Automated de-identification and unstructured textual electronic medical record data in Manitoba

by

Katelin Elizabeth Amy McDermott

A Thesis Submitted to the Faculty of Graduate Studies of

The University of Manitoba

In partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Community Health Sciences

University of Manitoba

Winnipeg, Manitoba, Canada

Abstract

Introduction

Unstructured textual electronic medical record (EMR) data, originally intended for monitoring and treatment, contain valuable patient details that can benefit health research. Personal health information (PHI) must be de-identified for EMR data to be used for secondary purposes; however, de-identifying unstructured data to meet ethical and legal guidelines is challenging. A considerable amount of de-identification research has been conducted using existing synthetic, de-identified, and annotated data sets, such as the Integrating Biology and the Bedside (i2b2) corpus. With its wide use in de-identification research, it has become the accepted environment for testing de-identification approaches. To date, little is known about how existing de-identification literature applies to unstructured EMR data in Manitoba.

Purpose & Objectives

This research examined unstructured textual EMR data in Manitoba and the applicability of existing de-identification literature to this setting. The objectives were to: 1) categorize the types and frequency of PHI in Manitoba EMR data, 2) assess the applicability of de-identification literature on Manitoba EMR data, and 3) test how NLM-Scrubber, an existing de-identification tool validated to be successful, redacts PHI in Manitoba EMR data. These objectives inform considerations for future de-identification of local unstructured textual EMR data.

Methods

The Manitoba data set was comprised of 750 unstructured textual encounter notes from 2003 to 2017 from the Manitoba Primary Care Research Network. In-scope PHI entities included name, personal health information number, address, phone number, and date (excluding year). Two annotators tagged PHI in the Manitoba EMR data using the Visual Tagging Tool. Cohen’s kappa assessed interrater reliability. Comparison of Manitoba EMR data and the 2014 i2b2 corpus examined encounter note compilation and PHI prevalence. NLM-Scrubber’s de-identification of Manitoba EMR data was assessed using performance measures and tested against the null hypothesis that NLM-Scrubber will recall 87% or more of PHI in Manitoba EMR data.

Results

The Manitoba EMR data contained 3,314 PHI instances, demonstrating 1.6% PHI prevalence. All in-scope PHI types were present. Interrater reliability was high (ƙ = 98.7; 95% CI, 92.0-105.4%). The Manitoba EMR data offered more independent notes and broader variety of note types than the i2b2 corpus. The Manitoba EMR data contained nearly twice as many name PHI instances as the i2b2 corpus (62% and 32%, respectively) but fewer date instances (31% and 55%, respectively). Overall, NLM-Scrubber showed weaker de-identification performance on Manitoba EMR data than seen on the i2b2 corpus. NLM-Scrubber's PHI recall was 75.4% (95% CI, 72.9-77.8%), leading to rejection of the null hypothesis.

Conclusion

Direct and indirect PHI exist within Manitoba EMR data, though they represent a small proportion of the data. De-identification literature may have limited applicability to Manitoba EMR data due to notable differences in percentage of name and date PHI, as well as generalizability of the data sets. NLM-Scrubber may not be acceptable for use on Manitoba EMR data due to its low recall performance. Attention should be directed to trained machine learning solutions that enable customization, adjustment of rule-based methods, and realistic surrogate PHI to help protect patient privacy.

Dedication

To Claire

May you always believe in yourself. You can accomplish whatever you set your mind to.

Believe you can and you're halfway there – Theadore Roosevelt

Acknowledgements

I would like to express my sincere gratitude to my advisor, Dr. Alan Katz. Thank you for your continued support and encouragement throughout the duration of my thesis, and most particularly through the final months. This thesis would not be possible without your knowledge, guidance, and creative problem solving. I am humbled by your commitment to me as a student and your willingness to work on a tight timeline.

I would also like to thank my committee members for their ongoing support and feedback throughout the initial development and revision of my thesis. Dr. Lisa Lix, thank you for your expertise and bringing rigor to my work, as well as expressing support for parental leave. Dr. Alexander Singer, thank you for your expertise in EMRs and support of in-kind Manitoba Primary Care Research Network resources.

A special thanks to Bill Peeler and Karim Asham. Bill, thank you for your time spent retrieving and preparing the Manitoba data, as well as your ongoing support accessing the data. Karim, I am grateful for your patience and support completing the painstaking process of annotating the Manitoba data set.

Thank you to Marcello Nesca and Nelson Prudencio for your willingness to share your knowledge and time with me. I am so appreciative to you both for providing a safe space to clarify gaps in my understanding.

To my close colleagues and friends… especially those that paved the way… thank you for encouraging me to complete my degree, the periodic check-ins, and celebrating the small wins along the way. I am incredibly lucky to be surrounded by such support.

To my parents, thank you for always believing in me, instilling self-confidence and self-belief in us kids, and demonstrating daily what discipline and hard work looks like. These qualities enabled me to persist. I am where I am today because of you.

And lastly, Ryan McDermott. This work would not have come to fruition without your encouragement, support, and patience. Thank you for persevering through the hard days to make sure this was checked off my list. I couldn’t have done it without you.

Table of Contents

Acronyms

APP	Australian Privacy Principles
CPCSSN	Canadian Primary Care Sentinel Surveillance Network
CI	Confidence Interval
CIHI	Canadian Institute for Health Information
CRF	Conditional Random Fields
DBMI	Department of Biomedical Informatics (Blavatnik Institute of Harvard Medical School)
EMR	Electronic Medical Record
EHR	Electronic Health Record
EPR	Electronic Patient Record
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FNR	False Negative Rate
HIPPA	Health Insurance Portability and Accountability Act
HL7	Health Level Seven
HMM	Hidden Markov Model
i2b2	Informatics for Integrating Biology and the Bedside
ICD-9	International Classification of Diseases, Ninth revision
MaPCReN	Manitoba Primary Care Research Network
MCHP	Manitoba Centre for Health Policy
MHRN	Manitoba Health Registration Number
n2c2	National NLP Clinical Challenges
NCBC	National Center for Biomedical Computing
NER	Named Entity Recognition
NLM	National Libraries of Medicine
NLP	Natural Language Processing
POS	Part-of-Speech
PHI	Personal Health Information

PHIA	Personal Health Information Act
PHIN	Personal Health Identification Number
PPV	Positive Predictive Value
SOAP	Subjective, Objective, Assessment, Plan
TN	True Negative
TP	True Positive
VTT	Visual Tagging Tool

Glossary

Corpus: A collection of data compiled to produce a data set.

De-identification: The process of removing information that can be used to identify a person. Identifying information can be in the form of a direct identifier (e.g., name) or indirect identifier (e.g., postal code).

Hybrid approach: Use of rule-based and machine learning approaches in a single de-identification method.

Instance: One or more strings of text that as a whole represent an occurrence (e.g., PHI instance).

Lexical feature: Algorithm condition(s) that use surrounding data to extract context and interpret the meaning of a string of text.

Machine learning: The application of algorithms to train and test data to achieve an outcome from the data set (e.g., categorize and de-identify data). Methods can be supervised by using labeled datasets to train algorithms or unsupervised where algorithms identify patterns without input or output data.

Named Entity Recognition (NER): Locate and classify a string of text as a named entity to enable categorization of textual unstructured data.

Natural Language Processing (NLP): The use of computer programming to process and analyze natural language data.

Orthographic features: Algorithm condition(s) that interpret the character type (e.g., special, capitalized, digit).

Parts-of-speech (POS): Identification of the part-of-speech for a word in natural language. For instance, recognition of a word as a noun or verb. Parts-of-speech is a lexical feature example.

Positional features: Algorithm condition(s) that assess the location of an entity within the data to support interpretation.

Pseudonymization: The process of replacing identifying data with realistic data of the same category (e.g., replace first name with realistic dummy first name, replace phone number with realistic dummy phone number) to produce a data set that protects patient privacy. A pseudonymized data set has been de-identified and contains pseudo data in place of existing PHI.

Regular expression: A specific sequence of characters that identifies a text pattern.

Rule-based approaches: Manually developed rules to support automated identification and categorization. Rules often incorporate data dictionary lookups, word patterns, and pre-defined word expressions (e.g., regular expression).

Semantic feature: Algorithm condition(s) that use lexical and syntactic information to interpret the relational information of the data.

Syntactic feature: Algorithm conditions(s) that analyze how data are arranged to apply a structure to text strings. An example is parsing text, the action of analyzing a string of text and separating the text strings into a data structure.

Token: A string of text that represents a named entity or component of natural language (e.g., definite article “the”).

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Background

Digital medical records have been widely adopted worldwide leading to rapid growth in the amount of health data about patients. These records contain detailed information collected during a healthcare encounter. Electronic medical records (EMRs), one type of digital medical record, contain structured and semi-structured data. Health-related research predominantly uses these data. An untapped source of rich information is unstructured EMR data such as clinician documentation in narrative notes. Use of machine learning models on unstructured data has demonstrated their power to predict disease, outbreaks, and drug misuse (Javeed et al., 2022; Mezzatesta et al., 2019; Panicacci et al., 2019; Sharma et al., 2020; Sung et al., 2021).

Patient privacy must be retained when accessing an individual's medical record, whether it is for primary or secondary uses. Legislation limits the use and disclosure of personal health information (PHI) to the original purposes for which it was collected, and requires patient consent if it is used for any other purpose (Canadian Institutes of Health Research, 2005; Manitoba Health Seniors and Active Living, 2022). In Manitoba, this legislation is called the Personal Health Information Act (PHIA). Collecting patient consent to use medical record data can however be a roadblock to completing valuable health-related research. Researchers report increased cost and time to complete research, an unnecessary burden placed on the patient, and a potential bias to population-based research where consent is not provided (Cavoukian, 2011; el Emam, 2011b; Harris et al., 2008; Kho et al., 2009; Ness, 2008).

Removing PHI from EMRs meets privacy legislation provided the patient's identity can not be re-identified. De-identification is a process that removes PHI while retaining the ability to link records to other data sets, and has been demonstrated to protect patient privacy (Cavoukian, 2011; Information and Privacy Commissioner of Ontario, 2014). Which data constitute PHI is broad; only the Health Insurance Portability and Accountability Act (HIPPA) in the United States of America (USA) outlines specific identifiers that must be de-identified to meet legal and ethical obligations. De-identification is clear for direct identifiers (e.g., name, health numbers). Indirect identifiers must also be de-identified, as they've been shown to provide data resulting in patient re-identification (Cavoukian & Khaled E, 2014; el Emam, Jonker, et al., 2011). The context, such as the rarity of the identifier or the content of the data set, influences whether an indirect identifier reveals a patient's identity with or without additional data, demonstrating the complexity of de-identifying context-rich unstructured textual data.

Manual de-identification is an onerous task and unrealistic for large data sets. Recognizing the need to automate de-identification, researchers have developed natural language processing (NLP) approaches to aid in de-identification of patient PHI. Approaches can be rule-based, machine learning-based, or a hybrid approach. Rule-based approaches use data dictionaries and regular expressions (a specified string of text used to define a search pattern) to identify strings of characters as PHI, whereas machine learning-based approaches rely on automated learning to apply algorithms, often with supervision from a researcher (Meystre et al., 2010). Automated de-identification is challenging. Methods must remove PHI while retaining data that are important to the analysis or data mining task. The context of words is critically important in assessing if a string is a PHI instance or not. For example, many names are the same as common words (e.g., mark, brown, black).

De-identification approaches have shown success with redacting PHI, however, there is not yet one approach that has shown consistent success across PHI categories and data types. Literature assessing the performance of de-identification approaches has relied heavily on pre-de-identified and annotated data sets known as corpuses in the computing world. The gold standard Informatics for Integrating Biology and the Bedside (i2b2) corpus is used widely in the literature (Stubbs & Uzuner, 2015), enabling a comparison of different de-identification approaches and providing indication of how an approach would perform on similar data. To date, little is known about unstructured textual EMR data in Manitoba and how existing de-identification literature applies to this setting. NLM-Scrubber, a de-identification tool validated to be successful, has been shown to recall 87.8% of PHI in the i2b2 corpus and up to 99% in other data sets (Kayaalp, Browne, Callaghan, et al., 2014). Recall, also referred to as sensitivity, demonstrates the proportion of PHI successfully redacted. It was hypothesized that NLM-Scrubber would demonstrate ≥87% recall on the Manitoba data set, an equivalent or better performance than demonstrated with the i2b2 data set.

1.2 Research Purpose and Objectives

The purpose of this research is to further understand unstructured textual EMR data in Manitoba and the applicability of de-identification literature on Manitoba data. The research aims to meet the following objectives:

1. To categorize the types and frequency of PHI in Manitoba unstructured textual EMR data.
2. To assess the applicability of de-identification literature on Manitoba unstructured textual EMR data.
3. To test how NLM-Scrubber redacts PHI in Manitoba unstructured textual EMR data.

NLM-Scrubber will be tested against the null hypothesis that PHI recall is ≥87%, demonstrating equivalent or better performance than its 87% recall on the i2b2 data set. The above objectives will inform considerations for de-identification of unstructured data in Manitoba.

Chapter 2: Literature Review

This chapter describes the literature relevant to electronic records commonly found in healthcare practices and the data structures found within the electronic systems. An overview of privacy legislation for personal data in health systems, and what constitutes as PHI, follows. The chapter concludes with a discussion on de-identification of PHI, including an overview of rule-based methods and machine learning techniques.

2.1 Electronic Records

Electronic records capture patient demographics and health information with the primary purpose to provide medical or health-related care. With the adoption of computer-based records on the rise worldwide (Protti, 2007), digital collection of medical record data has become standard in most clinical care settings.

There are multiple types of digital medical records including EMRs, electronic patient records (EPRs), and electronic health records (EHRs; eChart Manitoba, n.d.). An EMR is an electronic system used by healthcare professionals to track patient health history and visit details

(Canada Health Infoway, 2013). It contains detailed information collected during healthcare encounters. EMR records are typically accessible to one clinic or group only, therefore, is a clinic-centric health record for a patient (Shared Health, n.d.). As of 2020, over 90% of primary care providers are reported to have an EMR, demonstrating wide use of EMRs in primary care settings in Manitoba (Winnipeg Sun, 2020). An EPR is similar to an EMR. It is facility-based and typically only accessible to one group (eChart Manitoba, n.d.). In Manitoba, EPRs are used in hospital or specialty settings and have varying degrees of functionality, ranging from scheduling only to features that support documenting clinical encounters. An EHR stores information collected in other electronic records and systems to present a lifetime record of a patient's medical history (eChart Manitoba, n.d.). It typically does not reflect every healthcare encounter, but rather highlights key information and results that would be most relevant to provide episodic or emergent care. An EHR is patient-centric and is considered the most comprehensive record for a patient (eChart Manitoba, n.d.; Juhn & Liu, 2020).

In a Canadian Institute for Health Information (CIHI) National Physician Survey, 86% of Canadian physicians reported using an electronic record and it is estimated that 19 million Canadians have a health record within an EMR (Canadian Institute for Health Information, 2019). Given its detailed capture of patient medical encounters, EMRs are rich with information (Lee et al., 2017); data can be used to understand health and medical conditions, to assess care provided to patients, and for population health planning (Azam et al., 2016; Tolar & Balka, 2012). How this information is captured, and the quality of the recorded data, are important. Both influence the ability to retrieve and use the information, to support clinical decision making at a patient and population level.

2.2 Health Data Structure

Data in EMRs can be captured in a structured, semi-structured, or unstructured format. Structured data are collected in discrete data fields, follow a standardized structure, and are intended for a specified purpose (Taylor, 2021). For example, an EMR has discrete fields to capture numeric test results, dates, diagnoses, and unique patient health identifiers. Data entry is completed by following a specified structure or by using controlled features such as a picklist, as commonly seen with International Classification of Diseases, Ninth revision (ICD-9) diagnostic code selection.

Semi-structured data are collected in discrete fields but without functionality to force structure standardization (Kristianson et al., 2007). Data are entered using free text and thus can follow whatever formatting and standardization, or lack thereof, that an end user chooses. Though there are typically clear expectations of what is conventionally inputted in the field, semi-structured data requires data validation and processing. The processing of semi-structured data is more complex than structured data due to variation in expected content (e.g., numbers can be entered in fields where an alpha string is expected) and data quality issues (*Data Quality and Machine Learning: What's the Connection?*, n.d.).

Unstructured data lack a specified structure or organized data model (Taylor, 2021; Tayefi et al., 2021). A common example of unstructured text data is an encounter note. An encounter note documents a patient encounter by providing details of patient state and rationale for care provided. Its flexibility results in unstructured data capturing detail-rich information that cannot be obtained from structured data alone (Abhyankar et al., 2014; Kayaalp, Browne, Callaghan, et al., 2014).

Singer et al. (2017) demonstrated that some structured EMR data fields may lack completeness and be unreliable for research until they are shown to be more comprehensive and accurately represent true state. Additionally, recent research suggests improved patient outcomes may not be realized through clinical decision support functionality as expected (Heselmans et al., 2020). This highlights the need to begin to use unstructured data in research to support patient care, for example to detect disease or disease risk sooner and support prevention and early management.

Historically, structured data have been the main source of medical record data used for research; however, unstructured data are becoming of special interest to researchers because of the detailed, context-specific information they can provide. Unstructured data has been leveraged to study drug reactions (Banerji et al., 2020), identify drug misuse (Sharma et al., 2020), predict and detect disease (Javeed et al., 2022; Mezzatesta et al., 2019; Panicacci et al., 2019), and support COVID-19 efforts (Meystre et al., 2021). Nevertheless, privacy guidelines must be met to use unstructured data for this secondary purpose.

2.3 Health Information Privacy Legislation

Protecting patient privacy is essential when accessing an individual's medical record, whether it is for primary or secondary use. Legislation exists worldwide to manage this. In Manitoba, PHIA dictates what information can be collected, how it can be used, and when and with whom it can be shared (Manitoba Health Seniors and Active Living, 2022).

Similar provisions exist federally and in other Canadian jurisdictions (Canadian Institute for Health Information, 2010; el Emam, 2011a; Federal Register, 2002; Milieu Ltd, 2014; Ontario Cancer Care, 2014; Personal Health Information Protection Act, 2004). Each jurisdiction has its own legislation to protect patient PHI and meet Canadian healthcare privacy legislation.

In 2018, Europe implemented new legislation, General Data Protection Regulation (GDPR; What Is GDPR, the EU's New Data Protection Law?, n.d.). It aims to strike a balance between addressing privacy of patient identity, data security, and use of data for secondary purposes. It acknowledges technological advances, for both direct patient benefit (e.g., accessing their own electronic record) and indirect patient care through advancements supporting research. Still, at the root of Europe's data protection law is the emphasis that data must not identify a person. Australia's legislation, Australian Privacy Principles (APP), follows the same model as seen in Canada and Europe (Office of the Australian Information Commissioner, n.d.). While this legislation is comprehensive, there is a lack of consensus of when data is de-identified and thus not traceable back to a patient (Chevrier et al., 2019).

2.3.1 What constitutes PHI

There is a wide variety of information that may need to be de-identified to meet legal and ethical obligations. As outlined in the Manitoba Consolidated Act P33.5, Manitoba defines PHI as:

" recorded information about an identifiable individual that relates to:

(a) the individual's health, or health care history, including genetic information about the individual,

(b) the provision of health care to the individual, or

(c) payment for health care provided to the individual, and includes:

> (d) the PHIN and any other identifying number, symbol or particular assigned to an individual, and
>
> (e) any identifying information about the individual that is collected in the course of, and is incidental to, the provision of health care or payment for health care". (Government of Manitoba, 1997)

In addition, the Manitoba Government notes "this definition even includes general information about a person (such as name, address, gender and date of birth) if that information is collected during a health care service or if it is used to administer payment for a health care service" (Manitoba Health, n.d.). Other jurisdictions have similar definitions of PHI. For example, Nova Scotia defines PHI as "information that identifies an individual, or where it is reasonably foreseeable could identify an individual when used alone or with other information" (Department of Health and Wellness, 2013). In Europe, PHI is defined as "personal data is any information that relates to an identified or identifiable living individual. Different pieces of information, which collected together can lead to the identification of a particular person, also constitute personal data" (*What Is Personal Data? | European Commission*, n.d.). Demonstrating how expansive interpretation of information that requires de-identification can be, Australia's definition of PHI acknowledges that what constitutes as PHI is situation dependant and variable: "any information or an opinion about an identified individual, or an individual who is reasonably identifiable whether the information or opinion is true or not; and, whether the information or opinion is recorded in a material form or not"(Chapter B: Key Concepts - Home, n.d.).

Application of legislation in practice lacks clarity and there are few guidelines or rules established that clearly outline which information should be de-identified or scenarios in which they require deidentification. The Working Party in Europe, an advisory body on data protection and privacy, released an opinion paper analyzing the concept of personal data to establish agreement on the application of the legislation, including when it should not apply (Article 29 Data Protection Working Party, 2007). The USA Department of Health & Human Services provides some guidance on methods for de-identification in the HIPPA Privacy Rule (*Health Insurance Portability and Accountability Act of 1996 (HIPAA) | CDC*, n.d.; U.S. Department of Health & Human Services, n.d.-a). The guidelines were developed in consultation with practical, technical, and policy subject matter experts. Two methods were identified as acceptable: Expert Determination and Safe Harbor. Expert determination requires an expert in statistical or scientific principles to deem the information non-identifiable. The Safe Harbour method provides the most detail on which identifiers must be removed in order for data to be deemed sufficiently de-identified and anonymized. There are 18 types of identifiers that must be removed, as stated by the Department of Health & Human Services:

> (A) Names
>
> (B) All geographic subdivisions smaller than a state, including street address, city, county, precinct, ZIP code, and their equivalent geocodes, except for the initial three digits of the ZIP code if, according to the current publicly available data from the Bureau of the Census:
>
> > (1) The geographic unit formed by combining all ZIP codes with the same three initial digits contains more than 20,000 people; and
> >
> > (2) The initial three digits of a ZIP code for all such geographic units containing 20,000 or fewer people is changed to 000
>
> (C) All elements of dates (except year) for dates that are directly related to an individual, including birth date, admission date, discharge date, death date, and all ages over 89 and all elements of dates (including year) indicative of such age, except that such ages and elements may be aggregated into a single category of age 90 or older
>
> (D) Telephone numbers
>
> (E) Fax numbers
>
> (F) Email addresses
>
> (G) Social security numbers
>
> (H) Medical record numbers
>
> (I) Health plan beneficiary numbers
>
> (J) Account numbers
>
> (K) Certificate/license numbers

(L) Vehicle identifiers and serial numbers, including license plate numbers

(M) Device identifiers and serial numbers

(N) Web Universal Resource Locators (URLs)

(O) Internet Protocol (IP) addresses

(P) Biometric identifiers, including finger and voice prints

(Q) Full-face photographs and any comparable images; and,

(R) Any other unique identifying number, characteristic, or code.

The above guidelines were developed to support the clause that use of de-identified data is unrestricted, thus enabling de-identified data to be used for secondary purposes. This use exists for researchers; however, the ease of de-identification varies across data forms. It is easier to find PHI for de-identification in structured data because it is typically known where the PHI exists in the data, whereas unstructured data sets lack standardization making de-identification a difficult task.

2.4 De-identification

Medical records can be accessed without patient consent if PHI is removed from the data set so that a patient's identity cannot be determined (Canadian Institute for Health Information, 2013; el Emam, 2011a; Manitoba Health Seniors and Active Living, 2022; U.S. Department of Health & Human Services, n.d.-b; *What Is Personal Data? | European Commission*, n.d.). Anonymization and de-identification are both methods that remove patient identifying information. Anonymization of data removes the ability to link data to a patient; an anonymized record cannot be re-identified (Kushida et al., 2012). In health research, this method limits the value of the EMR data as linking multiple data sets can be beneficial (e.g., linking EMR and administrative health data). De-identification is a process that finds and redacts (i.e., removes or modifies) PHI, however, it retains the ability to link records to other data sets without revealing the patient identity. This linkage can be achieved by using a non-identifying unique patient identifier known only to authorized individuals (Kushida et al., 2012). Data are considered de-identified when they do not identify a patient, nor can be manipulated to re-identify a patient without appropriate approval (Canadian Institute for Health Information, 2013; Cavoukian &

Khaled E, 2014; el Emam, 2013; Lo, 2012). De-identification has been assessed as a sound way to protect patient privacy (Cavoukian K., 2011; Cavoukian K., 2014; Information and Privacy Commissioner of Ontario, 2014; Lafky, 2010; Uzuner et al., 2007).

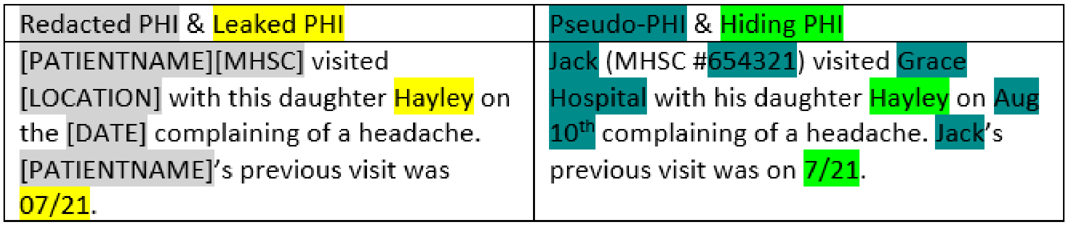
Thanks to their standardized structure, structured data enable easy identification of PHI for the purpose of de-identification. De-identifying unstructured data to meet legal and ethical guidelines is challenging (Kayaalp, Browne, Dodd, et al., 2014). Manual de-identification requires multiple annotators to ensure precision and reliability (Douglass et al., 2004). Studies on a large population would be impossible or prohibitively time consuming and costly (Dorr et al., 2006; Douglass et al., 2004, 2005), and is not a viable option as data volumes continue to grow exponentially.

Further challenging the task of de-identification, an identifier can be a direct or an indirect identifier (Cavoukian K., 2014; Fraser R. & Willison D., 2009). Examples of direct identifiers are name and health number. These data elements identify a patient without additional data. Indirect identifiers can also be used to identify a patient. The context, such as the rarity of the indicator or the content of the data set, influences whether an indirect identifier reveals a patient's identity with or without additional data (Cavoukian K., 2014). Examples of indirect identifiers are postal code, occupation, and rare health conditions. Direct and indirect indicators may exist throughout a patient's medical record, in either structured or unstructured fields. It is possible for indirect identifiers to be used to breach patient privacy, as demonstrated by studies that re-identified individuals with data that had been deemed de-identified (Barbaro & Zeller Jr., 2006; Barth-Jones, 2012; Cavoukian & Khaled E, 2014; Emam et al., 2009; Narayanan & Shmatikov, 2008; Sweeney, 2002). Further demonstrating the complexity of de-identifying data, even data de-identified using the safe-harbour method, the method with clear guidelines specifying 18 health information identifiers to remove, have been re-identified (el Emam, Buckeridge, et al., 2011; el Emam, J, et al., 2011). This challenge is exacerbated as new data sources become available with advancing technologies (Shin, 2018). Given the challenges with manual de-identification, ever-growing data sets, and the importance of patient privacy, NLP approaches have been developed to automate de-identification of unstructured health data.

2.5 De-identification Approaches

NLP has various applications in healthcare, such as supporting information extraction, categorization, and decision support (Negro-Calduch et al., 2021). De-identification of PHI using NLP has been shown to be as successful as manual de-identification (Deleger et al., 2013). NLP for de-identification is considered a named entity recognition (NER) task, though it typically requires more categories than traditional NER (Stubbs et al., 2015). Essentially it is the ability to identify a word or string of characters from a body of text and characterize it as a named entity (Tayefi et al., 2021). For de-identification, named entities are PHI categories. A common example of a named entity is name. There can also be subcategories of a named entity, such as patient name and clinician name (Stubbs et al., 2015).

When a string is identified, an approach can redact the PHI by replacing the characters in the PHI string with a symbol (e.g., *), with a tag (e.g., [PERSONALNAME]), or pseudo data (e.g., replace John with Michael). The latter is strongly encouraged as it helps hide true data missed during de-identification (Carrell et al., 2013; Murugadoss et al., 2021). Strategies have been developed to support replacement of PHI with pseudo data (Chen et al., 2019), a process that has been referred to as pseudonymization. Figure 1 adapted from Murugadoss et al. (2021) visually demonstrates the power of pseudonymization. A pseudonymized data set is a de-identified data set that replaced identifying data with realistic data of the same category.



Redacted PHI & Leaked PHI	Pseudo-PHI & Hiding PHI
[PATIENTNAME][MHSC] visited [LOCATION] with this daughter Hayley on the [DATE] complaining of a headache. [PATIENTNAME]'s previous visit was 07/21.	Jack (MHSC #654321) visited Grace Hospital with his daughter Hayley on Aug 10th complaining of a headache. Jack's previous visit was on 7/21.

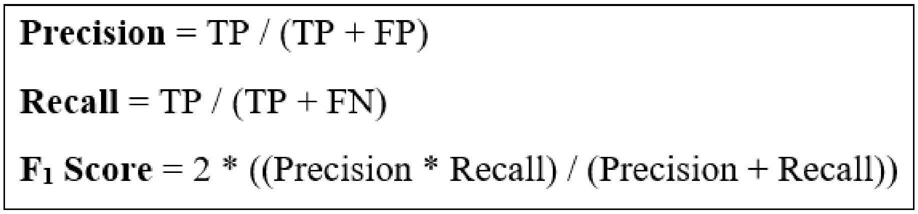
Figure 1. Leaked vs Hidden PHI Post De-identification

Despite sophisticated replacement techniques, automating de-identification is a challenging task. Identifying and removing PHI while retaining data that are important to the analysis or data mining is difficult. PHI are often not clearly marked and unnecessary removal or modification of common words may compromise the usefulness of the data set, especially when the common word is a clinical sign or characteristic (Berman, 2003; Garfinkel, 2015). For instance, many names are the same as common words (e.g., mark, black, brown, white).

Automated de-identification approaches have shown success, however, there is not yet an approach that's generalizable across PHI categories and sources of medical documentation. Over the years, NLP approaches for de-identification have continued to evolve. Initially primarily rule-based, approaches have advanced to leverage the sophistication of machine learning, leading to the development of hybrid approaches that incorporate rule-based methods into machine learning techniques, and more recently, deep learning (Sheikhalishahi et al., 2019).

2.5.1 Evaluation of de-identification approaches

When assessing de-identification, the decision on each token is considered a true positive (TP), false negative (FN), false positive (FP), or true negative (TN). A TP result identifies PHI is tagged as PHI, whereas a FN result occurs when a tag is missed for a PHI token. A FP identifies a non-PHI instance tagged as PHI. Lastly, a TN result occurred when non-PHI token was not tagged as PHI. Evaluation of de-identification performance is routinely assessed using the values shown in Figure 2: precision, recall, and F_1 Score (Stubbs et al., 2015).



Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F_1 Score = 2 * ((Precision * Recall) / (Precision + Recall))

Figure 2. Common De-identification Performance Metrics

2.5.2 Rule-based methods

Rule-based approaches date as far back as 1996 (Sweeney, 1996). They can use various methods to identify strings of characters as PHI such as: word patterns, data dictionaries, regular expressions, predictive markers (e.g., Dr.), and proximity matching (Chylek et al., 2011; Stubbs et al., 2015; Trienes et al., 2020). Data dictionaries can extend to any entity requiring identification, including but not limited to, person names (first and last names), locations (names and postal or ZIP codes), hospital names, medical terms and phrases, and common words that are non-PHI (Stubbs et al., 2015).

Rule-based redaction approaches are trained using the above-mentioned methods. The approach scans a body of text for strings that match the requirements outlined in the rules. This matching can be in the form of word patterns using regular expressions. Regular expressions identify the text and character patterns likely to represent a PHI instance. Similarly, using data dictionaries, a rule-based redaction approach looks through the body of text for strings that match data found in the dictionaries (Chylek et al., 2011).

Rule-based de-identification systems have been developed for a range of free-text medical documentation including clinical reports (Douglass et al., 2005; Friedlin & Mcdonald, 2008; Neamatullah et al., 2008; Shin et al., 2015) , discharge summaries (Aramaki et al., 2006), speciality-specific reports such as pathology (Beckwith et al., 2006; Berman, 2003), and health level seven (HL7) messages (Friedlin & McDonald, 2008). Rule-based approaches can be useful when there are clear named entity categories and available dictionary lists, but typically require manipulation for each unique data set to be effective at identifying PHI (Hartman et al., 2020). It can be difficult to maintain data dictionaries and ensure all possible data scenarios are accounted for (e.g., formatting, structure, errors such as spelling mistakes). Additionally, rule-based approaches are limited by their inability to consider the context of the words in question, a critical step to ensure that non-PHI, context-important text is retained (Dehghan et al., 2015; Norgeot et al., 2020). Additionally, rule-based approaches lack in generalizability across medical document types (Meystre et al., 2010).

One of the earliest rule-based methods to be developed was NLM-Scrubber. This method was developed by the US National Library of Medicine with the goal to minimize the burden of manual de-identification (Kayaalp, Browne, Dodd, et al., 2014). This method groups PHI entities into four main categories: name, address, date, and alphanumerical strings (any string of characters with one or more digits). NLM-Scrubber relies on pattern matching and regular expressions, as well as some data dictionaries (Kayaalp, Browne, Dodd, et al., 2014). The PHI instance is replaced with the entity name (e.g., Michael is replaced with [PERSONALNAME]). Using clinical notes from a random sample of patient records from the National Institutes of Health Clinical Center, NLM-Scrubber was shown to be successful at redacting names, with a 99.9% recall (Kayaalp, Browne, Callaghan, et al., 2014). Kayaalp, Browne, and Callaghan et. al (2014) attributed this high recall to fulsome name lists. On the i2b2 data set, NLP Scrubber achieved 76.3% precision and 87.8% recall overall (Norgeot et al., 2020).

2.5.3 Machine learning

Machine learning approaches rely on automated learning to apply algorithms, often with some supervision from a researcher (Meystre et al., 2010). Supervised machine learning approaches are on the rise (Sheikhalishahi et al., 2019) as they can overcome challenges seen with rule-based methods and have demonstrated better performance at identifying PHI than rule-based approaches (Meystre et al., 2010; Steinkamp et al., 2020). Supervised machine learning can employ various methods to identify and remove PHI. These include algorithms that consider:

- lexical features, such as part-of-speech (POS) tag and surrounding tokens;
- orthographic features, such as patterns like consistent capitalization or use of acronyms;
- semantic features, including phrases that can indicate that an identifier follows (e.g., workplace or profession is likely to follow the phrase "works for"); and,
- positional features (Dehghan et al., 2015).

Machine learning can also leverage contextual cues to identify PHI and distinguish between types of PHI, such as phone number versus fax number or clinician vs patient name. Previous research has shown that it is common for the textual formatting to cause PHI identification errors, such as 13% of PHI tagging errors due to boundary detection issues (Deleger et al., 2013). It is thus common to pre-process the data set prior to de-identification. Pre-processing can involve manipulating the data set to support de-identification. For example, to parse the data into tokens (i.e., tokenize) or sections using data dictionaries (He et al., 2015), or complete POS tagging which identifies a word in a body of text based on word type and neighbouring terms (Lee et al., 2017).

Numerous supervised machine learning de-identification approaches leverage rule based methods, a blended approach that can be used for hybrid or ensemble de-identification approaches (Dehghan et al., 2015; Lee et al., 2017). To accurately identify PHI, a mix of rules from the above features and data dictionaries can be incorporated to train a supervised machine learning tool. Research has shown that incorporating lists can lead to better results than methods that do not include lists (Kayaalp, Browne, Callaghan, et al., 2014). An added benefit is that machine learning enables a unique model to be deployed for each PHI type (Dehghan et al., 2015). The following section highlights how this blended approach has shown to be successful on a gold standard corpus, the i2b2 data set.

2.6 Machine Learning on i2b2 Corpus

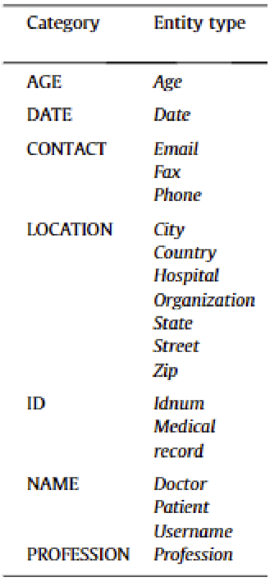
I2b2 at the National Center for Biomedical Computing (NCBC) created a gold standard corpus of de-identified and annotated clinical records (Uzuner & Stubbs, 2015). This i2b2 data set contained pseudo PHI with the intent of supporting automated de-identification research. The corpus was shared publicly, and researchers were invited to participate in a challenge to test de-identification approaches on the corpus.

Ten teams participated in the i2b2 2014 challenge. Of the eight teams for which there is an approach description, six used a combination of machine learning techniques and rule-based features, one used machine learning techniques only, and one relied solely on rule-based methods. The most successful approaches combined Conditional Random Fields (CRF), a machine learning statistical modelling method, and rule-based methods (Stubbs et al., 2015) which aligns with other work that demonstrated CRF performing well on de-identification tasks (Deleger et al., 2013; Uzuner et al., 2007). In de-identification work, CRF is a probabilistic discriminative model that uses contextual information to predict if a string of text is a PHI or non-PHI entity (Chavan, 2019). A review of the seven approaches that used machine learning follows, starting with the approach with the highest F_1 Score through to the approach with the lowest F_1 Score.

Yang & Garibaldi (2015) submitted an approach that used a CRF-based machine learning technique that involved word-token, context, orthographic, sentence-level, and task-specific features, as well as data dictionaries and regular expression rule-based methods. They argue that the range of PHI sub-categories have a variety and complexity of features which require a hybrid de-identification model involving machine learning techniques and rule-based approaches (Yang & Garibaldi, 2015). They pre-processed their data for parsing, tokenization, POS tagging, and document-level features such as headers, and incorporated a processing step that established a white-list of PHI terms in the data set and further parsed PHI tokens from surrounding data to identify PHI. Their approach achieved 96.5% precision, 90.9% recall, and a F_1 Score of 93.6% on the i2b2 PHI categories (Stubbs et al., 2015). Yang & Garibaldi (2015) report “regular expression template patterns, when combined with other orthographic features, can be quite effective in predicting PHI [that rely heavily on regular expression]” (p. s35).

The next best performing team also used a hybrid system involving machine learning and rule-based approaches. Liu et al.’s (2015) approach incorporated two CRFs, one based on token-level features and the other on character-level features, as well as rule-based methods using regular expressions. Their output was created from a rule-based decision system involving a review of overlapping instances and hierarchical decision points. Their approach achieved 92.6% precision, 89.9% recall, and a F_1 Score of 91.2% on the i2b2 PHI categories (Stubbs et al., 2015).

Dehghan et al. (2015) attempted to de-identify clinical notes using machine learning techniques with rule-based methods, referred to by the authors as data-driven and knowledge-driven methods. Leveraging the i2b2 2014 challenge corpus, they attempted to de-identify 19 PHI entity types (see Figure 3). In their approach, they pre-processed the corpus with lexical and terminological features, incorporated data dictionaries and feature-type rules, and trained CRF models for their supervised machine learning technique (Dehghan et al., 2015). Their method resulted in using an average of five rules on each entity type and 279 machine learning features for each token in the corpus, including a different machine learning model for each of the following entity types: city, date, hospital, organization, profession, and patient. Their method was unique; following the pre-processing and initial de-identification, they employed a second-pass aimed to remove ambiguous terms and known false positives.



Category	Entity type
AGE	*Age*
DATE	*Date*
CONTACT	*Email* *Fax* *Phone*
LOCATION	*City* *Country* *Hospital* *Organization* *State* *Street* *Zip*
ID	*Idnum* *Medical record*
NAME	*Doctor* *Patient* *Username*
PROFESSION	*Profession*

Figure 3. PHI Categories and Sub-categories De-identified by Dehghan et al. (2015)

Their two-pass method resulted in a precision range of 33.3% to 100.0% on sub-categories, with >80.0% precision on 16 out of the 19 sub-categories. The lowest precision occurred for fax numbers (33.3%) and organization (40.5%), where as the highest precision was achieved for doctor name (96.6%) and usernames (100.0%). Recall performance ranged from 20.7% to 100.0%; eleven of the 19 sub-categories had >80.0%. Lastly, F_1 Score performance ranged from 27.4% to 100.0%. Thirteen out of the 19 sub-categories had >80.0% F_1 Score (Dehghan et al., 2015).

Dehghan et al. (2015) analyzed the FN and FP errors encountered during de-identification attempts. FPs and FNs were common on data quality concerns within the data set (e.g., missing spaces), as well as organization and profession sub-categories. For the latter, the errors occurred on terms that were uncommon in the data set or due to the context in the neighbouring words. Other FNs commonly seen across all sub-categories were in PHI strings where one or more tokens were tagged but the remaining tokens were not. FPs were also seen for medical abbreviations being wrongly identified as hospital abbreviations, as well as identifying the wrong sub-category between doctor name and patient name. Their work resulted in proposed combinations of rule-based and machine learning methods based on the entity type (see Table 1).

Table 1.

Proposed Methods by Entity Sub-Category (Dehghan et al., 2015)

Method(s) to employ	**Entity sub-category**
Dictionaries + Supervised Machine Learning	City, Hospital, Organization, Profession
Dictionaries + Rule	Country, State
Rule + Supervised Machine Learning	Date, Patient
Rule	Age, Doctor, Username, Email, Fax, Phone, Id Number, Medical Record Number, Street, ZIP Code

In a different approach using machine learning techniques, He et al., (2015) achieved 92.3% precision, 85.1% recall, and a F_1 Score of 88.5% on the i2b2 PHI categories (Stubbs et al., 2015). The authors trained CRF on lexical, orthographic, and syntactic features (Stubbs et al., 2015). Prior to running the CRF, data was pre-processed for tokenization and sentence parsing. They used regular expressions to identify the string of characters, and then manipulated the data so that components of the data that were not PHI are separated from the PHI. For example, “45-year-old” becomes “45 -year old” after pre-processing; the added space helped the “45” be identified as an age in the subsequent de-identification step (He et al., 2015).

Another approach modified an existing tool, MIST, by adding their own lexicon rules for profession and location, regular expressions for phone, ZIP Code and organization, as well as added their own annotated data (Stubbs et al., 2015). This team from Kaiser Permanente achieved 87.3% precision, 77.0% recall, and a F_1 Score of 81.8% on the i2b2 PHI categories.

The LIMSI-CNRS team also trained a CRF. They incorporated surface features including token length and punctuation, syntactic features, and leveraged profession lists to identify semantic types. Their approach explored the number of times a string appeared in the corpus and 77 regular expressions (Stubbs et al., 2015). They found that their model using CRF and rules with lexicon was the most successful compared to CRF and CRF with rules but no lexicon. The CRF and rules with lexicon model achieved 89.4% precision, 73.3% recall, and a F_1 Score of 80.6% on the i2b2 PHI categories.

The final team results reviewed in this paper come from a Canadian team. Memorial University of Newfoundland's approach achieved 79.4% precision, 71.9% recall, and a F_1 Score of 75.5% using a Bayesian Hidden Markov Model (HMM) for de-identification (Chen et al., 2015), which is a statistical model that observes outputs to make predictions. Their model allows for “an infinite number of latent variables by implementing a Dirichlet [number theory] process … to identify PHI” (Stubbs et al., 2015, p. s14). They were the only team to exclude rule-based methods in their approach.

Lee et al. (2017) deployed a hybrid approach on psychiatric notes after the challenge was over. Their method to PHI de-identification involved pre-processing (tokenized and POS-tagged the data), a conditional random fields (CRF) tagger that uses “lexical, syntactic, semantic, and discourse level feature types” (Lee et al., 2017), and post-processing attempts to correct errors. Of note, section headers were excluded from their data. Errors encountered include fax numbers being tagged as phone numbers and occupations tagged as hospitals. Their hybrid method achieved an overall performance of 95.0% precision, 89.2% recall, and 92.0% F_1 Score on the 2014 i2b2 data set.

Lastly, Norgeot et al. (2020) developed a hybrid model that approaches de-identification from a privacy-centric lens. It incorporates rule-based methods and machine learning, including pattern matching, statistical prediction modeling, blacklists (i.e., named entities to redact), and whitelists (i.e., named entities to retain; Norgeot et. al, 2020, p. 2). This is an open-source product that can be run on secure environments. Their approach, called Philter, achieved 78.6% precision and 99.9% recall on the i2b2 data set. This recall performance was higher than any team's results in the 2014 i2b2 de-identification challenge (Stubbs et al., 2015).

As demonstrated, researchers have attempted de-identification using a multitude of unique approaches. Despite the extensive and ever-growing list of de-identification approaches, there has yet to be a single gold standard approach recommended for such work. Additionally, it is unclear how this body of research can apply to Manitoba unstructured textual EMR data.

Chapter 3: Methods

This chapter reviews the materials and tools used to support this research. An overview of the Manitoba and i2b2 data sets is provided. The PHI annotation tool and de-identification scrubber are both described. The chapter concludes with an overview of the procedures followed to complete the research.

3.1 Materials and Tools

3.1.1 Manitoba data set

The Manitoba Primary Care Research Network (MaPCReN) of the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) collects EMR data bi-annually from consenting primary care sites in Manitoba. A MaPCReN data analyst performed the data extract; the researchers did not have access to the original EMR data. Unstructured data in image, video, audio, and biometric form are considered out of scope.

A pre-de-identification was necessary for ethics requirements and to permit the researchers to access the data. Once extracted, the data analyst de-identified the data by employing an adaptation of a Bayesian classification scheme, a probabilistic approach to identifying categories, that incorporated pattern matching with regular expressions. This de-identification program is run using Microsoft SQL Server. It generates an output file that mimics the original unstructured data set; however, instances of PHI are replaced with realistic dummy data to create a pseudonymized data set. For example, John Smith may be replaced with Jennifer Bell. The MaPCReN approach replaces the following identifiers with pseudo data: name, phone number, and email. The following PHI are removed by MaPCReN: personal health information number (PHIN), Manitoba Health Registration Number (MHRN), credit card number, and social insurance number. This redacted and replaced data comprised the Manitoba dataset for this study. All pseudo data were treated as true PHI instances, including cases where pseudonymization occurred on non-PHI data (e.g., SHYLA was considered a name PHI instance in “I SHYLA’t really hear any specific murmur”).

An original component of this research that was later removed sought to assess scalability of different rule based de-identification methods. To meet this original need, a sample of 50,000 de-identified encounter notes was provided by MaPCReN. The one inclusion criteria required each encounter note in the sample to contain at least one instance of PHI, as determined by the MaPCReN data analyst. There were no exclusion criteria. This initial sample included encounter notes retrieved in 2018 from two EMR products, Jonoke and Accuro. The date of the encounter note entry was not provided; however, the dataset text has recorded dates for approximately one third of the records, showing that encounter note entry spanned 2003 to 2017 at minimum. The sample was stored on a password-protected computer in a locked office.

The data set with 750 encounter notes was created for the purpose of analyzing Manitoba unstructured textual EMR data and comparing it to de-identification literature. The initial analysis plan involved a comparison between de-identification methods. Based on that plan, a power analysis was conducted in SAS to ensure the sample contained enough PHI instances to achieve adequate statistical power (0.80) for comparing the methods. Allowing a 3% margin of error, 95% confidence interval, and $\alpha = 0.05$, the minimum number of PHI instances required to be in the data set was 2,750. Encounter notes were estimated to contain four PHI instances per note based on MaPCReN pre-processing and in-scope PHI, resulting in a selected data set of 750 encounter notes. A random number generator was used to select these encounter notes. This ensured random selection across EMR products, patient encounter notes, and frequency of PHI instances within the encounter notes.

3.1.2 i2b2 NLP research data set

I2b2 at NCBC produced realistic de-identified and annotated clinical data sets to support research focused on unstructured data. They retrieved unstructured notes from the Research Patient Data Registry at Partners Healthcare and thoroughly de-identified the PHI by the consensus of two experts (Kumar et al., 2015a; N2c2 NLP Research Data Sets, n.d.-a; Stubbs & Uzuner, 2015). The PHI instances were replaced with pseudo PHI and each instance was tagged with a PHI entity.

Their data sets are deemed the gold standard of training data and have been made available for research (Stubbs et al., 2015). Though publicly available, use of data sets is limited to permissioned access through the Department of Biomedical Informatics (DBMI) Portal (*N2c2 NLP Research Data Sets*, n.d.). Researchers register, declare a research purpose, and complete a data use agreement. This research project was granted access to the data sets in 2019.

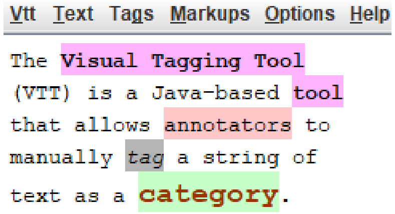
At the time of release, the data set was the first publicly available dataset with pseudo PHI (Stubbs & Uzuner, 2015). Additionally, i2b2 encouraged research on their data sets by facilitating challenges among researchers. For these reasons, i2b2 data sets are used extensively within de-identification literature (Informatics for Integrating Biology & the Bedside, n.d.).

One such challenge was the 2014 De-identification Challenge Data (Kumar et al., 2015; Stubbs et al., 2015; Uzuner & Stubbs, 2015). For this NLP shared task, the annotation guidelines extended beyond HIPPA requirements; among the replaced and annotated data were years, states, countries, patient professions, and any other indirect identifier detailing a hospital or clinician (Stubbs et al., 2015; Stubbs & Uzuner, 2015). The data set consists of approximately 805,000 words or tokens and includes 28,872 instances of PHI. The total corpus contains 1,304 medical notes from 296 patients; the corpus was separated into 790 files for training and 514 for testing the de-identification tools.

Of note, this work migrated to the Department of Biomedical Informatics (DBMI) in the Blavatnik Institute of Harvard Medical School (*DBMI Portal*, n.d.) and the data are now referred to as the National NLP Clinical Challenges (n2c2) Research Data Sets.

3.1.3 Visual Tagging Tool

Visual Tagging Tool (VTT; Lu & Browne, 2010a) is a publicly available Java-based tool that allows annotators to manually tag a string of text as a category (i.e., named entity). It is the annotation tool used by NLM when creating and testing the i2b2 and n2c2 gold corpuses (*De-Identification Tools*, n.d.). The tool allows a user to establish a named entity, in this case PHI, to support tagging of the data set. This is achieved by configuring Tag Properties in the Tags Menu. Further set-up, including visual style displayed on screen (as see in Figure 4), can be completed in the Tag Property Editing section. VTT is a user-friendly tool, requiring few directions that are supplied by the Lexical Systems Group publicly (Lu & Browne, 2010b).



The Visual Tagging Tool
(VTT) is a Java-based tool
that allows annotators to
manually tag a string of
text as a category.

Figure 4. Non-clinical VTT Tagging Example

3.1.4 NLM-Scrubber

NLM-Scrubber is a free de-identification tool designed and developed at the National Libraries of Medicine (NLM) to de-identify unstructured clinical text (U.S. National Library of Medicine, n.d.). It can be downloaded for Linux or Windows and executed with simple configuration in a command-line interface. Configuration options are available to the end user. Customization includes providing a list of terms to preserve, a list of terms to redact, and the ability to turn redaction off for dates, addresses, and age. There is no ability to specify formats of PHI categories, such as date formats. A user manual is freely available online (Raja et al., 2019). Support contact information is provided upon opening of the application.

NLM-Scrubber is widely used in the de-identification literature and in early testing it scored high on sensitivity (Kayaalp, Browne, Dodd, et al., 2014). It identifies and redacts names, addresses, dates, and alphanumerical strings such as telephone numbers and personal health identifiers. Redaction is completed by NLM-Scrubber by replacing the PHI with the name of the PHI categorization (e.g., Doug is replaced by [PersonalName]).

3.2 Procedure

3.2.1 Data set annotation

VTT.2010.0 (Lexical Systems Group, 2010) was installed on Windows. Using the Tag section in the menu bar, Tag Properties were configured for the following PHI: name, PHIN, phone number, address, and date. Tag Property Editing was set up to visually distinguish each PHI category from all other PHI categories. Mapping was completed in the Quick Keys section to enable use of short cut keys for annotation. To support analysis, sub-categories were set-up so that the type of name and address field was distinguishable. The final set up included nine sub-categories: Name – patient, Name – clinician, Name – user, PHIN, Phone number, Address – street, Address – city, Address – postal code, and Date. Instructions for using VTT were developed (see Appendix A).

The primary researcher and an undergraduate pre-medicine student annotated the unstructured Manitoba EMR data. Informed by Stubbs and Uzuner (2015), tagging guidelines were created to support annotation (see Appendix B). Names were separated into three categories: patient, clinician, and username. The context of surrounding words was used to differentiate patient and clinic named entities (e.g., Medical Corporation), as well as identify usernames (e.g., user). The clinician name category included all clinic staff, whether the role is clinical in nature or not, as well as clinician first and last names commonly found in a clinic name.

To complete the tagging, each annotator scanned the encounter notes, highlighted PHI text, and used the quick key function to assign a PHI entity tag (e.g., Name – clinician). The tagging was compared after each annotator tagged the first 60 encounter notes, consistent with annotation efforts seen in the literature (Stubbs & Uzuner, 2014). Following this comparison, the tagging guidelines were updated to include the following:

- Supplementary username detail: “Usernames typically indicated by an initial and name put together in a single string”
- Update to Phone number section: “Fax numbers do not need to be tagged”
- Update to Address section: “P.O. Boxes do not need to be tagged”

Tagging of the first 60 encounters notes were updated to reflect the revised tagging guidelines. The revised tagging guidelines were followed by each annotator for the remaining 690 encounter notes. Tagging results from both annotators were extracted directly from VTT using the Markups Log Details. This detail report outlines the location of the tag, length of the tag, tag name, tag category, the tagged text. The Markups Log Details from each annotator were entered into Excel and compared against location of the tag and tagged category. Conflicts were agreed upon between annotators, tagging of the data was updated accordingly, and a master tagged file was compiled in VTT. A final Markups Log Details Report with all PHI instances was exported from VTT and imported into Excel.

The action of the annotator on each token was compared to the final Markups Log Details Report. Actions were categorized as TP, FN, FP, or TN. For annotation analysis, a TP result identifies when PHI is tagged as PHI, a FN result identifies when a tag is missed for a PHI token, a FP identifies a non-PHI token tagged as PHI, and a TN result identifies when non-PHI token was not tagged as PHI.

Accuracy, the estimated overall per cent of correct PHI tags, was calculated by reviewing the total number of TPs and TNs divided by all PHI and non-PHI instances (i.e., total number of tokens) in the master annotated file. This measure was examined for each annotator at the PHI category level. The proportion of PHI successfully tagged, was calculated by assessing the total number of TPs over all PHI instances (i.e., TPs and FNs). This measure was examined for each annotator at the PHI category and sub-category level. A summary of missed PHI instances (i.e., FNs) is provided in Figure 8 in the results section.

Interrater reliability is reported as percent agreement, the number of tags agreed upon divided by the total number of tokens, as well as Cohen's kappa (ƙ) to enable examination of agreement with consideration of chance agreement. Though there is debate over acceptable Cohen's kappa for health research, a ƙ ≥ 0.80 is generally accepted as a high amount of agreement (McHugh, 2012).

3.2.2 Comparison of i2b2 and Manitoba data set

To assess the applicability of de-identification literature on Manitoba data, differences between the i2b2 corpus and Manitoba data set were examined. Selection method of notes that comprised each data set and independence of the notes were reviewed to assess generalizability of the data sets. Independence of notes within each data set were assessed by examining the proportion of patients that had multiple notes. The variety of note types present in each data set was documented and compared for the i2b2 corpus and Manitoba data set. Prevalence of PHI (i.e., PHI density) was calculated for each data set by determining the number of PHI instances and dividing it by the total number of tokens in the data set. The number of tokens in the Manitoba data set was calculated by summing the number of PHI instances identified during annotation and the remaining strings of text separated by one or more space. The per cent of PHI by category was provided for the i2b2 corpus and Manitoba data set. These percentages were compared between data sets in Table 7 in the results section.

3.2.3 NLM-Scrubber de-identification

NLM-Scrubber v.19.0411W was downloaded for Windows (Raja et al., 2019). The Configuration Editor was set to redact date and address terms and exclude redaction of ages. No redaction or preservation lists were uploaded during configuration. The Manitoba data set was set as the input directory and an output folder was established for the output directory.

Once configuration was complete, NLM-Scrubber was executed. The output results in a redacted data set in .txt format. Adobe Acrobat was used to compare the original Manitoba data set and the redacted data set to produce a report of NLM-Scrubber's actions on each token in the Manitoba data set. To achieve objective three of this research, NLM-Scrubber's actions on the Manitoba data set were assessed by examining how NLM-Scrubber identified a token (i.e., PHI or non-PHI) and compared against the Markups Log Details Report from VTT to determine if it correctly identified the tokens as PHI or non-PHI. Incorrectly redacted name instances from pre-de-identification were treated as PHI for NLM-Scrubber results (e.g., SHYLA was considered a name PHI instance in "I SHYLA't really hear any specific murmur").

The action of NLM-Scrubber on each token was categorized as a TP, FN, FP, or TN. Using the values for TP, FN, FP, and TN, the following performance statistics were computed for each PHI category and overall performance:

Recall: TP/(TP + FN). This statistic, also known as the true positive rate (TPR) or sensitivity, is estimated as the proportion of PHI successfully redacted.

False Negative Rate (FNR): FN/(TP+FN). This statistic is estimated as the proportion of PHI not redacted.

Precision: TP/(TP+FP). This statistic, also known as the positive predictive value (PPV), is estimated as the proportion of redacted data that are PHI.

False Positive Rate (FPR): FP/(FP+TN). This statistic is estimated as the proportion of non-PHI data that have been incorrectly redacted.

Accuracy: (TP+TN)/(TP+TN+FP+FN). This statistic estimates the overall per cent of correct actions completed by the approach.

F_1 Score: 2*((PPV*TPR)/(PPV+TPR)). This statistic is an equally weighted average of the precision and recall completed by the approach.

F_2 Score: 5*((PPV*TPR)/((4*PPV)+TPR)). This statistic weighs recall twice as much as precision.

Additionally, 95% confidence intervals based on the binomial distributions were produced to assess sampling error. The recall statistic was used to test the null hypothesis of 87% or more PHI redacted. For the purpose of this analysis, PHIN and phone number were classified as alphanumeric string categories and all name categories were classified as the personal name string category to align with NLM-Scrubber redaction categories.

Chapter 4: Results

To address objective one, this chapter begins by providing an overview of the Manitoba data set, including PHI types and frequencies. Annotation results are reviewed, including inter-rater reliability, annotator agreement, and individual annotator accuracy and recall. The next section of the chapter addresses objective two by providing an overview of the i2b2 data set, particularly the PHI present in the corpus, and offers a comparison to the Manitoba data set to better understand the applicability of de-identification literature in this local setting. The chapter concludes with a review of NLM-Scrubber’s de-identification performance on the Manitoba data set to meet objective three.

4.1 Manitoba Data Set

The 750 randomly selected encounter notes included a range of clinical documentation such as prescription refill requests, phone call notes, study trial follow-up notes, referral letters and responses, insurance claim inquiry responses, patient vitals and lab result history, and standard visit notes (e.g., Subjective, Objective, Assessment, Plan (SOAP) structure) for a wide range of visit concerns (e.g., wellness check, child development, unresolved and/or worsening medical concerns, mental health, assault follow-up, vaccination/medication administered on site). All notes met the required inclusion criteria of at least one PHI instance per encounter note.

The encounter notes were from 574 unique patients. The mean number of encounter notes per patient was 1.31 (*SD = 0.71)* and ranged from one to six (see Table 2). The mode and median number of encounter notes were both 1.0. The data set contained 190,333 tokens.

Table 2.

Frequency of Encounter Note Occurrences

Number of encounter notes	Number of patients
1	456
2	77
3	30
4	6
5	4
6	1

4.1.1 PHI instances in Manitoba data set

The Manitoba data set contained 3,314 instances of PHI, demonstrating PHI made of a small proportion (1.6%) of the total data. As shown in Figure 5, of the tokens that contained PHI, the *name* category made up 63% (*n* = 2,084) of PHI instances, the *date* category made up 31% (*n* = 1,029), and the *address* category made up 5% (*n* = 154). *Phone number* and *PHIN* represented 1% (*n* = 46*)* and <1% *(n* = 1), respectively.

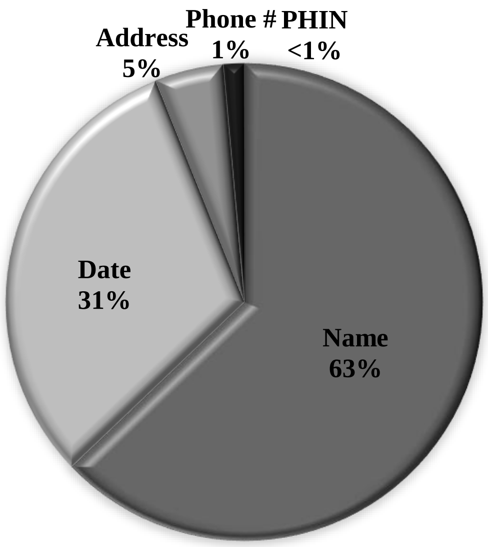


Figure 5. Percentage of PHI Instance types in Manitoba Data Set

The most common PHI subtype was clinician name (*n = 1,511)*, followed by date (*n* = 1,028), as shown in Table 3.

Table 3.

PHI Type and Subtype Frequency in Manitoba Data Set

PHI Type and Subtype	Frequency
Address	**154**
City	92
Postal Code	36
Street	26
Date	**1,029**
Name	**2,084**
Clinician	1,515
Patient	389
User	180
PHIN	**1**
Phone Number	**46**
Total	**3,314**

Nearly one third (31.3%) of PHI instances in the Clinician Name sub-category are due to how the data was exported from the EMR by the EMR vendor. There was a consistent format at the beginning of 237 encounter notes where the clinician first and last name appeared at the beginning of the note, resulting in a total of 474 PHI instances in the Clinician Name sub-category. It is probable that this format was due to how one EMR product exports the unstructured data. Username structure varied, demonstrating a mix of identifying and non-identifying strings. The identifying strings were most often the clinical user's first name (e.g., Carol) or first initial and last name (e.g., bduff). The most common non-identifying PHI tokens were initials (e.g., rjm).

4.1.2 Annotation results

Annotation results are reported for the last 690 encounter notes; the first 60 encounter notes supported training and finalization of the tagging guidelines and thus have been excluded. The 690 encounter notes included 190,333 tokens; of these tokens, 3,023 were PHI. As shown in Table 4, both annotators tagged the same 2,455 items as PHI and the same 187,289 items as non-PHI. Annotator 1 tagged more items as PHI than Annotator 2, though both annotators identified PHI that the other annotator did not. The two annotators had a high amount of agreement. Percent agreement was 99.7% and Cohen's kappa (ꝁ) was also 0.997 (95% CI, 0.92-1.05).

Table 4.

Annotation Results

		Annotator 2		
		PHI	Non-PHI	Row Totals
Annotator 1	PHI	2,455	415	2,870
	Non-PHI	174	187,289	187,463
	Column Totals	2,629	187,704	190,333

Both annotators had high accuracy overall, 99.9% for Annotator 1 and 99.8% for Annotator 2. As shown in Figure 6, Annotator 1 had higher recall than Annotator 2, 93.7% (95% CI, 93.2-94.1%) and 86.4% (95% CI, 85.96-86.8), respectively. Recall results are in line with manual de-identification by clinicians (Neamatullah et al., 2008).

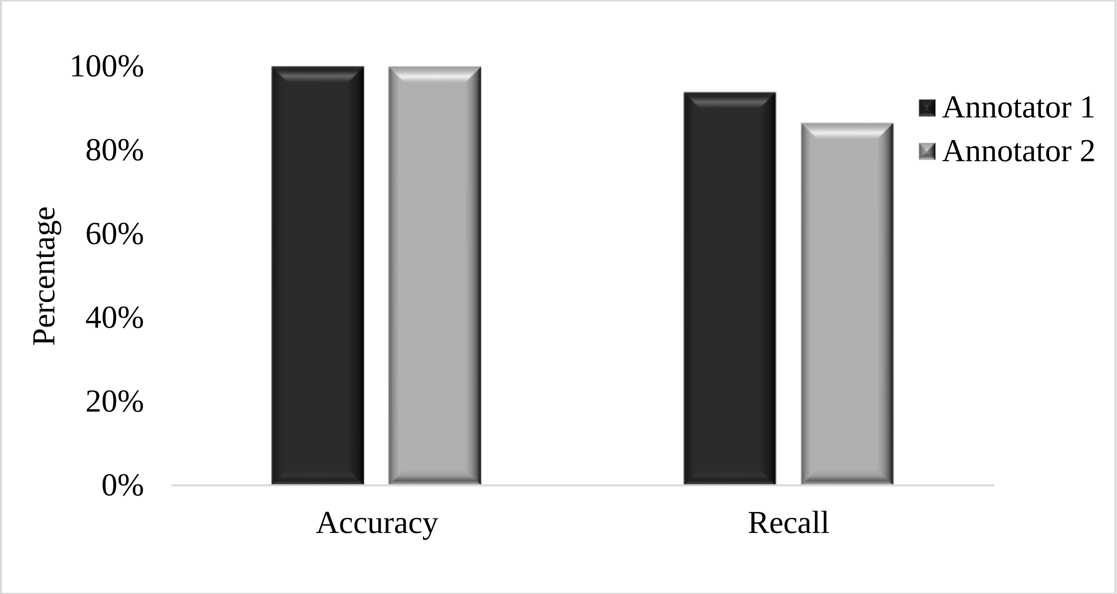


Figure 6. Accuracy and Recall by Annotator

Annotator recall across PHI categories was fairly consistent (see Figure 7). Annotator 1’s recall ranged from 92.1% to 97.3%, with exception to 0% on the single PHIN instance. Recall ranged from 86.0% to 100% across PHI categories for Annotator 2. The annotators had the same recall for phone numbers (91.1%). With exception to this and the one PHIN instance, Annotator 1 consistently had higher recall than Annotator 2 for the remaining categories, ranging from 5.6% to 11.3% higher category recall. Names had the highest number of PHI instances missed by the annotators. Annotator 1 tagged over 100 more instances for both name and date entities than Annotator 2.

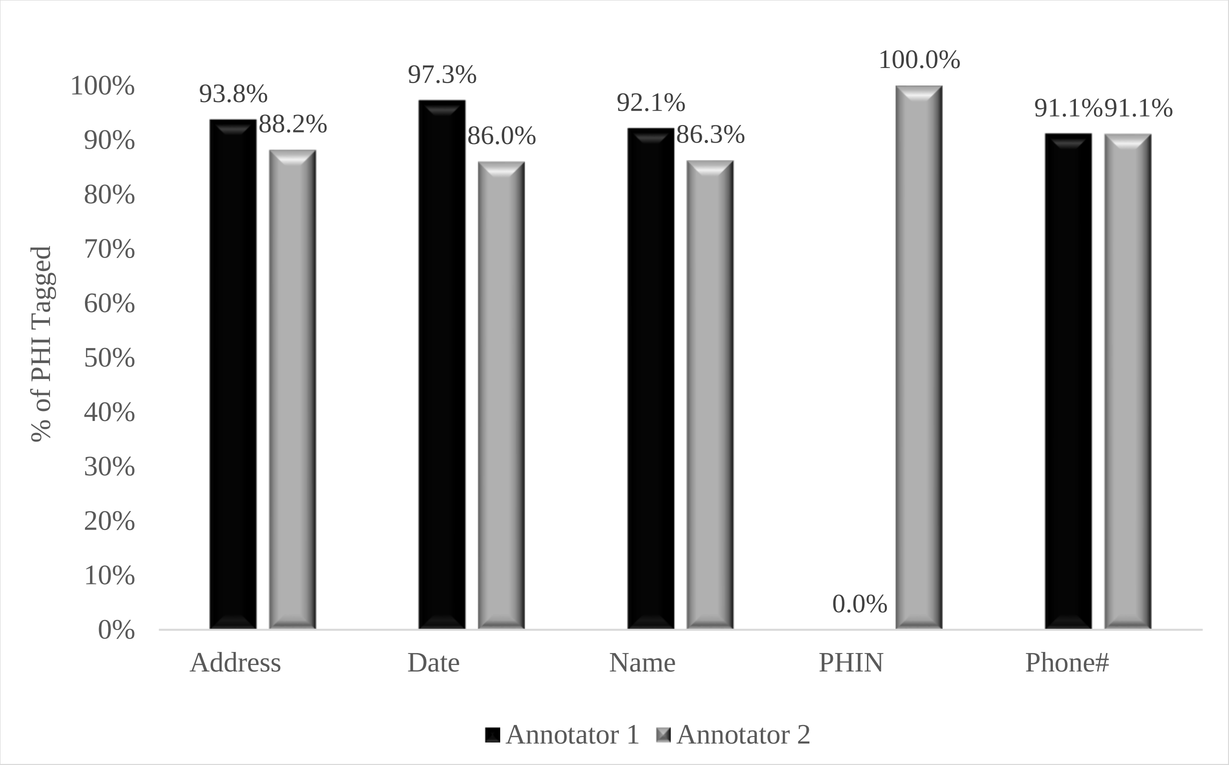


Figure 7. Annotator Recall by PHI Category

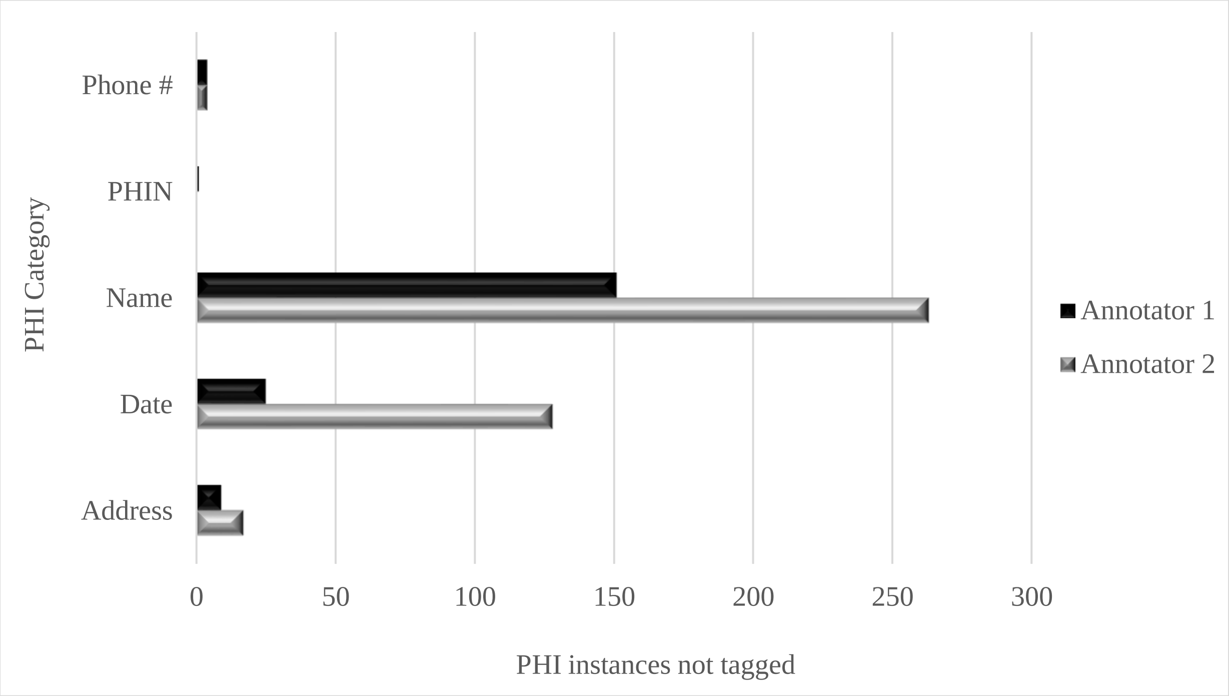


Figure 8. PHI Instances Not Tagged by Annotator

4.2 i2b2 and Annotated Data Set Comparison

The i2b2 corpus was compiled by selecting only records from patients with a documented diabetes diagnosis (Kumar et al., 2015). This was done to support another i2b2 challenge track taking place alongside the de-identification track (Stubbs et al., 2015). This differs from the approach taken to compile the Manitoba data set. Encounter notes in the Manitoba data set were randomly selected from the MaPCReN data that contained one or more instance of PHI. Neither data set was pre-tokenized.

As previously mentioned, the i2b2 data set training is a corpus of 1,304 medical notes from 296 patients; the testing corpus contained 514 notes. Each patient had three to five records within the data by choice to allow for longitudinal analysis. The mean number of records per patient was not provided. Use of random selection to comprise the Manitoba data set resulted in some patient records having more than one record. As shown in Table 2, 118 patients had more than one encounter note included in the Manitoba data set. Of the encounter notes in the Manitoba data set, 60.8% (*n* = 456) were from unique patients and 39.2% (*n* = 294) were from patients with more than one encounter note in the data set.

The i2b2 corpus included unstructured note types of encounter notes and correspondence between medical professionals in the form of discharge summaries, admission notes, transfer notes, and letters. The range of clinical documentation included in the Manitoba data set is previously described in section 4.1 Manitoba Data Set. There are a total of 805,118 tokens in the i2b2 corpus (Stubbs & Uzuner, 2015). Of these, 9,036 were PHI instances for a total PHI prevalence of 1.1%. As previously mentioned in section 4.1.1, the prevalence of PHI in the Manitoba data set was 1.6%. There were seven PHI categories annotated in the i2b2 corpus: name, profession, location, age, data, contact, and IDs. See Table 5 for details on what data was included in each category (Stubbs & Uzuner, 2014).

Table 5.

PHI Categories Annotated in i2b2 Data Set

PHI Category	**Sub-Category**
Name	Patient, Doctor, Username
Profession	-
Location	Hospital, Organization, Street, City, State, Country, ZIP, Other
Age	-
Date	-
Contact	Phone, Fax, Email, URL, IP Address
IDs	Social Security Number, Medical Record Number, Health Plan Number, Account Number, License Number, Vehicle ID, Device ID, Biometric ID, ID Number

Only in-scope PHI instances (i.e., Name – patient, doctor, username; Location – street, city, postal code; date; phone numbers; PHIN) are included in analysis. Identification performance on ZIP Code will be compared to postal codes and performance on Medical Record IDs will be compared to PHINs. The distribution of PHI instances in the i2b2 training data are shown in Figure 9 (Stubbs & Uzuner, 2015). Table 6 shows the frequencies of each sub-category (Stubbs & Uzuner, 2015) .

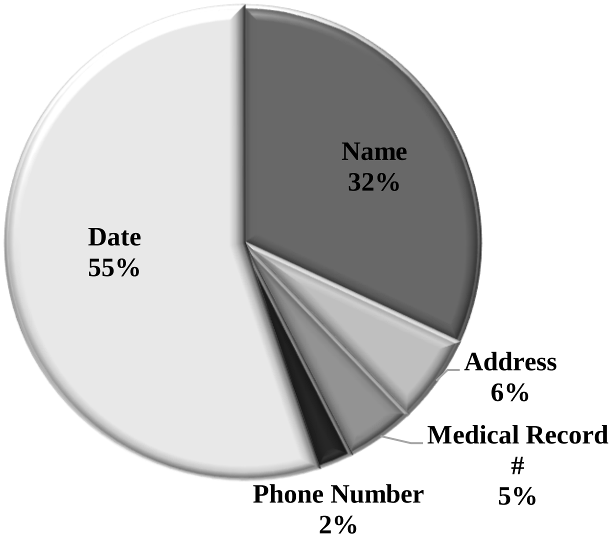


Figure 9. Percentage of PHI Instances by PHI Category in i2b2 Data Set

Table 6.

PHI Type and Subtype Occurrences in i2b2 Data Set

PHI Type and Subtype	Frequency
Address	**536**
City	260
ZIP Code	140
Street	136
Date	**4,980**
Name	**2,883**
Clinician	1,912
Patient	879
User	92
Medical Record	**422**
Phone Number	**215**

Of the PHI instances, the Manitoba data set had a higher proportion of name instances than seen in the i2b2 data, 63% versus 32%, respectively (see Table 10). Dates made up over half (55%) of PHI instances in the i2b2 data set, whereas they made up nearly a third (31%) of PHI instances in the Manitoba data set. As shown in Table 7, the percentage of PHI in the address and phone number categories is comparable.

Table 7.

Percentage of PHI Instances by Category in Manitoba Data Set vs. i2b2 Data Set

PHI Category	Manitoba Data Set	I2b2 Data Set[1]
Name	63%	32%
Address	5%	6%
PHIN/Med Rec #	<1%	5%
Phone #	1%	2%
Date	31%	55%

[1] The sum of i2b2 categories will not equal 100 due to out-of-scope PHI being excluded from the table.

4.3 NLM-Scrubber Redaction

NLM identified and redacted EMR text 4,724 times (see Table 8). Of the 4,724 redactions, 2,499 were PHI instances and 2,207 redactions were non-PHI instances. The remaining 18 redactions were PHI instances redacted more than once (i.e., one PHI instance with two unique redactions).

Table 8.

NLM-Scrubber Redaction by Category on PHI and non-PHI instances

NLM-Scrubber Category	**Redaction Frequency**
Alphanumeric	1,001
Personal Name	2,255
Address	277
Date	1,191
Total	**4,724**

As shown in Table 9, running the NLM-Scrubber tool on the Manitoba data set resulted in 2,499 TPs, 2,207 FPs, 815 FNs, and 197,608 TNs. The overall accuracy of PHI redaction was 98.5% (95% CI, 98.0-98.9%) and the proportion of PHI successfully redacted, the recall, was a 75.4% (95% CI, 72.9-77.8%). NLM-Scrubber demonstrated 53.1% (95% CI, 52.3-53.8%) precision on the Manitoba data set.

Table 9.

NLM-Scrubber Redaction Results

		Data Set		
		PHI	Non-PHI	Row Totals
NLM-Scrubber Tags	PHI	2,499	2,207	**4,706**
	Non-PHI	815	197,608	**198,405**
	Column Totals	**3,314**	**199,815**	**203,129**

Examining redaction by category reveals NLM-Scrubber's recall ranged from 58.9% for date to 100% for PHIN (See Figure 10).

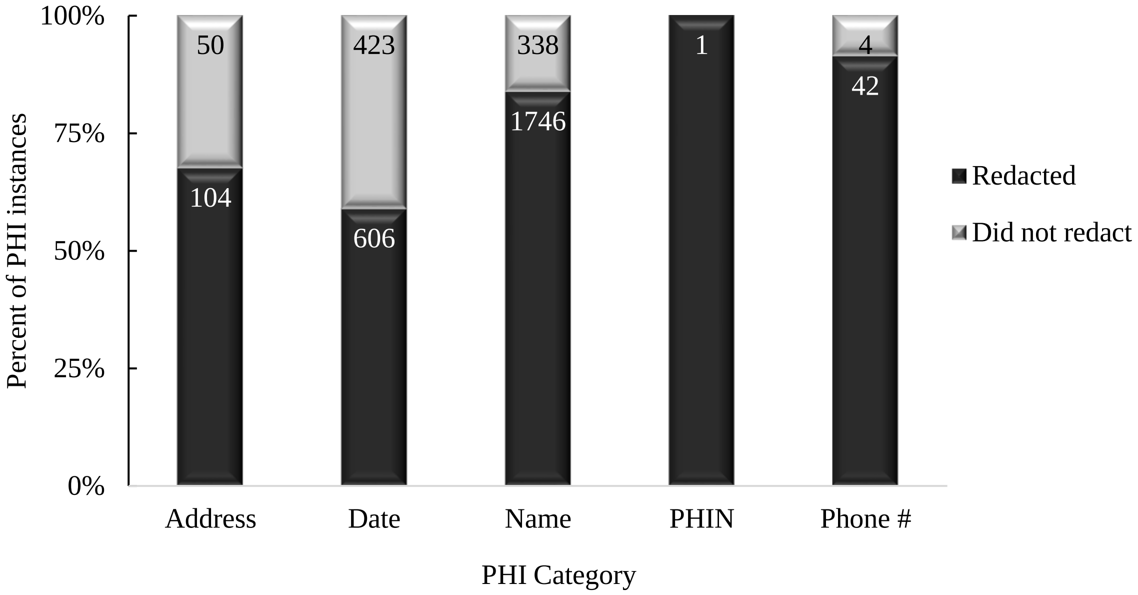


Figure 10. NLM-Scrubber's Redaction Results by PHI Category

NLM-Scrubber's accuracy for each category ranged from 98.5% to 100% (see Table 10). With exception to the single PHIN instance, it most accurately redacted address instances.

Table 10.

NLM-Scrubber Performance Statistics Overall and by Category

	Statistic (95% CI)					
	Overall	**PHIN**[1]	**Phone #**	**Name**	**Address**	**Date**
Recall	**75.4%** (73.9-76.9%)	100.0%	**91.3%** (90.9-91.6%)	**83.8%** (83.7-83.8%)	**67.5%** (67.2-67.7%)	**58.9%** (58.8-58.9%)
False Negative Rate	**24.6%** (10.0-39.2%)	0.0%	**8.7%** (8.4-9.0%)	**16.2%** (16.1-16.2%)	**32.5%** (32.2-32.7%)	**41.1%** (.410-.411%)
Precision	**53.1%** (51.7-54.5%)	100.0%	**4.2%** (4.1-4.2%)	**77.5%** (77.4-77.5%)	**39.0%** (38.8-39.2%)	**51.1%** (51.0-51.1%)
False Positive Rate	**1.1%** (1.0-1.2%)	0.0%	**0.5%** (0.5-0.5%)	**0.3%** (0.3-0.3%)	**0.1%** (0.1-0.1%)	**0.3%** (0.3-0.3%)
Accuracy	**98.5%** (98.4-98.5%)	100%	**99.5%** (99.4-99.5%)	**99.6%** (99.5-99.7%)	**99.9%** (99.8-99.9%)	**99.5%** (99.4-99.5%)
F_1 Score	**62.32** (61.0-63.6)	100.0	**8.0** (7.9-8.0)	**80.5** (80.4-80.5)	**49.4** (49.2-49.6)	**54.7** (54.7-54.7)
F_2 Score	**69.56** (67.9-71.2)	100.0	**17.7** (17.6-17.7)	**82.4** (82.3-82.4)	**58.9** (58.7-59.0)	**57.2** (57.1-57.2)

[1] CIs not applicable

Overall, recall for instances in the Name category was 83.8%. Recall performance at the sub-category level for Clinician, Patient, and username was 92.2%, 85.1%, and 10.0%, respectively. The majority of usernames (90.0%) were not redacted by NLM-Scrubber. Excluding usernames, recall performance for Clinician and Patient instances was 90.8%. There was no consistent theme for name instances that were redacted, nor instances missed. NLM-Scrubber consistently redacted and replaced the Name instances with the [PERSONALNAME] category, demonstrating the correct categorical redaction. The name category had the second lowest FNR (16.2%) and a lowest FPR (0.3%).

Date had the highest FNR. NLM-Scrubber consistently redacted dates in the following format: YYYY (e.g., 2007), MMM (e.g., Jul, July), MMM DD, YYYY (e.g., Dec 19, 2007), MMM DD (e.g., Jan 10), and MMM DDth (e.g., Feb 6th). Formats consistently missed were: MMDDYY (e.g., 10272002), DD-MMM (18-Apr), MMM.D (Nov.4), DD-MMM-YYYY (24-Aug-2009), and strings follow DDth of Month (e.g., 26th of December). NLM-Scrubber demonstrated 51.1% precision; it redacted instances of years, values out of scope where no other date information accompanies the year.

Overall, recall for instances in the Address category was 67.5% and precision was 39%. Recall performance at the sub-category level ranged from 0% to 100%. NLM-Scrubber failed to redact postal codes but successfully redacted 100% of street instances. The most common missed PHI instance in the Address category was postal codes. This accounted for 70.0% (*n* = 35*)* of missed Address PHI instances. NLM-Scrubber demonstrated an 84.8% recall of the city names. Carmen, Winkler, and Winnipeg were inconsistently redacted from the encounter notes. Although they were redacted, 42.9% *(n = 66)* of instances in the Address category were tagged with the wrong category, often redacting an address and replacing it with [PersonalName]. This occurred for 69.6% (*n* = 64*)* of city names instances and 7.7% (*n* = 2) of street instances.

A single PHIN instance was detected in the data set. This string was identified as an Alphanumeric value and redacted. Phone numbers were also treated as Alphanumeric strings by NLM-Scrubber. It demonstrated 91.3% recall and 4.2% precision. Phone number category had the lowest precision of all categories (4.2%) due to NLM-Scrubber redacting fax numbers, values out of scope of this project.

Incorrect redaction of non-PHI data occurred on various string types. Examples of incorrect redaction as appeared in the data are shown in Table 11. In examining redaction errors, both FPs and FNs were observed. There was no identifiable consistency in the type of NLM-Scrubber redaction errors on Manitoba data.

Table 11.

Non-PHI Redaction Examples

PHI Category	**Non-PHI Redactions**
Name	Day, CREA, FBS, Foodguideline, Gluco-FASTING, Hawkins, Jobes, Lipitor, Needs, Neers, Please, Rash, Smoker, TAB, TAB ID, Tab Refills:, Tab QD Vapo-rub
Date	5, 12:
Address	S:, Height, :22 PM
Alphanumeric	NAPROXEN-375, BP[120]/[70], LDL[3.06]

Chapter 5: Discussion & Conclusion

This chapter discusses the types and frequencies of PHI found in Manitoba unstructured textual EMR data, the applicability of existing supervised machine learning de-identification literature on Manitoba unstructured EMR data, and NLM-Scrubber's de-identification performance on the Manitoba data set. Study strengths and limitations will be discussed. The chapter concludes with directions for future research.

5.1 PHI in Manitoba Data Set

Though the Manitoba data set was altered through pre-de-identification by the MaPCReN data analyst, PHI was replaced with pseudo PHI. This replacement made the Manitoba data set a suitable reflection of the true data. An exception to this is that PHINs were removed from the data. It is probable that the one PHIN in the data was missed due to unusual spacing between characters and is likely that some types of typical unstructured Manitoba EMR data contain more PHINs (e.g., consult letters).

Name instances were the most common PHI type present in the Manitoba data making up over half of all PHI instances. Patient name and names of their family members were mostly documented within standard visit notes or correspondence between clinical staff (e.g., to address patient phone call or prescription refill request), reflecting standard capture of information to support patient care. The direct identifier categories represented 11.0% and 1.4% of all PHI instances, respectively. Pre-de-identification processing resulted in a small number of FN name instances which were treated as PHI instances. Consequently, the number of name PHI instances may be slightly lower in typical unstructured EMR data. Most of the remaining PHI instances were indirect identifiers of the care provider or location of care.

The Manitoba data showed the variety of content that can be found within an encounter note. The range of information and structure within the notes reveals that where and how content is stored is influenced by EMR product configuration and provider and/or clinic approach. For example, it is likely that health numbers in unstructured data would be due to semi-structured data like lab results and referral letters being included in an unstructured manner in the data export. It is not necessary, nor standard practice, for a clinician to type a patient's health numbers into free text encounter notes as this data is usually captured in structured data elsewhere in the patient's record.

The Manitoba data set was chosen to include encounter notes with one or more instances of PHI. While the prevalence of PHINs in a complete Manitoba unstructured EMR data set is unknown, the overall prevalence of PHI is likely to be less than this research found once encounter notes with no instances of PHI are included.

5.2 Applicability of Literature to Manitoba Setting and De-identification Considerations

5.2.1 Applicability of i2b2 literature

The i2b2 2014 de-identification challenge was one of the first of its kind, resulting in a large proportion of de-identification literature focused on redaction of their gold standard corpus. This allows us to consider the applicability of much of the existing de-identification literature to Manitoba unstructured EMR data.

The inclusion criteria for selecting records that comprised the i2b2 corpus may limit its generalizability to general unstructured textual EMR data. All records were required to have a recorded diabetes diagnosis and mention of CAD risk factors (Kumar et al., 2015). The random selection method chosen to compile the Manitoba data set may offer a data set that contains encounter notes more representative of what is currently found in EMR unstructured textual data.

Most of the encounter notes in the Manitoba data set come from unique patients and are independent of other notes in the data set. Only 39.2% of encounter notes come from patients with more than one note in the data set, whereas all encounter notes in the i2b2 corpus are from patients with other data in the corpus. The notes included in the i2b2 corpus include encounter notes and some correspondence types between medical professionals. The Manitoba data set included a broader range of clinical documentation, such as the inclusion of prescription refills and lab results. Additionally, the data was collected from encounters that span a long range of time (at least 14 years) and from more than one EMR source. These factors are likely to provide a wider variety of health concerns and diagnoses to be included in the dataset. The above factors provide a more realistic view of typical unstructured EMR data and offers a more comprehensive data set for testing how a de-identification approach works on differing text than the i2b2 corpus offers.

The Manitoba data set had a higher PHI density than the i2b2 corpus, 1.6% and 1.1%, respectively. The higher prevalence in the Manitoba data may have been due to the inclusion criteria that an unstructured note must include one or more instance of PHI. It is possible that a lower prevalence of PHI would be seen if unstructured notes without PHI were included; however, the density found in this Manitoba data set aligns with PHI density observed in Swedish clinical text (Henriksson et al., 2017).

Of the PHI in data sets, Manitoba has a higher proportion of names requiring de-identification; the percentage of name PHI was nearly double than observed in the i2b2 corpus. Among the name sub-categories, the i2b2 data set had a higher proportion of patient names. This distribution of names in both data sets demonstrates the importance of de-identification approaches to focus on name tokens. The i2b2 corpus had a higher percentage of PHI in the date category than in the Manitoba data set (55% and 31%, respectively). Based on the second objective findings, caution should be used when interpreting de-identification literature for application in the Manitoba setting due to the notable difference in PHI category percentage for name and dates, as well as difference in generalizability of the data sets.

5.2.2 Supplemental de-identification considerations

Knowing Manitoba's data has a high proportion of names and that the current literature recommends supplementing supervised machine learning with data dictionaries where possible (Stubbs et al., 2015; Uzuner et al., 2007), de-identification efforts are likely to be more successful when de-identifying Manitoba unstructured data using a hybrid approach with name lists. Customizable de-identification approaches may allow for use of location-specific lists, support domain adaptation, and enable use of the method across data sets (i.e., non-EMR data; Uzuner & Stubbs, 2015).

One such approach that has been shown to be most successful is a trained CRF model that incorporates pre-processing of the data. Pre-processing prepares the data for de-identification, cleaning up common data entries that may influence de-identification performance (e.g., spacing between tokens), as well as tokenizing the unstructured data (Uzuner & Stubbs, 2015). The Manitoba data set contained one PHIN, likely missed due to the unusual spacing between characters. Pre-processing Manitoba data may correct unusual spacing and reduce missed PHIN instances in future local work. The supervised component of the machine learning approaches incorporated data dictionaries and rule-based methods. Uzuner and Stubbs (2015) noted that the successful approaches also performed well on an earlier challenge, suggesting that the trained CRF model performance is maintained across data sets of varying data and task complexity.

NLM-Scrubber's de-identification has shown poorer performance in recent literature (Steinkamp et al., 2020) compared to earlier research available at the onset of this study. The same lower performance was demonstrated on Manitoba's data. As de-identification continues to be attempted on a broader variety of data, these findings reinforce Norgeot et al.'s (2020) recommendations that de-identification approaches should be customizable. They argue that the unique structure and context of data sets, as well as the purpose and intended use of de-identification, vary greatly and thus require a customizable approach to be of value to researchers. A customizable approach that incorporates local nuances (e.g., exported structure specific to local digital medical record, local name lists) will further efforts to develop a method that successfully de-identifies unstructured text to meet legislated privacy requirements. Privacy subject matter experts could provide information on the nuances and considerations unique to the location and data set in question, informing customization and configuration; this engagement may relieve concerns regarding use of unstructured text for secondary purposes. Lower performance may also be a reality with off the shelf de-identification models. This may be a consideration if off the shelf models are explored for future de-identification efforts in Manitoba.

5.3 NLM-Scrubber Redaction on Manitoba Data Set

The initial intent of this research was to look at rule-based de-identification methods. At the onset of this research, NLM-Scrubber was a promising rule-based de-identification tool. It had been shown to have high recall and precision de-identifying PHI. It was anticipated that this same performance could be achieved using this model on Manitoba data providing a supplemental indication of the likelihood that findings from the literature may be transferable to the Manitoba context. The field has since evolved, demonstrating that sophisticated machine learning de-identification approaches are more successful than rule-based methods alone. The de-identification performance of NLM-Scrubber on Manitoba data is still applicable to the current state however as rule-based methods are still predominately used for de-identification.

The null hypothesis that NLM-Scrubber will de-identify ≥87% of PHI in Manitoba unstructured textual EMR data was rejected; NLM-Scrubber achieved an overall recall of 75.4%. Overall recall was lower on the Manitoba data set than seen on the i2b2 data (87.8%; Norgeot et al., 2020). The tool showed successful redaction on some PHI types in the Manitoba data, such as over 90% recall and accuracy on phone numbers and high recall on clinician and patient name sub-types (92.1% and 85.1%, respectively). Despite these positive results, NLM-Scrubber had challenges de-identifying other PHI types, such as date and address. NLM-Scrubber's de-identification performance on postal code was low, missing 70% of addresses, and identifying the correct PHI category for addresses was a challenge. Additionally, NLM-Scrubber failed to de-identify the Manitoba data as well as other data sets. As previously mentioned, Kayaalp et al. (2014) reported NLM-Scrubber achieved 99.9% recall on names, whereas the tool achieved weaker recall on names overall in the Manitoba data (83.8%). This is due to its low recall on usernames (10%). These results highlight that off the shelf de-identification tools are unlikely to yield the performance required to meet local privacy standards. Using local name lists may help yield higher recall results. For instance, a list of EMR usernames at primary care sites submitting data to MaPCReN could be compiled and used as a blacklist for de-identification methods. A data dictionary may also help increase identification and redaction of addresses and postal codes. Adjusting training algorithms to recognize the typical character string for postal codes is another approach that may easily result in higher de-identification performance; character strings have been shown to be easier for redaction than non-character string PHI types (Bui et al., 2017).

Additionally, the main challenge with de-identification is removing PHI while retaining relevant data. On the Manitoba data, nearly half (*n* = 47%) of NLM-Scrubber's redactions were on non-PHI instances. NLM-Scrubber demonstrated much lower precision on the Manitoba data than on the i2b2 data, with estimates of 53.1% and 76.3%, respectively (Norgeot et al., 2020). NLM-Scrubber's over-redaction limits the usability of the redacted data set and creates a scenario with high risk of removing contextual and clinical information important to data mining and analysis. This was seen in the redacted data. Vital signs, lab values, details of prescribed medicine, and clinical descriptors and other supporting information such as rash, smoker, and height are examples of non-PHI redacted by NLM-Scrubber.

Though early research showed good de-identification performance by NLM-Scrubber, more recent research suggests its performance on different types of data is not as strong. Steinkamp et al. (2020) found that NLM-Scrubber achieved 87.5% overall but only 57.8% on patient names on radiology reports. Recall reported for other PHI categories were as follows: addresses 68.3%, phone numbers 95.6%, and dates 97.5% (Steinkamp et al., 2020). When comparing NLM-Scrubber's performance on Manitoba's data to its performance on radiology reports, it achieved higher recall on the Manitoba data for names but lower recall for addresses, phone numbers, and dates. NLM-Scrubber has been shown to miss redacting usernames (Norgeot et al., 2020; Steinkamp et al., 2020); this was also observed on the Manitoba data. NLM-Scrubber achieved a 10.0% recall on usernames in the Manitoba data, which made up 8.6% of PHI name instances. This demonstrates that de-identification performance on first or last names can vary greatly from performance on usernames. Future de-identification approaches should consider methods to identify full or partial named entities strung together to increase the likelihood of de-identifying usernames.

Further hindering the data set, NLM-Scrubber redacted PHI using the wrong category. This means the de-identified data set would show the wrong PHI type in a sentence. This loss of sentence context influences interpretability of the data; examination of the data set post-redaction could lead researchers to misinterpretation or findings that do not reflect true state. Not all incorrect categorization has the same impact; the impact of incorrect categorization would be greater for some PHI categories than others. For instance, the accuracy of PHI types for both the redacted PHI and the replacement tag is more important for Name than other PHI types (e.g., phone number, occupation). NLM-Scrubber did consistently replace name instances with the correct category. Additionally, approaches that replace PHI with pseudo data help mask missed PHI and are recommended over NLM-Scrubber's redaction approach (Murugadoss et al., 2021).

A review of the literature revealed that approaches with acceptable de-identification performance incorporate data pre-processing prior to performing de-identification actions (Stubbs et al., 2015). The NLM-Scrubber tool does not incorporate data pre-processing (Raja et al., 2019; U.S. National Library of Medicine, n.d.). To support comparison of the NLM-Scrubber results on Manitoba data to performance on the i2b2 corpus, no pre-processing was completed on the Manitoba data set. Introducing a pre-processing step on the Manitoba data set may have interfered with the reliability of these results. Future de-identification efforts should explore data pre-processing to improve the structure and quality of data, such as incorporating methods that organize the data through tokenization or POS tagging, or reorganize text where punctuation or character spacing may interfere with de-identification.

The F_1 Score can be a helpful metric for machine learning binary classification models. Compared to other performance statistics such as accuracy, it is a statistic that maintains focus on the most essential data, PHI, rather than the majority of data which are non-PHI. The F_1 Score equally weights recall and precision and is important as a standardized metric to compare across studies. NLM-Scrubber's overall F_1 Score on Manitoba data was 62.3%, reflecting the balance between 75.4% recall and 53.1% precision. A higher F_1 Score would need to be achieved to assure both researchers and privacy experts that the data set is suitable for use. There is a need to balance patient privacy and retention of important clinical context. For instance, a method that continually identifies strings as PHI would not be helpful to researchers. It is worth noting that there is no clear standard as to what constitutes as an acceptable de-identification performance. While Ferrández et al. (2012) state that recall is the most important performance measure, there seems to be consistent acknowledgement that the importance and achievement of performance statistics like recall and precision depend on the context of the research or intended use of the pseudonymized data set.

5.4 Limitations and Strengths

This research had some limitations. In order to meet the requirements of PHIA legislation, the Manitoba data set is scrubbed of PHIN and MHRNs. With exception to the one PHI instance, the Manitoba data set did not include health record numbers. Consequently, assessment of NLM-Scrubber performance on Manitoba health numbers was not completed. The Manitoba PHIN and MHRN would meet the previously mentioned definition of an alphanumerical string, which appears to be among the easiest type of PHI entity to identify.

This study was unable to assess performance using data dictionaries configured for Manitoba-specific PHI, such as postal codes, city names, and hospital names. Much of the literature involves customizable approaches and current data dictionaries; it is likely that redaction would have seen higher success if these location-specific lists were incorporated into the redaction method.

HIPPA advises to exclude year as PHI except where patients are over 89 (U.S. Department of Health & Human Services, n.d.-b). Following these guidelines, year was considered non-PHI when it existed without mention of the associated month. NLM-Scrubber redacted years; this redaction was counted as a FP. Recording these redactions as a FP may result in an over-inflation of FPs and lower performance on metrics involving the FPs. Literature varies on inclusion of year in isolation as a PHI entity and some researchers choose to de-identify it to err on the side of caution. When determining the inclusion or exclusion of date PHI fields, privacy subject matter experts should be consulted for local guidance and use of the de-identified data should be considered. For instance, a more risk-adverse approach may be preferred for data released to the public.

NLM-Scrubber's redaction results export without a location identifier. Analyzing NLM-Scrubber's de-identification results involved a mix of software and manual effort due to this limitation. As with any manual task, there is risk of human error in recording of the results. Tools are beginning to emerge that will support evaluation of redaction results in the future (Heider et al., 2018).

Despite these limitations, there were a number of strengths of this research. Firstly, the Manitoba data set contained pseudo PHI. This ensured the privacy of patients with encounter notes in the data set. Secondly, the data set proved to be robust and is anticipated to reflect true state. It contained a broad variety of note types, across multiple EMR products, and encounter notes collected over a span of more than a decade, as well as a high number of encounter notes independent from other data. The approach to tagging the Manitoba data set was also a strength of this research. The tool used to tag the data set was the same tool used by NLM to test their gold standard corpuses. Best practices were followed for the annotation procedure including development and revision of the Tagging Guidelines, and the process closely aligned with the approach taken to annotate i2b2 corpus. Lastly, the approach chosen to assess de-identification performance on Manitoba unstructured EMR data was a widely accepted rule-based tool viewed to be a standard at the onset of this research.

5.5 Directions for Future Research

Review of the literature and results from this research identify opportunities for future research that would further efforts to de-identify unstructured textual EMR data and enable its use for secondary purposes. These opportunities include focused efforts to improve de-identification of indirect PHI, strengthening the body of de-identification literature by detailing redaction errors and not just redaction successes, better understanding factors that influence the likelihood of identifying residual PHI and establish acceptable limits of residual PHI, and understanding the usability of deep learning for de-identification. There is also an opportunity to learn how to minimize data quality concerns that impact de-identification and the best pre-processing techniques to address them when they exist. Lastly, there should be continued efforts to share data between researchers to support de-identification work, expand the data sources in existing data repositories, and broadened access to, or sharing of, repository data. This will position health researchers to maximize the benefits of linking unstructured EMR and administrative data once unstructured EMR data becomes more readily accessible.

Data trustees and custodians have been cautious when releasing de-identified data sets to the public, as seen with the i2b2 corpuses. Often profession is present in unstructured EMR data, a PHI entity that fits under the generic HIPPA category “any other unique identifying number, characteristic, or code” (U.S. Department of Health & Human Services, n.d.-b). Redaction of organizations and professions has shown to be one of the most challenging entities to de-identify (Dehghan et al., 2015; Liu et al., 2017). Furthermore, indirect profession PHI may be more prevalent in unstructured data than earlier studies suggest. Stubbs, Filannino, and Uzuner (2017) expanded the definition of professions for the 2016 de-identification challenge on psychiatric intake notes. The profession PHI category included patient and family member professions, as well as tagging of generically named profession (e.g., gas station) to reduce risk of patient re-identification when combined with other information. This category expansion resulted in the 2016 corpus containing six times as many professions as the 2014 corpus (2481 versus 413, respectively). Literature on the challenges of de-identifying these indirect PHI is limited, likely in part due to the entity not being a specified PHI category in HIPPA. Liu et al. (2017) acknowledge the difficulty of detecting profession due to the lack of standard formatting and organizations due to the use of abbreviations. The authors suggest use of a module to support handling abbreviations. They did not offer a solution for detecting professions and noted that dictionary-matching failed to help detect profession instances (Liu et al., 2017). The above demonstrates a need to better understand the challenges of de-identifying these PHI entities and possible solutions to maximize detection in data.

Another gap in de-identification literature is the lack of detail on redaction errors such as missed PHI or incorrect redaction. Articles provide de-identification results, often focusing on an approach's performance such as the proportion of redactions and related performance metrics. There is little discussion on missed PHI instances and potential causes for the missed PHI are rarely detailed. Describing the errors in a fulsome manner would enable researchers to identify common themes encountered and focus on areas requiring improvement. It would also allow researchers to better identify where other approaches succeed and fail, enabling a collaborative approach that leverages a combination of methods to balance weak areas to increase the likelihood of developing one approach that's highly successful across varying domains and applications. Future work could include a systematic review of missed redactions by de-identification approach. This would provide further insight into errors that need to be addressed to meet privacy subject matter experts' concerns, as well as help identify information gaps in the de-identification domain.

Additionally, research should further explore identification and acceptability of residual PHI, which are PHI that remain in a data set after de-identification processing has been completed. Carrell et al. (2013) noted that while 87-91% of residual identifiers could not be identified, factors found to influence identification of the residual indicator included contextual information such as reader's prior experience with the information and other content within the data, and sophistication of replacement technique (e.g., replacing first and last names with names from discordant national origins). Further research could strive to better understand the factors that lead to identification of residual PHI and how they can be mitigated, such as more robust replacement techniques, expansion of data dictionaries, or review of data that should be treated as indirect PHI. Furthermore, residual identifiers are likely to always remain in de-identified data sets. Establishing acceptable limits of residual identifiers based on the research purpose and data use may help guide decision makers in deeming a data set sufficiently de-identified for secondary use.

There is additional opportunity to collaborate through sharing of patient-level data from clinical trials, an intriguing area with increasing interest (Committee on Strategies for Responsible Sharing of Clinical Trial Data, 2015). Pseudonymized data sets would be made accessible in a database for researchers with authorized access. This would provide researchers with detailed data from a vast range of clinical research including unstructured notes, data not readily available to researchers today. Machine learning techniques could be applied to more extensive data than a single clinical trial allows, thus allowing more comprehensive analysis to uncover new clinical findings, such as linkage between early signs and symptoms to outcomes and prevention strategies. These findings could also be linked to administrative data or other forms of structured data and applied to cross-jurisdictional corpuses. Though a population research data repository exists in Manitoba for Manitoba residents (Manitoba Centre for Health Policy, n.d.-b), it does not yet include data from clinical trials or unstructured textual EMR data. Additionally, a single national data repository is not yet available. There is tremendous opportunity if this idea comes to fruition on a large scale, as evidenced by Canadian Institute for Health Research (CIHR)'s support to create a national data platform (Dahl et al., 2020); however, challenges with sharing data cross-jurisdictionally need to be addressed (Katz et al., 2018). The untapped opportunity of sharing of new data set types and across multiple jurisdictions could greatly further population health research and health policy efforts. It is also worth noting that there are benefits of sharing data sets for de-identification efforts. Lee et al. (2017) suggest that leveraging an additional data set can strengthen redaction results on PHI entities by supporting domain adaptation, further supporting the idea of sharing health data sets between researchers.

Nesca (2021) discusses various data quality concerns in unstructured data. Addressing the data quality of unstructured textual EMR data to support de-identification can be approached by minimizing the quality errors at time of entry in the EMR, extract of data from EMR, or post-extract using pre-processing techniques. Nesca (2021) suggests future work should define data quality dimensions for unstructured EMR data. This would provide a framework for examining data quality of unstructured EMR data and enable researchers to assess if quality concerns can be addressed at time of entry (e.g., establish best practices for data entry) or during the export process (e.g., EMR vendor adjustments). The third opportunity to address quality concerns is through data pre-processing after unstructured data are extracted from the EMR. This helps organize the data and minimize common data quality concerns that may impact de-identification processing. As previously mentioned, the most successful supervised machine learning de-identification approaches incorporate data pre-processing (Kristianson et al., 2007; Uzuner & Stubbs, 2015). Future research should explore the impact of different types of pre-processing and if the order of pre-processing steps influences de-identification performance. It would also be useful to understand how this varies based on the de-identification approach and what combination of pre-processing order and de-identification technique yields the best de-identification performance.

Although use of unstructured data is exciting for future research, structured data fields will continue to be widely used in healthcare. This is important because structured fields will be used alongside unstructured data, yet research has demonstrated that some may lack the completeness required to be reliable for research. Unstructured data can be leveraged to identify and improve data quality issues in these structured data fields. For instance, NLP methods can be used to detect potential missed diagnoses or medications. This creates an opportunity for an additional source to identify quality concerns in structured data (Bowen, 2012). Additionally, context-rich unstructured data offers an opportunity to gain a deeper understanding of structured data.

Lastly, an emerging trend in de-identification is the use of deep learning. Data-driven deep learning approaches eliminate the need for researchers to manually train approach algorithms (Xiao et al., 2018). They have shown better de-identification performance than machine learning methods (Janiesch et al., 2021; Trienes et al., 2020). The lack of required training enables deep learning methods to be used across a variety of tasks; however, not pre-training the data leads to uncertainty on how a model will perform on a new data set, limiting the transferability of each approach (Janiesch et al., 2021; Xiao et al., 2018). More research is necessary to understand the usability of deep learning methods for de-identification. Deep learning requires an extensive amount of data to learn and achieve maximum performance, and thus, may not be suitable for some research. Like humans, biases have been noted in deep learning algorithms (Janiesch et al., 2021); there is a need to better understand this to prevent missed PHI and ultimately, protect patient privacy.

5.6 Conclusion

Manitoba unstructured EMR data contain a variety of direct and indirect PHI. Differences between it and the widely used gold standard i2b2 data set suggest that de-identification literature may have limited applicability to Manitoba EMR data. Use of NLP approaches to automate de-identification is a rapidly growing field. Customizable hybrid models are showing greater success than rule-based methods alone. The NLM-Scrubber tool showed weaker de-identification performance on the data than seen on i2b2 data. This research confirmed that an off the shelf rule-based de-identification tool may not be acceptable for use.

Manitoba is well positioned to leverage unstructured EMR data with existing access to MaPCReN data, a data set with extensive history of patient clinical encounters retrieved from multiple EMR sources. Linkage of the unstructured EMR data from MaPRCeN to other types of administrative data creates opportunity for population health and health services research. For example, it could provide new insights into disease risk and population surveillance, prevention strategies, or where best to invest public funds to achieve improved population health outcomes. Attention should be directed to trained, hybrid machine learning solutions that enable customization, including adjustment of rule-based methods and data dictionaries, and pseudonymization to help protect patient privacy. Automated de-identification should be actively pursued by researchers and health policy makers to further advance health research and improve population health.

References

Abhyankar, S., Demner-Fushman, D., Callaghan, F. M., & McDonald, C. J. (2014). Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. *Journal of the American Medical Informatics Association, 21*(5), 801–807. https://doi.org/10.1136/AMIAJNL-2013-001915

Aramaki, E., Imai, T., Miyo, K., & Ohe, K. (2006). Automatic deidentification by using sentence features and label consistency. *I2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*. http://www.chasen.org/

Article 29 Data Protection Working Party. (2007). *Opinion 4/2007 on the concept of personal data*.

Azam, L. S., Jackson, T. A., Knudson, P. E., Meurer, J. R., & Tarima, S. S. (2016). Use of secondary clinical data for research related to diabetes self-management education. *Research in Social and Administrative Pharmacy*. https://doi.org/10.1016/j.sapharm.2016.07.002

Banerji, A., Lai, K. H., Li, Y., Saff, R. R., Camargo, C. A., Blumenthal, K. G., & Zhou, L. (2020). Natural language processing combined with ICD-9-CM codes as a novel method to study the epidemiology of allergic drug reactions. *Journal of Allergy and Clinical Immunology: In Practice, 8*(3), 1032-1038.e1. https://doi.org/10.1016/j.jaip.2019.12.007

Barbaro, M., & Zeller Jr., T. (2006). *A face Is exposed for AOL searcher No. 4417749* (p. A1).

Barth-Jones, D. C. (2012). *The "re-identification" of Governor William Weld's medical information: a critical re-examination of health data identification risks and privacy protections, then and now*.

Beckwith, B. A., Mahaadevan, R., Balis, U. J., & Kuo, F. (2006). Development and evaluation of an open-source software tool for deidentification of pathology reports. *BMC Medical Informatics and Decision Making, 6*. https://doi.org/10.1186/1472-6947-6-12

Berman, J. J. (2003). Concept-match medical data scrubbing. How pathology text can be used in research. *Archives of Pathology & Laboratory Medicine, 127*(6), 680–686. https://doi.org/10.5858/2003-127-680-CMDS

Bowen, M. (2012). *EMR Data Quality Evaluation Guide*.

Bui, D. D. A., Wyatt, M., & Cimino, J. J. (2017). The UAB Informatics Institute and 2016 CEGS N-GRID de-identification shared task challenge. *Journal of Biomedical Informatics, 75,* S54–S61. https://doi.org/10.1016/J.JBI.2017.05.001

Canada Health Infoway. (2013). *The emerging benefits of electronic medical record use in community-based care*. www.infoway-inforoute.ca

Canadian Institute for Health Information. (2010). *"Best practice" guidelines for managing the disclosure of de-identified health information*. www.cihi.ca

Canadian Institute for Health Information. (2013). *Privacy policy on the collection, use, disclosure and retention of personal health information and de-Identified data*.

Canadian Institute for Health Information. (2019). *Commonwealth Fund Survey*. https://www.cihi.ca/en/commonwealth-fund-survey-2019?utm_medium=social-organic&utm_source=twitter&utm_campaign=CMWF-2019&utm_content=product-en-public-page

Canadian Institutes of Health Research. (2005). *CIHR best practices for protecting privacy in health research.* Canadian Institutes of Health Research. 29/05/2022https://cihr-irsc.gc.ca/e/29072.html

Carrell, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B., & Hirschman, L. (2013). Hiding in plain sight: use of realistic surrogates to reduce exposure of protected health information in clinical text. *Journal of the American Medical Informatics Association*, *20*(2), 342–348. https://doi.org/10.1136/AMIAJNL-2012-001034

Cavoukian, A. (2011). Dispelling the myths surrounding de-identification anonymization remains a strong tool for protecting privacy. In *Anonymization remains a strong tool for protecting privacy*. Toronto, Ontario: Information and Privacy Commissioner of Ontario, Canada.

Cavoukian, A., & Khaled E. (2014). *De-identification protocols: essential for protecting privacy*. Toronto, Ontario: Information and Privacy Commissioner of Ontario, Canada.

Chapter B: Key concepts. (2019). https://www.oaic.gov.au/privacy/australian-privacy-principles-guidelines/chapter-b-key-concepts#personal-information

Chavan A. (2019). *Introduction to conditional random fields (CRFs)*. AI Time Journal. https://www.aitimejournal.com/@akshay.chavan/introduction-to-conditional-random-fields-crfs

Chen, A., Jonnagaddala, J., Nekkantti, C., & Liaw, S. T. (2019). Generation of surrogates for de-identification of electronic health records. *Studies in Health Technology and Informatics, 264*, 70–73. https://doi.org/10.3233/SHTI190185

Chen, T., Cullen, R. M., & Godwin, M. (2015). Hidden Markov model using dirichlet process for de-identification. *Journal of Biomedical Informatics, 58*, S60–S66. https://doi.org/10.1016/J.JBI.2015.09.004

Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., & Lovis, C. (2019). Use and understanding of anonymization and de-identification in the biomedical literature: Scoping review. *Journal of Medical Internet Research*, *21*(5). https://doi.org/10.2196/13484

Christine Taylor. (2021, May). *Structured vs unstructured data*. Datamation. https://www.datamation.com/big-data/structured-vs-unstructured-data/

Chylek, L. A., Hu, B., Blinov, M. L., Emonet, T., Faeder, J. R., Goldstein, B., Gutenkunst, R. N., Haugh, J. M., Lipniacki, T., Posner, R. G., Yang, J., & Hlavacek, W. S. (2011). Guidelines for visualizing and annotating rule-based models. *Molecular BioSystems*, *7*(10), 2779–2795. https://doi.org/10.1039/c1mb05077j

Committee on Strategies for Responsible Sharing of Clinical Trial Data. (2015). Sharing clinical trial data: maximizing benefits, minimizing risk. In *National Center for Biotechnology Information*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK285994/

Dahl, L. T., Katz, A., McGrail, K., Diverty, B., Ethier, J. F., Gavin, F., McDonald, J. T., Alison Paprica, P., Schull, M., Walker, J. D., & Wu, J. (2020). The SPOR-Canadian Data Platform: a national initiative to facilitate data rich multi-jurisdictional research. *International Journal of Population Data Science*, *5*(1). https://doi.org/10.23889/IJPDS.V5I1.1374

Data quality and machine learning: what's the connection? (n.d.). Talend. Retrieved May 16, 2022, from https://www.talend.com/resources/machine-learning-data-quality/

DBMI Portal. (n.d.). Retrieved June 2, 2022, from https://portal.dbmi.hms.harvard.edu/

Dehghan, A., Kovacevic, A., Karystianis, G., Keane, J. A., & Nenadic, G. (2015). Combining knowledge-and data-driven methods for de-identification of clinical narratives. *Journal of Biomedical Informatics*, *58*, S53–S59. https://doi.org/10.1016/j.jbi.2015.06.029

De-Identification Tools. (n.d.). National Library of Medicine. Retrieved July 11, 2022, from https://lhncbc.nlm.nih.gov/LHC-research/LHC-projects/NLP/de-identification.html

Deleger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q., Marsolo, K., Jegga, A., Kaiser, M., Stoutenborough, L., & Solti, I. (2013). Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *Journal of the American Medical Informatics Association*, *20*(1), 84–94. https://doi.org/10.1136/AMIAJNL-2012-001012

Department of Health and Wellness. (2013). *Toolkit for custodians: A guide to the Personal Health Information Act*. www.novascotia.ca/DHW/PHIA

Dorr, D. A., Phillips, W. F., Phansalkar, S., Sims, S. A., & Hurdle J F. (2006). Assessing the difficulty and time cost of de-identification in clinical narratives. *Methods of Information in Medicine*, *45*(3), 246–252.

Douglass, M., Cliffford, G., Reisner, A., Long, W., Moody, G., & Mark, R. (2005). *De-identification algorithm for free-text nursing notes* (pp. 331–334). https://doi.org/10.1109/CIC.2005.1588104

Douglass, M., Clifford, G. D., Reisner, A., Moody, G. B., & Mark Rg, G. B. (2004). *Computer-assisted de-identification of free text in the MIMIC II database* (pp. 341–344). https://doi.org/10.1109/CIC.2004.1442942

eChart Manitoba. (n.d.). *Clinical system definitions*. Retrieved June 9, 2022, from https://echartmanitoba.ca/hcp/training/frequently-asked-questions/clinical-system-definitions/

el Emam, K. (2011a). *Pan-Canadian de-identification guidelines for personal health information*. Ottawa, Ontario: CHEO Research Institute.

el Emam, K. (2011b). *The case for de-identifying personal health information*. Ottawa, Ontario: CHEO Research Institute.

el Emam, K. (2013). *Guide to the de-identification of personal health information*. Boca Raton: CRC Press/Taylor & Francis Group.

el Emam, K., Buckeridge, D., Tamblyn, R., Neisa, A., Jonker, E., & Verma, A. (2011). The re-identification risk of Canadians from longitudinal demographics. *BMC Medical Informatics and Decision Making*, *11*, 46.

el Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). A systematic review of re-identification attacks on health data. *PLoS ONE*, *6*(12), e28071. https://doi.org/10.1371/journal.pone.0028071

Emam, K. el, Dankar, F. K., Vaillancourt, R., Roffey, T., & Lysyk, M. (2009). Evaluating the risk of re-identification of patients from hospital prescription records. *The Canadian Journal of Hospital Pharmacy*, *62*(4), 307.

Federal Register (Ed.). (2002). Standards for privacy of individually identifiable health information. Final rule. In *67* (Issue 53181, pp. 53181–53273).

Ferrández, O., South, B., Shen, S., Friedlin, F., Samore, M., & Meystre, S. (2012). Evaluating current automatic de- identification methods with Veteran's health administration clinical documents. *BMC Medical Research Methodology*, *12*, 109. https://doi.org/10.1186/1471-2288-12-109

Fraser, R., & Willison, D. (2009). *Tools for de-identification of personal health information*.

Friedlin, F., & McDonald, C. (2008). A software tool for removing patient identifying information from clinical documents. *Journal of the American Medical Informatics Association*, *15*, 601–610. https://doi.org/10.1197/jamia.M2702

Garfinkel, S. L. (2015). *De-identification of personal information*. https://doi.org/10.6028/NIST.IR.8053

Government of Manitoba. (1997). *The Personal Health Information Act - Consolidated Act P33.5*. https://web2.gov.mb.ca/laws/statutes/ccsm/p033-5e.php

Harris, M., Levy, A., & Teschke, K. (2008). Personal privacy and public health: potential impacts of privacy legislation on health research in Canada. *Canadian Journal of Public Health*, *99*(4), 293.

Hartman, T., Howell, M. D., Dean, J., Hoory, S., Slyper, R., Laish, I., Gilon, O., Vainstein, D., Corrado, G., Chou, K., Po, M. J., Williams, J., Ellis, S., Bee, G., Hassidim, A., Amira, R., Beryozkin, G., Szpektor, I., & Matias, Y. (2020). Customization scenarios for de-identification of clinical notes. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-020-1026-2

He, B., Guan, Y., Cheng, J., Cen, K., & Hua, W. (2015). CRFs based de-identification of medical records. *Journal of Biomedical Informatics*, *58*, S39–S46. https://doi.org/10.1016/J.JBI.2015.08.012

Health Insurance Portability and Accountability Act of 1996 (HIPAA) | CDC. (n.d.). Retrieved March 9, 2021, from https://www.cdc.gov/phlp/publications/topic/hipaa.html

Heider, P. M., Accetta, J., & Meystre, S. (2018). ETUDE for easy and efficient NLP application evaluation. *AMIA NLP-WG Pre-Symposium*.

Henriksson, A., Kvist, M., & Dalianis, H. (2017). Prevalence estimation of protected health information in Swedish clinical text. *Studies in Health Technology and Informatics*, *235*, 216–220. https://doi.org/10.3233/978-1-61499-753-5-216

Heselmans, A., Delvaux, N., Laenen, A., van de Velde, S., Ramaekers, D., Kunnamo, I., & Aertgeerts, B. (2020). Computerized clinical decision support system for diabetes in primary care does not improve quality of care: A cluster-randomized controlled trial. *Implementation Science*, *15*(1). https://doi.org/10.1186/s13012-019-0955-6

Informatics for Integrating Biology & the Bedside (i2b2). (n.d.). *Natural language processing research data sets*. Retrieved January 3, 2021, from https://www.i2b2.org/NLP/DataSets/Main.php

Information and Privacy Commissioner of Ontario. (2014). *Big data and innovation, setting the record straight: de-identification does work*.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*, 685–695. https://doi.org/10.1007/s12525-021-00475-2/Published

Javeed, A., Khan, S. U., Ali, L., Ali, S., Imrana, Y., & Rahman, A. (2022). Machine learning-based automated diagnostic systems developed for heart failure prediction using different types of data modalities: a systematic review and future directions. In *Computational and Mathematical Methods in Medicine* (Vol. 2022). Hindawi Limited. https://doi.org/10.1155/2022/9288452

Juhn, Y., & Liu, H. (2020). Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *Journal of Allergy and Clinical Immunology*, *145*(2), 463–469. https://doi.org/10.1016/j.jaci.2019.12.897

Katz, A., Enns, J. E., Wong, S. T., Williamson, T., Singer, A., McGrail, K., Bakal, J. A., Taylor, C., & Peterson, S. (2018). Challenges associated with cross-jurisdictional analyses using administrative health data and primary care electronic medical records in Canada. *International Journal of Population Data Science*, *3*(3). https://doi.org/10.23889/IJPDS.V3I3.437

Kayaalp, M., Browne, A. C., Callaghan, F. M., Dodd, Z. A., Divita, G., Ozturk, S., & McDonald, C. J. (2014). The pattern of name tokens in narrative clinical text and a comparison of five systems for redacting them. *Journal of the American Medical Informatics Association*, *21*(3), 423–431. https://doi.org/10.1136/AMIAJNL-2013-001689

Kayaalp, M., Browne, A. C., Dodd, Z. A., Sagan, P., & McDonald, C. J. (2014). De-identification of address, date, and alphanumeric identifiers in narrative clinical reports. *AMIA Annual Symposium Proceedings*, *2014*, 767.

Kho, M. E., Duffett, M., Willison, D. J., Cook, D. J., & Brouwers, M. C. (2009). Written informed consent and selection bias in observational studies using medical records: systematic review. *British Medical Journal*, *338*. https://doi.org/10.1136/bmj.b866

Kristianson, K. J., Ljunggren, H., & Gustafsson, L. L. (2007). 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 Journal*, *15*(4), 305–319. https://doi.org/10.1177/1460458209345889

Kumar, V., Stubbs, A., Shaw, S., & Uzuner, Ö. (2015). Creation of a new longitudinal corpus of clinical narratives. *Journal of Biomedical Informatics, 58,* S6–S10. https://doi.org/10.1016/j.jbi.2015.09.018

Kushida, C., Nichols, D., Jadrnicek, R., Miller, R., Walsh, J., & Griffin, K. (2012). Strategies for de-identification and anonymization of electronic health record data for use in multicenter research studies. *Medical Care, 50*, S82.

Lafky, D. (2010). The Safe Harbor method of de-identification: an empirical test. In *Fourth National HIPAA Summit West*. http://www.ehcca.com/presentations/HIPAAWest4/lafky_2.pdf

Lee, H. J., Wu, Y., Zhang, Y., Xu, J., Xu, H., & Roberts, K. (2017). A hybrid approach to automatic de-identification of psychiatric notes. *Journal of Biomedical Informatics*, *75*, S19–S27. https://doi.org/10.1016/J.JBI.2017.06.006

Lexical Systems Group. (2010). *Visual Tagging Tool 2010*. https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/vtt/current/docs/userDoc/vttFileFormat/vttFileFormat.2010.0.html

Liu, Z., Chen, Y., Tang, B., Wang, X., Chen, Q., Li, H., Wang, J., Deng, Q., & Zhu, S. (2015). Automatic de-identification of electronic medical records using token-level and character-level conditional random fields. *Journal of Biomedical Informatics*, *58*, S47–S52. https://doi.org/10.1016/J.JBI.2015.06.009

Liu, Z., Tang, B., Wang, X., & Chen, Q. (2017). De-identification of clinical notes via recurrent neural network and conditional random field. *Journal of Biomedical Informatics*, *75*, S34–S42. https://doi.org/10.1016/j.jbi.2017.05.023

Lo, J. (2012). *Consumers anonymous? The privacy risks of de-identified and aggregated consumer data*. Ottawa, Ontario: Public Interest Advocacy Centre.

Lu, C. J., & Browne, A. C. (2010a). *Visual Tagging Tool*.

Lu, C. J., & Browne, A. C. (2010b, December). Vtt 2010. *Library Associates Talk*. https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/vtt/current/docs/userDoc/presentations/index.html

Manitoba Centre for Health Policy. (n.d.-a). *Term: Personal Health Identification Number (PHIN)*. Retrieved April 2, 2018, from http://mchp-appserv.cpe.umanitoba.ca/viewDefinition.php?definitionID=103292

Manitoba Centre for Health Policy, U. of M. (n.d.-b). *The Manitoba Population Research Data Repository*. Retrieved July 12, 2022, from https://umanitoba.ca/manitoba-centre-for-health-policy/data-repository

Manitoba Health. (n.d.). *Frequently asked questions | PHIA*. Retrieved May 2, 2022, from https://www.gov.mb.ca/health/phia/faq.html#j

Manitoba Health Seniors and Active Living. (2022). *Personal Health Information Act*. The Government of Manitoba. http://www.gov.mb.ca/health/phia/

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3), 276. https://doi.org/10.11613/bm.2012.031

Meystre, S. M., Friedlin, F. J., South, B. R., Shen, S., & Samore, M. H. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology*, *10*, 70. https://doi.org/10.1186/1471-2288-10-70

Meystre, S. M., Heider, P. M., Kim, Y., & Davis, M. (2021). Natural language processing enabling COVID-19 predictive analytics to support data-driven patient advising and pooled testing clinical NLP and AI applications. *Journal of the American Medical Informatics Association*. https://doi.org/10.1093/jamia/ocab186/6355588

Mezzatesta, S., Torino, C., de Meo, P., Fiumara, G., & Vilasi, A. (2019). A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis. *Computer Methods and Programs in Biomedicine*, *177*, 9–15. https://doi.org/10.1016/J.CMPB.2019.05.005

Milieu Ltd. (2014). *Overview of the national laws on electronic health records in the EU Member States and their interaction with the provision of cross-border eHealth services. Final report and recommendations* (p. 65). http://ec.europa.eu/health//sites/health/files/ehealth/docs/laws_report_recommendations_en.pdf

Murugadoss, K., Rajasekharan, A., Malin, B., Agarwal, V., Bade, S., Anderson, J. R., Ross, J. L., Faubion, W. A., Halamka, J. D., Soundararajan, V., & Ardhanari, S. (2021). Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns*, *2*(6). https://doi.org/10.1016/J.PATTER.2021.100255

N2c2 NLP Research Data Sets. (n.d.). Retrieved June 13, 2021, from https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. *IEEE Symposium on Security and Privacy*, 111–125. https://doi.org/10.1109/SP.2008.33

Neamatullah, I., Douglass, M. M., Lehman, L. W. H., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G., & Clifford, G. D. (2008). Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making*, *8*. https://doi.org/10.1186/1472-6947-8-32

Negro-Calduch, E., Azzopardi-Muscat, N., Krishnamurthy, R. S., & Novillo-Ortiz, D. (2021). Technological progress in electronic health record system optimization: Systematic review of systematic literature reviews. *International Journal of Medical Informatics*, *152*. https://doi.org/10.1016/J.IJMEDINF.2021.104507

Nesca, M. (2021). *Measuring the quality of unstructured text in routinely collected electronic health data: a review and application (Master's thesis)*. https://mspace.lib.umanitoba.ca/handle/1993/36163

Ness, B. R. (2008). Influence of the HIPAA Privacy Rule on health research. *Obstetrical &amp; Gynecological Survey*, *63*(4), 236–237. https://doi.org/10.1097/01.ogx.0000310359.63794.a8

Norgeot, B., Muenzen, K., Peterson, T. A., Fan, X., Glicksberg, B. S., Schenk, G., Rutenberg, E., Oskotsky, B., Sirota, M., Yazdany, J., Schmajuk, G., Ludwig, D., Goldstein, T., & Butte, A. J. (2020). Protected health information filter (Philter): accurately and securely de-identifying free-text clinical notes. *Nature Portfolio Journal*, *3*(1). https://doi.org/10.1038/s41746-020-0258-y

Office of the Australian Information Commissioner. (n.d.). *Australian Privacy Principles*. Retrieved June 9, 2022, from https://www.oaic.gov.au/privacy/australian-privacy-principles

Ontario Cancer Care. (2014). *Principles and policies for the protection of personal health information at Cancer Care Ontario*. https://www.cancercare.on.ca/common/pages/UserFile.aspx?fileId=13632

Panicacci, S., Donati, M., Fanucci, L., Bellini, I., Profili, F., & Francesconi, P. (2019). Exploring machine learning algorithms to identify heart failure patients: The Tuscany region case study. *Proceedings - IEEE Symposium on Computer-Based Medical Systems*, *2019-June*, 417–422. https://doi.org/10.1109/CBMS.2019.00088

Personal Health Information Protection Act, SO 2004 (2004). https://www.ontario.ca/laws/statute/04p03

Protti, D. (2007). Comparison of information technology in general practice in 10 countries. *Electronic Healthcare*, *10*(2), 107–116.

Raja, N., Sagan, P., Jones, J., Way, M., & Kayaalp, M. (2019). *NLM-Scrubber user manual*.

Shared Health. (n.d.). *Provincial EMR*. Retrieved June 9, 2022, from https://sharedhealthmb.ca/health-providers/digital-health/pcis-office/provincial-emr/

Sharma, B., Dligach, D., Swope, K., Salisbury-Afshar, E., Karnik, N. S., Joyce, C., & Afshar, M. (2020). Publicly available machine learning models for identifying opioid misuse from the clinical notes of hospitalized patients. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/S12911-020-1099-Y

Sheikhalishahi, S., Miotto, R., Dudley, J. T., Lavelli, A., Rinaldi, F., & Osmani, V. (2019). Natural language processing of clinical notes on chronic diseases: systematic review. *Journal of Medical Internet Research Medical Informatics*, *7*(2). https://doi.org/10.2196/12239

Shin, S., Park, Y., Shin, Y., Choi, H., Park, J., Lyu, Y., Lee, M., Choi, C., Kim, W., & Lee, J. (2015). A de-identification method for bilingual clinical texts of various note types. *J Korean Med Sci*, *30*, 7–15. https://doi.org/10.3346/jkms.2015.30.1.7

Shin, S.-Y. (2018). Issues and solutions of healthcare data de-identification: the case of South Korea. *Journal of Korean Medical Science*, *33*(5), 41. https://doi.org/10.3346/jkms.2018.33.e41

Singer, A., Kroeker, A. L., Yakubovich, S., Duarte, R., Dufault, B., & Katz, A. (2017). Data quality in electronic medical records in Manitoba: do problem lists reflect chronic disease as defined by prescriptions? *Canadian Family Physician*, *63*(5), 382.

Steinkamp, J. M., Pomeranz, T., Adleberg, J., Kahn, C. E., & Cook, T. S. (2020). Evaluation of automated public de-identification tools on a corpus of radiology reports. *Radiology: Artificial Intelligence, 2*(6), e190137. https://doi.org/10.1148/ryai.2020190137

Stubbs, A., Filannino, M., & Uzuner, Ö. (2017). De-identification of psychiatric intake records: overview of 2016 CEGS N-GRID shared tasks Track 1. *Journal of Biomedical Informatics*, *75*, S4–S18. https://doi.org/10.1016/J.JBI.2017.06.011

Stubbs, A., Kotfila, C., & Uzuner, Ö. (2015). Automated systems for the de-identification of longitudinal clinical narratives: overview of 2014 i2b2/UTHealth shared task Track 1. *Journal of Biomedical Informatics*, *58*, S11–S19. https://doi.org/10.1016/J.JBI.2015.06.007

Stubbs, A., & Uzuner, O. (2014). *Annotation guidelines for de-identification of medical records*.

Stubbs, A., & Uzuner, Ö. (2015). Annotating longitudinal clinical narratives for de-identification: the 2014 i2b2/UTHealth corpus. *Journal of Biomedical Informatics*, *58*, S20–S29. https://doi.org/10.1016/J.JBI.2015.07.020

Sung, S. F., Hung, L. C., & Hu, Y. H. (2021). Developing a stroke alert trigger for clinical decision support at emergency triage using machine learning. *International Journal of Medical Informatics*, *152*. https://doi.org/10.1016/j.ijmedinf.2021.104505

Sweeney, L. (1996). Replacing personally-identifying information in medical records, the Scrub system. *AMIA Annual Fall Symposium*, 333–337. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2233179/

Sweeney, L. (2002). K-anonymity: a model for protecting privacy. *International Journal Of Uncertainty Fuzziness And Knowledge-Based Systems*, *10*(5), 557–570.

Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., & Godtliebsen, | Fred. (2021). *Challenges and opportunities beyond structured data in analysis of electronic health records*. https://doi.org/10.1002/wics.1549

Tolar, M., & Balka, E. (2012). Caring for individual patients and beyond: enhancing care through secondary use of data in a general practice setting. *International Journal of Medical Informatics*, *81*(7), 461–474. https://doi.org/10.1016/j.ijmedinf.2012.01.003

Trienes, J., Trieschnigg, D., Seifert, C., & Hiemstra, D. (2020). *Comparing rule-based, feature-based and deep neural methods for de-identification of Dutch medical records*. http://arxiv.org/abs/2001.05714

U.S. Department of Health & Human Services. (n.d.-a). *Methods for De-identification of PHI*. Retrieved February 4, 2022, from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

U.S. Department of Health & Human Services. (n.d.-b). *Summary of the HIPAA Privacy Rule*. Retrieved February 4, 2022, from http://www.hhs.gov/ocr/hipaa.

U.S. National Library of Medicine. (n.d.). *Clinical text de-identification using NLM-Scrubber*. Retrieved February 1, 2018, from https://scrubber.nlm.nih.gov

Uzuner, Ö., Luo, Y., & Szolovits, P. (2007). Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association*, *14*(5), 550–563. https://doi.org/10.1197/JAMIA.M2444

Uzuner, Ö., & Stubbs, A. (2015). Practical applications for natural language processing in clinical research: the 2014 i2b2/UTHealth shared tasks. *Journal of Biomedical Informatics*, *58*(Suppl), S1. https://doi.org/10.1016/J.JBI.2015.10.007

What is GDPR, the EU's new data protection law? (n.d.). Retrieved February 4, 2022, from https://gdpr.eu/what-is-gdpr/

What is personal data? | European Commission. (n.d.). Retrieved June 3, 2022, from https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en

Winnipeg Sun. (2020, January 29). *Manitoba makes big strides in digital health care: CIHI report*. https://winnipegsun.com/news/news-news/manitoba-makes-big-strides-in-digital-health-care-cihi-report

Xiao, C., Choi, E., & Sun, J. (2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. In *Journal of the*

American Medical Informatics Association (Vol. 25, Issue 10, pp. 1419–1428). Oxford University Press. https://doi.org/10.1093/jamia/ocy068

Yang, H., & Garibaldi, J. M. (2015). Automatic detection of protected health information from clinic narratives. *Journal of Biomedical Informatics*, *58*, S30–S38. https://doi.org/10.1016/J.JBI.2015.06.015

Appendices

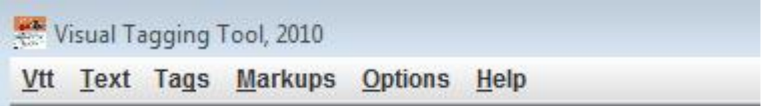
Appendix A. Visual Tagging Tool (VTT) Instructions

General notes

- Use the *Tagging guidelines* document for directions on what data needs to be tagged.
- We will each tag 60 encounter notes, compare our tagging, and update the tagging guidelines as necessary.
- Once the above is completed, we will each tag ~ 600 encounter notes (including the first 60).
- Please do not edit or change the tags.
- Discrepancies between tagging will be discussed as needed.

Accessing VTT and the dataset

1. Go to **bin** folder located here C:\Users\KatelinM\Documents\vtt2010\bin
   a. Tip: a shortcut to the bin folder has been created in the documents folder.
2. Open the **vtt2** file.
3. The software should open in a new window. You will see a white screen and the menu bar along the top looks like this:

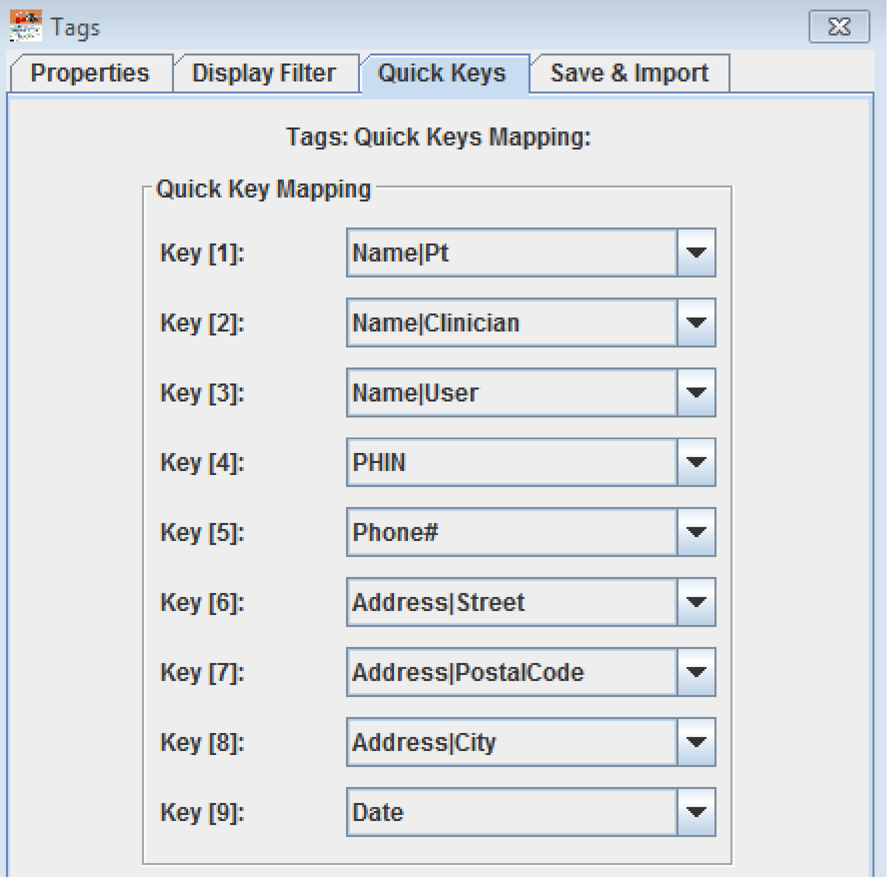


4. Next you need to open the file that contains our local dataset. From the VTT menu bar, select **Open (o)**.
5. The Open a file box will display. Near the top of the display box, you will see a **down arrow** (located to the right of the Look In section). Use this arrow to open the navigator and select the **Shared** folder.
6. Select **Files**, **Local**, and then your tagging file that ends with *txt.vtt*.
7. You should now see data display in the VTT pop-up box. This is the dataset that needs to be tagged.

Tips

- The cdm box will be open in the background. It must stay open while using VTT. Minimizing the screen is fine.
- **Options** in the VTT menu bar allows you to change the font size and other settings.

- Here are details of how to **move** through the dataset easily using your keyboard.
- Here is a list of operations that can be used to **tag** the dataset. It also includes quick keys which help you use keyboard shortcuts instead of your mouse.
  - This is a cheat sheet of the same information, in a concise, tabular format.
- Below is the list of quick keys for tagging:



Tags: Quick Keys Mapping:

Quick Key Mapping

Key [1]:	Name|Pt
Key [2]:	Name|Clinician
Key [3]:	Name|User
Key [4]:	PHIN
Key [5]:	Phone#
Key [6]:	Address|Street
Key [7]:	Address|PostalCode
Key [8]:	Address|City
Key [9]:	Date

Appendix B. Tagging Guidelines[1]

Definitions

Term	Description
Personal Health Information (PHI)	Data that identifies a person. Can be direct (e.g., name) or indirect when combined with other indirect data (e.g., postal code).
Personal Health Identification Number (PHIN)	Unique nine-digit numeric identifier assigned by Manitoba Health to every person registered for health insurance in Manitoba.[2]
String	A sequence of characters (e.g., letters, numbers, punctuation). For example, the name John is a string of four characters.
Tag	The process of associating a type to a token. In VTT, this is done by marking a string as a PHI type. For example, the token "John" would be associated to type: Name|Pt.
Token	A meaningful unit of information in a sequence of characters. For example, in the phrase "John Smith ate an apple", "John" would count as one token, Smith would count as one token and so on.

General guidelines

- Do not include the "|" character in a tag. Any characters after it will be ignored.
- Unless otherwise noted, do not include a space in a tag.
- When something looks out of place (e.g., "MELINDA of cough") consider it to be PHI

Names

- Tag patient names (first, middle, last) with **Name|Pt**
- Tag clinician names with **Name|Clinician**
    - All clinician types are in scope (e.g., name of Pharmacist must be tagged)
- Tag usernames (e.g., jsmith, johns) with **Name|User**
    - Usernames typically indicated by an initial and name put together in a single string

[1] Informed by (Stubbs & Uzuner, 2014, 2015)

[2] Definition adapted from the Manitoba Centre for Health Policy (MCHP; Manitoba Centre for Health Policy, n.d.)

- Titles (e.g., Dr., Mr., Mrs.) do not need to be tagged
- Initials do not need to be tagged
- Designations (e.g., MD, RN) do not need to be tagged
- If name is possessive (e.g., John's) do not tag the 's
- Names separated by spaces should be tagged as two separate tokens
- Names joined by a hyphen should be tagged as one token

Personal Health Identifiers

- Tag Personal Health Identification Numbers (PHIN) with **PHIN**
    - For this dataset, this includes strings of nine numbers (e.g., 80002222, 123456789)
- All other identifiers do not need to be tagged. This includes other jurisdictional health numbers, private insurance numbers, and unique record IDs assigned by an EMR.

Phone numbers

- Tag phone numbers with Phone#
    - This includes phone numbers with an area code and without an area code
- Phone numbers separated by spaces, periods or hyphens should be tagged as one token
- Fax numbers do not need to be tagged

Address

- Tag the house number and street with **Address|Street**
    - The entire number and street should be tagged as one token (e.g., 123 First Street is one token)
- Tag the city or town with **Address|City**
    - City or town names that consist of two separate strings should be tagged as one token (e.g., Rapid City)
    - If clinic or hospital name contains city or town name, tag the city/town portion of the site name only (e.g., for Brandon Regional Health Centre – only tag "Brandon")
- Tag the postal code with **Address|PostalCode**

- P.O. Boxes do not need to be tagged
- Zip codes do not need to be tagged

Dates

- Tag all dates with **Date**
  - Include all dates regardless of context (e.g., Date of Birth, encounter note entered, signed off date, date mail sent)
  - Include all dates regardless of format (e.g., 01/01/2018; January 1st, 2018; Jan 1, 2018)
- Where month and day are recorded together, tag as one token
- Years do not need to be tagged
- Where year is recorded with month and day, tag only the month and day (e.g., for 01/01/2018 – only tag 01/01)
- Mentions of age do not need to be tagged (e.g., X years old, X years of age)
- Days of the week (e.g., Wednesday) do not need to be tagged
- Seasons (e.g., Spring) do not need to be tagged