Published in final edited form as:
J Biomed Inform. 2017 November ; 75 Suppl: S1–S3. doi:10.1016/j.jbi.2017.10.005.

A Natural Language Processing Challenge for Clinical Records: Research Domains Criteria (RDoC) for Psychiatry

Ozlem Uzuner,
Department of Information Sciences and Technology, George Mason University, Fairfax, VA, USA

Amber Stubbs, and
School of Library and Information Science, Simmons College, Boston, MA, USA

Michele Filannino
Department of Computer Science, State University of New York at Albany, Albany, NY, USA

Society increasingly treats mental health as an integral part of overall well-being. As our understanding of mental health issues improves, so does our ability to diagnose and evaluate the effects of such issues on our overall physical health. Psychiatric records are a gold mine of information for understanding and managing mental health problems. However, access to these records is tightly controlled because of privacy concerns. Although society seeks to protect all aspects of our general health information, mental health records are treated with extra sensitivity because of the types of personal information they contain and because of potential stigma associated with mental illness. As a result, mental health records have never before been shared with the research community.

The Neuropsychiatric Genome-Scale and Research Domains Criteria (RDoC) Individualized Domains (N-GRID) project is one of the Centers of Excellence in Genomic Science (CEGS) based in Harvard Medical School, funded by National Institute of Mental Health and co-funded by National Human Genome Research Institute. N-GRID's leaders had the foresight to open up a set of psychiatric evaluation records to the research community, supporting a study of their contents through natural language processing (NLP). This allowed the development of the very first shared task on mental health notes, building on the former i2b2 shared tasks organized since 2006 [1–11].

The guest editors of this issue de-identified these mental health records, referred to as RDoC data, and shared them with the research community, along with gold standard annotations, for a shared task that we named the **RDoC for Psychiatry** challenge. Following the model of the past i2b2 challenges, we started the shared task by releasing annotated gold standard training data to the research community for system development and training. Following a development period, we released a held-out test set to the shared task participants without the matching gold standard annotations. We collected system outputs on the test set within

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

three days of test data release. We evaluated systems based on their performance on the test set.

Following the shared task, we organized a workshop that met as an Allied meeting of the Fall Symposium of the American Medical Informatics Association in November 2016. During this meeting, the participants shared their novel solutions to the RDoC for Psychiatry challenge and the insights gained from the challenge with the broader research community through plenary presentations and posters.

The RDoC for Psychiatry challenge was organized around three tracks:

> Track 1: De-identification: The goal of de-identification is to remove protected health information (PHI) from the records. We ran two versions of this track. One of the de-identification tracks evaluated existing systems on the RDoC data without any training or modification, as a way of measuring how well the existing systems generalize to brand new data. The second one evaluated systems after they were trained on the RDoC data. An overview of the de-identification track can be found in Stubbs, et al. [12].
>
> Track 2: Symptom severity classification: The goal of this track was to determine symptom severity in a patient, using an ordinal scale of 0–3, in one RDoC domain, based on information included in the initial psychiatric evaluation of the patient. This task represents the first ever attempt at classifying mental health records on an ordinal scale using natural language processing. An overview of the symptom severity classification track can be found in Filannino, et al. [13].
>
> Track 3: Novel Data Use: The data released for the 2016 challenge were the first set of mental health records released to the research community. These data could be used for mental health-related research questions that go beyond what is posed by the challenge organizers. The novel data use track was for participants who wanted to utilize the RDoC data to address new research questions, e.g., [14–16].

This supplement to the *Journal of Biomedical Informatics* gives an overview and presents a select set of peer-reviewed papers that participated in the three tracks of the RDoC for Psychiatry challenge.

Track 1: De-identification

One of the questions left open after the 2014 i2b2/UTHealth de-identification challenge [11] was “Does success on this challenge problem extrapolate to similar performance on other, untested data sets?” The use of psychiatric records for the CEGS NGRID shared task presented an opportunity to answer this question, as previous de-identification challenges used other forms of medical records, such as discharge summaries and clinical narratives.

The differences between the psychiatric records and previous shared task data included content and vocabulary, as psychiatric records deal primarily with mental states, not physical ones; organization, as the psychiatric records present different sections in terms of headings and content; and formatting, as the majority of the psychiatric records contain standard question and answer sections.

In addition to such expected differences, the patient records available for analysis contained some artifacts of the system from which they were drawn. For example, as a consequence of how the data were stored in the Partners Health Care database, many line breaks were missing after the answer to the questions in the question/answer sections, resulting in the last word of the answer being combined with the first word of the following question. These kinds of complications represent the realities of working with real-life data. Therefore, the organizers made no attempts to correct such problems and shared the data with the research community without clarifying edits. This required the NLP methods to be robust in working with real data and to perform well on their target applications regardless.

The shared task split Track 1 into two de-identification subtasks: the first was a “sight unseen” subtask, for which the challenge participants ran their systems over the new psychiatric records without making any modifications to their existing systems. The results of this subtask provided information on whether new data could be accurately de-identified based on existing models. The top-performing system for this subtask achieved an F1 score of 0.7985 [12], suggesting that out-of-the-box solutions fail to provide reliable results on the new data but provide a good start at building models that can be tuned to the new data.

The second subtask for Track 1 provided the participants with two months of development time to tune their models to the new data, and to experiment with different systems for approaching this subtask. Nearly all the teams in this subtask used hybrid systems of a variety of algorithms, training each component for a specific subset of private health information (PHI) and then combining the outputs of all components so as to cover all PHI. Some teams used combinations of Conditional Random Fields (CRFs) each trained for different subsets of PHI, for example using one CRF to identify patterns in dates, phone numbers, and other PHI with standardized alphanumeric patterns, and another CRF for text-based PHI such as names and locations [17, 18]. Others used combinations of CRFs and bi-directional long-short term memory systems (BI-LSTM) [19, 20]. Another team used a “multi-pass sieve” system, with different sieves focusing on pattern-matching, dictionary matching, or CRFs [21]. Overall, these hybrid systems proved to be largely effective, with the top-performing system achieving an F1 score of 0.9422 [19].

Track 2: Symptom Severity Classification

The second track of the 2016 CEGS N-GRID shared task introduced an application infrequently visited by the scientific community but of high practical impact: psychiatric symptom severity classification [13]. Psychiatric evaluation records are rich in medical signs and symptoms that are essential to diagnose the severity of psychiatric disorders. In this track of RDoC for psychiatry shared task, we focused on a specific subset of disorders called *positive valence*. Typical positive valence disorders, as described by RDoC, are substance abuse, dependence, mania, gambling, and obsessive-compulsive disorders. To study positive valence as an NLP application, we represented symptom severity on an ordinal discrete scale from 0–3: absent (0), mild (1), moderate (2), and severe (3). We defined an error measure, Inverse Normalized Mean Absolute Error Macro-averaged (INMAE^M), that takes into account ordinality of the classes. We released both annotated and unannotated records to the shared task participants.

The teams that participated in this track tackled the symptom severity classification problem using supervised machine learning approaches. They experimented with different classifiers and approaches. The classifiers most prominently used were Support Vector Regressors [22], Decision Trees [23], Random Forests [24] and Gradient Tree Boosting [25]. The approaches included an ensemble of Convolutional Neural Networks (CNN) with word embeddings [26] and a mixture of Regularized Multinomial Logistic Regression classifiers and Neural Networks [27].

Due to the availability of unannotated data, in addition to their supervised solutions, four teams experimented with semi-supervised approaches, e.g., [22, 24]. Only three teams involved medical experts.

The top 10 submissions scored higher than 0.80 INMAE^M, with the top performing system scoring 0.863 [13]. The error analysis performed on the top 10 submissions revealed three interesting facts. First, systems tended to misclassify ambivalent records: the ones in which patients showed signs of both positive and negative valence (e.g., depression). Second, systems misclassified records with very few (although crucial) positive valence signs. Finally, the top performing system achieved a level of accuracy close to the one recorded by the least experienced annotator [13]. Overall, the results of this track frame the classification of symptoms severity task as effectively accomplishable within the campus of data-driven approaches, although a space for improvement is still left.

Track 3: Novel Data Use

Often, data can serve purposes beyond those for which they are designed. This was true of the RDoC data as well. Given the scarcity of mental health records available to the research community, and as a way of encouraging researchers to pursue their own research questions on these data, we included in this issue peer-reviewed articles on the uses of RDoC data outside of de-identification and symptom severity classification. In the novel data use track, Zhang et al. [14] took an unsupervised approach to finding psychiatric disorder symptoms in the RDoC data. They used external sources as “seed” terms for symptoms and built a system for recognizing symptoms based on their syntactic characteristics and semantic similarity to seed terms. Tran and Kavuluru [15] study the “History of Present Illness” section of the RDoC data in order to predict a set of mental health conditions from which the patient may suffer. They used deep learning methods, achieving micro-averaged F-measures in 60’s for eleven common mental health conditions. Dai et al. [16] studied the relationship between violence and “clinical and social parameters” finding that important determinants of violent behavior included family “history of violent behavior”, suicidal tendencies, “presence of financial stress”, and most importantly, use of stimulants.

Conclusions

The CEGS NGRID challenge provided a unique learning opportunity for the research community in many ways. It allowed us to evaluate de-identification methods out-of-the box and showed that while systems perform better when they are trained on the new target data, they provide a good start at solving the problem on these data even without training. Depending on the use in mind, ~80% F-measure is a respectable performance level for many internal uses that can utilize partially de-identified data. Even in circumstances where better de-identification is absolutely necessary, off-the-shelf application of systems can provide a solid base, which can then be tuned to the new data.

On predicting symptom severity on an ordinal scale as well, systems show results above .80. In many real-world applications, the distinction between classes is a matter of degree. The fact that systems can capture the nuanced differences of the ordinal scale is promising for future work.

As we write this editorial, the continued availability of the RDoC data for future research remains a topic of discussion with institutional review boards. Our hope and expectation is that the data will remain available for future research, fostering creativity and helping to advance the state of the art in both the designed tracks of the RDoC for Psychiatry challenge and in novel uses.

Acknowledgments

We would like to thank the participants of the 2016 CEGS N-GRID shared tasks. Special thanks go to program committee members: Tianxi Cai, Harvard School of Public Health; Susanne Churchill, Harvard Medical School; Isaac Kohane, Harvard Medical School; Thomas H. McCoy, MGH, Harvard; Roy H. Perlis, MGH, Harvard; Peter Szolovits, MIT; Uma Vaidyanathan, NIMH; and Philip Wang, American Psychiatric Association. Funding for this project was provided by the grants NIH P50 MH106933 (PI: Isaac Kohane) and NIH 4R13LM011411 (PI: Ozlem Uzuner).

References

1. Uzuner, Özlem, Luo, Yuan, Szolovits, Peter. Evaluating the State-of-the-Art in Automatic De-identification. Journal of the American Medical Informatics Association. Sep; 2007 14(5):550–563. [PubMed: 17600094]
2. Uzuner, Özlem, Goldstein, Ira, Luo, Yuan, Kohane, Isaac. Identifying Patient Smoking Status from Medical Discharge Records. Journal of the American Medical Informatics Association. Jan; 2008 15(1):14–24. [PubMed: 17947624]
3. Uzuner, Özlem. Recognizing Obesity and Co-morbidities in Sparse Data. Journal of the American Medical Informatics Association. Jul; 2009 16(4):561–570. [PubMed: 19390096]
4. Uzuner, Özlem, Solti, Imre, Cadag, Ethon. Extracting Medication Information from Clinical Text. Journal of the American Medical Informatics Association. 2010; 17:514–518. DOI: 10.1136/jamia.2010.003947 [PubMed: 20819854]
5. Uzuner, Özlem, South, Brett, Shen, Shuying, DuVall, Scott. 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. Journal of the American Medical Informatics Association. 2011; 18:552–556. Published Online First: 16 June 2011. DOI: 10.1136/amiajnl-2011-000203 [PubMed: 21685143]
6. Uzuner, Özlem, Bodnari, Andreea, Shen, Shuying, Forbush, Tyler, Pestian, John, South, Brett. Evaluating the State of the Art in Coreference Resolution for Electronic Medical Records. Journal of the American Medical Informatics Association. 2012; 19:786–791. Published Online First: 24 February 2012. DOI: 10.1136/amiajnl-2011-000784 [PubMed: 22366294]

7. Sun, Weiyi, Rumshisky, Anna, Uzuner, Özlem. Evaluating Temporal Relations in Clinical Text: 2012 i2b2 Challenge Overview. Journal of the American Medical Informatics Association. 2013; 20:806–813. Published Online First: 5 April 2013. DOI: 10.1136/amiajnl-2013-001628 [PubMed: 23564629]

8. Stubbs, Amber, Kotfila, Chris, Uzuner, Özlem. Automated Systems for the De-identification of Longitudinal Clinical Narratives: Overview of 2014 i2b2/UTHealth Shared Task Track 1. Journal of Biomedical Informatics. 2015 Dec; 58(Suppl):S11–9. Epub 2015 Jul 28. DOI: 10.1016/j.jbi.2015.06.007 [PubMed: 26225918]

9. Stubbs, Amber, Kotfila, Chris, Xu, Hua, Uzuner, Özlem. Identifying Risk Factors for Heart Disease over Time: Overview of 2014 i2b2/UTHealth Shared Task Track 2. Journal of Biomedical Informatics. 2015 Dec; 58(Suppl):S67–77. Epub 2015 Jul 22. DOI: 10.1016/j.jbi.2015.07.001 [PubMed: 26210362]

10. Uzuner, Özlem, Stubbs, Amber, Sun, Weiyi. Chronology of Your Health Events: Approaches to Extracting Temporal Relations from Medical Narratives. Journal of Biomedical Informatics. 2013 Dec; 46(0):S1–S4. DOI: 10.1016/j.jbi.2013.11.005 [PubMed: 24286753]

11. Uzuner, Özlem, Stubbs, Amber. Practical Applications for Natural Language Processing in Clinical Research: the 2014 i2b2/UTHealth shared tasks. Journal of Biomedical Informatics. 2015 Dec; 58(Suppl):S1–S5. DOI: 10.1016/j.jbi.2015.10.007 [PubMed: 26515500]

12. Stubbs, Amber, Filannino, Michele, Uzuner, Ozlem. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID shared tasks Track 1. This issue.

13. Filannino, Michele, Stubbs, Amber, Uzuner, Ozlem. Symptom severity prediction from neuropsychiatric clinical records: Overview of 2016 CEGS N-GRID shared tasks Track 2. This issue.

14. Zhang, Yaoyun, Zhang, Olivia, Wu, Yonghui, Lee, Hee-Jin, Xu, Jun, Xu, Hua, Roberts, Kirk. Psychiatric symptom recognition without labeled data using distributional representations of phrases and on-line knowledge. This issue.

15. Tran, Tung, Kavuluru, Ramakanth. Predicting Mental Conditions Based on "History of Present Illness" in Psychiatric Notes with Deep Neural Networks. This issue.

16. Dai, Hong-Jie, Su, Emily Chia-Yu, Uddin, Mohy, Jonnagaddala, Jitendra, Wu, Chi-Shin, Syed-Abdul, Shabbir. Exploring Associations of Clinical and Social Parameters with Violent Behaviors among Psychiatric Patients. This issue.

17. Roberts, Kirk, Wu, Yonghui, Zhang, Yaoyun, Xu, Jun, Xu, Hua, Lee, Hee-Jin. A hybrid approach to automatic de-identification of psychiatric notes. This issue.

18. Nenadic, Goran, Dehghan, Azad, Kovacevic, Aleksandar, Karystianis, George, Keane, John A. Learning to identify protected health information by integrating knowledge- and data-driven algorithms: a case study on psychiatric evaluation notes. This issue.

19. Tang, Buzhou, Liu, Zengjian, Wang, Xiaolong, Chen, Qingcai. De-identification of Clinical Notes via Recurrent Neural Network and Conditional Random Field. This issue.

20. Guan, Yi, Jiang, Zhipeng, He, Bin, Jiang, Jingchi. De-identification of medical records using conditional random fields and long short-term memory networks. This issue.

21. Duc, Duy, Bui, An, Wyatt, Matthew, Cimino, James J. The UAB Informatics Institute and 2016 CEGS N-GRID De-Identification Shared Task Challenge. This issue.

22. Goodwin, Travis, Maldonado, Ramon, Harabagiu, Sanda M. Automatic Recognition of Symptom Severity from Psychiatric Evaluation Records. This issue.

23. Tsui, Fuchiang (Rich), Posada, Jose D., Barda, Amie J., Shi, Lingyun, Xue, Diyang, Ruiz, Victor, Kuan, Pei-Han, Ryan, Neal D. Predictive Modeling for Classification of Positive Valence System Symptom Severity from Initial Psychiatric Evaluation Records. This issue.

24. Scheurwegs, Elyne, Sushil, Madhumita, Tulkens, Stéphan, Daelemans, Walter, Luyckx, Kim. Counting trees in Random Forests: Predicting symptom severity in psychiatric intake reports. This issue.

25. Liu, Yang, Gu, Yu, Nguyen, John C., Li, Haodan, Zhang, Jiawei, Gao, Yuan, Yang, Huang. Symptom Severity Classification with Gradient Tree Boosting. This issue.

26. Kavuluru, Ramakanth, Rios, Anthony. Ordinal Convolutional Neural Networks for Predicting RDoC Positive Valence Psychiatric Symptom Severity Scores. This issue.

27. Clark, Cheryl, Wellner, Ben, Davis, Rachel, Aberdeen, John, Hirschman, Lynette. Automatic Classification of RDoC Positive Valence Severity with a Neural Network. This issue.