Learning Effective Representations from Clinical Notes

Sebastien Dubois sdubois@stanford.edu
Stanford Center for Biomedical Informatics Research
Stanford University
Stanford, CA 90035, USA

Nathanael Romano naromano@stanford.edu
Stanford Center for Biomedical Informatics Research
Stanford University
Stanford, CA 90035, USA

David C. Kale davekale@cs.stanford.edu
Stanford Center for Biomedical Informatics Research
Stanford University
Stanford, CA 90035, USA

Nigam Shah nigam@stanford.edu
Stanford Center for Biomedical Informatics Research
Stanford University
Stanford, CA 90035, USA

Kenneth Jung kjung@stanford.edu
Stanford Center for Biomedical Informatics Research
Stanford University
Stanford, CA 90035, USA

Abstract

Clinical notes are a rich source of information about patient state. However, using them effectively presents many challenges. In this work we present two methods for summarizing clinical notes into patient-level representations. The resulting representations are evaluated on a range of prediction tasks and cohort sizes. The new representations offer significant predictive performance gains over the common baselines of Bag of Words and topic model representations across all tested tasks and cohort sizes.

1. Introduction

Longitudinal clinical data are complex and high dimensional. Researchers must make challenging decisions that can have a significant impact on predictive performance and clinical findings when preparing the data for analysis. *Finding a good representation for clinical data* is a critical step in designing studies and constitutes a grand challenge for machine learning researchers interested in health care.

In this paper, we study how to learn good representations of clinical notes. Notes record crucial details about patient care and expert insights that cannot be found elsewhere in electronic health records (EHRs), but they also present a number of challenges to health care researchers, including high dimensionality, sparsity, and complex structure. A good representation of clinical text should directly mitigate or admit simple strategies for overcoming these challenges. Moreover, it should be *useful across a wide range of clinical prediction tasks and research questions* and require minimal task-specific customization.

We take a deep learning approach to constructing *data-driven representations of clinical text* that address these challenges and overcome the shortcomings of text representations that are commonly used in clinical informatics research. In particular, we describe two approaches based on deep learning techniques commonly used in natural language processing (NLP): the first approach, *embed and aggregate*, uses GloVe (Pennington et al., 2014) to learn low-dimensional, dense embeddings of clinical terms. We then derive patient-level representations from the term embeddings using simple operations, e.g., max pooling, across individual notes and sequences of notes. In the second approach, we use recurrent neural nets (RNNs) (Elman, 1990) with an embedding layer to directly learn hidden representations of terms, notes, and patients simultaneously.

We trained both approaches on a large corpus of clinical notes from a major US research hospital. We evaluated the resulting patient representations using simple linear classifiers on cohort sizes ranging from 125 to 4,000 patients to predict complex clinical outcomes and diagnoses. Across all outcomes and cohort sizes, our patient embeddings outperformed baseline representations based on bag of words (BOW) and topic models by large margins.

The remainder of this paper is organized as follows: we discuss the challenges and desiderata of effective representations of clinical text in greater detail in section 2. We describe our data, methods, and experimental results in section 3, section 4, and section 5, respectively. We conclude with a discussion of related work (section 6) and the implications and limitations of our work with an eye toward future work (section 7).

2. Background

Clinical text is one of the richest sources of information about patients and health care that can be found in EHRs. Researchers utilize the text in clinical notes for a variety of applications: phenotyping (Ford et al., 2016; Agarwal et al., 2016), pharmacovigilance (LePendu et al., 2013; Harpaz et al., 2014), and modeling risk (Ghassemi et al., 2014; Halpern et al., 2012). A number of studies have shown that adding text as an input alongside physiologic measurements boosts the performance of clinical prediction models by significant margins (Ghassemi et al., 2014; Caballero Barajas and Akella, 2015).

The standard representation for clinical text is bag of words (BOW), which represents a document as counts over a predefined vocabulary. BOW is intuitive and easy to construct and manipulate. It admits simple heuristics for dimensionality reduction (e.g., discarding words that are too frequent or infrequent) and schemes for incorporating information about relative frequency, e.g., TF-IDF. However, BOW fails to fully address many of the challenges presented by clinical text, leading health informatics researchers to seek alternatives (Halpern et al., 2012; Jung et al., 2015; Choi et al., 2016b; Miotto et al., 2016).

First, *clinical notes are intrinsically high dimensional* due to their diversity of topics, specialized vocabulary, and combinatorial nature. Naive representations such as BOW can yield input (feature) vectors with hundreds of thousands of dimensions. Patient cohorts, on the other hand, commonly number in the tens to hundreds, even at hospitals with millions of digital records. This invokes the "curse of dimensionality," making statistical models difficult to train without aggressive dimensionality reduction or regularization.

In a related problem, *clinical text data are also inherently sparse*: phrases describing important clinical concepts may occur in only a handful of records. This makes it difficult to discover a reliable statistical relationship between an outcome and such phrases, as well as the concepts they denote. This problem is exacerbated by *noise and ambiguity in clinical text*: "tachycardia" may alternatively be described as a "rapid heart rate" or "arrhythmia." One can easily imagine cases where the common BOW dimensionality reduction heuristic of discarding rare terms actually removes all mentions of an important concept.

Clinical notes also exhibit a hierarchical sequential structure: a longitudinal patient record includes a time series of notes, each itself consisting of a sequence of words. This poses a trade-off between simplicity and fidelity when choosing a representation: directly modeling this "sequence of sequences" structure requires a complex model architecture and most likely a large amount of training data. In contrast, representing a patient history as a BOW is simple and flexible but discards information about time and order. There is a spectrum of choices between these extremes, but they often involve only slightly less complexity (Choi et al., 2016a) or ad hoc decisions about, e.g., windowing time (Schuler et al., 2016).

In this paper, we present two methods for representing clinical text that occupy different points on the simplicity-richness spectrum. We evaluate the representations within the framework of transfer learning —we first learn how to represent patients using a *source task* with no or relatively plentiful labels, and then use the learned representations for new *target tasks* in which we learn to predict future events from small training samples. The resulting models are then evaluated on held out test patients. The first method, *embed-and-aggregate*, is very simple and requires no external labels for supervision. The second method uses recurrent neural nets to explicitly model the sequential nature of the notes, and is supervised using labels derived from contemporaneous diagnosis codes.

3. Data

Our data set includes records collected over a five year period for patients in a major US research hospital. These records were extracted into a clinical data warehouse, where they were cleaned and de-identified before being made available for research purposes. Each patients data comprises a timestamped series of encounters with care providers over the five year period. Encounters generate both structured and unstructured data. The structured data typically consists of discrete codes from controlled vocabularies for diagnoses, medications, and procedures, along with quantitative measurements such as vitals and laboratory measurements. The unstructured data consists of the free text of clinical notes written by care providers documenting each encounter and other aspects of patient care. In all, the warehouse contains tens of millions of encounters and clinical notes for several million patients. However, the effective size of the data is much smaller (roughly a million notes for a hundred thousand patients) after it has been filtered for usable data as described below.

Text processing

The text of the clinical notes was first processed using the National Center for Biomedical Ontology (NCBO) Annotator (LePendu et al., 2013), which extracts occurrences of terms in an expansive vocabulary of biomedical terms compiled from a collection of controlled terminologies and biomedical ontologies. The NCBO Annotator further flags each mention for negation (e.g., "atrial fibrillation was ruled out") and family history (e.g., "The patients father had hypertension"). Following annotation, each note is converted into a BOW representation and any terms that are not recognized as biomedical concepts are discarded. This is per institutional policy and not a specific research choice.

Finally, terms are normalized to unique biomedical concepts using the Unified Medical Language System MetaThesaurus, which provides a mapping of strings to Concept Unique Identifiers (CUIs) (Bodenreider, 2004). We retained negated concepts as distinct elements in our vocabulary because negative findings are often informative. Finally, we removed concepts that appeared in fewer than 50 or more than 10 million notes. This resulted in a final vocabulary of 107,388 CUIs. For the remainder of this paper, we will refer to CUIs as "words" for convenience. We discuss implications of this text processing in section 4 and section 7.

Data splits

Each patient's longitudinal record was split into two "eras," a label era that includes the last six months of the record and an input era comprising the 12 months prior to the label era. *We derived labels for target learning tasks, e.g., mortality, from structured data in the label era only.* Clinical notes from the input era form the inputs for the target tasks. *We trained embedding models (and topic model baselines) on data from the input era exclusively.*

We retained only those patients with usable labels and one or more non-empty annotated clinical notes in their input era, yielding a final data set of 115,232 patients and 2,735,6487 notes. This final set was then split into training, validation, and test sets with 69,417, 11,290, and 34,525 patients and 1,644,841, 275,684, and 815,122 notes, respectively.

4. Methods

We approach learning to represent patient clinical notes in the framework of transfer learning, in which we learn to represent a patient's clinical notes on a *source task*, and use the resulting representations for new *target tasks*.

Embed and Aggregate

In this approach, we break up the source task into two phases. First, we learn word embeddings using methods such as GloVe or word2vec. These embeddings are then aggregated into patient level embeddings using simple functions —min, mean, and max —applied elementwise to the embedding dimensions. Note that this aggregation is over two levels —word embeddings are first combined into note representations, and then into patient level representations. Neither phase uses external labels, and are unsupervised. We explored word embeddings learned from both clinical text and from biomedical research article abstracts as detailed below.

GloVe embeddings learned from clinical notes

We used GloVe to learn word embeddings from clinical text using a range of dimensionality and context sizes. In order to overcome the loss of the ordering of words in our input corpus, we adopted a technique used for learning embeddings of medical codes (Choi et al., 2016a,b), which are naturally unordered, and represent each note using 2 random orderings of its words. We ran GloVe for 25 iterations.

Word embeddings from research article abstracts

We also evaluated word embeddings learned from 350,000 biomedical journal article abstracts in (De Vine et al., 2014). Terms in the abstract text were mapped to UMLS CUIs, and the embeddings were learned using the word2vec skip-gram model with a window size of 5 and dimensionality of 200. Embeddings for 52,000 CUIs are publicly available,[1] and approximately 28,000 of those are in our vocabulary. We refer to these embeddings as MCEMJ (Medical Concept Embeddings from Medical Journals).

Aggregation of word embeddings

We constructed fixed length representations of the patients using simple functions – min, mean, and max – applied element-wise to the word embeddings within each note. We explored a range of embedding sizes, along with concatenation of embeddings resulting from different aggregation methods, and concatenation of embeddings learned from different data sources.

Recurrent neural nets

In this approach, we use recurrent neural nets whose inputs are a sequence of bag of words vectors representing the clinical notes for a given patient's input era. This approach jointly learns word-, note-, and patient-level embeddings. Models are trained using input era data only and are then used as fixed feature extractors for the target tasks.

Neural Net Supervision

Supervision for the recurrent neural net source task was provided by Clinical Classification Software (CCS)[2] codes derived from structured diagnosis codes from the input era. CCS maps 12,856 diagnosis codes into 254 broad disease categories such as *lung cancer* or *cardiac dysrhythmias*, and have previously been used as supervision labels for the training of recurrent neural nets (Choi et al., 2016a) because it dramatically reduces the label space. We formulated learning as a multi-task sequence classification problem in which the final hidden state of the net is used as a representation for a multi-task problem with CCS labels aggregated over the patient's input era. In preliminary experiments, this approach consistently outperformed providing supervision at each step using only that encounter's codes.

Architecture

As in other work using recurrent nets to model patient data, the input to our recurrent nets were many-hot vectors representing the unordered bag of words for each note (Choi et al., 2016a). Typically, the first layer of the network multiplies an embedding matrix by this vector, effectively summing the embeddings to arrive at a note level representation that is

1. `https://github.com/clinicalml/embeddings`
2. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

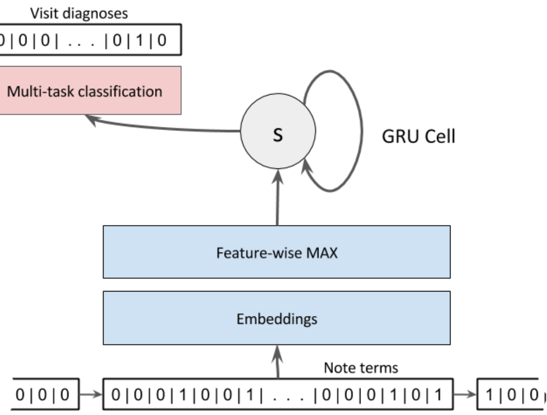


Figure 1: Model architecture

input to recurrent cells. However, we found that applying an element-wise maximum to the embeddings at each time-step significantly improved performance in the target tasks.

Finally, we explored using Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Gated Recurrent Units (GRU) (Cho et al., 2014) as recurrent cells. We found that, consistent with (Choi et al., 2016a), GRUs consistently outperformed LSTMs. Thus, our final experiments used GRUs with max aggregation of word embeddings.

Learning details

We used `tanh` activations, dropout with drop probability 0.2, and a dense layer with binary cross-entropies for the multi-task supervision. Networks were trained with RMSProp (Tieleman and Hinton, 2012) for 100 epochs using a batch size of 32 and maximum sequence length of 50, with longer sequences truncated to keep the most recent notes. The models were implemented in Keras 1.2.0 (Chollet, 2015) with a Tensorflow 0.12.1 backend (Abadi et al., 2016) on an NVIDIA GeForce GTX TITAN X GPU.

We explored initializing the embedding layer of our recurrent nets using word embeddings learned from the clinical notes or from from research article abstracts. We found that random initialization consistently outperformed initialization with pre-trained embeddings.

Baselines - Bag of words and topic models

We compared the learned representations against BOW and topic model representations of words and notes respectively. For the BOW experiments, we limited the vocabulary to the 15,000 most frequent words in each training subsample. Word counts were then summed over each patient's input notes. Latent Dirichlet Allocation was used to fit a topic model on training notes using Gensim (Řehůřek and Sojka, 2010). The model was then used to estimate topic distributions for validation and test set notes. 300 topics were fit over two passes through the training notes. The topic distributions for each note were aggregated into patient-level representations using mean and max pooling.

5. Experiments

Our primary evaluation consists of 9 prediction tasks in which representations derived from clinical notes from the input eras of training patients are used to fit linear models predicting the occurrence of clinical outcomes in the patients' label eras. The resulting models are evaluated on held out test patients. The prediction targets include three complex outcomes – all causes mortality, inpatient admissions, and emergency room visits – in addition to the six more prevalent CCS diagnosis code categories - cardiac dysrhythmias; diabetes mellitus without complication; disorders of lipid metabolism; essential hypertension; spondylosis, invertebral disc disorders, other back problems; and thyroid disorders. Note that the former targets were not used in any source task, while the latter targets were used in supervision of the RNN models, although in they were contemporaneous with the inputs instead of occurring in the future in that case.

For each task and representation, we take 20 random subsamples of training patients, balanced to 20% prevalence of positive cases. The various patient representations for these subsamples are used to fit L2 regularized logistic regression models, tuned by 5-fold cross validation on each training subsample. This process is repeated with N = 125, 250, 500, 1000, 2000, and 4000 random training patients in each subsample. Models fit to each subsample are evaluated on a set of test patients with the natural prevalence of the outcomes. Figure 2 and Figure 3 show plots of performance on the three complex outcomes and one CCS code target (cardiac dysrhythmias), as function of training set size. We include representative examples of class of representation, including baselines based on topic models and BOW. Results for all experiments are presented in table form in Appendix A.

For the embed-and-aggregate representations, it was advantageous to concatenate the min, mean and max aggregated embeddings derived from both the clinical notes and research article abstracts. We report results for this representation along with the best performing embed-and-aggregate representation using just clinical notes. Among the RNN representations, we found that the max aggregated GRU with 600 hidden units and random initialization generally performed best across all tasks and subsample sizes.

For all outcomes, we see the same general pattern: the RNN-based representation was dominant at small sample sizes (N less than 1000), but for larger training sets, the embed-and-aggregate features and BOW baseline rapidly caught up and, in some cases, surpassed the RNN (ER visits for the embeddings, mortality for both). This suggests that the difference in performance is attributable at least in part to dimensionality: the embeddings have 3-4 times as many features as the RNN, BOW another order of magnitude more. For very small values of N, these high dimensional data make learning very difficult and require greater regularization. As the number of examples approaches the number of features, (1000 examples or more), the classifier is better able to exploit the the higher dimensional representations.

For the CCS code outcomes, the RNN representation maintained a consistent advantage over the other methods, even as the sample size increases. This is likely because supervision of the RNN models includes these outcomes, albeit in the input era instead of in the future. The source task training encourages the RNN to encode information about past diagnoses, which is unsurprisingly useful for predicting future diagnostic codes. One could easily imagine closing the gap with the RNN by simply appending past diagnoses (or some representation thereof) to the embeddings, which already perform quite well at small values of N.

Both of the proposed representations dominate the topic model-based features, which have been a popular representation for clinical text in recent years (Halpern et al., 2012; Ghassemi et al., 2014; Miotto et al., 2016). What is more, the topic model baseline is outperformed by the BOW baseline at even modest sample sizes (500 examples or more). One could imagine a variety of schemes for trying to improve topic model-based performance, such as using more topics or including supervision, but previous work casts doubt on whether this will be successful (Halpern et al., 2012).

Intrinsic evaluation of word embeddings

Word embeddings are typically evaluated using analogy-like tasks that test whether or not the geometry of the word embeddings recapitulates known relationships between words. We evaluated the word level embeddings learned from clinical text, research article abstracts, and through the recurrent net models using the Medical Relatedness Property framework developed in (Choi et al., 2016b). In this framework, we test whether word embeddings express known *may treat* and *may prevent* relationships between drugs and indications using vector arithmetic. For instance, if d_1, d_2, m_1, and m_2 denote two drugs and indications respectively, and the relation r holds for (d_1, m_1) and (d_2, m_2), we check whether or not the following holds for their respective embeddings:

$$e_{d_1} - e_{m_1} + e_{m_2} \approx e_{d_2}$$

More precisely, the quality metric is the ratio of concepts for which at least one of the top-40 neighbors of $e_{d_1} - e_{m_1} + e_{m_2}$ satisfies the relation r with m_2. Results are shown in Table 1.

Embedding method	**May-Treat (%)**	**May-Prevent (%)**
GloVe	9.23	10.21
MCEMJ	8.25	6.81
GRU	1.26	0.43

Table 1: Evaluation of the Medical Relatedness Property for two relations: May-Treat and May-Prevent.

Interestingly, we found that GloVe embeddings performed better than the MCEMJ embeddings, despite randomized word orders and a significantly larger vocabulary. In addition, embeddings from randomly initialized neural models perform poorly at this task despite their effectiveness in the patient level prediction tasks, suggesting that word level evaluations are not particularly predictive of utility in such tasks.

6. Related Work

There has been a great deal of research showing the utility of clinical text for a variety of applications, including pharmacovigilance (LePendu et al., 2013), coding (Perotte et al., 2014), predictive modeling (Ghassemi et al., 2014; Yu et al., 2015), and phenotyping (Ford et al., 2016). BOW is the standard representation of choice, though often in combination with substantial preprocessing. One of the longstanding goals of clinical informaticists who work with text is to eliminate some or all of these preprocessing steps. For example, Jung et al. (2015) found that across a number of clinical data mining questions, using NCBO Annotator used in this work yielded comparable performance to a more computationally intensive commercial NLP pipeline including syntactic parsing and part-of-speech tagging.

Although our results suggest they may be superseded by word embeddings, topic models such as latent Dirichlet allocation (Blei et al., 2003) have been a popular choice for learning low-dimensional representations of clinical text. Halpern et al. (2012) found that topics underperformed BOW when used as features in a prediction model. Our results contradict theirs, but that is likely explained by our much larger training corpus. Chen et al. (2015) showed a high level of agreement between topic models trained on corpora from different institutions, indicating that they may provide a generic representation that can cross boundaries. Our results using the MCEMJ embeddings (De Vine et al., 2014) suggest the same may be true of embeddings.

Miotto et al. (2016) use a topic model to preprocess clinical notes before combining them with counts over structured clinical data. This combined representation is then fed into an autoencoder, followed by a random forest classifier to predict future diagnoses. This complex pipeline makes it difficult to evaluate the individual contributions of any one part.

Inspired by recent successes of word embedding models like word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014), a number of health care researchers have sought to learn embeddings of medical words or concepts (Minarro-Giménez et al., 2013). De Vine et al. (2014) trained the MCEMJ embeddings we used in our experiments, while Choi et al. (2016b) focused on designing quantitative evaluations of how well embeddings capture known medical concepts. None of this work, however, evaluated the utility of their embeddings as features in predictive models. Liu et al. (2016) described an innovative way to leverage embeddings to perform transfer learning in low data regimes like ours. They provide a useful initial bias by rescaling each feature proportional to its similarity with the prediction target based on text descriptions of each.

There is a growing interest in directly modeling sequential clinical data without ad hoc feature engineering. Lipton et al. (2015) and Choi et al. (2016a) apply RNNs to time series of physiology and codes, respectively, while Razavian et al. (2016) applies a convolutional network to longitudinal lab results. Several recent works have modeled sequences of clinical notes as time series of topics (Ghassemi et al., 2014, 2015; Caballero Barajas and Akella, 2015). Our work stands out in two ways: first, we focus on the small patient cohort regime, whereas these prior works use far larger training sets. Second, we find that a conceptually simple approach, embed-and-aggregate, is competitive with a more complex architecture.

7. Discussion

We have presented two methods for representing clinical notes that offer compelling advantages over BOW and topic model representations across a range of tasks and cohort sizes. The *embed-and-aggregate* method in particular represents a sweet spot in the performance-complexity trade-off because it is remarkably simple to learn, distribute and use, requiring little in the way of specialized hardware or software.

Our work has important limitations. First,the evaluation tasks are not entirely in the spirit of the work; targets like future inpatient admissions and diagnoses are actually plentiful since they can be derived from structured EHR data. It would be preferable to evaluate on prediction targets derived from clinician chart review for rare or subtle diagnoses, and to evaluate them in cohorts of patients that reflect intended usage instead of the general population. It is for such targets that improved patient representations can have the greatest impact.

Second, our methods start with highly processed abstracts of the raw text. This processing trades off information about the text for simplicity and ease of use. However, it may also leave a lot of performance on the table. We found that embeddings learned from randomized orderings of the processed notes performs better in our evaluations than embeddings learned from the full text of journal abstracts. In fairness, however, the latter embeddings were tuned through intrinsic evaluations rather than prediction performance.

Our results also beg the question of why recurrent neural net representations, which more explicitly model the structure of the data, do not offer a more significant performance advantage than the embed-and-aggregate. We suspect this is because we do not in fact have enough data and good supervision to support the training of a higher capacity model. We note that the best embed-and-aggregate representations have significantly higher dimensionality than the best RNN representation. Finally, we have not addressed interpretability and portability of the representations across institutions.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation*, OSDI'16, pages 265–283, Berkeley, CA, USA, 2016. USENIX Association. ISBN 978-1-931971-33-1. URL `http://dl.acm.org/citation.cfm?id=3026877.3026899`.

Vibhu Agarwal, Tanya Podchiyska, Juan M Banda, Veena Goel, Tiffany I Leung, Evan P Minty, Timothy E Sweeney, Elsie Gyang, and Nigam H Shah. Learning statistical models of phenotypes using noisy labeled training data. *Journal of the American Medical Informatics Association*, page ocw028, 2016.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet Allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue):D267–70, 2004.

Karla L Caballero Barajas and Ram Akella. Dynamically modeling patient's health state from electronic medical records: A time series approach. In *Proceedings of the 21th ACM*

SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 69–78. ACM, 2015.

You Chen, Joydeep Ghosh, Cosmin Adrian Bejan, Carl A. Gunter, Siddharth Gupta, Abel Kho, David Liebovitz, Jimeng Sun, Joshua Denny, and Bradley Malin. Building bridges across electronic health record systems through inferred phenotypic topics. *Journal of Biomedical Informatics*, 55:82 – 93, 2015. ISSN 1532-0464. doi: http://doi.org/10.1016/j.jbi.2015.03.011. URL `http://www.sciencedirect.com/science/article/pii/S1532046415000544`.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Edward Choi, Mohammad Taha Bahadori, and Jimeng Sun. Doctor AI: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, 2016a.

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016: 41, 2016b.

François Chollet. Keras, 2015.

Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1819–1822. ACM, 2014.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007, 2016. doi: 10.1093/jamia/ocv180. URL `+http://dx.doi.org/10.1093/jamia/ocv180`.

Marzyeh Ghassemi, Tristan Naumann, Finale Doshi-Velez, Nicole Brimmer, Rohit Joshi, Anna Rumshisky, and Peter Szolovits. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 75–84. ACM, 2014.

Marzyeh Ghassemi, Marco AF Pimentel, Tristan Naumann, Thomas Brennan, David A Clifton, Peter Szolovits, and Mengling Feng. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, volume 2015, page 446. NIH Public Access, 2015.

Yoni Halpern, Steven Horng, Larry A Nathanson, Nathan I Shapiro, and David Sontag. A comparison of dimensionality reduction techniques for unstructured clinical text. In *ICML Workshop on Clinical Data Analysis*, 2012.

Rave Harpaz, Alison Callahan, Suzanne Tamang, Yen Low, David Odgers, Sam Finlayson, Kenneth Jung, Paea LePendu, and Nigam H Shah. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug safety*, 37(10):777–790, 2014.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Kenneth Jung, Paea LePendu, Srinivasan Iyer, Anna Bauer-Mehren, Bethany Percha, and Nigam H Shah. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. *Journal of the American Medical Informatics Association*, 22(1): 121, 2015. doi: 10.1136/amiajnl-2014-002902. URL `+http://dx.doi.org/10.1136/amiajnl-2014-002902`.

Paea LePendu, Srinivasan V Iyer, Anna Bauer-Mehren, Rave Harpaz, Jonathan M Mortensen, Tanya Podchiyska, Todd A Ferris, and Nigam H Shah. Pharmacovigilance using clinical notes. *Clinical pharmacology & therapeutics*, 93(6):547–555, 2013.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzell. Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

Yun Liu, Collin M Stultz, John V Guttag, Kun-Ta Chuang, Fu-Wen Liang, and Huey-Jen Su. Transferring knowledge from text to predict disease onset. 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

José Antonio Minarro-Giménez, Oscar Marín-Alonso, and Matthias Samwald. Exploring the application of deep learning techniques on medical text corpora. *Studies in health technology and informatics*, 205:584–588, 2013.

Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6, 2016.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. Diagnosis code assignment: models and evaluation metrics. *Journal of the American Medical Informatics Association*, 21(2):231–237, 2014.

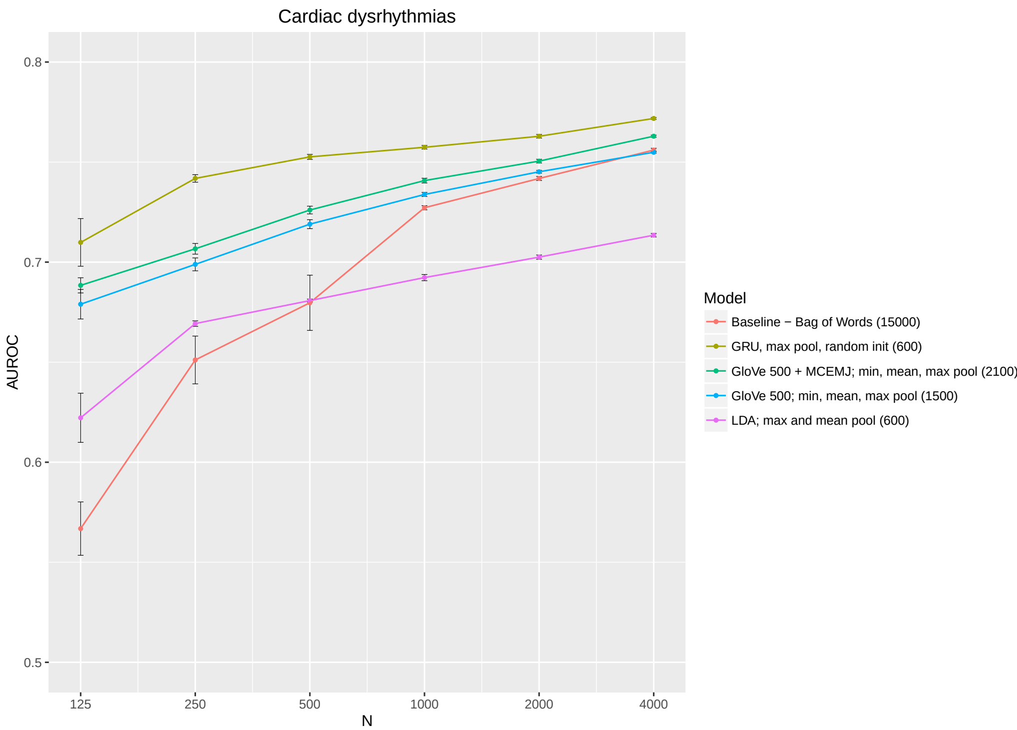
Narges Razavian, Jake Marcus, and David Sontag. Multi-task prediction of disease onsets from longitudinal lab tests. In *Machine Learning for Healthcare Conference*, 2016.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

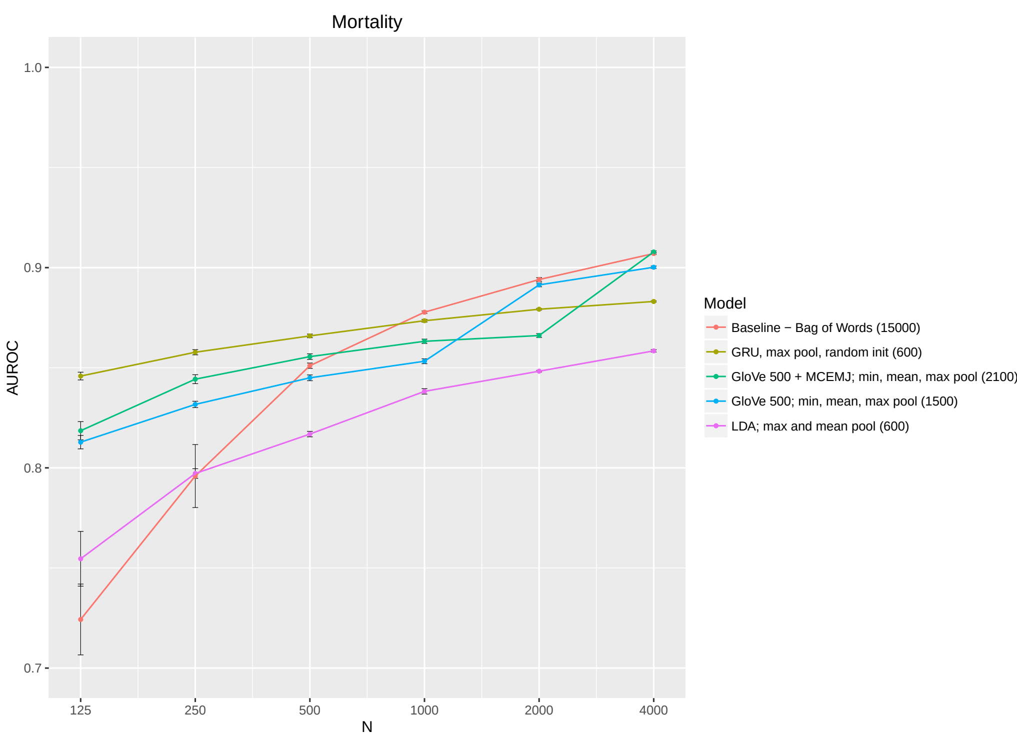
Alejandro Schuler, Vincent Liu, Joe Wan, Alison Callahan, Madeleine Udell, David E Stark, and Nigam H Shah. Discovering patient phenotypes using generalized low rank models. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 21, page 144. NIH Public Access, 2016.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012.

Sheng Yu, Katherine P Liao, Stanley Y Shaw, Vivian S Gainer, Susanne E Churchill, Peter Szolovits, Shawn N Murphy, Isaac S. Kohane, and Tianxi Cai. Toward high-throughput phenotyping: unbiased automated feature extraction and selection from knowledge sources. *Journal of the American Medical Informatics Association*, 22(5):993, 2015. doi: 10.1093/jamia/ocv034. URL +http://dx.doi.org/10.1093/jamia/ocv034.

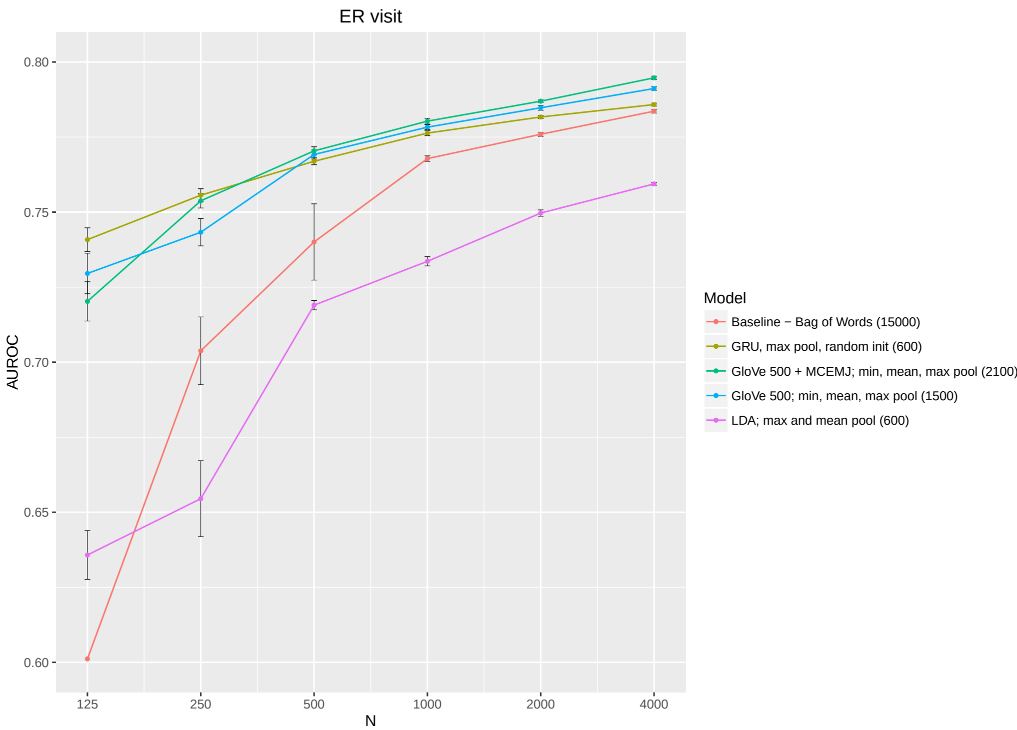


(a)

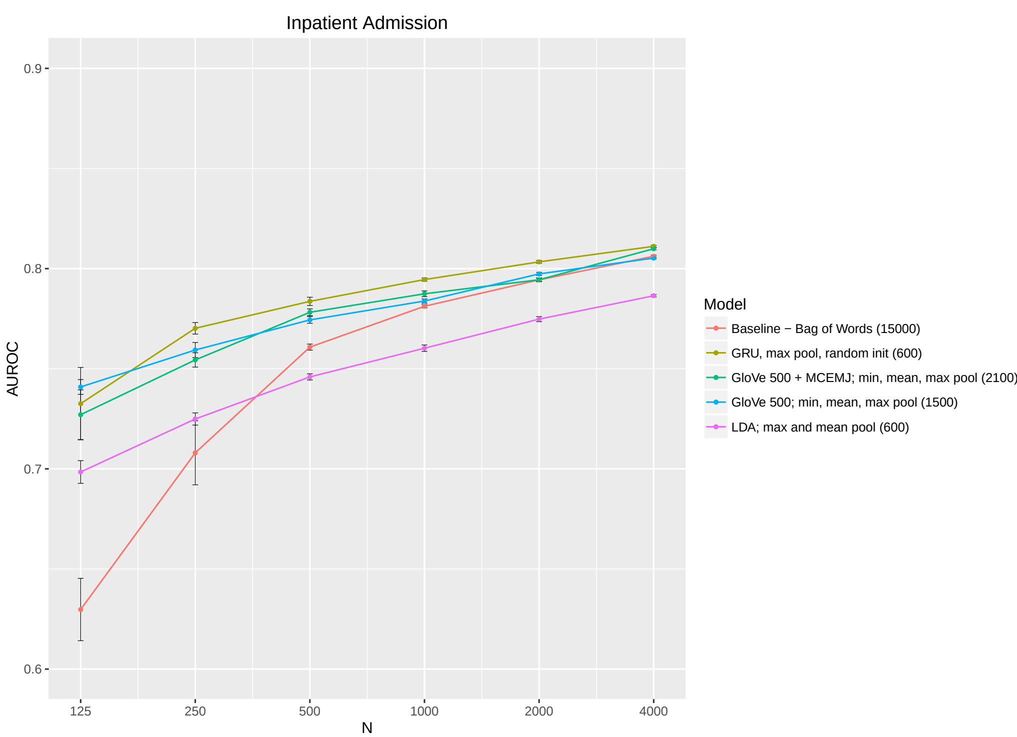


(b)

Figure 2: Target task performance as a function of training set size. (a) Cardiac dysrhythmia diagnosis. (b) Mortality. The dimensionality of each representation is in parentheses.



(a)



(b)

Figure 3: Target task performance as a function of training set size. (a) Future ER visit. (b) Future admission. The dimensionality of each representation is in parentheses.

Appendices

A. Full Results

N	Task	Model	AUROC (SEM)
125	Mortality	GRU, max pool, PubMed init (200)	0.8401 (0.0028)
125	Mortality	GRU, max pool, random init (600)	0.8458 (0.0019)
125	Mortality	GRU, max pool, random init (300)	0.8531 (0.002)
125	Mortality	GRU, max pool, PubMed init (500)	0.8338 (0.0022)
125	Mortality	Baseline - Bag of Words (15000)	0.7243 (0.0177)
125	Mortality	MCEMJ; min, mean, max pool (600)	0.8108 (0.0038)
125	Mortality	GloVe 300; min, mean, max pool (900)	0.8041 (0.0043)
125	Mortality	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.8242 (0.0019)
125	Mortality	GloVe 500; min, mean, max pool (1500)	0.8128 (0.0034)
125	Mortality	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.8185 (0.0046)
125	Mortality	LDA; max pool (300)	0.7774 (0.0024)
125	Mortality	LDA; max and mean pool (600)	0.7546 (0.0137)
125	Mortality	LDA; mean pool (300)	0.7112 (0.0035)
125	ER visit	GRU, max pool, PubMed init (200)	0.7285 (0.0037)
125	ER visit	GRU, max pool, random init (600)	0.7408 (0.0039)
125	ER visit	GRU, max pool, random init (300)	0.7258 (0.0124)
125	ER visit	GRU, max pool, PubMed init (500)	0.7102 (0.0134)
125	ER visit	Baseline - Bag of Words (15000)	0.6011 (0.0144)
125	ER visit	MCEMJ; min, mean, max pool (600)	0.6884 (0.013)
125	ER visit	GloVe 300; min, mean, max pool (900)	0.7184 (0.0124)
125	ER visit	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7336 (0.0065)
125	ER visit	GloVe 500; min, mean, max pool (1500)	0.7296 (0.0068)
125	ER visit	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7203 (0.0065)
125	ER visit	LDA; max pool (300)	0.6396 (0.0045)
125	ER visit	LDA; max and mean pool (600)	0.6358 (0.0081)
125	ER visit	LDA; mean pool (300)	0.6068 (0.0117)
125	Inpatient Admission	GRU, max pool, PubMed init (200)	0.7496 (0.005)
125	Inpatient Admission	GRU, max pool, random init (600)	0.7326 (0.0181)
125	Inpatient Admission	GRU, max pool, random init (300)	0.7645 (0.0042)
125	Inpatient Admission	GRU, max pool, PubMed init (500)	0.7394 (0.0132)
125	Inpatient Admission	Baseline - Bag of Words (15000)	0.6298 (0.0156)
125	Inpatient Admission	MCEMJ; min, mean, max pool (600)	0.7363 (0.0026)
125	Inpatient Admission	GloVe 300; min, mean, max pool (900)	0.7434 (0.0037)
125	Inpatient Admission	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7306 (0.0126)
125	Inpatient Admission	GloVe 500; min, mean, max pool (1500)	0.7409 (0.0037)
125	Inpatient Admission	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7271 (0.0124)
125	Inpatient Admission	LDA; max pool (300)	0.7121 (0.0019)
125	Inpatient Admission	LDA; max and mean pool (600)	0.6984 (0.0057)
125	Inpatient Admission	LDA; mean pool (300)	0.6317 (0.0132)

N	Task	Model	AUROC (SEM)
125	Cardiac dysrhythmias	GRU, max pool, PubMed init (200)	0.7196 (0.0028)
125	Cardiac dysrhythmias	GRU, max pool, random init (600)	0.7099 (0.0119)
125	Cardiac dysrhythmias	GRU, max pool, random init (300)	0.7301 (0.0038)
125	Cardiac dysrhythmias	GRU, max pool, PubMed init (500)	0.7239 (0.0026)
125	Cardiac dysrhythmias	Baseline - Bag of Words (15000)	0.5668 (0.0133)
125	Cardiac dysrhythmias	MCEMJ; min, mean, max pool (600)	0.6921 (0.004)
125	Cardiac dysrhythmias	GloVe 300; min, mean, max pool (900)	0.6829 (0.0037)
125	Cardiac dysrhythmias	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6844 (0.0102)
125	Cardiac dysrhythmias	GloVe 500; min, mean, max pool (1500)	0.679 (0.0074)
125	Cardiac dysrhythmias	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6884 (0.0038)
125	Cardiac dysrhythmias	LDA; max pool (300)	0.6554 (0.0097)
125	Cardiac dysrhythmias	LDA; max and mean pool (600)	0.6222 (0.0123)
125	Cardiac dysrhythmias	LDA; mean pool (300)	0.5976 (0.0078)
125	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (200)	0.7365 (0.0026)
125	Diabetes mellitus w/o complic.	GRU, max pool, random init (600)	0.7241 (0.0122)
125	Diabetes mellitus w/o complic.	GRU, max pool, random init (300)	0.74 (0.0024)
125	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (500)	0.72 (0.0052)
125	Diabetes mellitus w/o complic.	Baseline - Bag of Words (15000)	0.5747 (0.0096)
125	Diabetes mellitus w/o complic.	MCEMJ; min, mean, max pool (600)	0.6756 (0.0095)
125	Diabetes mellitus w/o complic.	GloVe 300; min, mean, max pool (900)	0.6587 (0.0083)
125	Diabetes mellitus w/o complic.	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6617 (0.0094)
125	Diabetes mellitus w/o complic.	GloVe 500; min, mean, max pool (1500)	0.6648 (0.0035)
125	Diabetes mellitus w/o complic.	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6693 (0.0046)
125	Diabetes mellitus w/o complic.	LDA; max pool (300)	0.609 (0.0142)
125	Diabetes mellitus w/o complic.	LDA; max and mean pool (600)	0.6098 (0.0109)
125	Diabetes mellitus w/o complic.	LDA; mean pool (300)	0.5638 (0.0084)
125	Spondylosis (back pain)	GRU, max pool, PubMed init (200)	0.7099 (0.0168)
125	Spondylosis (back pain)	GRU, max pool, random init (600)	0.7345 (0.0034)
125	Spondylosis (back pain)	GRU, max pool, random init (300)	0.6955 (0.0193)
125	Spondylosis (back pain)	GRU, max pool, PubMed init (500)	0.7075 (0.006)
125	Spondylosis (back pain)	Baseline - Bag of Words (15000)	0.589 (0.0095)
125	Spondylosis (back pain)	MCEMJ; min, mean, max pool (600)	0.678 (0.0042)
125	Spondylosis (back pain)	GloVe 300; min, mean, max pool (900)	0.6679 (0.0057)
125	Spondylosis (back pain)	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6602 (0.0097)
125	Spondylosis (back pain)	GloVe 500; min, mean, max pool (1500)	0.6587 (0.0091)
125	Spondylosis (back pain)	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6797 (0.0046)
125	Spondylosis (back pain)	LDA; max pool (300)	0.6073 (0.0121)
125	Spondylosis (back pain)	LDA; max and mean pool (600)	0.5632 (0.0128)
125	Spondylosis (back pain)	LDA; mean pool (300)	0.5483 (0.0061)

N	Task	Model	AUROC (SEM)
125	Disorders of lipid metabolism	GRU, max pool, PubMed init (200)	0.6972 (0.0062)
125	Disorders of lipid metabolism	GRU, max pool, random init (600)	0.7089 (0.0033)
125	Disorders of lipid metabolism	GRU, max pool, random init (300)	0.6944 (0.0151)
125	Disorders of lipid metabolism	GRU, max pool, PubMed init (500)	0.697 (0.0026)
125	Disorders of lipid metabolism	Baseline - Bag of Words (15000)	0.5598 (0.0071)
125	Disorders of lipid metabolism	MCEMJ; min, mean, max pool (600)	0.6197 (0.0077)
125	Disorders of lipid metabolism	GloVe 300; min, mean, max pool (900)	0.6011 (0.0124)
125	Disorders of lipid metabolism	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6243 (0.0086)
125	Disorders of lipid metabolism	GloVe 500; min, mean, max pool (1500)	0.6259 (0.007)
125	Disorders of lipid metabolism	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6224 (0.0103)
125	Disorders of lipid metabolism	LDA; max pool (300)	0.571 (0.0104)
125	Disorders of lipid metabolism	LDA; max and mean pool (600)	0.5839 (0.0063)
125	Disorders of lipid metabolism	LDA; mean pool (300)	0.5392 (0.0061)
125	Essential hypertension	GRU, max pool, PubMed init (200)	0.7173 (0.0116)
125	Essential hypertension	GRU, max pool, random init (600)	0.7076 (0.0114)
125	Essential hypertension	GRU, max pool, random init (300)	0.7324 (0.0025)
125	Essential hypertension	GRU, max pool, PubMed init (500)	0.7162 (0.003)
125	Essential hypertension	Baseline - Bag of Words (15000)	0.5563 (0.0071)
125	Essential hypertension	MCEMJ; min, mean, max pool (600)	0.6444 (0.0115)
125	Essential hypertension	GloVe 300; min, mean, max pool (900)	0.6452 (0.0083)
125	Essential hypertension	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6499 (0.0087)
125	Essential hypertension	GloVe 500; min, mean, max pool (1500)	0.6443 (0.0069)
125	Essential hypertension	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6409 (0.0118)
125	Essential hypertension	LDA; max pool (300)	0.6206 (0.0073)
125	Essential hypertension	LDA; max and mean pool (600)	0.5951 (0.0095)
125	Essential hypertension	LDA; mean pool (300)	0.5514 (0.0059)
125	Thyroid disorders	GRU, max pool, PubMed init (200)	0.6574 (0.0094)
125	Thyroid disorders	GRU, max pool, random init (600)	0.6902 (0.0151)
125	Thyroid disorders	GRU, max pool, random init (300)	0.6907 (0.0057)
125	Thyroid disorders	GRU, max pool, PubMed init (500)	0.6452 (0.0052)
125	Thyroid disorders	Baseline - Bag of Words (15000)	0.5346 (0.0055)
125	Thyroid disorders	MCEMJ; min, mean, max pool (600)	0.6388 (0.0054)
125	Thyroid disorders	GloVe 300; min, mean, max pool (900)	0.6141 (0.0063)
125	Thyroid disorders	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6238 (0.008)
125	Thyroid disorders	GloVe 500; min, mean, max pool (1500)	0.6124 (0.0096)
125	Thyroid disorders	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6265 (0.0079)
125	Thyroid disorders	LDA; max pool (300)	0.5737 (0.0077)
125	Thyroid disorders	LDA; max and mean pool (600)	0.553 (0.0079)
125	Thyroid disorders	LDA; mean pool (300)	0.525 (0.0054)

N	Task	Model	AUROC (SEM)
250	Mortality	GRU, max pool, PubMed init (200)	0.849 (0.0012)
250	Mortality	GRU, max pool, random init (600)	0.8577 (0.0013)
250	Mortality	GRU, max pool, random init (300)	0.8628 (0.001)
250	Mortality	GRU, max pool, PubMed init (500)	0.8429 (0.0014)
250	Mortality	Baseline - Bag of Words (15000)	0.7959 (0.0157)
250	Mortality	MCEMJ; min, mean, max pool (600)	0.8241 (0.0014)
250	Mortality	GloVe 300; min, mean, max pool (900)	0.8278 (0.0025)
250	Mortality	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.8379 (0.0021)
250	Mortality	GloVe 500; min, mean, max pool (1500)	0.8317 (0.0016)
250	Mortality	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.8443 (0.0022)
250	Mortality	LDA; max pool (300)	0.8008 (0.0019)
250	Mortality	LDA; max and mean pool (600)	0.7972 (0.0024)
250	Mortality	LDA; mean pool (300)	0.7359 (0.0127)
250	ER visit	GRU, max pool, PubMed init (200)	0.7487 (0.0028)
250	ER visit	GRU, max pool, random init (600)	0.7556 (0.0022)
250	ER visit	GRU, max pool, random init (300)	0.7452 (0.0024)
250	ER visit	GRU, max pool, PubMed init (500)	0.7447 (0.0036)
250	ER visit	Baseline - Bag of Words (15000)	0.7038 (0.0113)
250	ER visit	MCEMJ; min, mean, max pool (600)	0.7426 (0.0028)
250	ER visit	GloVe 300; min, mean, max pool (900)	0.7446 (0.0069)
250	ER visit	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7593 (0.0028)
250	ER visit	GloVe 500; min, mean, max pool (1500)	0.7433 (0.0046)
250	ER visit	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7538 (0.0024)
250	ER visit	LDA; max pool (300)	0.6591 (0.0093)
250	ER visit	LDA; max and mean pool (600)	0.6545 (0.0126)
250	ER visit	LDA; mean pool (300)	0.6516 (0.0089)
250	Inpatient Admission	GRU, max pool, PubMed init (200)	0.7631 (0.0042)
250	Inpatient Admission	GRU, max pool, random init (600)	0.7702 (0.0029)
250	Inpatient Admission	GRU, max pool, random init (300)	0.7832 (0.002)
250	Inpatient Admission	GRU, max pool, PubMed init (500)	0.7641 (0.0025)
250	Inpatient Admission	Baseline - Bag of Words (15000)	0.708 (0.016)
250	Inpatient Admission	MCEMJ; min, mean, max pool (600)	0.7504 (0.0028)
250	Inpatient Admission	GloVe 300; min, mean, max pool (900)	0.7613 (0.003)
250	Inpatient Admission	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7572 (0.0044)
250	Inpatient Admission	GloVe 500; min, mean, max pool (1500)	0.7594 (0.0038)
250	Inpatient Admission	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7544 (0.0036)
250	Inpatient Admission	LDA; max pool (300)	0.7287 (0.0022)
250	Inpatient Admission	LDA; max and mean pool (600)	0.7249 (0.0031)
250	Inpatient Admission	LDA; mean pool (300)	0.6865 (0.0025)

N	Task	Model	AUROC (SEM)
250	Cardiac dysrhythmias	GRU, max pool, PubMed init (200)	0.7267 (0.0026)
250	Cardiac dysrhythmias	GRU, max pool, random init (600)	0.7418 (0.0019)
250	Cardiac dysrhythmias	GRU, max pool, random init (300)	0.7434 (0.0019)
250	Cardiac dysrhythmias	GRU, max pool, PubMed init (500)	0.7238 (0.0018)
250	Cardiac dysrhythmias	Baseline - Bag of Words (15000)	0.6511 (0.012)
250	Cardiac dysrhythmias	MCEMJ; min, mean, max pool (600)	0.707 (0.0032)
250	Cardiac dysrhythmias	GloVe 300; min, mean, max pool (900)	0.7052 (0.0025)
250	Cardiac dysrhythmias	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7 (0.0109)
250	Cardiac dysrhythmias	GloVe 500; min, mean, max pool (1500)	0.6989 (0.0032)
250	Cardiac dysrhythmias	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7067 (0.0027)
250	Cardiac dysrhythmias	LDA; max pool (300)	0.6762 (0.0014)
250	Cardiac dysrhythmias	LDA; max and mean pool (600)	0.6693 (0.0014)
250	Cardiac dysrhythmias	LDA; mean pool (300)	0.6286 (0.0022)
250	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (200)	0.7437 (0.0047)
250	Diabetes mellitus w/o complic.	GRU, max pool, random init (600)	0.7579 (0.0024)
250	Diabetes mellitus w/o complic.	GRU, max pool, random init (300)	0.7541 (0.0017)
250	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (500)	0.7393 (0.0022)
250	Diabetes mellitus w/o complic.	Baseline - Bag of Words (15000)	0.6135 (0.0133)
250	Diabetes mellitus w/o complic.	MCEMJ; min, mean, max pool (600)	0.696 (0.0029)
250	Diabetes mellitus w/o complic.	GloVe 300; min, mean, max pool (900)	0.6775 (0.0047)
250	Diabetes mellitus w/o complic.	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6905 (0.004)
250	Diabetes mellitus w/o complic.	GloVe 500; min, mean, max pool (1500)	0.6776 (0.0098)
250	Diabetes mellitus w/o complic.	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6887 (0.004)
250	Diabetes mellitus w/o complic.	LDA; max pool (300)	0.6584 (0.0016)
250	Diabetes mellitus w/o complic.	LDA; max and mean pool (600)	0.6431 (0.0078)
250	Diabetes mellitus w/o complic.	LDA; mean pool (300)	0.5694 (0.0086)
250	Spondylosis (back pain)	GRU, max pool, PubMed init (200)	0.7396 (0.0028)
250	Spondylosis (back pain)	GRU, max pool, random init (600)	0.7462 (0.0029)
250	Spondylosis (back pain)	GRU, max pool, random init (300)	0.7426 (0.0028)
250	Spondylosis (back pain)	GRU, max pool, PubMed init (500)	0.7264 (0.0027)
250	Spondylosis (back pain)	Baseline - Bag of Words (15000)	0.651 (0.0035)
250	Spondylosis (back pain)	MCEMJ; min, mean, max pool (600)	0.6862 (0.0034)
250	Spondylosis (back pain)	GloVe 300; min, mean, max pool (900)	0.6838 (0.0036)
250	Spondylosis (back pain)	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6811 (0.0102)
250	Spondylosis (back pain)	GloVe 500; min, mean, max pool (1500)	0.6897 (0.0043)
250	Spondylosis (back pain)	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6969 (0.0036)
250	Spondylosis (back pain)	LDA; max pool (300)	0.6475 (0.0036)
250	Spondylosis (back pain)	LDA; max and mean pool (600)	0.6367 (0.0075)
250	Spondylosis (back pain)	LDA; mean pool (300)	0.5649 (0.008)

N	Task	Model	AUROC (SEM)
250	Disorders of lipid metabolism	GRU, max pool, PubMed init (200)	0.714 (0.0021)
250	Disorders of lipid metabolism	GRU, max pool, random init (600)	0.7204 (0.0033)
250	Disorders of lipid metabolism	GRU, max pool, random init (300)	0.7212 (0.0027)
250	Disorders of lipid metabolism	GRU, max pool, PubMed init (500)	0.7099 (0.0023)
250	Disorders of lipid metabolism	Baseline - Bag of Words (15000)	0.6057 (0.0042)
250	Disorders of lipid metabolism	MCEMJ; min, mean, max pool (600)	0.6522 (0.0038)
250	Disorders of lipid metabolism	GloVe 300; min, mean, max pool (900)	0.6506 (0.006)
250	Disorders of lipid metabolism	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6505 (0.0047)
250	Disorders of lipid metabolism	GloVe 500; min, mean, max pool (1500)	0.6555 (0.0045)
250	Disorders of lipid metabolism	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6576 (0.0044)
250	Disorders of lipid metabolism	LDA; max pool (300)	0.604 (0.0075)
250	Disorders of lipid metabolism	LDA; max and mean pool (600)	0.5942 (0.0092)
250	Disorders of lipid metabolism	LDA; mean pool (300)	0.5618 (0.0043)
250	Essential hypertension	GRU, max pool, PubMed init (200)	0.7373 (0.0023)
250	Essential hypertension	GRU, max pool, random init (600)	0.7399 (0.003)
250	Essential hypertension	GRU, max pool, random init (300)	0.7413 (0.0027)
250	Essential hypertension	GRU, max pool, PubMed init (500)	0.731 (0.0028)
250	Essential hypertension	Baseline - Bag of Words (15000)	0.5942 (0.0113)
250	Essential hypertension	MCEMJ; min, mean, max pool (600)	0.6749 (0.0027)
250	Essential hypertension	GloVe 300; min, mean, max pool (900)	0.668 (0.0031)
250	Essential hypertension	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6726 (0.0037)
250	Essential hypertension	GloVe 500; min, mean, max pool (1500)	0.6629 (0.0092)
250	Essential hypertension	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6776 (0.0037)
250	Essential hypertension	LDA; max pool (300)	0.6378 (0.0075)
250	Essential hypertension	LDA; max and mean pool (600)	0.6173 (0.0097)
250	Essential hypertension	LDA; mean pool (300)	0.5715 (0.0074)
250	Thyroid disorders	GRU, max pool, PubMed init (200)	0.682 (0.0048)
250	Thyroid disorders	GRU, max pool, random init (600)	0.7261 (0.003)
250	Thyroid disorders	GRU, max pool, random init (300)	0.7215 (0.0027)
250	Thyroid disorders	GRU, max pool, PubMed init (500)	0.6602 (0.0038)
250	Thyroid disorders	Baseline - Bag of Words (15000)	0.5773 (0.0092)
250	Thyroid disorders	MCEMJ; min, mean, max pool (600)	0.6631 (0.0092)
250	Thyroid disorders	GloVe 300; min, mean, max pool (900)	0.6421 (0.004)
250	Thyroid disorders	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6511 (0.0092)
250	Thyroid disorders	GloVe 500; min, mean, max pool (1500)	0.6506 (0.0047)
250	Thyroid disorders	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6584 (0.0038)
250	Thyroid disorders	LDA; max pool (300)	0.6048 (0.0012)
250	Thyroid disorders	LDA; max and mean pool (600)	0.5802 (0.0078)
250	Thyroid disorders	LDA; mean pool (300)	0.5415 (0.004)

N	Task	Model	AUROC (SEM)
500	Mortality	GRU, max pool, PubMed init (200)	0.8534 (0.0009)
500	Mortality	GRU, max pool, random init (600)	0.8659 (0.0009)
500	Mortality	GRU, max pool, random init (300)	0.866 (0.0007)
500	Mortality	GRU, max pool, PubMed init (500)	0.8493 (0.0008)
500	Mortality	Baseline - Bag of Words (15000)	0.851 (0.0013)
500	Mortality	MCEMJ; min, mean, max pool (600)	0.8273 (0.0018)
500	Mortality	GloVe 300; min, mean, max pool (900)	0.8336 (0.0018)
500	Mortality	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.8521 (0.0017)
500	Mortality	GloVe 500; min, mean, max pool (1500)	0.845 (0.0015)
500	Mortality	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.8556 (0.0014)
500	Mortality	LDA; max pool (300)	0.8261 (0.0019)
500	Mortality	LDA; max and mean pool (600)	0.8169 (0.0014)
500	Mortality	LDA; mean pool (300)	0.7711 (0.0024)
500	ER visit	GRU, max pool, PubMed init (200)	0.7577 (0.0016)
500	ER visit	GRU, max pool, random init (600)	0.7669 (0.0012)
500	ER visit	GRU, max pool, random init (300)	0.7565 (0.0024)
500	ER visit	GRU, max pool, PubMed init (500)	0.76 (0.0014)
500	ER visit	Baseline - Bag of Words (15000)	0.7401 (0.0127)
500	ER visit	MCEMJ; min, mean, max pool (600)	0.7588 (0.0027)
500	ER visit	GloVe 300; min, mean, max pool (900)	0.7676 (0.0012)
500	ER visit	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.771 (0.0013)
500	ER visit	GloVe 500; min, mean, max pool (1500)	0.7692 (0.0015)
500	ER visit	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7704 (0.0014)
500	ER visit	LDA; max pool (300)	0.6982 (0.0034)
500	ER visit	LDA; max and mean pool (600)	0.719 (0.0016)
500	ER visit	LDA; mean pool (300)	0.6942 (0.0024)
500	Inpatient Admission	GRU, max pool, PubMed init (200)	0.7786 (0.0021)
500	Inpatient Admission	GRU, max pool, random init (600)	0.7836 (0.0021)
500	Inpatient Admission	GRU, max pool, random init (300)	0.7878 (0.002)
500	Inpatient Admission	GRU, max pool, PubMed init (500)	0.7757 (0.0013)
500	Inpatient Admission	Baseline - Bag of Words (15000)	0.7608 (0.0015)
500	Inpatient Admission	MCEMJ; min, mean, max pool (600)	0.7632 (0.0016)
500	Inpatient Admission	GloVe 300; min, mean, max pool (900)	0.7695 (0.0021)
500	Inpatient Admission	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7774 (0.002)
500	Inpatient Admission	GloVe 500; min, mean, max pool (1500)	0.7744 (0.0017)
500	Inpatient Admission	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7782 (0.0017)
500	Inpatient Admission	LDA; max pool (300)	0.7423 (0.0027)
500	Inpatient Admission	LDA; max and mean pool (600)	0.7459 (0.0015)
500	Inpatient Admission	LDA; mean pool (300)	0.7126 (0.0024)

N	Task	Model	AUROC (SEM)
500	Cardiac dysrhythmias	GRU, max pool, PubMed init (200)	0.7336 (0.0025)
500	Cardiac dysrhythmias	GRU, max pool, random init (600)	0.7526 (0.0013)
500	Cardiac dysrhythmias	GRU, max pool, random init (300)	0.7493 (0.0014)
500	Cardiac dysrhythmias	GRU, max pool, PubMed init (500)	0.737 (0.0011)
500	Cardiac dysrhythmias	Baseline - Bag of Words (15000)	0.6797 (0.0138)
500	Cardiac dysrhythmias	MCEMJ; min, mean, max pool (600)	0.7215 (0.0023)
500	Cardiac dysrhythmias	GloVe 300; min, mean, max pool (900)	0.7184 (0.0028)
500	Cardiac dysrhythmias	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7272 (0.0023)
500	Cardiac dysrhythmias	GloVe 500; min, mean, max pool (1500)	0.719 (0.0022)
500	Cardiac dysrhythmias	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.726 (0.0019)
500	Cardiac dysrhythmias	LDA; max pool (300)	0.6819 (0.0012)
500	Cardiac dysrhythmias	LDA; max and mean pool (600)	0.6808 (0.0008)
500	Cardiac dysrhythmias	LDA; mean pool (300)	0.6422 (0.001)
500	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (200)	0.7557 (0.0014)
500	Diabetes mellitus w/o complic.	GRU, max pool, random init (600)	0.773 (0.002)
500	Diabetes mellitus w/o complic.	GRU, max pool, random init (300)	0.764 (0.0013)
500	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (500)	0.7537 (0.0018)
500	Diabetes mellitus w/o complic.	Baseline - Bag of Words (15000)	0.6697 (0.0133)
500	Diabetes mellitus w/o complic.	MCEMJ; min, mean, max pool (600)	0.7094 (0.0015)
500	Diabetes mellitus w/o complic.	GloVe 300; min, mean, max pool (900)	0.7049 (0.0026)
500	Diabetes mellitus w/o complic.	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7119 (0.0025)
500	Diabetes mellitus w/o complic.	GloVe 500; min, mean, max pool (1500)	0.7081 (0.0022)
500	Diabetes mellitus w/o complic.	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7152 (0.003)
500	Diabetes mellitus w/o complic.	LDA; max pool (300)	0.6677 (0.0014)
500	Diabetes mellitus w/o complic.	LDA; max and mean pool (600)	0.6561 (0.0085)
500	Diabetes mellitus w/o complic.	LDA; mean pool (300)	0.5949 (0.0098)
500	Spondylosis (back pain)	GRU, max pool, PubMed init (200)	0.7481 (0.002)
500	Spondylosis (back pain)	GRU, max pool, random init (600)	0.76 (0.0011)
500	Spondylosis (back pain)	GRU, max pool, random init (300)	0.7523 (0.0019)
500	Spondylosis (back pain)	GRU, max pool, PubMed init (500)	0.739 (0.0011)
500	Spondylosis (back pain)	Baseline - Bag of Words (15000)	0.6713 (0.0132)
500	Spondylosis (back pain)	MCEMJ; min, mean, max pool (600)	0.7052 (0.0019)
500	Spondylosis (back pain)	GloVe 300; min, mean, max pool (900)	0.7002 (0.0029)
500	Spondylosis (back pain)	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7078 (0.0035)
500	Spondylosis (back pain)	GloVe 500; min, mean, max pool (1500)	0.7086 (0.0021)
500	Spondylosis (back pain)	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7137 (0.0021)
500	Spondylosis (back pain)	LDA; max pool (300)	0.6686 (0.0018)
500	Spondylosis (back pain)	LDA; max and mean pool (600)	0.6655 (0.0015)
500	Spondylosis (back pain)	LDA; mean pool (300)	0.6046 (0.0038)

N	Task	Model	AUROC (SEM)
500	Disorders of lipid metabolism	GRU, max pool, PubMed init (200)	0.7284 (0.002)
500	Disorders of lipid metabolism	GRU, max pool, random init (600)	0.7389 (0.0016)
500	Disorders of lipid metabolism	GRU, max pool, random init (300)	0.739 (0.0013)
500	Disorders of lipid metabolism	GRU, max pool, PubMed init (500)	0.7275 (0.0017)
500	Disorders of lipid metabolism	Baseline - Bag of Words (15000)	0.6276 (0.0126)
500	Disorders of lipid metabolism	MCEMJ; min, mean, max pool (600)	0.6631 (0.0039)
500	Disorders of lipid metabolism	GloVe 300; min, mean, max pool (900)	0.6664 (0.0032)
500	Disorders of lipid metabolism	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6728 (0.0038)
500	Disorders of lipid metabolism	GloVe 500; min, mean, max pool (1500)	0.6698 (0.004)
500	Disorders of lipid metabolism	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.68 (0.0026)
500	Disorders of lipid metabolism	LDA; max pool (300)	0.6233 (0.0034)
500	Disorders of lipid metabolism	LDA; max and mean pool (600)	0.6199 (0.0066)
500	Disorders of lipid metabolism	LDA; mean pool (300)	0.5779 (0.0051)
500	Essential hypertension	GRU, max pool, PubMed init (200)	0.7517 (0.0031)
500	Essential hypertension	GRU, max pool, random init (600)	0.7592 (0.0019)
500	Essential hypertension	GRU, max pool, random init (300)	0.7547 (0.0022)
500	Essential hypertension	GRU, max pool, PubMed init (500)	0.7479 (0.0016)
500	Essential hypertension	Baseline - Bag of Words (15000)	0.66 (0.0019)
500	Essential hypertension	MCEMJ; min, mean, max pool (600)	0.6893 (0.0021)
500	Essential hypertension	GloVe 300; min, mean, max pool (900)	0.692 (0.0021)
500	Essential hypertension	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6967 (0.0021)
500	Essential hypertension	GloVe 500; min, mean, max pool (1500)	0.6911 (0.0022)
500	Essential hypertension	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6958 (0.0024)
500	Essential hypertension	LDA; max pool (300)	0.6456 (0.0082)
500	Essential hypertension	LDA; max and mean pool (600)	0.6533 (0.002)
500	Essential hypertension	LDA; mean pool (300)	0.575 (0.0094)
500	Thyroid disorders	GRU, max pool, PubMed init (200)	0.7046 (0.0019)
500	Thyroid disorders	GRU, max pool, random init (600)	0.7447 (0.0028)
500	Thyroid disorders	GRU, max pool, random init (300)	0.7344 (0.003)
500	Thyroid disorders	GRU, max pool, PubMed init (500)	0.6779 (0.0024)
500	Thyroid disorders	Baseline - Bag of Words (15000)	0.6319 (0.002)
500	Thyroid disorders	MCEMJ; min, mean, max pool (600)	0.6958 (0.0023)
500	Thyroid disorders	GloVe 300; min, mean, max pool (900)	0.6663 (0.0019)
500	Thyroid disorders	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6923 (0.0032)
500	Thyroid disorders	GloVe 500; min, mean, max pool (1500)	0.6685 (0.0043)
500	Thyroid disorders	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.6914 (0.003)
500	Thyroid disorders	LDA; max pool (300)	0.6128 (0.0015)
500	Thyroid disorders	LDA; max and mean pool (600)	0.6011 (0.008)
500	Thyroid disorders	LDA; mean pool (300)	0.544 (0.0063)

N	Task	Model	AUROC (SEM)
1000	Mortality	GRU, max pool, PubMed init (200)	0.8588 (0.0007)
1000	Mortality	GRU, max pool, random init (600)	0.8735 (0.0007)
1000	Mortality	GRU, max pool, random init (300)	0.8728 (0.0005)
1000	Mortality	GRU, max pool, PubMed init (500)	0.8566 (0.0008)
1000	Mortality	Baseline - Bag of Words (15000)	0.8777 (0.0006)
1000	Mortality	MCEMJ; min, mean, max pool (600)	0.8641 (0.0012)
1000	Mortality	GloVe 300; min, mean, max pool (900)	0.8408 (0.0015)
1000	Mortality	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.8578 (0.0014)
1000	Mortality	GloVe 500; min, mean, max pool (1500)	0.8532 (0.0012)
1000	Mortality	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.8632 (0.0011)
1000	Mortality	LDA; max pool (300)	0.8432 (0.0013)
1000	Mortality	LDA; max and mean pool (600)	0.8382 (0.0013)
1000	Mortality	LDA; mean pool (300)	0.7933 (0.0015)
1000	ER visit	GRU, max pool, PubMed init (200)	0.7663 (0.001)
1000	ER visit	GRU, max pool, random init (600)	0.7763 (0.0009)
1000	ER visit	GRU, max pool, random init (300)	0.7678 (0.0009)
1000	ER visit	GRU, max pool, PubMed init (500)	0.7649 (0.0011)
1000	ER visit	Baseline - Bag of Words (15000)	0.7678 (0.0009)
1000	ER visit	MCEMJ; min, mean, max pool (600)	0.7719 (0.0014)
1000	ER visit	GloVe 300; min, mean, max pool (900)	0.7765 (0.0007)
1000	ER visit	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7793 (0.0008)
1000	ER visit	GloVe 500; min, mean, max pool (1500)	0.7783 (0.0008)
1000	ER visit	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7803 (0.001)
1000	ER visit	LDA; max pool (300)	0.7211 (0.0015)
1000	ER visit	LDA; max and mean pool (600)	0.7336 (0.0015)
1000	ER visit	LDA; mean pool (300)	0.7148 (0.002)
1000	Inpatient Admission	GRU, max pool, PubMed init (200)	0.7896 (0.0009)
1000	Inpatient Admission	GRU, max pool, random init (600)	0.7945 (0.0008)
1000	Inpatient Admission	GRU, max pool, random init (300)	0.7952 (0.0009)
1000	Inpatient Admission	GRU, max pool, PubMed init (500)	0.7831 (0.0008)
1000	Inpatient Admission	Baseline - Bag of Words (15000)	0.7812 (0.0008)
1000	Inpatient Admission	MCEMJ; min, mean, max pool (600)	0.7804 (0.0014)
1000	Inpatient Admission	GloVe 300; min, mean, max pool (900)	0.7798 (0.0012)
1000	Inpatient Admission	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7854 (0.0009)
1000	Inpatient Admission	GloVe 500; min, mean, max pool (1500)	0.7838 (0.0011)
1000	Inpatient Admission	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7874 (0.0015)
1000	Inpatient Admission	LDA; max pool (300)	0.7589 (0.0017)
1000	Inpatient Admission	LDA; max and mean pool (600)	0.7603 (0.0016)
1000	Inpatient Admission	LDA; mean pool (300)	0.7338 (0.0016)

N	Task	Model	AUROC (SEM)
1000	Cardiac dysrhythmias	GRU, max pool, PubMed init (200)	0.7435 (0.0012)
1000	Cardiac dysrhythmias	GRU, max pool, random init (600)	0.7574 (0.0009)
1000	Cardiac dysrhythmias	GRU, max pool, random init (300)	0.7549 (0.001)
1000	Cardiac dysrhythmias	GRU, max pool, PubMed init (500)	0.7425 (0.002)
1000	Cardiac dysrhythmias	Baseline - Bag of Words (15000)	0.7272 (0.001)
1000	Cardiac dysrhythmias	MCEMJ; min, mean, max pool (600)	0.7401 (0.001)
1000	Cardiac dysrhythmias	GloVe 300; min, mean, max pool (900)	0.7325 (0.0011)
1000	Cardiac dysrhythmias	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7394 (0.0016)
1000	Cardiac dysrhythmias	GloVe 500; min, mean, max pool (1500)	0.7338 (0.001)
1000	Cardiac dysrhythmias	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7408 (0.0012)
1000	Cardiac dysrhythmias	LDA; max pool (300)	0.6941 (0.0016)
1000	Cardiac dysrhythmias	LDA; max and mean pool (600)	0.6923 (0.0015)
1000	Cardiac dysrhythmias	LDA; mean pool (300)	0.6587 (0.0017)
1000	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (200)	0.7645 (0.0015)
1000	Diabetes mellitus w/o complic.	GRU, max pool, random init (600)	0.7885 (0.0009)
1000	Diabetes mellitus w/o complic.	GRU, max pool, random init (300)	0.7734 (0.0013)
1000	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (500)	0.7587 (0.0012)
1000	Diabetes mellitus w/o complic.	Baseline - Bag of Words (15000)	0.7226 (0.0018)
1000	Diabetes mellitus w/o complic.	MCEMJ; min, mean, max pool (600)	0.7384 (0.0018)
1000	Diabetes mellitus w/o complic.	GloVe 300; min, mean, max pool (900)	0.7169 (0.0015)
1000	Diabetes mellitus w/o complic.	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7341 (0.0018)
1000	Diabetes mellitus w/o complic.	GloVe 500; min, mean, max pool (1500)	0.7256 (0.0019)
1000	Diabetes mellitus w/o complic.	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7384 (0.0019)
1000	Diabetes mellitus w/o complic.	LDA; max pool (300)	0.6806 (0.0018)
1000	Diabetes mellitus w/o complic.	LDA; max and mean pool (600)	0.6834 (0.0018)
1000	Diabetes mellitus w/o complic.	LDA; mean pool (300)	0.6325 (0.0019)
1000	Spondylosis (back pain)	GRU, max pool, PubMed init (200)	0.7556 (0.0013)
1000	Spondylosis (back pain)	GRU, max pool, random init (600)	0.7667 (0.0009)
1000	Spondylosis (back pain)	GRU, max pool, random init (300)	0.7599 (0.0009)
1000	Spondylosis (back pain)	GRU, max pool, PubMed init (500)	0.7469 (0.0011)
1000	Spondylosis (back pain)	Baseline - Bag of Words (15000)	0.7115 (0.0016)
1000	Spondylosis (back pain)	MCEMJ; min, mean, max pool (600)	0.7252 (0.001)
1000	Spondylosis (back pain)	GloVe 300; min, mean, max pool (900)	0.7178 (0.0015)
1000	Spondylosis (back pain)	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.729 (0.0012)
1000	Spondylosis (back pain)	GloVe 500; min, mean, max pool (1500)	0.7229 (0.0015)
1000	Spondylosis (back pain)	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7283 (0.0016)
1000	Spondylosis (back pain)	LDA; max pool (300)	0.6802 (0.0018)
1000	Spondylosis (back pain)	LDA; max and mean pool (600)	0.6827 (0.0017)
1000	Spondylosis (back pain)	LDA; mean pool (300)	0.6281 (0.0025)

N	Task	Model	AUROC (SEM)
1000	Disorders of lipid metabolism	GRU, max pool, PubMed init (200)	0.7443 (0.0015)
1000	Disorders of lipid metabolism	GRU, max pool, random init (600)	0.7509 (0.0011)
1000	Disorders of lipid metabolism	GRU, max pool, random init (300)	0.7481 (0.0013)
1000	Disorders of lipid metabolism	GRU, max pool, PubMed init (500)	0.7331 (0.0016)
1000	Disorders of lipid metabolism	Baseline - Bag of Words (15000)	0.6787 (0.0022)
1000	Disorders of lipid metabolism	MCEMJ; min, mean, max pool (600)	0.6954 (0.0016)
1000	Disorders of lipid metabolism	GloVe 300; min, mean, max pool (900)	0.6905 (0.0018)
1000	Disorders of lipid metabolism	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.6989 (0.0017)
1000	Disorders of lipid metabolism	GloVe 500; min, mean, max pool (1500)	0.6904 (0.0023)
1000	Disorders of lipid metabolism	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7014 (0.0019)
1000	Disorders of lipid metabolism	LDA; max pool (300)	0.6375 (0.002)
1000	Disorders of lipid metabolism	LDA; max and mean pool (600)	0.6442 (0.002)
1000	Disorders of lipid metabolism	LDA; mean pool (300)	0.6001 (0.0056)
1000	Essential hypertension	GRU, max pool, PubMed init (200)	0.7659 (0.001)
1000	Essential hypertension	GRU, max pool, random init (600)	0.7741 (0.0017)
1000	Essential hypertension	GRU, max pool, random init (300)	0.7662 (0.0008)
1000	Essential hypertension	GRU, max pool, PubMed init (500)	0.7637 (0.0011)
1000	Essential hypertension	Baseline - Bag of Words (15000)	0.6902 (0.0019)
1000	Essential hypertension	MCEMJ; min, mean, max pool (600)	0.7086 (0.0014)
1000	Essential hypertension	GloVe 300; min, mean, max pool (900)	0.7045 (0.0016)
1000	Essential hypertension	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7104 (0.0013)
1000	Essential hypertension	GloVe 500; min, mean, max pool (1500)	0.7057 (0.0019)
1000	Essential hypertension	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7141 (0.0016)
1000	Essential hypertension	LDA; max pool (300)	0.6681 (0.0016)
1000	Essential hypertension	LDA; max and mean pool (600)	0.6703 (0.0016)
1000	Essential hypertension	LDA; mean pool (300)	0.6133 (0.0062)
1000	Thyroid disorders	GRU, max pool, PubMed init (200)	0.7148 (0.0017)
1000	Thyroid disorders	GRU, max pool, random init (600)	0.7667 (0.0023)
1000	Thyroid disorders	GRU, max pool, random init (300)	0.7532 (0.0019)
1000	Thyroid disorders	GRU, max pool, PubMed init (500)	0.6998 (0.0015)
1000	Thyroid disorders	Baseline - Bag of Words (15000)	0.6687 (0.0021)
1000	Thyroid disorders	MCEMJ; min, mean, max pool (600)	0.7268 (0.0016)
1000	Thyroid disorders	GloVe 300; min, mean, max pool (900)	0.6823 (0.0019)
1000	Thyroid disorders	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7128 (0.0019)
1000	Thyroid disorders	GloVe 500; min, mean, max pool (1500)	0.6945 (0.0018)
1000	Thyroid disorders	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7134 (0.0015)
1000	Thyroid disorders	LDA; max pool (300)	0.6219 (0.0019)
1000	Thyroid disorders	LDA; max and mean pool (600)	0.6293 (0.0023)
1000	Thyroid disorders	LDA; mean pool (300)	0.5721 (0.0036)

N	Task	Model	AUROC (SEM)
2000	Mortality	GRU, max pool, PubMed init (200)	0.8624 (0.0004)
2000	Mortality	GRU, max pool, random init (600)	0.8792 (0.0004)
2000	Mortality	GRU, max pool, random init (300)	0.877 (0.0004)
2000	Mortality	GRU, max pool, PubMed init (500)	0.8612 (0.0005)
2000	Mortality	Baseline - Bag of Words (15000)	0.894 (0.001)
2000	Mortality	MCEMJ; min, mean, max pool (600)	0.8744 (0.0006)
2000	Mortality	GloVe 300; min, mean, max pool (900)	0.8839 (0.0006)
2000	Mortality	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.8934 (0.0006)
2000	Mortality	GloVe 500; min, mean, max pool (1500)	0.8914 (0.0009)
2000	Mortality	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.8661 (0.001)
2000	Mortality	LDA; max pool (300)	0.8486 (0.0007)
2000	Mortality	LDA; max and mean pool (600)	0.8483 (0.0004)
2000	Mortality	LDA; mean pool (300)	0.8089 (0.0009)
2000	ER visit	GRU, max pool, PubMed init (200)	0.7737 (0.0005)
2000	ER visit	GRU, max pool, random init (600)	0.7817 (0.0005)
2000	ER visit	GRU, max pool, random init (300)	0.7757 (0.0008)
2000	ER visit	GRU, max pool, PubMed init (500)	0.7723 (0.0005)
2000	ER visit	Baseline - Bag of Words (15000)	0.7759 (0.0007)
2000	ER visit	MCEMJ; min, mean, max pool (600)	0.7805 (0.0009)
2000	ER visit	GloVe 300; min, mean, max pool (900)	0.782 (0.0006)
2000	ER visit	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7867 (0.0007)
2000	ER visit	GloVe 500; min, mean, max pool (1500)	0.7848 (0.0008)
2000	ER visit	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.787 (0.0004)
2000	ER visit	LDA; max pool (300)	0.7368 (0.0014)
2000	ER visit	LDA; max and mean pool (600)	0.7497 (0.0011)
2000	ER visit	LDA; mean pool (300)	0.7334 (0.0011)
2000	Inpatient Admission	GRU, max pool, PubMed init (200)	0.7938 (0.0006)
2000	Inpatient Admission	GRU, max pool, random init (600)	0.8034 (0.0007)
2000	Inpatient Admission	GRU, max pool, random init (300)	0.8008 (0.0008)
2000	Inpatient Admission	GRU, max pool, PubMed init (500)	0.7882 (0.0009)
2000	Inpatient Admission	Baseline - Bag of Words (15000)	0.7944 (0.0009)
2000	Inpatient Admission	MCEMJ; min, mean, max pool (600)	0.7906 (0.0006)
2000	Inpatient Admission	GloVe 300; min, mean, max pool (900)	0.7942 (0.0007)
2000	Inpatient Admission	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7996 (0.0007)
2000	Inpatient Admission	GloVe 500; min, mean, max pool (1500)	0.7974 (0.0008)
2000	Inpatient Admission	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7944 (0.0008)
2000	Inpatient Admission	LDA; max pool (300)	0.7701 (0.0012)
2000	Inpatient Admission	LDA; max and mean pool (600)	0.7748 (0.0013)
2000	Inpatient Admission	LDA; mean pool (300)	0.7503 (0.0011)

N	Task	Model	AUROC (SEM)
2000	Cardiac dysrhythmias	GRU, max pool, PubMed init (200)	0.7506 (0.001)
2000	Cardiac dysrhythmias	GRU, max pool, random init (600)	0.7629 (0.0009)
2000	Cardiac dysrhythmias	GRU, max pool, random init (300)	0.7638 (0.0008)
2000	Cardiac dysrhythmias	GRU, max pool, PubMed init (500)	0.7499 (0.0009)
2000	Cardiac dysrhythmias	Baseline - Bag of Words (15000)	0.7418 (0.001)
2000	Cardiac dysrhythmias	MCEMJ; min, mean, max pool (600)	0.7495 (0.0006)
2000	Cardiac dysrhythmias	GloVe 300; min, mean, max pool (900)	0.7431 (0.0007)
2000	Cardiac dysrhythmias	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7544 (0.0007)
2000	Cardiac dysrhythmias	GloVe 500; min, mean, max pool (1500)	0.7452 (0.0007)
2000	Cardiac dysrhythmias	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7505 (0.001)
2000	Cardiac dysrhythmias	LDA; max pool (300)	0.6993 (0.0009)
2000	Cardiac dysrhythmias	LDA; max and mean pool (600)	0.7025 (0.0011)
2000	Cardiac dysrhythmias	LDA; mean pool (300)	0.6734 (0.0013)
2000	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (200)	0.7741 (0.0007)
2000	Diabetes mellitus w/o complic.	GRU, max pool, random init (600)	0.7994 (0.0008)
2000	Diabetes mellitus w/o complic.	GRU, max pool, random init (300)	0.7829 (0.0006)
2000	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (500)	0.7672 (0.0007)
2000	Diabetes mellitus w/o complic.	Baseline - Bag of Words (15000)	0.7499 (0.002)
2000	Diabetes mellitus w/o complic.	MCEMJ; min, mean, max pool (600)	0.7592 (0.0011)
2000	Diabetes mellitus w/o complic.	GloVe 300; min, mean, max pool (900)	0.7408 (0.0011)
2000	Diabetes mellitus w/o complic.	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7625 (0.0009)
2000	Diabetes mellitus w/o complic.	GloVe 500; min, mean, max pool (1500)	0.7452 (0.0011)
2000	Diabetes mellitus w/o complic.	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7481 (0.0017)
2000	Diabetes mellitus w/o complic.	LDA; max pool (300)	0.6921 (0.0014)
2000	Diabetes mellitus w/o complic.	LDA; max and mean pool (600)	0.6984 (0.0013)
2000	Diabetes mellitus w/o complic.	LDA; mean pool (300)	0.6464 (0.0012)
2000	Spondylosis (back pain)	GRU, max pool, PubMed init (200)	0.7609 (0.0008)
2000	Spondylosis (back pain)	GRU, max pool, random init (600)	0.7732 (0.0005)
2000	Spondylosis (back pain)	GRU, max pool, random init (300)	0.7662 (0.0008)
2000	Spondylosis (back pain)	GRU, max pool, PubMed init (500)	0.7519 (0.0004)
2000	Spondylosis (back pain)	Baseline - Bag of Words (15000)	0.7286 (0.0009)
2000	Spondylosis (back pain)	MCEMJ; min, mean, max pool (600)	0.7342 (0.0011)
2000	Spondylosis (back pain)	GloVe 300; min, mean, max pool (900)	0.7393 (0.0009)
2000	Spondylosis (back pain)	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7453 (0.0009)
2000	Spondylosis (back pain)	GloVe 500; min, mean, max pool (1500)	0.7409 (0.0011)
2000	Spondylosis (back pain)	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7403 (0.0009)
2000	Spondylosis (back pain)	LDA; max pool (300)	0.6971 (0.0007)
2000	Spondylosis (back pain)	LDA; max and mean pool (600)	0.7027 (0.0009)
2000	Spondylosis (back pain)	LDA; mean pool (300)	0.6482 (0.0011)

N	Task	Model	AUROC (SEM)
2000	Disorders of lipid metabolism	GRU, max pool, PubMed init (200)	0.7523 (0.001)
2000	Disorders of lipid metabolism	GRU, max pool, random init (600)	0.7624 (0.0009)
2000	Disorders of lipid metabolism	GRU, max pool, random init (300)	0.7596 (0.0009)
2000	Disorders of lipid metabolism	GRU, max pool, PubMed init (500)	0.7428 (0.001)
2000	Disorders of lipid metabolism	Baseline - Bag of Words (15000)	0.7106 (0.0015)
2000	Disorders of lipid metabolism	MCEMJ; min, mean, max pool (600)	0.7149 (0.0011)
2000	Disorders of lipid metabolism	GloVe 300; min, mean, max pool (900)	0.7086 (0.0013)
2000	Disorders of lipid metabolism	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7221 (0.0011)
2000	Disorders of lipid metabolism	GloVe 500; min, mean, max pool (1500)	0.7107 (0.0012)
2000	Disorders of lipid metabolism	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7163 (0.0011)
2000	Disorders of lipid metabolism	LDA; max pool (300)	0.6545 (0.0013)
2000	Disorders of lipid metabolism	LDA; max and mean pool (600)	0.6628 (0.0017)
2000	Disorders of lipid metabolism	LDA; mean pool (300)	0.6219 (0.0014)
2000	Essential hypertension	GRU, max pool, PubMed init (200)	0.7758 (0.0008)
2000	Essential hypertension	GRU, max pool, random init (600)	0.7878 (0.0009)
2000	Essential hypertension	GRU, max pool, random init (300)	0.7759 (0.0011)
2000	Essential hypertension	GRU, max pool, PubMed init (500)	0.7737 (0.0005)
2000	Essential hypertension	Baseline - Bag of Words (15000)	0.7176 (0.0013)
2000	Essential hypertension	MCEMJ; min, mean, max pool (600)	0.7272 (0.0011)
2000	Essential hypertension	GloVe 300; min, mean, max pool (900)	0.7204 (0.0011)
2000	Essential hypertension	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7328 (0.0008)
2000	Essential hypertension	GloVe 500; min, mean, max pool (1500)	0.7213 (0.0008)
2000	Essential hypertension	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7259 (0.0012)
2000	Essential hypertension	LDA; max pool (300)	0.678 (0.0012)
2000	Essential hypertension	LDA; max and mean pool (600)	0.6824 (0.0014)
2000	Essential hypertension	LDA; mean pool (300)	0.6355 (0.0014)
2000	Thyroid disorders	GRU, max pool, PubMed init (200)	0.7281 (0.001)
2000	Thyroid disorders	GRU, max pool, random init (600)	0.7819 (0.0012)
2000	Thyroid disorders	GRU, max pool, random init (300)	0.769 (0.001)
2000	Thyroid disorders	GRU, max pool, PubMed init (500)	0.713 (0.0011)
2000	Thyroid disorders	Baseline - Bag of Words (15000)	0.7008 (0.0022)
2000	Thyroid disorders	MCEMJ; min, mean, max pool (600)	0.7482 (0.0013)
2000	Thyroid disorders	GloVe 300; min, mean, max pool (900)	0.7043 (0.0008)
2000	Thyroid disorders	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7401 (0.0012)
2000	Thyroid disorders	GloVe 500; min, mean, max pool (1500)	0.7135 (0.001)
2000	Thyroid disorders	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7264 (0.0015)
2000	Thyroid disorders	LDA; max pool (300)	0.6343 (0.0015)
2000	Thyroid disorders	LDA; max and mean pool (600)	0.6451 (0.0014)
2000	Thyroid disorders	LDA; mean pool (300)	0.5896 (0.0049)

N	Task	Model	AUROC (SEM)
4000	Mortality	GRU, max pool, PubMed init (200)	0.8661 (0.0002)
4000	Mortality	GRU, max pool, random init (600)	0.8831 (0.0003)
4000	Mortality	GRU, max pool, random init (300)	0.8801 (0.0003)
4000	Mortality	GRU, max pool, PubMed init (500)	0.866 (0.0003)
4000	Mortality	Baseline - Bag of Words (15000)	0.907 (0.0005)
4000	Mortality	MCEMJ; min, mean, max pool (600)	0.882 (0.0005)
4000	Mortality	GloVe 300; min, mean, max pool (900)	0.8921 (0.0006)
4000	Mortality	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.9015 (0.0006)
4000	Mortality	GloVe 500; min, mean, max pool (1500)	0.9002 (0.0006)
4000	Mortality	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.9078 (0.0006)
4000	Mortality	LDA; max pool (300)	0.8558 (0.0007)
4000	Mortality	LDA; max and mean pool (600)	0.8584 (0.0006)
4000	Mortality	LDA; mean pool (300)	0.8189 (0.0006)
4000	ER visit	GRU, max pool, PubMed init (200)	0.7763 (0.0005)
4000	ER visit	GRU, max pool, random init (600)	0.7858 (0.0005)
4000	ER visit	GRU, max pool, random init (300)	0.7791 (0.0005)
4000	ER visit	GRU, max pool, PubMed init (500)	0.7761 (0.0003)
4000	ER visit	Baseline - Bag of Words (15000)	0.7836 (0.0006)
4000	ER visit	MCEMJ; min, mean, max pool (600)	0.7871 (0.0005)
4000	ER visit	GloVe 300; min, mean, max pool (900)	0.7873 (0.0007)
4000	ER visit	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7927 (0.0006)
4000	ER visit	GloVe 500; min, mean, max pool (1500)	0.7912 (0.0006)
4000	ER visit	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7947 (0.0006)
4000	ER visit	LDA; max pool (300)	0.7443 (0.0006)
4000	ER visit	LDA; max and mean pool (600)	0.7594 (0.0005)
4000	ER visit	LDA; mean pool (300)	0.7414 (0.0008)
4000	Inpatient Admission	GRU, max pool, PubMed init (200)	0.7995 (0.0004)
4000	Inpatient Admission	GRU, max pool, random init (600)	0.8111 (0.0005)
4000	Inpatient Admission	GRU, max pool, random init (300)	0.8069 (0.0004)
4000	Inpatient Admission	GRU, max pool, PubMed init (500)	0.7939 (0.0003)
4000	Inpatient Admission	Baseline - Bag of Words (15000)	0.8062 (0.0006)
4000	Inpatient Admission	MCEMJ; min, mean, max pool (600)	0.7989 (0.0005)
4000	Inpatient Admission	GloVe 300; min, mean, max pool (900)	0.8027 (0.0005)
4000	Inpatient Admission	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.8085 (0.0005)
4000	Inpatient Admission	GloVe 500; min, mean, max pool (1500)	0.8053 (0.0005)
4000	Inpatient Admission	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.8099 (0.0005)
4000	Inpatient Admission	LDA; max pool (300)	0.7801 (0.0006)
4000	Inpatient Admission	LDA; max and mean pool (600)	0.7864 (0.0006)
4000	Inpatient Admission	LDA; mean pool (300)	0.7615 (0.0005)

N	Task	Model	AUROC (SEM)
4000	Cardiac dysrhythmias	GRU, max pool, PubMed init (200)	0.7559 (0.0005)
4000	Cardiac dysrhythmias	GRU, max pool, random init (600)	0.7718 (0.0005)
4000	Cardiac dysrhythmias	GRU, max pool, random init (300)	0.7675 (0.0005)
4000	Cardiac dysrhythmias	GRU, max pool, PubMed init (500)	0.7536 (0.0005)
4000	Cardiac dysrhythmias	Baseline - Bag of Words (15000)	0.756 (0.0009)
4000	Cardiac dysrhythmias	MCEMJ; min, mean, max pool (600)	0.7577 (0.0007)
4000	Cardiac dysrhythmias	GloVe 300; min, mean, max pool (900)	0.7523 (0.0006)
4000	Cardiac dysrhythmias	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7617 (0.0007)
4000	Cardiac dysrhythmias	GloVe 500; min, mean, max pool (1500)	0.7549 (0.0005)
4000	Cardiac dysrhythmias	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7629 (0.0005)
4000	Cardiac dysrhythmias	LDA; max pool (300)	0.7086 (0.001)
4000	Cardiac dysrhythmias	LDA; max and mean pool (600)	0.7135 (0.0009)
4000	Cardiac dysrhythmias	LDA; mean pool (300)	0.6808 (0.0008)
4000	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (200)	0.7816 (0.0006)
4000	Diabetes mellitus w/o complic.	GRU, max pool, random init (600)	0.8076 (0.0007)
4000	Diabetes mellitus w/o complic.	GRU, max pool, random init (300)	0.7909 (0.0004)
4000	Diabetes mellitus w/o complic.	GRU, max pool, PubMed init (500)	0.7746 (0.0006)
4000	Diabetes mellitus w/o complic.	Baseline - Bag of Words (15000)	0.7741 (0.0011)
4000	Diabetes mellitus w/o complic.	MCEMJ; min, mean, max pool (600)	0.775 (0.0009)
4000	Diabetes mellitus w/o complic.	GloVe 300; min, mean, max pool (900)	0.7557 (0.0009)
4000	Diabetes mellitus w/o complic.	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7796 (0.0008)
4000	Diabetes mellitus w/o complic.	GloVe 500; min, mean, max pool (1500)	0.7638 (0.0008)
4000	Diabetes mellitus w/o complic.	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7814 (0.0008)
4000	Diabetes mellitus w/o complic.	LDA; max pool (300)	0.7037 (0.0009)
4000	Diabetes mellitus w/o complic.	LDA; max and mean pool (600)	0.7129 (0.001)
4000	Diabetes mellitus w/o complic.	LDA; mean pool (300)	0.6578 (0.0014)
4000	Spondylosis (back pain)	GRU, max pool, PubMed init (200)	0.7644 (0.0003)
4000	Spondylosis (back pain)	GRU, max pool, random init (600)	0.7776 (0.0003)
4000	Spondylosis (back pain)	GRU, max pool, random init (300)	0.7708 (0.0003)
4000	Spondylosis (back pain)	GRU, max pool, PubMed init (500)	0.7556 (0.0004)
4000	Spondylosis (back pain)	Baseline - Bag of Words (15000)	0.7429 (0.0008)
4000	Spondylosis (back pain)	MCEMJ; min, mean, max pool (600)	0.7442 (0.0007)
4000	Spondylosis (back pain)	GloVe 300; min, mean, max pool (900)	0.749 (0.0008)
4000	Spondylosis (back pain)	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7544 (0.0005)
4000	Spondylosis (back pain)	GloVe 500; min, mean, max pool (1500)	0.753 (0.0007)
4000	Spondylosis (back pain)	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7577 (0.0008)
4000	Spondylosis (back pain)	LDA; max pool (300)	0.7046 (0.0006)
4000	Spondylosis (back pain)	LDA; max and mean pool (600)	0.7127 (0.0006)
4000	Spondylosis (back pain)	LDA; mean pool (300)	0.6624 (0.0011)

N	Task	Model	AUROC (SEM)
4000	Disorders of lipid metabolism	GRU, max pool, PubMed init (200)	0.7591 (0.0006)
4000	Disorders of lipid metabolism	GRU, max pool, random init (600)	0.7698 (0.0004)
4000	Disorders of lipid metabolism	GRU, max pool, random init (300)	0.7659 (0.0007)
4000	Disorders of lipid metabolism	GRU, max pool, PubMed init (500)	0.7491 (0.0006)
4000	Disorders of lipid metabolism	Baseline - Bag of Words (15000)	0.7365 (0.001)
4000	Disorders of lipid metabolism	MCEMJ; min, mean, max pool (600)	0.7294 (0.0008)
4000	Disorders of lipid metabolism	GloVe 300; min, mean, max pool (900)	0.7227 (0.0006)
4000	Disorders of lipid metabolism	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.739 (0.0007)
4000	Disorders of lipid metabolism	GloVe 500; min, mean, max pool (1500)	0.7262 (0.0007)
4000	Disorders of lipid metabolism	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7381 (0.0008)
4000	Disorders of lipid metabolism	LDA; max pool (300)	0.6674 (0.0014)
4000	Disorders of lipid metabolism	LDA; max and mean pool (600)	0.6789 (0.0013)
4000	Disorders of lipid metabolism	LDA; mean pool (300)	0.6388 (0.001)
4000	Essential hypertension	GRU, max pool, PubMed init (200)	0.7825 (0.0005)
4000	Essential hypertension	GRU, max pool, random init (600)	0.7961 (0.0006)
4000	Essential hypertension	GRU, max pool, random init (300)	0.7838 (0.0004)
4000	Essential hypertension	GRU, max pool, PubMed init (500)	0.7815 (0.0006)
4000	Essential hypertension	Baseline - Bag of Words (15000)	0.7432 (0.0011)
4000	Essential hypertension	MCEMJ; min, mean, max pool (600)	0.7364 (0.0008)
4000	Essential hypertension	GloVe 300; min, mean, max pool (900)	0.7325 (0.0007)
4000	Essential hypertension	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7466 (0.0007)
4000	Essential hypertension	GloVe 500; min, mean, max pool (1500)	0.7355 (0.0008)
4000	Essential hypertension	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7462 (0.0008)
4000	Essential hypertension	LDA; max pool (300)	0.6913 (0.0007)
4000	Essential hypertension	LDA; max and mean pool (600)	0.6988 (0.0009)
4000	Essential hypertension	LDA; mean pool (300)	0.6479 (0.0012)
4000	Thyroid disorders	GRU, max pool, PubMed init (200)	0.7392 (0.0006)
4000	Thyroid disorders	GRU, max pool, random init (600)	0.7934 (0.0008)
4000	Thyroid disorders	GRU, max pool, random init (300)	0.7773 (0.0007)
4000	Thyroid disorders	GRU, max pool, PubMed init (500)	0.7234 (0.0006)
4000	Thyroid disorders	Baseline - Bag of Words (15000)	0.7301 (0.0021)
4000	Thyroid disorders	MCEMJ; min, mean, max pool (600)	0.7641 (0.001)
4000	Thyroid disorders	GloVe 300; min, mean, max pool (900)	0.7169 (0.0008)
4000	Thyroid disorders	GloVe 300 + MCEMJ; min, mean, max pool (1500)	0.7592 (0.0008)
4000	Thyroid disorders	GloVe 500; min, mean, max pool (1500)	0.7281 (0.0006)
4000	Thyroid disorders	GloVe 500 + MCEMJ; min, mean, max pool (2100)	0.7585 (0.0008)
4000	Thyroid disorders	LDA; max pool (300)	0.6455 (0.0012)
4000	Thyroid disorders	LDA; max and mean pool (600)	0.6585 (0.001)
4000	Thyroid disorders	LDA; mean pool (300)	0.6061 (0.0013)