EVALUATING THE EFFECTS OF NONINTERACTIVE AND MACHINE-ASSISTED INTERACTIVE MANUAL CLINICAL TEXT ANNOTATION APPROACHES ON THE QUALITY OF REFERENCE STANDARDS

by

Brett Ray South

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

The University of Utah

May 2014

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Brett Ray South**

has been approved by the following supervisory committee members:

Stephane Meystre	, Chair	**03/03/2014** Date Approved
Wendy W. Chapman	, Member	**03/03/2014** Date Approved
Charlene Weir	, Member	**03/03/2014** Date Approved
Bruce Bray	, Member	**03/03/2014** Date Approved
Matthew Samore	, Member	**03/03/2014** Date Approved

and by **Wendy W. Chapman**, Chair/Dean of the Department/College/School of **Biomedical Informatics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

Manual annotation of clinical texts is often used as a method of generating reference standards that provide data for training and evaluation of Natural Language Processing (NLP) systems. Manually annotating clinical texts is time consuming, expensive, and requires considerable cognitive effort on the part of human reviewers. Furthermore, reference standards must be generated in ways that produce consistent and reliable data but must also be valid in order to adequately evaluate the performance of those systems. The amount of labeled data necessary varies depending on the level of analysis, the complexity of the clinical use case, and the methods that will be used to develop automated machine systems for information extraction and classification. Evaluating methods that potentially reduce cost, manual human workload, introduce task efficiencies, and reduce the amount of labeled data necessary to train NLP tools for specific clinical use cases are active areas of research inquiry in the clinical NLP domain.

This dissertation integrates a mixed methods approach using methodologies from cognitive science and artificial intelligence with manual annotation of clinical texts. Aim 1 of this dissertation identifies factors that affect manual annotation of clinical texts. These factors are further explored by evaluating approaches that may introduce efficiencies into manual review tasks applied to two different NLP development areas – semantic annotation of clinical concepts and identification of information representing Protected Health Information (PHI) as defined by HIPAA. Both experiments integrate

different priming mechanisms using noninteractive and machine-assisted methods. The main hypothesis for this research is that integrating pre-annotation or other machine-assisted methods within manual annotation workflows will improve efficiency of manual annotation tasks without diminishing the quality of generated reference standards.

For Sarah, Charlotte and Kate.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Committee members Drs. Wendy Chapman, Stephane Meystre and Matthew Samore have been a constant source of support and encouragement. It was also an honor to have Drs. Charlene Weir and Bruce Bray as committee members. Without the positive attitudes, guidance and patience of all committee members it would have been difficult to continue and complete this work. The insight, professionalism and feedback have significantly benefited my research. I owe them all a debt of gratitude. I am thankful for my friends and colleagues Scott DuVall, Shuying Shen, Tyler Forbush, Danielle Mowery, Chris Leng, Ying Suo and Matthew Maw. Especially Chris Leng for his devoted effort invested in developing eHOST and Ying Suo and Matthew Maw for their patience with ever-changing data analysis strategies. I would also like to thank my family and extended family for their encouragement. Finally, I would like to thank my beautiful patient wife, Sarah, and my daughters, Charlotte and Kate who encouraged me to finish this dissertation even during times of great frustration.

Development of the Extensible Human Oracle Suite of Tools (eHOST) was supported by the VA Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374 and VA Informatics and Computing Infrastructure (VINCI), VA HIR 08-204 grants. Annotation of the 2010 i2b2/VA corpus was supported by the NIH Roadmap for Medical Research, Grant U54LM008748 and the VA CHIR VA HSR HIR 08-374 Grants. Annotation of the MTSamples corpus was supported by NIH Grant U54-

HL108460 for Integrating Data for Analysis, Anonymization and Sharing (iDASH), NIGMS-7R01GM090187.

PREFACE

"This is your last chance. After this, there is no turning back. You take the blue pill – the story ends, you wake up in your bed and believe whatever you want to believe. You take the red pill – you stay in Wonderland, and I show you how deep the rabbit hole goes. Remember, all I'm offering is the truth – nothing more."

Morpheus to Neo. The Matrix, 1999.

Indeed this quote relates to the topic of this dissertation – integrating efficiencies with reference standard generation for clinical applications of Natural Language Processing (NLP). When I began this dissertation work I excitedly took the "red pill" and as a result learned more about reference standard generation, human cognition, statistics, and the painfully flawed reality of subjective human review than I ever imagined. These are lasting lessons that I have integrated with my professional experience. The Wachowski Brothers imagined dystopian science fiction world of the Matrix (1999) where machines rule over man is just that – science fiction. However, despite our best attempts to train machines to do the more mundane tasks of human review there will always be a place for an "oracle" or many "oracles" that silently build truth labels however flawed they may be. This is very true even of the recent advances in question answering and in 2011, IBM's supercomputer Watson's victory over Ken Jennings, the all time top Jeopardy champion.

CHAPTER 1

INTRODUCTION

1.1 Main Objectives

Thanks to science fiction and the 2011 victory of IBM's Watson over all time Jeopardy champion Ken Jennings, it is easy to imagine a world where machines have the capability to capture all the richness and nuance of human written or spoken language. The reality, however, is different from imagination since the current state of the art in Natural Language Processing (NLP) can produce structured representations of only part of the meaning of human language [1]. This conversion often relies on supervised learning approaches that use the output of some "oracle" or many "oracles" to generate truth labels that are used as a "ground truth" or "reference standard." Manual text annotation is often one approach to labeling data that is used for training and evaluation of NLP systems for information extraction or classification tasks. Annotation is a schema based manual review process that is used to identify spans of text that represent instances of particular target information classes. Guidelines and rule sets driven by specific NLP development goals are written defining what to annotate and what not to annotate. Manual annotation is therefore a human information extraction task. These efforts may also include classification steps such as assignment of various attributes or identifying relations between information classes [2].

Manually annotating clinical texts is time consuming, expensive, and requires considerable effort on the part of human reviewers to identify and classify information of interest [3]. Clinical texts are unique in that they contain specialized sublanguages, they are context dependent, and are often written by experts to be read and interpreted by other experts. Often what is not mentioned in a clinical document is as important as what is clearly stated. Furthermore, reference standards must be generated in a way that is reliable and must also be valid in order to adequately evaluate the performance of those machine systems for information extraction or classification. The amount of labeled data necessary for system training and evaluation varies depending on the level of analysis, the complexity of the clinical use case, and methods used. Evaluating manual review methods that potentially reduce cost, workload, introduce task efficiencies, and reduce the amount of labeled data necessary to train NLP tools for specific clinical use cases, are active areas of research inquiry in the clinical NLP community.

This dissertation addresses some of the issues that affect efficiencies of manual annotation of clinical texts by human reviewers. The main hypothesis for these research efforts is that integrating pre-annotation or other machine-assisted methods within manual annotation workflows will improve efficiency of manual annotation tasks without diminishing the quality of generated reference standards. Each aim in this dissertation builds upon prior aims and starts in Aim 1 that first identifies factors that affect manual annotation of clinical texts. These factors are further explored by evaluating approaches that may introduce efficiencies into manual review tasks applied to two different NLP development areas – semantic annotation of clinical concepts and identification of information representing Protected Health Information (PHI) as defined by HIPAA [4].

Aims 1 and 2 deal with a scenario where a noninteractive pre-annotation approach is used to prime human reviewers to identify clinical mentions. Aim 3 defines an annotation task that deals with information extraction and classification of PHI and non-PHI information using an interactive machine-assisted annotation interface coupled with pre-annotation of target information.

1.2 Aim 1

Identify factors that could affect reliability and validity of manual annotation of clinical texts:

- *Research question 1.1* What factors affect manual annotation of clinical texts?
- *Research question 1.2* How do these factors relate to human review tasks when potential efficiencies are integrated with manual annotation?

As described in Chapter 3, this aim attempts to generate an understanding of cognitive processes involved with manual human annotation and generation of reference standards for NLP systems development.

Limited scientific evidence currently exists about factors that affect annotation reliability or reference standard validity when annotating clinical texts. Moreover, there has been limited research within the clinical NLP community regarding the effect that noninteractive or interactive approaches have on annotation task reliability and reference standard validity. The 2010 i2b2/VA Challenge [2, 5] on medical problems, treatments, and tests, assertion classification and relations extraction from clinical texts provided the use case for this analysis. Qualitative methods adapted from Patton [6] and also used by Campbell [7] were employed to identify key constructs and themes related to annotation tasks and generate hypotheses integrated with the experiments in Aims 2 and 3.

Qualitative analysis of semistructured interviews of annotators from the 2010 i2b2/VA Challenge provided a useful assessment of human adaptation to efficiency attempts, such as modified task workflows or use of machine-assisted or pre-annotation approaches and annotation outcomes [2]. These analyses used the Atlas ti [8] software and involved many hours of iterative review.

1.3 Aim 2

Evaluate a noninteractive pre-annotation approach for manual annotation of semantic concepts, assertions and relations found in clinical texts.

- *Research Question 2.1* What are the effects of noninteractive pre-annotation on reliability, validity, and task efficiency?

As discussed in Chapter 4, the 2010 i2b2/VA Challenge [2, 5] on medical problems, treatments, and tests, assertion classification and relations extraction provided an opportunity to assess the impact of integrating pre-annotation applied to raw full clinical texts and sentence segments conditioning on information scope and context. This experiment also integrated a modified annotation workflow that incorporated additional review levels that go beyond traditional double annotation with adjudication. For a community task like i2b2 these efforts are often done using a large corpus of clinical documents. The reference standard must be generated quickly and in the most efficient way possible given constraints on available resources and time [3, 9].

For this experiment it was assumed that noninteractive pre-annotation would be most beneficial for information that can be identified using combinations of regular expressions and dictionaries and where the prevalence of annotations, relationships, or attributes will be high across a given clinical document. Annotation data generated from

the 2010 i2b2/VA Challenge on medical problems, treatments, and tests, assertion classification and relations extraction from clinical texts provide the use case for this analysis [2, 5]. Pre-annotations were derived from data obtained from the Unified Medical Language System (UMLS) [10] using Apache Lucene [11] to create an index of semantic concept terms [12] mapped to their associated broader information class (i.e., medical problems, treatments, and tests).

From a cognitive perspective, noninteractive pre-annotation functions as a semantic priming mechanism [13-15] helping annotators to more easily "*identify*" certain categories of information more rapidly. The annotator task is changed slightly with the annotator examining existing annotations and modifying them if needed, adding missed annotations, or deleting spurious annotations. It was unknown how much training an annotator must complete to reach an acceptable performance threshold or plateau, and how factors such as domain expertise and the way data are presented impact human annotator performance.

1.4 Aim 3

Evaluate an interactive machine-assisted interface added to noninteractive pre-annotations for manual annotation of PHI found in clinical texts.

- *Research Question 3.1* What are the effects of a combined noninteractive and machine-assisted interactive annotation interface on reliability, validity, and task efficiency?

As described in Chapter 5, given the expectations of HIPAA and the shared ethical and legal responsibility to protect patient confidentiality, consistent and accurate identification and classification of protected health information (PHI) is an important

consideration. High sensitivity is often required to ensure adequate redaction of PHI. This potentially makes de-identification of clinical documents a good candidate to test an approach that combined pre-annotation with a machine-assisted interactive annotation interface.

For this research aim annotation data were generated using a publicly available clinical document corpus available from the MTSamples transcription company [16]. Outputs representing information classes for PHI from a de-identification system called BoB [17] (best-of-breed clinical text de-identification system) developed in the Consortium for Health Care Informatics Research (CHIR) were used as pre-annotations provided to annotators. An annotation tool called the Extensible Human Oracle Suite of Tools (eHOST) [18] provided the machine-assisted interactive annotation interface. From a cognitive perspective, pre-annotations once again functioned as a semantic priming mechanism [13-15] that annotators reacted to by adding missed annotations, correcting or modifying annotated spans, or deleting spurious annotations [19]. It was assumed that additional task efficiencies could be introduced by coupling pre-annotation with a machine-assisted interactive annotation interface that allows reviewers to mark the same spans of text found elsewhere in the documents, and accept or reject the option of annotating the same candidate phrase with the same information class. Furthermore, it was unknown how much training an annotator would have to complete to reach an acceptable performance threshold or plateau, how many annotators would have to review the same documents to achieve acceptable coverage for differing classes of PHI, and how factors such as domain expertise and the way data are presented would impact human annotator performance.

1.5 Hypotheses Under Test

- **Aim 2 Hypothesis:** The one-sided alternative hypothesis (H_a) is that non-interactive pre-annotation of semantic concepts (conditioned on the amount and scope of text provided) improves efficiency of manual clinical text annotation without diminishing reliability and validity metrics.
- **Aim 3 Hypothesis:** The one-sided alternative hypothesis (H_a) is that a combined interactive annotation interface plus pre-annotation approach improves efficiency of manual clinical text annotation without diminishing reliability and validity metrics.

1.6 References

1. Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S., Samore, M.H. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Medical Research Methodology. 10, 70.

2. South, B.R., Shen, S., Barrus, R., DuVall, S.L., Uzuner, Ö., Weir, C. Qualitative analysis of workflow modifications used to generate the reference standard for the 2010 i2b2/VA challenge. In: AMIA Annu Symp Proc. 2011.

3. Settles, B., Craven, M., Friedland. L. Active learning with real annotation costs. In: Proceedings of the NIPS Workshop on Cost-Sensitive Learning. 2008.

4. Health Insurance Portability and Accountability Act (HIPAA). Standards for privacy of individually identifiable health information: final rule. 67 Federal Register 53181 (2002) (codified at 45 CFR 160 and 164).

5. Uzuner, Ö., South, B.R., Shen, S., DuVall, S.L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. In: J Am Med Inform Assoc, 2011. 18(5): 552-556.

6. Patton, M.Q. Qualitative Research and Evaluation Methods. 2002. Sage Publications.

7. E.M. Campbell, D.F. Sittig, W.W. Chapman, B.L. Hazlehurst, A. M. Cohen. Understanding inter-rater disagreement: a mixed methods approach. In: AMIA Annu Symp Proc. 2010: p. 81-5.

8. ATLAS, ti v5.7.1, GmbH, Berlin, Germany. 1993-2011.

9. Uzuner, Ö., Solti, I. Xia, F. Cadag, E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. In: J Am Med Inform Assoc, 2010. 17(5):519-23.

10. Unified Medical Language System. 1996. Available: http://www/nlm.nih.gov/archive/20040831/pubs/cbm/umlscbm.html

11. Apache Lucene. 2011-12. Available: http://lucene.apache.org/core/

12. Friedman, C. Liu, H., Shagina, L. Johnson, S., Hripcsak, G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. In: AMIA Annu Symp Proc. 2001: 189-93.

13. Ferrand, L. New, B. Semantic and associative priming in the mental lexicon. In Bonin (Ed), The Mental Lexicon. New York: Nova Science Publishers. 2003. pp 25-43.

14. Underwood, G. Implicit Cognition. 1996, Oxford: Oxford Science Publications.

15. Wyer, R.S. Social Comprehension and Judgment: The Role of Situation Models, Narratives, and Implicit Theories. 2004, Mahwah, NJ: Erlbaum.

16. Medical Transcription Samples Document Corpus (MTSamples). Available from: http://www.mtsamples.com.

17. Ferrández, O., South, B.R., Shen, S., Friedlin, F.J., Samore, M.H., Meystre, S.M. BoB, a best-of breed automated text de-identification system for VHA clinical documents. J. Am. Med. Inform. Assoc. 2013. 20(1), 77–83.

18. South, B. Shen, S., Leng, J. Forbush, T., DuVall, S., Chapman, W.W. A prototype tool set to support machine-assisted annotation. Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. BioNLP '12, Stroudsburg, PA, USA, Association for Computational Linguistics. 2012. 130-139.

19. South, B.R., Shen, S., Friedlin, F.J., Samore, M., Meystre. S.M. Enhancing annotation of clinical text using pre-annotation of common PHI. In: AMIA Annu Symp Proc. 2010:126.

CHAPTER 2

BACKGROUND

2.1 Main Objectives

Generating reference standards is a costly and resource intensive endeavor and funding and resources are almost always limited. These resources should be used wisely and new tools and methods that potentially introduce task efficiencies should be evaluated and integrated with reference standard generation workflows. Employing many human reviewers (i.e., "oracles") to carry out manual clinical text annotation involves considerable human effort. Methods to minimize workload may compromise annotation reliability and reference standard validity. Manual annotation of clinical texts often requires a tradeoff between annotator reliability, reference standard validity, and task workload. Developing and evaluating potential efficiencies for generation of reference standards is an area of ongoing research in domains of artificial intelligence and more specifically the clinical NLP community. Building reference standards in ways that allow practical, efficient and scalable approaches should be the end goal for any annotation campaign. Testing methods that can be easily implemented that may improve efficiencies in manual human annotation is often missing in the clinical NLP domain.

Although other studies have addressed similar methods in recent years, the experiments conducted as part of this dissertation research are unique in the way they

were implemented and in the context under which they were used. There are three areas this research addresses. First, this research will identify factors related to annotation task reliability, and reference standard validity in general. Second, evaluate the effects of pre-annotation of semantic concepts on reliability, and reference standard quality and efficiency of manual annotation tasks. Third, assess reliability, reference standard quality and efficiency of manual annotation tasks when pre-annotation is combined with an interactive machine-assisted interface integrated with manual annotation of PHI found in clinical documents. The ultimate goal of this research will be to make generalizable recommendations and demonstrate methods of creating reference standards in more efficient ways while controlling for different sources of bias that noninteractive pre-annotation, or a combined interactive machine-assisted interface coupled with pre-annotation approach may introduce.

It is possible that efficiencies in manual annotation can be achieved but it is important to first define how these tasks are typically done, how they are evaluated, and briefly discuss alternatives strategies, infrastructure required and limitations.

2.2 Typical Annotation Strategies

A reference standard is often developed and used to train and evaluate NLP systems for information extraction or classification tasks. One typical annotation strategy involves double annotation in which two independent human reviewers annotate the same documents with a third clinician reviewer either arbitrating disagreements or breaking ties between the first two annotators (Figure 2.1) [1, 2]. These reviewers function as the "oracles" whose task is to generate labels that are used as a ground truth or reference

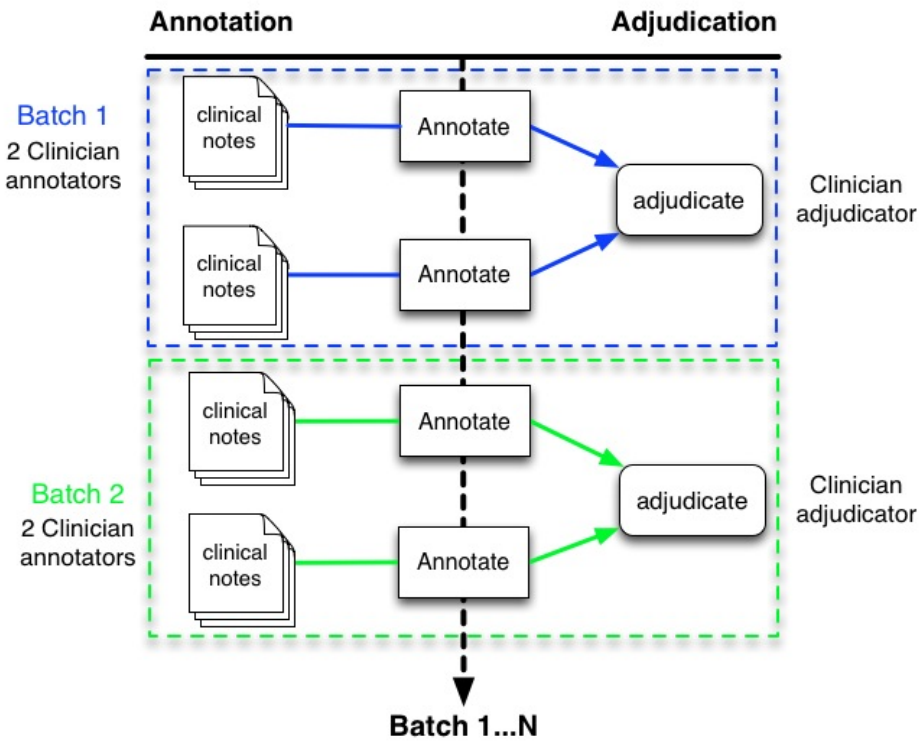

Figure 2.1: One Commonly Used Annotation Strategy

standard used to train and evaluate an NLP system for a specific use-case. Annotation guidelines are often iteratively developed defining inclusion and exclusion criteria for information to include in annotated spans and what to exclude. These guidelines often include linguistic or syntactic rules to aid annotation consistency. A more formal representation of the annotation task is created as part of this iterative process and realized in an annotation tool. Annotators receive training on guidelines, annotation schema and annotation tools during an open annotation phase. This training continues until reviewers meet or exceed some predefined threshold of agreement that adequately reflects the complexity of the annotation campaign. This is followed by a closed annotation phase where annotators are expected to review documents and generate labeled data independently. It is widely assumed that this rigorous approach produces the

highest quality data possible [1-5]. There are specific characteristics of the annotation process that may introduce bias and the effects of introducing methods to reduce annotator workload have not been adequately explored.

Depending on the complexity of the particular NLP task or clinical use case, more efficient and less costly annotation strategies could be implemented [6, 7]. It is also not uncommon to employ an annotation strategy that only uses double annotation early when defining annotation guidelines and for annotator training. Once annotators have reached an acceptable level of agreement only a small proportion of the overall document corpus is doubly annotated [1, 2]. It is generally assumed that the consistency and accuracy of human review tasks is directly related to domain expertise or annotator training [1-13]. However, it may also not be necessary to use clinician annotators for all tasks and nonclinician annotators may produce annotations of similar quality to clinicians [8, 9]. Other potentially more efficient methods could include providing machine or other outputs as pre-annotations to human reviewers who then accept, reject, or modify candidate annotations.

Balancing the goals of task reliability with reference standard validity is further complicated by the complexity of the use case and the goals for NLP system development, nuances of human language, inconsistencies in information quality, differences in how the information was entered or recorded in clinical texts by health care providers, and the cognitive processes involved in review and labeling [10, 13]. For these reasons, there will always be some amount of disagreement between reviewers regardless of attempts to explicitly define review tasks, provide adequate training, or ensure appropriate domain expertise. Adjudication of disagreements may reduce, but not

completely eliminate, the limitations of subjective human review even in situations where tasks are explicitly defined.

2.3 Cognitive Attributes and Their Interplay with Annotation

Annotating clinical text is a complex task characterized by high cognitive load and manual human workload. High reliability and accuracy in human annotation tasks begins with adequate training on specific methods and tools used [8, 11, 14-16]. Adequate training must include some explanation and clear definition of the information categories human reviewers are expected to identify and the tools that will be used [1, 2, 14]. Many attributes of human cognition contribute to the complexity of an annotation campaign. First among these is prior knowledge and training of the reviewers. Prior knowledge can affect reviewer performance since schemas in long-term memory can reduce intrinsic cognitive load. Human reviewers who have more extensive knowledge structures are able to more quickly classify spans of text representing particular information classes [17]. Information scope and context provided to reviewers are also important considerations. However, because annotation is a schema-driven process this may result in more heuristic processing and introduction of reviewer error because reviewers may rely on prior frequently used categories to guide their classification of unknown information, thereby leaving out the underlying meaning, the semantics of what is being annotated [18].

Conscious or unconscious reviewer priming affects categorization accuracy [19, 20]. In the experiments integrated with this dissertation, we use different approaches to "prime" annotators to find information of interest. Specifically these experiments rely on one technique borrowed from the cognitive science literature called "semantic priming"

[21]. Different priming mechanisms may prove useful when human reviewers annotate clinical texts. These effects are most apparent when reviewers are exposed to full documents and full information context [22]. When used as a priming tool, semantic priming functions on implicit memory through which exposure to the priming mechanism influences annotator response. Under the experimental conditions in this dissertation research semantic priming is used to provide the annotator a "primed" candidate spans of information either via the annotation scheme or a machine-generated pre-annotation Aim 2 (Chapter 4) or a pre-annotation coupled with machine-assisted annotation and an interactive annotation interface Aim 3 (Chapter 5).

Human cognition is almost always goal-based [23]. Human decision-making is affected by the underlying task perception of the reviewer and their ability to recall specific facts and access knowledge resources [17, 20]. When humans annotate spans of information found in clinical texts they must deal with the underlying competing goals of speed and accuracy, balanced with the goals of the underlying annotation use case typically explained using annotation guidelines or specific training examples. These factors all compete for limited cognitive resources of the reviewer and may interfere with identification of information and classification [24, 25]. Humans have the ability to regulate and adapt to the complexity of a given task that requires high cognitive resources by making attempts to control their information environment. Reviewers use both conscious and unconscious strategies as a way to acquire information, enhance motivation, reduce uncertainty and minimize cognitive load [26]. It is not uncommon for new rules to be induced, and for guidelines to be revised over the course of an annotation campaign in an attempt to improve annotator consistency, and reduce annotator

uncertainty and task ambiguity. Iterative modifications to rules increase learning curves and may add more task uncertainty resulting in the reviewer reverting to heuristic processing and relying on prior training and internal knowledge resources, which may introduce unintentional bias into the review process.

Annotation has a tendency to be a highly social process and human reviewers often attempt to control for uncertainty by seeking feedback or opinions from others via social support mechanisms that include discussion boards, stand-alone rule sets that may differ from accepted guidelines, chat sessions, moderated question answering, or direct feedback between reviewers. These are all attempts used to resolve uncertainty and reduce task ambiguity. Personalities ultimately affect the views of other reviewers [27]. This reduces task independence and may introduce unintended bias into a review task. No one reviewer can be seen as an expert on every single domain and social factors may inadvertently increase task bias. During an annotation campaign it is not uncommon that one or more reviewers are perceived as an expert. This perception likely has more influence on other reviewers who may be more or less experienced at a particular task. A human reviewer must balance these attributes of human cognition with the goals of minimizing effort, maximizing productivity and accuracy, and personal motivation. For any annotation campaign individual reviewer goals must be balanced at an administrative level with the deadlines of the current NLP development effort, the workload, and costs of hiring multiple reviewers for a given project.

2.4 Qualitative Analysis Used to Generate Hypotheses

Qualitative analysis is an effective method often used to generate hypotheses that provide insight into generalizable factors having an effect on annotation task reliability

and reference standard validity [16, 28]. These analyses consist of an iterative process where the research team identifies themes and constructs related to the annotator's response to semistructured interview questions. The goal behind applying qualitative analysis methods is to provide a representation of annotator perspectives and experience including the annotation process, issues of annotation strategy, usefulness and choice of the tools, uncertainty reduction, effort and decision quality used for an annotation project using clinical documents.

2.5 Alternative Workflows Integrated with Annotation Tasks

Hripcsak [13, 29, 30] suggests several alternative options that may introduce efficiencies into manual review efforts including reducing or controlling the context and scope of information provided for review, reducing the amount of annotation and adjudication required, modifying the number of judges, changing annotation task workflows, and adding or removing reviewers. These options do not represent an exhaustive list of possibilities that could be implemented to carry out annotation tasks and various combinations or modifications to these options could be used depending on the needs of each particular annotation project.

2.6 Assessments of Clinical Text Pre-Annotation

One of the simplest approaches to introduce efficiencies in manual text annotation may include integrating machine-assisted methods that are used to pre-annotate (also referred to as pretagging) relevant spans of text. This changes the annotation workflow slightly allowing the annotator to add missing annotations, modify spans, or delete spurious annotations. Neveol et al. [12], evaluated use of automatic semantic pre-

annotation of PubMed queries showing a significant reduction in the number of required annotations when using pre-annotations, reduction in annotation time and higher interannotator agreement. Other approaches for pre-annotating include using dictionaries coupled with regular expressions [31, 32] based on the Unified Medical Language System (UMLS) Metathesaurus [33] as a source of lexical domain knowledge [34]. For example, a study by Lingren et al. [32] that integrated iteratively developed dictionaries combined with regular expressions to generate pre-annotation of clinical named entities found in a corpus of clinical documents and clinical trial announcements. Pre-annotation has also been used as a priming mechanism to identify information representing PHI [35]. In these cases there was improvement in observed efficiency and reference standard quality when pre-annotations were provided. These results were contradicted in a study by South [31] that suggests annotators produce the highest quality and are more efficient when annotating on raw clinical texts and that pre-annotation of clinical entities provided no observable improvement in data quality or efficiency. These results are also congruent with a study by Ogren [9].

Others have proposed combining pre-annotation with active learning approaches to identify named entities in clinical texts [36]. Still others have used third-party tools to pre-annotate clinical texts to identify UMLS concepts [9] as well as algorithmic approaches to pre-annotation [10] combined with domain expert annotations reused for temporal relation annotation [11]. Ogren [9] suggests limited utility when a third party tool is used for pre-annotation of clinical named entities and Mowery et al. [11] suggest that even with domain expert pre-annotations, additional features are required to discern context dependent labels such as temporality. A study by Ganchev [37] used a semi-

automated approach for named entity recognition that used only binary decisions from annotators to determine true positive or false positive pre-annotations. A study by Rosset et al. [38] demonstrated gains in quality and reduction in annotation times were observed even when an out-of-domain system was used to generate pre-annotations. Rehbein and colleagues [39] demonstrate improvements in data quality for a semantic frame-labeling task using semi-automatic annotation. Finally, Fort and Sagot [40] evaluated using pre-annotation for part-of-speech tagging on the Penn Tree bank corpus and demonstrate a gain in quality and annotation speed even when a not so accurate tagger is used to provide pre-annotations.

Several annotation tools geared towards integrating noninteractive or interactive machine-assisted approaches have been developed. These tools include the BRAT annotation tool developed by Stenetorp [41], eHOST (extensible Human Oracle Suite of Tools) [42] developed as part of this dissertation research (Chapter 5), RapTAT (Rapid Text Annotation Tool) [43, 44], the MITRE Identification Scrubber Toolkit (MIST) system for pre-annotation of PHI [45], and ABNER developed by Settles [46]. Various approaches are integrated with these tools to speed up annotation using web 2.0 technology to support collaborative annotation [41], use of dictionaries, regular expressions, and an interactive annotation interface as a means to provide pre-annotations [42], machine learning coupled with naïve Bayes modeling [43, 44], simple bootstrapping [45], or statistical machine learning using linear-chain Conditional Random Fields (CRFs) [46].

Each of the above mentioned experiments and tools support some utility (even if limited), in using various approaches to provide pre-annotations to annotators as a means

of improving efficiency and data quality of manual human annotation tasks. Even though there have been a number of preliminary studies integrating pre-annotation in the domains of computational linguistics and clinical NLP, the corpora and annotation tasks in these studies are difficult to compare. Although the results of some of these studies are encouraging, the effects of pre-annotation on efficiency and data quality are still worth exploring when applied to different domains and annotation that supports various NLP development goals.

There are several potential biases that are possible when pretagging is used as a means to reduce the time or costs associated with annotation of texts. First, humans may concentrate only on pre-annotated information correcting pre-annotations without adding what is missing. On the other hand, human annotators may also concentrate too much on what is missing but not correct pre-annotations due to the volume or complexity of what is pre-annotated. Second, it is difficult (if not impossible) for some types of pre-taggers to produce high enough quality pre-annotations because the tools are difficult to build or might not exist (i.e., coreference resolution would be a good example). The quality of pre-annotation is directly related to training and domain adaptation on the document types or clinical domain to which it is applied. Pre-annotation may also lead to classification error when the pre-annotated span is assigned to some incorrect class in an annotation scheme. These types of errors can easily be missed in cases where the number of pre-annotations becomes overwhelming to human reviewers.

2.7 Evaluation Metrics

Measures of annotation reliability, validity, and annotation task efficiency were calculated using data produced from the annotation efforts related to Aims 2 and 3 of this

dissertation. These metrics are introduced below and folded into the discussion as they relate to each experiment discussed in Chapters 3-5.

2.7.1 Reliability Metrics: Inter-Annotator Agreement (IAA)

One commonly used metric to assess annotation task reliability is an assessment of agreement between annotators that labeled the same clinical document. This is often measured by calculating interannotator agreement (IAA) using the general formula described in the clinical NLP domain by Hripcsak [29, 30] and Roberts [1, 2], and also explored in the computational linguistics domain by Artstein [47, 48]. In Figure 2.2 IAA (Equation 2.1) is plotted for a hypothetical annotation task showing how interannotator agreement should increase incrementally by some iterative unit of analysis which could be documents distributed into annotator specific batches or at the individual document level or at the annotated span. Interannotator agreement should increase relative to some human annotator-training goal (blue sigmoid curve shown as "goal"). The goal is to reach an acceptable training plateau (black circle) within a few batches or iterations (red sigmoid curve shown as "reality") from the resulting annotations. Depending on task complexity and the interplay between factors affecting the underlying cognitive task, this training plateau may be reached within only a few iterations or it may never be reached. Estimation of IAA is normally conducted in the absence of a "ground truth."

One possible way to limit annotation bias is to use many reviewers. It is also possible that annotators can be highly consistent (reliable), but not accurate. Low interrater reliability (low interannotator agreement) is not always a methodological weakness for a given review task. Artstein suggests [48] as the number of annotators increases, the effect of their individual preferences starts to approximate random noise.

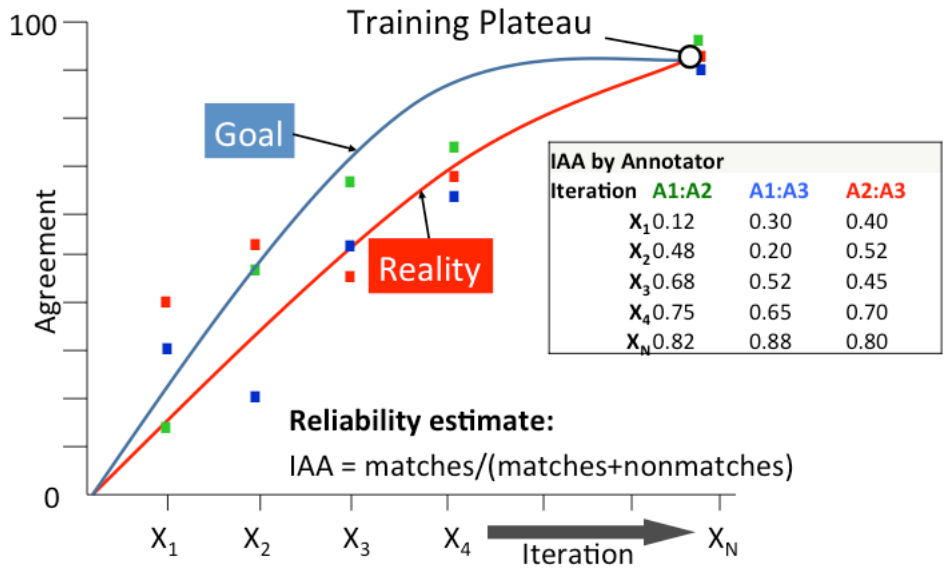

Figure 2.2: Reliability Metrics: Inter-Annotator Agreement by Iteration

The general equation for IAA is as follows:

$$IAA = \frac{2 \times Matches}{(2 \times Matches + Non\text{-}Matches)} \tag{2.1}$$

For the experiments in Aims 2 and 3 both exact match (annotation offset start and end match exactly) and inexact match (annotation offset start and end at least overlap) IAAs were calculated at the annotation class, and any attribute levels (Equation 2.2). Batch averages and overall batch IAAs were calculated.

$$Batch\ Average\ IAA = \frac{\sum_{i=1}^{N} IAA_i}{N},$$

$$Batch\ Overall\ IAA = \frac{\sum_{i=1}^{N}(2\times Matches_i)}{\sum_{i=1}^{N}(2\times Matches_i + Non\text{-}Matches_i)} \quad (2.2)$$

where N is the number of records in a batch.

2.7.2 Validity Metrics: Recall, Precision, F-Measure

At the end of an annotation campaign a final reference standard is assembled from the resulting annotator labels and is used to assess task validity. True Positives (TP) occur when the annotator has the same annotation as the reference standard. True positives are often calculated for two levels; exact and inexact. For exact calculations the annotations from both the annotator and reference standard were identical. For inexact calculations the annotations match at least partially. False Positives (FP) occur when the annotator has annotated something that is not in, has a different class, different attribute value, or different relationship than the reference standard. False Negatives (FN) occur when the annotator failed to annotate something that is in the reference standard (Figure 2.3). These typical validations are performed at various levels:

- Span – Based on the start and end of the annotation.
- Class – Based on the class attributed to the annotation.
- Attribute – Based on the values assigned as an attribute of an annotation class.
- Relation level – Based on a relationship between two annotations.

Formulae used for these calculations rely on true positives (TP), false positives (FP), and false negatives (FN) shown in the 2x2 contingency table in Figure 2.3.

Typical metrics of recall, precision and F-measure are used to assess annotator performance and reference standard validity. With beta weighted accordingly where F_1-measure represents the mean of recall and precision and the F_2-measure that weights recall twice as high as precision. Recall, precision, and a corresponding F-measure are calculated using the formulae shown in Equation 2.3. For outcomes that place higher emphasis on recall it is not uncommon to use the F_2-measure (which uses a β weight of 2) along with the F_1-measure.

$$Recall = \frac{TP}{(TP + FN)}, \quad Precision = \frac{TP}{(TP + FP)},$$
$$F\beta - \text{measure} \; = \; [(1 + \beta 2)(\text{P} * \text{R})] \; / \; (\beta 2\text{P} \; + \; \text{R}) \tag{2.3}$$

with β being the weight placed on recall (R) or precision (P).

		Reference Standard	
		Yes	No
Annotator	Yes	TP	FP
	No	FN	TN

Figure 2.3: 2x2 Contingency Table

In Figure 2.4 F-measure is plotted for a hypothetical annotation task showing the incremental increase in annotation accuracy across some iterative unit of analysis. Under these conditions, a reference standard ("ground truth") is used to measure annotator accuracy in the reference standard's representation of the task. Like IAA, F-measure should increase relative to some human annotator-training goal (blue sigmoid curve shown as "goal"). The goal is to reach an acceptable training plateau (black circle) within a few batches or iterations (red sigmoid curve shown as "reality") from the resulting annotations. F_1-measure is often used as a surrogate for Kappa since in a large heterogeneous document corpus the number of possible mentions corresponding with true negatives is unknown or could be very large [29, 30, 47].

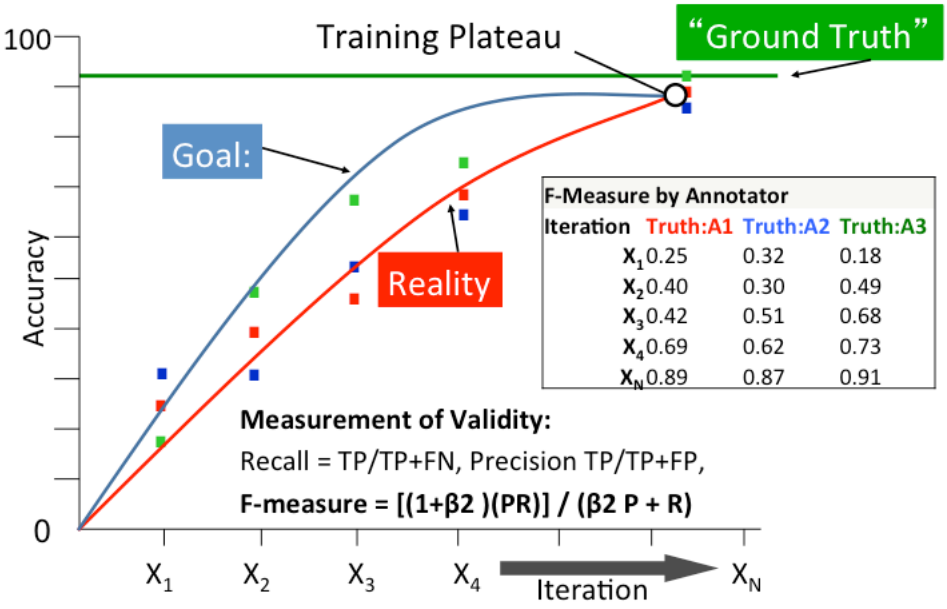

Figure 2.4: Validity Metrics: Annotator Performance by Iteration

2.7.3 Micro- and Macroaveraging

Both micro- and macroaveraged metrics are often calculated when IAA and F-measures are the dependent variables used to assess annotator performance. Microaveraged performance metrics are calculated based on the instance level annotations for the entire batch then combined across all classes calculating measures using these summed totals. Microaveraging counts each instance of annotation with equal weight, whereas macroaveraging calculates metrics for each category then takes the average across all classes. Because microaveraging assigns equal weight to each annotated class – even those that rarely occur – performance is generally lower when using micro-averaging. Micro- and macroaverages were calculated across classes at the batch level (Equation 2.4). Finally, micro- and macroaverages are often calculated across annotation classes at the batch level for standard validity metrics of recall, precision and F-measure (Equation 2.5).

$$IAA\ Micro\ Average = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} 2\times Matches_{ij}}{\sum_{i=1}^{M}\sum_{j=1}^{N}(2\times Matches_{ij} + Non\text{-}Matches_{ij})}$$

$$IAA\ Macro\ Average = \frac{\sum_{i=1}^{M}\left(\frac{\sum_{j=1}^{N}(2\times Matches_{j})}{\sum_{j=1}^{N}(2\times Matches_{j} + Non\text{-}Matches_{j})}\right)_{i}}{M} \quad (2.4)$$

where M is the total number of classes and N is the total number of records in a batch.

$$Recall\ Micro\ Average = \frac{\sum_{i=1}^{M}\left(\sum_{j=1}^{N} TP_j\right)_i}{\sum_{i=1}^{M}\left(\sum_{j=1}^{N}\left(TP_j + FN_j\right)\right)_i},$$

$$Precision\ Micro\ Average = \frac{\sum_{i=1}^{M}\left(\sum_{j=1}^{N} TP_j\right)_i}{\sum_{i=1}^{M}\left(\sum_{j=1}^{N}\left(TP_j + FP_j\right)\right)_i},$$

$$F\text{-}measure\ Micro\ Average$$

$$= \frac{2\times Recall\ Micro\ Average\times Precision\ Micro\ Average}{(Recall\ Micro\ Average + Precision\ Micro\ Average)}$$

$$Recall\ Macro\ Average = \frac{\sum_{i=1}^{M}\left(\frac{\sum_{j=1}^{N}\left(TP_j\right)}{\sum_{j=1}^{N}\left(TP_j + FN_j\right)}\right)_i}{M},$$

$$Precision\ Macro\ Average = \frac{\sum_{i=1}^{M}\left(\frac{\sum_{j=1}^{N}\left(TP_j\right)}{\sum_{j=1}^{N}\left(TP_j + FP_j\right)}\right)_i}{M},$$

$$F\text{-}measure\ Macro\ Average = \frac{\sum_{i=1}^{M}\left(\frac{2\times\frac{\sum_{j=1}^{N}\left(TP_j\right)}{\sum_{j=1}^{N}\left(TP_j + FN_j\right)}\times\frac{\sum_{j=1}^{N}\left(TP_j\right)}{\sum_{j=1}^{N}\left(TP_j + FP_j\right)}}{\frac{\sum_{j=1}^{N}\left(TP_j\right)}{\sum_{j=1}^{N}\left(TP_j + FN_j\right)} + \frac{\sum_{j=1}^{N}\left(TP_j\right)}{\sum_{j=1}^{N}\left(TP_j + FP_j\right)}}\right)_i}{M}$$

(2.5)

where M is the total number of classes and N is the total number of records in a batch.

2.7.4 Annotation Task Efficiency

Annotation task efficiency can be estimated based on workload metrics measured using time for annotation, adjudication, and reported annotator effort as real monetary cost. Efficiency metrics should exhibit an inverse trend to reliability and validity. In Figure 2.5, time is plotted for a hypothetical annotation task showing the incremental decrease in time as efficiency improves across some iterative unit of analysis. In this example time decreases relative to some human annotator-training goal (blue sigmoid curve shown as "goal"). Once an acceptable training plateau (black circle) is achieved these measurements exhibit a decreasing trend (red sigmoid curve shown as "reality").

When machine outputs are used as pre-annotations it is also necessary to calculate additional metrics of missing annotations (i.e., where one annotator made an annotation and the other annotator did not), as well as matching, modified (i.e., span or class

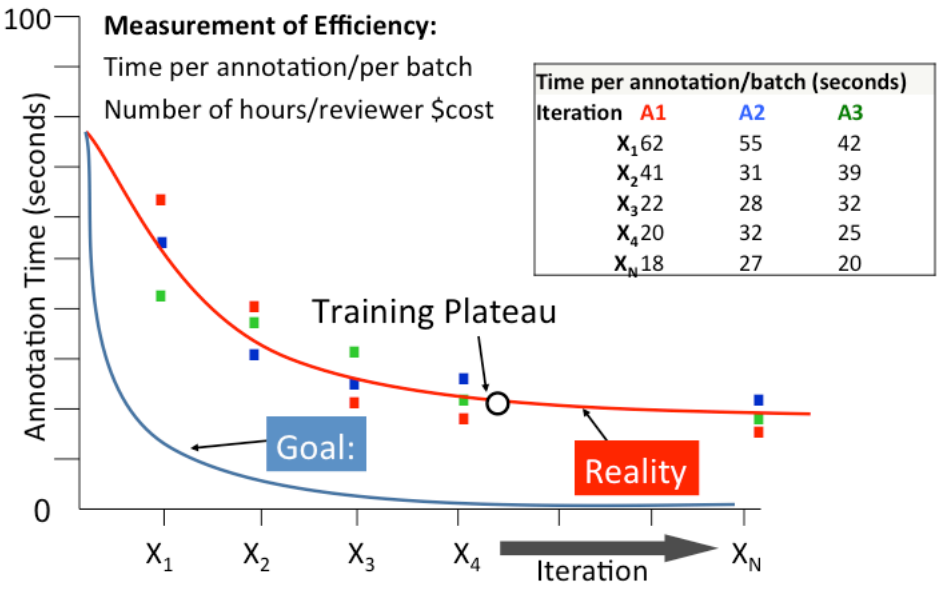

Figure 2.5: Efficiency Metrics: Annotator Task Time/Cost

modifications), deleted or added annotations. Where pre-annotations were provided and generated annotations were compared to the pre-annotated version of the same document. Percentages were calculated for matching, modified, deleted, and added annotations at the span and annotation class level according to the following criteria and Equation 2.6.

- Matching – Annotations were identical to the pre-annotation.
- Modified – Annotations were changed somehow from the pre-annotation.
- Deleted – Annotations were in the pre-annotation but not the annotated file.
- Added – Annotations were in the annotated file but not the pre-annotation.

In order to identify where the modifications took place (i.e., either at the annotated span or annotation class level) the modified category was further broken down into the following subgroups:

- Span – If the modification took place in the annotation itself.
- Class – If the pre-annotated class was changed by the annotator.

$$Matching\ \% = \frac{Total\ Matching\ Annotations}{Total\ Annotations},$$
$$Modified\ \% = \frac{Total\ Modified\ Annotations}{Total\ Annotations},$$
$$Deleted\ \% = \frac{Total\ Deleted\ Annotations}{Total\ Annotations},$$
$$Added\ \% = \frac{Total\ Added\ Annotations}{Total\ Annotations} \quad (2.6)$$

2.7.5 Annotation Task Difficulty

Evaluation of annotation task difficulty can be measured and visualized in several ways. Task difficulty can be calculated in a post hoc fashion where metrics of reliability and validity can be seen as a function of annotator time, proportional to the number of items and classifications included in the annotation scheme and the density of annotations (annotation prevalence versus the number of words, lines, or tokens in a given document). Task difficulty can also be assessed using post hoc regression techniques to predict the number of annotations required to achieve adequate coverage or reach performance thresholds for information classes defined in an annotation scheme. Task difficulty can be visualized using locally estimated scatterplot smoothing (LOESS) [49] to model annotator training plateaus using readily calculated evaluation metrics. Estimates for annotator training plateaus will be steeper or shallower depending on the cognitive load associated with an annotation task.

2.7.6 Annotation Quality

IAA can be used to assess annotation task reliability and consistency with guidelines. Metrics of validity (i.e., recall, precision and appropriately weighted F-Measure) can be used to assess annotator accuracy with some reference standard. Annotation time or cost associated with reviewing is also one dimension quality. Together, the dimensions of IAA, metrics of validity, and annotation workload based on annotation time or direct cost can be used to assess the quality of both the annotation task and the generated reference standard. When generating reference standards these three dimensions must be balanced with NLP development expectations and administrative overhead.

2.7.7 Time Estimates

Mean, variance, 95% confidence intervals, minimum, maximum, median, and sum for time between annotations were calculated by annotator batch, the overall batch (all annotators included), and the overall source (all annotators and batches included). For analyses in experiments 1 and 2, outliers with annotation times greater than 900 seconds were excluded from the analysis.

2.7.8 Estimating Intervention Effects

Locally estimated scatter-plot smoothing (LOESS) [49] is useful not only to visualize annotator training plateaus for data that are nonnormally distributed, but as a means to estimate the ith document or batch where performance has plateaued for an annotation task as a whole or an annotation subtask within an annotation campaign. Once it is determined where a given training plateau exists, generalized estimating equations (GEE) [50] can be used to determine the effects of interventions that are integrated with annotation workflows for Aims 2 and 3. Paired T-tests can also be used to estimate intervention effects where subjects are matched in some way but caution should be exercised since it is not uncommon for data generated from an annotation campaign to exhibit a nonnormal distribution. For this reason, Wilcoxon rank sum test (Mann-Whitney U) or GEE [50] may offer a better method to estimate intervention effects since these are nonparametric techniques that can be used where data are nonnormally distributed or there is some unknown correlation between outcomes.

2.8 Infrastructure Development

Many human annotators were recruited to carry out annotation tasks for the experiments in Aims 2 and 3 of this dissertation project. For Aim 2, 12 annotators (8 clinician and 4 nonclinician), and for Aim 3, 7 annotators (3 clinician and 4 non-clinician) were recruited. These annotators were trained on the annotation tasks and tools used and were expected to achieve a predefined performance threshold before being allowed to annotate documents included in the closed annotation phase for each experiment.

This dissertation utilized two different annotation tools. One tool called Knowtator [51, 52] was used for the annotation tasks described in Chapters 3 and 4. Knowtator is one of the most widely used open source annotation tools. Enhancements were made to the Knowtator tool used for annotation tasks in Aims 1 and 2 (Chapters 3, 4). One modification allowed side-by-side comparison of annotations, allowing an adjudicator to easily resolve disagreements as part of consensus set generation. Another modification made assertion annotation compulsory, reducing the potential for annotators to miss assigning an assertion. Further modifications included reducing the number of actions required to create an annotation. The use of a complex slot value was added to relations allowing an annotator to create a relation between one concept and another in a single mouse click. Another annotation tool called the extensible Human Oracle Suite of Tools (eHOST) [42], introduced in the Chapter 5, was developed to test the experimental conditions evaluated and reported in Chapter 6. These development efforts were supported by the VA CHIR, VINCI and iDASH collaborations.

2.9 References

1. Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J., Roberts, I., Setzer, A., Tapuria, A., Wheeldin, B. The CLEF corpus: semantic annotation of clinical text. In: AMIA Annu Symp Proc, 2007. 625-9.

2. Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., Setzer, A. Building a semantically annotated corpus of clinical texts. In: J Biomed Inform, 2009. 42(5): 950-66.

3. Grishman, R., Sundheim, B. Message Understanding Conference-6: a brief history. 16th Conference on Computational Linguistics (COLING), 1996: p. 466-71.

4. Hersh, W., Bhupatiraju, R.T., Corley, S. Enhancing access to the Bibliome: the TREC Genomics Track. Stud Health Technol Inform, 2004. 107(Pt 2): p. 773-7.

5. Sparck Jones, K. Reflections on TREC Information Processing Management. 1995. 31(3): p. 291-314.

6. Settles, B., Craven, M., Friedland, L. Active learning with real annotation costs. In: Proceedings of the NIPS Workshop on Cost-Sensitive Learning. 2008.

7. Settles, B. Active learning literature survey. In: Computer Sciences Technical Report 1648. University of Wisconsin-Madison. 2009.

8. Chapman, W.W, Dowling, J.N., Hripcsak, G. Evaluation and training with an annotation scheme for manual annotation of clinical conditions from emergency department reports. Int J Med Inform. 2008;77(2):107-113.

9. Ogren, P.V., Savova, G.K., Chute, C.G. Constructing evaluation corpora for automated clinical named entity recognition. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC 2008: 3143-3150.

10. Chapman, W.W., Chu, D., Dowling, J.N. ConText: An algorithm for identifying contextual features from clinical text. In: ACL-07 2007.

11. Mowery, D., Harkema, H., Chapman, W.W. Temporal annotation of clinical text. In: ACL-08 2008.

12. Névéol, A., Islamaj-Doğan, R., Lu, Z. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. In: J Biomed Inform. 2011 Apr; 44(2):310-8.

13. Hripcsak, G., Wilcox, A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. In: J Am Med Inform Assoc. 2002. Jan-Feb;9(1):1-15.

14. Dandapat, S., Biswas, P., Choudhury, M., Bali, K. Complex linguistic annotation - no easy way out! A case from Bangla and Hindi POS labeling tasks. In Proceedings of the Third ACL Linguistic Annotation Workshop, Singapore. 2009.

15. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y. Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of EMNLP 2008, pages 254-263.

16. Campbell, E.M., Sittig, D.F., Chapman, W.W., Hazlehurst, B.L., Cohen, A.M. Understanding inter-rater disagreement: a mixed methods approach. In: AMIA Annu Symp Proc. 2010: p. 81-5.

17. Estes, W. Classification and Cognition. 1996. New York: Oxford University Press.

18. Koehler, D., Brenner, L., Griffin, D. The calibration of expert judgment: heuristics and biases beyond the laboratory. In: Heuristics and Biases: The Psychology of Intuitive Judgment. T. Gilovich, D. Griffin, and K. D, Editors. 2002, Cambridge University Press: New York.

19. Underwood, G. Implicit Cognition. 1996, Oxford: Oxford Science Publications.

20. Heckhausen, J., Heckhausen, H. Motivation and Action. Cambridge University Press, Cambridge, UK., 2008.

21. Ferrand, L., New. B. (2004) Semantic and associative priming in the mental lexicon. In Bonin (Ed), The Mental Lexicon. New York: Nova Science Publishers. 2003. pp 25-43.

22. MacCoon, D., Wallace, J.F., Newman, J.P. Self-regulation: context-appropriate balanced attention. In: R. Baumeister and K. Vohs (Eds) Handbook of Self-Regulation: Research, Theory and Applications. New York: 2004 (Guilford Press): p. 422-446.

23. Wyer, R.S. Social Comprehension and Judgment: The Role of Situation Models, Narratives, and Implicit Theories. 2004, Mahwah, NJ: Erlbaum.

24. Kruglanski, A. Lay Epistemics and Human Knowledge: Cognitive and Motivational Bases. New York: Plenum. 1989.

25. Hollnagel, E. The ETTO Principle: Efficiency-Thoroughness Trade-off: Why Things that Go Right Sometimes Go Wrong. Cornwall, Britain: Ashgate. 2009.

26. Weir, C.R., Nebeker, J.R., Hicken, B.L., Campo, R., Drews, F., LeBar, B. (2007). A cognitive task analysis of information management strategies in a computerized provider order entry environment. J Am Med Inform Assoc, 2007. 14(1): p. 65-75.

27. Harton, H.C., Green, L.R., Jackson, C., Latane, B. Demonstrating dynamic social impact: consolidation, clustering, correlation, and (sometimes) the correct answer. Teaching of Psychology, 1998. 25:p. 31-34.

28. Patton, M.Q. Qualitative Research and Evaluation Methods. 2002. Sage Publications.

29. Hripcsak, G., Heitjan, D.F. Measuring agreement in medical informatics reliability studies. In: J Biomed Inform. 2002. Apr;35(2):99-110.

30. Hripcsak, G., Rothschild, A.S. Agreement, the f-measure, and reliability in information retrieval. In: J Am Med Inform Assoc. 2005. May-Jun;12(3):296-8.

31. South, B.R., Shen, S., Barrus, R., DuVall, S.L., Uzuner, Ö., Weir, C. Qualitative analysis of workflow modifications used to generate the reference standard for the 2010 i2b2/VA challenge. In: AMIA Annu Symp Proc. 2011.

32. Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., Stoutenborough, L., Li, Q., Solte, I. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. J. Am. Med. Inform. Assoc doi: 10.1136/amiajnl-2013-001837.

33. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium: American Medical Informatics Association; 2001:17.

34. Friedman, C., Liu, H., Shagina, L., Johnson, S., Hripcsak, G. Evaluating the UMLS as a source of lexical knowledge for medical language processing. In: Proc AMIA Symp, 2001: 189-93.

35. South, B.R., Shen, S., Friedlin, F.J., Samore, M.H., Meystre, S.M. Enhancing annotation of clinical text using pre-annotation of common PHI. In: AMIA Annu Symp Proc. 2010.

36. Skeppstedt, M. Annotating named entities in clinical text by combining pre-annotation and active learning. Proceedings of the ACL Student Research Workshop. pp 74-80. Sofia, Bulgaria, August 4-9 2013. Association for Computational Linguistics. 2013.

37. Ganchev, K., Pereira, F., Mandel, M., Carrol, S., White. Semi-automated named entity annotation. Proceedings of the Linguistic Annotation Workshop; Prague, Czech Republic: Association for Computational Linguistics, 2007:53–6.

38. Rosset, S., Grouin, C., Lavergne, T., Jannet, M.B., Leixa, J., Galibert, O., Zweigenbaum, P. Automatic named entity pre-annotation for out-of-domain human annotation. In: Proceedings of the 7th Linguistic Annotation Workshop &

Interoperability with Discourse, pages 168-177, Sofia Bulgaria, August 8-9, 2013. ACL 2013.

39. Rehbein, I., Ruppenhofer, J., Sporleder, C. Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. In: Proceedings of the Third Linguistic Annotation Workshop, ACL 2009.

40. Fort, K., Saggot, B. Influence of pre-annotation on POS-tagged corpus development. In: Proceedings of the Fourth Linguistic Annotation Workshop. ACL 2010: 56-63.

41. Stenetorp, P., Pyysalo, S., Topic, G., Ananiadou, S., Tsujii. J. BRAT: a web-based tool for NLP-assisted text annotation. EACL 2012;2012:102.

42. South, B.R., Shen, S., Leng, J., Forbush, T.B., DuVall, S.L., Chapman, W.W. A prototype tool set to support machine-assisted annotation. In: BioNLP 2012. Montreal, Canada.

43. Gobbel, G.T., Reeves, R., Jayaramaraja, S., Giuse, D., Speroff, T., Brown, S.H., Elkin, P.L., Matheny, M.E. Development and evaluation of RapTAT: a machine learning system for concept mapping of phrases from medical narratives. Journal of Biomedical Informatics. December, 2013.

44. Gobbel, G.T., Garvin, J., Reeves, R., Cronin, R.M., Heavirland, J., Williams, J., Matheny, M.E. Assisted annotation of medical free text using RapTAT. J Am Med Inform Assoc. doi: 10.1136/amiajnl-2013-002255

45. Hanauer, D., Aberdeen, J., Bayer, S., Wellner, B., Clark, C., Zheng, K., Hirschman, L. Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. Int. J. Med. Inform. 82 (2013) 821-831.

46. Settles, B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 2005;21:3191–2.

47. Artstein, R., Poesio, M. Inter-coder agreement for computational linguistics. Computational Linguistics 34(4): 555-596, 2008.

48. Artstein, R., Poesio, M. Bias decreases in proportion to the number of annotators. In: Gerhard Jaeger, Paola Monachesi, Gerald Penn, James Rogers, and Shuly Wintener (Eds.), Proceedings of FG-MoL 2005, pp. 141-150. Edingurgh, August 2005.

49. Cleveland, W.S. LOWESS: a program for smoothing scatterplots by robust locally weighted regression. The American Statistician. 1981. 35 (1): 54.

50. Hardin, J., Hilbe, J. Generalized Estimating Equations. London: Chapman and Hall/CRC. 2003.

51. Ogren, P.V. Knowtator a protege plug-in for annotated corpus construction. In: Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2006: 273-5.

52. Musen, M.A., Gennari, J., Eriksson, H., Tu, S.W., Puerta, A.R. PROTEGE-II: computer support for development of intelligent systems from libraries of components. In: Medinfo 1995.

CHAPTER 3

AIM 1: QUALITATIVE ANALYSIS OF WORKFLOW MODIFICATIONS USED TO GENERATE THE REFERENCE STANDARD FOR THE 2010 I2B2/VA CHALLENGE

Originally published in

B.R. South, S. Shen, R. Barrus, S.L. DuVall, Ö. Uzuner, C. Weir.

2011. AMIA Annu Symp Proc. 2011.

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243132/pdf/1243_amia_2011_proc.pdf

Qualitative Analysis of Workflow Modifications Used to Generate the Reference Standard for the 2010 i2b2/VA Challenge

Brett R South, MS[1,2,3], Shuying Shen, MStat[1,2,3], Robyn Barrus, MS[1], Scott L. DuVall, PhD[1,2], Ozlem Uzuner, PhD[4], Charlene Weir, PhD[1,2,3]

[1]IDEAS Center SLCVA Healthcare System, Salt Lake City, Utah,
[2]Department of Internal Medicine, Division of Epidemiology, University of Utah,
[3]Department of Biomedical Informatics, University of Utah
[4] University at Albany, SUNY, Albany, New York, USA

ABSTRACT

The Department of Veterans Affairs (VA) and the Informatics for Integrating Biology and the Bedside (i2b2) team partnered to generate the reference standard for the 2010 i2b2/VA challenge task on concept extraction, assertion classification, and relation classification. The purpose of this paper is to report an in-depth qualitative analysis of the experience and perceptions of human annotators for these tasks. Transcripts of semi-structured interviews were analyzed using qualitative methods to identify key constructs and themes related to these annotation tasks. Interventions were embedded with these tasks using pre-annotation of clinical concepts and a modified annotation workflow. From the human perspective, annotation tasks involve an inherent conflict between bias, accuracy, and efficiency. This analysis deepens understanding of the biases, complexities and impact of variations in the annotation process that may affect annotation task reliability and reference standard validity that are generalizable for other similar large-scale clinical corpus annotation projects.

INTRODUCTION

A reference standard is required to conduct the supervised learning methods used for natural language processing (NLP) and functions as a ground truth (or gold standard) for training and evaluation data. Manual annotation processes are usually used to create this reference standard. The process usually involves double annotation with strict adjudication to resolve discrepant annotations. Generating coded data in this way is expensive and involves high levels of cognitive workload for human reviewers. Evaluating methods that reduce human workload without introducing task bias, or reducing reference standard reliability and validity is one area of active research in the clinical NLP domain.

An assessment of annotator experience and perceptions related to the process of generating a reference standard is rarely addressed in published clinical NLP studies. Understanding the cognitive processes involved in annotation is particularly important when those reference standards are dependent upon human information extraction and classification. A qualitative analysis of the subjective experience of annotators is particularly informative and could provide additional data as well as lead to new hypotheses to support recommendations for other large-scale clinical corpus annotation efforts. The overarching goal is to create generalizable and evidence-based annotation methods.

The Department of Veterans Affairs (VA) Consortium for Healthcare Informatics Research partnered with the informatics for Integrating Biology and the Bedside (i2b2) team to generate the reference standard for the 2010 i2b2/VA challenge. Annotation tasks for the 2010 i2b2/VA challenge focused on concepts, assertions, and relations and were particularly complex, including both information extraction and classification. The annotation workflow was intentionally manipulated in order to test specific hypotheses. These manipulations included: 1) use of concept level pre-annotation versus none; 2) conditioning on sentence level context versus document level context and; 3) modification of adjudication steps.

This paper reports the results of a qualitative analysis of the experiences and perceptions of annotators for the 2010 i2b2/VA challenge related to these variations. Using qualitative methods adapted from Patton [1] we identified key constructs and themes related to annotation tasks. Data for our analyses were derived from semi-structured annotator interviews developed to explore task characteristics that might impact the reliability and validity of a reference standard created to support NLP systems development and evaluation for the 2010 i2b2/VA challenge. We also

address issues related to administrating a large-scale clinical annotation effort in the context of a community shared-task challenge.

BACKGROUND

We define human annotation tasks as a schema based manual review process that is used to identify spans of text that represent instances of particular target information classes; a human information extraction task. Annotation tasks may also include classification steps such as assignment of assertions, or identifying relations between information classes.

Common Clinical Corpus Annotation Processes. One commonly used workflow to generate a reference standard involves double annotation with strict adjudication to resolve discrepant annotations and break ties. Guidelines are often iteratively developed and specify target information classes and how much information to annotate along with specific examples of inclusions and exclusions related to each information class or attribute [2-6]. During the iterative process, an annotation schema that functions as a more formal representation of the information classes and their attributes is developed. Annotators receive training on the annotation task, as well as the annotation schema, guidelines, and tools. Often syntactic rules are incorporated with these tasks to improve annotation consistency and reduce differences in annotated spans. Annotator training usually continues until a set threshold of agreement is achieved that adequately reflects the complexity of the annotation task. It is widely assumed that this rigorous approach produces the highest quality reference standard possible. However, the characteristics of the process that might lead to systematic bias and the effects of methods to reduce annotator workload have not been fully explored.

Cognitive Attributes of the Annotation task. Annotating free texts is a complex cognitive task, characterized by high cognitive load and workload. Several key attributes of human cognition contribute to the complexity of the annotation process. First, appropriate categorization is highly dependent on prior knowledge. The more extensive the knowledge structures, the faster and more automatic the classification process [7]. The downside of extensive knowledge and well-developed schematic structures is that prior well-used categories might interfere with the classification process. The result may be more heuristic processing and introduction of possible errors [8]. Therefore, for any annotation task, the context and the information scope are important considerations from the perspective of the human reviewer. Second, categorization will vary as a function of priming, both conscious and unconscious [9] [10]. The impact of priming is expected to be the strongest when reading full documents and the task involves substantial inference. Third, almost all human cognition is goal-based [11]. Information processing goals drive perception, recall and decision-making [7, 10]. The annotation process includes two prevailing goals; speed and accuracy, plus the goals of the annotation task itself, and they all compete for attentional resources [12, 13]. These competing goals may occur automatically and interfere with identification of information and classification. Humans regulate and adapt to this complexity by actively manipulating and attempting to control their information environment. They use a wide variety of strategies to acquire information, reduce stress related to uncertainty, minimize cognitive load and enhance self-motivation [14]. Over the course of an annotation project, rules may be induced to improve task consistency, clarify or attempt to resolve ambiguity in project guidelines, or reduce annotator uncertainty. These changes in the rules result in increased learning curves and more uncertainity.

Social influence of other annotators may inadvertently increase bias. Annotators who are perceived by others as being more expert will likely have more influence on others than an annotator who has little experience. Personalities may also have differential impact on the views of others [15]. Annotation is a social process and often involves some type of support mechanism such as discussion boards, moderated question answering, or direct annotator feedback to help improve annotation consistency and resolve uncertainty. Annotators must strike a balance between the conflicting human goals of minimizing effort, maximizing task accuracy, and maintaining motivation. These annotator specific goals must also be balanced at an administrative level with project deadlines, task workload, and annotation costs that are dependent upon fixed budgets.

Qualitative Analysis as a Method of Hypothesis Generation. Qualitative analysis is an effective method that can be used to generate hypotheses that provide insight into generalizable factors that have an effect on annotation reliability and validity. Our goal behind applying qualitative analysis methods is to provide a representation of annotator perspectives on the experience of the annotation process, the usefulness of tools, and the impact of various approaches used for a large-scale clinical corpus annotation project.

METHODS

For the 2010 i2b2/VA challenge, we modified the workflow and integrated interventions with annotation tasks used to generate the reference standard. This section begins with a description of the data sources, study subjects, and types of interventions associated with the 2010 i2b2/VA Challenge. We also describe the interventions in order to adequately frame the discussion of constructs and themes identified from qualitative analysis of semi-structured interviews of all participating annotators.

The i2b2 Challenge Tasks. Previous i2b2 challenges [16-19] have provided a valuable source of labeled data for the medical informatics community across various challenge tasks. For all but the 2009 i2b2 medication extraction challenge [19] these reference standards were generated by two domain experts who were allowed to discuss annotations as they were made, disagreements were arbitrated by a third domain expert. For the 2009 i2b2 medication extraction challenge the i2b2 team conducted a community annotation experiment [20] where annotations were submitted and arbitrated by participating teams. The 2010 i2b2/VA challenge on concepts, assertions, and relations [21] involved both information extraction and classification. An information extraction step was necessary to identify spans of text representing medical problems, tests, and treatments. Classification tasks included classifying assertions for medical problems; and another to classify relations between medical problems and other medical problems, medical problems and tests, and medical problems and treatments.

Data Sources. A total of 826 documents obtained from three healthcare institutions were annotated and released to the community for the 2010 i2b2/VA challenge. Documents were de-identified and released only after appropriate data use agreements (DUA) were signed by annotators and all participants of challenge teams. These included 230 discharge summaries from Partners Healthcare, 196 discharge summaries from Beth Israel Deaconess Medical Center, and 200 progress notes, and 200 discharge summaries from University of Pittsburgh Medical Center. Redacted PHI elements for BIDMC, and Partners data sources were resynthesized with realistic surrogates.

Annotator Training and Methods of Feedback. Generating the reference standard for the 2010 i2b2/VA challenge involved the efforts of 12 part-time annotators over six months. Eight clinician and four non-clinician annotators were recruited for these tasks. All non-clinician annotators had at least a Bachelor's degree and one non-clinician annotator was a non-native English speaker. We made no assumption about previous experience with annotation tools or familiarity with syntactic rules that were used to guide annotated information. Clinician annotators included nurses, nursing students, a respiratory therapist, and a Bachelor's in Pharmacy. We purposefully recruited both clinician and non-clinician annotators believing this would provide us with the most optimal reference standard. Annotation tasks were accomplished using an annotation schema created using Knowtator version 1.9 beta2 [22] an annotation plug-in for Protégé 3.3.1 [23]. An initial annotator training of 1-2.5 hours was provided individually to each annotator on the annotation task, the guidelines, and use of the annotation tool. Over the course of the annotation project, feedback was provided to annotators using a variety of methods. Similar to the approach used for the 2009 i2b2 community annotation task [20] annotators were encouraged to actively participate in an online Google groups discussion board. Annotators could post questions and receive answers from the challenge organizing team or their peers regarding annotation guidelines, use of the Knowtator tool, or clarification on particularly difficult annotation instances.

Interventions and Modifications to Annotation Workflow. For the 2010 i2b2/VA challenge we integrated different intervention types with annotation tasks to assess the effects of a modified workflow that used adjudication steps that went beyond strict adjudication, additional review levels for each challenge annotation task, and pre-annotation approaches applied to raw clinical texts conditioning on information scope and context. All document sources were annotated in the order in which they were received from source institutions.

Semi-structured Interviews. All annotators (n=12) participated in post completion semi-structured interviews. The challenge co-leads (DuVall, and Uzuner) also participated in semi-structured interviews that were tailored to address issues related to administration of the challenge and development of the reference standard. The semi-structured interview included open-ended questions about individual annotator perceptions of the intervention to which they were assigned, the approaches used, the annotation tasks, and tools (see Appendix A). The interviews were tape-recorded and transcribed.

Qualitative Analysis Process. The qualitative analysis was iterative, involving multiple hours of discussion. We used the ATLAS ti [24] software and the following steps:

1) ***Pre-coding step:*** The first step in the review process involved pre-coding relevant text and information from the transcript that reflected the experiences of the annotators. The pre-coding step was done initially together by all members of the research team. Once all reviewers agreed upon the type of information and a sense of understanding of the task was achieved, the remainder of the documents were reviewed independently.
2) ***Category construction:*** The output from pre-coding steps usually consisted of a large number of identified quotations and memos from all members of the team. For this second process, the research team reviewed the quotations associated with the constructs and begin to identify categories through discussion and comparison.
3) ***Completion of categorical assignments:*** Significant team discussion occurs at this stage to aggregate memos and quotes that were associated with each reviewed category. Some items were moved and new categories may have also been created. Significant discussion was involved at this stage.
4) ***Thematic identification:*** In the final stages of review categorical contents were again reviewed and aggregated. Thematic categories were named and reviewed for consistency using a group process that involved re-reviewing the content of the interviews and discussing each identified theme looking for new material.

RESULTS

From the semi-structured interviews, a total number of 452 quotes were generated by three independent reviewers on transcripts from the semi-structured interviews. These pre-coded quotes and concepts were consolidated and combined into 62 constructs. Final consolidation of pre-codes produced 6 overarching themes with 60 sub themes. These themes are harmonious with those suggested by Campbell [25]. We expanded on these relating to issues associated with complex cognitive tasks and human factors. The overarching themes identified are described below and representative quotes provided in Table 1.

Theme 1: Managing the effort between efficacy and accuracy. Annotators have the dual goals of maximizing accuracy and enhancing efficiency or speed. The result is that manual reviewers are constantly balancing these goals in a variety of ways, with more or less success. The ways in which annotators manage this tension is idiosyncratic and results in process variations.

Theme 2: The power of motivational and social forces. Interviewees often referred to strategies used for motivation, to minimize uncertainty and to seek social support. Any opportunity to talk to other annotators was appreciated and sought after, using message boards or informal conversations. The amount and intensity of interactions varied across annotators, resulting in process variations, differences in knowledge and levels of frustration.

Theme 3: The inherent difficulty of managing uncertainty. Annotation tasks deal with natural language and it is therefore impossible for guidelines to cover every possibility. As a result, rules are often induced as part of an annotation task as new instances are encountered where guidelines don't fit or ambiguity remains in inclusion or exclusions for each information class. Changing rules increases cognitive load as annotators seek information to reconcile inconsistencies and to assuage their anxiety.

Theme 4: Document readability and its affects on the annotation process. The structure of the notes themselves varied and contributed to the difficulty of the annotation process. There may have also been issues related to de-identification approaches used. Annotators noted that the interpretability of the text varied as a function of using either resynthesized identifiers, or tokens representing PHI categories inserted into the texts. Document readability was also higher for those annotators reading the full document versus those reading sentences.

Theme 5: The complexity of the annotation project. The 2010 i2b2/VA Challenge annotation project was significantly more complicated than prior years and involved a high level of cognitive workload. Organizing the documents, the annotation process, and annotation outputs was time-consuming and often frustrating. The amount of instruction and communication required varied in significant ways throughout the process.

Theme 6: The effects of interventions integrated with annotation workflow. Integrating interventions with an annotation task is an ambitious goal. In general, annotators preferred to annotate on full documents. Some annotators were also favorable to pre-annotation and viewed it as a time saving tool, or a way to reduce their workload. Other annotators had a more negative impression since they perceived pre-annotation increased their cognitive workload. The effects of interventions are folded into discussion of the first five themes.

Table 1. Constructs and Themes identified from Semi-structured Interviews

Sub theme	Example transcript quotes
Theme 1: Managing the effort between efficacy and accuracy	
a) The inherent tension between task bias, accuracy and annotator uncertainty	**Quote: "**I second guessed myself a lot more because when something was marked and I didn't normally mark that, then I was like oh, maybe I should be marking it like that and so it made me just really hesitant to unmark things that I originally had done…"
b) Differences between the annotation on raw documents compared with review tasks	**Quote: "**I pretty much did it all at the same time, trying to do the problems and the tests and the treatments and then doing assertions at the same time…"
c) Use of linguistic rules or syntax to guide annotation	**Quote:** "I would say the linguistic rules were the worst for myself. So like it was hard to look at. The verbs versus the adjectives and nouns, verbal phrases."
d) Induction of rules across the annotation task	**Quote: "**I think probably the biggest problem that I can think of is just some of the rules changing and so it was hard to keep track when you were marking one way for awhile and then having to change that or maybe not being aware of a change that happened and so we're marking one way for awhile and then having to switch that train of thought…"
e) Rapid changes and learning effects	**Quote:** "No I felt like I was still learning, even at the end, I mean, because I don't think there was ever any, I mean everything was black and white. Like sometimes I would still come across things and I'm not quite sure and so then you sort of ask the group and then their experience helped me..."
2. The power of motivational and social forces	
a) Controlling for behavior and motivational factors	**Quote:** "I really liked having the group being able to go to and read because there's a spreadsheet also that we had written down some of the problems and made the list of things, like a reference list. I thought that was very helpful. I thought having the group was very helpful and being able to reference both of those things helped me toward the end become more efficient in it."
b) Strategies for Motivation	**Quote:** "There were some you know, tricky situations or tricky little things that we didn't know, but I was able to either go online and look up the section where to find my answer or contact one of the people in charge to find the answer or one of the other annotators. And that was really helpful, to help in the places where I got stuck."
c) Social support mechanisms	**Quote:** "But in a discussion like that it was a lot more prone to stick in my brain than just reading the guidelines, because the guideline is very technical and it's hard for technical things to stick in my brain, so when it was a dialogue in an open communication then I would remember what we talked about and what we decided, and I used the information from that a lot."
3. The inherent difficulty of managing uncertainty	
a) Issues of independence	**Quote:** "….it's hard when you hear from another annotator, I still didn't feel sure until I heard from the committee or something."
b) Guidelines and resolution ambiguity	**Quote:** But the guidelines, at first I looked at them quite a bit, but as I got used to it and I was more familiar with the guidelines obviously as time went on and so by the end I hardly ever referred back to the guidelines, especially because I knew that a lot of things in there had changed and so I tried not to look at that

	quite as often because it's hard to remember which ones had changed."
4. Document readability and its effects on the annotation process	
a) Medical sublanguage and its effects on review annotation tasks	**Quote:** "So yeah some of them did get a little long and some of them were a little bit difficult because of the abbreviations and because of punctuation and things like that."
5) The complexity of the annotation process	
a) Administrative processes	**Quote:** "In terms of preparing batches, in terms of just coordinating those types of things and answering questions to the annotators and then also at the same time fielding questions and trying to bring people on as competitors, it was a significant amount of my time."
6. The effects of interventions integrated with annotation workflow	
a) Effects of pre-annotation methods	**Quote:** "I thought it was a lot worse. Like I said, it made me second guess myself and so I was confused a lot more and it just took a lot more time because I had to delete a lot of things that shouldn't have been there, so I thought it was a lot more tiring and a lot more time consuming." **Quote:** "…instead of just handing me a blank piece of paper and saying draw...draw a picture, it was like handing me a dot-to-dot and telling me to connect the dots…For some reason it just seemed to me that I felt like I learned better with the pre-annotation because if there was something highlighted, I had to find out whether it was wrong or not..."
b) Effects of document vs. sentence level review	**Quote:** "I really liked the whole document better because first of all, I could see like how much they had left, like I felt like it was easier to like stay motivated to get through the document, you know?" **Quote:** "…it kind of made you feel like you were accomplishing more work because, you know, you went through 500 sentences instead of just five documents…"

DISCUSSION

A valid reference standard directly impacts use of NLP for information extraction and classification. However, because human annotators create the reference standard, there are potentially many factors that may impact task reliability and reference standard validity. It is essential that validation of these methods be based on sound empirical grounds. We identified five generalizable themes and a sixth theme specifically related to interventions that are particularly important for future research and other large-scale clinical corpus annotation projects. The effects of interventions are folded into discussion of each of the first five themes where appropriate.

Theme 1: Managing the effort between efficacy and accuracy. There is often tension between the twin goals of efficiency and accuracy. From the interviewee's point of view, any tools that could help them gauge the degree to which they were meeting these goals would support effective allocation of attention [12, 13, 26]. Much of the idiosyncratic behaviors reported were around the active attempts of annotators to improve speed and accuracy. The interplay between these two interdependent goals was the most visible when comparing responses to the interventions. For example, the basic process of annotating on raw documents is very different from the process of arbitrating discrepancies during adjudication or subsequent review levels. In both situations, the majority of annotators reported double checking annotations on each document before moving on to the next document. Most annotators preferred full document annotation to annotation on sentences. When pre-annotation was used annotators expected the pre-annotation to improve across the annotation task in a way that would reduce modifying existing or deleting spurious annotations. Documents were annotated in the order they were obtained from source institutions, and therefore, improvements in pre-annotation may have been diminished or were not as obvious to annotators, thereby creating frustration regarding the accuracy/speed trade-off.

Recommendation: Future work in this area should focus on designing tools that annotators can use to monitor themselves as well as to improve training in a way that facilitates real-time performance feedback. Pre-annotation may also increase annotator workload, especially in situations where a high number of false positive annotations are generated. Further investigation is needed to determine the effects of pre-annotation on annotation task reliability and reference standard validity.

Theme 2: The power of motivational and social forces. The second theme focused on the importance of motivational and social factors and was a recurring topic through out the semi-structured interviews. Maintaining annotator interest and commitment is important for all annotation projects. Low motivation impacts attention and the accuracy of performance significantly [27] and monetary rewards are not likely sufficient. Factors known to enhance interest include feedback [28, 29] enhanced control over the process [29] and social interaction [30]. A key activity appeared to be creating common "task-relevant views" of the problem space that has been noted by others [31]. The use of a Google groups discussion board provided one from of social interaction that annotators mentioned helped motivation. All annotators commented that getting an answer to help resolve ambiguity and identify the correct annotation was a powerful motivator to continue. When answers were not immediately provided annotators would seek some type of social support mechanism to help resolve ambiguity. This type of interaction should help increase task consistency [15]. Social interactions helped manage uncertainty by reinforcing rules and inclusions or exclusions specified in the guidelines. However, uncontrolled interactions run the significant risk of non-independence and bias [32]. Sometimes the discussion evolved into a group decision-making process where the group consensus could be biased by a stronger personality or someone perceived as an expert due to more clinical training. These discussions may have influenced the annotators' individual decision making and at times actually caused confusion. This may result in higher reliability that is neither generalizable nor replicable (lower validity).

Recommendation: A social component should be incorporated during the initial learning process so that the annotation group can get a sense of "co-presence" and then to offer controlled access later in the project. A sense of "co-presence" is known to improve learning in online environments and similar interventions could facilitate this for annotation. Annotation is an activity that requires high cognitive load, attention to detail and may benefit from a social support environment that still preserves annotation task independence. Bias could be minimized by incorporating some of the principles of nominal group judgment, such as in the Delphi process [33].

Theme 3: The inherent difficulty of managing uncertainty. A third theme involved the problem of managing uncertainty, both in terms of task itself and because of the intense learning curve required to develop expertise. Interviewees commented that they never stopped learning across the annotation task. As a result, there may be some degree of bias in that the documents annotated at the beginning are different in some unknown metric than those annotated towards the end. The expectation is often that once guidelines are internalized, and given enough examples of inclusions and exclusions, learning effects should be diminished. However, all annotators commented that the guidelines, as well as the induction of rules across the annotation task, might have added additional complexity and cognitive load. Researchers may not appreciate the intensity of the learning curve and the high degree of continuous uncertainty. The most common questions asked about guidelines were often about linguistic forms used, how much to include in an annotation span, disambiguating between problems, treatments, and tests and assigning attributes for each. These learning effects were observed when a new document source was introduced. Some of the interventions differentially impacted the degree of uncertainty. Not surprisingly, most clinician and non-clinician annotators preferred reviewing full documents instead of sentences. Also, correctly classifying assertions or relations between information classes requires informational context to understand the entire clinical document [26]. Whereas simply identifying concepts may not require context at the level of the full document [34]. It is important to include and make available the complete context to prevent attribute assignment from occurring in a vacuum. When pre-annotation is used, it is possible that annotators with less domain knowledge or experience may be more influenced by pre-annotation than other experienced annotators [7].

Recommendation: Substantially enhanced decision-support, pattern recognition and self-learning tools may be a necessary component for future work. More systemized training may also support greater generalizability.

Theme 4: Document readability and its effects on the annotation process. Depending on the type of document reviewed there may be issues related to readability of documents. Use of abbreviations, acronyms, ambiguous section headings, template document structures, or other imbedded elements may introduce ambiguity in the documents for reviewers when they are not familiar with the source document format. Another issue that can affect document readability involves the de-identification process itself. It is unknown what effects different de-identification approaches have on document readability.

Recommendation: Improving methods of maintaining the context of redacted information should be explored for those situations where PHI will be redacted and realistic surrogates introduced. In situations where surrogates or tokens are introduced it may be more difficult for annotators to track these across the document. Use of both realistic

surrogates and tokens may increase cognitive load on the part of annotators who may inadvertently believe documents were not fully de-identified and feel ethically compelled to report potentially missed PHI.

Theme 5: The complexity of the annotation project. Annotation done in the context of a shared-task community challenge presents particular challenges related to processes and administration. These challenges include, the amount of time and workload involved, communication issues, annotator training, and management of document batches for annotation. Important questions to ask include the amount of time and resources required and the manpower needed, the type and amount of annotator training required to achieve some threshold of acceptable accuracy for a community challenge task, how partnerships will be conducted across an organizing team, and what types of administrative issues may pose particular hurdles to carrying out these tasks. These include obtaining permissions from document source institutions, regulatory issues related to data use agreements, and data stewardship.

Recommendation: A tradeoff often occurs between balancing available resources with the quality of the reference standard measured in terms of task reliability and validity. Evaluation of annotator perceptions of these tasks provides an important perspective that is often overlooked in clinical NLP studies. In general when building reference standards the question should ultimately be; "who will be the end user of the generated data"? This is an important question where development efforts for one particular use case could be repurposed or modularized as part of a larger NLP system possibly implemented as one component of an EMR. This end goal should be kept in mind when using data generated to support development of NLP systems. Scientific generalizability is enhanced with adequate reporting of methods used for annotator training, and publishing of annotation guidelines and schema.

CONCLUSION

Our qualitative analysis focuses on the human task of annotating clinical documents recognizing that there are many different methods that may improve task efficiencies and reference standard validity that can be incorporated with annotation tasks. Developing methods for reference standard generation that are reliable and at the same time highly accurate is an active area of research in the clinical NLP domain. The qualitative analyses we present in this paper inform annotation efforts that may use methods to potentially reduce annotation workload. These results also inform research efforts focusing on development of reference standards that support NLP systems development for information extraction and classification in the context of large-scale clinical corpus annotation.

ACKNOWLEDGEMENTS

Prior to conducting this study appropriate IRB approvals were received. This study was supported using resources and facilities at the VA Salt Lake City Health Care System, the VA Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374, and the NIH Roadmap for Medical Research, Grant U54LM008748. We also wish to acknowledge the efforts of human annotators responsible for generating the reference standard and qualitative analyses discussed in this paper.

REFERENCES

1. Patton, M.Q., *Qualitative research and evaluation methods*. 2002: Sage Publications.
2. Grishman, R., Sundheim, B., *Message Understanding Conference-6: A Brief History.* 16th Conference on Computational Linguistics (COLING), 1996: p. 466-71.
3. Hersh, W., Bhupatiraju, R.T., and Corley, S., *Enhancing access to the Bibliome: the TREC Genomics Track.* Stud Health Technol Inform, 2004. 107(Pt 2): p. 773-7.
4. Sparck Jones K., *Reflections on TREC* Information Processing Management. , 1995. 31(3): p. 291-314.
5. Roberts, A., et al., *The CLEF corpus: semantic annotation of clinical text.* AMIA Annu Symp Proc, 2007: p. 625-9.
6. Roberts, A., et al., *Building a semantically annotated corpus of clinical texts.* J Biomed Inform, 2009. 42(5): p. 950-66.
7. Estes, W., *Classification and Cognition.* 1996. New York: Oxford University Press.
8. Koehler, D., Brenner, L.S., and Griffin, D., *The calibration of expert judgment : Heuristics and biases beyond the laboratory*, in *Heuristics and Biases: The Psychology of Intuitive Judgment.*, T. Gilovich, D. Griffin, and K. D, Editors. 2002, Cambridge University Press: New York.

9. Underwood, G., *Implicit Cognition*. 1996, Oxford: Oxford Science Publications.
10. Heckhausen, J. and Heckhausen, H., *Motivation and Action.* Cambridge University Press, Cambridge, UK., 2008
11. Wyer, R.S., *Social Comprehension and Judgment: The Role of Situation Models, Narratives, and Implicit Theories*. 2004, Mahwah, NJ: Erlbaum.
12. Kruglanski, A., *Lay epistemics and human knowledge: Cognitive and motivational bases.* . New York: Plenum., 1989.
13. Hollnagel, E., *The ETTO Principle: Efficiency-Thoroughness Trade-Off: Why things that go right sometimes go wrong.* . Cornwall, Britian: Ashgate., 2009.
14. Weir, C.R., et al., *A cognitive task analysis of information management strategies in a computerized provider order entry environment.* J Am Med Inform Assoc, 2007. 14(1): p. 65-75.
15. Harton, H.C., Green, L.R., Jackson, C, Latané, B., *Demonstrating dynamic social impact: consolidation, clustering, correlation, and (sometimes) the correct answer.* . Teaching of Psychology, , 1998. 25: p. 31-34.
16. Uzuner, O., *Recognizing obesity and comorbidities in sparse data.* J Am Med Inform Assoc, 2009. 16(4): p. 561-70.
17. Uzuner, O., et al., *Identifying patient smoking status from medical discharge records.* J Am Med Inform Assoc, 2008. 15(1): p. 14-24.
18. Uzuner, O., Y. Luo, and P. Szolovits, *Evaluating the state-of-the-art in automatic de-identification.* J Am Med Inform Assoc, 2007. 14(5): p. 550-63.
19. Uzuner, O., I. Solti, and E. Cadag, *Extracting medication information from clinical text.* J Am Med Inform Assoc, 2010. 17(5): p. 514-8.
20. Uzuner, O., et al., *Community annotation experiment for ground truth generation for the i2b2 medication challenge.* J Am Med Inform Assoc, 2010. 17(5): p. 519-23.
21. Uzuner, O., South, BR, DuVall S, *2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text.* 2010.
22. Ogren, P., *Knowtator a protege plug-in for annotated corpus construction.* Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2006: p. 273-5.
23. Musen, M.A., et al., *PROTEGE-II: computer support for development of intelligent systems from libraries of components.* Medinfo, 1995. 8 Pt 1: p. 766-70.
24. *ATLAS, ti v5.7.1, GmbH, Berlin Germany.* 1993-2011.
25. Campbell, E.M., et al., *Understanding Inter-rater Disagreement: A Mixed Methods Approach.* AMIA Annu Symp Proc. 2010: p. 81-5.
26. MacCoon, D., Wallace, J.F. and Newman, J.P., *Self-regulation: Context-appropriate balanced attention.* . In R. Baumeister and K Vohs (Eds) Handbook of Self-Regulation: Research, theory and applications. New York: , 2004 (Guilford Press.): p. 422-446.
27. Zimmerman, B.J. and Kitsantas A., *The hidden dimension fo personal competence: Self-regulated learning and practice.*, in *Handbook of Competence and Motivation.*, A.J. Elliott and C.S. Dweck, Editors. 2006, The Guilford Press: New York.
28. Sansone, C., Sachau, D.A., and Weir, C.R., *Effects of instruction on intrinsic interest. The importance of context.* Journal of Personality and Social Psych, 1989.
29. Gollwitzer, P.M., K. Fujita, and G. Oettingen, *Planning and the Implementation of Goals.* , in *Handbook fo Self-Regulation.*, R.F. Baumeister and K.D. Vohs, Editors. 2004, Guliford Press: New York.
30. Baumeister, R.F. and Leary, M.R., *The need to belong: Desire for interpersonal attachments as a fundamental human motivation.* J Pers Soc Psychol, 2000. 31(2): p. 43-64.
31. Middleton, D., *Talking work: Argument, common knowledge, and improvisation in teamwork. In Engestrom, Y and David Middleton (Eds) Cognition and Communication at Work.* . Cambridge University Press: Cambridge, UK, 1998: p. 233-256.
32. Baron, R., *So right it's wrong: Groupthink and the ubiquitous nature of polarized group decision making.* . Advances in Experimental Social Psychology, 2005. 37: p. 35.
33. Linstone, H.A. and Turoff, M., *The Delphi Method: Techniques and Applications*. 1975, Reading, Mass.: Adison-Wesey.
34. Hirschman, L., Yeh, A, Blaschke, C, Valencia, A, *Overview of BioCreAtIvE: Critical Assessment of Information Extraction for Biology* BMC Bioinformatics, 2005. 6(S1).

CHAPTER 4

AIM 2: EVALUATING THE EFFECTS OF NONINTERACTIVE PRE-ANNOTATION OF CLINICAL NAMED ENTITIES AND ITS IMPACT ON ANNOTATION EFFICIENCY AND ANNOTATOR PERFORMANCE

Brett R. South, MS[1-3], Scott L. DuVall, PhD[1,2],

Shuying Shen, M.Stat[1-3], Ying Suo, MS[1,2], Ozlem Uzuner, PhD[4]

[1] VA Salt Lake City Health Care System, Salt Lake City, Utah, USA

[2] Department of Internal Medicine, University of Utah, Salt Lake City, Utah, USA

[3] Department of Biomedical Informatics, University of Utah, Salt Lake City, Utah, USA

[4] University at Albany, SUNY, Albany, New York, USA

(To be submitted)

4.1 Abstract

The shared-task challenges organized by Informatics for Integrating Biology and the Bedside (i2b2) have provided a valuable source of reference standard clinical corpora. In a collaborative effort under the 2010 i2b2/VA challenge, i2b2 and the Department of Veterans Affairs (VA) Consortium for Healthcare Informatics Research (CHIR) generated the reference standard for three tasks: concept extraction; assertion classification; and relation classification.

A typical workflow for creating a reference standard involves two reviewers who independently annotate the same clinical texts. Annotations that both reviewers label are added to the reference standard with disagreements resolved by a third independent reviewer. We experimented with pre-annotating clinical entities for problems, treatments, and tests to estimate the effects of noninteractive pre-annotation on clinical texts. We used double annotation, and a modified annotation workflow incorporating three review levels. The first review allowed adjudicators to modify the primary annotations instead of just breaking ties. The second review involved a re-examination of adjudicated annotations for accuracy. The final review was machine-assisted to check for inconsistencies. We also estimate learning effects as they relate to predicting the time required to train annotators for review tasks that include identifying and marking mentions of text representing clinical concepts, and classifying the assertional status and relations between them.

We measured the reliability of annotation tasks using interannotator agreement and calculate standard performance metrics using recall, precision and F_1-measures. At the level of primary annotation, interannotator agreement ranged from 0.89-0.92 for

concepts, 0.46-0.94 for assertions, and 0.59-0.96 for relations. Significant improvements in recall and precision were obtained as additional review layers were added. For concepts, F_1-measures increased from 0.93 at primary annotation to 1.0 at secondary review. Significant improvements were observed with additional review layers for assertions F_1-measures ranged from 0.87 primary annotation to 0.97 at secondary review, and from 0.68 primary annotation to 0.98 at secondary review on relations.

Finally, we present results of an experiment integrated with these manual annotation efforts used to evaluate the effects of a noninteractive machine generated outputs used as pre-annotations for annotation of medical problems, treatments and tests. We evaluate the impact of this approach highlighting experimental results for each intervention type. LOESS regression was used as a nonparametric technique to assess annotation task difficulty and visualize annotator training plateaus across document batches using F_1-measures for annotation of concepts, and assertions and relations classification. F_1-measure for concept annotation of medical problems, treatments and tests improved for the first 10 batches (from 0.82 to 0.87) and then plateaued. F_1-measure for assertion classification showed a similar pattern as that of concept annotations although it took longer before reaching its first plateau at batch 15 (from 0.73 to 0.83). F_1-measure for relations increased very slowly across the early half of the annotation task but decreased towards the end of the task with no obvious plateau. Across the 339 primary annotation batches the number of annotations ranged from 65-2,598 per document batch. The time per annotation decreased sharply from 25 seconds on batch 1 to 12 seconds on batch 10 before leveling off.

4.2 Introduction

Natural language processing (NLP) systems convert narrative texts into structured representations that capture part of their meaning [1]. This conversion is more accurate when a manually generated reference standard that supports both system development and evaluation is used. Machine systems developed can only be as good as the reference standard that supports them, so an adequate reference standard must meet some level of reliability and validity. Reliability reflects annotator agreement, and validity measures the accuracy in the reference standard's representation of the task.

A typical approach to reference standard generation follows a workflow process involving three trained reviewers: two reviewers independently annotate the same document and a third reviewer arbitrates any disagreements between the first two. This process involves considerable effort, but methods to minimize workload may compromise annotation reliability and reference standard validity. Manual annotation often requires a tradeoff between annotator reliability, reference standard validity, and human workload.

Within the context of the 2010 i2b2/VA challenge, Informatics for Integrating Biology and the Bedside (i2b2) and the Department of Veterans Affairs (VA) Consortium for Healthcare Informatics Research (CHIR) undertook a large-scale annotation effort on concept extraction, assertion classification, and relation classification. Concept extraction focused on medical problems, tests, and treatments that were to be extracted from patient reports. Assertion classification required the interpretation of whether a medical problem was present, absent, uncertain, conditionally present, or hypothetically present in the patient at some future time, or mentioned in the patient report, but present in someone

other than the patient. Relation classification determined the nature of the clinician-stated interaction of the medical problems, tests, and treatments with each other [2]. We refer to concept extraction, and classification of assertions and relations as the 2010 i2b2/VA *challenge tasks*.

For the 2010 i2b2/VA challenge tasks, we experimented with a simple noninteractive pre-annotation approach along with a modified annotation workflow that incorporated three levels of review. This experiment conditioned on full document versus sentence level annotation, and use of the pre-annotation techniques discussed above. We were interested in assessing the effects of each experiment type on primary annotation.

All annotation tasks used an open-source annotation tool called Knowtator [30]. The first review level allowed adjudicators to go beyond arbitration of the primary annotations by adding, deleting, or modifying annotations. The second review level re-examined the annotations from each challenge task, inspecting each adjudicated annotation separately. The reference standard was finalized after completing a third and final machine-assisted review level. We report metrics of annotator workload, reliability, and validity estimates at each of these review levels demonstrating improvements that were achieved as a result of the modified annotation workflow.

4.3 Background

Since 2006, the i2b2 shared-task challenges have provided reference standards for the medical informatics community for finding personally identifiable information [3], for classifying smoking status of patients [4], for determining obesity comorbidities exhibited in patients [5], and for extracting medication mentions [6]. With the exception

of extraction of medication mentions, the past i2b2 reference standards were created by three domain experts who provided annotations for the task. These reference standards were created by two domain experts who discussed annotations as they made them, disagreements between reviewers were arbitrated by a third domain expert whose sole responsibility was to break ties. For the medication extraction challenge, the i2b2 team ran a community annotation experiment in which they asked each participating team to annotate documents. Given the lack of domain expertise of the challenge participants, and the fact that some of the participants were not native English speakers, the i2b2 team asked two independent challenge teams to annotate each record. They followed these annotations with arbitration from a third challenge team to break ties. Finally, disagreements and suggested tie-breaks were reviewed by the organizing team before finalizing the reference standard. They found this reference standard to be comparable in reliability and validity to an expert-generated counterpart.

4.3.1 Paradigms and Approaches used for Clinical Corpus Annotation

Previous large scale annotation efforts, including those by Message Understanding Conference (MUC) [7], Text Retrieval Conference (TREC) [8, 9], Clinical E-Science Framework (CLEF) [10, 11], GENIA [12, 13], and Penn Treebank [14], have established procedures for reference standard generation that are often implemented. The workflow followed to generate the reference standard for CLEF was the first of its kind focusing on semantic annotation of clinical texts. It included an annotation workflow describing particular steps for annotators to follow and also used double annotation with a strict adjudication step, allowing tie-breaks only. Hripcsak [15-17] proposes various experiments that modify the type of information provided to

reviewers, the roles of expert reviewers in reference standard creation, and use of various metrics for evaluation of reference standard reliability and validity. Work by Chapman [18] suggests that once trained, lay annotators can perform as well as clinician annotators on specific tasks that use linguistic form and identification of modifiers. Mayer [19] and South [20] evaluate the effects of annotator training and induction on rules and guidelines supporting annotation tasks. Their work discusses the related difficulties of designing annotation schema, including determining the appropriate level of information to be annotated. Work by Roberts et al. [10, 11], Savova [21, 22], Chapman [18, 23, 24] Uzuner [25], and previous i2b2 challenges [3-5, 25, 26] provides the context and motivation for the 2010 i2b2/VA challenge, for development of annotation guidelines and schema, and for development of the resulting reference standard.

We compare and contrast IAA used to assess annotation task reliability and consistency with guidelines and performance metrics (i.e., recall, precision and F_1-measure) used to assess annotator accuracy with the generated reference standard. Together, these metrics provide a measurement of the quality of generated data. The effects of pre-annotation have been investigated in many studies that include annotation of medical literature [20], POS tagging [19], Named Entity Recognition (NER) [22] and clinical named entities [23, 24], as well as common PHI types [25]. Other studies have employed semi-automated annotation methods that produce machine-generated candidate spans presented in such a way that the human reviewer must either modify incorrect annotations, delete spurious annotations, or add missed annotations [26, 27, 28]. There are several issues with providing pre-annotations to human reviewers. First, a pre-annotation tool that is not good enough may slow human annotators down and reduce

consistency. Second, a pre-annotation system that is too precise may cause the annotator to lose concentration, trusting the suggestions too much, resulting in a biased reference standard.

For the 2010 i2b2/VA challenge tasks, we were interested in assessing the effects of a noninteractive pre-annotation approach along with addition of various review levels on reference standard generation. Our approach to these tasks also allows measurement of the tradeoffs between reliability, validity and annotator workload. We accept that there may be more efficient approaches that can be used for reference standard generation. We believe our methods are generalizable for other human review tasks used to generate a reference standard and provide data for NLP systems training and evaluation.

4.3.2 Data

The documents used for the 2010 i2b2/VA challenge included 230 de-identified discharge summaries from Partners Healthcare, 196 de-identified discharge summaries from Beth Israel Deaconess Medical Center [27], and 200 de-identified progress notes and 200 de-identified discharge summaries from the University of Pittsburgh Medical Center (UPMC) [28]. The data were released to challenge participants under data use agreements [26]; 349 reports were used for training and 477 were reserved for testing. There were 826 documents in total that were distributed to 12 part-time annotators representing data from 339 raw annotation batches having a minimum of two reviewers. Annotator performance metrics reported in this paper are based on a set of 729 from the original 826 documents (136 document batches) that were subjected to the modified workflow.

4.4 Methods

The workflow used to generate the reference standard used double annotation at a primary annotation level followed by three different levels of review: the first level allowed modification to primary annotations during a modified adjudication step, the second level re-examined the adjudicated annotations, and the third was machine-assisted for the identification of inconsistencies in the annotations. This workflow was followed across document batches, and by all annotators trained on annotation guidelines, tools, and schema.

4.4.1 Annotation Guidelines, Schema and Tools

Annotation guidelines 2010 i2b2/VA challenge were written by the two coleads of the challenge (Uzuner and DuVall), and then opened for comments by the challenge organizing committee. We created the annotation schema using Knowtator [29] version 1.9 beta 2 [30], an open-source plug-in tool for the Protégé [31] knowledge management system.

The annotation schema included classes representing problems, tests, and treatments. Class attributes, referred to as "slots," were created to represent assertions and relations. Because the 2010 i2b2/VA challenge tasks involved both information extraction (IE) and classification, annotation was done in steps. The annotators: 1) selected the span of concepts; 2) classified the concept as a problem, test, or treatment; 3) assigned assertion values to medical problems; 4) linked concepts using the slot representing relations of problems and problems, problems and treatments, or problems and tests; and 5) assigned the appropriate relation to pairs of concepts (see Figure 4.1). Knowtator provides a method to build consensus sets based on annotation projects from

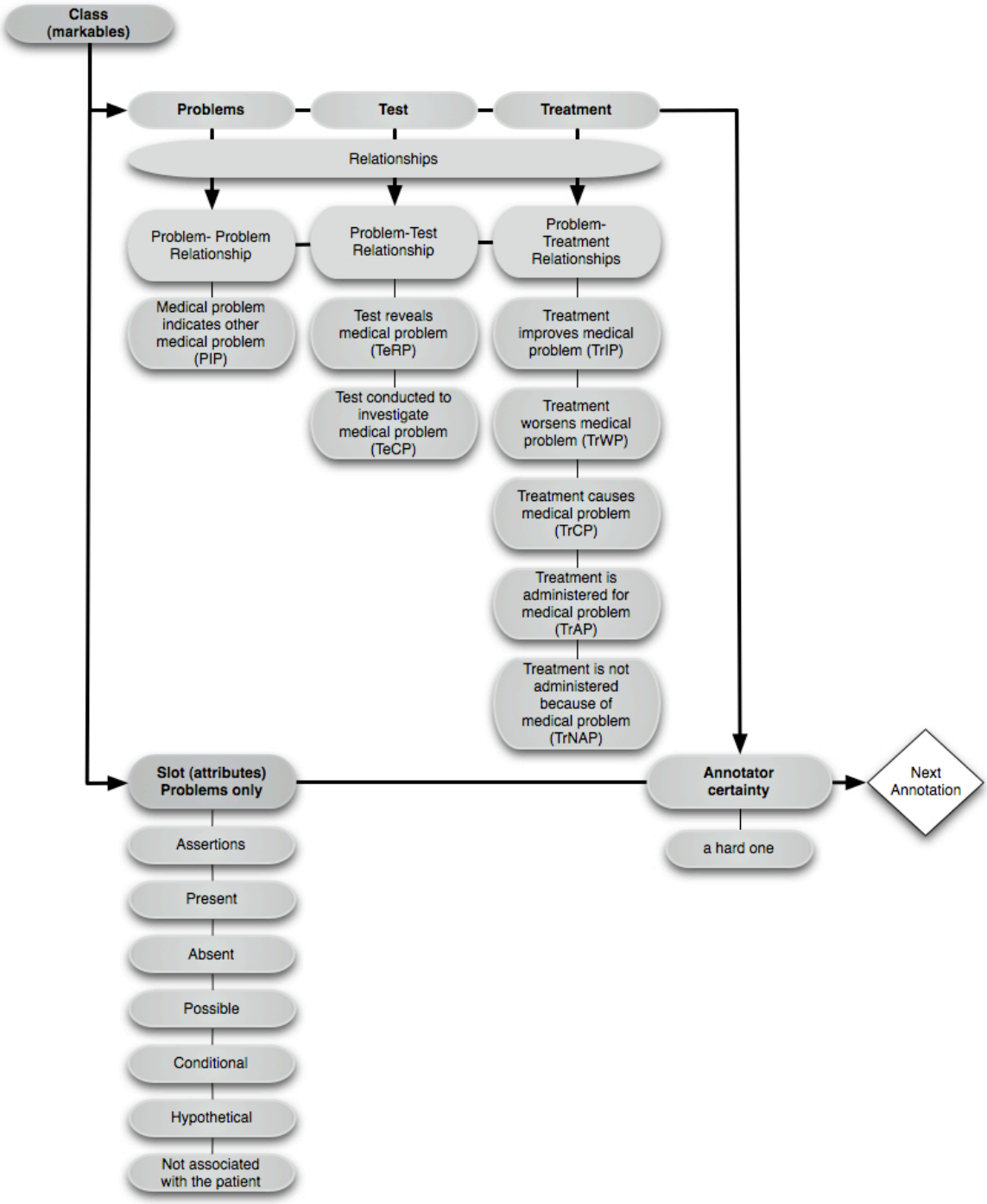

Figure 4.1: Annotation Schema 2010 i2b2/VA Challenge

multiple annotators by consolidating matching annotations between sets. The remaining conflicting annotations can then be adjudicated. Annotations on each document can be exported from the Knowtator tool as stand-off xml files.

4.4.2 Annotator Training and Methods of Feedback

Creating the reference standard for the 2010 i2b2/VA challenge required the combined efforts of 12 part-time annotators over six months. We recruited six clinician and three non-clinician annotators. We assumed no previous experience on the part of annotators with the tools or challenge tasks. Furthermore, we made no assumption about annotator experience with the use of linguistic rules that were incorporated with challenge annotation tasks. All annotators were provided with an initial one-on-one training session of 1-2.5 hours on the annotation guidelines and tool. Similar to the approach used for the 2009 i2b2 community annotation experiment, reviewers used an online discussion group to post their questions and share answers about annotation guidelines, specific annotation instances, and the annotation tool. The 2010 i2b2/VA organizing team provided answers to posted questions and provided annotator feedback throughout the annotation process.

4.4.3 Modifed Annotation Workflow

Initial testing of guidelines and schema revealed the complexity of the challenge annotation task. In the interest of keeping with our first data release for training data, once we had generated a predefined number of annotated documents, we implemented an annotation workflow that went beyond the typical double annotation and adjudication.

This annotation workflow incorporated three review levels and is summarized in Figure 4.2.

Our goal for integrating an annotation experiment with a modified annotation workflow was to assess performance differences by introducing additional review levels, while balancing expectations of data release deadlines with annotator workload and available resources. For **primary annotation**, we used a double annotation approach in which two annotators label the same clinical documents. Pre-annotations at the

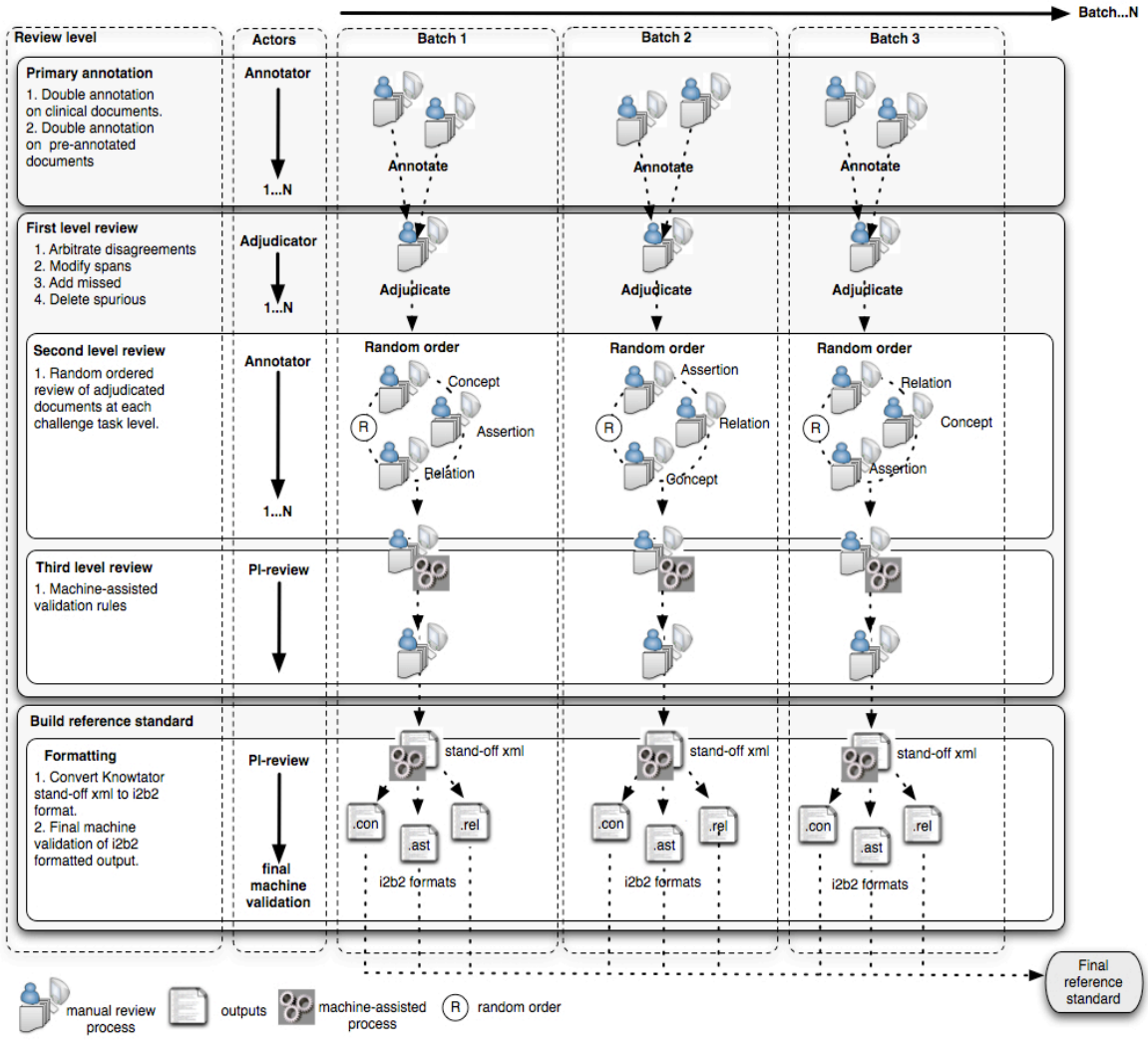

Figure 4.2: Modified Workflow for Annotation Tasks

annotation class level were generated using an Apache Lucene [32] phrase lookup based on UMLS terminologies that included SNOMED-CT, MSH, NCBI, RxNorm, ICD10PCS, ICD9CM, LNC, MTH matching on text span mapped loosely to the appropriate UMLS concept. For primary annotation we assembled 4 intervention groups conditioning on the amount of information and document context. This experiment involved a factorial design so that annotators were randomized into 4 experimental arms: 1) Full document; 2) Full document + pre-annotations; 3) Sentence; 4) Sentence + pre-annotations. For these annotation tasks we did not experiment with pre-annotation of assertions or relations.

We followed the primary annotation with a **first level review** that diverged from strict adjudication, allowing an adjudicator to not only arbitrate and resolve disagreements on primary annotations, but also to add missed annotations and correct errors by modifying or deleting existing annotations. The **second review** involved re-examining all adjudicated annotations from the first level review. During this second review, the reviewer was asked to focus on one of the following review questions: 1) Are concept level assignments and any relations correct?; 2) Is the assigned assertion correct?; or 3) Does the annotation span follow the linguistic rules provided in the guidelines?

Batches of documents were circulated in random order among annotators, each of whom checked the adjudicated annotations against their specific question. Though this second review level did not use double annotation for each review question, annotators were allowed to make changes by modifying, adding, or deleting annotations if they found an error. **The third review** further refined output from the second review with a

machine-assisted consistency check against the annotation guidelines. Our guidelines were very specific about the syntactic phrase types and spans that were to be included in the annotation; human annotators could inadvertently deviate from these, while automated approaches caught the inconsistencies [26]. All annotators participated in primary annotation, modified adjudication, and the first and the second reviews. The challenge coleads conducted the third review and finalized the reference standard.

4.4.4 Annotation Experimental Conditions

For **primary annotation**, annotators were assigned to an intervention group in a way that would pair nonclinician annotators with clinician annotators. The primary annotation experiment was conditioned on the presence or absence of contextual information, full document versus sentence level annotation, and use of the pre-annotation techniques discussed in the previous section. We intentionally modified annotation workflow to include additional review levels that extend beyond typical double anntoation with adjudication. Woven into each of these experiments were general questions around factors that are related to reference standard quality and the effects of pre-annotation and workflow modification on annotator efficiency, annotation consistency and reference standard validity. We were interested in assessing the effects of each different type of intervention at each review level of the annotation task.

4.4.5 Modifications to Knowtator

We modified the Knowtator tool to better support reference standard generation for the challenge tasks [31]. One modification allowed side-by-side comparison of annotations, allowing the adjudicators to easily resolve disagreements as part of

consensus set generation. Other modifications made assertion classification compulsory, reducing the potential of annotators misassigning an assertion where a problem was identified. Additional modifications included reducing the number of actions required to create an annotation, and allowing annotators to create a relation between one concept and another in a single mouse click.

4.4.6 Fielding of Document Batches

Each annotator was provided with pre-compiled Knowtator projects that were linked to assigned document batches. Between levels of review, annotations on completed document batches were exported to stand-off xml files. The final reference standard was distributed in pipe-delimited text format, following the format used by previous i2b2 challenges. For these annotation tasks we did not experiment with pre-annotation of assertions or relations.

4.4.7 Annotator Performance Metrics

For assessment of reliability at the primary annotation level, we estimate two variations of interannotator agreement (IAA). First, all concepts, assertions, and relations were assessed for agreement using exact span matching for concepts, in which the character span of each annotation had to be the same, and using inexact matching, where spans between the annotations could overlap. We also report strict matching for assertions and relations in which concepts must overlap with the same classification and must also match on assertion and relation. This approach has the drawback that an annotator who failed to identify a concept or who does not include assertion or relation is penalized. A second calculation is used to overcome this by calculating an adjusted IAA,

including only those assertions or relations where both annotators identified a matching concept. This approach allows us to isolate reliability estimates for assertion and relations from reliability estimates for the concept identification task. Inter-annotator agreement (IAA) is calculated using the formula:

$$IAA\ Class = \frac{2 \times Matches}{(2 \times Matches + Non\text{-}Matches)}$$

We report IAA metrics for the primary annotation only to demonstrate areas of improvement in subsequent review levels and to show baseline annotator performance. We used validity estimates, as discussed below, to assess reference standard validity at each review level. We also constructed confusion matrices to demonstrate areas where improvements in annotator consistency were achieved for each review level. Evaluation metrics were computed by constructing a 2x2 table for each individual category. Overall metrics are reported using micro-averaging across all concepts, attributes, and relation types using the formula:

$$IAA\ Micro\ Average = \frac{\sum_{i=1}^{M} 2 \times Matches_i}{\sum_{i=1}^{M} (2 \times Matches_i + Non\text{-}Matches_i)}$$

where M is the total number of classes.

We measured validity in terms of F_1-measure, computed from estimates of precision and recall. Formulae used for these calculations rely on true positives (TP), false positives (FP), and false negatives (FN):

$$Recall = \frac{TP}{(TP + FN)}$$

$$Precision = \frac{TP}{(TP + FP)}$$

$$F1\text{-}measure = \frac{2 \times Recall \times Precision}{(Recall + Precision)}$$

These metrics are the same as those used to assess systems performance of 2010 i2b2/VA challenge team submissions [26].

4.4.8 Measuring the Effects of the Primary Annotation Experiment

For the annotation experiment we assessed whether human annotators provided with pre-annotations could generate similar quality data for span classification for the three information classes defined in the challenge annotation task (i.e., annotation of medical problems, treatments, tests). Annotators received different types of intervention for different batches of medical records and once assigned to a given intervention type (where possible) each annotator remained in that intervention group.

We include a complete assessment of each intervention type used at the level of primary annotation. For this evaluation, we were interested in techniques that can be used to evaluate and visualize annotator learning patterns for the concept level challenge annotation task. For these analyses we use a statistical technique of locally estimated scatterplot smoothing (LOESS) as a nonparametric method to model and visualize annotator learning patterns using standard performance metrics (i.e., F_1-measure) for

concept, assertion, or relation and the time per annotation and its change with batch ordering. We make the assumption that the annotator "training" period begins with the first document annotated until performance plateuas (i.e., the ith batch). Once it is determined where the ith batch occurs for a given annotator, Generalized Estimating Equations (GEE) can be used to ascertain the intervention effects at the primary anntotation level. We estimate intervention effects for annotation efficiency (time) by each intervention and by anotation class for each intervention using the differences in least squares menas for exact F_1-measure and report p-values adjusted for multiple testing.

4.5 Results: Annotation Volume and Workload by Challenge Task

The total number of annotations generated by all 12 reviewers at the primary annotation level was 179,170 compared with the 72,846 annotations in the final reference standard. Annotation time (seconds/annotation) varied by document source from 15.9 for Partners data, 11.5 for Beth Israel, 10.5 for UPMC discharge summaries, and 9.0 UPMC progress notes. Based on time stamps obtained from the Knowtator tool the total time taken across all document sources for primary annotation was 593 hours. However, creating the reference standard required a total of 1,378 billed hours for completion of all annotation tasks across all review levels. Which approximates to just over 1.5 hours per document for all challenge annotation tasks. Annotation prevalence in the final reference standard across all challenge tasks and document sources can be found in Uzuner [26].

4.5.1 Annotator Agreement Estimates

After annotator training, guidelines were refined in an attempt to reduce ambiguity by including more examples of what should be marked and what should not be marked. These refinements also provided detail on how much text should be included in annotated spans, and addressed applying syntactic rules used to guide annotators [26].

Interannotator agreement (Table 4.1) metrics at the primary annotation level are adjusted to approximate metrics used for system evaluation [10]. Missing assertions or

Table 4.1: Inter-Annotator Agreement (IAA): Primary Annotation

IAA – Primary Annotation		
Concepts	**Exact**	**Inexact**
Problem	0.84	0.91
Treatment	0.85	0.92
Test	0.84	0.89
Overall	**0.85**	**0.91**
Assertions	**Strict**	**Adjusted**
Absent	0.89	0.94
Conditional	0.43	0.46
Hypothetical	0.76	0.79
Possible	0.59	0.61
Present	0.84	0.9
AWSE	0.85	0.88
Overall	**0.86**	**0.91**
Relationships	**Strict**	**Adjusted**
PIP	0.37	0.79
TeRP	0.7	0.96
TeCP	0.45	0.75
TrIP	0.45	0.62
TrWP	0.32	0.59
TrAP	0.69	0.95
TrCP	0.49	0.83
TrNAP	0.46	0.78
Overall	**0.59**	**0.94**

relationships are included in nonmatches when calculating strict IAA. For adjusted IAA, missing values are excluded and only misclassified values are considered nonmatches. Concept level exact IAA and inexact IAA were 0.85 and 0.91, respectively. Strict IAA for assertion was 0.86, and adjusted IAA was slightly higher at 0.91. The lowest adjusted IAA was 0.46 for the conditional category, and the highest IAA was 0.94 for the absent category. Relations had a strict IAA of 0.59 and adjusted IAA of 0.94. This result suggests that human annotators had difficulty in agreeing on the presence of a relation. It also suggests that once a relation is recognized, the classification is usually easier. Even in situations where these metrics were adjusted, there is a clear opportunity for improvement in reliability, further justifying additional review levels, particularly for assertions and relations classification. The wide ranges of agreement scores at primary annotation for both assertions and relations prompted the additional review levels we implemented in the modified annotation workflow and demonstrated in the confusion matrices and validity estimates we report.

4.5.2 Annotator Performance Metrics

Primary review level microaveraged inexact F_1-measures for concepts, assertions, and relationships were 0.93, 0.87, 0.68, respectively (Tables 4.2-4.4). These estimates are comparable to IAAs reported in Table 4.1. Not surprisingly, the adjudication level results had higher validity than the primary review, with F_1-measures ranging from 0.82 (relationships) to 0.97 (concepts). However, despite the overall improvement, some of the categories still had less than satisfactory accuracy. These included conditional and possible assertions, the "Test conducted to investigate Problem" (TeCP), "Treatment worsens Problem" (TrWP), and "Treatment not administered for Problem" (TrNAP)

Table 4.2: Task: Information Extraction (IE) - Problems, Treatments, Tests

Review level	Recall		Precision		F_1-Measure	
	Exact	**Inexact**	**Exact**	**Inexact**	**Exact**	**Inexact**
Primary annotation						
Problems	0.83	0.92	0.96	0.96	0.89	0.94
Treatments	0.82	0.91	0.94	0.95	0.88	0.93
Tests	0.82	0.88	0.94	0.94	0.88	0.91
Overall (Micro)	**0.83**	**0.91**	**0.95**	**0.95**	**0.88**	**0.93**
First level review						
Problems	0.92	0.97	0.98	0.98	0.95	0.97
Treatments	0.91	0.96	0.98	0.98	0.94	0.97
Tests	0.91	0.95	0.96	0.97	0.94	0.96
Overall (Micro)	**0.92**	**0.96**	**0.97**	**0.98**	**0.94**	**0.97**
Second level review						
Problems	1	1	1	1	1	1
Treatments	1	1	1	1	1	1
Tests	1	1	1	1	1	1
Overall (Micro)	**1**	**1**	**1**	**1**	**1**	**1**

** Estimates > =0.995 are approximated to 1.00*

relations. Secondary review resulted in further improvements, particularly for problematic categories that tend to have low prevalence. The F_1-measures for each category exceeded 0.89, and microaveraged F_1-measures for concepts, assertions, and relationships all exceeded 0.97. Additional areas of improvement are illustrated in the confusion matrices in Tables 4.5-4.7 and are shown as percent of annotations compared with the reference standard. Significant improvements occurred for assertional classification and relations classification. With overall improvement in less prevalent assertion types for each review level for conditional (51.4%, 70.9%, 99.6%), hypothetical (42.7%, 87.7%, 100%), and 57.3%, 74.3%, 100% for possible assertions. Other categories exhibited similar gains at each review level.

Table 4.3: Task: Information Extraction (IE) Plus Classification - Assertions

Review level	**Assertions**						
	Present	**Absent**	**Conditional**	**AWSE**	**Hypothetical**	**Possible**	**Overall (Micro)**
Primary annotation							
Recall	0.81	0.87	0.41	0.8	0.57	0.49	**0.79**
Precision	0.96	0.98	0.84	0.97	0.99	0.98	**0.96**
F_1-Measure	**0.88**	**0.92**	**0.55**	**0.87**	**0.72**	**0.65**	**0.87**
First level review							
Recall	0.87	0.94	0.54	0.87	0.79	0.64	**0.87**
Precision	0.98	0.98	0.98	0.98	0.99	0.99	**0.98**
F_1-Measure	**0.92**	**0.96**	**0.7**	**0.92**	**0.88**	**0.78**	**0.92**
Second level review							
Recall	0.93	0.99	0.96	1	0.95	0.9	**0.94**
Precision	1	1	1	1	1	1	**1**
F_1-Measure	**0.97**	**1**	**0.98**	**1**	**0.97**	**0.95**	**0.97**

***** Estimates >= 0.995 are approximated to 1.00

Table 4.4: Task: Information Extraction (IE) Plus Classification - Relations

Review level	**Relations**								
	PIP	**TeRP**	**TeCP**	**TrIP**	**TrWP**	**TrAP**	**TrCP**	**TrNAP**	**Overall (Micro)**
Primary annotation									
Recall	0.37	0.68	0.4	0.47	0.37	0.66	0.46	0.49	**0.56**
Precision	0.76	0.91	0.82	0.78	0.74	0.89	0.83	0.74	**0.86**
F_1-Measure	**0.5**	**0.78**	**0.54**	**0.59**	**0.49**	**0.76**	**0.59**	**0.59**	**0.68**
First level review									
Recall	0.67	0.84	0.58	0.71	0.56	0.81	0.63	0.63	**0.75**
Precision	0.84	0.92	0.88	0.87	0.85	0.92	0.87	0.79	**0.9**
F_1-Measure	**0.74**	**0.88**	**0.7**	**0.78**	**0.68**	**0.86**	**0.73**	**0.7**	**0.82**
Second level review									
Recall	0.99	0.98	0.89	0.91	0.91	0.98	0.92	0.95	**0.97**
Precision	0.99	0.99	0.99	0.95	0.98	0.99	0.96	0.99	**0.99**
F_1-Measure	**0.99**	**0.98**	**0.84**	**0.93**	**0.94**	**0.98**	**0.94**	**0.97**	**0.98**

* Estimates >= 0.995 are approximated to 1.00

Table 4.5: Confusion Matrices: Problems, Treatments, Tests

Review level			
Primary Annotation	**Reference standard (%)**		
	Problem	Test	Treatment
Problem	**92**	1	1
Test	1	**89**	1
Treatment	1	1	**91**
Missing	**7**	**8**	**91**
First level review			
	Problem	Test	Treatment
Problem	**98**	1	1
Test	<1	**96**	<1
Treatment	<1	1	**97**
Missing	**2**	**2**	**2**
Second level review			
	Problem	Test	Treatment
Problem	**99.9**	0	0
Test	0	**99.9**	<1
Treatment	0	<1	**99.9**
Missing	**<1**	**<1**	**<1**

Table 4.6: Confusion Matrices: Assertions

Review level						
Primary annotation	**Reference standard (%)**					
	Absent	AWSE	Conditional	Hypothetical	Possible	Present
Absent	**88.9**	4.6	1.9	1	1.9	0.4
AWSE	0	**88.5**	0	0.1	0	0
Conditional	0	0	**51.4**	0.4	0.2	0.3
Hypothetical	0	0	0.5	**42.7**	1	0.1
Possible	0.2	0	0.8	8	**57.3**	0.3
Present	3.1	3.1	36.9	38.1	31.4	**89.2**
Missing	**7.7**	**3.8**	**8.5**	**9.6**	**8.2**	**9.8**
First level review						
	Absent	AWSE	Conditional	Hypothetical	Possible	Present
Absent	**96.5**	2	1.8	0.4	2.1	0.3
AWSE	0	**90.7**	0	0.1	0	0
Conditional	0	0	**70.9**	0.8	0	0.2
Hypothetical	0	0	0.4	**87.7**	1.6	0.2
Possible	0.1	0	0.9	1.1	**74.3**	0.2
Present	1.1	6.8	23.8	8.2	20	**96.7**
Missing	**2.3**	**0.5**	**2.2**	**1.7**	**2**	**2.4**
Second level review						
	Absent	AWSE	Conditional	Hypothetical	Possible	Present
Absent	**99.9**	0	0	0	0	0
AWSE	0	**99.5**	0	0	0	0
Conditional	0	0	**99.6**	0	0	0
Hypothetical	0	0	0	**100**	0	0
Possible	0	0	0	0	**100**	0
Present	0	0	0	0	0	**99.8**
Missing	**0**	**0.5**	**0.4**	**0**	**0**	**0.1**

Table 4.7: Confusion Matrices: Information Extraction (IE) Plus Classification - Relations

Review level								
Primary annotation								
	Reference standard (%)							
	TrAP	TrCP	TrIP	TrNP	TrWP	TeCP	TeRP	PIP
TrAP	**68.1**	2.9	19.8	6.5	14.4	0	0	0
TrCP	0.4	**48.4**	1	6.4	2.7	0	0	0
TrIP	0.5	0.3	**49.8**	0.4	0.2	0	0	0
TrNP	1	2.5	0	**56.1**	0.8	0	0	0
TrWP	0.1	1.6	1.2	0.8	**40.4**	0	0	0
TeCP	0	0	0	0	0	**43.5**	1.8	0
TeRP	0	0	0	0	0	7.1	**68.2**	0
PIP	0	0	0	0	0	0	0	**36.5**
Missing	**30**	**44.2**	**28.3**	**29.9**	**41.4**	**49.5**	**30**	**63.5**
First level review								
	TrAP	TrCP	TrIP	TrNP	TrWP	TeCP	TeRP	PIP
TrAP	**83.6**	3.4	12.8	6.1	13	0	0	0
TrCP	0.6	**67.3**	1	5.7	2.1	0	0	0
TrIP	0.3	0.6	**74.7**	0.4	0	0	0	0
TrNP	1.5	0.6	0	**77.8**	1	0	0	0
TrWP	0.1	2.8	2.1	0	**64.8**	0	0	0
TeCP	0	0	0	0	0	**63.5**	1.2	0
TeRP	0	0	0	0	0	7.8	**95.8**	0
PIP	0	0	0	0	0	0	0	**68.2**
Missing	**13.9**	**25.3**	**9.4**	**10**	**19.2**	**28.7**	**13**	**31.8**
Second level review								
	TrAP	TrCP	TrIP	TrNP	TrWP	TeCP	TeRP	PIP
TrAP	**98.9**	3.3	0.7	2.5	1	0	0	0
TrCP	0.5	**95.5**	1.7	0.7	1	0	0	0
TrIP	0.2	0.6	**92.2**	0.4	0	0	0	0
TrNP	0.1	0	0	**96.4**	0	0	0	0
TrWP	0.1	0.3	2.7	0	**97.9**	0	0	0
TeCP	0	0	0	0	0	**94.2**	1.1	0
TeRP	0	0	0	0	0	5.5	**98.8**	0
PIP	0	0	0	0	0	0	0	**100**
Missing	**0.1**	**0.3**	**2.7**	**0**	**0**	**0.4**	**0.1**	**0**

4.5.3 Analysis of Primary Annotation Experiment

LOESS was used to visualize F_1-measures for concept identification, assertion classification, and relations. Across the 339 raw annotation batches for the 9 annotators the number of annotations ranged from 65 to 2,598 per document batch. The time decreased sharply from 25 seconds on document batch 1 to 12 seconds on document batch 10 before leveling off (Figure 4.3). F_1-measures for concept annotation on medical problems, treatments, and tests improved for the first 10 batches (from 0.82 to 0.87) and then plateaued (Figure 4.4). F_1-measures for assertion classification showed a similar trend as that of concept annotations reaching its first palateau at batch 15 (from 0.73 to 0.83) (Figure 4.5). F_1-measures for relations increased very slowly across the early half of the annotation task but decreased towards the end of the task with no obvious plateau (Figure 4.6).

Further analyses used GEE to compare F_1-measures at the annotated batch, document and task level for the same documents reviewed by annotators assigned to intervention groups. From these analyses (Tables 4.8, 4.9), significant differences were observed between documents that were pre-annotated at the semantic level versus documents annotated on raw full documents. Broken out by semantic class these differences were significant for medical treatments and tests but not for annotation of medical problems. Significant differences were also observed for sentence level, pre-annotated sentence, and pre-annotated full document interventions at the individual semantic class level and in situations where context is required such as assertion classification, or relations. In all cases, annotation on raw full documents produced the highest F_1-measures with the lowest annotator time.

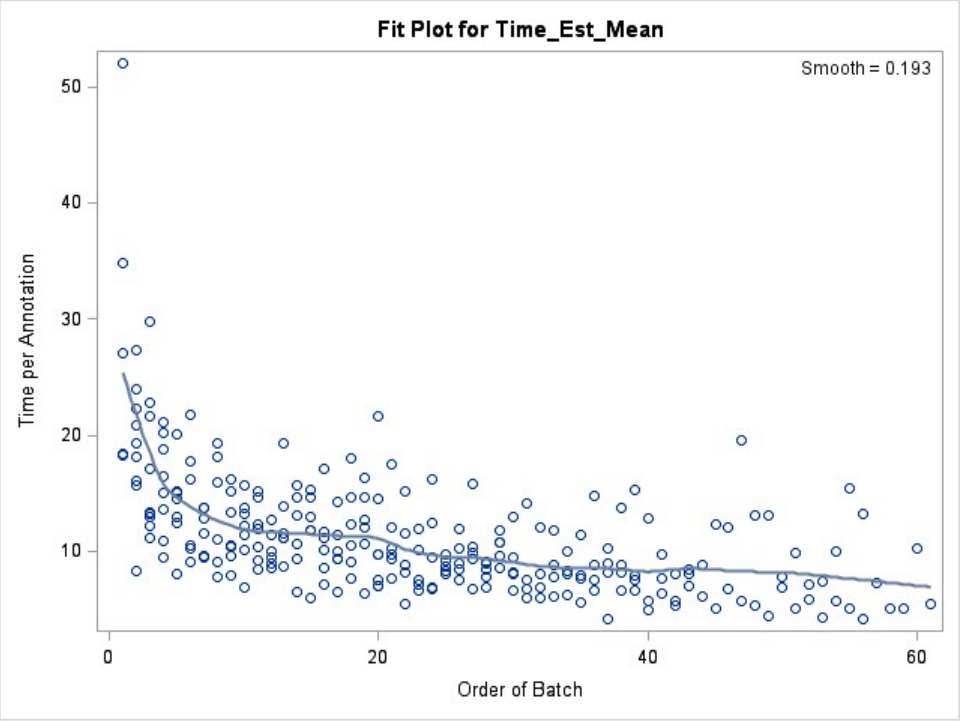

Figure 4.3: LOESS Plots - Mean Time Estimates

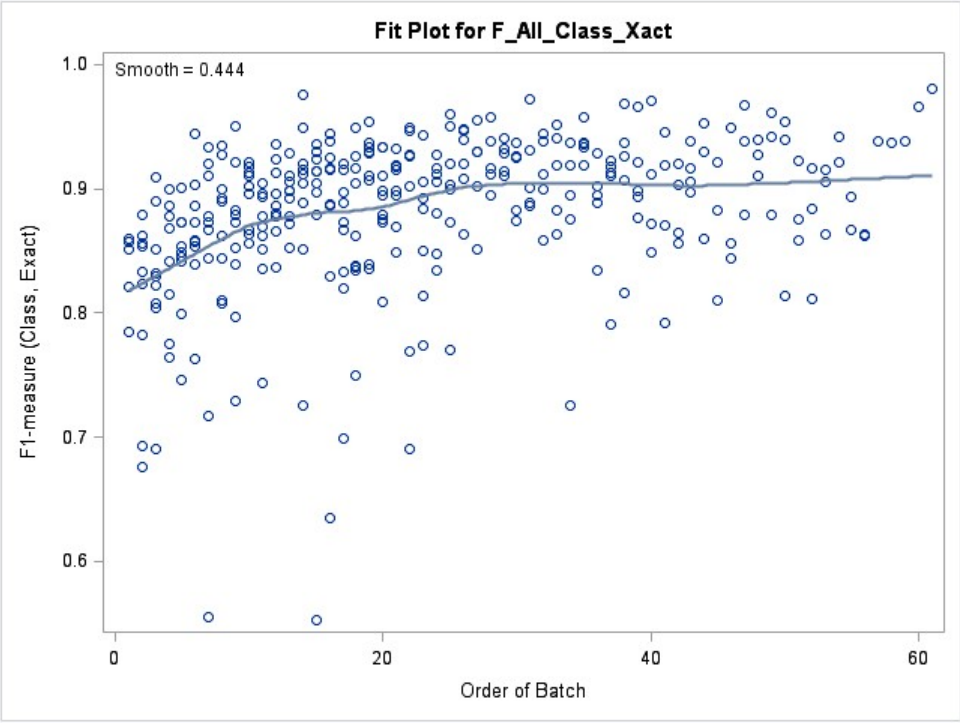

Figure 4.4: LOESS Plots - Class Exact F_1-Measure

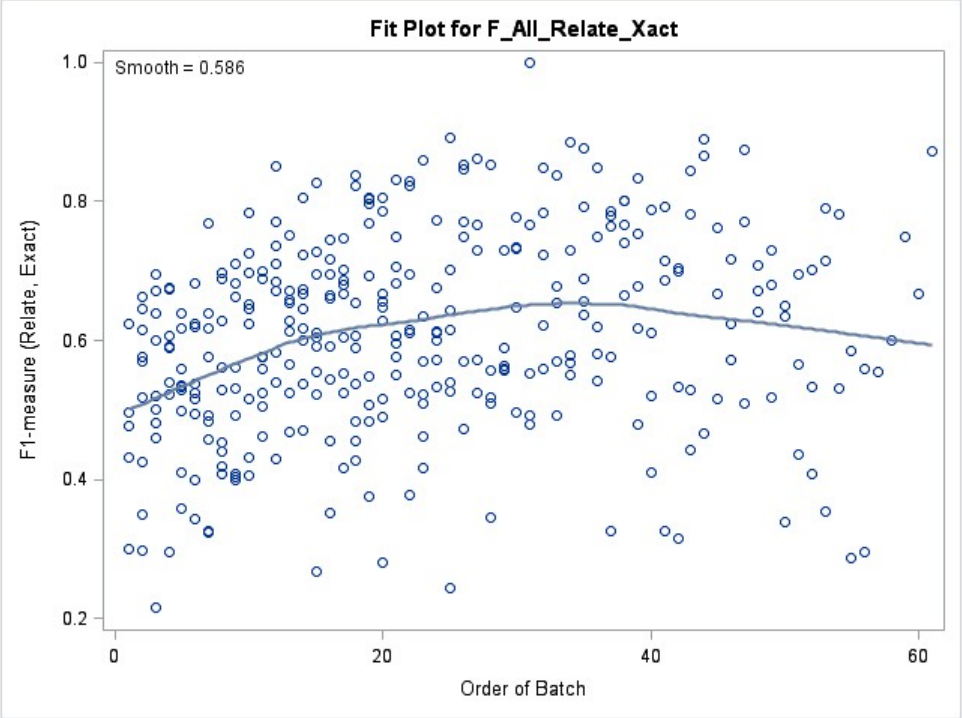

Figure 4.5: LOESS Plots - Relations Exact F_1-Measure

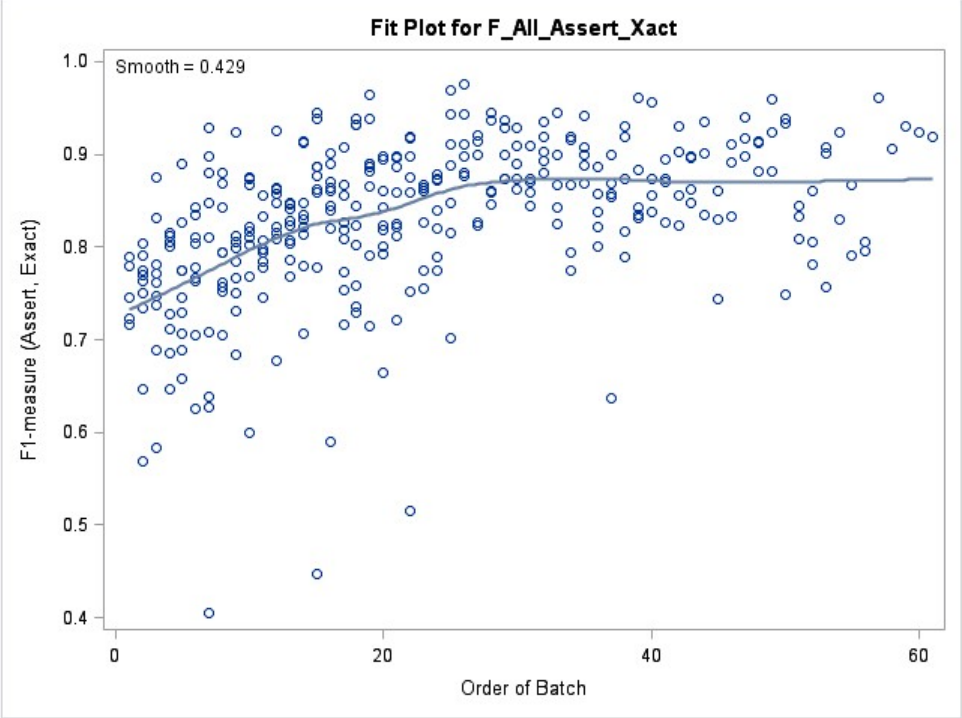

Figure 4.6: LOESS Plots - Assertions F_1-Measure

Table 4.8: Results of GEE and Primary Annotation Intervention Effects

Overall Class	**Intervention**	**_Intervention**	**Diff LSM**	**Pr > |z|**	**Adj P**
	f	pf	0.01847	0.0209	0.0715
	f	ps	0.01775	0.0018	0.0071
	f	s	0.01729	0.0001	0.0007
	pf	ps	-0.00072	0.8741	0.9971
	pf	s	-0.00118	0.8777	0.9973
	ps	s	-0.00046	0.9141	0.9992
Problem	**Intervention**	**_Intervention**	**Diff LSM**	**Pr > |z|**	**Adj P**
	f	pf	0.009782	0.2863	0.6448
	f	ps	0.005864	0.2557	0.6006
	f	s	0.01538	0.0216	0.077
	pf	ps	-0.00392	0.526	0.8923
	pf	s	0.0056	0.585	0.9269
	ps	s	0.009518	0.082	0.2525
Test	**Intervention**	**_Intervention**	**Diff LSM**	**Pr > |z|**	**Adj P**
	f	pf	0.02422	0.0296	0.1134
	f	ps	0.03465	0.0004	0.0023
	f	s	0.00295	0.6482	0.9597
	pf	ps	0.01043	0.087	0.2896
	pf	s	-0.02127	0.0314	0.1198
	ps	s	-0.0317	<.0001	0.0002
Treatments	**Intervention**	**_Intervention**	**Diff LSM**	**Pr > |z|**	**Adj P**
	f	pf	0.02585	0.0119	0.0479
	f	ps	0.0285	<.0001	<.0001
	f	s	0.0252	0.0023	0.0099
	pf	ps	0.002652	0.6896	0.9735
	pf	s	-0.00065	0.955	1
	ps	s	-0.0033	0.6648	0.9664
Assertions	**Intervention**	**_Intervention**	**Diff LSM**	**Pr > |z|**	**Adj P**
	f	pf	0.01928	0.212	0.4698
	f	ps	0.0156	0.0183	0.0561
	f	s	0.02871	<.0001	<.0001
	pf	ps	-0.00369	0.7316	0.9637
	pf	s	0.009428	0.5717	0.8808
	ps	s	0.01311	0.0388	0.1125

Table 4.8: Continued

Relations	**Intervention**	**_Intervention**	**Diff LSM**	**Pr > |z|**	**Adj P**
	f	pf	0.04773	<.0001	<.0001
	f	ps	0.05155	0.0004	0.0014
	f	s	0.01271	0.5999	0.9445
	pf	ps	0.003818	0.6349	0.9573
	pf	s	-0.03501	0.1437	0.4105
	ps	s	-0.03883	0.1496	0.4234
Time	**Intervention**	**_Intervention**	**Diff LSM**	**Pr > |z|**	**Adj P**
	f	pf	-0.3327	0.4463	0.8324
	f	ps	-2.84	<.0001	<.0001
	f	s	-3.5382	<.0001	<.0001
	pf	ps	-2.5073	<.0001	<.0001
	pf	s	-3.2055	<.0001	<.0001
	ps	s	-0.6982	0.2429	0.5712

Interventions: **f** = full document, **pf** = pre-annotated full documents, **s** = sentence, **ps** = pre-annotated sentence.

Table 4.9: Exact Estimates for Least Mean Squares by Intervention

Overall Class	**Intervention**	**Exact estimate LSM**	**lower bound**	**upper bound**
	f	0.9045	0.8863	0.9226
	pf	0.886	0.8636	0.9084
	ps	0.8867	0.8708	0.9027
	s	0.8872	0.8767	0.8977
Problem	**Intervention**	**Exact Estimate LSM**	**lower bound**	**upper bound**
	f	0.9147	0.8987	0.9307
	pf	0.9049	0.8829	0.9269
	ps	0.9088	0.8973	0.9204
	s	0.8993	0.8895	0.9091
Test	**Intervention**	**Exact Estimate LSM**	**lower bound**	**upper bound**
	f	0.8918	0.8688	0.9147
	pf	0.8676	0.8415	0.8936
	ps	0.8571	0.8359	0.8783
	s	0.8888	0.8767	0.9009
Treatment	**Intervention**	**Exact Estimate LSM**	**lower bound**	**upper bound**
	f	0.8982	0.8762	0.9201
	pf	0.8723	0.8453	0.8993
	ps	0.8697	0.8532	0.8861
	s	0.873	0.8535	0.8925
Assertions	**Intervention**	**Exact Estimate LSM**	**lower bound**	**upper bound**
	f	0.8641	0.8428	0.8853
	pf	0.8448	0.8098	0.8798
	ps	0.8485	0.8313	0.8656
	s	0.8353	0.823	0.8477
Relations	**Intervention**	**Exact Estimate LSM**	**lower bound**	**upper bound**
	f	0.673	0.625	0.721
	pf	0.6253	0.5867	0.6638
	ps	0.6214	0.5721	0.6707
	s	0.6603	0.6055	0.715
Time	**Intervention**	**Exact Estimate LSM**	**lower**	**upper bound**
	f	9.6602	8.4829	10.8375
	pf	9.9929	9.0722	10.9136
	ps	12.5002	11.7087	13.2917
	s	13.1984	11.6113	14.7855

Interventions: **f** = full document, **pf** = pre-annotated full documents, **s** = sentence, **ps** = pre-annotated sentence.

4.6 Discussion

There are specific benefits to modifying annotation workflows when generating reference standards. At the same time, available resources must be balanced with the need to evaluate and develop new approaches of generating adequate labeled data and support the operational goals of a shared community challenge. Our experimental results suggest that for tasks such as identification of concepts additional review layers beyond the typical use of double annotation with adjudication may not be necessary. However, for tasks such as assertion classification and more complex tasks such as relation classification, additional review layers result in significant improvements in reference standard validity.

We learned several important lessons from the annotation efforts for this challenge that can be generalized to other similar clinical corpus annotation projects. First, there are clear opportunities to implement more efficient methods of annotating texts, by modifying workflows, and fielding and processing completed batches of annotated documents. Second, annotation guidelines and schema must be developed, adequately pilot tested, and made available long before the actual annotation begins. Annotation guidelines must also remain constant throughout in order to ensure consistency and accuracy. When modifications are necessary they must be made in a way that preserves annotation task reliability. Finally, use of additional review levels introduced significant improvements in validity for human information extraction and classification.

Relying only on time stamping from the Knowtator tool significantly underestimates the time required for guidelines review, discussions and/or clarification,

additional training and other activities, such as checking in review batches, exporting annotations, importing pre-annotations, or obtaining newly assigned batches. Also, time stamping estimation does not include the programming time required to implement machine-assisted approaches for verification nor administrative support for annotation activities. Clear opportunities exist to streamline these processes using automated methods, particularly at the level of batch assignment and data management using some application that allows human annotators to pull documents from some common data pool rather than using a push mechanism. Future research directions include determining under what situations modified workflows provide optimal benefits in terms of annotator reliability, reference standard validity and tradeoff with annotator workload.

4.7 Conclusion

The 2010 i2b2/VA challenge task provided an excellent opportunity to examine some of the issues related to large-scale annotation of clinical document corpora. The level of task complexity, the number of annotators, and the high prevalence of annotations made this an interesting use case to evaluate new approaches for clinical corpus annotation. We experimented with noninteractive pre-annotation and integrated a modified workflow that included various levels of review, incorporating machine-assisted approaches to validate annotated information and check on compliance with our guidelines. These data have been used as the starting point for other expanded and more complex tasks or as part of other annotation efforts. This has greatly expanded the reusability and utility of the data generated for this challenge task to other related information extraction or classification efforts.

4.8 Acknowledgments

This work was supported in part by the NIH Roadmap for Medical Research, Grant U54LM008748 and the VA Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374. Views expressed are those of the authors and not necessarily those of the Department of Veterans Affairs. We gratefully commend the efforts of all annotators, our study coordinators, Robyn Barrus and Neil Nokes, Matthew Maw and Ying Suo for program, data management, and statistical programming support. We wish to thank the 2010 i2b2/VA challenge organizers for their participation and helpful consultation on annotation guidelines and schema.

4.9 References

1. Meystre, S.M., Savova, G.K., Kipper-Schuler, K.C., Hurdle, J.F. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform, 2008: p. 128-44.

2. Uzuner, O., Mailoa, J., Ryan R., Sibanda, T. Semantic relations for problem-oriented medical records. Artif Intell Med, 2010. 50(2): p. 63-73.

3. Uzuner, O., Luo, Y., Szolovits, P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc, 2007. 14(5): p. 550-63.

4. Uzuner, O., Goldstein, I., Luo, Y., Kohane, I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc, 2008. 15(1): p. 14-24.

5. Uzuner, O. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc, 2009. 16(4): p. 561-70.

6. Uzuner, O., Solti, I., Cadag, E. Extracting medication information from clinical text. J Am Med Inform Assoc, 2010. 17(5): p. 514-8.

7. Grishman, R., Sundheim. B. Message Understanding Conference-6: a brief history. 16th Conference on Computational Linguistics (COLING), 1996: p. 466-71.

8. Hersh, W., Bhupatiraju, R.T., Corley, S. Enhancing access to the Bibliome: the TREC Genomics Track. Stud Health Technol Inform, 2004. 107(Pt 2): p. 773-7.

9. Sparck Jones, K. Reflections on TREC Information Processing Management. In: TREC-2 Proceedings of the Second conference on text retrieval. 1995. 31(3): p. 291-314.

10. Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., Setzer, A. Building a semantically annotated corpus of clinical texts. J Biomed Inform, 2009. 42(5): p. 950-66.

11. Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J., Roberts, I., Setzer, A., Tapuria, A., Wheeldin, B. The CLEF corpus: semantic annotation of clinical text. AMIA Annu Symp Proc, 2007: p. 625-9.

12. Kim, J.D., Ohta, T., Tateisi, Y., Tsujii, J. GENIA corpus--semantically annotated corpus for bio-textmining. Bioinformatics, 2003. 19 Suppl 1: p. i180-2.

13. Kim, J.D., Ohta, T., Tsujii, J. Corpus annotation for mining biomedical events from literature. BMC Bioinformatics, 2008. 9: p. 10.

14. Marcus, M.P., Santorini, B., Marcinkiewicz, M.A. Building a large annotated corpus of english; penn treebank. Computational Linguistics, 1993. 19: p. 313-330.

15. Hripcsak, G., Wilcox, A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. J Am Med Inform Assoc, 2002. 9(1): p. 1-15.

16. Hripcsak, G., Heitjan, D.F. Measuring agreement in medical informatics reliability studies. J Biomed Inform, 2002. 35(2): p. 99-110.

17. Hripcsak, G., Rothschild, A.S. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc, 2005. 12(3): p. 296-8.

18. Chapman, W.W., Dowling. J.N., Hripcsak, G. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. Int J Med Inform, 2008. 77(2): p. 107-13.

19. Mayer, J., Shen, S., South, B.R., Meystre, S., Friedlin, F.J., Ray, W.R., Samore, M.H. Inductive creation of an annotation schema and a reference standard for de-identification of VA electronic clinical notes. AMIA Annu Symp Proc, 2009. 2009: p. 416-20.

20. South, B.R., Shen, S., Jones, M., Garvin, J., Samore, M.H., Chapman, W.W., Gundlapalli, A.V. Developing a manually annotated clinical document corpus to identify phenotypic information for inflammatory bowel disease. BMC Bioinformatics, 2009. 10 Suppl 9: p. S12.

21. Savova, G.K., Coden, A.R., Sominsky, I.L., Johnson, R., Ogren, P.V., De Groen, P.C., Chute, C.G. Word sense disambiguation across two domains: biomedical literature and clinical notes. J Biomed Inform, 2008. 41(6): p. 1088-100.

22. Savova, G.K., Ogren, P.V., Duffy, P.H., Buntrock, J.D., Chute, C.G. Mayo clinic NLP system for patient smoking status identification. J Am Med Inform Assoc, 2008. 15(1): p. 25-8.

23. Chapman, W.W., Bridewell, W., Hanbury, P., Cooper, G.F., Buchanan, B.G. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform, 2001. 34(5): p. 301-10.

24. Chapman, W.W., Haug, P.J. Comparing expert systems for identifying chest x-ray reports that support pneumonia. Proc AMIA Symp, 1999: p. 216-20.

25. Uzuner, O., Solti, I., Xia, F., Cadag, E. Community annotation experiment for ground truth generation for the i2b2 medication challenge. J Am Med Inform Assoc, 2010. 17(5): p. 519-23.

26. Uzuner, O., South, B.R., Shen, S., DuVall, S. 2010 i2b2/VA Challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc, 2011. Sept-Oct, 18(5):552-6.

27. Saeed, M., Lieu, C, Raber, G., Mark, R.G. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Computers in Cardiology, 2002. 29: p. 641-644.

28. Chapman, W.W., et al. Creation of a Repository of Automatically De-Identified Clinical Reports: Processes, People, and Permission. Summit on Clinical Research Informatics, San Francisco: American Medical Informatics Association (AMIA). 2011.

29. Ogren, P.V. Knowtator a protege plug-in for annotated corpus construction. Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2006: p. 273-5.

30. Knowtator 1.9 beta 2. Available from: http://sourceforge.net/projects/knowtator/files/Knowtator/.

31. Musen, M.A., Gennari, J.H., Eriksson, H., Tu, S.W., Puerta, A.R. PROTEGE-II: computer support for development of intelligent systems from libraries of components. Medinfo, 1995. 8 Pt 1: p. 766-70.

32. Apache Lucene. 2011-12. Available: http://lucene.apache.org/core/

CHAPTER 5

AIM 3: A PROTOTYPE TOOL SET TO SUPPORT MACHINE-ASSISTED ANNOTATION

Originally published in

B.R. South, S. Shen, J. Leng, T.B. Forbush, S.L. DuVall, W.W. Chapman. 2012. Association for Computational Linguistics (ACL) Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP) 2012.

http://aclweb.org/anthology/W/W12/W12-2416.pdf

A Prototype Tool Set to Support Machine-Assisted Annotation

Brett R. South[1,2], Shuying Shen[1,2], Jianwei Leng[2], Tyler B. Forbush[4], Scott L. DuVall[3,4], Wendy W. Chapman[5]

Departments of [1]Biomedical Informatics, [2]Internal Medicine, and [3]Radiology University of Utah, Salt Lake City, Utah, USA

[4]IDEAS Center SLCVA Healthcare System, Salt Lake City, Utah, USA

[5]University of California, San Diego, Division of Biomedical Informatics, La Jolla, California, USA

brett.south@hsc.utah.edu, shuying.shen@hsc.utah.edu, jianwei.leng@utah.edu, tyler.forbush@utah.edu, scott.duvall@utah.edu, wendy.w.chapman@gmail.com

Abstract

Manually annotating clinical document corpora to generate reference standards for Natural Language Processing (NLP) systems or Machine Learning (ML) is a time-consuming and labor-intensive endeavor. Although a variety of open source annotation tools currently exist, there is a clear opportunity to develop new tools and assess functionalities that introduce efficiencies into the process of generating reference standards. These features include: management of document corpora and batch assignment, integration of machine-assisted verification functions, semi-automated curation of annotated information, and support of machine-assisted pre-annotation. The goals of reducing annotator workload and improving the quality of reference standards are important considerations for development of new tools. An infrastructure is also needed that will support large-scale but secure annotation of sensitive clinical data as well as crowdsourcing which has proven successful for a variety of annotation tasks. We introduce the Extensible Human Oracle Suite of Tools (eHOST) http://code.google.com/p/ehost that provides such functionalities that when coupled with server integration offer an end-to-end solution to carry out small or large scale as well as crowd sourced annotation projects.

1 Introduction

Supervised learning methods benefit from a reference standard that is used to train and evaluate the performance of Natural Language Processing (NLP) or Machine Learning (ML) systems for information extraction and classification. Ideally, generating a reference standard involves the review of more than one annotator with an accompanying adjudication step to resolve discrepancies (Roberts et al., 2007; Roberts et al., 2009). However, manual annotation of clinical, texts is time-consuming, expensive, and requires considerable effort. Reducing the time and costs required for manual annotation could be achieved by developing new tools that integrate methods to more efficiently annotate clinical texts and integrate a management interface that allows administration of large or small scale annotation projects. Such a tool could also integrate methods to pre-annotate entities such as noun phrases or clinical concepts mapped to a standard vocabulary. Efficiencies could be realized via reduction in human workload, modification of annotation tasks that could include crowd sourcing, and implementation of machine-assisted approaches.

Typically annotation of clinical texts requires human reviewers to identify information classes of interest called *"markables"*. These tasks may also require reviewers to assign attributes to those information classes and build relations between spans of annotated text. For each annotation task there may be one or many types of markables and each markable class may be associated with one or more spans of text and may include single or even multiple tokens. These tasks may occur simultaneously, or may also be done in different steps and by multiple reviewers. Furthermore, these activities require written guidelines that clearly explicate what infor-

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 130–139,
Montréal, Canada, June 8, 2012.

mation to annotate, specifics about each markable class, such as how much information to include in annotated spans, or syntactic rules to provide further guidance on annotated spans. Annotation tasks may benefit by incorporating rules or guidelines as part of the annotation task itself in the form of machine-assisted verification.

There are many annotation tools available, and the majority of them were designed for linguistic or gene annotation. Linguistic annotation tools such as Callisto and WordFreak are standalone clients suitable for small to medium scale tasks where collaborative effort is not emphasized. Functionality integrated with eHOST was inspired by existing features of these tools with the intent of providing a more efficient means of reference standard generation in a large collaborative environment. One annotation tool called Knowtator, a plug-in for Protégé (Musen, M.A., et al, 1995) developed by Ogren (2006) has been widely used to annotate clinical texts and generate reference standards. However, no standalone system exists that can provide end users with the ability to manually or semi-automatically edit, curate, and easily navigate annotated information. There are also specific functionalities that are missing from open source annotation tools in the clinical and biomedical domains that would introduce efficiencies into manual annotation tasks. These functionalities include: annotation of clinical texts along with database storage of stand-off annotations, the ability to interactively annotate texts in a way that allows users to react to either pre-annotations imported from NLP or ML systems or use exact string matching across an active corpus to identify similar spans of text to those already annotated. Additionally, these systems do not generally support crowd sourcing, machine-assisted pre-annotation or verification approaches integrated directly with the annotation tool.

This paper discusses development of a prototype open source system designed to provide functionality that supports these activities and offers an end-to-end solution when coupled with server integration to reduce both annotator and administrative workload associated with reference standard. We introduce the Extensible Human Oracle Suite of Tools (eHOST) created with these expectations in mind.

2 Background

Our goal for these development efforts was to build a prototype open source system that improves upon existing tools by including new functions and refining capabilities available in other annotation tools. The resulting GUI interface provides a means of visually representing annotated information, its attributes, and relations between annotated mentions. These efforts also focused integrating various machine-assisted approaches that can be used to easily curate and navigate annotated information within a document corpus, pre-annotate information, and also verify annotations based on rules checks that correspond with annotation guidelines or linguistic and syntactic cues.

The eHOST provides basic functionality including manual annotation of information representing markable classes and assignment of information attributes and relationships between markable classes. Annotations exported from eHOST are written using the XML format as Knowtator thus allowing integration of inputs and outputs to and from Knowtator and indirectly to Protégé 3.3.1. Coupling eHOST with an integrated server package such as the one under development by the VA Informatics and Computing Infrastructure (VINCI) called the Chart Administration Server for Patient Review (CASPR) provides one method of increasing efficiencies for small or large-scale annotation efforts that could also include crowd sourcing.

2.1 System Features Development

In the domains of computational linguistics and biomedical informatics various approaches that can be used to improve annotation efficiencies have been evaluated for a variety of tasks including information extraction and classification. While several methods may help reduce the time and costs required to create reference standards, one of the simplest approaches may include integrating machine-assisted methods to pre-annotate relevant spans of text allowing the annotator to add missing annotations, modify spans, or delete spurious annotations. Neveol (2011) evaluated use of automatic semantic pre-

annotation of PubMed queries. This study showed a significant reduction in the number of required annotations when using pre-annotations, reduction in annotation time with higher inter-annotator agreement. Pre-annotation using simple approaches such as regular expressions coupled with dictionaries (South et al., 2010a) based on the UMLS as a source of lexical knowledge (Friedman, 2001) and pre-annotation of information representing protected health information (South et al., 2010b). In both cases finding that annotators preferred particular types of pre-annotation over others, but improvements in reference standard quality occur when pre-annotation was provided. Others have explored the use of third party tools for the pre-annotation task for UMLS concepts (Savova, 2008) and pre-annotation using an algorithmic approach (Chapman, et al., 2007) combined with domain expert annotations reused for temporal relation annotation (Mowery, 2008). Savova (2008) suggests limited utility when a third party tool is used for pre-annotation and Mowery (2008) suggest that even with domain expert pre-annotations, additional features are required to discern temporality. Finally, Fort and Sagot (2008) evaluated using pre-annotation for part-of-speech tagging on the Penn Tree bank corpus and demonstrate a gain in quality and annotation speed even with a not so accurate tagger.

Semi-automated curation has been explored as a means to build custom dictionaries for information extraction tasks (Riloff, 1993). More recently this approach was spurred on by the BioCreative II competition (Yeh et al., 2003). Alex et al., (2008), explored the use of NLP-assisted text mining to speed up curation of biomedical texts. Settles et al., (2008) estimates true labeling costs and provides a review of active and interactive learning approaches as a means of providing labels and reducing the cost of obtaining training data (Settles, 2010). Although eHOST does not yet include an active learning module it does provide one means of interactive annotation so these are important considerations for future development efforts.

In the biomedical informatics domain crowd sourcing has been evaluated as part of the 2009 i2b2 Medication Challenge (Uzuner, 2010). Nowak and Ruger (2010) provide estimates of annotation reliability from crowd sourcing of image annotation. Hsueh et al., (2009) provide estimates of the quality of crowd sourcing for sentiment classification using both experts and non-expert annotators. In all three cases the resulting annotation set was of comparable quality to that derived from expert annotators. Wang et al., (2008) make general recommendations for best approaches to crowd sourcing that include closer interactions between human and machine methods in ways that more efficiently connect domain expertise with the annotation task.

Subsequent sections in this paper walk the reader through the various basic and advanced features eHOST provides. These features have been developed in a way that provides flexibility to add additional modules that support improvements in annotation workflow and efficiency for a variety of annotation scenarios applicable to computational linguistics and biomedical informatics. Some of these features may be useful for crowd-sourced efforts whereas others may simply represent an improvement in the way annotation is visualized or how manual effort can be reduced. Figures in this paper use a set of synthetic clinical documents and a demonstration annotation project based on the 2010 and 2011 i2b2/VA annotation tasks as examples available from http://code.google.com/p/ehost.

2.2 Systems Architecture

The eHOST is a client application that can run on most operating systems that supports Java including, most Microsoft Windows x86/x64 platforms, Apple Mac OS X, Sun Solaris, and Linux. The application uses standardized formats including a file folder system, and structured XML inputs and outputs. These capabilities also support integration with other open source tools for annotation and knowledge management including Knowtator and Protégé. An Extract-Transform-Load process (ETL) is used by the system to import concept information from different sources, such as XML or Protégé PINS files. These inputs sources are normalized for loading into eHOST. All data that exists in the data pool can be transformed into various output formats. Raw input data documents in a single text file or sequential text files in a file folder system.

Information representing an annotation in-

cluding concept attributes such as the annotated span, attributes, and relationships between annotations are inserted into a common data pool using a dynamic structured storage space. The data pool ensures that eHOST has capabilities to add new functions easily without making major changes to system architecture.

2.3 Annotation Project Workspace

In eHOST each project has its own user assigned workspace that includes an annotation schema and document corpus. Annotation schema can also be imported from an existing Protégé PINS file. Project settings can be inherited from existing projects for similar annotations tasks using eHOST. Other workspace functions include quickly switching between up to five of the most recently used workspaces. A workspace can be assigned for each annotation layer or document batch. In these situations, an annotator would receive a pre-compiled project that specifies all settings including any text documents and the annotation schema. Defining a workspace is a particularly useful function in situations where annotations may be crowd sourced and there may be multiple layers of annotation that are potentially fielded to many annotators.

2.3.1 Corpus Management

For any annotation task, the end user must manage the document corpus, which can originate from a server or a file folder system that contains individual text files. Using the stand-alone eHOST client tool, corpus management is accomplished via the current workspace (Figure 1). When the user initializes a new project, documents are placed in a "corpus" folder that is associated with the newly created annotation project. All text files, are copied to the "corpus" folder at the time of workspace assignment. Therefore, there is no risk of deleting the original documents associated with each new annotation project. This feature makes distribution of projects easier, because of the consistency between the workspace, corpus assignment and annotation output folders. For crowd-sourced projects eHOST can be integrated with a backend server via web services using an administrative module called CASPR.

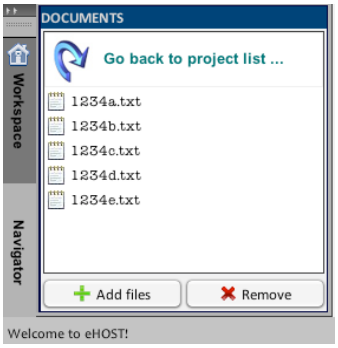

Figure 1. eHOST corpus management

2.3.2 Viewer/Editor Panels

Figure 2 shows an annotation for *"full body pain"*, (shown with black bar above and below the active annotation) and information for that annotation including the annotated span, the class assignment and an assertion for the 2010 and 2011 i2b2/VA Challenge annotation tasks (Uzuner et al., 2011 and Uzuner et al., 2012). The result editor tab and its associated panels serve as the central place for basic annotation features. These functionalities include: assigning an annotator, creating new annotations or adjusting annotated spans of text and assignment of attributes or creating relationships between annotated spans of text. Other functions in the results editor tab include navigation between documents in the active corpus, resizing the text displayed in the document viewer, and *"save"* and *"save as"* functions that assigns a path for XML output files. The end user can easily remove all annotations in a document or remove specific kinds of annotations by deleting a "markable" class as well as remove attributes, and relationships between all annotations.

From the navigator screen in the stand-alone eHOST client tool a user can build annotation schema specifying markable classes, their associated attributes, and any allowed relationships. The navigator interface allows the user to review all annotated spans either within the current document or across the entire document corpus, toggle the view of each class on or off, see counts for all unique annotations and all annotations for each class, and choose a class for a fast annotate mode.

An annotation editor panel allows the user to view more detailed information for each selected

annotation. This includes the time stamp of when the annotation was created, annotator assignment, comments on the annotation and class, attribute and relationship information.

Annotations can be created using several approaches from the result editor. In the normal mode, a class assignment window appears when the user selects a span of text, new annotations are generated by selecting any one of the markable classes. Activating a *"one click annotate"* mode is possible by checking the box next to a class of markables. Under this mode, any text selected is automatically annotated as that markable class. This feature improves task efficiencies when categories of markables are low or annotations of the same category cluster in small sections. Keyboard shortcuts have also been integrated with eHOST to reduce annotator click burden and dependence on a mouse. These shortcuts are available for tasks such as modification of spans, deletion of annotations, and navigation between annotations.

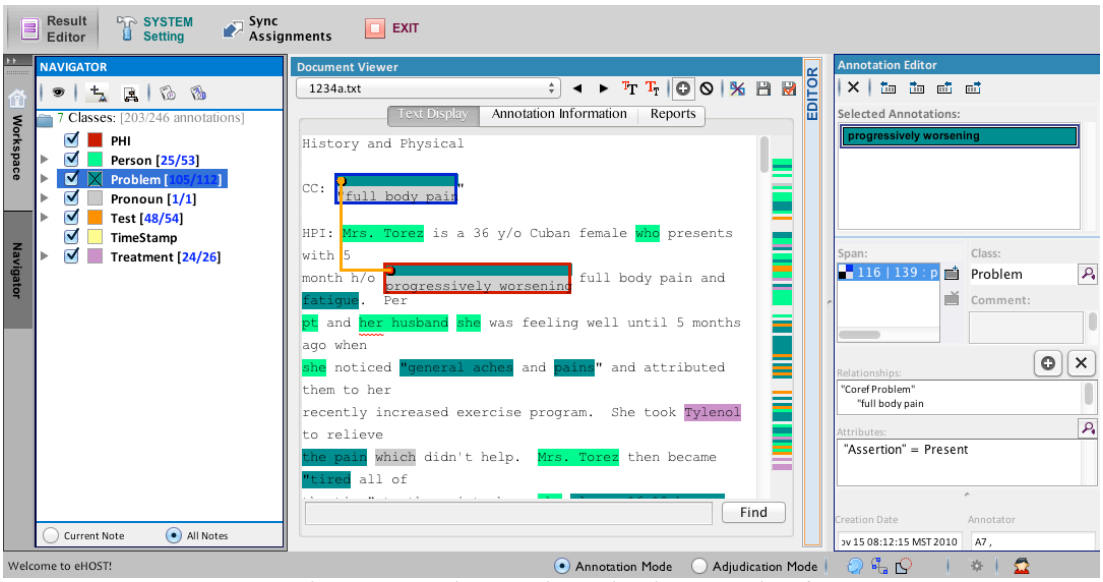

Figure 2. Example annotations using the eHOST interface

2.3.3 Server Integration

Annotation projects of any scale benefit from an automated means of building and distributing batches of texts to annotators, managing standoff XML files generated from annotation tasks or written directly to a database and getting and submitting assignments with minimal user input. Coupling eHOST with server components that comply with the web services API defined for eHOST allows these functionalities. The CASPR module under development by VINCI provides a means to automate the administration of annotation efforts that could include crowdsourced annotation projects.

Clicking on the sync assignments tab in the eHOST client (Figure 2) brings up a GUI that allows annotators to sync with a server location, enter credentials, see documents assigned, and designate documents as on hold, in process, or completed. When a user syncs and gets assignments from CASPR, a project folder is created that contains the annotation schema, text documents, and annotations sent from the server. The CASPR module allows an annotator to open the project and complete their task without needing to manage files or folders. Once completed, annotations can be synced to the server, and the next assignment will be loaded. The CASPR module allows iterative distribution of annotation batches without sending large sets of documents to annotators that may contain sensitive data, decreasing the risk of breaches in privacy and data security.

2.3.4 Additional Features

The document viewer panel employs visual cues to display relationships between annotations using color coding representing a parent and child node and line indicator between them showing the relationship. An "annotation profiler" to the right of the scroll bar shows the density of annotations color-coded to their categories, as well as relative to their positions in the document. This type of data visualization is useful to see the relative location of annotations within a single document or across an en tire document corpus.

An adjudication mode is also included in the stand-alone eHOST client that allows difference matching and side-by-side comparison of annotations for efficient adjudication of discrepancies between annotations. Standard reporting metrics can be calculated including Inter-Annotator Agreement (IAA), Recall, Precision and F_1-Measure.

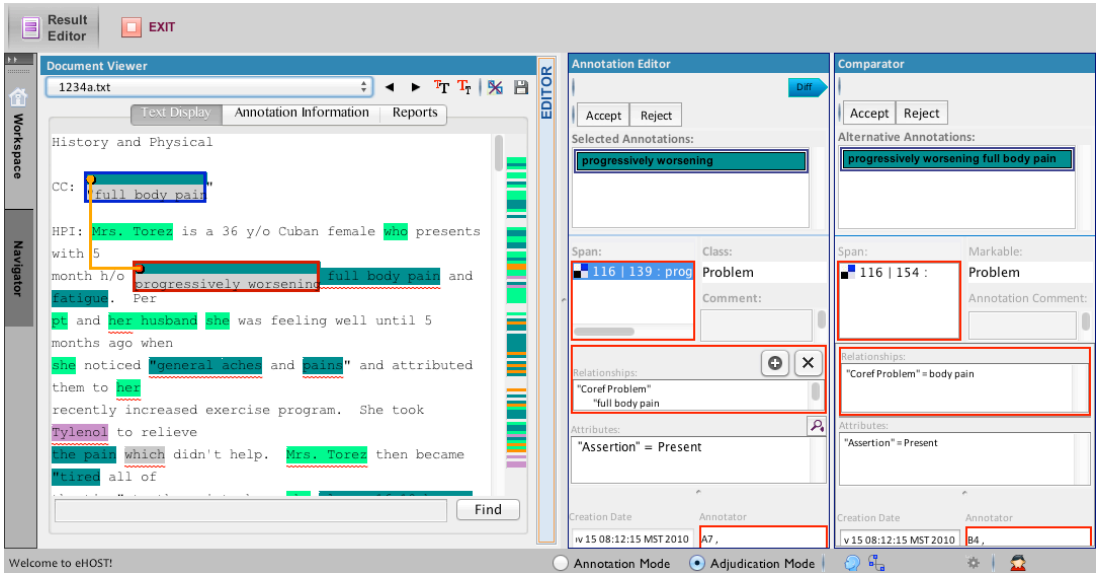

Figure 3. eHOST adjudication mode showing discrepant annotations between annotators A7 and B4

In Adjudication mode discrepant annotations are shown using a wavy red underline in the editor window and by a red bolded outline in a side by side two panel view between the annotation editor and comparator (Figure 3). These metrics and comparison tables between annotator results on the same documents can be output as HTML formatted reports that can be used by an adjudicator to quickly identify discrepancies between annotators (Figure 4). These reports and the editor window display can also be used to quickly train annotators on new clinical domains using a reference standard created by domain experts for training purposes. Using these features error analysis can also be done by importing outputs from an NLP system that have been converted into the XML format used by eHOST.

File: 1234a.txt
h 5 month h/o progressively worsening full body pain and fatigue. Per pt and her husband she was feeling well until 5 months ago when she noticed "general aches and pains" and attributed them to her recently increased exercise program. She took Tylenol to relieve the pain which didn't help. Mrs. Torez the

	Annotator:[A7,]	Annotator:[B4,]
Annotation Text	"general aches	aches
Span	(251,265)	(260,265)
Class	Problem	Problem
Relationship	linked to "progressively worsening" with relationship:[Coref Problem]	linked to "progressively worsening full body pain" with relationship:[Coref Problem]
Attributes	Assertion = Conditional;	Assertion = Conditional;

Figure 4. HTML Formatted report showing discrepant annotations between annotators A7 and B4

3 Advanced eHOST Features

There are also other more advanced features that have been integrated with eHOST. These include an "Oracle" mode that allows semi-automated annotation of similar spans of text across a document corpus, a means to easily and quickly curate annotated spans of text to create custom dictionaries, and machine-assisted pre-annotation integrated with the annotation tool itself.

3.1 Oracle Mode

Also implemented with eHOST is an "Oracle" mode which uses exact string matching allowing the user to annotate all spans of text that are identical to a new annotation. The oracle lists where these candidate annotations are found along with the surrounding context. The annotator can then accept or reject candidate spans annotated with the same markable class. Oracle mode can run within the current document or across the entire document corpus. This type of functionality is useful for annotation tasks that may involve identifying and marking spans of text that are repetitive or follow the same format For example, the 2011 i2b2/VA annotation task in which annotation of pronominal information was required for co-reference resolution (Figure 5).

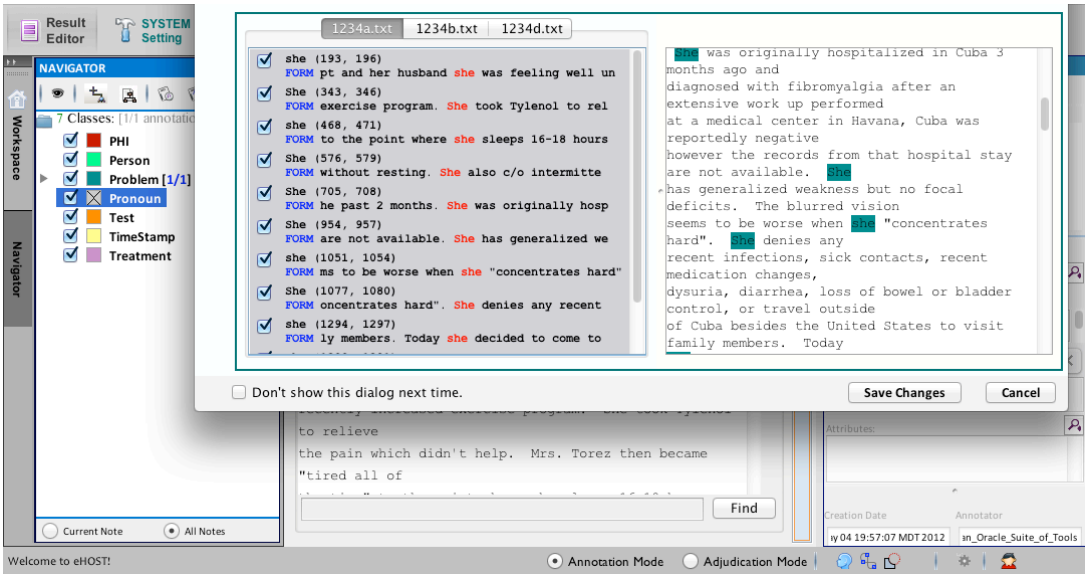

Figure 5. Example annotations generated using the eHOST "Oracle" mode

3.2 Semi-Automated Curation and Dictionary Management

Using the navigator window users can navigate to all annotations in either a single document or across an entire document corpus (Figure 6). The end user can curate annotations directly, create classes on the fly, or add attributes to annotations found from the navigator pane. These functions also allow users to easily identify spurious annotations introduced from machine-assisted approaches correct misclassification errors, and quickly curate all annotations within a single document or across an entire document corpus.

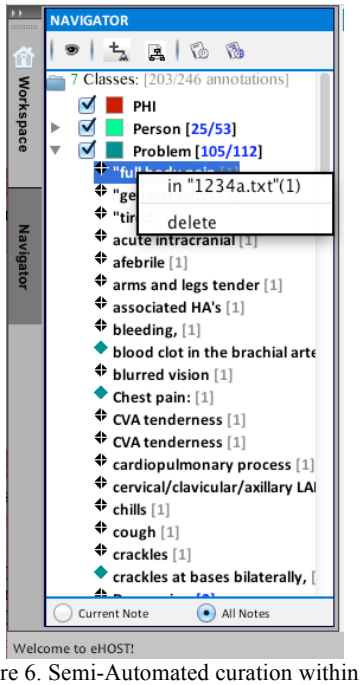

Figure 6. Semi-Automated curation within the document corpus

One task often associated with development of NLP systems involves manually creating or enhancing some existing representation of lexical knowledge that can be used as a domain specific dictionary. Using eHOST users can export annotations to create a dictionary of terms, phrases, or individual tokens that have been identified by human annotators and assigned to markable information classes. Once curated, annotated information can be exported as a new dictionary. User created dictionaries can be integrated with a database or exported and used in the creation of some ontologic representation of information using Protégé. Output from a dictionary manager is in the form of a delimited text file and can therefore be modified to fit any standardized information model or used to pre-annotate subsequent document batches.

3.3 Machine-Assisted Pre-Annotation

An interface is provided in eHOST that can be used for machine-assisted pre-annotation of documents in the active project corpus using either dictionaries or regular expressions based approaches. Users can import libraries of regular expressions or build their own regular expressions using a custom regular expression builder. Users can build and modify dictionaries created as part of annotation tasks that may include semi-automated curation steps. Dictionaries and regular expressions can also be coupled with the ConText algorithm (Chapman et al., 2007) to identify concept attributes such as negation, experiencer, and temporality. Pre-annotations derived from some external third party source such as an NLP system written as Knowtator XML outputs may also be imported into eHOST or passed to eHOST using CASPR.

Computational speed required for pre-annotation can be improved by selecting an option to use an internal statistical dictionary indexing function. This feature is particularly useful in situations where pre-annotation dictionaries are extremely large, such as where a subset of some standard vocabulary may be used to pre-annotate documents. Using the result editor and its associated functions annotators can add missed annotations, modifying existing annotations and delete spurious annotations. Handling pre-annotations in this way allows troubleshooting and error analysis of NLP system outputs imported into eHOST that can be shown to a reviewer in context and also facilitates interactive annotator training.

3.4 Machine-Assisted Verification

One of the more innovative features integrated with eHOST is the ability to verify and produce recommendations that help human annotators comply with syntactic and lexical rules that are specified by annotation task guidelines. Ma-

chine-Assisted verification is most useful when used on lexical or syntax rules to ensure that candidate phrases generated by automated systems are similar to those marked by humans. These rules rely more on adherence to patterns than on decision-making, so the strengths of human review with machine approaches to semi-automated verification can be leveraged. When identifying medical concepts, it is common that noun phrases are marked as candidates. The determination of how much of a noun phrase to mark (inclusion of articles, adjectives, noun-modifiers, prepositional phrases) and at what granularity (simple nouns or complex noun phrases) may vary with each project.

The verifier allows portions of an annotation guideline to be programmed into rules that check for consistency. Rules check whether a word appears within a user-defined window before and after an annotation. Each rule can be linked to text that describes why the annotation was flagged. Annotators are then provided suggestions on the correct span based on the rule. Using the surrounding text, the guideline text, and the suggestion, the annotator can determine the final span for an annotation. These machine-assisted verifier functions help support reference standard generation by providing the context of annotations that seem to fail syntactic and lexical rules while allowing human annotators to focus on domain expertise required to identify and classify information found in clinical texts.

Conclusion

Our prototype system provides functionalities that have been created to more efficiently support reference standard generation including machine-assisted annotation approaches. It is our hope that these system features will serve as the basis for the further development efforts that will be part of an enterprise level system. Outputs of such an annotation tool could be used as inputs for pipeline NLP systems or as one component of a common workbench of tools used for clinical NLP development tasks.

We have implemented and tested eHOST for the 2010 and 2011 i2b2/VA challenge annotation tasks and annotation projects for the Consortium for Healthcare Informatics Research (CHIR). The stand-alone eHOST client tool is available from http://code.google.com/p/ehost along with a demonstration project, a users guide, API documentation, and source code. The eHOST/CASPR interfaces will be used to support a large-scale crowd sourced annotation task used for annotation of disorders, temporal expressions, uncertainty, and negation along with data standardization. These efforts will include more rigorous analysis and usability assessment of eHOST/CASPR for crowd sourcing and other small and large-scale annotation projects.

Acknowledgments

Support and funding was provided by the VA Salt Lake City HealthCare System and the VA Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374, the VA Informatics and Computing Infrastructure (VINCI), VA HIR 08-204, and NIH Grant U54 HL 108460 for integrating Data for Analysis, Annonymization and Sharing (iDASH), NIGMS 7R01GM090187.

References

Beatrice Alex, Claire Grover, Barry Haddow, Mijail Kabadjov, Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin, and Xinglong Wang. 2008. Assisted curation: does text mining really help? In: *Proceedings of the Pacific Symposium on Biocomputing*.

Wendy W. Chapman, David Chu, John N. Dowling. 2007. ConText: An Algorithm for Identifying Contextual Features from Clinical Text. In: *ACL-07 2007*.

Carol Friedman, Hongfang Liu, Lyudmila Shagina, Stephen Johnson, George Hripcsak. 2001. Evaluating the UMLS as a source of lexical knowledge for medical language processing. In: *Proc AMIA Symp, 2001: 189-93.*

Karen Fort and Saggot B. 2010. Influence of Pre-Annotation on POS-tagged Corpus Development. In: *Proceedings of the Fourth Linguistic Annotation Workshop. ACL 2010: 56-63.*

Pei-Yun Hsueh, Prem Melville, Vikas Sindhwani. 2009. Data Quality from Crowdsourcing: A study of Annotation Selection Criteria. In: *Proceedings of the NAACL HLT Workshop on Active Learning for Natural Language Processing. June 2009: 27-35*.

Danielle Mowery, Henk Harkema, Wendy W. Chapman. 2008. Temporal Annotation of Clinical Text. In: *ACL-08 2008*.

Mark A. Musen, John Gennari, Henrik Eriksson, Samson W. Tu, and Angel R. Puerta. 1995. PROTEGE-II: computer support for development of intelligent systems from libraries of components. In: *Medinfo 1995*.

Aurélie Névéol, Rezarta Islamaj-Doğan, Zhiyong Lu. 2011. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. In: *J Biomed Inform. 2011 Apr; 44(2):310-8*.

Stefanie Nowak and Stefan Ruger. 2010. How Reliable are Annotations via Crowdsourcing? A Study about Inter-Annotator Agreement for Multi-label Image Annotation. In: *MIR 10 2010*.

Philip V. Ogren. 2006. Knowtator a protege plug-in for annotated corpus construction. In: *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2006: 273-5*.

Philip V. Ogren, Guergana K. Savova, Christopher G. Chute. 2008. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. In: *Proceedings of the sixth international conference on Language Resources and Evaluation LREC 2008: 3143-3150*.

Angus Roberts, Robert Gaizauskas, Mark Hepple, Neil Davis, George Demetriou, Yikun Guo, Jay Kola, Ian Roberts, Andrea Setzer, Archana Tapuria, Bill Wheeldin. 2007. The CLEF corpus: semantic annotation of clinical text. In: *AMIA Annu Symp Proc, 625-9*.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, Andrea Setzer. 2009. Building a semantically annotated corpus of clinical texts. In: *J Biomed Inform, 42(5): 950-66*.

Ellen Riloff. 1993. Automatically constructing a dictionary for information extraction tasks. In: *Proceedings of the Eleventh National Conference on Artificial Intelligence, 811-816*.

Burr Settles, Mark Craven, and Lewis Friedland. 2008. Active Learning with Real Annotation Costs. In: *Proceedings of the NIPS Workshop on Cost-Sensitive Learning*. 2008.

Burr Settles. 2009. Active Learning Literature Survey. In: *Computer Sciences Technical Report 1648. University of Wisconsin-Madison. 2009*.

Brett R. South, Shuying Shen, F. Jeff Friedlin, Matthew H. Samore, and Stephane M. Meystre. 2010. Enhancing Annotation of Clinical Text using Pre-Annotation of Common PHI. In: *AMIA Annu Symp Proc. 2010*.

Brett R. South, Shuying Shen, Robyn Barrus, Scott L. DuVall, Ozlem Uzuner, and Charlene Weir. 2011. Qualitative analysis of workflow modifications used to generate the reference standard for the 2010 i2b2/VA challenge. In: *AMIA Annu Symp Proc. 2011*.

Özlem Uzuner, Imre Solti, Fei Xia, and Eithon Cadag. 2010. Community annotation experiment for ground truth generation for the i2b2 medication challenge. In: *J Am Med Inform Assoc, 2010. 17(5):519-23*.

Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2011. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. In: *JAMIA 18(5): 552-556*.

Özlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler B. Forbush, John Pestian, and Brett R. South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. In: *JAMIA doi: 10.1136/amiajnl-2011-000784*.

Aobo Wang, Cong Duy Vu Hoang, and Min-Yen Kan. 2010. Perspectives on Crowdsourcing Annotations for Natural Language Processing. In: *CSIDM Project No. CSIDM-200805*.

Alexander S. Yeh, Lynette Hirschman, and Alexander A. Morgan. 2003. Evaluation of text data mining for database curation: Lessons learned from the KDD challenge cup. In: *Bioinformatics, 19(Suppl 1): i331–339, 2003*.

CHAPTER 6

AIM 3: EVALUATING THE EFFECTS OF MACHINE PRE-ANNOTATION AND AN INTERACTIVE ANNOTATION INTERFACE ON MANUAL DE-IDENTIFICATION OF CLINICAL TEXT

[1,2,4]Brett R. South, MS, [5,6,7]Danielle Mowery, MS,

[2,4]Ying Suo, MS, [2,4]Jianwei Leng, MS, [3]Óscar Ferrández, PhD,

[1,2] Stephane M. Meystre, MD, PhD, [1,2,5,6,7] Wendy W. Chapman, PhD

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
[2] VA Health Care System, Salt Lake City, UT, USA
[3] Nuance Communications Inc., Burlington, MA, USA
[4] Department of Internal Medicine, University of Utah, Salt Lake City, UT, USA
[5] Department of Biomedical Informatics, University of Pittsburgh, PA, USA
[6] VA Health Care System, San Diego, CA, USA
[7] Division of Biomedical Informatics, University of California San Diego, La Jolla, CA, USA

Journal of Biomedical Informatics (submitted August, 2013, under review April, 2014)

6.1 Abstract

The Health Insurance Portability and Accountability Act (HIPAA) *Safe Harbor* method requires removal of 18 types of protected health information (PHI) from clinical documents to be considered "de-identified" prior to use for research purposes. Human review of PHI elements from a large corpus of clinical documents can be tedious and error-prone. Indeed, multiple annotators may be required to consistently redact information that represents each PHI class. Automated de-identification may improve annotation quality and reduce annotation time. For instance, using machine-assisted annotation by combining de-identification system outputs used as pre-annotations and an interactive annotation interface to provide annotators with PHI annotations for "curation" rather than manual annotation from "scratch" on raw clinical documents. In order to assess whether machine-assisted annotation improves the reliability and accuracy of the reference standard quality and reduces annotation effort, we conducted an annotation experiment. In this annotation study, we assessed the generalizability of the VA Consortium for Healthcare Informatics Research (CHIR) annotation schema and guidelines applied to a corpus of publicly available clinical documents called MTSamples. Specifically, our goals were to 1) characterize a heterogeneous corpus of clinical documents manually annotated for risk-ranked PHI and other annotation types (clinical eponyms and person relations), 2) evaluate how well annotators apply the CHIR schema to the heterogeneous corpus, 3) compare whether machine-assisted annotation (experiment) improves annotation quality and reduces annotation time compared to manual annotation (control), and 4) assess the change in quality of reference standard coverage with each added annotator's annotations.

6.2 Introduction

In most electronic medical record (EMR) systems, a great deal of relevant clinical information is stored in clinical documents. Clinical documents and other medical records data are rich in protected health information (PHI). Preserving a patient's privacy and confidentiality of PHI is fundamental to the physician-patient relationship. In order to use patient medical records for purposes other than providing health care (e.g., clinical research), informed consent from the patient is required. Indeed, use of patient medical record data is subject to the ethical and legal considerations defined by the Health Insurance Portability and Accountability Act (HIPAA) codified as 45 CFR §160 and 164 and the Common Rule [1]. However, obtaining the informed consent of a large population of patients, especially for retrospective research is difficult, time-consuming, and costly. This requirement can be waived if clinical documents are de-identified (i.e., all information identifying the patient has been redacted). Although de-identification of clinical documents remains a significant challenge, fulfilling these ethical and legal requirements is often a necessary step prior to using them for clinical research. However, manually de-identifying clinical documents represents a considerable expense in terms of time and human workload.

Automated methods that apply natural language processing (NLP) techniques may reduce the time and effort required to manually de-identify clinical documents, especially for large-scale projects applied to tens of thousands of patient records in which manual redaction of PHI is impractical. An NLP de-identification system must accurately remove the 18 types of PHI identifiers specified under the HIPAA *Safe Harbor* method for clinical documents to be considered "de-identified." NLP systems that de-identify

clinical documents are readily available [2-17], but are often developed and evaluated using specific document types. The approaches used by these systems may not be generalizable to all document types due to document specific formatting, clinical sublanguages, and prevalence of PHI [2]. Indeed, there is always the possibility that even with "de-identified" documents a PHI identifier may slip by and not be removed by all review methods [18].

A combined approach may reduce the likelihood of missing PHI identifiers and achieve acceptable coverage for certain PHI types by combining the efforts of many human reviewers with the outputs of an NLP system used as pre-annotations [19, 20, 21]. By leveraging NLP system outputs, this approach could offer a lower cost solution by pre-annotating potential PHI identifiers that human annotators review i.e., modifying existing, adding missing, or deleting spurious machine annotations. However, with any human review task relying on understanding of guidelines and tools, the cost of manual effort is high and may produce marginal returns of improved coverage as additional reviewers are added. The number of judges required to achieve acceptable coverage may also correlate with the risk of re-identification for different PHI types. In this study, we evaluate the effects of a combined machine pre-annotation plus interactive annotation interface used to de-identify clinical documents from a publicly accessible document corpus called MTSamples. This heterogeneous clinical document corpus was selected for this study because it is a publicly available data source that could be easily obtained without a rigorous institutional data release process and contains replaced PHI mentions in context ("Dr. Sample Doctor...") that are useful for de-identification research.

6.3 Background

Creating a reference standard that adequately identifies all HIPAA PHI identifier types and provides accurate training and evaluation data is imperative for developing rule-based or machine-learning-based de-identification systems. A few NLP researchers have championed efforts to facilitate the creation of state-of-the-art de-identifications for clinical documents and evaluate such systems against a standard corpus [16]. In 2006, NLP researchers from the University of Albany and MIT CSAIL sponsored the 2006 i2b2 Challenge task for automatic de-identification of clinical documents. A corpus of 889 discharge summaries from Partners Healthcare was annotated in two phases. In phase 1, PHI of eight types – patient names, doctor names, hospital names, IDs, dates, locations, phone numbers, and ages – were pre-annotated using an automated de-identification system that applied machine learning approaches [17]. In phase 2, three annotators sequentially annotated each report using a serial review method and achieved consensus after each review round. The interannotator agreement (IAA) between annotators and the performance of the NLP de-identification system was not reported as part of the 2006 i2b2 Challenge [16].

In contrast to the 2006 i2b2 Challenge, the goal for our manual de-identification task was to estimate the effects of machine pre-annotations and an interactive annotation interface on human annotator performance and quality of generated data for a heterogeneous clinical document corpus. We compare and contrast between annotators and the generated reference standard using IAA and standard performance metrics (i.e., recall, precision, and F1-measure) to assess annotator task consistency and accuracy. The effects of pre-annotation on the quality of annotated data has been investigated in many

studies that include annotation of medical literature [20], POS tagging [19], Named Entity Recognition (NER) [22] and Clinical Named Entities [23, 24], as well as common PHI types [25]. Other studies have employed semi-automated annotation methods that produce machine-generated candidate spans presented in such a way that the human reviewer must either modify incorrect annotations, delete spurious annotations, or add missed annotations [26, 27, 28]. It was our goal to produce a corpus of clinical documents annotated for PHI that maximized annotation quality while minimizing annotation effort.

6.4 Methods

We begin by describing the annotated MTSamples corpus. Next, we describe an annotation experiment including the annotation schema and training process. We further detail our annotation training, experiment, and evaluation approaches.

6.4.1 Medical Transcription Samples Corpus

A document sample consisting of 2,330 unique clinical documents was obtained from a publicly available resource of clinical documents called MTSamples (Medical Transcription Samples at www.mtsamples.com). These clinical documents were originally created to train medical coders and transcriptionists. The sample corpus contains document samples from 40 different medical specialties – consults, discharge summaries, and specialized medical services – including some uncommon formats. Although the MTSamples corpus does include data representing most of the 18 types of PHI identifiers specified under the HIPAA regulation, names and dates that remain have been changed (or removed) to preserve confidentiality of the users providing the data.

6.4.2 Annotation Schema

We build upon previous efforts [29] by expanding PHI types defined as part of the 2006 i2b2 challenge [16] and definitions for the Veteran Affair's (VA) setting using an annotation schema and guidelines originally developed as part of the VA Consortium for Healthcare Informatics Research (CHIR) De-identification project [8,11]. These annotation guidelines go beyond the PHI types annotated from the 2006 i2b2 Challenge. We include annotation types representing clinical eponyms, organization names, military deployments, health care units, and coreferring-paired relationships between annotations for person names (Table 6.1). For example, "Patient **Joe Smith**…and Mr. **Smith…**", "**Joe Smith**" and "**Smith**" might refer to the same person, in which case they would be linked in a paired relationship.

Our motivation to include annotation of clinical eponyms was twofold. First, we wished to measure human performance identifying clinical information that machine systems may misclassify as PHI. Second, we wished to enrich available data sources training classifiers and methods to identify these information types. Human reviewers more easily identify this type of information than machines because the reviewer can take into account contextual cues that may not be integrated with machine learned systems. We show a logical representation of these annotation types in Figure 6.1. Our annotation schema defines annotation types categorized by PHI privacy risk ranking: high risk, medium risk, low risk, and non-PHI. As mentioned previously, co-referring paired relationships were created between annotations for person names (Patients, Relatives, Health Care Providers, Other Persons).

Table 6.1: Annotation Type Definitions Between i2b2 and Extended CHIR Schema

i2b2 PHI Types [16]	**Definitions**	**CHIR Annotation Types** [8,11]	**Definition**
Dates	all elements of a date except for the year	*Dates*	date, including year and/or time, and specific time of day. Ex. "clinic on **Jul 4, 2001@01:00**"
Patients	first and last names of patients, their health proxies, and family members	*Patient Names**	patient's first name, last name, middle name, and initials excluding salutations. Ex. "Mr. **Smith** complains of cough"
		*Relative Names**	proper name of relatives. Ex. "patient's daughter **Jennifer**"
		*Other Person Names**	other persons mentioned or patient proxy. Ex. "lived in his friend **Mike's** place"
Doctors	medical doctors and other practitioners as well as transcriber's name and initials	*Health Care Provider Names**	health care worker's first name, last name, middle name, and initials excluding salutations Ex. "**JONES, JANE** MD"
Ages	ages above 90	*All mentions of age*	ages above 90 (expanded to include all mentions of age including duration of time. Ex. "**52-year-old** man")
IDs	any combination of numbers, letters, and special characters identifying medical records, patients, doctors, or hospitals	*Other ID Numbers*	all combinations of numbers and letters that could represent a medical record number, lab test number, or other patient or provider identifier such as driver's license number. Ex. "Driver's license: **S-012-34567**"
		Electronic Addresses	electronic mail addresses and references to personal Websites, Facebook pages, Twitter. Ex.
			"CC: **smarty@yahoo.com**"
		Social Security Numbers	numbers and/or characters, that could represent a social security reference. Ex. "SSN is **000-00-0000**"

*Annotation types having co-referring relationships.

Table 6.1: Continued

i2b2 PHI Types [16]	**Definitions**	**CHIR Annotation Types** [8,11]	**Definition**
Locations	geographic locations such as cities, states, street names, zip codes, building names, and numbers	*Street City*	street or city names excluding name as part of organization name. Ex. "lived on **5 Main Street**"
		State Country	state or country. Ex. "lived in **Alaska**"
		Zip codes	all digits acting as a zipcode. Ex. "works in **08536** area"
		Deployments	a specific geographic location, or mention of unit, battalion, regiment, brigade etc. Ex. "deployed with the **81st infantry unit**"
Hospitals	names of medical organizations and nursing homes where patients are treated and may also reside including room numbers of patients, buildings and floors related to doctors' affiliations	*Other Organizations Names*	affiliation with companies such as employment that are not related to health care. Ex. "employed by **WalMart**"
		Health Care Unit Names	sub-specialty clinics, consults or referral to services, or recommendations from services where health care was or will be provided to a patient. Ex "Care provided at **VA SALT LAKE CITY HCS**"
Phone Numbers	telephone, pager, and fax numbers	*Phone Numbers*	phone/fax/pager numbers including phone number extensions. Ex. "Fax no: **381-7777**"
Non-PHI	Not annotated as part of i2b2	*Clinical eponyms part of medical procedure names*	medical procedures that contain proper names of persons, places, or locations. Ex. "**DeLuca** criteria was used"
Non-PHI	Not annotated as part of i2b2	*Clinical eponyms part of medical device names*	medical devices that contain proper names of persons, places, or locations excluding brand names. Ex. "**Foley** catheter"
Non-PHI	Not annotated as part of i2b2	*Clinical eponyms part of medical disease names*	diseases that contain proper names of persons, places, or locations. Ex. "history of **Crohn's** disease"
Non-PHI	Not annotated as part of i2b2	*Clinical eponyms part of anatomic structures*	anatomic locations that have proper names of persons, places, or locations. Ex. "**Achilles** tendon"

*Annotation types having co-referring relationships.

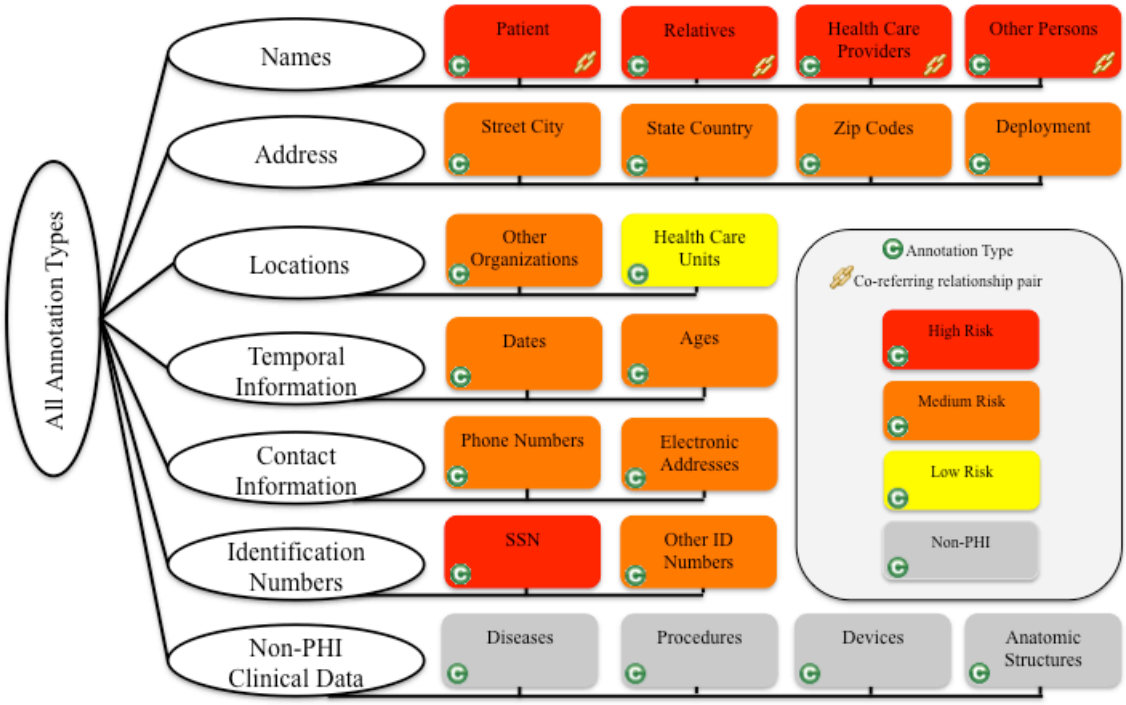

Figure 6.1: Annotation Schema De-Identification of Clinical Texts

For each annotation type, we developed detailed guidelines specifying inclusion and exclusion criteria regarding what to mark and not mark, which tokens to include, and what type of information should be marked. Annotations were defined using a contiguous span, beginning at the start of a phrase and ending at the completion of the phrase to capture instances rather than individual word tokens.

6.4.3 Experimental Design

Manual annotation can be a slow, laborious process. We performed an experiment to determine the effects of combining machine pre-annotations with an interactive annotation interface. It was our goal to maximize annotation quality while minimizing

manual annotation effort. We also wished to limit confusion or uncertainty related to annotator training on the guidelines, schema, and tool while maximizing the number of documents annotated from the original 2,330 MTSamples document corpus. This was achieved by separating the annotation of the MTSamples corpus into annotator training and experiment. A stratified random sample was obtained for both training and experiment based on document type and the number of lines, words and tokens found in each clinical document. During the annotator training, 7 reviewers annotated a random sample of 350 documents divided into 15 batches of 20-25 documents. Annotator training continued until a reviewer either exhausted their supply of training documents/batches or achieved a predefined IAA performance threshold of 75% or greater when compared to other annotators on the training corpus. During the annotator experiment, the same 7 reviewers annotated another random sample of 1,535 documents divided into 15 batches of 35 documents. Each annotator reviewed a total of 525 documents with 1,229 (80%) of these annotated by 2 or more reviewers. For both annotator training and experiment, annotators applied the same guidelines and annotation schema using an annotation tool called the extensible Human Oracle Suite of Tools (eHOST) [27]. Following the annotator training and experiment, a final reference standard was created after adjudicating discrepancies and consensus review of the resulting annotations from all reviewers.

During the annotator experiment, we employed two types of machine-assisted annotation: 1) machine pre-annotations (pre-annotations generated using an out-of-the-box de-identification system) and 2) interactive annotation (interactive annotation interface integrated with the annotation tool) (Figure 6.2). We hypothesized that a

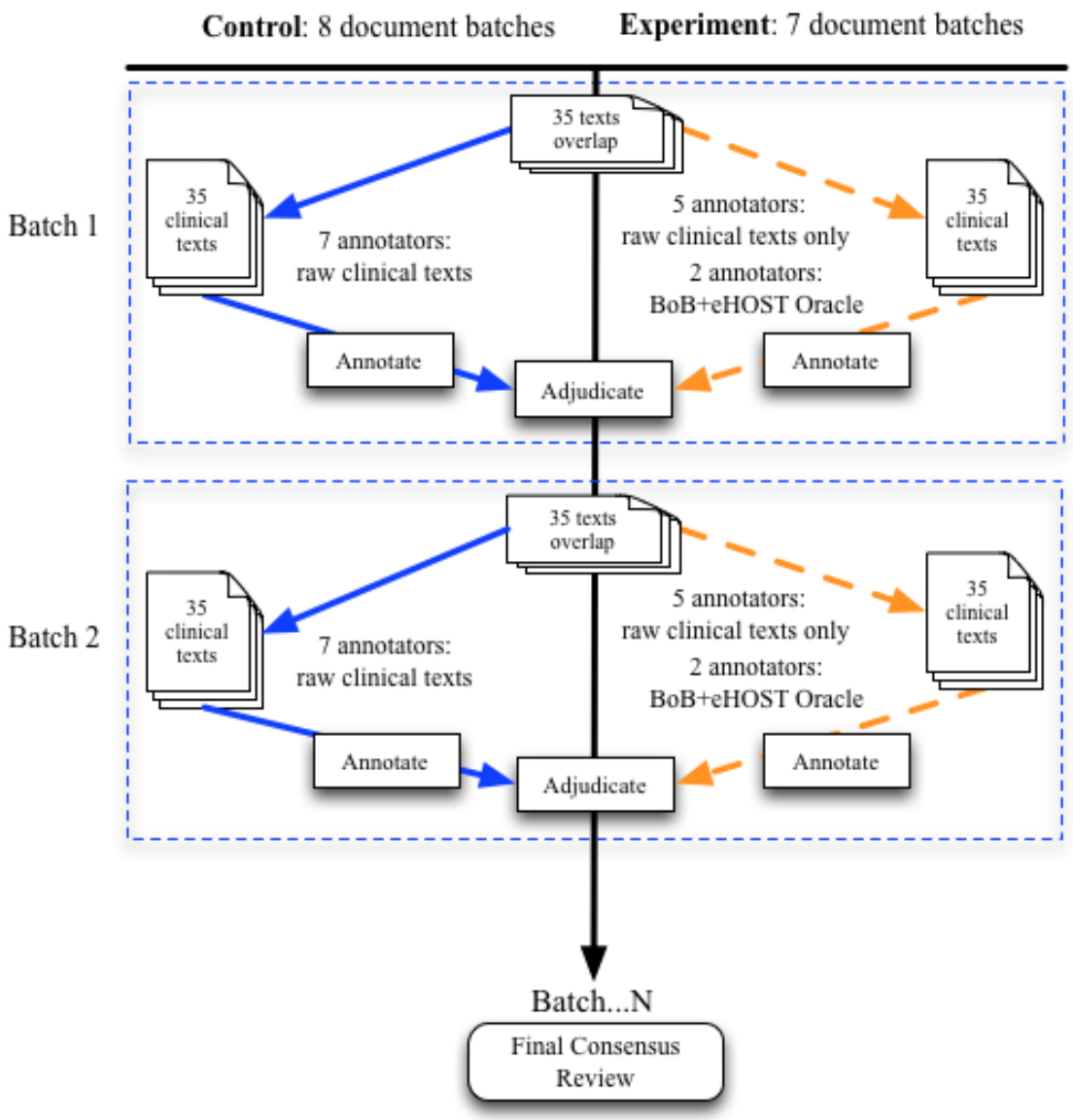

Figure 6.2: Annotation Experiment Conditions De-Identification of Clinical Texts

combined approach using machine pre-annotations and an interactive annotation interface would reduce the time required for manual annotation of annotation types defined by our schema and found in clinical texts and would not reduce the quality of the data annotated. We used an "out-of-the-box" version of a de-identification system called BoB to generate pre-annotations, and a function integrated with the eHOST tool called "the Oracle" to provide the interactive annotation interface.

6.4.4 BoB: De-Identification System Pre-Annotations

One automated de-identification system designed for clinical documents is the Veterans Affairs "Best-of-Breed" (BoB) de-identification system [8]. BoB is a hybrid system that integrates known high-performing approaches specific to each particular PHI type from existing rule-based and machine learning systems. BoB is developed on a UIMA framework and processes documents using two main components: a high-sensitivity extractor and a false positive filterer. The high-sensitivity extractor applies a conditional random field classifier and rules to identify all potential PHI annotations maximizing recall. The false positive filterer leverages a support vector machine classifier to reclassify incorrectly tagged PHI annotations maximizing precision. For instance, the filterer may reclassify clinical eponyms such as *Anatomical Structures* e.g., "Circle of **Willis**" as non-PHI. We processed the MTSamples corpus using an out-of-the-box version of BoB originally trained on VA document types to generate pre-annotations provided to annotators during the experiment. Under these conditions, the annotation task is modified slightly and the human annotator accepts correct pre-annotations, modifies incorrect spans and deletes incorrect pre-annotations. We evaluate how helpful the system outputs could be without additional training on the MTSamples document corpus,

a reality most researchers face when obtaining any open-source, de-identification software. We report recall, precision and F1-measure and provide baseline "out-of-the-box" performance of BoB without domain adaptation on the MTSamples corpus.

6.4.5 The eHOST Oracle: Machine-Assisted Interactive Annotation

One function integrated with the eHOST tool is a module called "the Oracle." When enabled, the Oracle provides new annotation suggestions to the annotator based on an exact string match of the last human reviewer-produced annotation corresponding with that annotation type. For instance, if an annotator spans "**Jane**" as a *Patient Name*, the Oracle can search either the current document or across an entire batch of documents for other spans of "**Jane**" and then present these as potential candidate spans representing *Patient Name*. The annotator can choose to accept or reject these candidate PHI annotations. Annotators completed the annotator training using the eHOST Oracle module and were accustomed to its functionality before starting the annotator experiment discussed later. We report annotator utilization of the eHOST Oracle module in comparison to the total number of annotations generated during the experiment.

6.4.6 Annotation Prevalence

We characterized the final reference standards generated from the annotator training and experiment. We report prevalence and performance metrics for all annotation types according to the following ranking of re-identification risk in the case where PHI is potentially missed.

- **High Risk:** Social Security Numbers, Patient Names, Relative Names, Other Person Names, and Health Care Provider Names.

- **Medium Risk:** Dates, Street City, State Country, Zip Codes, Deployments, Other Organization Names, Other ID Numbers, Phone Numbers, Electronic Addresses and Ages.
- **Low Risk:** Health Care Units.
- **Non-PHI:** Clinical Eponyms *(Anatomic Structures, Devices, Diseases, Procedures*) and Person Relations.

6.4.7 Annotator Performance Metrics

We evaluated interannotator agreement (IAA) using F1-measure as a surrogate for Kappa since the number of document strings not annotated as a PHI annotation or true negatives (TN) are unknown [30]. We applied three types of annotator comparisons using standard performance metrics:

- **BoB-Reference Standard:** compare BoB-generated pre-annotations and the reference standard generated during the annotator training using average exact performance metrics (i.e., recall, precision, F_1-measure).
- **Annotator-Annotator:** compare average paired exact and partial IAA between annotators.
- **Annotator-Reference Standard:** compare average exact and partial performance metrics (i.e., recall, precision, F_1-measure) between annotators and the annotator experiment reference standard.

F1-measure was calculated using the harmonic mean of recall (TP/TP+FN) and precision (TP/TP+FP) defined as 2 ((recall * precision)/ (recall + precision)). For example, during Annotator-Reference Standard comparisons we defined:

- **True Positive (TP):** an annotation that exactly (exact) or partially (partial) overlapped a reference standard annotation for the same annotation type.
- **False Positive (FP):** an annotator's annotation that did not occur as a reference standard annotation.
- **False Negative (FN):** a reference standard annotation that did not occur in the annotator's annotations.

6.4.8 Annotation Experiment

For the experiment, we assessed whether human annotators provided with machine pre-annotations and an interactive interface could generate similar quality data for span and classification of annotation type than without machine pre-annotations plus an interactive interface. We created two versions of the corpus – one with and one without BoB pre-annotated machine annotations and two versions of the eHOST tool – one with and one without the eHOST Oracle module. Annotators were randomly assigned 7 batches with pre-annotations plus the interactive annotation interface and 8 batches without (Figure 6.2). For each annotation type and PHI risk of re-identification ranking, we evaluated whether human annotators receiving pre-annotations plus the interactive interface (experiment = BoB+eHOST Oracle) were able to generate data of similar quality with human annotators that did not receive the experiment (control = raw annotation). For this evaluation, we used the Wilcoxon Rank Sum test (Mann-Whitney U) [31]. The Wilcoxon Rank Sum test is a nonparametric test equivalent to a parametric 2-sample T-test for determining whether median F_1-measures for the experiment are different than medians of the control (calculated for each annotation type and stratified by

PHI risk ranking). For significance testing, independent T-tests were used to determine if there were differences in averaged F_1-measure between control and experiment for each annotator on each clinical document. For all statistical analyses, we used a null hypothesis stating there was no difference between the control and experiment using a significance level of 0.05. We calculated statistics (mean, median, and quartiles) and significance tests using SAS version 9.3.

6.4.9 Time Comparison

Next, we hypothesized that human annotators receiving the experimental condition (BoB+eHOST Oracle) could produce annotations in less time (seconds) than human annotators under the control condition (annotation on raw clinical documents). We compared the average time per annotation for the experiment and control conditions. These calculations were made using the mean time between annotation spans using the timestamp for each annotation type classification within each document.

6.4.10 Coverage Differences with Added Annotators

Finally, since our goal was to maximize annotation quality while minimizing annotation effort, we wanted to estimate how adding additional reviewers would affect recall, precision, and F_1-measure. During the annotator training, we assessed the effects of annotation coverage as a function of adding additional reviewers. All 7 reviewers annotated the same 350 documents. Discrepant annotations were adjudicated and a final consensus review was conducted to create a reference standard after the completion of both annotator training and experiment.

6.5 Results

We characterized prevalence of each PHI risk ranking and annotation type by training and experiment. For the annotator experiment, we report performance metrics for BoB compared with the reference standard generated for annotator training. We also report averaged IAA for annotators during the annotator experiment, and performance metrics for annotators compared with the reference standard generated at the completion of the annotator experiment. We compared distributions of the final annotations produced by experimental and control conditions and report the effects of the experiment applying the Wilcoxon Rank Sum test. Time savings introduced by the experiment were also calculated. We also determined the coverage differences for each added annotator based on the annotation training reference standard. Finally, we report the distribution of each annotation type generated during the training and experiment using the complete annotated corpus.

6.5.1 Annotation Prevalence

The majority of documents were annotated during the experiment. Discordant annotations generated from the training and experiment were adjudicated and subjected to a final consensus review. We characterized the prevalence of annotations by PHI risk category and annotation type found in the final reference standard in Table 6.2. PHI categorized as medium risk had the highest prevalence for both annotator training and experiment; PHI categorized as high risk had the lowest prevalence for training and experiment. Counts are expanded by annotation type for each collapsed risk ranking. For each PHI risk ranking, the most common PHI types represented *Health Care Provider Names* for high risk, *Dates* for medium risk, and *Healthcare Unit Names* for low risk.

Table 6.2: Prevalence of Annotation Types and PHI Risk Categories

Annotation Prevalence Training and Experiment				
	Annotator Training		**Annotator Experiment**	
	N	**%**	**N**	**%**
Documents reviewed	**350**	**18.6**	**1,535**	**81.43**
Annotation Type				
High Risk	**311**	**12.8**	**1,135**	**11.2**
Social Security Numbers	--	--	--	--
Patient Names	86	3.5	248	2.5
Health Care Provider Names	204	8.4	860	8.5
Relative Names	17	<1.0	12	<1.0
Other Person Names	4	<1.0	15	<1.0
Medium Risk	**1,220**	**50.2**	**4,357**	**43.2**
Dates	630	26	2,305	22.8
Street City	24	1	119	1.2
State Country	33	1.4	95	1
Zip codes	--	--	--	--
Phone Number	2	<1.0	6	<1.0
Deployments	2	<1.0	1	<1.0
Other Organization Names	49	2	109	1.1
Electronic Addresses	--	--	--	--
Other ID Numbers	4	<1.0	178	1.8
Ages	476	19.6	1,544	15.3
Low Risk	**110**	**9**	**469**	**4.6**
Health Care Unit Names	110	9	469	4.6
Non-PHI	**661**	**27.1**	**2762**	**27.4**
Clinical Eponyms	**661**	**27.1**	**2762**	**27.4**
Anatomic Structures	44	1.8	164	1.6
Devices	412	16.9	1,622	16.1
Diseases	48	2	263	2.6
Procedures	157	6.5	713	7.1
Person Relations	**129**	**5.3**	**456**	**4.5**
Health Care Provider Names relations	66	2.7	287	2.8
Patient Names relations	61	2.5	167	1.66
Relative Names relations	2	<1.0	2	<1.0
Total Annotations	2,431	19.4	10,091	80.6
Overall	**12,522**			

The most prevalent clinical eponyms were medical *Devices*. It is important to note that paired relations between person relations were quite common (5% within the entire annotated corpus); the most prevalent were *Health Care Provider Names* and *Patient Names*. Some PHI types, *Social Security Numbers, Zip Codes,* and *Electronic Addresses,* did not occur in the MTSamples data.

6.5.2 BoB-Reference Standard Performance Metrics

Baseline performance for out-of-the-box BoB pre-annotations on the MTSamples experiment corpus using standard performance metrics (recall, precision and F_1-measure) was low when microaveraged across all annotation types (0.20, 0.42, 0.27), moderate on medium risk (0.44 0.48, 0.46), but very low for high risk (0.17, 0.04, 0.07) and low risk (0.10, 0.76, 0.18) PHI types. This is in contrast to the published overall microaveraged performance of BoB trained on VA clinical documents averaged across all PHI types (0.92, 0.86, 0.86) [8]. Highest performance on BoB pre-annotations on the MTSamples corpus was for *Dates* (0.78, 0.80, 0.79), followed by *Other ID Numbers* (0.34, 0.25, 0.29) and *State Country* (0.85, 0.18, 0.29). BoB's lowest performance was on *Other Person Names* (1.0, 0.04, 0.09). There were a total of 8,181 BoB pre-annotations provided to annotators across the experiment document corpus and over half of these were false positive annotations, 67% with only 16% (2,899) of these left unmodified prior to final adjudication and consensus review. Indeed, human annotators were more likely to delete BoB pre-annotations than modify or accept them. The majority of false positive annotations introduced by BoB pre-annotations were clinical eponyms that were incorrectly classified as *Health Care Unit Names* 21% (1,740) and *Other Person Names* 27% (2,237). The majority of false negative annotations corresponded with *Ages* 10%

(850) and *Dates* 3.5% (285).

Annotators used the eHOST Oracle for only 640 (3.6%) annotations out of the total 17,643 annotations generated by all 7 annotators in the experiment. Out of these annotations the eHOST Oracle was used to mark 243 clinical eponyms, 145 *Ages*, 120 proper names of persons, and 104 *Dates*. Which is not surprising since these types of annotations can easily be found using exact string matching and some are highly prevalent (Clinical Eponyms, *Ages*, *Dates*) in the MTSamples corpus. The eHOST Oracle produced only 16 false positive annotations (<1%), on those annotations where it was used.

6.5.3 Annotator-Annotator Agreement

For all annotation types (Table 6.3), agreement was moderate for exact IAA (control 0.75; experiment 0.66) and slightly higher for partial IAA (control 0.79; experiment 0.69). For each PHI risk ranking, both exact and partial IAA was higher for annotation on raw documents, ranging from moderate IAA for low risk PHI to high IAA for medium and high risk PHI. For *Person Relations*, the experiment condition produced higher IAA than the control. Interannotator agreement on raw document annotation ranged from low (*Other ID Numbers, Deployments,* and *Other Person Names*) to moderate (*Phone Numbers, Other Organization Names, Health Care Unit Names*, and all clinical eponyms) to high (all other types). Agreement on experiment documents ranged from low (*Relative Names, Phone Numbers, Other Organization Names,* and *Other Person Names*) to moderate (*Street City, State Country, Other ID Numbers, Health Care Unit Names*, and most clinical eponyms) to high (all other types). It is worth noting that

Table 6.3: Inter-Annotator Agreement

Inter-Annotator Agreement (IAA) (Experiment)				
	Exact (IAA)		**Partial (IAA)**	
	Control: Raw Annotation	Experiment: BoB+eHOST Oracle	Control: Raw Annotation	Experiment: BoB+eHOST Oracle
Annotation Type	**0.75**	**0.66**	**0.79**	**0.69**
High Risk	**0.9**	**0.73**	**0.95**	**0.75**
Social Security Numbers	--	--	--	--
Patient Names	0.87	0.4	0.91	0.8
Health Care Provider Names	0.9	0.91	0.95	0.92
Relative Names	0.8	0	0.8	0
Other Person Names	0.33	0.1	0.33	0.11
Medium Risk	**0.81**	**0.76**	**0.85**	**0.6**
Dates	0.84	0.75	0.86	0.76
Street City	0.82	0.44	0.84	0.44
State Country	0.78	0.35	0.79	0.46
Zip codes	--	--	--	--
Phone Numbers	0.5	0	0.5	0
Deployments	0.33	--	0.33	--
Other Organization Names	0.61	0.3	0.64	0.39
Electronic Addresses	--	--	--	--
Other ID Numbers	0.07	0.6	0.15	0.6
Ages	0.84	0.83	0.92	0.89
Low Risk	**0.5**	**0.5**	**0.54**	**0.55**
Health Care Unit Names	0.5	0.5	0.54	0.55
Non-PHI	**0.64**	**0.63**	**0.67**	**0.65**
Clinical Eponyms	**0.64**	**0.63**	**0.67**	**0.65**
Anatomic Structures	0.67	0.55	0.68	0.59
Devices	0.68	0.76	0.72	0.77
Diseases	0.62	0.67	0.65	0.67
Procedures	0.55	0.4	0.56	0.45
Person Relations	**0.6**	**0.91**	**0.62**	**0.95**

both exact (control 0.60; experiment 0.91) and partial (control 0.62; experiment 0.95) IAA was higher for person relations generated under the experimental condition.

6.5.4 Annotator-Reference Standard Performance Metrics

We report performance metrics (recall, precision, and F1-measure) using the reference standard generated during the annotation experiment (Table 6.4). We observed high exact recall (control 0.82, experiment 0.80), precision (control 0.91, experiment 0.81), and F1-measure (control 0.86, experiment 0.81) between annotators, with improved partial recall (control 0.84, experiment 0.84), precision (control 0.94, experiment 0.85), and F1-measure (control 0.89, experiment 0.84). For each PHI risk category, similar to Annotator-Annotator performance, both exact and partial metrics were higher when annotating on raw clinical documents. These differences were statistically significant for all annotation types between control (0.84, +/- 0.211) and experiment (0.81, +/- 0.255), $t(3.13) = 1363.5$, $p=0.0018$.

6.5.5 Annotation Experiment

We evaluated whether annotators provided with machine pre-annotations plus an interactive interface (experiment) produced annotations and annotation type classification of similar quality as compared to annotators reviewing raw clinical texts (control). In Table 6.5, we show summary statistics for the control and experimental conditions by annotation type stratified by PHI risk category computed from the Wilcoxon Rank Sum test. Significant differences were observed when comparing raw annotation (control) and annotation using BoB+eHOST Oracle (experiment) for *Patient Names*, *Other Person*

Table 6.4: Performance Metrics for Control and Experimental Conditions

Performance Metrics Annotator (Experiment)				
	Exact (recall, precision, F1-measure)		**Partial (recall, precision, F1-measure)**	
	Control: Raw Annotation	**Experiment: BoB+eHOST Oracle**	**Control: Raw Annotation**	**Experiment: BoB+eHOST Oracle**
Annotation Type	**0.82, 0.91, 0.86**	**0.80, 0.81, 0.81**	**0.84, 0.94, 0.89**	**0.84, 0.85, 0.84**
High Risk	**0.94, 0.96, 0.95**	**0.87, 0.74, 0.80**	**0.96, 0.98, 0.97**	**0.93, 0.78, 0.85**
Social Security Numbers	--	--	--	--
Patient Names	0.95, 0.98, 0.96	0.78, 0.85, 0.81	0.96, 0.99, 0.98	0.91, 0.99, 0.95
Health Care Provider Names	0.94, 0.96, 0.95	0.90, 0.96, 0.93	0.97, 0.98, 0.97	0.93, 0.99, 0.96
Relative Names	0.82, 0.93, 0.88	0.50, 0.50, 0.50	0.88, 1.0, 0.94	1, 1, 1
Other Person Names	0.50, 0.80, 0.62	0.69, 0.06, 0.11	0.50, 0.80, 0.62	0.81, 0.07, 0.13
Medium Risk	**0.85, 0.92, 0.88**	**0.82, 0.86, 0.84**	**0.88, 0.96, 0.92**	**0.86, 0.91, 0.88**
Dates	0.86, 0.95, 0.90	0.84, 0.93, 0.88	0.88, 0.97, 0.92	0.86, 0.94, 0.90
Street City	0.88, 0.92, 0.90	0.92, 0.50, 0.65	0.89, 0.93, 0.91	0.93, 0.51, 0.66
State Country	0.80, 0.94, 0.86	0.83, 0.50, 0.62	0.80, 0.95, 0.87	0.96, 0.57, 0.72
Zip codes	--	--	--	--
Phone Numbers	0.50, 0.71, 0.59	1, 1, 1	0.70, 1.0, 0.82	1, 1, 1
Deployments	0.67, 0.67, 0.67	--	0.67, 0.67, 0.67	--
Other Organization Names	0.69, 0.81, 0.74	0.61, 0.53, 0.57	0.72, 0.84, 0.77	0.67, 0.58, 0.62
Electronic Addresses	--	--	--	--
Other ID Numbers	0.37, 0.46, 0.41	0.36, 0.54, 0.44	0.54, 0.69, 0.61	0.53, 0.80, 0.64
Ages	0.90,0.93,0.91	0.89, 0.93, 0.91	0.94, 0.98, 0.96	0.93, 0.98, 0.95

Table 6.4: (Continued)

	Exact (recall, precision, F1-measure)	**Partial (recall, precision, F1-measure)**		
	Control: Raw Annotation	**Experiment: BoB+eHOST Oracle**	**Control: Raw Annotation**	**Experiment: BoB+eHOST Oracle**
Low Risk	**0.69, 0.75, 0.72**	**0.76, 0.54, 0.63**	**0.73, 0.80, 0.76**	**0.83, 0.59, 0.69**
Health Care Unit Names	0.69, 0.75, 0.72	0.76, 0.54, 0.63	0.73, 0.80, 0.76	0.83, 0.59, 0.69
Non-PHI	**0.75, 0.89, 0.81**	**0.74, 0.84, 0.96**	**0.76, 0.91, 0.83**	**0.75, 0.86, 0.96**
Clinical Eponyms	**0.75, 0.89, 0.81**	**0.74, 0.84, 0.96**	**0.76, 0.91, 0.83**	**0.75, 0.86, 0.96**
Anatomic Structures	0.77, 0.83, 0.80	0.64, 0.82, 0.72	0.78, 0.84, 0.81	0.65, 0.83, 0.73
Devices	0.77, 0.91, 0.83	0.79, 0.88, 0.83	0.79, 0.94, 0.86	0.81, 0.91, 0.85
Diseases	0.76, 0.87, 0.81	0.81, 0.79, 0.80	0.79, 0.91, 0.84	0.83, 0.81, 0.82
Procedures	0.69, 0.85, 0.76	0.62, 0.73, 0.67	0.69, 0.85, 0.76	0.63, 0.75, 0.68
Person Relations	**0.75, 0.93, 0.83**	**0.74, 0.89, 0.81**	**0.76, 0.94, 0.84**	**0.74, 0.90, 0.81**

Table 6.5: Experimental Effects Estimated Using the Wilcoxon Rank Sum Test

Wilcoxon Rank Sum Test					
	Control: Raw Annotation		Experiment: BoB+eHOST Oracle		Significance
	Median F1-measure	N	Median F1-Measure	N	Pr>|Z|
All Annotation Types	0.91	1156	0.91	741	0.296
High Risk	1	365	1	274	***<0.0001**
Patient Names	1	78	1	32	***0.0389**
Health Care Provider Names	1	338	1	201	0.278
Relative Names	1	8	0.5	2	0.553
Other Person Names	0.5	11	0	106	***<0.0001**
Medium Risk	1	879	1	579	0.0748
Street City	1	68	0	72	***<0.0001**
State Country	0.96	48	0	64	***0.0009**
Zip Codes	--	--	--	--	--
Deployments	0.33	2	--	--	--
Other Organization Names	0.5	72	0	65	***0.0319**
Dates	1	533	1	342	0.195
Ages	1	764	1	493	0.992
Phone Numbers	0.58	4	1	2	0.14
Electronic Addresses	--	--	--	--	--
Other ID Numbers	0.2	47	0	37	0.553
Low Risk	0.667	277	0	221	***0.0002**
Health Care Unit Names	0.667	277	0	221	***0.0002**
Non-PHI	0.857	729	0.995	459	0.7103
Clinical Eponyms					
Anatomic Structures	0.933	101	0.8	61	0.6
Devices	0.872	485	1	303	0.103
Diseases	0.919	116	1	66	0.784
Procedures	0.667	347	0.667	211	0.929
Person Relations	1	141	1	72	0.458

Names, *Relative Name*s, *Street City*, *State Country*, *Other Organization Names*, and *Health Care Unit Names*.

6.5.6 Time Comparison

There was no statistically significant difference when comparing times for annotation of raw clinical documents compared with annotation of documents annotated under the experimental conditions across all annotation types. Observed mean time in seconds per annotation was 13.7 seconds for annotation on raw clinical documents and 13.6 seconds for documents annotated using BoB+eHOST Oracle. Although these differences were not significant across all annotation types, the mean time between annotations generated using only the eHOST Oracle was 5.24 seconds.

6.5.7 Coverage Differences with Added Annotators

In Figure 6.3, we show the change in performance metrics as logical combinations of reviewers are compared. Recall ranged from 0.66 (1 reviewer) plateauing at a high of 0.92 (7 reviewers). Alternatively, precision decreased from 0.82 (1 reviewer), to a low of 0.61 for the union of all 7 judges. F_1-measures ranged from 0.73 (1 reviewer), 0.79 (2), 0.78 (3), 0.77 (4), 0.75 (5), 0.74 (6), and 0.73 (7 reviewers). Document level F_1-measure (not shown) by PHI risk ranking ranged from 0.20-1.00 (mean=0.96, std=0.12) for high risk, 0.11-1.00 (mean=0.89, std=0.17) for medium risk, and 0.07-1.0 (mean=0.81, std=0.22) for low risk.

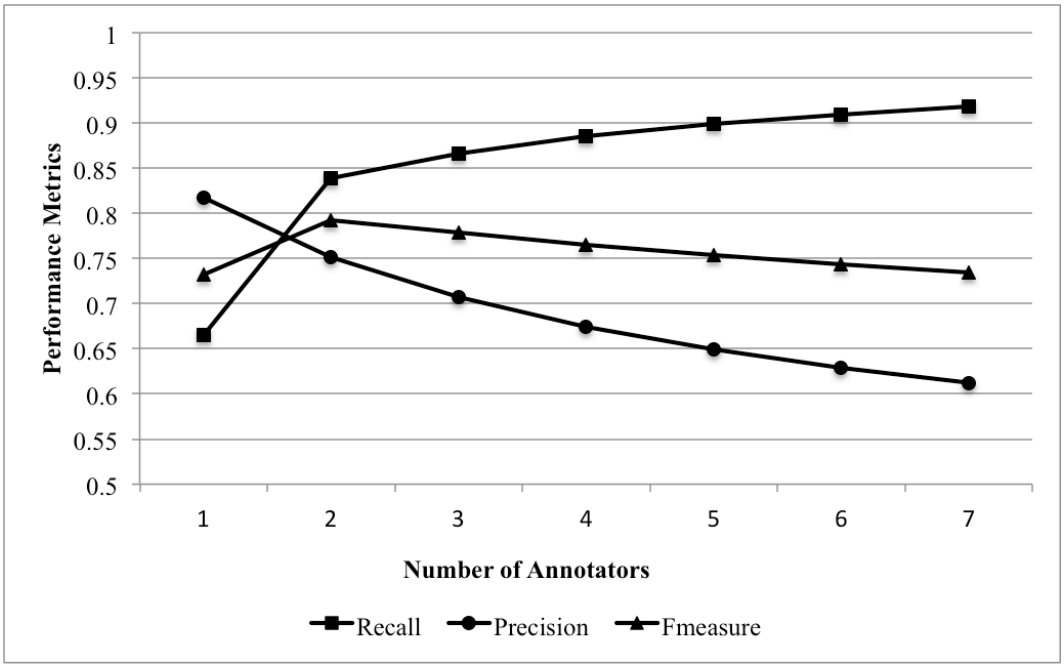

Figure 6.3: PHI Coverage Differences as a Function of Annotator Number

6.6 Discussion

6.6.1 Annotation Prevalence

For the annotator training and experiment the most prevalent PHI category included those PHI types categorized as medium risk and the least prevalent PHI types were those in the high risk. At the corpus-level, these prevalence estimates are difficult to compare with other published studies [8, 14, 16, 20, 27] due to the large variety of report types used in our study and the differences in annotation schema between studies. Our average prevalence of 4.0 PHI annotated per document for the MTSamples corpus is lower than those reported using other clinical corpora such as 26 per document from the VA [8], 22 from the 2006 i2b2 De-identification challenge [16], 8.79 per document from the 2012 Deleger et al. study [14], 49 from the 2013 Hanauer et al. study [20], and 7.9 per document from a 2006 study by Dorr et al. [32]. This difference could be due to the

varied PHI types and prevalence of document types annotated in these other studies. For instance, a general clinical document containing instructions for how a patient should continue to treat "**Athlete's foot**" would not contain PHI.

We should note that our corpus does contain clinical information that can be mistaken for PHI. Clinical eponyms are one such example accounting for 27.3% of the total corpus, which is significantly higher than previously observed using the same annotation schema on VA clinical documents 3.5% [33].

6.6.2 BoB-Reference Standard Performance Metrics

We observed low to moderate F_1-measure for predicting low to high risk PHI mentions. This is not surprising since this performance is based on pre-annotations produced by BoB previously trained using VA clinical documents and not on MTSamples documents. Even though the performance of the baseline pre-annotation system was poor we would expect (particularly for medium and low risk PHI types), that a combined approach using machine-generated pre-annotations plus the interactive annotation interface would result in improvements in the quality of annotated data and result in gains in annotation efficiency. This expectation was not borne out for the majority of annotation types in our study. Furthermore, annotators used the eHOST Oracle for only a small proportion of their annotations because they found it easier to annotate on raw clinical texts without the interactive machine suggestions. It is not clear whether this preference was due to the high number of false positives introduced by BoB or related to usability issues with the eHOST Oracle. However, annotator usability ratings of the eHOST Oracle based on the system usability scale (SUS) [34] were slightly above average. Despite this preference the number of false positives introduced using the

eHOST Oracle was very small compared with the number of false positives introduced by BoB. Although the Oracle was not specifically designed for relations it was used to annotate one or both entities in coreferring relation pairs for annotation types representing proper names of persons. This is interesting since identifying a coreferring pair first involves identifying the entities that should be linked.

6.6.3 Annotator-Annotator Agreement

When viewed in aggregate for each PHI risk category, raw annotation on clinical texts produced the highest interannotator agreement. The combination of BoB+eHOST Oracle introduced false positives producing less reliable annotation between annotators. However, these false positives were introduced in the majority of cases due to the low baseline performance of the BoB outputs used as pre-annotations and not via annotator interaction with the eHOST Oracle. Moreover, even though not statistically significant, use of the eHOST Oracle produced higher quality data when building relation pairs between person names. Person relations IAA was higher where BoB+eHOST was used due to the high IAA for *Health Care Provider Names* and *Patient Names* (particularly for partial IAA) and their high prevalence in the corpus. We were not surprised to observe less prevalent PHI types like *Relative Names* and *Deployments* had the lowest IAA. Introducing more training instances could boost IAA performance for these types.

6.6.4 Annotator-Reference Standard Performance Metrics

Standard performance metrics demonstrated similar results with the control condition producing higher quality data among all PHI risk categories as demonstrated in Table 6.4 and the Wilcoxon Rank Sum test in Table 6.5.

6.6.5 Annotation Experiment

There are several lessons we learned from integrating a combined approach using outputs from an untrained de-identification system along with an interactive interface. First, the experimental condition did not introduce significant gains in recall, precision, and F_1-measure. This is surprising since particular annotation types including clinical non-PHI can easily and consistently be found using the eHOST Oracle since they follow standard naming conventions and were often flagged as false positive BoB pre-annotations (i.e., clinical eponyms and *Other Organization Names*). Annotation on raw clinical texts produced higher quality data across all annotation types when compared with the experiment. For some annotation types (i.e., *Other Person Names*, *Health Care Unit Names*), annotator agreement remained lower than expected throughout the experiment and never plateaued. In the best of all possible experiments annotators would train until their agreement meets or exceeded some pre-defined threshold. This brings us to several remaining questions reserved for future experimentation. First, we did not explore how applying a "tag a little, learn a little" approach could be implemented in a practical way [20]. Second, we did not explore "how high" system performance should be to optimize annotator performance e.g., would higher performing pre-annotation with precision and/or recall greater than 50% produce better results instead of the out-of-the-box application of BoB?

The methods used for this annotation task could be modified to fit annotation of other types of information commonly found in clinical texts including clinical entities. However, caution should be used when pre-annotation or machine-assisted methods are employed as a means to improve the quality of generated data or reduce the time required to generate annotated data. This is particularly true when an untrained system is used out-

of-the-box to produce pre-annotations with no domain adaptation. On the one hand, providing pre-annotated information derived from system outputs may result in human annotators either trusting the pre-annotations too much in the case where system outputs are highly precise or missing incorrect annotations when system outputs produce results of high recall. This is a limitation in the way BoB outputs were used as pre-annotations in our study since they are derived using both rules and machine learning approaches. High performing machine learning based systems usually require training on similar documents to those being de-identified [20].

6.6.6 Time Comparison

Across all annotation types, we observed no statistical differences for annotation times between the experiment and control conditions. Lack of time difference may be due to time added for deleting false positives that could equate to the same amount of time required to identify a PHI span in the same document that is not reviewed using BoB+eHOST. This result is contrary to a study by Fort and Saggot [19] that used machine pre-annotations for POS tagging in which significant reduction in time was observed for the experiment. As well as a more recent study by Lingren et al. [24] in which machine pre-annotation was employed to annotate clinical named entities resulting in significant reduction in annotation time and no effect on IAA or standard performance metrics. However, our experimental results are congruent with findings by Ogren and colleagues [23] that outputs generated from a third-party system used as pre-annotations decreased efficiency and produced little gain in data quality.

Although annotations using only the eHOST Oracle were generated faster than the control condition alone, the lack of time difference between experimental and control

conditions may be a consequence of combining pre-annotations with the interactive annotation interface. Higher quality pre-annotations may introduce efficiencies compared with annotation on raw clinical texts. On the other hand, lower quality pre-annotations certainly do not offer a net gain in efficiency or annotator performance due to the added task of modifying existing, adding missed, or deleting spurious annotations. It is likely that the ratio of correct to incorrect pre-annotations must be small in order for there to be any efficiency gains offered by the machine-assisted approach [35].

6.6.7 Coverage Differences with Added Annotators

The number of annotators needed to achieve adequate recall and precision may be dependent on various factors that should be explored in future annotation studies. First, different clinical documents may require more reviewers as compared with fewer. Second, a privacy risk ranking of PHI types should be one consideration for these tasks. Third, there are policy implications for the redaction of PHI from clinical texts that extend beyond simply removing personally identifiable information. A reference standard generated by human reviewers is never perfect and the ability of humans to reliably annotate for PHI and generate an accurate reference standard is a difficult goal to achieve. Even though annotators trained on the de-identification task and tools until they achieved a pre-defined performance threshold in the training, IAA never plateaued across either annotator training or experiment for both control and experimental conditions for some annotation types in our study. This indicates human annotators were still "learning" to correctly identify and classify some annotation types through both the training and experiment. There are two tasks that must go on simultaneously in the reviewers mind: first, the reviewer must read the clinical text and second, the reviewer must apply the

guidelines and annotation schema. This observation speaks to the complexity of a manual de-identification task, the difficulty of providing enough examples of each annotation type, and the ability of human annotators to consistently apply an annotation schema written in the spirit of the HIPAA *Safe Harbor* method.

6.7 Conclusions

We have demonstrated the generalizability of a manual de-identification task on a publicly available, heterogeneous corpus of clinical documents, MTSamples, using an annotation schema and guidelines originally developed for a similar annotation effort on VHA clinical documents. Based on this schema and the resulting annotations, we determined most PHI annotations represent expressions of medium risk of re-identification Overall, we observed that PHI classes can be annotated with high average interannotator agreement. In this experiment, machine-assisted annotation did not improve annotation quality for most PHI classes and did not provide statistically significant time savings compared to manual annotation of raw documents. However, we determined that two annotators perform PHI annotation with highest F1-measure and observed diminishing PHI coverage with each added annotator. This could be an important finding for institutions creating a de-identification service where humans would be hired to manually redact PHI from clinical texts. Finally, we have produced a de-identified clinical document corpus and a reference standard that can be used for future experimentation on NLP de-identification methods.

In the case of building a reference standard that will be used to train automated systems for de-identification, it is better to err on the side of high recall considering the implications and negative impacts of HIPAA violations on the institution providing the

data. These issues should be considered in the context of patient privacy, potential information loss, and the workload associated with manual de-identification of clinical texts. Balancing the expectations of existing ethical and legal responsibility with practicality and the burdens of human review is paramount for any sound implementation of automated machine methods used for clinical text de-identification. This study contributes to the ongoing analysis of human review methods used for de-identification of clinical texts.

6.8 Acknowledgments

Development of the eHOST annotation tool was supported by the VA Consortium for Healthcare Informatics Research (CHIR), VA HSR HIR 08-374. Annotation of the MTSamples corpus was supported by NIH Grant U54 HL 108460 for integrating Data for Analysis, Anonymization and Sharing (iDASH), NIGMS 7R01GM090187. We also wish to acknowledge the efforts of all annotators involved in pilot testing annotation schema and annotating this document corpus. The fully de-identified MTSamples document corpus is available from iDASH via a data use agreement.

6.9 References

1. GPO US. CFR Title 45 Subtitle A Part 46: Protection of Human Subjects [Internet]. GPO; Oct 1, 2008. Available from: http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr46_08.html.

2. Meystre, S.M., Friedlin, F.J., South, B.R., Shen, S., Samore, M.H. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med. Res. Methodol. 2010. 10, 70.

3. Aberdeen, J., Bayer, S. Yeniterzi, R. Wellner, B. Clark, C. Hanauer, D. The MITRE Identification Scrubber Toolkit: design, training, and assessment. Int. J. Med. Inform. 2010. 79(12), 849–859.

4. Friedlin, F.J., McDonald, C.J. A software tool for removing patient identifying information from clinical documents. J. Am. Med. Inform. Assoc. 2008. 15(5), 601–10.

5. Beckwith, B.A., Mahaadevan, R., Balis, U.J. Kuo, F. Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med. Inform. Decis. Mak. 2006. 6, 12.

6. Gardner, J. Xiong, L. HIDE: An integrated system for health information de-identification. IEEE Symposium On Computer Based Medical Systems. Proceedings. 2008. 254–259.

7. Neamatullah, I., Douglass, M.M., Lehman, L.W., Reisner, A., Villarroel, M., Long, W.J. Automated de-identification of free-text medical records. BMC Med. Inform. Decis. Mak. 2008. 8(1), 32.

8. Ferrández, O., South, B.R., Shen, S., Friedlin, F.J., Samore, M.H., Meystre, S.M. BoB, a best-of breed automated text de-identification system for VHA clinical documents. J. Am. Med. Inform. Assoc. 2013. 20(1), 77–83.

9. Sweeney, L. Replacing personally-identifying information in medical records, the Scrub system. Proc. AMIA. Annu. Fall. Symp. 1996. 333–337.

10. Taira, R.K., Bui, A.A., Kangarloo, H. Identification of patient name references within medical documents using semantic selectional restrictions. AMIA Annu Symp Proc. 2002. 757–761.

11. Ferrández, O., South, B.R., Shen, S., Friedlin, F.J., Samore, M.H., Meystre, S.M. Evaluating current automatic de-identification methods with Veteran's Health Administration clinical documents. BMC Med. Res. Methodol. 2012. 12(1), 109.

12. Berman, J.J. Concept-match medical data scrubbing. How pathology text can be used in research. Arch. Pathol. Lab. Med. 2003. 127(6), 680–586.

13. Ruch, P., Baud, R.H., Rassinoux, A.M., Bouillon, P., Robert, G. Medical document anonymization with a semantic lexicon. AMIA Ann. Symp. (2000) 729–733.

14. Deleger, L., Molnar, K., Savova, G., Xia, F., Lingren, T., Li, Q. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. J. Am. Med. Inform. Assoc. Aug 2, 2012.

15. GPO US. CFR 45 Subtitle A Part 164: Security and Privacy [Internet]. GPO; Oct 1, 2008. Available from: http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr164_08.html.

16. Uzuner, Ö., Luo, Y., Szolovits, P. Evaluating the state-of-the-art in automatic de-

identification. J. Am. Med. Inform. Assoc. 2007. 14(5), 550–563.

17. Sibanda, T., Uzuner, Ö. Role of Local Context in De-identification of Ungrammatical, Fragmented Text. NAACL-HLT. 2006. 65-73.

18. Carrel, D., Malin, B., Aberdeen, J., Bayer, S., Clark, C., Wellner, B., Hirschman, L. Hiding in plain sight: use of realistic surrogates to reduce exposure of protected helath information in clinical text. J. Am. Med. Inform. Assoc. 2013. Mar-Apr;20(2):342-8.

19. Fort, K., Sagot, B. Influence of pre-annotation on pos-tagged corpus development. In: Proceedings of the Fourth Linguistic Annotation Workshop, LAW IV. 2010:56-63.

20. Hanauer, D., Aberdeen, J., Bayer, S., Wellner, B., Clark, C., Zheng, K., Hirschman, L. Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. Int. J. Med. Inform. 2013. 821-831.

21. Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Symposium: American Medical Informatics Association; 2001:17.

22. Ganchev, K., Pereira, F., Mandel, M., Carrol, S., White, P. Semi-automated named entity annotation. Proceedings of the Linguistic Annotation Workshop; Prague, Czech Republic: Association for Computational Linguistics, 2007. 53–6.

23. Ogren, P.V., Savova, G.K., Chute, C.G. Constructing Evaluation Corpora for Automated Clinical Named Entity Recognition. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC 2008. 3143-3150.

24. Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., Stoutenborough, L., Li, Q., Solte, I. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. J. Am. Med. Inform. Assoc doi: 10.1136/amiajnl-2013-001837.

25. South, B.R., Shen, S., Friedlin, F.J., Samore, M.H., Meystre, S.M. Enhancing annotation of clinical text using pre-annotation of common PHI. Proceedings of the AMIA Annual Symposium; 2010:1267.

26. Stenetorp, P., Pyysalo, S., Topic, G., Ananiadou, S., Tsujii, J. BRAT: a web-based tool for NLP-assisted text annotation. EACL 2012; 2012:102.

27. South, B.R., Shen, S., Leng, J., Forbush, T., DuVall, S., Chapman, W.W. A prototype tool set to support machine-assisted annotation. Proceedings of the 2012 Workshop

on Biomedical Natural Language Processing. BioNLP '12, Stroudsburg, PA, USA, Association for Computational Linguistics. 2012. 130-139.

28. Settles, B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 2005;21:3191–2.

29. Mayer, J.M., Shen, S., South, B.R., Meystre, S.M., Friedlin, F.J., Ray, W.R., Samore, M.H. Inductive creation of an annotation schema and a reference standard for de-identification of VA electronic clinical notes. AMIA Annu. Symp. Proc. 2009. Nov 14; 2009:416-20.

30. Hripcsak, G., Rothschild, A. Agreement, the F-measure, and reliability in information retrieval. J. Am. Med. Inform. Assoc. 12(3) 296-8.

31. Pappas, P.A., DePuy, V., An overview of nonparametric tests in SAS: when, wh, and how. http://analytics.ncsu.edu/sesug/2004/TU04-Pappas.pdf. Accessed: 21 August 2013.

32. Dorr, D.A., Phillips, W.F., Phansalkar, S., Sims, S.A., Hurdle, J.F. Assessing the difficulty and time costs of de-identification in clinical narratives. Method Inform. Med. 2006. 45(3); 246-252.

33. South, B.R., Shen, S., Maw, M., Ferrández, O., Friedlin, F.J., Meystre, S.M. Prevalence estimates of clinical eponyms in de-identified clinical documents. AMIA Summits Transl. Sci. Proc., CRI. 2012; 136.

34. Brooke, J. SUS: A "quick and dirty" usability scale. In: Jordan, P. W., Thomas, B., Weerdmeester, B. A., McClelland (Eds.) Usability Evaluation in Industry pp. 189-194. Taylor & Francis, London UK. 1996.

35. Felt P., Ringger, E., Seppi, Heal, K., Haertel, R., Londsdale, D. First results in a study evaluating pre-annotation and correction propagation for machine-assisted syriac morphological analysis. In: Proceedings of the Tenth International Conference on Language Resources and Evalauation. LREC 2012: 878-885.

CHAPTER 7

DISCUSSION

7.1 Summary

This dissertation research hypothesized that integrating pre-annotation or other machine-assisted methods within manual annotation workflows would improve efficiency of manual annotation tasks without diminishing the quality of generated reference standards. A mixed methods approach was used for this research integrating methodologies from cognitive science and artificial intelligence [1-8]. These methods were used to ascertain qualitative aspects related to manual annotation of clinical texts along with quantitative methods used to estimate quantitative aspects related to two annotation experiments that tested two approaches to machine-assisted manual annotation of clinical texts used as a semantic priming mechanism. Generating reference standards is costly and resource intensive endeavor so it makes sense to evaluate methods that could potentially increase efficiency and reduce human workload.

There are several important implications the research in this dissertation addresses that are useful lessons when machine-assisted approaches are integrated with workflows used to generate labels and build reference standards for NLP systems development. First, under most practical applications of NLP, and more specifically supervised learning, development efforts are reliant on some reference standard that is generated by

subjective human review (the annotation task). Human beings are not infallible reviewers and therefore there can never be an absolute ground truth. Even in cases where best attempts are made to explicate the annotation task and annotators receive task specific training on tools, guidelines. Second, errors have a tendency to creep into the reference standard even with the best efforts to minimize ambiguity and uncertainty. These errors ultimately affect the validity of the generated reference standard. Errors result where an annotator misinterprets guidelines, misunderstands the nuances or context of the information they are reviewing and the interplay between these factors and the underlying information quality of a given clinical document. Subtle errors can sneak past even the most rigorous adjudication effort and become part of the final reference standard that is then used for NLP system development [9].

It is also possible that despite the best efforts annotators will continue "learning" throughout the course of an annotation campaign [2, 10, 11]. This is precisely what happened in both Aims 1 and 2 for some annotation types, which resulted in interannotator agreement being lower than some expected threshold for a given annotation task or subtask that was part of some larger NLP development effort. This is not necessarily a weakness in study design and is more of a reality for any task where subjective manual human review is used to generate data. Interrater reliability or its measurement in terms of annotation studies (interannotator agreement), is an important indicator of task consistency and a useful measure of the "correctness" or agreement between multiple reviewers for a given annotation task relative to a desired "ground truth." It can also be used as a measure of task complexity when integrated early in an annotation campaign and provides valuable insight as to the quality of instructions sets in

the form of annotation guidelines and annotator training, and may also be used to help induce new rules [12] and identify problems with these. Interannotator agreement should not be viewed as a hypothesis test. It is simply a measurement of how well reviewers agree with each other [13] (or themselves [14]) for a given annotation task. In addition, the effects of individual biases are proportional to the number of annotators [15], and many factors affect the outcome of interrater agreement metrics across an annotation campaign [2, 10]. Some of these factors are listed below and examined in greater detail in the discussion section below for Aim 1 and were discussed in the context of the qualitative analysis in Chapter 3.

1) Interactions between the human annotator, the tools used to produce the labels, and the document sources that are labeled;
2) Managing uncertainty and appropriateness of reviewer training and learning;
3) Managing annotation task complexity;
4) Balancing annotator efficiency and accuracy;
5) Social and motivational forces.

When annotators label texts for a given annotation task there are several assumptions that are made by the researcher. First, agreement must be measured on the same texts. Second, documents will be multiply annotated by different annotators. Third, annotators work independently on these labeling tasks. If these assumptions are met and annotators mark texts in the same way then it is also assumed that:

1) The annotators have successfully internalized the annotation scheme (instruction set) and use of annotation tools in the same way.

2) The annotators can apply the annotation scheme and annotation tools consistently when given new data that differs from what they were trained on.
3) The annotations generated provide "correct" labels that can be used to train machines for the equivalent task.

Even when these assumption are met it is possible in situations where there is low agreement between annotators there may be underlying consistency in disagreements. Therefore, in these cases high agreement may not necessarily be required to infer a "ground truth." Annotators may also disagree in consistent ways allowing one to correct for these systematic disagreements. One danger in this case, however, is it may be more difficult to understand how the annotator is interpreting the task. Like any other statistical measure however, interannotator agreement alone should not be relied on as a sole measure of the strength or weakness of study methods. It simply provides a measure of task consistency (reproducibility) compared with other reviewers for the same task. For complex annotation tasks where there are many items defined in an annotation schema, and where the annotation task is inherently difficult and requires some level of inference beyond what can be explicated, interannotator agreement may always be low regardless of how well a study design has integrated attempts to control for confounding, annotator uncertainty, or systematic bias introduced into the reference standard.

There are some specific limitations and threats to internal validity regarding Aim 2 and Aim 3 of this dissertation that are worth noting. These include data sparsity, attrition, maturation, and instrumentation bias. For example, certain types of information prevalence may be lower than expected (data sparsity), annotators may drop out of the study before completion (attrition), annotator performance may take many iterations to

reach a plateau or it may simply continually improve (maturation). Use of different tools, pre-annotations, or annotation workflows may introduce bias over time (instrumentation bias). Prior differences between groups may also affect the outcome when any of these options are introduced. In the worst case scenario the conclusion may be made that the intervention and use of pre-annotation did not make a difference when in reality some difference was achieved, or that the intervention did make a difference when it may not have. Improving methods of training, will only help overcome but not completely eliminate these study limitations. These issues are further discussed below as they relate to each dissertation aim and respective experiments.

7.2 Aim 1

Recommendations from the qualitative analysis of semistructured interviews discussed in Chapter 3 include developing new methods to predict required annotator training, increasing annotator performance feedback, and integrating interactive approaches via new annotation tools [10]. Annotation is an excellent example of human information processing in practice. Generating a reference standard is one task that focuses on balancing cognitive load with the expectations of the NLP development team in a way that produces data in a timely and cost effective manner. Often these goals are at odds with each other. Balancing cognitive load with the factors that affect the actual task of manually generating labels is one area where annotation research could clearly benefit. Cognitive load is an important aspect related to annotation tasks for several reasons. First, annotation is a schema driven process that relies on appropriate categorization that is dependent on prior knowledge/training or expertise gained while learning a new review task. Some items defined in an annotation scheme or that are deemed necessary for some

NLP development use case are simply easier to annotate than other items in an annotation scheme.

There are many factors that come into play when human beings label texts to generate reference standards that are used to train machine systems for information extraction or classification tasks. For annotation tasks there is an interplay between these factors and the cognitive load required to successfully carry out an annotation campaign. The qualitative analysis related to Aim I and discussed in Chapter 3 uncovered 5 specific factors that come into play for any annotation task.

The first of these factors involves interactions between the human annotators, the tools used to produce the labels, and the document sources that are labeled. These are important interactions since the tools used by humans to annotate texts can generate biases or introduce systematic error. This is certainly true where machine outputs are used as pre-annotations and their cues, both conscious and subconscious, function as a semantic priming mechanism [16]. For example, in the situation where the machine performance used to generate pre-annotations does not meet or exceed the underlying interannotator agreement threshold for an annotation task or annotation type as occurred for some annotation types in Aims 2 and 3. In this case the annotator will spend more time correcting or adding missed annotations. Or in the case where machine outputs generate too many false positives the annotator will spend more time deleting spurious annotations and may become overwhelmed by the prevalence of pre-annotations and miss spans of text that represent true positives. Certainly the information quality of the document being annotated also comes into play as does the reviewers ability to understand the underlying meaning of the clinical information [9]. Determining task

specific performance thresholds where the balance exists between pre-annotation precision and recall is one open area of research in the clinical NLP domain.

A second factor involves managing annotator uncertainty and gauging appropriateness of reviewer training and learning. Annotation is a process driven by expert knowledge required to help define rules incorporated into guidelines, develop annotation schema, and produce reliable and valid information labels. Annotator training and learning is an important factor when generating reference standards since experts must be hired or trained to generate labels. Several studies have shown that it is possible to use nonexperts for some annotation tasks but they have a much steeper learning curve than experts [17, 18]. However, training and identifying experts is a costly process. Human reviewers are driven by the goal produce high quality labels [19] while at the same time they may also spend time attempting to correct errors. Furthermore, the guidelines and documents used for an annotation task may also further complicate and introduce bias if they contain inconsistencies or ambiguities. Humans react to this by attempting to control their information environment using a wide variety of strategies to cope with uncertainty, reduce cognitive load, and increase motivation [20].

The third factor involves managing annotation task complexity. All annotation tasks involve some level of complexity, this goes without saying, but reducing the complexity of instructions sets, the number of items in an annotation scheme and defining what is to be annotated in explicit way helps minimize cognitive load and reduces the possibility that annotators will revert to heuristic processing. The danger with relying on heuristic processing is that it leaves out the underlying meaning, the semantics, of what is being annotated. This obviously affects the quality of data produced. As the number of

items in an annotation scheme increases the demands on attentional resources increase [21], it becomes more and more difficult for the annotator to recall rules sets and definitions that correspond with each information class and the reliability of the annotation task decreases. The reviewer will need to rely more and more on heuristic processing and explicitly defined rules found in annotation guidelines or from other experts annotating the same data types, rather than prior knowledge. motivation [20].

The fourth factor that comes into play is balancing annotator efficiency, speed and accuracy. Annotators must balance their inherent desire to be reliable and produce valid data with the competing goal of being efficient [22, 23]. Annotation is a context dependent task. This is particularly true in real-world production environment where an annotator is expected to produce a minimal number of annotations or reviews within a given time period. Different methods can be used to regulate these goals and increase motivation, but balancing them places attentional resources at risk.

Finally, the fifth and final factor involves social and motivational forces related to the annotation task. Annotation as it turns out is a highly social task. No one person is an expert on every given item specified by an annotation scheme or defined within annotation guidelines. Expertise for any annotation task is obtained over time, after an annotator has been provided many examples of what it is they are to annotate. This also requires individual feedback [11, 14, 24] and a social component that relies on motivational forces that reinforce "correct" labels [25]. For example, more attention may be paid to certain things like inconsistencies in guidelines, or individual annotator performance on an annotation task or subtask. It is possible that personalities and social processing can lead to bias where one expert influences the decisions of a more novice

annotator. These social factors are not all bad and they do lead to refinements in guidelines and sharing of expert knowledge. However, with more complex annotation schemes more social processing will occur.

Each of these factors is important since each has some effect on reliability of an annotation task and the validity of generated reference standards. These factors are also related to the cognitive load associated with an annotation task. Often the development goals, the cognitive load, the degree of task independence are also issues where the trade-off between accuracy and speed becomes more apparent. The relation of these factors with the annotation tasks integrated with Aims 2 and 3 of this dissertation is important and should not be discounted for other annotation studies.

7.3 Aim 2

Chapter 4 of this dissertation details the annotation experiment integrated with the 2010 i2b2/VA Challenge. This experiment modified a typical annotation workflow integrating different intervention types conditioning on full document versus sentence level annotation and use of noninteractive pre-annotation at the annotation class level. We hypothesized that integrating noninteractive pre-annotation with a modified annotation workflow improves efficiency of manual clinical text annotation without diminishing reliability and validity metrics. Our modified workflow also integrated review steps going beyond strict adjudication with arbitration adding additional review levels for each challenge annotation subtask. Significant improvements in recall and precision for annotation of semantic classes and assertions were observed as additional review layers were added.

This is in contrast to the analysis that used LOESS to visualize F_1-measures for concept identification, assertion classification, and relations as a method of evaluating annotator efficiency and data quality. In these analyses F_1-measure for concept annotation of medical problems, treatments and tests improved (from 0.82 to 0.87) and then plateaued across the first 10 document batches. F_1-measure for assertion classification demonstrated a similar pattern reaching a plateau at batch 15 (from 0.73 to 0.83). For relations F_1-measure increased slightly but never reached an obvious plateau. Further analyses used GEE to compare F_1-measures at the annotated batch, document and task level for the same documents reviewed by annotators assigned to Aim 2 intervention groups. At the document level, and where semantic classes are aggregated, no significant differences were observed between documents that were pre-annotated at the semantic level versus documents annotated on raw full documents. However, significant differences were observed for sentence level, pre-annotated sentence, and pre-annotated full document interventions at the semantic class level and in situations where context is required such as assertion classification, or relations [26]. In the end, highest task efficiency and data quality was achieved when annotators reviewed raw clinical texts and did not receive pre-annotations.

For the experiment integrated with Aim 2 and discussed in Chapter 4, providing pre-annotations using a simple noninteractive approach did not adversely affect annotation quality, but it did not improve annotation quality. Furthermore, this experiment shows that noninteractive pre-annotation can be integrated effectively with tasks that can be clearly explicated at the level of concept identification and classification. However, annotation on raw clinical texts produced the highest quality

data in comparison with the condition where reviewers annotated pre-annotated full or sentence level documents. Moreover, results from the experiment integrated with Aim 2 are contradictory to many published studies, discussed as part of the background in Chapter 2, that suggest use of pre-annotation may increase task consistency, and overall reference standard validity [27, 28]. Indeed, the effect on annotation task efficiency is dependent on how pre-annotation was done and the underlying performance of the pre-annotation system.

There are clearly some limitations to this research. First, it is possible that confounding was introduced as a result of the study design. This experiment utilized a factorial design including 4 intervention types distributed among 9 part-time annotators and not all annotators reviewed the same number of documents. Annotator feedback on the noninteractive pre-annotation approach used it was a split between annotators either believing the pre-annotations were useful, or believing it was not. This sentiment was not evenly distributed between clinician and nonclinician annotators; even though clinician annotators achieved a training plateau (i.e., the ith batch) much faster than non-clinician annotators on average. In a follow-up study that involved annotators for the subsequent 2011 i2b2/VA Challenge on coreference resolution, annotators rated semantic priming as helpful for their decision making processes when it was not provided in the form of a pre-annotated class but an attribute value associated with uncommon clinical concepts, definitions and synonyms. Semantic priming was rated as less helpful when it was associated with common diseases [29].

In regards to use of a noninteractive pre-annotation approach there are several potential biases that were explained in Chapter 2 but we state them again here to remind

the reader. First, when pre-annotations are highly precise human reviewers may concentrate only on pre-annotated information simply correcting pre-annotations without adding relevant spans that were missed in the pre-annotations. On the other hand, when pre-annotations are overly sensitive and there are many false positives human annotators may concentrate too much on what is missing but not correct pre-annotations due to the volume or complexity of what has been pre-annotated. Second, it is difficult for some types of pretaggers to produce high enough quality pre-annotations because the tools are simply too difficult to build or might not yet exist.

Contrary to the recommendations made by Lingren et al. [28], unless the case can strongly be made to use a noninteractive pre-annotation approach it is a balancing act to integrate its use with the tradeoffs with dimensions of annotation quality (i.e., time, cost, task reliability and reference standard validity). Especially in the case as occurred in Aim 2 where the pretagger produced lower quality pre-annotations for some annotation types. This conclusion is congruent with the recommendation by Ogren and colleagues [30] that pre-annotation resulted in little benefit in terms of efficiencies and performance gains. Further work is needed to assess where use of pre-annotation is optimal for clinical annotation tasks and propose generalizable estimates of the performance threshold tradeoffs that occur between the quality of pre-annotations and annotator efficiency. These types of generalizable estimates are currently lacking in the clinical NLP literature.

7.4 Aim 3

For the experiment integrated with Aim 3 (Chapter 6), integrating pre-annotations from the BoB de-identification system coupled with an interactive annotation interface via the eHOST "Oracle mode" did not produce statistically significant time savings.

Furthermore, coupling the outputs of a de-identification system with an interactive annotation interface improved annotation quality for only two annotation types found in the MTSamples corpus. These included *Other Organization Names* and clinical eponyms (*Diseases* only). It is reasonable to assume that some annotation types can be more easily identified than others and require less contextual information to classify. Annotation of raw clinical texts in the absence of both BoB pre-annotations and the eHOST Oracle mode produced the highest quality data. High performance metrics were observed for partial span matching criteria – recall (control 0.84, experiment 0.84), precision (control 0.94, experiment 0.85), and F_1-measure (control 0.89, experiment (0.84). With each added annotator from 1-7, recall increased from 0.66 to 0.92 as precision decreased from 0.82 to 0.61. F_1-measure peaked at 0.79 with 2 annotators. Finally, LOESS analysis suggests that annotators were "learning" throughout the annotation task and unlike the experiment integrated with Aim 2, annotators never reached a training plateau for many annotation types in the annotation schema despite achieving a predefined agreement threshold prior to moving on to annotating documents reserved for the closed annotation phase of the study.

There are some obvious limitations to the annotation experiment integrated with Aim 3. First, the quality of pre-annotations used may have significant bearing on whether or not pre-annotation provides a net gain in terms of annotation efficiency. Indeed, a very well pre-annotated corpus may introduce efficiencies compared to annotation on raw clinical texts. On the other hand, a poorly pre-annotated corpus clearly does not offer a net gain in efficiency or performance since the tasks of adding, modifying and deleting spurious annotations more than offsets the benefits of providing correct pre-annotations.

Baseline performance (recall, precision, F_1-measure) for out-of-the-box BoB pre-annotations was quite low (0.20, 0.42, 0.27) across all annotation types and only in the moderate range for annotation types that were most prevalent in the MTSamples corpus (0.44, 0.48, 0.46). It is likely that the ratio of incorrect to correct annotations must be small in order for there to be any efficiency gains offered by the system pre-annotations and the interaction of human annotator and machine-assisted interface [31]. As mentioned in the discussion for Aim 2, discovering where this balance exists is an open question for de-identification methods and many other applications of NLP to clinical documents.

A second limitation worth noting involves implementation of the interactive machine-assisted annotation approach in this experiment. For this experiment we used noninteractive pre-annotation and out-of-the-box outputs from BoB coupled with the eHOST "Oracle." Candidate spans were provided to the annotator via the eHOST Oracle based on their own previous annotations using a simple string matching and regular expressions approach. When candidate spans are presented to the reviewer he/she can choose to accept or reject these. A better implementation of the interactive machine-assisted annotation approach could involve leveraging the input of the expert reviewer that is then used to produce machine-generated suggestions on subsequent annotation batches as suggested by Settles [32] and implemented in the MIST tool [33]. This approach implies using the outputs from prior annotation batches that are used as training inputs for the machine with machine pre-annotations provided on subsequent document batches.

There are two scenarios for improvement on our application of the methods implemented in this experiment. First, given a simple use case and under a real-time

scenario such as a manual de-identification task, the training and underlying expert knowledge of the reviewer is the limiting factor. For simple annotation tasks one must assume that the annotator can be quickly trained to reach an acceptable performance level and produce "correct" labels. This did not occur with the approach used for this experiment. Second, for more complex cases, one must assume that the annotation task can be explicated in such a way that inexperienced annotators can be trained in a short amount of time or expert reviewers can be readily recruited to produce reliable and "correct" labels. Adapting a truly interactive machine-assisted annotation approach for use on a more complex use case would require significant modifications to eHOST. However, eHOST could be modified to implement the algorithm used by RapTAT developed by Gobbel and colleagues [34]. Under this scenario, the limiting factor is the ability of the system to provide candidate spans that are appropriate and reflect the complexity of the data being reviewed. One threat to this approach when applied to a variety of document types or more complex clinical use cases is the number of items returned via the interactive annotation interface may increase to the point where cognitive load is actually higher than it might be if annotators reviewed raw documents using a smaller more constrained annotation schema. This is a compelling proposition and is one area reserved for future experimentation.

7.5 References

1. Patton, M.Q. Qualitative Research and Evaluation Methods. 2002. Sage Publications.
2. Campbell, E.M., Sittig, D.F., Chapman, W.W., Hazlehurst, B.L., Cohen, A.M. Understanding inter-rater disagreement: a mixed methods approach. In: AMIA Annu Symp Proc. 2010: p. 81-5.

3. Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J., Roberts, I., Setzer, A., Tapuria, A., Wheeldin, B. The CLEF corpus: semantic annotation of clinical text. In: AMIA Annu Symp Proc. 2007. 625-9.

4. Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., Roberts, I., Setzer, A. Building a semantically annotated corpus of clinical texts. In: J Biomed Inform, 2009. 42(5): 950-66.

5. Grishman, R. Sundheim, B. Message Understanding Conference-6: a brief history. 16th Conference on Computational Linguistics (COLING), 1996: p. 466-71.

6. Hripcsak, G., Wilcox, A. Reference standards, judges, and comparison subjects: roles for experts in evaluating system performance. In: J Am Med Inform Assoc. 2002. Jan-Feb;9(1):1-15.

7. Hripcsak, G., Heitjan, D.F. Measuring agreement in medical informatics reliability studies. In: J Biomed Inform. 2002. Apr;35(2):99-110.

8. Hripcsak, G., Rothschild, A.S. Agreement, the f-measure, and reliability in information retrieval. In: J Am Med Inform Assoc. 2005. May-Jun;12(3):296-8.

9. Shen, S., South, B., Butler, J., Barrus, R., Weir, C. The relationship between structural characteristics of 2010 challenge documents and ratings of document quality. AMIA Annu Symp Proc 2012: 848-855.

10. South, B.R., Shen, S., Barrus, R., DuVall, S.L., Uzuner, Ö., Weir, C. Qualitative analysis of workflow modifications used to generate the reference standard for the 2010 i2b2/VA challenge. In: AMIA Annu Symp Proc. 2011.

11. Dandapat, S., Biswas, P., Choudhury, M., Bali, K. Complex linguistic annotation - no easy way out! A case from Bangla and Hindi POS labeling tasks. In: Proceedings of the Third ACL Linguistic Annotation Workshop, Singapore. 2009.

12. Mayer, J.M., Shen, S., South, B.R., Meystre, S., Friedlin, F.J., Ray, W.R., Samore, M. Inductive creation of an annotation schema and a reference standard for de-identification of VA electronic clinical notes. AMIA Annu. Symp. Proc. 2009 Nov 14; 2009:416-20.

13. Artstein, R., Poesio, M. Inter-coder agreement for computational linguistics. Computational Linguistics 34(4): 555-596, 2008.

14. Gut, U., Saskia-Bayerl, P. Measuring the reliability of manual annotations of speech corpora. Poster at the Second International Conference for Speech Prosody 2004 (SP2004), Nara-Ken New Public Hall, Nara, Japan, March 23-26.

15. Artstein, R., Poesio, M. Bias decreases in proportion to the number of annotators. In: Gerhard Jaeger, Paola Monachesi, Gerald Penn, James Rogers, and Shuly Wintener (Eds.), Proceedings of FG-MoL 2005, pp. 141-150. Edingurgh, August 2005.

16. Ferrand, L. New, B. Semantic and associative priming in the mental lexicon. In Bonin (Ed), The Mental Lexicon. New York: Nova Science Publishers. 2003. pp 25-43.

17. Chapman, W.W., Dowling, J.N., Hripcsak, G. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. Int J Med Inform, 2008. 77(2): p. 107-13.

18. Snow, R., O'Connor, B., Jurafsky, D., Ng. A.Y. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In Proceedings of EMNLP 2008, pages 254-263.

19. Wyer, R.S. Social Comprehension and Judgment: The Role of Situation Models, Narratives, and Implicit Theories. 2004, Mahwah, NJ: Erlbaum.

20. Weir, C.R., Nebeker, J.R., Hicken, B.L., Campo, R., Drews, F., LeBar, B. A cognitive task analysis of information management strategies in a computerized provider order entry environment. J Am Med Inform Assoc, 2007. 14(1): p. 65-75.

21. Estes, W. Classification and Cognition. 1996. New York: Oxford University Press.

22. Kruglanski, A. Lay Epistemics and Human Knowledge: Cognitive and Motivational Bases. New York: Plenum. 1989.

23. Hollnagel, E. The ETTO Principle: Efficiency-Thoroughness Trade-off: Why Things that Go Right Sometimes Go Wrong. Cornwall, Britain: Ashgate. 2009.

24. Voormann, H., Gut, U. (2008). Agile corpus creation. Corpus Linguistics and Linguistic Theory, 4(2):235-251.

25. Harton, H.C., Green, L.R., Jackson, C., Latane, B. Demonstrating dynamic social impact: consolidation, clustering, correlation, and (sometimes) the correct answer. Teaching of Psychology, 1998. 25:p. 31-34.

26. Suo, Y., Shen, S., DuVall, S.L., Uzuner, O., South, B.R. Evaluation and visualization of human annotator learning patterns. In: AMIA Annu Symp Proc. 2012.

27. Aurélie, N., Islamaj-Doğan, R., Zhiyong, L. Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. In: J Biomed Inform. 2011 Apr; 44(2):310-8.

28. Lingren, T., Deleger, L., Molnar, K., Zhai, H., Meinzen-Derr, J., Kaiser, M., Stoutenborough, L., Li, Q., Solte, I. Evaluating the impact of pre-annotation on annotation speed and potential bias: natural language processing gold standard development for clinical named entity recognition in clinical trial announcements. J.

Am. Med. Inform. Assoc doi: 10.1136/amiajnl-2013-001837.

29. Forbush, T.B., Shen, S., Thibault, J.C., Weir, C., Uzuner, O., South, B.R. Using the UMLS as a semantic priming mechanism for co-reference resolution in annotation of clinical texts. In: AMIA Symp Proc. 2011.

30. Ogren, P.V., Savova, G.K., Chute, C.G. Constructing evaluation corpora for automated clinical named entity recognition. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation LREC 2008: 3143-3150.

31. Felt, P., Ringger, E., Seppi, K., Heal, K., Haertel, R., Lonsdale, D. First results in a study evaluating pre-annotation and correction propagation for machine-assisted syriac morphological analysis. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2012: 878-885.

32. Settles, B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics 2005;21:3191–2.

33. Hanauer, D., Aberdeen, J., Bayer, S., Wellner, B., Clark, C., Zheng, K., Hirschman, L. Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. Int. J. Med. Inform. 82 (2013) 821-831.

34. Gobbel, G.T., Reeves, R., Jayaramaraja, S., Giuse, D., Speroff, T., Brown, S.H., Elkin, P.L., Matheny, M.E. Development and evaluation of RapTAT: a machine learning system for concept mapping of phrases from medical narratives. Journal of Biomedical Informatics. December, 2013.

CHAPTER 8

CONCLUSION

8.1 Significance to the Field

Use of the pre-annotation approaches studied as part of this dissertation should be implemented with caution as the machine performance was not high enough in both experiments integrated with Aims 1 and 2 to generate good candidate spans for some of the annotation types specified under these research conditions. The interplay of these approaches and other factors affecting integration of these methods with annotation workflow was discussed in depth in Chapter 6. However, these experiments are useful since they add to the growing number of studies where pre-annotation approaches have been implemented in an attempt to produce efficiency gains without reducing the quality of generated annotation data.

In application of NLP methods to the clinical domain there is an increased need to explore methods that introduce efficiencies into annotation of clinical texts in ways that are simple, practical, economical, and scalable. As large amounts of unstructured clinical texts become available with adoption of Electronic Medical Records (EMRs) there will be a greater demand to use information extraction methods to extract data from unstructured information resources common to clinical narratives. This leads to a problem of scalability when these methods are possibly applied to tens of thousands of

clinical documents and the resulting annotations are used to develop an NLP system. Indeed the quality of pre-annotations has significant impact on the net gain in annotation efficiency and quality of the reference standard generated. Furthermore, annotation workflows that use outputs from some system simultaneously being trained for the same task should be tightly scrutinized for bias.

8.2 Opportunities for Future Directions

The methods evaluated as part of this dissertation research could be implemented and scaled in a way that lends itself to annotation of larger clinical corpora. These methods are agile, could be used in ways that are collaborative, and implemented using the tools that were developed as part of this research. Further research expanding these methods is needed to assess interactive approaches that can be scaled to larger annotation efforts. Use of active or interactive learning that Settles et al. [1, 2] suggest may provide faster and more scalable solutions instead of simple noninteractive pre-annotation. However, these methods have not yet been evaluated in the clinical domain in any practical manner. Other methods that may introduce efficiencies include crowd-sourced or distributed annotation. This could even go so far as an agile annotation approach suggested by Voorman et al. [3].

There are several recommendations and potential areas of additional exploration that the discerning reader should be aware of. These include:

- (Re)Design annotation tools to allow real-time performance feedback and self-monitoring during annotator training.

- Evaluate interactive annotation using different annotation tools [4, 5]. These tools should integrate some kind of context aware approaches to deal with clinical sub-language differences.
- Develop and evaluate annotation tools or workflows that incorporate a social component in a way that supports distributed or collaborative annotation [6, 7, 8].
- Systematize training to ensure generalizability integrating the "Games With a Purpose Approach" (GWAP) originally suggested and implemented by Von Ahn [9, 10] in other domains,
- Integrate interventions in a way that controls for learning and training effects, bias, and confounding.
- Develop and evaluate new methods to estimate annotation task complexity and efficiency tradeoffs.
- Extend pre-annotation methods to attributes, relations, and mapping clinical entities to some standardized vocabulary.
- Integrate interoperability standards and relevant vocabularies with methods used to normalize data and make it more computable.
- Balance the best distribution of annotation task(s) and the role of experts [11, 12] with NLP system development goals.
- Reuse previously annotated corpora for development of out-of-domain applications.

These are all topics that could be explored as part of some other dissertation research project or grant and we pass them to others wishing to study this domain. Returning once

more to the preface of this dissertation. The question you have to ask yourself is do you take the blue pill or the red one? The choice is yours.

8.3 References

1. Settles, B., Craven, M., Friedland, L. Active learning with real annotation costs. In: Proceedings of the NIPS Workshop on Cost-Sensitive Learning. 2008.

2. Settles, B. Active Learning literature survey. In: Computer Sciences Technical Report 1648. University of Wisconsin-Madison. 2009.

3. Voormann, H., Gut, U. Agile corpus creation. Corpus Linguistics and Linguistic Theory. 2008. 4(2):235-251.

4. Settles, B. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. Bioinformatics. 2005; 21:3191–2.

5. Gobbel, G.T., Reeves, R., Jayaramaraja, S., Giuse, D., Speroff, T., Brown, S.H., Elkin, P.L., Matheny, M.E. Development and evaluation of RapTAT: a machine learning system for concept mapping of phrases from medical narratives. Journal of Biomedical Informatics. December, 2013.

6. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In: Proceedings of EMNLP 2008, pages 254-263.

7. Le, J., Edmonds, A., Hester, V., Biewald, L. Ensuring quality in crowdsourced search relevance evaluation: the effects of training question distribution. In: Proceedings of the SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010) – July 23, 2010.

8. Zhai, H., Lingren, T., Deleger, L., Li, Q., Kaiser, M., Stoutenborough, L., Solti, I. Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. 2013. J Med Internet Res, 15(4), e73. doi: 10.2196/jmir.2426

9. Von Ahn, L., Dabbish, L. Labeling images with a computer game. Proceedings of the 2004 Conference on Human Factors in Computing Systems – CHI '04. 2004. pp. 319-326.

10. Von Ahn, L. Games with a Purpose. Computer. 2006. 39 (6): 92-94.

11. Dandapat, S., Biswas, P., Choudhury, M., Bali, K. Complex linguistic annotation - no easy way out! A case from Bangla and Hindi POS labeling tasks. In: Proceedings of the Third ACL Linguistic Annotation Workshop, Singapore. 2009.

12. Chapman, W.W., Dowling, J.N., Hripcsak, G. Evaluation of training with an annotation schema for manual annotation of clinical conditions from emergency department reports. Int J Med Inform, 2008. 77(2): p. 107-13.