Empirical Findings on the Role of Structured Data, Unstructured Data, and their Combination for Automatic Clinical Phenotyping

Asher Moldwin, Dina Demner-Fushman, MD, PhD, Travis R. Goodwin, PhD
U.S. National Library of Medicine, Bethesda, MD, USA

Abstract

The objective of this study is to explore the role of structured and unstructured data for clinical phenotyping by determining which types of clinical phenotypes are best identified using unstructured data (e.g., clinical notes), structured data (e.g., laboratory values, vital signs), or their combination across 172 clinical phenotypes. Specifically, we used laboratory and chart measurements as well as clinical notes from the MIMIC-III critical care database and trained an LSTM using features extracted from each type of data to determine which categories of phenotypes were best identified by structured data, unstructured data, or both. We observed that textual features on their own outperformed structured features for 145 (84%) of phenotypes, and that Doc2Vec was the most effective representation of unstructured data for all phenotypes. When evaluating the impact of adding textual features to systems previously relying only on structured features, we found a statistically significant ($p < 0.05$) increase in phenotyping performance for 51 phenotypes (primarily involving the circulatory system, injury, and poisoning), one phenotype for which textual features degraded performance (diabetes without complications), and no statistically significant change in performance with the remaining 120 phenotypes. We provide analysis on which phenotypes are best identified by each type of data and guidance on which data sources to consider for future research on phenotype identification.

Introduction

In recent years the use of Electronic Health Records (EHRs) has become standard practice across hospitals in the United States,[1] with medical professionals regularly recording patient information digitally and using the recorded information to aid in diagnosis and clinical decision making.[2] In addition to the direct clinical benefits of easily being able to look up individual patient records, large collections of EHRs are also useful to researchers who wish to understand medical conditions, disease processes, and hospitalization patterns based on retrospective records. Data documenting patient care is recorded in EHRs through (a) structured measurements such as lab results, vital signs, demographic information, etc., as well as (b) unstructured clinical narratives such as those found in admission reports, nursing notes, or surgical reports. Clinical notes are routinely recorded to provide relevant contextual information about the patient's medical background, condition, and care. While there can be overlap between the information included in the structured and unstructured portions of EHRs, the unstructured nature of clinical notes is uniquely suited for extracting unique or contextual information that may not conform to a preset field or measurement.[3] Both text and structured data have been used to model medical processes for Clinical Decision Support[4,5] and Disease Prediction[6] applications.

Denny (2012)[7] points out that phenotype identification is one task for which clinical notes are particularly useful, noting that this is often because salient observations in text documents such as pathology and radiology reports are often not also included in tabular data. However, while clinical notes are clearly a useful data source, it is still important to know specifically where (i.e. for which phenotypes) notes are most likely to be beneficial and to identify whether there are situations where they are not worthwhile to include at all. Previous studies have used custom-engineered search terms in clinical notes to identify specific clinical phenotypes,[8] while others used a Bag of Words or Bag of Concepts representation to identify clinical phenotypes based on the words that appear in clinical notes.[9,10] However, it is difficult to determine based on these studies whether clinical notes, structured data, or a combination of both should be used to identify a given phenotype that has not previously been automatically identified. Helpfully, Scheurwegs et al. (2015)[11] show that the best performance can be consistently achieved by combining structured data with a Bag of Words representation of clinical notes when predicting ICD-9 billing codes from fourteen different medical specialties based on EHR data. They conclude that adding structured data is consistently beneficial when compared with using only a Bag of Words representation of clinical notes. However, it is not clear if the performance increase observed by Scheurwegs et al. (2015)[11] when adding structured data is due to information that is present in structured data but missing in clinical notes, or if the information is in fact present in clinical notes but is simply not captured using a Bag of Words representation.

For the purposes of this study, we use the term *clinical phenotype* to refer to a clinically significant group of medical

abnormalities that are characteristic of a single disease (sometimes referred to as a disease phenotype).[12] We determine the individual clinical phenotypes of a patient by means of the ICD-9 diagnostic codes assigned to the patient upon discharge from the hospital, as demonstrated in previous phenotyping work.[13,14] Specifically, we rely on the clinically meaningful groupings of ICD-9 codes as defined by the Agency for Healthcare Research and Quality (AHRQ)'s Clinical Classification Software (CCS). We consider the presence of any ICD-9 code in each of these CCS groupings as evidence for the corresponding clinical phenotype.

In this paper, we train a Long Short-Term Memory network (LSTM)[15] to identify 172 phenotypes using structured data features, textual features, and their combination. We explore three methods for representing textual features: Bag of Words, Bag of Concepts, and Doc2Vec. Finally, we apply statistical analysis on phenotyping performance (measured in AUC) to investigate which categories of phenotypes are most likely to benefit from each data source, and which benefit from their combination. Our results indicate that the combination of text and structured features provides a statistically significant ($p < 0.05$) increase in performance compared to using structured features alone for 51 phenotypes, a decrease in performance for a single phenotype, and no statistically significant benefit for the remaining 117 phenotypes. We analyse the categories of phenotypes that are best identified with each data set, provided analyses which we hope can serve as a practical guide for determining which data sources are most likely to be of value for a given phenotype.

Background and Related Work

The task of automatic clinical phenotype identification consists of computationally processing EHRs to determine whether a given patient's clinical phenotype indicates a particular disease. This can be useful when a disease may not have been explicitly documented in the EHR despite the presence of markers that are characteristic of the disease in the patient's record. In addition to being useful for Clinical Decision Support, research into automatic phenotype identification can lead to a better understanding of disease mechanisms and outcomes.[16] Automatic phenotype identification is also a preliminary step in clinical research that requires selecting patient cohorts based on their diseases or medical conditions.[17,18]

The relationship between structured and unstructured data for phenotyping has been explored in the past. For example, Liao et al. (2010)[10] show that rheumatoid arthritis can be automatically identified most effectively when using a combination of both clinical notes and structured data, and Nunes et al. (2016)[19] reached a similar conclusion when identifying hypoglycemic patients based on EHRs. However, neither of these studies compares systems relying on clinical notes with systems relying on structured data for a wide range of different phenotypes, meaning that their findings may not be completely generalizable to the broader task of phenotype identification. Gehrmann et al. (2018)[9] compared different Natural Language Processing approaches for using clinical notes to identify 10 different clinical phenotypes, but did not focus on comparing notes with structured data (though they do express interest in doing this in their *"Future Extensions"* section). Consequently, this study can be viewed as a generalization of these approaches wherein we test several different formulations of text, structured data, and a combination of both to predict a broad range of phenotypes, thus providing insight into whether text is useful for identifying all phenotypes, or just a specific subset of them.

Data

In this work, we extracted general structured data features from patients' laboratory and chart values in the MIMIC-III Critical Care Database.[20] The MIMIC-III Critical Care Database[20] contains de-identified structured data and timestamped clinical notes for *46*, *520* patients and *61*, *532* ICU stays, based on data collected between 2001 and 2012. This includes clinical notes, lab results, demographic information, logs of hospital admission and discharge, prescriptions, and admission-level ICD-9 discharge diagnoses and procedures. For the sake of reproducibility and comparison, we used the 38 structured features extracted from the MIMIC-III Benchmark[21] to identify phenotypes. The MIMIC-III Benchmark is a set of four clinical prediction tasks using a consistent set of structured data features for all tasks: capillary refill rate, diastolic blood pressure, fraction inspired oxygen, Glasgow coma scale eye opening, Glasgow coma scale motor response, Glasgow coma scale total, Glasgow coma scale verbal response, glucose, heart rate, height, mean blood pressure, oxygen saturation, respiratory rate, systolic blood pressure, temperature, weight, and pH. We adapt the same set of structured features in this work. In addition, we used all types of clinical notes present in MIMIC-III (e.g. admission and discharge reports, nursing notes, radiology reports) in our experiments using notes.

As in Harutyunyan et al. (2019)[21], we considered each ICU stay as an independent episode and used discharge

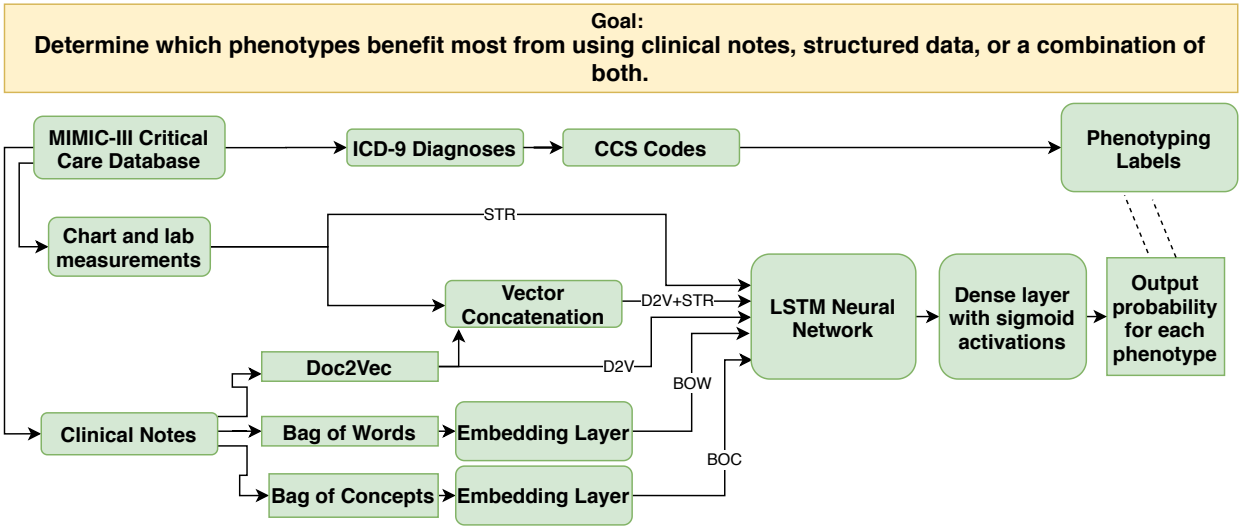

Figure 1: Schematic of the our pipeline including data sources and neural network.

Table 1: Top-level CCS phenotype categories with their size (i.e., the number of phenotypes in the group after filtering phenotypes with < 30 episodes) as well as the prevalence of that group in the testing data.

CCS Category Name	CCS Codes	Size	Prevalence
Infectious and parasitic diseases	1-10	8	26.6 %
Neoplasms	11-47	17	22.2 %
Endocrine; nutritional; and metabolic diseases and immunity disorders	48-58	9	66.9 %
Diseases of the blood and blood-forming organs	59-64	5	35.4 %
Mental illness	650-663, 670	10	35.6 %
Diseases of the nervous system and sense organs	76-95	15	28.1 %
Diseases of the circulatory system	96-121	24	81.6 %
Diseases of the respiratory system	122-134	9	47.3 %
Diseases of the digestive system	135-155	15	41.2 %
Diseases of the genitourinary system	156-175	9	39.7 %
Diseases of the skin and subcutaneous tissue	197-200	4	9.6 %
Diseases of the musculoskeletal system and connective tissue	201-212	11	20.7 %
Congenital anomalies	213-217	2	2.7 %
Injury and poisoning	225-244	16	43.7 %
Symptoms; signs; and ill-defined conditions and factors influencing health status	245-258	9	25.3 %
Residual codes; unclassified; all E codes	259-260	9	41.4 %

diagnoses to determine phenotype labels. Specifically, phenotypes were defined by CCS (Clinical Classifications Software) codes. Developed by the Agency for Healthcare Research and Quality (AHRQ), CCS codes are groups of ICD-9 codes that correspond to specific diseases; these CCS codes are further grouped into a hierarchy based on organ systems and disease categories. Table 1 shows the different top-level CCS categories and the number of phenotypes in each category that we evaluated in this study. Note: unlike Harutyunyan et al. (2019)[21] which considered only 25 phenotypes, we considered all phenotypes associated with at least 30 episodes in MIMIC-III resulting in the 172 phenotypes shown in Table 1. Moreover, we filtered episodes such that all episodes had at least one data point from both the textual and structured data sources. In this study we used the same 14:3:3 splits for training, validation, and testing used by Harutyunyan et al. (2019)[21]. Due to the potential for different disease manifestation, we omitted patients under the age of 18 from this study.

Methods

As shown in Figure 1, our phenotype-identification pipeline consists of preparing training, validation, and testing data based on MIMIC-III and then repeatedly training and evaluating the same deep neural network using different combinations of structured data and three different representations of clinical notes. Our data preparation pipeline consists primarily of (1) extracting clinical episodes based on ICU stays in MIMIC-III, (2) selecting a data source or combination thereof, (3) producing fixed-length continuous vectors capturing the information in the selected data source, and (4) using these fixed-length vectors as the input to train a shared deep neural network for joint phenotype identification.

A. Data Representations

To compare the impact of different data sources, it was necessary to represent the structured data and clinical notes recorded at each timestep with fixed-length, continuous vector encodings. We considered the following methods for representing data in each episode:

1. **Structured Data:** As in Harutyunyan et al. (2019)[21], measurements were aggregated every hour. For measurements consisting of a continuous number (such as height, weight, and temperature), we dedicate one vector dimension to the raw numerical measurement, and for categorical measurements (such as the Glasgow coma scale fields) we use a one-hot encoding with one vector dimension reserved for each possible value of the given measurement. All values were then normalized by subtracting the mean value of each field and dividing by the standard deviation. The resultant structured data vector included a total of 76 elements corresponding to these continuous and categorical variables.
2. **Bag of Words:** We generated vocabularies based on all of the clinical notes in MIMIC-III. For words, the vocabulary size was 1,891,434 words. For each clinical note, we created a Bag of Words vector representation by using a vocabulary-length vector with ones for all words that occur in the note and zeros for all words that do not occur.
3. **Bag of Concepts:** We used MetaMap Lite[22] to identify concepts in clinical notes based on UMLS[23]. Defining the vocabulary to consist only of concepts occurring in clinical notes, we then created a vocabulary-length vector to represent the medical concepts in each note, similarly to the Bag-of-Words representation above. The vocabulary size for concepts was 51,893 concepts.
4. **Doc2Vec:** Because the order in which words appear is not considered by either the Bag of Words nor the Bag of Concepts approaches, we considered a more sophisticated representation – Doc2Vec[24] – for generating document-level representations of each clinical note. We used a vector dimension of 300, an initial learning rate of 0.025, and 100 iterations.*
5. **Structured and Doc2Vec:** We considered a final representation in which the Doc2Vec and Structured Data encodings were concatenated together to form a single multi-datasource representation. Because structured data was available every hour while only 3.2 notes were produced per day (on average), the features corresponding to Doc2Vec were left as zero for hours in which no notes were generated.

Note: an important advantage of the three text encoding schemes is that they can all be applied to entire documents regardless of the document length, unlike other text-based neural network systems such as the Universal Sentence Encoder[25] or BERT[26].

B. Neural Network Architecture

We identify phenotypes by training a Long-Short-Term-Memory (LSTM) network similar to that used in Harutyunyan et al. (2019)[21]. Specifically, we train the LSTM to sequentially process the feature representations from each time-step for a patient such that the output of the LSTM after processing the last time-step can be used to predict the clinical phenotypes of the patient. Because our goal was to test the effect of different data sources (structured vs various forms of text), our approach was to keep the neural network architecture as constant as possible across all of our experiments while varying only the input data. However, it was necessary to modify the network's architecture slightly by adding a fully-connected layer to embed the inputs when using a Bag of Concepts or Bag of Words representation for clinical note text, before input to the first LSTM layer. Inputs are then passed into a bidirectional LSTM network using hyperbolic tangent activation functions. Finally, all outputs are passed through a dense layer using 172 parallel sigmoid activation functions to produce the probability of identifying each of the 172 phenotypes considered in this work.

*Determined empirically

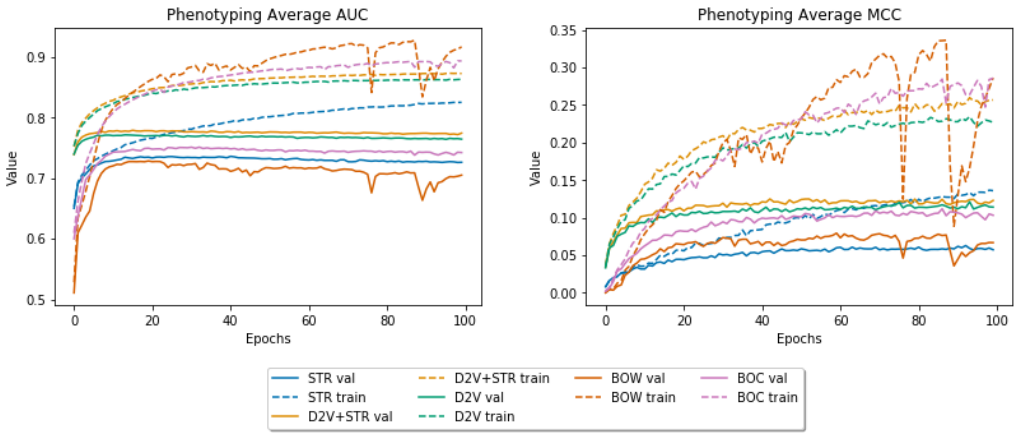

Figure 2: Validation MCC score and AUC-ROC score at each epoch of training systems on the 172-class phenotyping task with the standard LSTM architecture.

Experiments

We compared the LSTM's phenotyping performance when provided with the five different data source representations described earlier: (1) structured data from the patient's chart and lab results only (STR), (2) clinical notes only, encoded as a Bag of Words (BOW); (3) clinical notes only, encoded as a Bag of UMLS Concepts (BOC); (4) clinical notes only, encoded using Doc2Vec (D2V); and (5) a combination of structured data and Doc2Vec-encoded clinical notes (STR+D2V). In all experiments, models were trained for up to 100 epochs, using early stopping based on validation MCC (described below). The AUC and MCC on the validation set are shown at all epochs of training in Figure 2, where it is clear that, without early stopping, all data representations enable the LSTM to overfit well before reaching 100 epochs.

A. Evaluation Metrics

While accuracy is commonly used to evaluate the performance of classification systems, it is not particularly useful for imbalanced classes (in our case, less-prevalent phenotypes). To account for this, we reports metrics that are robust to class imbalance by taking into account the number of false positives associated with each class. We primarily used the Matthews Correlation Coefficient (MCC) as our metric for selecting the best epoch and combaring models during development. Equation (1) shows the Matthews Correlation Coefficient in terms of the number of False Negatives (FN), True Negatives (TN), False Positives (FP), and True Positives (TP):

$$\text{MCC} = \left(\frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \right) \tag{1}$$

The advantage of using MCC for phenotype identification is that it gives a reliable measure of the model's performance on both common and uncommon phenotypes, whereas Recall, Precision, and AUC are hard to interpret when the class frequencies are imbalanced. Because we are evaluating joint-phenotype identification, we report the macro-average MCC across all 172 phenotypes (using a classification threshold of 0.5).

For evaluating the performance of our models we report the Area under the Receiver Operating Characteristic curve (AUC). This curve is obtained by plotting the model's True Positive Rate (TPR) against the False Positive Rate (FPR) for different values of the binary classification threshold. To determine whether different data sources resulted in statistically significantly different AUCs, we relied on the AUC confidence interval obtained using DeLong's Test.[27]. DeLong's Test is an asymptotically exact method to evaluate the uncertainty of an AUC plot, allowing us to determine 95% confidence intervals for the AUC produced by each data source for each phenotype.

Results

We report the aggregate phenotyping performance for all five data representations in Table 2. Specifically, we report the AUC, MCC, average Precision and average F_1 scores for 172 phenotypes as evaluated using a held-out testing set

of 5424 episodes.

For the sake of comparison, we also report the impact of each data representation for the other three tasks evaluated in Harutyunyan et al. (2019)[21] – In-hospital Mortality, Length of Stay, and Decompensation Prediction – as well as the 25 CCS phenotypes considered in that work. Table 3 reports these results, using the same metrics reported in Harutyunyan et al. (2019)[21]. For the Decompensation and Length-of-Stay tasks, we used the "deep supervision" formulation of the task in which the targets are predicted after every timestep. Note, for Length of Stay, MCC could not be computed; thus, we always tested the Length of Stay model that had the best kappa score when evaluated using the validation set. We found that the only task for which the clinical notes were definitively helpful was phenotyping.

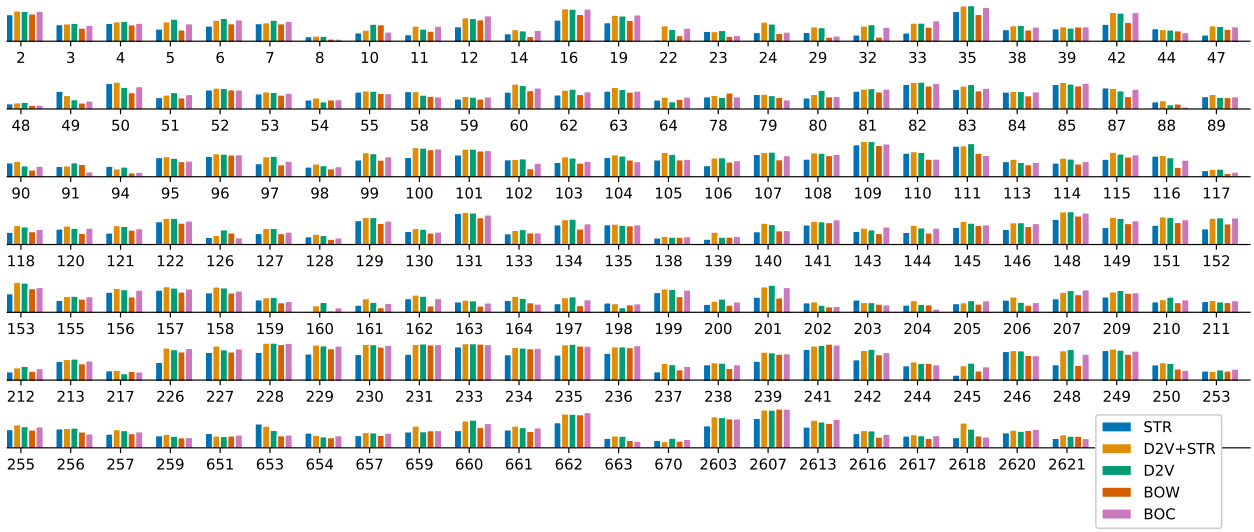

Figure 3: AUC with each data source for all 172 phenotypes (labeled by their CCS codes); the vertical axis represents AUC between 0.55 and 1.0.

Discussion

As shown in Figure 3, clinical notes had a strong positive impact on phenotyping performance. These results indicate that, on average, the combination of structured data and clinical notes was best, with regard to AUC-ROC, MCC, and average precision. In addition, clinical notes alone are superior to structured data alone.[†] Nonetheless, this trend is not true for every phenotype. Analyzing the performance on individual phenotypes and clinically-relevant categories of phenotypes can offer a more fine-grain explanation of why some phenotypes are highly amenable to being identified through clinical notes while others are not. We make use of the multi-level CCS categories (shown in Table 1) to group the clinical phenotypes and determine if phenotypes within certain categories are more easily identified using text (i.e., clinical notes) or text and structured data. Figure 3 shows the AUC-ROC obtained for each of the 172 phenotypes using

[†]We conducted additional experiments re-sampling structured data only during timesteps with notes, and found that even accounting for the frequency of structured data, clinical notes were still superior.

Table 2: Performance when identifying 172 clinical phenotypes from MIMIC-III hospital stays in the test set. Clinical phenotyping performance; Precision, Recall, and F_1 are weighted macro-averages based on phenotype prevalence.

Data	AUC (Micro)	AUC (Macro)	AUC (Weighted)	MCC	Precision	Recall	F_1
STR	85.55%	72.67%	73.04%	6.67%	30.56%	30.56%	16.69%
D2V+STR	87.84%	77.52%	77.37%	13.23%	37.20%	37.20%	25.05%
D2V	87.54%	76.83%	76.55%	12.58%	36.31%	36.31%	23.83%
BOW	84.50%	72.54%	72.48%	7.95%	30.87%	30.87%	20.94%
BOC	86.22%	74.85%	74.59%	10.46%	34.05%	34.05%	23.43%

Table 3: Performance on In Hospital Mortality, Decompensation, Length of Stay, and 25-Class Phenotyping using the metrics reported by Harutyunyan et al.[21]. The relative ranking of each approach is provided in parenthesis.

	In-Hospital Mortality			Decompensation		
Data	AUC-ROC	AUC-PRC	MCC	AUC-ROC	AUC-PRC	MCC
STR	0.854 (#1)	0.470 (#1)	0.383 (#1)	0.900 (#1)	0.309 (#1)	0.326 (#1)
BOW	0.716 (#5)	0.225 (#5)	0.168 (#5)	0.822 (#4)	0.167 (#4)	0.207 (#4)
BOC	0.770 (#4)	0.322 (#4)	0.263 (#4)	0.846 (#3)	0.211 (#3)	0.272 (#3)
D2V+STR	0.848 (#2)	0.461 (#2)	0.369 (#2)	0.885 (#2)	0.236 (#2)	0.273 (#2)
D2V	0.790 (#3)	0.345 (#3)	0.295 (#3)	0.821 (#5)	0.151 (#5)	0.192 (#5)

	Length of Stay			25-Class Phenotyping		
Data	MAD	MAPE	Kappa	AUC	MCC	Precision
STR	108.718 (#2)	185.011 (#2)	0.437 (#1)	0.739 (#4)	0.238 (#5)	0.448 (#4)
BOW	115.345 (#4)	161.749 (#1)	0.301 (#5)	0.732 (#5)	0.251 (#4)	0.445 (#5)
BOC	108.815 (#3)	260.538 (#4)	0.426 (#3)	0.752 (#3)	0.307 (#3)	0.481 (#3)
D2V+STR	108.348 (#1)	191.180 (#3)	0.427 (#2)	0.792 (#1)	0.369 (#1)	0.540 (#1)
D2V	118.734 (#5)	264.235 (#5)	0.404 (#4)	0.783 (#2)	0.352 (#2)	0.533 (#2)

each data representation.

A. How to Represent Clinical Notes

There where no phenotypes for which the Bag of Words had an AUC that was statistically-significantly higher than that of the Bag of Concepts. There were however 8 phenotypes for which the Bag of Concepts performed significantly better than the Bag of Words. Neither Bag of Words nor Bag of Concepts performs significantly better than Doc2Vec for any phenotype. For this reason, we did not consider combining structured data with BoW or BoC in our experiments.

B. When to Combine Clinical Notes with Structured Data

For 160 out of the 172 clinical phenotypes that we considered, we observed an increase in AUC when including clinical notes rather than solely using structured data, and for 51 of these the benefit is statistically significant. When breaking the phenotypes into the CCS categories shown in Table 1, the two categories in which the majority of the phenotypes exhibited a statistically significant increase in AUC when including text along with structured data were "diseases of the circulatory system" and "Injury and poisoning", indicating that it is generally worth including text when identifying phenotypes in these categories. These two categories of phenotypes accounted for 46% of the phenotypes for which combining clinical notes with structured data yielded a statistically significant improvement (diseases of the circulatory system accounting for 26% and injury and poisoning the remaining 20%).

The five phenotypes with the largest statistically significant increase in improvement when adding Doc2Vec to structured features were: "Fracture of neck of femur (hip)" (226), "Melanomas of skin" (22), "Gangrene" (248), "Secondary malignancies" (42), and "Gastrointestinal hemorrhage" (153). These conditions likely benefit from textual features because they all are primarily documented in radiology or surgery reports and are not directly indicated by vital signs or numerical lab measurements. Thus, we recommend combining clinical notes and structured data for phenotypes characterized by both discrete observations and continuous or structured measurements.

Figure 4 shows ROC curves for "Gastrointestinal hemorrhage", where clinical notes were particularly useful, alongside "Diabetes mellitus without complication", the phenotype for which notes were the most detrimental. The positions of the ROC curves for text and structured data sources are reversed between these two plots, with the ROC curve for structured data appearing below all other sources in the case of "Gastrointestinal hemorrhage", but above all the others in the case of "Diabetes mellitus without complication".

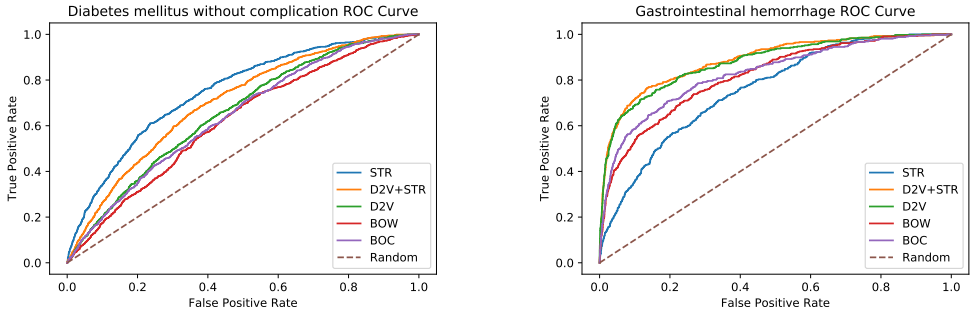

Figure 4: Receiver Operating Characteristic (ROC) curve when identifying phenotypes in the test set with the standard LSTM architecture trained on 172 phenotypes.

C. When to Use Only Structured Data

There were 121 phenotypes for which there was no statistically significant gain from combining clinical notes with structured data. Out of these, there were 12 phenotypes for which adding clinical notes to the structured data harmed the AUC performance at least slightly, but only one of these differences was statistically significant. The only phenotype for which the combination of the combination of text and structured data performed statistically-significantly worse than structured data alone was "Diabetes mellitus without complication" (49), from the "Endocrine; nutritional; and metabolic diseases and immunity disorders" category. A likely reason for this result is the prevalence of Glucose, which is the main measure used for diagnosing Diabetes,[28] in the chart and lab measurements and the inability of Doc2Vec to identify glucose values or severity indicators from the text.

In addition, there were 5 CCS categories for which no phenotypes had a statistically significant improvement when including clinical notes with structured data: "Congenital anomalies", "Diseases of the musculoskeletal system and connective tissue", "Diseases of the nervous system and sense organs", "Diseases of the skin and subcutaneous tissue", and "Endocrine; nutritional; and metabolic diseases and immunity disorders". Table 1 shows the prevalence of each of these groups in our data set. While it is possible that under-performance of text may be influenced by the fact that these phenotype categories are underrepresented in MIMIC-III compared with categories that are frequently treated in an ICU setting such as "Injury and Poisoning", we believe the low level of difference in performance for these categories between text-based and structured-data-based identification seems to indicate that text is less useful for these categories of phenotype. Moreover, because Doc2Vec learns a task-agnostic representation of clinical notes, it will naturally be better able to encode information about more commonly documented observations, suggesting that perhaps context-aware representations such as those learned by BERT may be better suited for recognizing less prevalent phenotypes.

D. When to Use Only Clinical Notes

While adding Doc2Vec-encoded text to structured data often improved results when compared with using structured data alone (as explained above), using the Doc2Vec encoded-text alone was never statistically significantly worse than using the combination of Doc2Vec and structured data, with the exception of two phenotypes: "Diabetes mellitus with complications" and "Diabetes mellitus without complication". This leads us to believe that while Scheurwegs et al. (2015)[11] found that clinical notes and structured data were generally complimentary when using a Bag of Words to represent clinical notes, switching to a better-performing representation of clinical notes such as Doc2Vec can render the addition of structured data mostly redundant.

E. Limitations and Future Work

It is possible that many of the advantages of clinical notes come from the fact that medication names, dosages, and administration instructions are often strongly associated with specific medical conditions. For this reason it would be worth considering including other types of structured data such as prescriptions rather than just laboratory and chart measurements. In addition, we would like to identify specific clinical and textual features that make some phenotypes particularly easy to identify based on note text. While our LSTM-based approach was effective for taking advantage of the chronological nature of EHR data, due to nonlinearities in the network, it does not allow us to easily determine

which features were most important when identifying each phenotype.

In future work, we would like to use a neural network that is more likely to take advantage of the linguistic complexities of clinical notes. For example, it is possible that other methods such as transformer models[29] would be more successful at extracting useful information from text, and could also potentially be useful for improving the timeseries processing of structured data. We would also like to explore in future work the extent to which the information contained in clinical notes and structured data tend to overlap to better inform our analysis and intuitions of whether using clinical notes will be helpful for a given task. We would also like to continue our work with the other three tasks from the MIMIC-III Benchmark, to determine whether there is a way to better harness text such that using clinical notes would improve rather than reduce performance on those tasks.

Code Availability

The code for this work can be found at `github.com/amoldwin/notes_benchmark`.

Conclusion

Exploring the effect of including clinical notes for EHR-based phenotyping, we found that phenotyping performance generally benefited greatly from the inclusion of clinical notes. We observed that while on average there was a significant difference in performance between systems that use structured data exclusively and those that use a combination of text and structured data, some groups of phenotypes, such as "diseases of the circulatory system" and "injury and poisoning" are most likely to benefit in a statistically significant way from the combination of structured data and text, while others can be detected reliably from text alone. When comparing different text representations, the Bag of Concepts tended to be more effective than the Bag of Words approach, but was consistently less effective than Doc2Vec, indicating that the word order and document structure are in fact important for clinical phenotyping. By utilizing a broad array of common phenotypes, we were able to compare the efficacy of these systems across a wide array of phenotypes, allowing us to determine how generalizable our findings were. By taking into account our findings, future researchers may decide to rely on text rather than structured data for phenotyping applications, if given a choice between the two. We hope that this study will help inform researchers and clinicians in situations where it is necessary to design task-specific phenotyping systems for cohort selection and clinical decision support applications.

Acknowledgements

This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health and utilized the computational resources of the NIH HPC Biowulf cluster (`http://hpc.nih.gov`).

References

1. Atasoy H, Greenwood BN, Mccullough JS. The Digitization of Patient Care: A Review of the Effects of Electronic Health Records on Health Care Quality and Utilization. Annual Review of Public Health. 2019;40(1):487–500.
2. Romano MJ, Stafford RS. Electronic Health Records and Clinical Decision Support Systems. Archives of Internal Medicine. 2011;171(10).
3. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. Data from clinical notes: a perspective on the tension between structure and flexible documentation. Journal of the American Medical Informatics Association. 2011;18(2):181–186.
4. Apostolova E, Wang T, Tschampel T, Koutroulis I, Velez T. Combining Structured and Free-text Electronic Medical Record Data for Real-time Clinical Decision Support. Proceedings of the 18th BioNLP Workshop and Shared Task. 2019;.
5. Rossetti SC, Knaplund C, Albers D, Tariq A, Tang K, Vawdrey D, et al. Leveraging Clinical Expertise as a Feature - not an Outcome - of Predictive Models: Evaluation of an Early Warning System Use Case. AMIA Annu Symp Proc. 2019;2019:323–332.
6. Sun M, Baron J, Dighe A, Szolovits P, Wunderink RG, Isakova T, et al. Early Prediction of Acute Kidney Injury in Critical Care Setting Using Clinical Notes and Structured Multivariate Physiological Measurements. Studies in Health Technology and Informatics. 2019;264:368–372.
7. Denny J. Chapter 13: Mining Electronic Health Records in the Genomics Era. PLoS computational biology. 2012 12;8:e1002823.
8. Ludvigsson JF, Pathak J, Murphy S, Durski M, Kirsch PS, Chute CG, et al. Use of computerized algorithm to identify individuals in need of testing for celiac disease. Journal of the American Medical Informatics Assn. 2013;.

9. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. Plos One. 2018;13(2).
10. Liao KP, Cai T, Gainer V, Goryachev S, Zeng-Treitler Q, Raychaudhuri S, et al. Electronic medical records for discovery research in rheumatoid arthritis. Arthritis Care & Research. 2010;62(8):1120–1127.
11. Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Bulcke TVD. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. Journal of the American Medical Informatics Association. 2015;23(e1).
12. Scheuermann RH, Ceusters W, Smith B. Toward an ontological treatment of disease and diagnosis. Summit Transl Bioinform. 2009 Mar;2009:116–120.
13. Sinnott JA, Cai F, Yu S, Hejblum BP, Hong C, Kohane IS, et al. PheProb: probabilistic phenotyping using diagnosis codes to improve power for genetic association studies. Journal of the American Medical Informatics Association. 2018 05;25(10):1359–1365. Available from: https://doi.org/10.1093/jamia/ocy056.
14. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. Journal of the American Medical Informatics Association. 2015 09;23(e1):e20–e27. Available from: https://doi.org/10.1093/jamia/ocv130.
15. Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Comput. 1997 Nov;9(8):1735–1780. Available from: https://doi.org/10.1162/neco.1997.9.8.1735.
16. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. npj Digital Medicine. 2019;2(1).
17. Alzoubi H, Alzubi R, Ramzan N, West D, Al-Hadhrami T, Alazab M. A Review of Automatic Phenotyping Approaches using Electronic Health Records. Electronics. 2019;8(11):1235.
18. Rethinking Clinical Trials;. Available from: https://sites.duke.edu/rethinkingclinicaltrials/informed-consent-in-pragmatic-clinical-trials/.
19. Nunes AP, Yang J, Radican L, Engel SS, Kurtyka K, Tunceli K, et al.. Assessing occurrence of hypoglycemia and its severity from electronic health records of patients with type 2 diabetes mellitus. U.S. National Library of Medicine; 2016. Available from: https://www.ncbi.nlm.nih.gov/pubmed/27744128.
20. Johnson AE, Pollard TJ, Shen L, Lehman LWH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Scientific Data. 2016;3(1).
21. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Scientific Data. 2019;6(1):96. Available from: https://doi.org/10.1038/s41597-019-0103-9.
22. Demner-Fushman D, Rogers WJ, Aronson AR. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. Journal of the American Medical Informatics Association. 2017;.
23. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research. 2004;32(90001).
24. Le Q, Mikolov T. Distributed Representations of Sentences and Documents. In: Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32; 2014. .
25. Cer D, Yang Y, Kong Sy, Hua N, Limtiaco N, John RS, et al. Universal Sentence Encoder for English. ACL Anthology;Available from: https://www.aclweb.org/anthology/D18-2029/.
26. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: NAACL-HLT; 2019. .
27. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837–845.
28. Sacks DB, Bruns DE, Goldstein DE, Maclaren NK, McDonald JM, Parrott M. Guidelines and Recommendations for Laboratory Analysis in the Diagnosis and Management of Diabetes Mellitus. Clinical Chemistry. 2002 03;48(3):436–472. Available from: https://doi.org/10.1093/clinchem/48.3.436.
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is All you Need. In: Advances in Neural Information Processing Systems 30. Curran Associates, Inc.; 2017. p. 5998–6008.