Distinguishing Clinical Sentiment: The Importance of Domain Adaptation in Psychiatric Patient Health Records

Eben Holderness[1,2], **Philip Cawkwell**[1], **Kirsten Bolton**[1],
James Pustejovsky[2] and **Mei-Hua Hall**[1]
[1]Psychosis Neurobiology Laboratory, McLean Hospital, Harvard Medical School
[2]Department of Computer Science, Brandeis University
{eholderness, mhall}@mclean.harvard.edu
{pcawkwell, kbolton}@partners.org
jamesp@cs.brandeis.edu

Abstract

Recently natural language processing (NLP) tools have been developed to identify and extract salient risk indicators in electronic health records (EHRs). Sentiment analysis, although widely used in non-medical areas for improving decision making, has been studied minimally in the clinical setting. In this study, we undertook, to our knowledge, the first domain adaptation of sentiment analysis to psychiatric EHRs by defining psychiatric clinical sentiment, performing an annotation project, and evaluating multiple sentence-level sentiment machine learning (ML) models. Results indicate that off-the-shelf sentiment analysis tools fail in identifying clinically positive or negative polarity, and that the definition of clinical sentiment that we provide is learnable with relatively small amounts of training data. This project is an initial step towards further refining sentiment analysis methods for clinical use. Our long-term objective is to incorporate the results of this project as part of a machine learning model that predicts inpatient readmission risk. We hope that this work will initiate a discussion concerning domain adaptation of sentiment analysis to the clinical setting.

1 Introduction

Psychotic disorders typically emerge in late adolescence or early adulthood (Kessler et al., 2007; Thomsen, 1996) and affect approximately 2.5-4% of the population (Perälä et al., 2007; Bogren et al., 2009), making them one of the leading causes of disability worldwide (Vos et al., 2015). A substantial proportion of psychiatric inpatients are readmitted after discharge (Wiersma et al., 1998). Readmissions are disruptive for both patients and families, and are a key driver of rising healthcare costs (Mangalore and Knapp, 2007; Wu et al., 2005). Reducing readmission risk is therefore a major unmet need of psychiatric care. Developing clinically implementable ML tools to enable accurate assessment of readmission risk factors offers opportunities to inform the selection of treatment interventions and to subsequently implement appropriate preventive measures.

Sentiment analysis (also known as opinion mining) has been used for capturing the subjective "feeling" (e.g. positive, negative, or neutral valence) of reviews and has recently been expanded to include other domains such as reactions to stock market prediction or political trends (Mäntylä et al., 2018). With the rise of social media and other user-generated web content, sentiment analysis has been adopted by many industries as a way of monitoring opinions towards their products, reputations, and for identifying opportunities for improvement. Traditionally, sentiment analysis has been approached with a lexicon-based majority vote approach, where a dictionary of terms and their associated sentiments (e.g. SentiWordnet, Pattern, SO-CAL, VADER) are queried to determine the sentiment of a given text (Taboada et al., 2011). However, this approach fails to account for many associated linguistic challenges such as negation handling, scope, sarcasm, qualified statements, and out-of-vocabulary terms. As such, research groups have moved towards approaching the problem from a corpus-based machine learning perspective. This approach has the added benefit of model flexibility depending on the training data and can capture more syntactic nuance. Most state-of-the-art performances on sentiment analysis benchmarks are currently achieved with deep learning sequence models that are trained on syntactically parsed corpora such as the Stanford Sentiment Treebank (Socher et al., 2013).

In clinical and medical domains, however, sentiment analysis has not yet been well studied. Yet retrieving subjective clinical attitudes (sentiment) from EHR narratives has the potential to facili-

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 117–123
Minneapolis, Minnesota, June 7, 2019. ©2019 Association for Computational Linguistics

tate identification of a patient's symptomatological worsening or increased readmission risk.

The concept of medical sentiment is complex and vocabularies differ from general-domain sentiment. In the field of psychiatry, this is especially true. Therefore, there is a need for domain adaptation of sentiment analysis that includes a richer array of attributes than can typically be found in off-the-shelf tools. In this work, we established an annotation scheme to characterize sentiment-related features in EHRs, and used this to carry out, to our knowledge, the first psychiatry-specific sentiment annotation project on EHRs. The resulting datasets are used to train and evaluate a classifier to predict clinical sentiment at the sentence level. This classifier, which in future works will be integrated in a pipeline for predicting readmission risk, is clinically useful for targeting treatments and aiding in decision making.

2 Related Works

Although there has been some work on clinical adaptation of sentiment analysis using healthcare-related data extracted from web forums, biomedical texts, or social media postings (See for example (Smith and Lee, 2012; Niu et al., 2005; Salas-Zárate et al., 2017; Nguyen et al., 2014)), there has been minimal work on sentiment analysis when applied to actual EHR data.

McCoy et al. (2015) used a corpus of psychosis patient discharge summaries and the 3,000 word Pattern lexical opinion mining dictionary (Smedt and Daelemans, 2012) to classify the associated sentiment of documents using a majority vote classifier. Results of their Cox regression models showed that greater positive sentiment was associated with a reduction in inpatient readmission risk. Waudby-Smith et al. (2018) applied the same Pattern sentiment lexicon to a corpus of ICU nursing notes to predict 30-day mortality risk. They found that stronger negative sentiment polarity was associated with an increased 30-day mortality risk. One of the limitations in both studies is that Pattern is a general-domain sentiment lexicon that contains few informative medical or psychiatry-specific terminology. Also, the authors did not manually annotate the datasets they worked with. As a result, they were not able to confirm that the predicted sentiment aligned with the sentiment from a clinical perspective.

(Deng et al., 2014) and (Denecke and Deng, 2015) systematically compared word usage and sentiment distribution between clinical narratives (nurse letters, discharge summary, and radiology reports) and medical social media (MedBlog, drug reviews). They concluded that off-the-shelf sentiment tools were not ideal for analyzing sentiment in medical documents and that EHRs were significantly more difficult in predicting sentiment, in particular neutral sentiment (Neutral F1=0.216 and 0.080 for nurse letters and radiology reports, respectively). They developed annotation guidelines and undertook a span-level annotation task on 300 ICU nurse letters to identify words related to clinical sentiment (Deng et al., 2016). Results of applying ML algorithms to these data are not available yet.

3 Methods

In this work, we define psychiatric clinical sentiment as a clinician's attitudes (positive, negative, or neutral) towards a patient's prognosis with regards to seven readmission risk factor domains (appearance, mood, interpersonal relations, substance use, thought content, thought process, and occupation) that were identified in prior work (Holderness et al., 2018). The scope of our current definition is intentionally narrow such that the sentiment of a given sentence is considered in isolation without any prior knowledge.

Three clinicians participated in an annotation project that focused on identifying the clinical sentiment associated with psychiatric EHR texts at the sentence level. In total, two corpora of clinical narratives from institutional EHRs, one containing 3,500 sentences (training dataset) and the other 1,650 (test dataset) were annotated using the definition established in the annotation scheme.

The training dataset consisted exclusively of sentence-length sequences that involved only one risk factor domain in each example. The examples in the dataset were identified from a large corpus of unannotated psychosis patient EHR data sourced from the psychiatric units of several Boston-area hospitals in the Partners HealthCare network, including Massachusetts General Hospital and Brigham & Women's Hospital. We used our risk factor domain topic extraction model to automatically identify relevant sentences, which were then manually validated by one of the clinicians involved in this project to ensure they did not involve multiple domains in the same exam-

Domain	Positive Example	Neutral Example	Negative Example
Appearance	Presents on time, dressed and groomed nicely, good hygiene.	Casually dressed and wearing knit vest and belt.	Notes that he wears the same clothes 2-3 days at a time, he doesn't care for his appearance–which is atypical for him.
Mood	Her depression and anxiety have improved immensely.	Mood is largely euthymic although he stated he gets depressed occasionally.	Tearful, presented very depressed with sad affect.
Interpersonal	Continues to be happy in her relationship with her boyfriend and school friends are stable as well.	She voiced no complaints about her primary relationship or other social relationships.	Poor social supports, abusive relationship.
Substance Use	Denies substance use or alcohol other than an occasional glass of wine.	Remote history of cocaine (smoked), marijuana and mescaline use many years ago.	He reports daily k2 use in addition to using crack cocaine about once a week.
Occupation	Pt reports having taken further steps toward employment – applied for two jobs and has interview lined up for Saturday.	Discusses new job as part time substitute teacher.	Recently has a new job that she hates and took a paycut.
Thought Content	She never had auditory hallucinations or delusions of thought broadcasting and thought insertion.	No overt hallucinations or delusions but expansive thinking.	Delusions and hallucinations continue.
Thought Process	Stable, slow speech with fewer word finding difficulties today, linear thought process, cooperative,attentive.	Slightly pressured speech but not as bad as some past visits.	Speech spontaneous and decreased in volume, rate, and rhythm; hard to understand at times because she barely opens her mouth when she talks.

Table 1: Example EHR sentences reflecting sentiment polarity for each risk factor domain.

ple. See Table 1 for example sentences for each domain.

The test dataset is an extension of the corpus used previously to evaluate our risk factor domain topic extraction model and is non-overlapping with the training data, consisting of discharge summaries, admission notes, individual encounter notes, and other clinical notes from 220 patients in the OnTrack^{TM} program at McLean Hospital. OnTrack^{TM} is an outpatient program, focusing on treating adults ages 18 to 30 who are experiencing their first episodes of psychosis. Because we are interested in identifying the clinical sentiment associated with each risk factor domain individually, the test dataset consists of examples that were intentionally selected to be challenging for our model: they are variable in length, wide-ranging in vocabulary, and can involve multiple risk factor domains (e.g. "Work functioning is impaired, but pt has good relationship w/ his girlfriend and is not engaging in substance use.").

These corpora are available to other researchers upon request. Table 2 details the distribution of the training and test data. The imbalance of training examples across the three sentiment classes reflects the natural distribution of sentiment reflected in EHRs, as certain risk factor domains (e.g. substance use) will rarely be reflected in a neutral or positive sense.

	Positive	Negative	Neutral
Appearance	290	69	141
Mood	100	322	77
Interpersonal	205	165	130
Substance Use	181	261	58
Occupation	250	143	150
Thought Process	150	266	84
Thought Content	183	253	64

Table 2: Distribution of training and test examples.

We evaluated three classification models. Our baseline model is a majority vote approach using the Pattern sentiment lexicon employed by McCoy (2015) and Waudby-Smith (2018). The second and third models use fully supervised and semi-supervised multilayer perceptron (MLP) architectures, respectively. Since positive and negative clinical sentiment can differ across each domain, we train a suite of seven models, one for each risk factor domain. The training and test data were vectorized at the sentence level using the pretrained Universal Sentence Encoder (USE) embedding module (Cer et al., 2018) that is available through TensorFlow Hub and is designed specifically for transfer learning tasks. Although USE is trained on a large volume of web-based, general-domain data, we have found in prior work that the embeddings lead to higher accuracy on down-

Model	Domain	Pos P	Pos R	Pos F1	Neg P	Neg R	Neg F1	Neu P	Neu R	Neu F1
Baseline (Pattern)	All	0.612	0.231	0.319	0.552	0.245	0.337	0.234	**0.736**	**0.348**
	Interpersonal	0.8	0.222	0.348	0.429	0.103	0.167	0.413	0.929	0.571
	Mood	0.511	0.233	0.32	0.558	0.352	0.432	0.266	0.672	0.381
	Occupation	0.75	0.129	0.22	0.328	0.188	0.265	0.329	0.917	0.484
	Substance Use	0.429	0.067	0.115	0.593	0.241	0.342	0.222	0.74	0.341
	Appearance	0.781	0.424	0.549	0.556	0.309	0.397	0.174	0.552	0.265
	Thought Content	0.556	0.19	0.283	0.723	0.29	0.414	0.055	0.6	0.101
	Thought Process	0.459	0.354	0.4	0.677	0.231	0.344	0.181	0.739	0.291
Fully Supervised MLP	All	**0.62**	0.416	**0.478**	**0.67**	0.652	**0.658**	**0.289**	0.437	0.329
	Interpersonal	0.632	0.667	0.649	0.731	0.656	0.691	0.567	0.607	0.558
	Mood	0.717	0.32	0.443	0.597	0.73	0.657	0.286	0.418	0.339
	Occupation	0.645	0.571	0.606	0.558	0.604	0.58	0.346	0.375	0.36
	Substance Use	0.423	0.244	0.31	0.674	0.714	0.693	0.344	0.42	0.378
	Appearance	0.705	0.525	0.602	0.69	0.605	0.645	0.241	0.448	0.313
	Thought Content	0.59	0.127	0.209	0.667	0.654	0.66	0.078	0.4	0.13
	Thought Process	0.629	0.458	0.53	0.775	0.604	0.679	0.161	0.391	0.228
Semi-Supervised MLP (Self-Training)	All	0.588	0.4	0.46	0.611	**0.733**	**0.658**	0.285	0.291	0.259
	Interpersonal	0.632	0.667	0.649	0.625	0.69	0.656	0.583	0.5	0.539
	Mood	0.645	0.301	0.411	0.502	0.885	0.641	0.233	0.105	0.144
	Occupation	0.671	0.671	0.671	0.539	0.583	0.56	0.364	0.333	0.348
	Substance Use	0.394	0.289	0.333	0.617	0.835	0.709	0.333	0.1	0.154
	Appearance	0.722	0.441	0.547	0.653	0.605	0.628	0.224	0.448	0.299
	Thought Content	0.5	0.139	0.218	0.689	0.753	0.72	0.088	0.333	0.139
	Thought Process	0.583	0.292	0.389	0.651	0.78	0.685	0.172	0.217	0.192

Table 3: Results of the clinical sentiment extraction task.

stream classification tasks than embedding models (e.g. ELMo, Doc2Vec, FastText) trained on smaller volumes of EHR data (Holderness et al., 2019).

Hyperparameters were tuned using grid search with 5-fold cross-validation on the training dataset and are specified in Table 4. Due to the relatively small amount of labeled training data, our proposed model architecture is designed to prevent overfitting by using a restricted view of the training data via a high rate of dropout in the hidden layers. Additionally, we use two hidden layers to extract a more abstracted form of the input. Additionally, because neutral sentiment is much broader in scope and has fewer training examples, resulting in covariate shift, we compute a threshold for classifying positive and negative sentiment using the formula min=avg(sim)+*(sim), where is standard deviation and is a constant, which we set to 0.2. If a given test sentence does not have positive or negative outputs that exceed this threshold, the sentence is classified as neutral even if neutral is not the maximal output.

We experimented with two semi-supervised learning configurations, Self-Training and K-Nearest Neighbors (KNN). The self-training approach involved first training our model on the labeled training data and then using this model to identify unlabeled examples from a large preprocessed corpus of unlabeled EHR data (2,100,000 sentences, 85,000,000 tokens). For the KNN approach, we projected all of the labeled and unlabeled examples into vector space and treated the labeled examples as centroids. For each centroid, we then used Euclidean distance to compute the five nearest unlabeled examples. Both models were trained using a 20:80 combination of the original labeled data and the additional unlabeled data.

Parameter	Value
Batch Size	28
Iterations	100
Hidden Units Per Layer	300
Dropout	0.75
Kernel Initializer	Uniform
Optimizer	Adam
Input/Hidden Layer Activations	ReLU
Output Layer Activations	Sigmoid

Table 4: Hyperparameters for sentiment model.

4 Results and Discussion

Inter-annotator agreement (IAA) was substantial on the first corpus (Scott's Pi=0.691, Cohen's Kappa=0.693) and higher on the second (Pi=0.768, Kappa=0.768) (Fleiss, 1971; Davies and Fleiss, 1982). This is expected as the first corpus contains many sentences involving multiple readmission risk factor domains and annotators were instructed to provide clinical sentiment labels for each, whereas the second corpus consists entirely of single domain sentences. In both cases, IAA surpasses that reported by Denecke and Deng (2016), primarily because of the clinical expertise of the annotators involved in this project.

Results of the three classifiers are shown in Table 3, with the highest score on each performance metric in bold. The 'All' row for each model configuration was computed by averaging the scores of the sentiment models for each risk factor domain. Applying the Pattern sentiment lexicon to our test corpus showed a strong trend towards underclassification of positive and negative examples, which led to poor recall scores while maintaining moderate precision. Neutral examples, however, were correctly classified significantly more often. This confirms that many of the most informative words in terms of clinical sentiment (e.g. 'hallucination', 'depressed', 'employed', etc.) do not hold significance in general-domain sentiment and are therefore not part of the Pattern lexicon.

Despite the relatively small size of the training corpus, the EHR data used for training captured much of the domain-specific vocabulary related to clinical sentiment and our suite of models achieved F-measures on classifying positive and negative sentiment that exceed those reported in prior literature (Deng et al., 2014). Although direct comparison between our EHR dataset and the EHR datasets used by other researchers is limited due to HIPAA restrictions, our training EHR data is sourced from the same EHR database as McCoy (2015). Therefore, a better performance of our models indirectly supports that our model can better capture the underlying clinical sentiment embedded in EHRs.

Because clinical documents are written for a specific purpose such as assessing the outcome of treatment, they contain less neutral content and as a result sentiment distributions are intrinsically biased to either positive or negative polarity. Thus, identifying training examples with neutral sentiment was challenging and consequently both the fully and semi-supervised models were poor at identifying neutral sentiment across all seven domains. In addition, unless the patient is markedly improved, clinicians tend to document continuing unresolved symptoms. leading to a greater amount of negative content. We hypothesize that this may be one reason for the lower overall F1 performance on positive versus negative sentiment.

We observed that per-domain performance of our models aligned with the natural distribution of positive vs. negative clinical sentiment in EHRs. Substance use, for example, had low positive F1 scores as the majority of references to substance use in EHRs involve negative sentiment unless the patient is noted to be abstaining from substance use. We also observed that sentiment distribution towards negative polarity is more evident in mood and thought content, which include, for example, delusions, depression, anxiety, and hallucinations.

When applying semi-supervised learning methods, we found self-training to marginally improve performance on negative clinical sentiment but the overall F1 score was not better than the fully supervised model due to lower precision. We observed minimal change in performance when using a k-nearest neighbors approach.

5 Conclusion and Future Work

We focused in this study on the clinical sentiment associated with readmission for seven risk factor domains identified in prior work by undertaking an annotation project and using the resultant gold standard to train semi-supervised ML algorithms to automatically infer this sentiment. Our results indicate that domain adaptation of sentiment analysis is necessary for aligning with clinician opinions.

We intend to improve our clinical sentiment classifier in future work by increasing the size of the annotated training corpus (in particular neutral examples) and by changing the model input to a sequence model as opposed to a full sentence vector representation. We also intend to modify our definition of clinical sentiment to include temporal linking of elements that involve clinical sentiment in an EHR to establish gradients of changes in patient status over time. Finally, we will incorporate our sentiment analysis model in a classifier that predicts inpatient readmission risk.

6 Acknowledgments

This work was supported by a grant from the National Institute of Mental Health (grant no. 5R01MH109687 to Mei-Hua Hall). We would also like to thank the Clinical NLP 2019 Workshop reviewers for their constructive and helpful comments.

References

Mats Bogren, Cecilia Mattisson, Per-Erik Isberg, and Per Nettelbladt. 2009. How common are psychotic and bipolar disorders? a 50-year follow-up of the lundby population. *Nordic journal of psychiatry*, 63(4):336–346.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Mark Davies and Joseph L Fleiss. 1982. Measuring agreement for multinomial data. *Biometrics*, pages 1047–1051.

Kerstin Denecke and Yihan Deng. 2015. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1):17–27.

Yihan Deng, Thierry Declerck, Piroska Lendvai, and Kerstin Denecke. 2016. The generation of a corpus for clinical sentiment analysis. In *European Semantic Web Conference*, pages 311–324. Springer.

Yihan Deng, Matthaeus Stoehr, and Kerstin Denecke. 2014. Retrieving attitudes: Sentiment analysis from clinical narratives. In *MedIR@ SIGIR*, pages 12–15.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Eben Holderness, Marie Meteer, James Pustejovsky, and Mei-Hua Hall. 2019. Evaluating the role of pretraining in natural language processing of clinical narratives. Poster presented at McLean Research Day, Belmont, MA. Results are available upon request.

Eben Holderness, Nicholas Miller, Kirsten Bolton, Philip Cawkwell, Marie Meteer, James Pustejovsky, and Mei Hua-Hall. 2018. Analysis of risk factor domains in psychosis patient health records. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 129–138.

Ronald C Kessler, G Paul Amminger, Sergio Aguilar-Gaxiola, Jordi Alonso, Sing Lee, and T Bedirhan Ustun. 2007. Age of onset of mental disorders: a review of recent literature. *Current opinion in psychiatry*, 20(4):359.

Roshni Mangalore and Martin Knapp. 2007. Cost of schizophrenia in england. *The journal of mental health policy and economics*, 10(1):23–41.

Mika V Mäntylä, Daniel Graziotin, and Miikka Kuutila. 2018. The evolution of sentiment analysisa review of research topics, venues, and top cited papers. *Computer Science Review*, 27:16–32.

Thomas H McCoy, Victor M Castro, Andrew Cagan, Ashlee M Roberson, Isaac S Kohane, and Roy H Perlis. 2015. Sentiment measured in hospital discharge notes is associated with readmission and mortality risk: an electronic health record study. *PloS one*, 10(8):e0136341.

Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, 5(3):217–226.

Yun Niu, Xiaodan Zhu, Jianhua Li, and Graeme Hirst. 2005. Analysis of polarity information in medical text. In *AMIA annual symposium proceedings*, volume 2005, page 570. American Medical Informatics Association.

Jonna Perälä, Jaana Suvisaari, Samuli I Saarni, Kimmo Kuoppasalmi, Erkki Isometsä, Sami Pirkola, Timo Partonen, Annamari Tuulio-Henriksson, Jukka Hintikka, Tuula Kieseppä, et al. 2007. Lifetime prevalence of psychotic and bipolar i disorders in a general population. *Archives of general psychiatry*, 64(1):19–28.

María del Pilar Salas-Zárate, Jose Medina-Moreira, Katty Lagos-Ortiz, Harry Luna-Aveiga, Miguel Angel Rodriguez-Garcia, and Rafael Valencia-Garcia. 2017. Sentiment analysis on tweets about diabetes: an aspect-level approach. *Computational and mathematical methods in medicine*, 2017.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research*, 13(Jun):2063–2067.

Phillip Smith and Mark Lee. 2012. Cross-discourse development of supervised sentiment analysis in the clinical domain. In *Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis*, pages 79–83. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

PH Thomsen. 1996. Schizophrenia with childhood and adolescent onseta nationwide register-based study. *Acta Psychiatrica Scandinavica*, 94(3):187–193.

Theo Vos, Ryan M Barber, Brad Bell, Amelia Bertozzi-Villa, Stan Biryukov, Ian Bolliger, Fiona Charlson, Adrian Davis, Louisa Degenhardt, Daniel Dicker, et al. 2015. Global, regional, and national incidence, prevalence, and years lived with disability for

301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the global burden of disease study 2013. *The Lancet*, 386(9995):743–800.

Ian ER Waudby-Smith, Nam Tran, Joel A Dubin, and Joon Lee. 2018. Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients. *PloS one*, 13(6):e0198687.

Durk Wiersma, Fokko J Nienhuis, Cees J Slooff, and Robert Giel. 1998. Natural course of schizophrenic disorders: a 15-year followup of a dutch incidence cohort. *Schizophrenia bulletin*, 24(1):75–85.

Eric Q Wu, Howard G Birnbaum, Lizheng Shi, Daniel E Ball, Ronald C Kessler, Matthew Moulis, and Jyoti Aggarwal. 2005. The economic burden of schizophrenia in the united states in 2002. *Journal of Clinical Psychiatry*, 66(9):1122–1129.