Multiple Sclerosis Diagnosis from Natural Language Processing of Patient Text Data

Anonymous ACL submission

Abstract

Multiple Sclerosis (MS) is an important and growing concern for public health. Therefore, there is a need for inexpensive yet effective methods for diagnosis of MS. One such diagnosis method involves NLP-based predictive classifiers, which take as input patient experiences in text format. However, in order to employ NLP for MS diagnosis, we require a large-scale, high-quality, labelled dataset containing experiences of MS patients and individuals who do not have MS. In this paper, we introduce the largest ever dataset for MS diagnosis from patient free text data, created by combining social media text data with free text data from the MS Register database, a database of clinically diagnosed MS patients in the UK. We examine differences in language between the MS-diagnosed group and the control group, and we explore text classification methods to identify individuals with MS.

1 Introduction

Multiple Sclerosis (MS) is a chronic, inflammatory, neurological condition that has presented itself as a significant challenge to healthcare. According to Walton et al. (2020), it is estimated that 2.8 million people worldwide suffer from MS in 2020. Furthermore, a cure for MS still does not exist and knowledge about the condition is currently incomplete (Bebo et al., 2022). Thus, there is a compelling need to improve our understanding of the condition.

Specifically, one of the challenges that primary healthcare providers face is the diagnosis of MS. There is no single test or laboratory marker for the diagnosis of MS, and diagnosis of MS is usually done by ruling out other similar conditions (Omerhoca et al., 2018). Since MS can potentially cause a vast variety of symptoms, including muscle spasms, vision problems and anxiety (Ghasemi et al., 2017), accurate diagnosis of the patient seems increasingly impossible without computerized diagnostic assistance.

A solution to improve the accuracy of multiple sclerosis diagnosis can be NLP-based predictive classifiers, which take the experiences of the patient – in text format – as input. This method can be inexpensive yet effective because a well-trained classifier should not only be able to detect physical symptoms of MS in the text, but also be able to capture the psychological differences – at least those that can be expressed through language – between people with and without MS. Therefore, the aim of this study was to investigate the accuracy of NLP-based predictive models to detect multiple sclerosis from patient free text data.

Our work has the following important improvements over existing literature. To create our dataset, we utilize a unique approach where we combine data from social media with the data from the MS Register database, a database of clinically diagnosed MS patients from the UK (Ford et al., 2012). Since the MS Register database has a relatively large number of MS patients, with this approach we were able to create the largest ever dataset for computerized MS diagnosis from free text. Moreover, existing research in free text MS diagnosis is also limited because most research relies on only one type of predictive classifier. Thus, another way we extend prior research is by training and testing multiple classification models on the dataset and comparing their performance to get the best classifier for our use-case.

Our contributions are as follows: *(i)* create the largest ever dataset for computerized MS diagnosis from free text by combining the data from the existing MS Register dataset with newly collected social media data *(ii)* perform data analysis on this newly created dataset *(iii)* build and train binary classifiers on the dataset to detect multiple sclerosis from patient free text.

2 Related work

Natural language processing has been extensively employed in medical diagnosis in the past decade. Swartz et al. (2017) was able to apply natural language processing to diagnose venous thromboembolism (VTE) from radiology reports with a high degree of accuracy. Kolanu et al. (2020) similarly uses NLP techniques to improve identification of patients with clinically significant fractures. Liang et al. (2019) utilized a deep learning NLP system to extract clinically important information from electronic health records for diagnosis of paediatric diseases. Coppersmith et al. (2014) relied on publicly available Twitter data to create NLP-based classifiers to detect mental health disorders, including post-traumatic stress disorder (PTSD), depression, bipolar disorder, and seasonal affective disorder (SAD). Yang et al. (2019) was able to develop a recurrent neural network with contextual word embeddings to get state-of-the-art accuracy in extracting symptoms from the electronic health record.

Although natural language processing has been pervasively applied in medical diagnosis, prior research in NLP-based diagnosis of MS is limited. While Chase et al. (2017) created a predictive model for diagnosing MS from patient clinical notes, they employed a relatively small dataset of only 165 MS patients, and they only use Naïve Bayes classification as a predictive model. Although Koss and Bohnet-Joschko (2022) employ NLP methods to investigate patterns in the unmet medical needs of MS patients, they do not explicitly build classification models for MS detection from text data. Schwab and Karlen (2021) utilise the multimodal, smartphone-collected data of MS patients to a train a model to diagnose MS. This data also includes questionnaires, but they only use simple sentiment analysis to calculate a mood score which is then used as an input to the model – there was no further NLP analysis of the questionnaire data. Furthermore, they only consider deep learning as a method for diagnosing MS patients.

We extend this existing body of work by generating the largest ever dataset for computerized MS diagnosis and conducting extensive natural language processing analysis on this dataset. As part of this analysis, we explore multiple methods of converting text in the dataset to word vectors and we train various machine learning models on these word vectors for the diagnosis of MS.

3 Data

To build binary classifiers for detecting multiple sclerosis from free text, we need a dataset containing experiences of both people with and people without MS, in text format. In this section, we describe how such a dataset was constructed by combining the data from social media platforms with data from the MS Register database (Ford et al., 2012). We then discuss the privacy and ethical issues surrounding the construction of our dataset.

3.1 Dataset construction

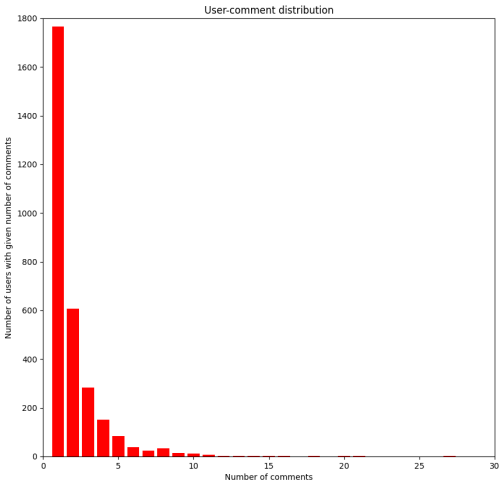

Figure 1: User-Comment Distribution

Diagnosed group. To get textual comments outlining the experiences of MS patients, we will utilize free text data from the MS Register database (Ford et al., 2012). We chose to use this database because it provides a large, rich resource of experiences of patients diagnosed with MS, in text format. However, one problem of using this dataset directly for our use-case is that the experiences of the patients are recorded post MS diagnosis, which means patients often refer to their condition (that is, multiple sclerosis), in their comments. However, since the classifier will be used pre-diagnosis, we cannot make this classifier rely on the term "multiple sclerosis", or any other term that only MS patients will know about. Therefore, we remove all clauses containing MS-related terms[1] from patient comments. If every clause in a patient comment includes some MS-related term, then we omit the

[1] https://www.doc.ic.ac.uk/~dj321/msnlp/msterms.txt

comment from our dataset. With this approach, we were able to get 6121 comments from 3037 diagnosed users in our dataset.

Control group. To get the experiences of users who do not have MS (that is, the control group), we chose social media platforms because, like the MS Register database, people share their experiences in text format on these platforms. Twitter and Reddit were chosen as the social media platforms because they are popular platforms where users often post in plain text about their general life experiences. Therefore, from these social media platforms, we construct two separate control groups. The explanation for the need for two control groups and the description of how these control groups were constructed are outlined below.

The first control group constitutes random users from Twitter and Reddit who have not mentioned the term "MS" or any related terms in their timeline. The dataset was constructed such that there is an equal probability for a potential control user to be selected from Twitter or from Reddit. Moreover, we ignored image-only posts, video-only posts, Twitter retweets and Reddit cross-posts. For Reddit, we added the additional restriction that the potential control group user has never posted on any MS-related subreddit[2]. After identifying the users, we select random posts in their timeline to be included in our dataset. The number of posts we select per user is controlled such that the final user-comment distribution is kept the same as the diagnosed users in the dataset, as illustrated in Figure 1. Finally, as a measure to prevent posts created by bots to be included in our dataset, no more than 5 duplicate posts were allowed by the same user.

This first control group is relevant because a classifier trained using this control group can potentially be used to automatically diagnose a social media user for MS. However, we observed a potential problem: since social media contains a large variety of text data, including news, satire and advertisements, a good classifier will likely learn that all this text data corresponds to a person not having MS. Thus, although a classifier resulting from control one can be used in social media for MS diagnosis, it may still be inaccurate in clinical settings. To solve this problem, we need to create another control group containing experiences of patients diagnosed with another illness.

Therefore, the second control group are comments of users who have given self-reported diagnoses of diabetes on social media (both type 1 and type 2), and who, like the first control group, have not mentioned "MS" or any related terms in their timeline. Diabetes was chosen as the illness for the second control group because firstly, it is an incurable, chronic condition like MS, and secondly, it is relatively prevalent, so we can find enough users on social media to exactly match the number of MS-diagnosed users from the MS Register dataset. The users were found by searching on Twitter and Reddit for high-precision diagnostic patterns[3]. Some examples of these patterns include *"i have been diagosed with diabetes"*, *"he diagnosed me with diabetes"* and *"i have a diagnosis of diabetes"*. After identifying these users, we only include posts where the user mentions their condition (that is, diabetes), because we would expect only those experiences to be presented to a primary healthcare provider. Then, like what we did for the diagnosed users, we remove all clauses with the word "diabetes" from the text, since the classifier will be used pre-diagnosis, and we cannot make the classifier rely on the term "diabetes" for classifying patients who do not have MS.

For both the control groups, the number of comments and the user-comment distribution was kept the same as the diagnosed users. However, the age and gender of the users could not be controlled for since APIs of social media platforms do not provide this data.

3.2 Ethics and privacy

The risks associated with our data collection and data access methods are minimal. All the posts collected from Twitter and Reddit were publicly available through the Twitter and Reddit API respectively. We replaced social media usernames with user identifier codes to anonymise their identity. Although the MS Register database contains identifiable information, including the gender and date of birth of the patients, this database was accessed through a secure virtual desktop environment, so the identifiable information remains private. For both the patients in the MS Register database and the social media users, we made no attempt identify the users with other external sources of information.

[2] https://www.doc.ic.ac.uk/~dj321/msnlp/mssubreddits.txt

[3] List of diagnosis patterns (https://www.doc.ic.ac.uk/~dj321/msnlp/diagnosispatterns.txt) was inspired by the work of Cohan et al. (2018) on NLP-based diagnosis of mental health disorders

Therefore, this research was conducted in compliance to the GDBR guidelines of data protection (GDPR, 2016).

4 Datset Analysis

To increase our understanding of the newly created dataset, we started by conducting some preliminary data analysis on our dataset.

To begin with, we created word clouds from the text given by the MS-diagnosed users (Figure 2a), the generic control one users (Figure 2b), and the diabetes control two users (Figure 2c). Before creating each of these word clouds, the text was made lowercase, stop words were removed from the text, and the text was tokenized.

Inspecting the word clouds, we see that it is consistent with our expectations of the various parts of the dataset. Words prevalent in the MS-diagnosed group that are absent in the other groups are 'fatigue', 'walking', 'pain', 'support' and 'wheelchair', which are words representing symptoms of MS (Gustavsen et al., 2021). Common words in the diabetes control group are 'diet', 'weight', 'eating', 'sugar', and 'blood', which is as expected given the symptoms of diabetes (Ramachandran, 2014). The generic control group contains common words used in social media, such as 'love', 'game', 'people' and 'one'. Words used commonly in the English language, like 'time' and 'get' were present in all three sections of the dataset. Another notable point is that the word 'ms' is not present in the MS-diagnosed group and the word 'diabetes' is not present in the diabetes control group. This is consistent with the dataset construction method because all clauses with the word 'ms' were removed from the MS-diagnosed dataset and similarly all clauses with the word 'diabetes' were removed from the diabetes control group, as explained in the 'Dataset construction' section above.

In addition to generating the word clouds for the dataset, we decided to count the average number of tokens per comment for each part of the dataset. The results of this investigation are shown in Table 1. We observe that the average number of tokens is the greatest for the diabetes control group, lower for the MS-diagnosed control group and lowest for the generic social media control group. Since sentence length varies significantly across the dataset, we need to ensure that the classifiers we train on the dataset do not rely only on the length of the comment for text classification. The implications of this difference are further discussed in the next section.

5 Experiments

In order to understand which classifier is the best for detecting MS, we experimented several ways to convert text to word vectors and we applied various classification models on these word vectors.

However, before conducting any experiments, we pre-processed the dataset as follows. We removed any punctuation and URLs from the free text and made all the characters lowercase. To handle hashtags, tags, and subreddit names in the text collected from social media, we removed the characters '#', '@' and 'r/' from the start of every word in the free text, if the word started with any of those characters. Finally, we tokenized, lemmatized and removed stop words[4] from the text in the database.

After pre-processing, we explored three ways of converting the pre-processed text into word vectors: bag of words (BoW), tf-idf weighted BoW, and word embeddings. For BoW and tf-idf weighted BoW, the vocabulary was set to 1600. Therefore, the dimension of every word vector was 1600. For word embeddings, we employed word2vec embeddings pretrained using the continuous bag-of-words (CBOW) method on the Google News dataset (Mikolov et al., 2013). The dimension of each of these pretrained word vectors was 300. For each of the classifiers below (except RNN and LSTM), we also average the word embeddings of all the tokens in every comment to get an embedding for each comment. We then use this embedding as input to the model.

The dataset was then split into 80% training dataset and a 20% testing dataset. For each of the three methods to convert text into word vectors, we train and test the dataset on the following classifiers:

Logistic Regression. We started by training a simple logistic regression model on the word vectors.

Naïve Bayes. For the BoW vectors and the tf-idf weighted BoW, we train the probabilistic multinomial Naïve Bayes classifier. Since the multinomial Naïve Bayes classifier expects positive word vectors and our word embeddings contain values from -1 to 1, we normalize the word embedding values to the range 0 to 1 and then train the classifier.

K-Nearest Neighbours (KNN). For the KNN

[4]List of stop words (https://www.doc.ic.ac.uk/~dj321/msnlp/stopwords.txt) was inspired by NLTK's list of English stop words

(a) Diagnosed Users (MS) (b) Control One (Generic) (c) Control Two (Diabetes)

Figure 2: Word Clouds For Each Part of the Dataset

Dataset	Average comment length (in number of tokens)
Diagnosed (MS)	29.1
Control One (Generic)	14.1
Control Two (Diabetes)	85.3

Table 1: Average number of tokens per comment for every part of the dataset

classifier, we set the hyperparameter k to 1, 2, 3, 5, 8, 10, and 100. The best accuracy was then recorded in Table 2.

Neural Network. We decided to investigate both a single-layer neural network and two-layer neural network. We set the number of epochs to 10 and batch size to 64. In the final layer, we use the sigmoid activation function, and for non-output layers, we tried both the ReLU and tanh activation function. For the two-layer neural network, we set the number of neurons in the hidden layer to 50, 100 and 200. After testing these hyperparameters, the best accuracy was then recorded in Table 2.

Recurrent Neural Network (RNN). Since we need an embedding for each word for the RNN, BoW and tfidf-weighted BoW cannot be used for this model. Furthermore, the RNN, by nature of being a sequential model, exemplifies the difference between input sequence lengths. Since we know that the average comment length varies considerably in different sections of the dataset, as shown in Table 1 of section 4, for RNNs we particularly need to make sure that the classifier does not rely only on the length of the text for classification. Therefore, we set the number of RNN units to minimum of the average comment lengths of the datasets in question (rounded up). Accordingly, from the values in Table 1 of section 4, we set the number of RNN units for diagnosed users vs. control one (generic) to be 15 and diagnosed users vs. control two (diabetes) to 30. Then, any comments with longer lengths were truncated to the number of RNN units. This strategy ensures that the classifier relies on the content of the text rather than its length. Lastly, we set the number of epochs to 10 and batch size to 64 before training.

Long Short-Term Memory Network (LSTM). The hyperparameters used for the LSTM model were kept the same as the RNN, and for the same reasons as the RNN. BoW and tfidf-weighted BoW could not be used for this model as well, and the number of LSTM units was again chosen to be 15 with control one and 30 with control two. The number of epochs was set to 10 and batch size to 64.

The results for each of these classification models are reported in Table 2.

Overall, for control one, word embeddings with a two-layer neural network obtained the best accuracy of 94.3%. With control two, tf-idf-weighted BoW with a Naïve Bayes classifier had the best accuracy of 92.4%. We believe that richer, more complicated models like RNNs and LSTMs did not do as well because they must have overfit on the training set, which eventually led to a lower accuracy on the test set.

In general, our experiments with control one had higher accuracy than with control two. Therefore, we conclude that classifiers find it easier to distinguish between generic social media text and the experiences of MS patients rather than distinguish

		MS-Diagnosed vs. Control One (Generic)			MS-Diagnosed vs. Control Two (Diabetes)		
		BoW	tfidf-BoW	Embeddings	BoW	tfidf-BoW	Embeddings
Logistic Regression	Precision	96.3	95.2	90.5	90.3	90.2	87.4
	Accuracy	92.9	93.3	91.8	91.3	91.7	86.4
	Recall	90.5	92.0	93.3	92.5	93.4	86.3
Naïve Bayes	Precision	88.6	89.2	85.4	93.4	93.8	86.8
	Accuracy	91.0	91.6	87.5	92.0	92.4	77.2
	Recall	93.6	94.2	89.7	91.2	91.6	73.5
KNN	Precision	70.8	66.6	85.0	71.5	83.0	94.6
	Accuracy	83.5	78.3	87.1	77.3	85.1	77.8
	Recall	96.1	88.1	89.2	81.3	87.2	71.4
1-Layer Neural Network	Precision	94.2	92.4	94.8	92.3	90.9	91.5
	Accuracy	92.3	92.6	93.9	91.5	90.8	89.8
	Recall	91.2	93.2	93.4	91.1	91.2	88.9
2-Layer Neural Network	Precision	94.1	95.0	95.8	90.2	90.6	88.7
	Accuracy	92.9	93.3	94.3	90.8	91.6	90.1
	Recall	92.3	92.2	93.2	91.7	92.9	91.7
RNN	Precision	N/A	N/A	80.2	N/A	N/A	84.9
	Accuracy	N/A	N/A	86.7	N/A	N/A	88.7
	Recall	N/A	N/A	92.8	N/A	N/A	92.4
LSTM	Precision	N/A	N/A	94.1	N/A	N/A	88.7
	Accuracy	N/A	N/A	94.1	N/A	N/A	90.8
	Recall	N/A	N/A	94.4	N/A	N/A	92.2

Table 2: Results for all the classification models

between experiences of diabetes patients and MS patients.

Another interesting observation is that for control two, word embeddings always performed worse than BoW or tf-idf-weighted BoW, even though word embeddings are usually known for having more semantic richness than BoW models (Mikolov et al., 2013). This result can be explained by the fact that BoW models give more weight to particular words, which likely correspond to symptoms of MS and diabetes in this case. On the other hand, word embeddings cannot distinguish that well between the MS-diagnosed user group and the diabetes control because comments from both these groups contain medical terminology and are thus likely to have similar word embeddings.

6 Conclusion

By employing a novel approach combining social media data with the MS Register dataset of MS patients in the UK, we were able to create the largest ever dataset for computerized diagnosis for MS. We trained multiple different models on this dataset, and we were able to achieve a maximum accuracy of 94.3% when classifying experiences of MS patients against generic social media text data and a maximum accuracy of 92.4% against experiences of diabetes patients. This result is significant because with the classifiers trained in this paper, it may be possible to not only detect multiple sclerosis from the patients' clinical notes, but also from their social media posts. Resulting early diagnosis of MS allows immediate measures to be taken and the symptoms of MS to be controlled.

However, we concede that there are still some limitations in our work that can be improved by future research on this topic. Firstly, there can be further improvements in the construction of the control group of users. With control group two for example, we essentially work with only two diseases – MS and diabetes – which is not like diagnosis in a clinical context, where there are more diseases to account for. In future research therefore, we can have a control group with more illnesses, so the experiments more closely resemble a situation that primary healthcare providers face. In addition, newer, transformer models like BERT and GPT-2 can be applied to this task in future

work. While constructing the dataset, although we remove any clause with the term "MS" to prevent the trivial text classification of post-diagnosis comments, we still assume that the experiences of a patient post-diagnosis are the same the same as the experiences of the patient pre-diagnosis. In future work, patients can be asked about their experiences before the diagnosis of MS to create a more reliable dataset for MS diagnosis.

References

Bruce F Bebo, Mark Allegretta, Douglas Landsman, Kathy M Zackowski, Fiona Brabazon, Walter A Kostich, Timothy Coetzee, Alexander Victor Ng, Ruth Ann Marrie, Kelly R Monk, Amit Bar-Or, and Caroline C Whitacre. 2022. Pathways to cures for multiple sclerosis: A research roadmap. *Multiple Sclerosis Journal*, 28(3):331–345.

Herbert S. Chase, Lindsey R. Mitrani, Gabriel G. Lu, and Dominick J. Fulgieri. 2017. Early recognition of multiple sclerosis using natural language processing of the electronic health record. *BMC Medical Informatics and Decision Making*, 17(1).

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. 2018. Smhd: A large-scale resource for exploring online language usage for multiple mental health conditions.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 1(1):51–60.

David V Ford, Kerina H Jones, Rod M Middleton, Hazel Lockhart-Jones, Inocencio DC Maramba, Gareth J Noble, Lisa A Osborne, and Ronan A Lyons. 2012. The feasibility of collecting information from people with multiple sclerosis for the uk ms register via a web portal: characterising a cohort of people with ms. *BMC Medical Informatics and Decision Making*, 12(1):73.

GDPR. 2016. General data protection regulation (gdpr).

Nazem Ghasemi, Shahnaz Razavi, and Elham Nikzad. 2017. Multiple sclerosis: Pathogenesis, symptoms, diagnoses and cell-based therapy. *Cell journal*, 19(1):1–10.

S. Gustavsen, A. Olsson, H. B. Søndergaard, S. R. Andresen, P. S. Sørensen, F. Sellebjerg, and A. Oturai. 2021. The association of selected multiple sclerosis symptoms with disability and quality of life: a large danish self-report survey. *BMC Neurology*, 21(1).

Nithin Kolanu, A Shane Brown, Amanda Beech, Jacqueline Center, and Christopher Patrick White. 2020. Or29-02 natural language processing of radiology reports improves identification of patients with fracture. *Journal of the Endocrine Society*, 4(1).

Jonathan Koss and Sabine Bohnet-Joschko. 2022. Social media mining in drug development decision making: Prioritizing multiple sclerosis patients' unmet medical needs. *Proceedings of the 55th Hawaii International Conference on System Sciences*.

Huiying Liang, Brian Y. Tsui, Hao Ni, Carolina C. S. Valentim, Sally L. Baxter, Guangjian Liu, Wenjia Cai, Daniel S. Kermany, Xin Sun, Jiancong Chen, Liya He, Jie Zhu, Pin Tian, Hua Shao, Lianghong Zheng, Rui Hou, Sierra Hewett, Gen Li, Ping Liang, and Xuan Zang. 2019. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Medicine*, 25(3):433–438.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Sami Omerhoca, Sinem Yazici Akkas, and Nilufer Kale Icen. 2018. Multiple sclerosis: Diagnosis and differrential diagnosis. *Archives of Neuropsychiatry*, 55(1).

A Ramachandran. 2014. Know the signs and symptoms of diabetes. *The Indian journal of medical research*, 140(5):579–81.

Patrick Schwab and Walter Karlen. 2021. A deep learning approach to diagnosing multiple sclerosis from smartphone data. *IEEE Journal of Biomedical and Health Informatics*, 25(4):1284–1291.

Jordan Swartz, Christian Koziatek, Jason Theobald, Silas Smith, and Eduardo Iturrate. 2017. Creation of a simple natural language processing tool to support an imaging utilization quality dashboard. *International Journal of Medical Informatics*, 101(1):93–99.

Clare Walton, Rachel King, Lindsay Rechtman, Wendy Kaye, Emmanuelle Leray, Ruth Ann Marrie, Neil Robertson, Nicholas La Rocca, Bernard Uitdehaag, Ingrid van der Mei, Mitchell Wallin, Anne Helme, Ceri Angood Napier, Nick Rijke, and Peer Baneke. 2020. Rising prevalence of multiple sclerosis worldwide: Insights from the atlas of ms, third edition. *Multiple Sclerosis Journal*, 26(14):1816–1821.

Jianliang Yang, Yuenan Liu, Minghui Qian, Chenghua Guan, and Xiangfei Yuan. 2019. Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding. *Applied Sciences*, 9(18):3658.