA: Springer US; 2009 [cited 2021 Jun 12]. p. 1502–8. Available from: https://doi.org/10.1007/978-0-387-39940-9_497

121. Government of Canada. Medical Abbreviations [Internet]. 2008 [cited 2021 Aug 28]. Available from: https://www.canada.ca/en/indigenous-services-canada/services/first-nations-inuit-health/health-care-services/nursing/clinical-practice-guidelines-nurses-primary-care/adult-care/abbreviations.html

122. Katz A, Kinew KA, Star L, Taylor C, Koseva I, Lavoie J, et al. The health status of and access to healthcare by registered First Nation peoples in Manitoba [Internet]. 2019. Available from: http://mchp-appserv.cpe.umanitoba.ca/reference/FN_Report_web.pdf

123. Kobayashi VB, Mol ST, Berkers HA, Kismihók G, Den Hartog DN. Text mining in organizational research. Organ Res Methods [Internet]. 2018 [cited 2018 Nov 21];21(3):733–65. Available from: https://doi.org/10.1177/1094428117722619

124. Batini C, Scannapieco M. Data Quality -- concepts, methodologies and techniques [Internet]. Springer, Berlin, Heidelberg; 2006. Available from: https://link-springer-com.uml.idm.oclc.org/book/10.1007%2F3-540-33173-5

125. Pantazos K, Lauesen S, Lippert S. De-identifying an EHR database-Anonymity, correctness and readability of the medical record [Internet]. Stud. Health Technol. Informatics. IOS Press; 2011. 862–866 p. (Health Technol. Informatics; vol. 169). Available from: https://ebooks.iospress.nl/publication/14293

Appendix A – Scoping review data extraction form

Obj 1 - Data Extraction Form

The data extraction form will extract data on the following topics:

1) Characteristics of the article
2) Characteristics of the text data that was proprocessed
3) Characteristics of any methods used to restructure/reorganize the text data
4) Characteristics of any techniques used to validate the text data
5) Characteristics of the software used to assess data quality.

1. Name of Data Extractor

   Mark only one oval.

   - ( ) Marcello Nesca
   - ( ) Dr Lisa Lix

1) Characteristics of the article

Please make sure to fill every item. If not applicable on text based questions please write "N/A"

2. Article Title

   ____

3. First Author

   Last name only

   ____

4. Publication Year

   ____

5. Journal Name

Full journal names - no abbreviations please. This is located in the "Publication" field in Zotero

6. What type of article was this?

Mark only one oval.

- empirical research
- case study
- review article
- Other:

7. What approach(es) was used to pre-process the text data in the article:

Only relevant if empirical research or case study was used

Mark only one oval per row.

	Yes	No
Restructuring or Reorganizing text data		
Validating the text data		
Neither approach was used		

8. If neither approach was used, what approach was adopted?

For example, If an unsupervised method was used, describe it here.

9. Were any of the following data quality dimensions described in the article?

Mark only one oval per row.

	Yes	No
Accessibility	◯	◯
Usability	◯	◯
Accuracy	◯	◯
Timeliness	◯	◯
Relevance	◯	◯
Completeness	◯	◯
Validity	◯	◯
Comparability	◯	◯
Readability	◯	◯

10. Disciplinary Area

Which disciplinary area is the article focused on?

Mark only one oval.

- ◯ Epidemiology Population Health or Health Services
- ◯ Basic Clinical Sciences
- ◯ Computer Sciences
- ◯ Engineering (e.g., Computer Engineering)
- ◯ Social Sciences (e.g., Sociology, Psychology, Economics)
- ◯ Business (e.g., Marketing, Finance)
- ◯ Humanities (e.g., Linguistics, English, Classics)
- ◯ Other: ______________________

2) Characteristics of the text data

Text data properties include: size, annotation, provenance

Please make sure to fill every item. If not applicable on text based questions please write "N/A"

11. Type of Text Data

Mark only one oval per row.

	Yes	No
Electronic Medical Records	⬭	⬭
Administrative Data	⬭	⬭
HTML/Web Data	⬭	⬭
PDFs	⬭	⬭
Social Media Data (E.g., Tweets, Facebook posts, etc)	⬭	⬭
Product Review Data (e.g., Amazon reviews, Google reviews, etc)	⬭	⬭
Lexical Data (e.g., Prose, treebanks, industry articles)	⬭	⬭

12. If another type of text data was described in this article, please list it here.

____

13. Are the following elements regarding the size of the text data described?

Mark only one oval per row.

	Yes	No
Number of words	◯	◯
Number of phrases	◯	◯
Number of records (rows)	◯	◯
Number of documents	◯	◯
Number of features (columns, variables)	◯	◯

14. if yes to number of Words how many words?

____

15. if yes to number of Phrases how many phrases?

____

16. if yes to number of Records (or rows) how many records?

____

17. if yes to number of Documents how many documents?

____

18. if yes to number of Features (columns, variables) how many features?

____

19. Was the text data annotated?

An annotation is a note, explanation, or tag added to text to enhance meaning (e.g., part of speech tagging where each word is tagged as noun, verb, adjective, etc). Text data can also be annotated with medical codes (e.g., ICD codes), please see reference below.

Mark only one oval.

- ( ) Yes
- ( ) No

20. If Yes, what level of annotation automation was adopted?

Manual annotation: labour intensive and involves subject matter experts to manually tag specific text elements such as words, documents, phrases. (e.g., a doctor tagging patients text notes in electronic medical records, as having a diagnosis in PTSD or not) Automatic annotation: No labour involved in tagging specific text elements. Automatic annotation can reduce time and financial costs, at the expense of accuracy.

Mark only one oval per row.

	Yes	No
Manual	( )	( )
Automatic	( )	( )

21. If the annotation was automatic, how was it implemented and used?

____

22. Is the text data publically available?

If Yes, do a Google search for the dataset

Mark only one oval.

- Yes
- No

23. If yes to above, what was the source(s)?

Report the URL

____

3) Characteristics of restructuring/reorganizing methods for text data

Please make sure to fill every item. If not applicable on text based questions please write "N/A"

24. What methods were used to restructure/reorganize the data?

Mark only one oval per row.

	Yes	No
Stop word removal	⬭	⬭
Word stemming	⬭	⬭
Lemmatization	⬭	⬭
Removal of punctuation	⬭	⬭
Text segmentation	⬭	⬭
Lowercase conversion	⬭	⬭
Removal of spelling errors	⬭	⬭
Expansion of abbreviations	⬭	⬭
Expansion of contractions	⬭	⬭
Tokenization	⬭	⬭
HTML tag conversion	⬭	⬭
Part of speech tagging	⬭	⬭
Removal of URLs	⬭	⬭
Removal of numbers	⬭	⬭

Hashtag/tag segmentation		
Term frequency - Inverse document frequency TF-IDF (assign weights to text)	⬭	⬭

25. If another method was used to restructure/reorganize the data, please describe it below.

____

26. Are quantitative measures used to describe the results of text pre-processing to restructure or reorganize the data?

Mark only one oval.

⬭ Yes

⬭ No

27. If yes, then what measure(s) were used to describe the results of text pre-processing to restructure or reorganize the data?

____

4) Characteristics of data validation techniques

Please make sure to fill every item. If not applicable on text based questions please write "N/A"

28. Was one or more text dataset features validated?

Mark only one oval.

- Yes
- No

29. If Yes, what text dataset feature(s) was validated?

____

30. Was a text dataset method validated?

Mark only one oval.

- Yes
- No

31. What supervised methods were used to validate the data

Mark only one oval per row.

	Yes	No
Support Vector Machine (SVM)	◯	◯
Statistical Text Mining (STM)	◯	◯
Latent Semantic Analysis (LSM)	◯	◯
Singular Value Decomposition (SVD)	◯	◯
Convolutional Neural Network (CNN)	◯	◯
Hidden Markov Model (HMM)	◯	◯
Conditional Random Fields (CRF)	◯	◯
Naive Bayes	◯	◯

32. If another method was used to validate the text data, please describe it below.

33. Were any of the restructuring/reorganizing methods or validation methods applied to subgroups or subsets of the data?

Mark only one oval.

- Yes
- No

34. If yes, what subgroups or subsections were analyzed?

35. How were the supervised methods evaluated?

Mark only one oval per row.

	Yes	No
Sensitivity/Recall/True Positive Rate	◯	◯
Specificity/Selectivity/True Negative Rate	◯	◯
Precision/Positive Predictive Value	◯	◯
Negative Predictive Value	◯	◯
Accuracy	◯	◯
F1 Score/F-Measure/F-Score (Harmonic Mean of PPV and Sensitivity)	◯	◯
Youden Index/J (combines specificity and sensitivity)	◯	◯
Area under the ROC curve	◯	◯
Mean Absolute Error / Mean Squared Error / Root Mean Squared Error	◯	◯

36. If another evaluation measure was used, please describe it here.

____

37. What reference database was used to validate the text data?

This item refers to the use of reference or gold standard corpora used in text data validation. List the name of the reference database / gold standard corpus. If no answer type "N/A". If a URL is available for the reference database, add it here.

____

5) Characteristics of the software.

Please make sure to fill every item. If not applicable on text based questions please write "N/A"

38. What software was used to restructure/reorganize and/or validate the text data?

Mark only one oval per row.

	Yes	No
R	◯	◯
Python	◯	◯
User Defined	◯	◯
No software stated	◯	◯

39. If user defined software was adopted, what was it used to do?

Complete this question if "User Defined" is checked as "Yes" in previous question otherwise "N/A"

____

40. If R or Python was used, what packages were used?

Mark only one oval per row.

	Yes	No
Keras - python/R	◯	◯
Tensorflow - python/R	◯	◯
Numpy - python	◯	◯
Scipy - python	◯	◯
Theano - python	◯	◯
PyTorch - python	◯	◯
Pandas - python	◯	◯
Matplotlib - python	◯	◯
NLTK - python	◯	◯
TextBlob - python	◯	◯
Stanford CoreNLP - python	◯	◯
spaCy - python	◯	◯
gensim - python	◯	◯
tidytext - R	◯	◯
openNLP - R	◯	◯

tm - R	⬭	⬭
quanteda - R	⬭	⬭
text2vec - R	⬭	⬭
qdap - R	⬭	⬭
textstem - R (for lemmatization)	⬭	⬭

41. If another package was used, please list it here.

if its in R or Python, please put a "- R" or " - Python" at the end of the package

____

Appendix B – Data dictionary for the scoping review data extraction form

Variable Name	Type of Variable	Instructions for Data Extraction
Article Title	Free Text	Fill out.
First Author	Free Text	Write the name of the first author last name only
Publication Year	Numeric	Fill out.
Journal Name	Free Text	Write the full journal name - no abbreviations please. This is located in the "Publication" field in Zotero
What type of article was this?	Categorical	**Empirical research article** – defines a research question, study population, behaviour or phenomena being studied, and description of method/process. **Case study article** – an intensive study about a person, a group of people, or a unit, which is aimed to generalize over several units. **Review article** – includes literature reviews, syntheses, and overviews of the current state of research on a given topic. **Other** – describe using free text if the article does not fall into one of the categories listed
What approach(es) was used to pre-process the text data in the article?	Categorical	**Restructuring or reorganizing text data** – methods applied to raw text to standardize or manipulate the data into a format that is suitable for research **Validating text data** – methods applied to verify text data against a reference standard dataset **Neither approach was used** NOTE: both restructuring or reorganizing AND validating responses can be selected for a single article. If neither approach was used select "Yes" for that option and select "No" for both restructuring or reorganizing text data and validating text data
If neither approach was used, what approach was adopted?	Free Text	If an unsupervised method was used, describe it here.
Were any of the following data quality dimensions described in the article?	Categorical	Check "yes" for all dimensions that are explicitly mentioned in the article. Use the search command in the PDF (Ctrl + F) and type the data quality dimension to help search.
Disciplinary Area	Categorical	What disciplinary area is best represented by the topic of the research or review article?

Type of Text Data	Categorical	**Electronic Medical Records** – text data from a computer-based patient record. Usually from a primary care setting. **Administrative Data** – text data that are collected by governments, agencies, or organizations (e.g., health, justice, social) for management or administrative functions **HTML/Web Data** – text data scraped from web pages **PDFs** – text data from a specific file format to present documents (i.e., *.pdf) **Social Media Data** – text data that comes from social media websites such as Facebook, Twitter, Instagram, etc. **Product Review Data** – text data that comes from product reviews such as Amazon.ca, Google reviews, or other online retail stores **Lexical Data -** text data in the form of prose (e.g., books), treebanks (e.g., news articles), or industry articles (e.g., organizational documents)
If another type of text data was described in this article, please list it here.	Free Text	If no other type of text data is described in the article, type "N/A"
Are the following elements regarding the size of the text data described?	Categorical	**Words** – individual words **Phrase** – a sentence or part of a sentence (e.g., cannot include a number, must be more than 1 word (e.g., a Tweet) **Records (rows)** – rows or units of a data frame **Documents** – a written, printed, or electronic paper furnishing information or evidence, such as a passport, deed, bill of sale, or bill of lading; a legal or official paper: **Features** –variables or columns of a data frame
if yes to number of words how many words?	Free Text (restricted to numbers only)	Fill out, if the article does not apply in this context, type "N/A"
if yes to number of phrases how many phrases?	Free Text (restricted to numbers only)	Fill out, if the article does not apply in this context, type "N/A"
if yes to number of records (or rows) how many records?	Free Text (restricted to numbers only)	Fill out, if the article does not apply in this context, type "N/A"

if yes to number of documents how many documents?	Free Text (restricted to numbers only)	Fill out, if the article does not apply in this context, type “N/A”
if yes to number of features (columns, variables) how many features?	Free Text (restricted to numbers only)	Fill out, if the article does not apply in this context, type “N/A”
Was the text data annotated?	Categorical	Yes/No An annotation is a note, explanation, or tag added to text to enhance meaning (e.g., part of speech tagging where each word is tagged as noun, verb, adjective, etc)
If Yes, what level of annotation automation was adopted?	Categorical	**Manual annotation** - labour intensive and involves subject matter experts to manually tag specific text elements such as words, documents, phrases. (e.g., a doctor tagging patients text notes in electronic medical records, as having a diagnosis in PTSD or not) **Automatic annotation** – No labour involved in tagging specific text elements. Automatic annotation can reduce time and financial costs, at the expense of accuracy. NOTE: Both Manual and Automatic annotation can be selected as an option
If the annotation was automatic, how was it implemented and used?	Free Text	Describe where automation was used and how was it used. If automation was manual, put N/A in this space.
Is the text data publicly available?	Categorical	Fill out as yes (i.e., publically available), no (i.e., not publically available)
If yes to above, what was the source(s)?	Free Text	Report URL for the source
What methods were used to restructure and/or reorganize the data?	Categorical	Fill out, choose an option, multiple methods can be selected.
If another method was used to restructure/reorganize the data, please describe it below.	Free Text	Fill out, if the article does not apply in this context, type “N/A”
Are quantitative measures used to describe the results of methods applied to restructure or reorganize the data?	Categorical	Fill out, yes/no

If yes, then what measure(s) were used to describe the results of text pre-processing to restructure or reorganize the data?	Free Text	Fill out, if the article does not apply in this context, type “N/A”
Was one or more text dataset features validated?	Categorical	Fill out, yes/no
If Yes, what text dataset feature(s) was validated?	Free Text	List the feature that was validated and how the validation was done
Was a text dataset method validated?	Categorical	Fill out, yes/no
What supervised methods were used to characterize or validate the data	Categorical	Fill out, choose yes or no for each method
If another supervised method was used to characterize or validate the text data, please describe it below.	Free Text	Fill out, if the article does not apply in this context, type “N/A”
Were any of the restructuring/reorganizing methods, validation methods, or descriptive methods applied to subgroups or subsets of the data?	Categorical	Fill out, yes/no
If yes, what subgroups or subsections were analyzed?	Free Text	Fill out, if the article does not apply in this context, type “N/A”
How were the supervised methods evaluated?	Categorical	Fill out, choose either yes or no for each option according to the article
If another evaluation measure was used, please describe it here.	Free Text	Fill out, if the article does not apply in this context, type “N/A”
What reference database was used to validate the text data?	Free Text	List the name of the reference database / gold standard corpus. If there is no reference database or gold standard, please type “N/A”. If a URL is available for the reference database, add it here.
What software was used to restructure/reorganize and/or validate the text data?	Categorical	Fill out, choose Yes/No for each type of software. The categories are not mutually exclusive.

If user defined software was adopted, what was it used to do?	Free Text	Complete this question if “User Defined” is checked as “Yes” in previous question otherwise “N/A”
If R or Python was used, what packages were used?	Categorical	Fill out; choose either yes or no for each listed package
If another package was used, please list it here.	Free Text	Fill out, if the article does not apply in this context, type “N/A”

Appendix C – Published articles selected for the scoping review

1. Abad ZSH, Karras O, Ghazi P, Glinz M, Ruhe G, Schneider K. What works better? A study of classifying requirements. In: 2017 IEEE 25th International Requirements Engineering Conference [Internet]. 2017. p. 496–501. Available from: https://arxiv.org/abs/1707.02358

2. Allen TT, Sui Z, Akbari K. Exploratory text data analysis for quality hypothesis generation. Qual Eng [Internet]. 2018 [cited 2020 Jun 5];30(4):701–12. Available from: https://doi.org/10.1080/08982112.2018.1481216

3. Alnajran N, Crockett K, McLean D, Latham A. A heuristic based pre-processing methodology for short text similarity measures in microblogs. In: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS) [Internet]. 2018. p. 1627–33. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/8623003

4. Assale M, Dui LG, Cina A, Seveso A, Cabitza F. The revival of the notes field: leveraging the unstructured content in electronic health records. Front Med [Internet]. 2019;6:66. Available from: https://www.frontiersin.org/articles/10.3389/fmed.2019.00066/full

5. Berndt DJ, McCart JA, Finch DK, Luther SL. A case study of data quality in text mining clinical progress notes. ACM Trans Manag Inf Syst [Internet]. 2015 [cited 2020 Jun 5];6(1):1:1-1:21. Available from: http://doi.org/10.1145/2669368

6. Briesch D, Hobbs R, Jaja C, Kjersten B, Voss C. Training and evaluating a statistical part of speech tagger for natural language applications using Kepler Workflows. Procedia Comput Sci [Internet]. 2012 [cited 2020 Jun 5];9:1588–94. Available from: http://www.sciencedirect.com/science/article/pii/S1877050912002955

7. Chen CC, Tseng Y-D. Quality evaluation of product reviews using an information quality framework. Decis Support Syst [Internet]. 2011 [cited 2020 Jun 5];50(4):755–68. Available from: http://www.sciencedirect.com/science/article/pii/S0167923610001478

8. Christen P, Gayler RW, Tran K-N, Fisher J, Vatsalan D. Automatic discovery of abnormal values in large textual databases. J Data Inf Qual [Internet]. 2016 [cited 2020 Jun 5];7(1–2):1–31. Available from: http://doi.org/10.1145/2889311

9. Doğan RI, Leaman R, Lu Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. J Biomed Inform [Internet]. 2014;47:1–10. Available from: https://www-sciencedirect-com.uml.idm.oclc.org/science/article/pii/S1532046413001974?via%3Dihub

10. Edwards B, Zatorsky M, Nayak R. Clustering and classification of maintenance logs using text data mining. In: Conf Res Pract Inf Technol Ser [Internet]. CRC for Integrated Engineering Asset Management, Faculty of Information Technology, Queensland University of Technology, PO Box 2434, Brisbane 4001, QLD, Australia: AusDM; 2008. p. 193–9. Available from: https://www.scopus.com/inward/record.uri?eid=2-s2.0-84870484659&partnerID=40&md5=f34d58eba076a357439884fae90dc336

11. Gero Z, Ho J. PMCVec: Distributed phrase representation for biomedical text processing. J Biomed Inform [Internet]. 2019 [cited 2020 Jun 5];100:100047. Available from: http://www.sciencedirect.com/science/article/pii/S2590177X19300460

12. Gharatkar S, Ingle A, Naik T, Save A. Review preprocessing using data cleaning and stemming technique. In: 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS) [Internet]. 2017. p. 1–4. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/8276011

13. HaCohen-Kerner Y, Miller D, Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation. PLoS One [Internet]. 2020 [cited 2021 May 23];15(5). Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0232525

14. He B, Dong B, Guan Y, Yang J, Jiang Z, Yu Q, et al. Building a comprehensive syntactic and semantic corpus of Chinese clinical texts. J Biomed Inform [Internet]. 2017 [cited 2020 Jun 5];69:203–17. Available from: http://www.sciencedirect.com/science/article/pii/S153204641730076X

15. Hoste V, Hendrickx I, Daelemans W, Bosch AVD. Parameter optimization for machine-learning of word sense disambiguation. Nat Lang Eng [Internet]. 2002 [cited 2020 Jun 6];8(4):311–25. Available from: http://www.cambridge.org/core/journals/natural-language-engineering/article/parameter-optimization-for-machinelearning-of-word-sense-disambiguation/ED5DAA7F32FEFBC5D1C81B22DEA5F43B#

16. Joopudi V, Dandala B, Devarakonda M. A convolutional route to abbreviation disambiguation in clinical text. J Biomed Inform [Internet]. 2018;86:71–8. Available from: https://www-sciencedirect-com.uml.idm.oclc.org/science/article/pii/S1532046418301552

17. Jung Y. A semantic annotation framework for scientific publications. Qual Quant [Internet]. 2017 [cited 2020 Jun 6];51(3):1009–25. Available from: https://doi.org/10.1007/s11135-016-0369-3

18. Kang N, van Mulligen EM, Kors JA. Comparing and combining chunkers of biomedical text. J Biomed Inform [Internet]. 2011 [cited 2020 Jun 6];44(2):354–60. Available from: http://www.sciencedirect.com/science/article/pii/S1532046410001577

19. Lai KH, Topaz M, Goss FR, Zhou L. Automated misspelling detection and correction in clinical free-text records. J Biomed Inform [Internet]. 2015 [cited 2020 Jun 6];55:188–95. Available from: http://www.sciencedirect.com/science/article/pii/S1532046415000751

20. Lauría EJM, March AD. Effect of dirty data on free text discharge diagnoses used for automated ICD-9-CM coding. In: AMCIS [Internet]. 2006. Available from: https://aisel.aisnet.org/amcis2006/188

21. Liang C, Gong Y. Enhancing patient safety event reporting by K-nearest neighbor classifier. In: Stud health Technol Inform [Internet]. University of Texas Health Science Center at Houston, United States: IOS Press; 2015. p. 93–9. Available from: https://ebooks.iospress.nl/publication/40603

22. Marzec M, Uhl T, Michalak D. Verification of text mining techniques accuracy when dealing with urban buses maintenance data. Diagnostyka [Internet]. 2017 [cited 2020 Jun 6];15(3):51–7. Available from: http://www.diagnostyka.net.pl/Verification-of-text-mining-techniques-accuracy-when-dealing-with-urban-buses-maintenance,81413,0,2.html

23. Moreno I, Boldrini E, Moreda P, Romá-Ferri MT. DrugSemantics: A corpus for named entity recognition in Spanish summaries of product characteristics. J Biomed Inform [Internet]. 2017 [cited 2020 Jun 6];72:8–22. Available from: https://doi.org/10.1016/j.jbi.2017.06.013

24. Nguyen P-T, Le A-C, Ho T-B, Nguyen V-H. Vietnamese treebank construction and entropy-based error detection. Lang Resour Eval [Internet]. 2015 [cited 2020 Jun 6];49(3):487–519. Available from: https://doi-org.uml.idm.oclc.org/10.1007/s10579-015-9308-5

25. Nguyen QT, Miyao Y, Le HTT, Nguyen NTH. Ensuring annotation consistency and accuracy for Vietnamese treebank. Lang Resour Eval [Internet]. 2018 [cited 2020 Jun 6];52(1):269–315. Available from: https://doi-org.uml.idm.oclc.org/10.1007/s10579-017-9398-3

26. Pantazos K, Lauesen S, Lippert S. Preserving medical correctness, readability and consistency in de-identified health records. Health Informatics J [Internet]. 2017 [cited 2020 Jun 6];23(4):291–303. Available from: https://doi.org/10.1177/1460458216647760

27. Qin kan, Yujiu Yang, Wenhuang Liu, Xiaodong Liu. An integrated approach for detecting approximate duplicate records. In: 2009 Asia-Pacific Conference on Computational Intelligence and Industrial Applications (PACIIA) [Internet]. IEEE; 2009. p. 381–4. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/5406409

28. Rianto, Mutiara Achmad Benny, Wibowo Eri Prasetyo, Santosa PI. Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. J Big Data [Internet]. 2021;8(1):26. Available from: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00413-1

29. Ruch P, Baud R, Geissbühler A. Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record. Artif Intell Med [Internet]. 2003 [cited 2020 Jun 6];29(1):169–84. Available from: http://www.sciencedirect.com/science/article/pii/S0933365703000526

30. Scharkow M. Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. Qual Quant [Internet]. 2013 [cited 2020 Jun 6];47(2):761–73. Available from: https://doi.org/10.1007/s11135-011-9545-7

31. Šnajder J. DerivBase.hr: A high-coverage derivational morphology resource for croatian. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) [Internet]. Reykjavik, Iceland: European Language Resources Association (ELRA); 2014 [cited 2020 Jun 6]. p. 3371–7. Available from: http://www.lrec-conf.org/proceedings/lrec2014/pdf/1090_Paper.pdf

32. Song B, Wu P, Zhang Q, Chai B, Gao Y, Yang H. Intelligent assessment of 95598 speech transcription text quality based on topic model. IOP Conf Ser: Mater Sci Eng [Internet]. 2019 [cited 2020 Jun 6];563(4):2001. Available from: https://doi.org/10.1088%2F1757-899x%2F563%2F4%2F042001

33. Stewart M, Liu W, Cardell-Oliver R, Wang R. Short-text lexical normalisation on industrial log data. In: 2018 IEEE International Conference on Big Knowledge (ICBK) [Internet]. IEEE; 2018. p. 113–22. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/document/8588782

34. Subha R, Palaniswami S. Quality factor assessment and text summarization of unambiguous natural language requirements. In: Unnikrishnan S, Surve S, Bhoir D, editors. Advances in Computing, Communication, and Control [Internet]. Berlin, Heidelberg: Springer; 2013. p. 131–46. Available from: https://doi-org.uml.idm.oclc.org/10.1007/978-3-642-36321-4_12

35. Tissot H, Del Fabro MD, Derczynski L, Roberts A. Normalisation of imprecise temporal expressions extracted from text. Knowl Inf Syst [Internet]. 2019 Dec 1 [cited 2020 Jun 6];61(3):1361–94. Available from: https://doi.org/10.1007/s10115-019-01338-1

36. Uejima H, Miura T, Shioya I. Improving text categorization by resolving semantic ambiguity. In: 2003 IEEE Pacific Rim Conference on Communications Computers and Signal Processing (PACRIM 2003) (Cat No03CH37490) [Internet]. IEEE; 2003. p. 796–9. Available from: https://ieeexplore-ieee-org.uml.idm.oclc.org/abstract/document/1235901

37. Vítovec P, Kléma J. Gene interaction extraction from biomedical texts by sentence skeletonization. In: Central Europe (CEUR) Workshop Proc [Internet]. Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic: NA; 2011. p. 230–41. Available from: http://ceur-ws.org/Vol-802/paper1.pdf

38. Westpfahl S, Schmidt T. FOLK-Gold ― A Gold Standard for Part-of-Speech-Tagging of Spoken German. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16) [Internet]. Portorož, Slovenia: European Language Resources Association (ELRA); 2016 [cited 2020 Jun 6]. p. 1493–9. Available from: https://www.aclweb.org/anthology/L16-1237

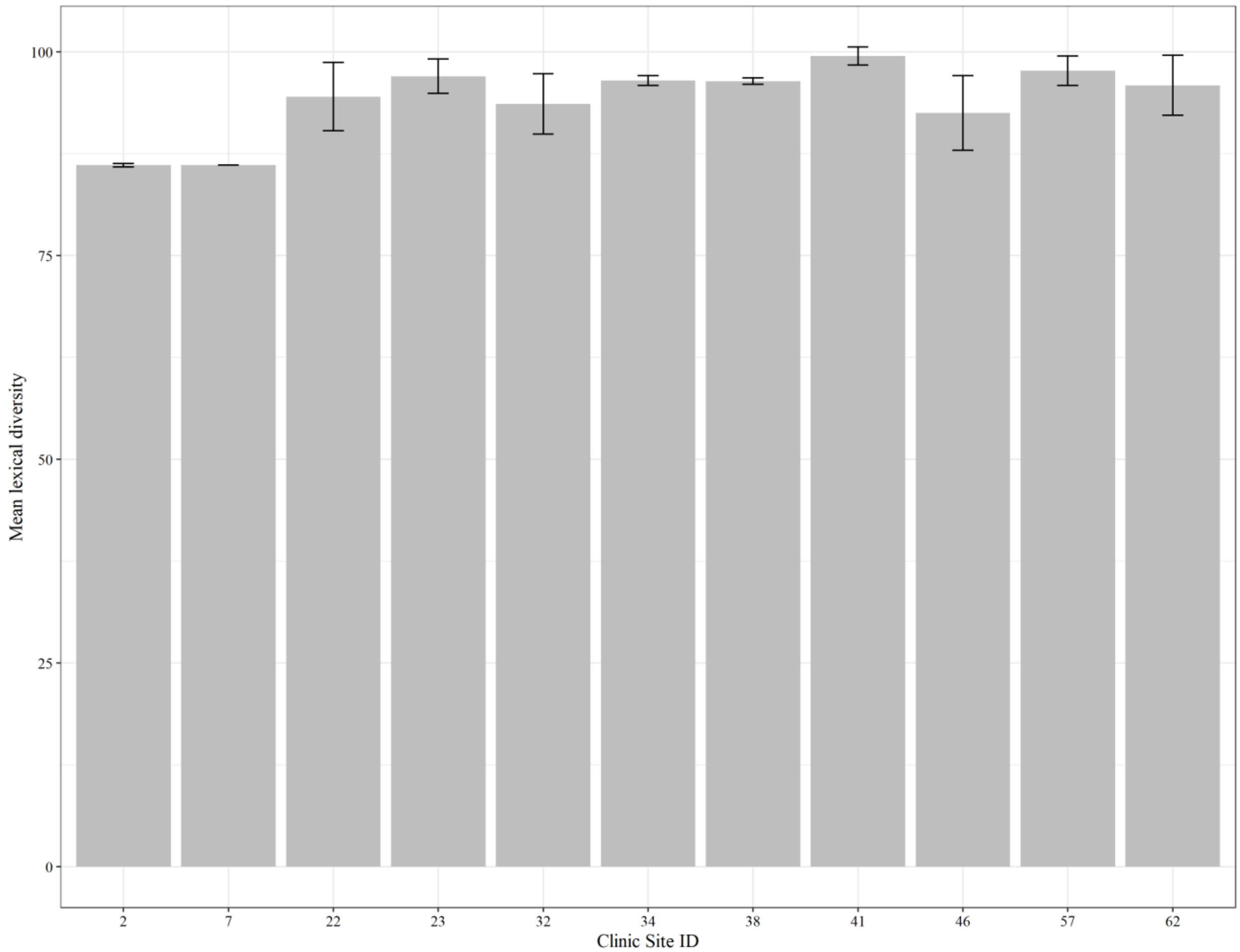
39. Xue N, Xia F, Chiou F, Palmer M. The Penn Chinese TreeBank: phrase structure annotation of a large corpus. Nat Lang Eng [Internet]. 2005 [cited 2020 Jun 6];11(2):207–38. Available from: http://doi.org/10.1017/S135132490400364X

40. Yang H-L, Chao AFY. Sentiment annotations for reviews: an information quality perspective. Online Inf Rev [Internet]. 2018 [cited 2020 Jun 5];42(5):579–94. Available from: https://doi.org/10.1108/OIR-04-2017-0114

41. Zurini M. Word sense disambiguation using aggregated similarity based on WordNet graph representation. Inform econ [Internet]. 2013 [cited 2020 Jun 6];17(3):169–80. Available from: https://ideas.repec.org/a/aes/infoec/v17y2013i3p169-180.html

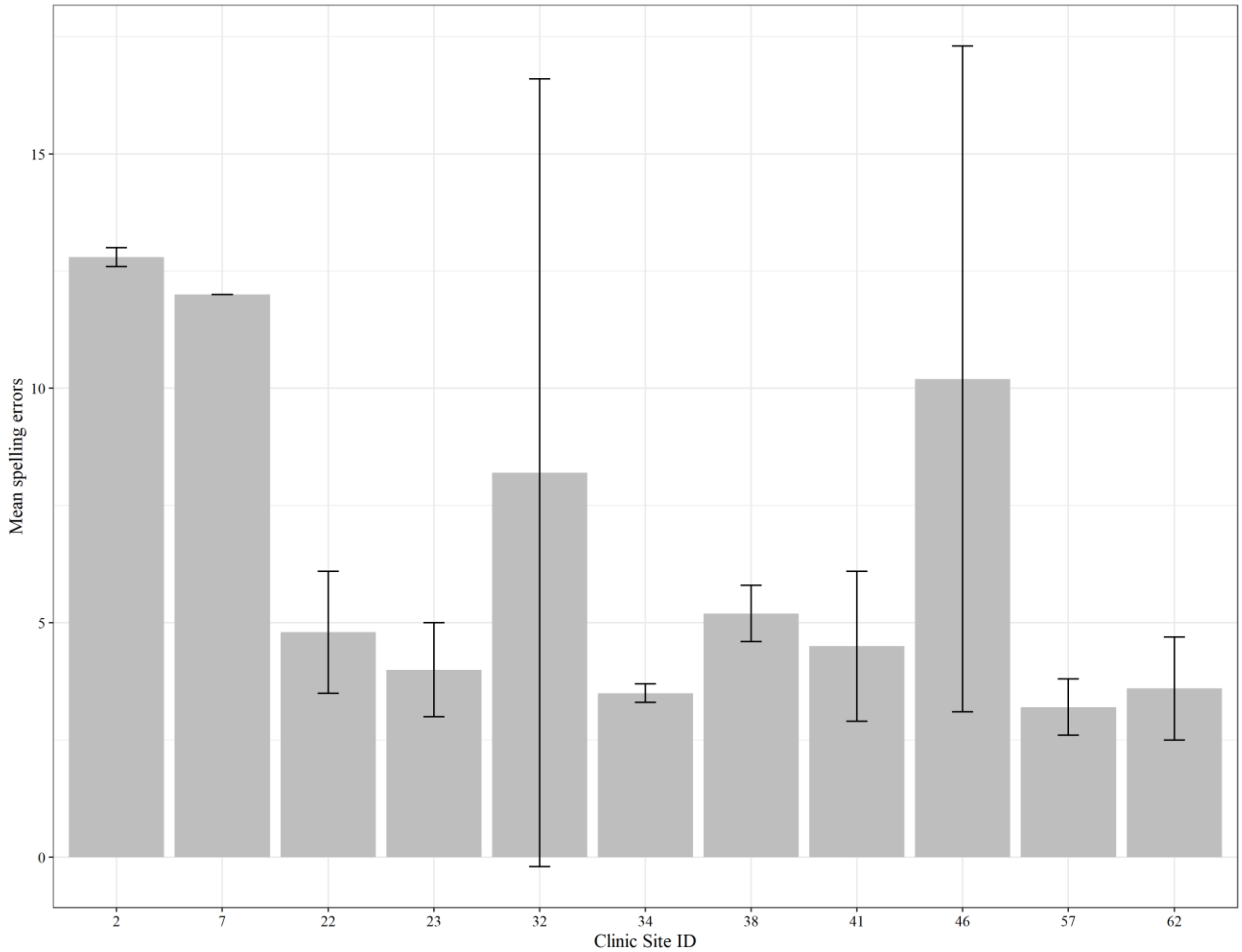
Appendix D – Preprocessing methods by clinic site

Figure 21 - Mean lexical diversity per note by clinic site



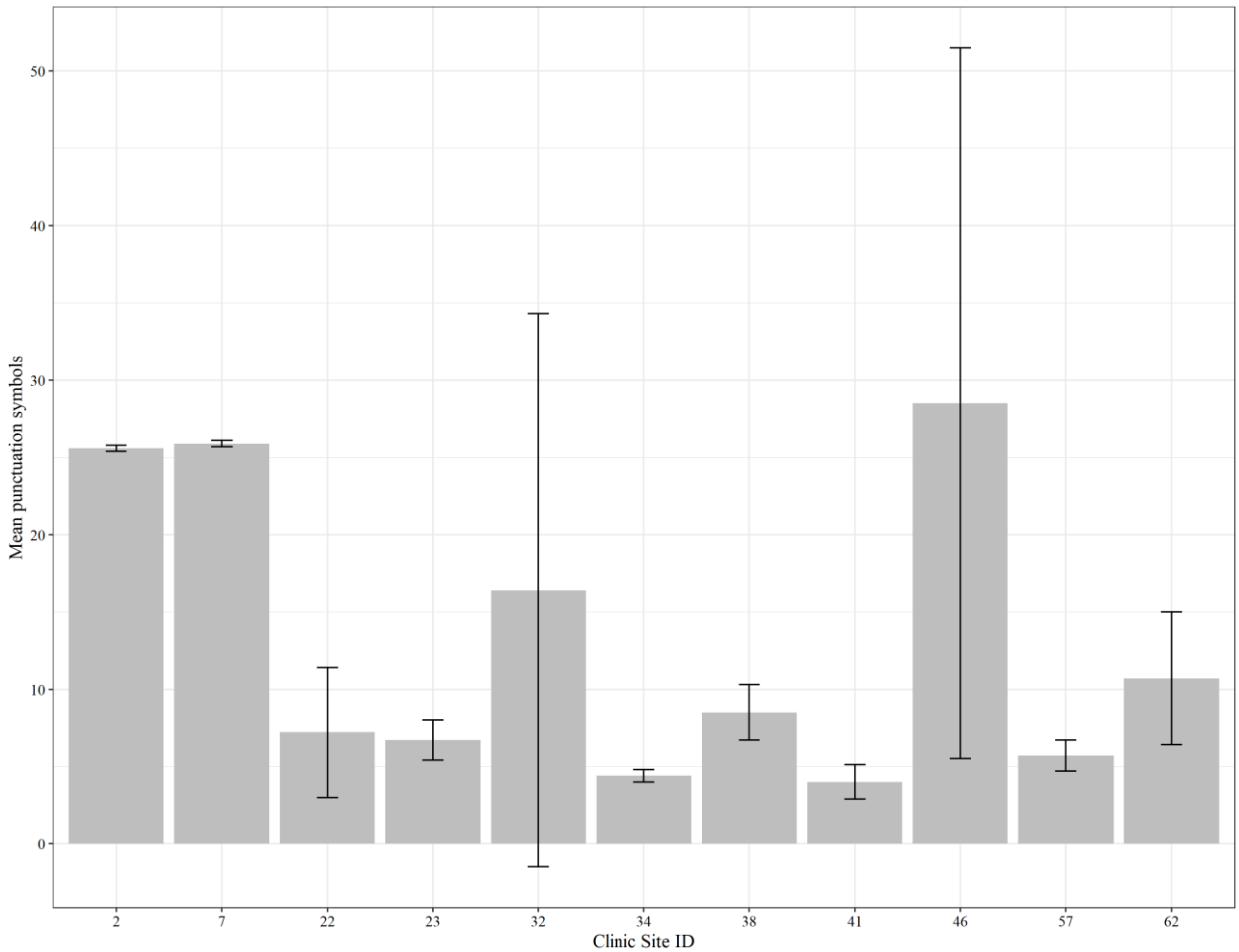
Note: Error bars represent the 95% confidence interval

Figure 22 - Mean spelling errors per note by clinic site



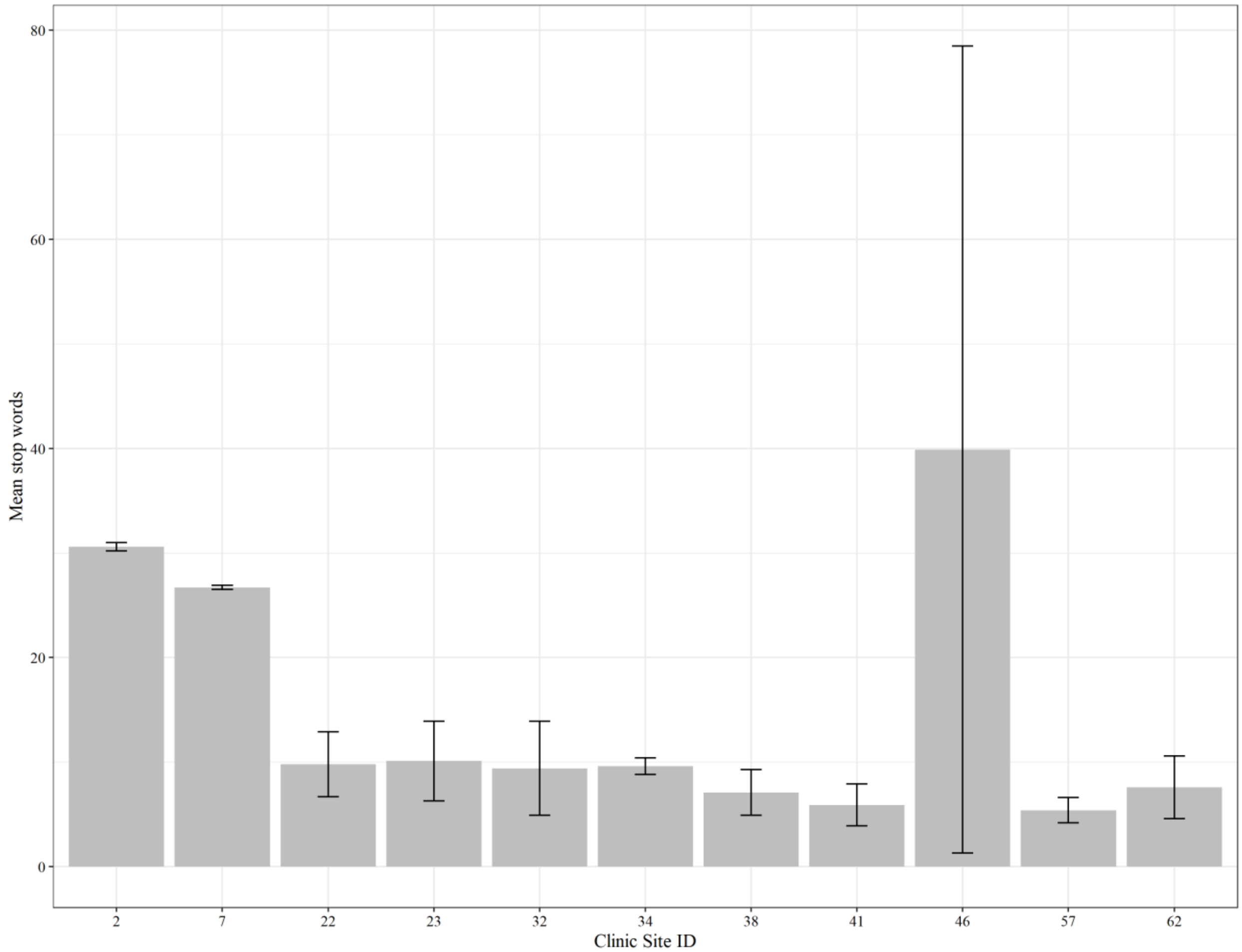
Note: Error bars represent a 95% confidence interval

Figure 23 - Mean punctuation per note by clinic site



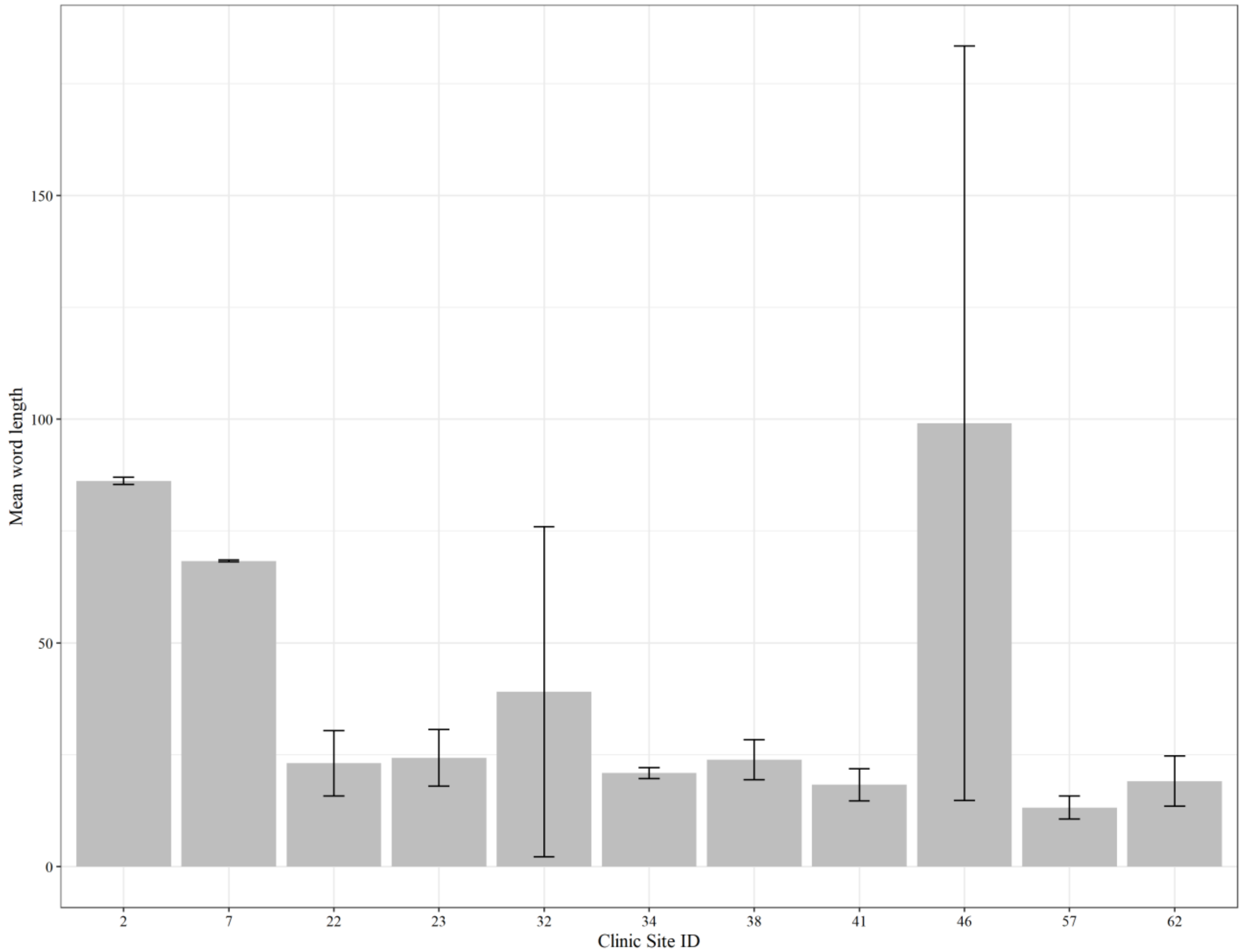
Note: Error bars represent the 95% confidence interval

Figure 24 - Mean stop words per note by clinic site



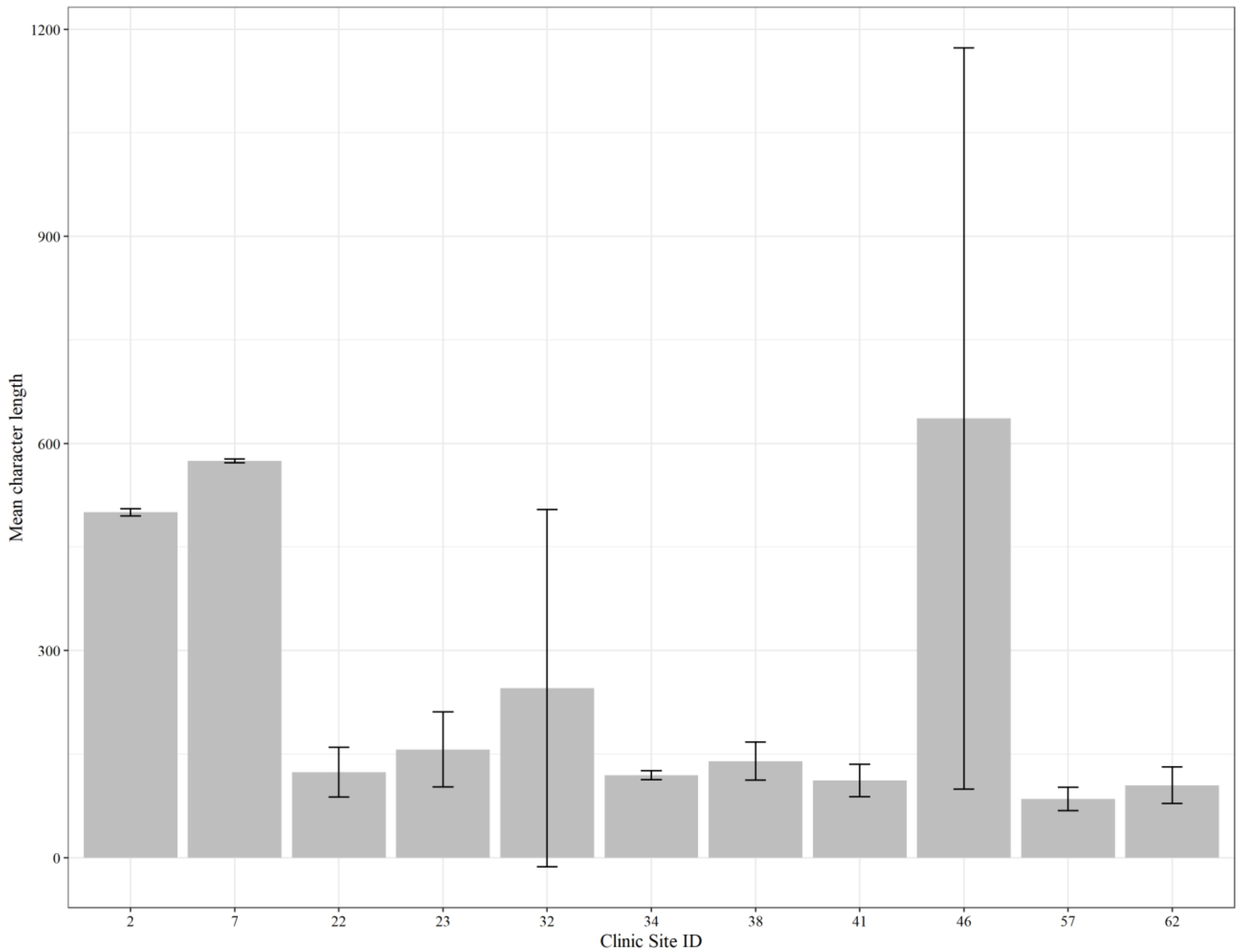
Note: Error bars represent the 95% confidence interval

Figure 25 - Mean word length per note by clinic site



Note: Error bars represent the 95% confidence interval

Figure 26 - Mean character length per note by clinic site



Note: Error bars represent the 95% confidence interval