A Shared Task Involving Multi-label Classification of Clinical Free Text

John P. Pestian[1], **Christopher Brew**[2], **Paweł Matykiewicz**[1,4],
DJ Hovermale[2], **Neil Johnson**[1], **K. Bretonnel Cohen**[3],
Włodzisław Duch[4]
[1]Cincinnati Children's Hospital Medical Center, University of Cincinnati,
[2]Ohio State University, Department of Linguistics,
[3]University of Colorado School of Medicine,
[4]Nicolaus Copernicus University, Toruń, Poland.

Abstract

This paper reports on a shared task involving the assignment of ICD-9-CM codes to radiology reports. Two features distinguished this task from previous shared tasks in the biomedical domain. One is that it resulted in the first freely distributable corpus of fully anonymized clinical text. This resource is permanently available and will (we hope) facilitate future research. The other key feature of the task is that it required categorization with respect to a large and commercially significant set of labels. The number of participants was larger than in any previous biomedical challenge task. We describe the data production process and the evaluation measures, and give a preliminary analysis of the results. Many systems performed at levels approaching the inter-coder agreement, suggesting that human-like performance on this task is within the reach of currently available technologies.

1 Introduction

Clinical free text (primary data about patients, as opposed to journal articles) poses significant technical challenges for natural language processing (NLP). In addition, there are ethical and social demands when working with such data, which is intended for use by trained medical practitioners who appreciate the constraints that patient confidentiality imposes. State-of-the-art NLP systems handle carefully edited text better than fragmentary notes, and clinical language is known to exhibit unique sublanguage characteristics (Hirschman and Sager, 1982; Friedman et al., 2002; Stetson et al., 2002) (e.g. verbless sentences, domain-specific punctuation semantics, and unusual metonomies) that may limit the performance of general NLP tools. More importantly, the confidentiality requirements take time and effort to address, so it is not surprising that much work in the biomedical domain has focused on edited journal articles (and the genomics domain) rather than clinical free text in medical records. The fact remains, however, that the automation of healthcare workflows can bring important benefits to treatment (Hurtado et al., 2001) and reduce administrative burden, and that free text is a critical component of these workflows. There are economic motivations for the task, as well. The cost of adding labels like ICD-9-CM to clinical free text and the cost of repairing associated errors is approximately $25 billion per year in the US (Lang, 2007). For these (and many other) reasons, there have been consistent attempts to overcome the obstacles which hinder the processing of clinical text (Uzuner et al., 2006). This paper discusses one such attempt—The 2007 Computational Medicine Challenge, hereafter referred to as "the Challenge". There were two main reasons for conducting the Challenge. One is to facilitate advances in mining clinical free text; shared tasks in other biomedical domains have been shown to drive progress in the field in multiple ways (see (Hirschman and Blaschke, 2006; Hersh et al., 2005; Uzuner et al., 2006; Hersh et al., 2006) for a comprehensive review of biomedical challenge tasks and their contributions). The other is a ground-

BioNLP 2007: Biological, translational, and clinical language processing, pages 97–104,
Prague, June 2007.

breaking distribution of useful, reusable, carefully anonymized clinical data to the research community, whose data use agreement is simply to cite the source. The remaining sections of this paper describe how the data were prepared, the methods for scoring, preliminary results [to be updated if submission is accepted—results are currently still under analysis], and some lessons learned.

2 Corpus collection and coding process

Supervised methods for machine learning require training data. Yet, due to confidentiality requirements, spotty electronic availability, and variance in recording standards, the requisite clinical training data are difficult to obtain. One goal of the challenge was to create a publicly available ”gold standard” that could serve as the seed for a larger, open-source clinical corpus. For this we used the following guiding principles: individual identity must be expunged to meet United States HIPAA standards, (U.S. Health, 2002) and approved for release by the local Institutional Review Board (IRB); the sample must represent problems that medical records coders actually face; the sample must have enough data for machine-learning-based systems to do well; and the sample must include proportionate representations of very low-frequency classes.

Data for the corpus were collected from the Cincinnati Children's Hospital Medical Center's (CCHMC) Department of Radiology. CCHMC's Institutional Review Board approved release of the data. Sampling of all outpatient chest x-ray and renal procedures for a one-year period was done using a bootstrap method (Walters, 2004). These data are among those most commonly used, and are designed to provide enough codes to cover a substantial proportion of pediatric radiology activity. Expunging patient identity to meet HIPAA standards included three steps: disambiguation, anonymization, and data scrubbing (Pestian et al., 2005).

Ambiguity and Anonymization. Not surprisingly, some degree of disambiguation is needed to carry out effective anonymization (Uzuner et al., 2006; Sibanda and Uzuner, 2006). The reason is that clinical text is dense with medical jargon, abbreviations, and acronyms, many of which turn out to be ambiguous between a sense that needs anonymization and a different sense that does not. For example, in a clinical setting, *FT* can be an abbreviation for *full-term*, *fort* (as in *Fort Bragg*), *feet*, *foot*, *field test*, *full-time* or *family therapy*. *Fort Bragg*, being a place name, and a possible component of an address, could indirectly lead to identification of the patient. Until such occurrences are disambiguated, it is not possible to be certain that other steps to anonymize data are adequate. To resolve the relevant ambiguities found in this free text, we relied on previous efforts that used expert input to develop clinical disambiguation rules (Pestian et al., 2004).

Anonymization. To assure patient privacy, clinical text that is used for non-clinical reasons must be anonymized. However, to be maximally useful for machine-learning, this must be done in a particular way. Replacing personal names with some unspecific value such as ”*” would lose potentially useful information. Our goal is to replace the sensitive fields with *like* values that obscure the identity of the individual (Cho et al., 2002). We found that the amount of sensitive information routinely found in unstructured free text data is limited. In our case, these data included patient and physician names and sometimes dates or geographic locations, but little or no other sensitive information turned up in the relevant database fields. Using our internally developed encryption broker software, we replaced all female names with “Jane”, all male names with ”John”, and all surnames with ”Johnson”. Dates were randomly shifted.

Manual Inspection. Once the data were disambiguated and anonymized, they were manually reviewed for the presence of any Protected Health Information (PHI). If a specific token was perceived to potentially violate PHI regulations, the entire record was deleted from the dataset. In some case, however, a general geographic area was changed and not deleted. For example if the data read ”patient lived near Mr. Roger's neighborhood” it would be deleted, because it may be traceable. On the other hand, if the data read ”patient was from Cincinnati” it may have been changed to read ”patient was from the Covington” After this process, a corpus of 2,216 records was obtained (See Table 2 for details).

ICD-9-CM Assignment. A radiology report has multiple components. Two parts in particular are essential for the assignment of ICD-9-CM codes:

clinical history—provided by an ordering physician before a radiological procedure, and *impression*—reported by a radiologist after the procedure. In the case of radiology reports, ICD-9-CM codes serve as justification to have a certain procedure performed. There are official guidelines for radiology ICD-9-CM coding (Moisio, 2000). These guidelines note that every disease code requires a minimum number of digits before reimbursement will occur; that a definite diagnosis should always be coded when possible; that an uncertain diagnosis should never be coded; and that symptoms must never be coded when a definite diagnosis is available. Particular hospitals and insurance companies typically augment these principles with more specific internal guidelines and practices for coding. For these reasons of policy, and because of natural variation in human judgment, it is not uncommon for multiple annotators to assign different codes to the same text. Understanding the sources of this variation is important; so too is the need to create a definite gold standard for use in the challenge. To do so, data were annotated by the coding staff of CCHMC and two independent coding companies: COMPANY Y and COMPANY Z.

Majority annotation. A single gold standard was created from these three sets of annotations. There was no reason to adopt any *a priori* preference for one annotator over another, so the democratic principle of assigning a majority annotation was used. The majority annotation consists of those codes assigned to the document by two or more of the annotators. There are, however, several possible problems with this approach. For example, it could be that the majority annotation will be empty. This will be rare (126 records out of 2,216 in our case), because it only happens when the codes assigned by the three annotators form disjoint sets. In most hospital systems, including our own, the coders are required to indicate a primary code. By convention, the primary code is listed as the record's first code, and has an especially strong impact on the billing process. For simplicity's sake, the majority annotation process ignores the distinction between primary and secondary codes. There is space for a better solution here, but we have not seriously explored it. We have, however, conducted an analysis of agreement statistics (not further discussed here) that suggests that the overall effect of the majority method is to create a coding that shares many statistical properties with the originals, except that it reduces the effect of the annotators' individual idiosyncrasies. The majority annotation is illustrated in Table 1.

Our evaluation strategy makes the simplistic assumption that the majority annotation is a true gold standard and a worthwhile target for emulation. This is debatable, and is discussed below, but for the sake of definiteness we simply stipulate that submissions will be compared against the majority annotation, and that the best possible performance is to exactly replicate said majority annotation.

3 Evaluation

Micro- and macro-averaging. Although we rank systems for purposes of determining the top three performers on the basis of micro-averaged F1, we report a variety of performance data, including the micro-average, macro-average, and a cost-sensitive measure of loss. Jackson and Moulinier comment (for general text classification) that: "No agreement has been reached...on whether one should prefer micro- or macro-averages in reporting results. Macro-averaging may be preferred if a classification system is required to perform consistently across all classes regardless of how densely populated these are. On the other hand, micro-averaging may be preferred if the density of a class reflects its importance in the end-user system" (Jackson and Moulinier, 2002):160-161. For the present medical application, we are more interested in the number of patients whose cases are correctly documented and billed than in ensuring good coverage over the full range of diagnostic codes. We therefore emphasize the micro-average.

A cost-sensitive accuracy measure. While F-measure is well-established as a method for ranking, there are reasons for wanting to augment this with a cost-sensitive measure. An approach that allows penalties for over-coding (a false positive) and under-coding (a false negative) to be manipulated has important implications. The penalty for under-coding is simple—the hospital loses the amount of revenue that it would have earned if it had assigned the code. The regulations under which coding is done enforce an automatic over-coding penalty of

Table 1: Majority Annotation

	Hospital	**Company Y**	**Company Z**	**Majority**
Document 1	AB	BC	AB	AB
Document 2	BC	ABD	CDE	BCD
Document 3	EF	EF	E	EF
Document 4	ABEF	ACEF	CDEF	ACEF

three times what is earned from the erroneous code, with the additional risk of possible prosecution for fraud. This motivates a generalized version of Jaccard's similarity metric (Gower and Legendre, 1986), which was introduced by Boutell, Shen, Luo and Brown (Boutell et al., 2003).

Suppose that Y_x is the set of correct labels for a test set and P_x is the set of labels predicted by some participating system. Define $F_x = P_x - Y_x$ and $M_x = Y_x - P_x$, i.e. F_x is the set of false positives, and M_x is the set of missed labels or false negatives. The score is given by

$$score(P_x) = \left(1 - \frac{\beta|M_x| + \gamma|F_x|}{|Y_x \cup P_x|}\right)^{\alpha} \qquad (1)$$

As noted in (Boutell et al., 2003), if $\beta = \gamma = 1$ this formula reduces to the simpler case of

$$score(P_x) = \left(1 - \frac{|Y_x \cap P_x|}{|Y_x \cup P_x|}\right)^{\alpha} \qquad (2)$$

The discussion in (Boutell et al., 2003) points out that constraints are necessary on β and γ to ensure that the inner term of the expression is non-negative. We do not understand the way that they formulate these constraints, but note that non-negativity will be ensured if $0 \leq \beta \leq 1$ and $0 \leq \gamma \leq 1$. Since overcoding is three times as bad as undercoding, we use $\gamma = 1.0$, $\beta = 0.33$. Varying the value of α would affect the range of the scores, but does not alter the rankings of individual systems. We therefore used $\alpha = 1$. This measure does not represent the possibility of prosecution for fraud, because the costs involved are incommensurate with the ones that are represented. With these parameter settings, the cost-sensitive measure produces rankings that differ considerably from those produced by macro-averaged balanced F-measure. For example, we shall see that the system ranked third in the competition by macro-averaged F-measure assigns a total of 1167 labels, where the second-ranked assigns 1232, and the cost-sensitive measure rewards this conservatism in assigning labels by reversing the ranking of the two systems. In either case, the difference between the systems is small (0.86% difference in F-measure, 0.53% difference in the cost-sensitive measure).

4 The Data

We selected for the challenge a subset of the comprehensive data set described above. The subset was created by stratified sampling, such that it contains 20% of the documents in each category. Thus, the proportion of categories in the sample is the same as the proportion of categories in the full data set. We included in the initial sample only those categories to which 100 or more documents from the comprehensive data set were assigned. After the process summarized in Table 2, the data were divided into two partitions: a training set with 978 documents, and a testing set with 976. Forty-five ICD-9-CM labels (e.g 780.6) are observed in these data sets. These labels form 94 distinct combinations (e.g. the combination 780.6, 786.2). We required that any combination have at least two exemplars in the data, and we split each combination between the training and the test sets. So, there may be labels and combinations of labels that occur only one time in the training data, but participants can be sure that no combination will occur in the test data that has not previously occurred at least once in the training data. Our policy here has the unintended consequence that any combination that appears exactly once in the training data is highly likely to appear exactly once in the test data. This gives unnecessary information to the participants. In future challenges we will drop the requirement for two occurrences in the data, but ensure that single-occurrence combinations are allocated to the training set rather than the test set. This maintains the guarantee that there will be no unseen combinations in the test data. The full data set may be downloaded from the official challenge web-site.

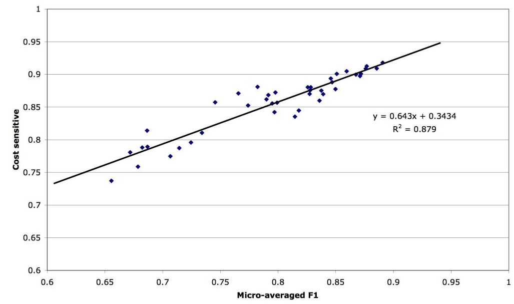
5 Results

Notice of the Challenge was distributed using electronic mailing lists supplied by the Association of Computational Linguistics, IEEE Computer Intelligence and Data Mining, and American Medical Informatics Association's Natural Language Processing special interest group. Interested participants were asked to register at the official challenge web-site. Registration began February 1, 2007 and ended February 28, 2007. Approximately 150 individuals registered from 22 countries and six continents. Upon completing registration, an automated e-mail was sent with the location of the training data. On March 1, 2007 participants received notice of the location of the testing data. Participants were encouraged to use the data for other purposes as long as it was non-commercial and the appropriate citation was made. There were no other data use restrictions. Participants had until March 18, 2007 to submit their results and an explanation of their model. Approximately 33% (50) of the participants submitted results. During the course of the Challenge participants asked a range of questions. These were posted to the official challenge web-site - www.computationalmedicine.org/challenge.

The figure below is a scatterplot relating micro-averaged F1 to the cost-sensitive measure described above. Each point represents a system. The top-performing systems achieved 0.8908, the minimum was 0.1541, and the mean was 0.7670, with a SD of 0.1340. There are 21 systems with a micro-averaged F1 between 0.81 and 0.90. Another 14 have $F1 > 0.70$. It is noticeable that the systems are not ranked identically by the cost-sensitive and the micro-averaged measure, but the differences are small in each case.

A preliminary screening using a two-factor ANOVA with system identity and diagnostic code as predictive factors for balanced F-measure revealed a significant main effect of both system and code. Pairwise t-tests using Holm's correction for multiple comparisons revealed no statistically significant differences between the systems performing at F=0.70 or higher. Differences between the top system and a system with a microaveraged F-measure of 0.66 do come out significant on this measure.

Figure 1: Scatter plot of evaluation measures



We have also calculated (Table 3) the agreement figures for the three individual annotations that went into the majority gold standard. We see that CCHMC outranks COMPANY Y on the cost-sensitive measure, but the reverse is true for micro- and macro-averaged F1, with the agreement between the hospital and the gold standard being especially low for the macro-averaged version. To understand these figures, it is necessary to recall that the gold standard is a majority annotation that is formed from the the three component annotations. It appears that for rare codes, which have a disproportionate effect on the macro-averaged F, the majority annotation is dominated by cases where company Y and company Z assign the same code, one that CCHMC did not assign.

The agreement figures are comparable to those of the best automatic systems. If submitted to the competition, the components of the majority annotation would not have outranked the best systems, even though the components contributed to the majority opinion. It is tempting to conclude that the automated systems are close to human-level performance. Recall, however, that while the hospital and the companies did not have the luxury of exposure to the majority annotation, the systems did have that access, which allowed them to explicitly model the properties of that majority annotation. A more moderate conclusion is that the hospital and the companies might be able to improve (or at least adjust) their annotation practices by studying the majority

Table 2: Characteristics of the data set through the development process.

Step	Removed	Total documents
One-year collection of documents		20,275
20 percent sample of one-year collection		4,055
Manual inspection for anonymization problems	1,839	2,216
Removal of records with no majority code	126	2,090
Removal of records with a code occurring only once	136	1,954

Table 3: Comparison of human annotators against majority.

Annotator	Cost-sensitive	Micro-averaged F1	Macro-averaged F1
HOSPITAL	0.9056	0.8264	0.6124
COMPANY Y	0.8997	0.8963	0.8973
COMPANY Z	0.8621	0.8454	0.8829

annotation and adapting as appropriate.

6 Discussion

Compared to other recent text classification shared tasks in the biomedical domain (Uzuner et al., 2006; Hersh et al., 2004; Hersh et al., 2005), this task required categorization with respect to a set of labels more than an order of magnitude larger than previous evaluations. This increase in the size of the set of labels is an important step forward for the field–systems that perform well on smaller sets of categories do not necessarily perform well with larger sets of categories (Jackson and Moulinier, 2002), so the data set will allow for more thorough text categorization system evaluations than have been possible in the past. Another important contribution of the work reported here may be the distribution of the data—the first fully distributable, freely usable data set of clinical text. The high number of participants and final submissions was a pleasant surprise; we attribute this, among other things, to the fact that this was an applied challenge, that real data were supplied, and that participants were free to use these data in other venues.

Participants utilized a diverse range of approaches. These system descriptions are based on brief comments entered into the submission box, and are obviously subject to revision. The three highest scorers all mentioned "negation," all seemed to be using the structure of UMLS in a serious way. The better systems frequently mentioned "hypernyms" or "synonyms," and were apparently doing significant amounts of symbolic processing. Two of the top three had machine-learning components, while one of the top three used purely symbolic methods. The most common approach seems to be thoughtful and medically-informed feature engineering followed by some variety of machine learning. The top-performing system used C4.5, suggesting that use of the latest algorithms is not a pre-requisite for success. SVMs and related large-margin approaches to machine learning were strongly represented, but did not seem to be reliably predictive of high ranking.

6.1 Observations on running the task and the evaluation

The most frequently viewed question of the FAQ was related to a script to calculate the evaluation score. This was supplied both as a downloadable script and as an interactive web-page with a form for submission. In retrospect, we realize that we had not fully thought through what would happen as people began to use this script. If we run a similar contest in the future, we will be better prepared for the confusion that this can cause.

A novel aspect of this task was that although we only scored a single run on the test data, we allowed participants to submit their "final" run up to 10 times, and to see their score each time. Note that although participants could see how their score varied on successive submissions, they did *not* have access to the actual test data or to the correct answers, and so there were no opportunities for special-purpose hacks to handle special cases in the test data. The average participant tried 5.27 (SD 3.17) submissions against the test data. About halfway through the submission period we began to realize that in a competitive situation, there are risks in providing the type of feedback given on the submission form. In future challenges, we will be judicious in selecting the number of attempts allowed and the provision of any type of feedback. As far as we can tell our general assumption that the scientific integrity of the participants was greater than the need to game the system is true. It is good policy for those administering the contest, however, to keep temptations to a minimum. Our current preference would be to provide only the web-page interface with no more than five attempts, and to tell participants only whether their submission had been accepted, and if so, how many items and how many codes were recognized.

We provided an XML schema as a precise and publicly visible description of the submission format. Although we should not have been, we were surprised when changes to the schema were required in order to accommodate small but unexpected variations in participant submissions. An even simpler submission format would have been good. The advantage of the approach that we took was that XML validation gave us a degree of sanity-checking at little cost. The disadvantage was that some of the necessary sanity-checking went beyond what we could see how to do in a schema.

The fact that numerous participants generated systems with high performance indicates that the task was reasonable, and that sufficient information about the coding task was either provided by us or inferred by the participants to allow them to do their work. Since this is a first attempt, it is not yet clear what the upper limits on performance are for this task, but preliminary indications are that automated systems are or will soon be viable as a component of deployed systems for this kind of application.

7 Acknowledgements

The authors thank Aaron Cohen of the Oregon Health and Science University for observations on the inter-rater agreement between the three sources and its relationship to the majority assignments, and also for his input on testing for statistically significant differences between systems. We also thank PERSON of ORGANIZATION for helpful comments on the manuscript. Most importantly we thank all the participants for their on-going commitment, professional feedback and scientific integrity.

References

[Boutell et al., 2003] Boutell M., Shen X., Luo J. and Brown C. 2003. *Multi-label Semantic Scene Classification*, Technical Report 813. Department of Computer Science, University of Rochester September.

[Cho et al., 2002] Cho P. S., Taira R. K., and Kangarloo H. 2002 Text boundary detection of medical reports. *Proceedings of the Annual Symposium of the American Medical Informatics Association*, 998.

[Friedman et al., 2002] Friedman C., Kra P., and Rzhetsky A. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35:222–235.

[Gower and Legendre, 1986] Gower J. C. and Legendre P. 1986. Metric and euclidean properties of dissimilarity coefficient. *Journal of Classification*, 3:5–48.

[Hersh et al., 2004] Hersh W., Bhupatiraju R. T., Ross L., Roberts P., Cohen A. M., and Kraemer D. F. 2004. TREC 2004 Genomics track overview. *Proceedings of the 13th Annual Text Retrieval Conference*. National Institute of Standards and Technology.

[Hersh et al., 2006] Hersh W., Cohen A. M., Roberts P., and Rekapalli H. K. 2006. TREC 2006 Genomics track overview. *Proceedings of the 15th Annual Text Retrieval Conference* National Institute of Standards and Technology.

[Hersh et al., 2005] Hersh W., Cohen A. M., Yang J., Bhupatiraju R. T., Roberts P., and Hearst M. 2005. TREC 2005 Genomics track overview. *Proceedings of the 14th Annual Text Retrieval Conference*. National Institute of Standards and Technology.

[Hirschman and Blaschke, 2006] Hirschman L. and Blaschke C. 2006. Evaluation of text mining in biology. *Text mining for biology and biomedicine*, Chapter 9. Ananiadou S. and McNaught J., editors. Artech House.

[Hirschman and Sager, 1982] Hirschman L. and Sager S. 1982. Automatic information formatting of a medical sublanguage. *Sublanguage: studies of language in restricted semantic domains*, Chapter 2. Kittredge R. and Lehrberger J., editors. Walter de Gruyter.

[Hurtado et al., 2001] Hurtado M. P, Swift E. K., and Corrigan J. M. 2001. Crossing the Quality Chasm: A New Health System for the 21st Century. Institute of Medicine, National Academy of Sciences.

[Jackson and Moulinier, 2002] Jackson P. and Moulinier I. 2002. *Natural language processing for online applications: text retrieval, extraction, and categorization*. John Benjamins Publishing Co.

[Lang, 2007] Lang, D. 2007. CONSULTANT REPORT - Natural Language Processing in the Health Care Industry. Cincinnati Children's Hospital Medical Center, Winter 2007.

[Moisio, 2000] Moisio M. 2000. *A Guide to Health Care Insurance Billing*. Thomson Delmar Learning, Clifton Park.

[Pestian et al., 2005] Pestian J. P., Itert L., Andersen C. L., and Duch W. 2005. Preparing Clinical Text for Use in Biomedical Research. *Journal of Database Management*, 17(2):1-12.

[Pestian et al., 2004] Pestian J. P., Itert L., and Duch W. 2004. Development of a Pediatric Text-Corpus for Part-of-Speech Tagging. *Intelligent Information Processing and Web Mining, Advances in Soft Computing*, 219–226 New York, Springer Verlag.

[Sammuelsson and Wiren, 2000] Sammuelsson C. and Wiren M. 2000. Parsing Techniques. *Handbook of Natural Language Processing*, 59–93. Dale R., Moisl H., Somers H., editors. New York, Marcel Deker.

[Sibanda and Uzuner, 2006] Sibanda T. and Uzuner O. 2006. Role of local context in automatic deidentification of ungrammatical, fragmented text. *Proceedings of the Human Language Technology conference of the North American chapter of the Association for Computational Linguistics*, 65–73.

[Stetson et al., 2002] Stetson P. D., Johnson S. B., Scotch M., and Hripcsak G. 2002. The sublanguage of cross-coverage. *Proceedings of the Annual Symposium of the American Medical Informatics Association*, 742–746.

[U.S. Health, 2002] U.S. Heath & Human Services. 2002. 45 CFR Parts 160 and 164 Standards for Privacy of Individually Identifiable Health Information *Final Rule Federal Register*, 67(157):53181–53273.

[Uzuner et al., 2006] Uzuner O., Szolovits P., and Kohane I. 2006. i2b2 workshop on natural language processing challenges for clinical records. *Proceedings of the Fall Symposium of the American Medical Informatics Association*.

[Walters, 2004] Walters S. J. 2004. Sample size and power estimation for studies with health related quality of life outcomes: a comparison of four methods using the SF-36 *Health and Quality of Life Outcomes*, 2:26.