Earlier Identification of Epilepsy Surgery Candidates Using Natural Language Processing

Pawel Matykiewicz[1], Kevin Bretonnel Cohen[2], Katherine D. Holland[1], Tracy A. Glauser[1], Shannon M. Standridge[1], Karin M. Verspoor[3,4], and John Pestian[1]§
[1] Cincinnati Children's Hospital Medical Center, Cincinnati OH USA
[2] University of Colorado, Denver, CO
[3] National ICT Australia and [4]The University of Melbourne, Melbourne, Australia
§corresponding author: john.pestian@cchmc.org

Abstract

This research analyzed the clinical notes of epilepsy patients using techniques from corpus linguistics and machine learning and predicted which patients are candidates for neurosurgery, i.e. have intractable epilepsy, and which are not. Information-theoretic and machine learning techniques are used to determine whether and how sets of clinic notes from patients with intractable and non-intractable epilepsy are different. The results show that it is possible to predict from an early stage of treatment which patients will fall into one of these two categories based only on text data. These results have broad implications for developing clinical decision support systems.

1 Introduction and Significance

Epilepsy is a disease characterized by recurrent seizures that may cause irreversible brain damage. While there are no national registries, epidemiologists have shown that roughly three million Americans require $17.6 billion USD in care annually to treat their epilepsy (Epilepsy Foundation, 2012; Begley et al., 2000). Epilepsy is defined by the occurrence of two or more unprovoked seizures in a year. Approximately 30% of those individuals with epilepsy will have seizures that do not respond to anti-epileptic drugs (Kwan and Brodie, 2000). This population of individuals is said to have intractable or drug-resistant epilepsy (Kwan et al., 2010).

Select intractable epilepsy patients are candidates for a variety of neurosurgical procedures that ablate the portion of the brain known to cause the seizure. On average, the gap between the initial clinical visit when the diagnosis of epilepsy is made and surgery is six years. If it were possible to predict which patients should be considered candidates for referral to surgery earlier in the course of treatment, years of damaging seizures, under-employment, and psychosocial distress may be avoided. It is this gap that motivates this research.

In this study, we examine the differences between the clinical notes of patients early in their treatment course with the intent of predicting which patients will eventually be diagnosed as intractable versus which will be amenable to drug-based treatment. The null hypothesis is that there will be no detectable differences between the clinic notes of patients who go on to a diagnosis of intractable epilepsy and patients who do not progress to the diagnosis of intractable epilepsy (figure 1). To further elucidate the phenomenon, we look at both the patient's earliest clinical notes and notes from a progression of time points. Here we expect to gain insight into how the linguistic characteristics (and natural language processing-based classification performance) evolve over treatment course. We also study the linguistic features that characterize the differences between the document sets from the two groups of patients. We anticipate that this approach will ultimately be adapted for various clinical decision support systems.

2 Background

2.1 Related work

Although there has been extensive work on building predictive models of disease progression and of mortality risk, few models take advantage of natural language processing in addressing this task.

(Abhyankar et al., 2012) used univariate analysis, multivariate logistic regression, sensitivity analyses, and Cox proportional hazards models to predict 30-day and 1-year survival of overweight

Proceedings of the 2013 Workshop on Biomedical Natural Language Processing (BioNLP 2013), pages 1–9,
Sofia, Bulgaria, August 4-9 2013.

and obese Intensive Care Unit patients. As one of the features in their system, they used smoking status extracted from patient records by natural language processing techniques.

(Himes et al., 2009) used a Bayesian network model to predict which asthma patients would go on to develop chronic obstructive pulmonary disease. As one of their features, they also used smoking status extracted from patient records by natural language processing techniques.

(Huang et al., under review) is the work most similar to our own. They evaluated the ability of a Naive Bayesian classifier to predict future diagnoses of depression six months prior and twelve months prior to the actual diagnoses. They used a number of feature types, including fielded data such as billing codes, ICD-9 CM diagnoses, and others, as well as data drawn from natural language processing.

In particular, they used an optimized version of the NCBO Annotator (Jonquet et al., 2009) to recognize terms from 22 clinically relevant ontologies and classify them additionally as to whether they were negated or related to the patient's family history. Their system demonstrated an ability to predict diagnoses of depression both six months and one year prior to the actual diagnoses at a rate that exceeds the success of primary care practitioners in diagnosing active depression.

Considering this body of work overall, natural language processing techniques have played a minor role, providing only a fraction of a much larger set of features—just one feature, in the first two studies discussed. In contrast, in our work natural language processing is the central aspect of the solution.

2.2 Theoretical background to the approaches used in this work

In comparing the document sets from the two patient populations, we make use of two lines of inquiry. In the first, we use information-theoretic methods to determine whether or not the contents of the data sets are different, and if they are different, to characterize the differences. In the second, we make use of a practical method from applied machine learning. In particular, we determine whether it is possible to train a classifier to distinguish between documents from the two sets of patients, given an appropriate classification algorithm and a reasonable set of features.

From information-theoretic methods, we take Kullback-Leibler divergence as a way to determine whether the contents of the two sets of documents are the same or different. Kullback-Leibler divergence is the relative entropy of two probability mass functions—"a measure of how different two probability distributions (over the same event space) are" (Manning and Schuetze, 1999). This measure has been previously used to assess the similarity of corpora (Verspoor et al., 2009). Details of the calculation of Kullback-Leibler divergence are given in the Methods section. Kullback-Leibler divergence has a lower bound of zero; with a value of zero, the two document sets would be identical. A value of 0.005 is assumed to correspond to near-identity.

From practical applications of machine learning, we test whether or not it is possible to train a classifier to distinguish between documents from the two document sets. The line of thought here is that provided that we have an appropriate classification algorithm and a reasonable feature set, then if clinic notes from the two document sets are indeed different, it should be possible to train a classifier to distinguish between them with reasonable accuracy.

3 Materials and methods

3.1 Materials

The experimental protocol was approved by our local Institutional Review Board (#2012-1646). Neurology clinic notes were extracted from the electronic medical record system. Records were sampled from two groups of patients: 1) those with intractable epilepsy referred for and eventually undergoing epilepsy surgery and 2) those with epilepsy who were responsive to medications and never referred for surgical evaluation. They were also sampled at three time periods before the "zero point", the date at which patients were either referred for surgery or the date of last seizure for the non-intractable group. Table 1 shows the distribution of patients and clinic notes.

3.2 Methods

As described in the introduction, we applied information-theoretic and machine learning techniques to determine whether the two document collections were different (or differentiable).

	Non-Intractable	Intractable
-12 to 0	355 (127)	641 (155)
-6 to +6	453 (128)	898 (155)
0 to +12 months	454 (132)	882 (149)

Table 1: Progress note and patient counts (in parentheses) for each time period. A minus sign indicates the period before surgery referral date for intractable epilepsy patients and before last seizure for non-intractable patients. A plus sign indicates the period after surgery referral for intractable epilepsy patients and after last seizure for non-intractable patients. Zero is the surgery referral date or date of last seizure for the two populations, respectively.

3.2.1 Feature extraction

Features for both the calculation of Kullback-Leibler divergence and the machine learning experiment were unigrams, bigrams, trigrams, and quadrigrams. We applied the National Library of Medicine stopword list `http://mbr.nlm.nih.gov/Download/2009/WordCounts/wrd_stop`. All words were lower-cased, all numerals were substituted with the string *NUMB* for abstraction, and all non-ASCII characters were removed.

3.3 Information-theoretic approach

Kullback-Leibler divergence compares probability distribution of words or n-grams between different datasets $D_{KL}(P||Q)$. In particular, it measures how much information is lost if distribution Q is used to approximate distribution P. This method, however, gives an asymmetric dissimilarity measure. **Jensen-Shannon divergence** is probably the most popular symmetrization of D_{KL} and is defined as follows:

$$D_{JS} = \frac{1}{2}D_{KL}(P||Q) + \frac{1}{2}D_{KL}(Q||P) \quad (1)$$

where

$$D_{KL}(P||Q) = \sum_{w \in P \cup Q} \left(p(w|c_P) \log \frac{p(w|c_P)}{p(w|c_Q)} \right) \quad (2)$$

By Zipf's law any corpus of natural language will have a very long tail of infrequent words. To account for this effect we use D_{JS} for the top N most frequent words/n-grams. We use Laplace smoothing to account for words or n-grams that did not appear in one of the corpora.

We also aim to uncover terms that distinguish one corpus from another. We use a metamorphic D_{JS} test, log-likelihood ratios, and weighted SVM features. Log-likelihood score will help us understand where precisely the two corpora differ.

$$n_{ij} = \frac{k_{ij}}{k_{iP} + k_{iA}} \quad (3)$$

$$m_{ij} = \frac{k_{Pj} + k_{Qj}}{k_{QP} + k_{PP} + k_{QA} + k_{PA}} \quad (4)$$

$$LL(w) = 2 \sum_{i,j} k_{ij} \log \frac{n_{ij}}{m_{ij}} \quad (5)$$

3.4 Machine learning

For the classification experiment, we used an implementation of the libsvm support vector machine package that was ported to R (Dimitriadou et al., 2011). Features were extracted as described above in Section 3.2.1. We used a cosine kernel. The optimal C regularization parameter was estimated on a scale from 2^{-1} to 2^{15}.

3.5 Characterizing differences between the document sets

We used a variety of methods to characterize differences between the document sets: log-likelihood ratio, SVM normal vector components, and a technique adapted from metamorphic testing.

3.5.1 Applying metamorphic testing to Kullback-Leibler divergence

As one of our methods for characterizing differences between the two document sets, we used an adaptation of metamorphic testing, inspired by the work of (Murphy and Kaiser, 2008) on applying metamorphic testing to machine learning applications. The intuition behind metamorphic testing is that given some output for a given input, it should be possible to predict in general terms what the effect of some alternation in the input should be on the output. For example, given some Kullback-Leibler divergence for some set of features, it is possible to predict how Kullback-Leibler divergence will change if a feature is added to or subtracted from the feature vector. We adapted this observation by iteratively subtracting all features one by one and ranking them according to how much of an effect on the Kullback-Leibler divergence its removal had.

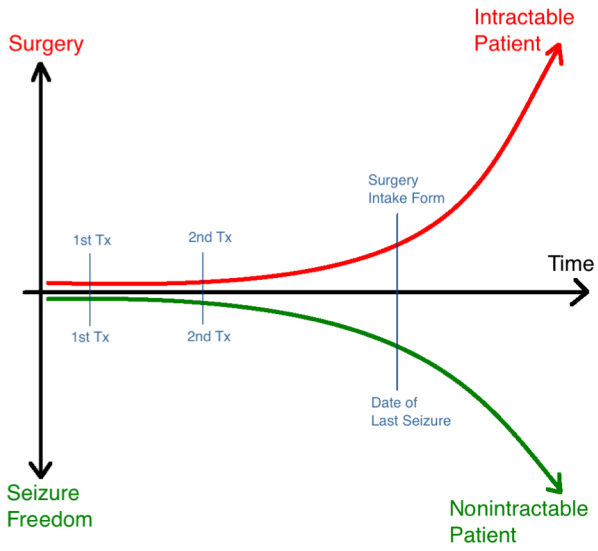

Figure 1: Two major paths in epilepsy care. At the begining of epilepsy care two groups of patients are indistinguishable. Subsequently, the two groups diverge.

4 Results

4.1 Kullback-Leibler (Jensen-Shannon) divergence

Table 2 shows the Kullback-Leibler divergence, calculated as Jensen-Shannon divergence, for three overlapping time periods—the year preceding surgery referral, the period from 6 months before surgery referral to six months after surgery referral, and the year following surgery referral, for the intractable epilepsy patients; and, for the non-intractable epilepsy patients, the same time periods with reference to the last seizure date.

As can be seen in the left-most column (-12 to 0), at one year prior, the clinic notes of patients who will require surgery and patients who will not require surgery cannot easily be discriminated by Kullback-Leibler divergence—the divergence is only just above the .005 near-identity threshold even when 8000 unique n-grams are considered. If the -6 to +6 and 0 to +12 time periods are examined, we see that the divergence increases as we reach and then pass the period of surgery (or move into the year following the last seizure, for the non-intractable patients), indicating that the difference between the two collections becomes more pronounced as treatment progresses. The divergence for these time periods does pass the assumed near-identity threshold for larger numbers of n-grams,

n-grams	-12 to 0 months	-6 to +6 months	0 to +12 months
125	0.00125	0.00193	0.00244
250	0.00167	0.00229	0.00286
500	0.00266	0.00326	0.00389
1000	0.00404	0.00494	0.00585
2000	0.00504	0.00618	0.00718
4000	0.00535	0.00657	0.00770
8000	0.00555	0.00681	0.00796

Table 2: Kullback-Leibler divergence (calculated as Jensen-Shannon divergence) for difference between progress notes of the two groups of patients. Results are shown for the period 1 year before, 6 months before and 6 months after, and one year after surgery referral for the intractable epilepsy patients and the last seizure for non-intractable patients. 0 represents the date of surgery referral for the intractable epilepsy patients and date of last seizure for the non-intractable patients.

largely accounted for by terms that are unique to one notes set or the other.

4.2 Classification with support vector machines

Table 3 shows the results of building support vector machines to classify individual notes as belonging to the intractable epilepsy or the non-intractable epilepsy patient population. Three time periods are evaluated, as described above. The number of features is varied by row. For each cell, the average F-measure from 20-fold cross-validation is shown.

As can be seen in the left-most column (-12 to 0), at one year prior to referral to surgery referral date or last seizure, the patients who will become intractable epilepsy patients can be distinguished from the patients who will become non-intractable epilepsy patients *purely on the basis of natural language processing-based classification* with an F-measure as high as 0.95. This supports the conclusion that the two document sets are indeed different, and furthermore illustrates that this difference can be used to predict which patients will require surgical intervention.

4.3 Characterizing the differences between clinic notes from the two patient populations

Tables 4 and 5 show the results of three methods for differentiating between the document col-

n-grams	-12 to 0 months	-6 to +6 months	0 to +12 months
125	0.8885	0.9217	0.9476
250	0.8928	0.9297	0.9572
500	0.9107	0.9367	0.9667
1000	0.9245	0.9496	0.9692
2000	0.9417	0.9595	0.9789
4000	0.9469	0.9661	0.9800
8000	0.9510	0.9681	0.9810

Table 3: Average F_1 for the three time periods described above, with increasing numbers of features. Values are the average of 20-fold cross-validation. See Figure 2 for an explanation of the time periods.

lections representing the two patient populations. The methodology for each is described above. The most strongly distinguishing features when just the 125 most frequent features are used are shown in Table 4, and the most strongly distinguishing features when the 8,000 most frequent features are used are shown in Table 5. Impressionistically, two trends emerge. One is that more clearly clinically significant features are shown to have strong discriminatory power when the 8,000 most frequent features are used than when the 125 most frequent features are used. This result is supported by the Kullback-Leibler divergence results, which demonstrated the most divergent vocabularies with larger numbers of n-grams. The other trend is that the SVM classifier does a better job of picking out clinically relevant features. This has implications for the design of clinical decision support systems that utilize our approach.

5 Discussion

5.1 Behavior of Kullback-Leibler divergence

Kullback-Leibler divergence varies with the number of words considered. When the vocabularies of two document sets are merged and the words are ordered by overall frequency, the further down the list we go, the higher the Kullback-Leibler divergence can be expected to be. This is because the highest-frequency words in the combined set will generally be frequent in both source corpora, and therefore carry similar probability mass. As we progress further down the list of frequency-ranked words, we include progressively less-common words, with diverse usage patterns, which are likely to reflect the differences between the two document sets, if there are any. Thus, the Kullback-Leibler divergence will rise.

To understand the intuition here, imagine looking at the Kullback-Leibler divergence when just the 50 most-common words are considered. These will be primarily function words, and their distributions are unlikely to differ much between the two document sets unless the syntax of the two corpora is radically different. Beyond this set of very frequent common words will be words that may be relatively frequent in one set as compared to the other, contributing to divergence between the sets.

In Table 2, the observed behavior for our two document collections follows this expected pattern. However, the divergence between the vocabularies remains close to the assumed near-identity threshold of 0.005, even when larger numbers of n-grams are considered. The divergence never exceeds 0.01; this level of divergence for larger numbers of n-grams is consistent with prior analyses of highly similar corpora (Verspoor et al., 2009).

We attribute this similarity to two factors. The first is that both document sets derive from a single department within a single hospital; a relatively small number of doctors are responsible for authoring the notes and there may exist specific hospital protocols related to their content. The second is that the clinical contexts from which our two document sets are derived are highly related, in that all the patients are epilepsy patients. While we have demonstrated that there are clear differences between the two sets, it is also to be expected that they would have many words in common. The nature of clinical notes combined with the shared disease context results in generally consistent vocabulary and hence low overall divergence.

5.2 Behavior of classifier

Table 3 demonstrates that classifier performance increases as the number of features increases. This indicates that as more terms are considered, the basis for differentiating between the two different document collections is stronger.

Examining the SVM normal vector components (SVMW) in Tables 4 and 5, we find that unigrams, bigrams and trigrams are useful in differentiation between the two patient populations. While no quadrigrams appear in this table, they may in fact contribute to classifier performance. We will perform an ablation study in future work to quantify

JS metamorphic test (JSMT)	**Log-likelihood ratio (LLR)**	**SVM normal vector components (SVMW)**
family = -0.000114	none = 623.702323	bilaterally = -19.009380
normal = -0.000106	family = -445.117177	age.NUMB = 17.981459
seizure = -0.000053	NUMB.NUMB.NUMB.NUMB = 422.953816	review = 17.250652
problems = -0.000053	normal = -244.603033	based = -14.846495
none = 0.000043	problems = -207.021130	family.history = -14.659653
detailed = -0.000037	left = 176.434519	NUMB = -14.422525
including = -0.000036	bid = 142.105691	lower = -13.553434
risks = -0.000033	NUMB = 136.255678	mother = -13.436694
NUMB = 0.000032	detailed = -133.012908	first = -13.001744
concerns = -0.000032	right = 120.453596	including = -12.800433
NUMB.NUMB.NUMB.NUMB = 0.000031	seizure = -120.047686	extremities = 11.709199
additional = -0.000029	including = -119.061518	documented = -11.441394
brain = -0.000026	risks = -116.543250	awake = -11.418535
NUMB.NUMB = 0.000022	concerns = -101.366110	hpi = 11.121019
minutes = -0.000021	additional = -95.880792	follow = -10.550802
NUMB.minutes = -0.000020	clear = 83.848170	neurology = -10.533895
reviewed = -0.000018	brain = -74.267220	call = -10.422606
history = -0.000017	seizures = 71.937757	effects = 10.298221
noted = -0.000017	one = 65.203819	brain = -9.900864
upper = -0.000017	epilepsy = 46.383564	weight = 9.819712
well = -0.000015	hpi = 45.932630	patient.s = -9.603531
side = -0.000015	minutes = -45.278770	discussed = -9.473544
bilaterally = -0.000014	NUMB.NUMB.NUMB = 43.320354	today = 9.390896
motor.normal = -0.000014	negative = 42.914770	allergies = -9.346146
notes = -0.000014	NUMB.minutes = -42.909968	NUMB.NUMB.NUMB.NUMB = 9.342800
Spearman correlation between JSMT and LLR = 0.912454	Spearman correlation between LLR and SVMW = 0.086784	Spearman correlation between SVMW and JSMT = 0.101965

Table 4: Comparison of three different methods for finding the strongest differentiating features. This table shows features for the -12 to 0 periods with the 125 most frequent features. The JSMT and LLR statistics give values greater than zero. We add sign to indicate which corpus has higher relative frequency of the feature: a positive value indicates that the relative frequency of the feature is greater in the intractable group, while a negative value indicates that the relative frequency of the feature is greater in the non-intractable group. The last row shows the correlation between two different ranking statistics.

JS metamorphic test (JSMT)	**Log-likelihood ratio (LLR)**	**SVM normal vector components (SVMW)**
family = -0.000118	family = -830.329965	john = -4.645071
normal = -0.000109	normal = -745.882086	lamotrigine = 4.320412
seizure = -0.000057	problems = -386.238711	surgery = 4.299546
problems = -0.000057	seizure = -369.342334	jane = 4.091609
none = 0.000047	none = 337.461504	epilepsy.surgery = 4.035633
including = -0.000040	detailed = -262.240496	janet = -3.970101
detailed = -0.000040	including = -255.076808	excellent.control = -3.946283
additional.concerns = -0.000038	additional.concerns.noted = -246.603655	excellent = -3.920620
additional.concerns.noted = -0.000038	concerns.noted = -246.603655	NUMB.seizure = -3.886997
concerns.noted = -0.000038	additional.concerns = -243.353912	mother = -3.801364
NUMB = -0.000036	NUMB.NUMB.NUMB.NUMB = 238.065700	jen = 3.568809
concerns = -0.000036	risks = -232.741511	back = -3.319477
risks = -0.000036	concerns = -228.805299	visit = -3.264600
NUMB.NUMB.NUMB.NUMB = 0.000035	additional = -204.462411	james = 3.174763
additional = -0.000033	brain = -182.413340	NUMB.NUMB.NUMB.normal = -3.024471
brain = -0.000030	NUMB = -162.992065	continue = -3.011293
NUMB.NUMB = -0.000026	surgery = 153.646067	idiopathic.localization = -2.998177
minutes = -0.000025	minutes = -142.761961	idiopathic.localization.related = -2.998177
surgery = 0.000024	NUMB.minutes = -134.048116	increase = 2.948187
NUMB.minutes = -0.000023	diff = -131.388230	diastat = -2.937431
diff = -0.000023	NUMB.NUMB = -125.067347	taking = -2.902673
history = -0.000021	reviewed = -116.013417	lamictal = 2.898987
reviewed = -0.000021	noted = -114.241532	going = 2.862764
noted = -0.000021	idiopathic = -112.331060	described = 2.844830
upper = -0.000020	shaking = -112.186858	epilepsy = 2.745872
Spearman correlation between JSMT and LLR = 0.782918	Spearman correlation between LLR and SVMW = 0.039860	Spearman correlation between SVMW and JSMT = 0.165159

Table 5: Comparison of three different methods for finding the strongest differentiating features. This table shows features for the -12 to 0 periods with the 8,000 most frequent features. The JSMT and LLR statistics give values greater than zero. We add sign to indicate which corpus has higher relative frequency of the feature: a positive value indicates that the relative frequency of the feature is greater in the intractable group, while a negative value indicates that the relative frequency of the feature is greater in the non-intractable group. The last row shows the correlation between two different ranking statistics.

the contribution of the different feature sets. In addition, we find that table 5 shows many clinically relevant terms, such as seizure frequency ("excellent [seizure] control"), epilepsy type ("localization related [epilepsy]"), etiology classification ("idiopathic [epilepsy]"), and drug names ("lamotrigine", "diastat", "lamictal"), giving nearly complete history of the present illness.

6 Conclusion

The classification results from our machine learning experiments support rejection of the null hypothesis of no detectable differences between the clinic notes of patients who will progress to the diagnosis of intractable epilepsy and patients who do not progress to the diagnosis of intractable epilepsy. The results show that we can predict from an early stage of treatment which patients will fall into these two classes based only on textual data from the neurology clinic notes. As intuition would suggest, we find that the notes become more divergent and the ability to predict outcome improves as time progresses, but the most important point is that the outcome can be predicted from the earliest time period.

SVM classification demonstrates a stronger result than the information-theoretic measures, uses less data, and needs just a single run. However, it is important to note that we cannot entirely rely on the argument from classification as the sole methodology in testing whether or not two document sets are similar or different. If the finding is positive, i.e., it is possible to train a classifier to distinguish between documents drawn from the two document sets, then interpreting the results is straightforward. However, if documents drawn from the two document sets are not found to be distinguishable by a classifier, one must consider the possibility of multiple possible confounds, such as selection of an inappropriate classification algorithm, extraction of the wrong features, bugs in the feature extraction software, etc. Having established that the two sets of clinical notes differ, we noted some identifying features of clinic notes from the two populations, particularly when more terms were considered.

The Institute of Medicine explains that "...to accommodate the reality that although professional judgment will always be vital to shaping care, the amount of information required for any given decision is moving beyond unassisted human capacity (Olsen et al., 2007)." This is surely the case for those who care for the epileptic patient. Technology like natural language processing will ultimately serve as a basis for stable clinical decision support tools. It, however, is not a decision making tool. Decision making is the responsibility of professional judgement. That judgement will labor over such questions as: what is the efficacy of neurosurgery, what will be the long term outcome, will there be any lasting damage, are we sure that all the medications have been tested, and how the family will adjust to a poor outcome. In the end, it is that judgement that will decide what is best; that decision will be supported by research like what is presented here.

7 Acknowledgements

This work was supported in part by the National Institutes of Health, Grants #1R01LM011124-01,and 1R01NS045911-01; the Cincinnati Children's Hospital Medical Center's: Research Foundation, Department of Pediatric Surgery and the Department of Paediatrics's divisions of Neurology and Biomedical Informatics. We also wish to acknowledge the clinical and surgical wisdom provided by Drs. John J. Hutton & Hansel M. Greiner, MD. K. Bretonnel Cohen was supported by grants XXX YYY ZZZ. Karin Verspoor was supported by NICTA, which is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council.

References

[Abhyankar et al.2012] Swapna Abhyankar, Kira Leishear, Fiona M. Callaghan, Dina Demner-Fushman, and Clement J. McDonald. 2012. Lower short- and long-term mortality associated with overweight and obesity in a large cohort study of adult intensive care unit patients. *Critical Care*, 16.

[Begley et al.2000] Charles E Begley, Melissa Famulari, John F Annegers, David R Lairson, Thomas F Reynolds, Sharon Coan, Stephanie Dubinsky, Michael E Newmark, Cynthia Leibson, EL So, et al. 2000. The cost of epilepsy in the united states: An estimate from population-based clinical and survey data. *Epilepsia*, 41(3):342–351.

[Dimitriadou et al.2011] Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer, and Andreas Weingessel, 2011. *e1071: Misc Functions of the Department of Statistics (e1071), TU*

Wien. http://CRAN.R-project.org/package=e1071. R package version 1.5.

[Epilepsy Foundation2012] Epilepsy Foundation, 2012. *What is Epilepsy: Incidence and Prevalence*. http://www.epilepsyfoundation.org/ aboutepilepsy /whatisepilepsy/ statistics.cfm.

[Himes et al.2009] Blanca E. Himes, Yi Dai, Isaac S. Kohane, Scott T. Weiss, and Marco F. Ramoni. 2009. Prediction of chronic obstructive pulmonary disease (copd) in asthma patients using electronic medical records. *Journal of the American Medical Informatics Association*, 16(3):371–379.

[Huang et al.under review] Sandy H. Huang, Paea LePendu, Srinivasan V Iyer, Anna Bauer-Mehren, Cliff Olson, and Nigam H. Shah. under review. Developing computational models for predicting diagnoses of depression. In *American Medical Informatics Association*.

[Jonquet et al.2009] Clement Jonquet, Nigam H. Shah, Cherie H. Youn, Mark A. Musen, Chris Callendar, and Margaret-Anne Storey. 2009. NCBO Annotator: Semantic annotation of biomedical data. In *8th International Semantic Web Conference*.

[Kwan and Brodie2000] Patrick Kwan and Martin J Brodie. 2000. Early identification of refractory epilepsy. *New England Journal of Medicine*, 342(5):314–319.

[Kwan et al.2010] Patrick Kwan, Alexis Arzimanoglou, Anne T Berg, Martin J Brodie, W Allen Hauser, Gary Mathern, Solomon L Moshé, Emilio Perucca, Samuel Wiebe, and Jacqueline French. 2010. Definition of drug resistant epilepsy: consensus proposal by the ad hoc task force of the ilae commission on therapeutic strategies. *Epilepsia*, 51(6):1069–1077.

[Manning and Schuetze1999] Christopher Manning and Hinrich Schuetze. 1999. *Foundations of statistical natural language processing*. MIT Press.

[Murphy and Kaiser2008] Christian Murphy and Gail Kaiser. 2008. Improving the dependability of machine learning applications.

[Olsen et al.2007] LeighAnne Olsen, Dara Aisner, and J Michael McGinnis. 2007. The learning healthcare system.

[Verspoor et al.2009] K. Verspoor, K.B. Cohen, and L. Hunter. 2009. The textual characteristics of traditional and open access scientific journals are similar. *BMC Bioinformatics*, 10(1):183.