original report

Clinical Data Extraction and Normalization of Cyrillic Electronic Health Records Via Deep-Learning Natural Language Processing

Boyang Zhao, PhD[1]

abstract

PURPOSE A substantial portion of medical data is unstructured. Extracting data from unstructured text presents a barrier to advancing clinical research and improving patient care. In addition, ongoing studies have been focused predominately on the English language, whereas inflected languages with non-Latin alphabets (such as Slavic languages with a Cyrillic alphabet) present numerous linguistic challenges. We developed deep-learning–based natural language processing algorithms for automatically extracting biomarker status of patients with breast cancer from three oncology centers in Bulgaria.

METHODS We used dual embeddings for English and Bulgarian languages, encoding both syntactic and polarity information for the words. The embeddings were subsequently aligned so that they were in the same vector space. The embeddings were used as input to convolutional or recurrent neural networks to derive the biomarker status of estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2.

RESULTS We showed that we can resolve ambiguity in highly variable medical text containing both Latin and Cyrillic text. Final models incorporating both English and Bulgarian syntax and polarity embeddings achieved F_1 scores of 0.90 or higher for all estrogen receptor, progesterone receptor, and human epidermal growth factor receptor 2 biomarkers. The models were robust against human errors originally found in the training set. In addition, such models can be extended for analyzing text containing words not seen during training.

CONCLUSION By using several techniques that incorporate dual-word embeddings encoding syntactic and polarity information in two languages followed by deep neural network architectures, we show that researchers can extract and normalize parameters within medical data. The principles described here can be used to analyze Cyrillic or Latin mixed medical text and extract other parameters.

JCO Clin Cancer Inform. © 2019 by American Society of Clinical Oncology

ASSOCIATED CONTENT

Appendix

Author affiliations and support information (if applicable) appear at the end of this article.

Accepted on August 13, 2019 and published at ascopubs.org/journal/cci on October 2, 2019: DOI https://doi.org/10.1200/CCI.19.00057

INTRODUCTION

Digitalization and extraction of medical records is critical in clinical research, patient recruitment for clinical trials, and improved patient care in the era of value-based care. However, more than 80% of data in electronic health records (EHRs) exists as unstructured text. Automated extraction of medical text into structured data is challenging. In addition, substantial variance in reporting because of different sources further compounds the complexity. Manual extraction from such text is extremely tedious and cost prohibitive because specialized medical knowledge is often required. It is also not scalable, given the high volume of medical data to be annotated. So addressing even simple epidemiologic questions can be a burdensome task.

In working with the hospital information systems in Eastern Europe, we observed additional challenges. We focus specifically on Bulgarian EHRs, although this discussion is relevant to other non-English languages. Here we provide examples of the biomarker status for patients with breast cancer. Estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) are important prognostic and predictive biomarkers in breast cancer. Their status is used for tumor classification and guides the use of hormonal, anti-HER2, or other appropriate therapies for treatment and management of the cancer. The challenges in extracting biomarker status include the following seven points: (1) Terms in both English (in Latin) and Bulgarian (in Cyrillic); for example, there is substantial heterogeneity in the way the progesterone biomarker is described: PR, **ПР**, **Прогестерон**, **Рецепторният**, **хормонални** (translation: PR [Latin], PR [Cyrillic], progesterone, receptor, hormonal); (2) misspellings, such as **Протестерон**

CONTEXT

Key Objective

This study aims to develop novel approaches in natural language processing for the extraction and normalization of unstructured data from electronic health records (EHRs) containing both English and Cyrillic text.

Knowledge Generated

We incorporated dual-word embeddings of two languages aligned into the same vector space and augmented with polarity information. The combined embedding and convolutional or recurrent neural network–based architecture can accurately retrieve the biomarker status of patients with breast cancer from unstructured medical data and is robust to human errors, unseen words, and data from different sources.

Relevance

This study is generalizable for extracting and normalizing unstructured EHR data with substantial heterogeneity in parameter values and language. The ability to obtain important patient information from unstructured EHRs can enhance the tracking of treatment outcome, eligibility for clinical trials, and epidemiology.

instead of Прогестерон; (3) multiple variants of the same word because Bulgarian and other Slavic languages are inflected languages; (4) heterogeneity in the position of the value relative to the target parameter: "негативна ексПресия на HER2" or "HER2-негативен" (translation: "negative expression of HER2" or "HER2-negative"); (5) heterogeneity in value types and values themselves: PR +, Прогестерон-негативен, PR75%, PR/4т./-Позит (translation: PR +, progesterone-negative, PR75%, PR/4t./- posit); (6) variable lengths of the parameter values and what delimiter can be used before and after the target parameter: for example, "ER и PR Позитивна" or "ER Позитивна и PR Позитивна" (translation: "ER and PR positive" or "ER positive and PR positive"); and (7) ambiguity with certain characters, such as "–", which can be used as the value "negative" or as a delimiter (eg, PR–/+/). Other challenges include human error in the labeling of such data sets, and substantial unbalanced data sets for parameter values.

Many methods have been studied for extracting data from clinical text, and they can be broadly classified into rule-based and machine-learning–based approaches. Rule-based methods[1,2] can be labor intensive regarding engineering the logic for parsing the data and are unlikely to capture all values in current and unseen future medical data. For example, earlier analysis showed that there were at least 124 ways for denoting "invasive ductal carcinoma" in breast cancer pathology reports, excluding misspellings and spacing errors.[2] One study achieved only 69% accuracy for extracting hormonal status.[3] Alternatively, machine learning or natural language processing (NLP)–based data extraction[4-6] has been applied with varying degrees of accuracy. Most use classical machine learning classifiers (eg, support vector machines [SVMs], decision trees) as individual models or as part of ensemble models. In light of the substantial variance we observed in Bulgarian EHRs, these approaches are likely not sufficient for accurate data extraction. Only a few limited studies focused on non-English medical text, and many of them used similar rules-based and machine learning–based approaches.[4,7] Studies of extraction from medical Cyrillic languages are even rarer.[8-10] Although applications of deep learning for the specific task of medical text extraction are limited, they have recently been used for extracting International Classification of Diseases for Oncology, 3rd revision (ICD-O-3) codes with limited accuracy (F_1 score of 0.722).[11]

Here we use unstructured medical data in EHRs from Bulgaria for patients with breast cancer to build models for the extraction and normalization of the status of biomarkers. We show that we can apply deep-learning NLP models on the basis of convolutional neural networks (CNNs) or recurrent neural networks that are superior to classical machine learning algorithms to accurately determine the status of biomarkers for patients with breast cancer.

METHODS

Data Set

We retrieved anonymized data from three oncology hospitals in Bulgaria. We filtered specifically for patients with breast cancer and retrieved medical records from procedure 38, which contains unstructured text with information pertaining to biomarker status of patients. A total of 1,235 records from two hospitals were retrieved from April 1, 2018, to January 31, 2019. The distributions of the data set were 86% PR-positive, 14% PR-negative, 99% ER-positive, 1% ER-negative, 8% HER2-positive, and 92% HER2-negative. In addition, a total of 1,011 records from three hospitals were retrieved from February 1, 2019, to April 31, 2019, and were withheld; the test data set was not examined, augmented, or modified in any way and was used only once at the end for final model evaluation. The distributions of the test data set were 88% PR-positive, 12% PR-negative, 99.7% ER-positive, 0.3% ER-negative, 8% HER2-positive, and 92% HER2-negative. The data sets were manually annotated with biomarker status for ER, PR, and HER2 by one lead physician, and four additional members were involved in the preparation of, assistance with, and review of the manually annotated records.

We augmented the data set to create a more balanced input data set for training and validation. For PR and ER, we had substantially more positive than negative receptor values. In contrast, for HER2, we had substantially more negative than positive receptor values. We extracted, wherever possible using regular expression, text segments pertaining to the biomarker value. New data were created by appending pre- and post-text segments drawn from other entries (eg, [pre-text segment] [PR segment] [post-text segment]). We confirmed the robustness of our data augmentation and balancing and confirmed that it did not result in an overly optimistic estimation of F_1 scores and area under the curve (AUC) values. Using the PR data set as the test set, we created a separate test set by first randomly withholding 40 positive and 40 negative PR data samples. The remaining data were augmented and balanced, followed by splitting that data into training and validation data sets. We observed that the models achieved high and consistent F_1 and AUC values in training and validation test sets (Appendix Table A1).

Identifying Region of Interest

For each medical record, a region of interest (ROI) was first identified to narrow down the specific text segment that pertained to biomarker information (Fig 1A). Each biomarker (ER, PR, HER2) was identified by using regular expressions and was found to capture nearly all patients for whom the biomarker information could be found. The text segment surrounding the target biomarker (30 characters to the left and 55 characters to the right) was then extracted. The target biomarker name was also homogenized to ER, PR, or HER2 (eg, Прогестерон was changed to PR). Overall, this step is merely an ROI proposal, and we subsequently relied on an NLP algorithm to resolve all variance and ambiguity in discerning the biomarker values. For cases in which multiple ROIs exist in a given medical text, the text was labeled as such and was not used by the subsequent NLP algorithm for training.

The text segment of interest was then preprocessed (Fig 1A). Extra spaces were removed. The obvious delimiter where "–" was in front of or following /[value]/ was removed. An extra space was added after the biomarker word PR, ER, or HER2 to facilitate accurate word embedding (eg, "PRнегат" is changed to "PR негат").

Word Embedding

The preprocessed text segment was tokenized and mapped onto word embeddings (Fig 1B). We used the pretrained fastText embeddings[12] for English and Bulgarian languages, on the basis of data from the free online Wikipedia encyclopedia and the Common Crawl project and trained using the CBOW method with position weights in 300 dimensions, character n-grams of length 5, window size of 5, and 10 negatives.[13] We chose fastText embedding, because unlike Word2Vec[14] or GloVe,[15] it is designed to represent words as character n-grams and thus can effectively derive embedding for out-of-vocabulary words (eg, additional misspelled words that are not in the training corpus). Because the two fastText embedding models for English and Bulgarian were trained independently, words in one language do not necessarily relate by distance to the corresponding translated word in the other vector space. In bilingual models, we transformed the Bulgarian embedding using singular value decomposition onto the English vector space such that the bilingual embeddings were both in the same shared vector space.[16]

The fastText embeddings pretrained on the large corpus enabled incorporation of syntactic and semantic information, which allows words with similar meaning to be close (as measured by cosine similarity) in the high-dimensional embedding space. However, because the objective function was not designed to capture sentiments, words that may be related but have opposite meanings can be close together in the vector space (eg, the words "positive" and "negative"). Thus, for tasks for which the objective is to predict sentiment, model accuracy can be improved by incorporating polarity information for the words. Past studies have tried to incorporate sentiment during the training process of the embedding,[17,18] concatenation of pretrained embedding with additional linguistic features,[19] and refinement of the pretrained embedding.[20] Here we incorporated a polarity one-dimensional vector (Fig 1B). We built the dictionary on the basis of a previous lexicon with known sentiments[21] and manually added the words "plus" and "minus." These added words do not exist in our medical data set and were later used to validate our out-of-vocabulary predictions. The English words were also translated into Bulgarian. Positive words and negative words in this dictionary had the values −1 and 1, respectively. Other words had the value of 0.

Model

We have tested several model architectures, including using CNNs and recurrent neural networks. In this article, we summarized the final optimized models using these two approaches. All text processing was performed using Python, with deep learning models built using the Keras library.

Baseline SVM classifiers were built using scikit-learn with radial basis function as kernel and a penalty parameter C of 0.5. In lieu of a pretrained embedding, the input to the baseline SVM classifiers was word encoded. All text preprocessing, ROI identification, and clean-ups were the same in both baseline and neural network models.

CNN

Our CNN architecture (Fig 1C; Appendix Fig A1) consisted of three separate one-dimensional convolutional layers with 32 filters, stride length of 1, and kernel sizes of 3, 5, and 7. Each was followed by a rectified linear unit (ReLU) as

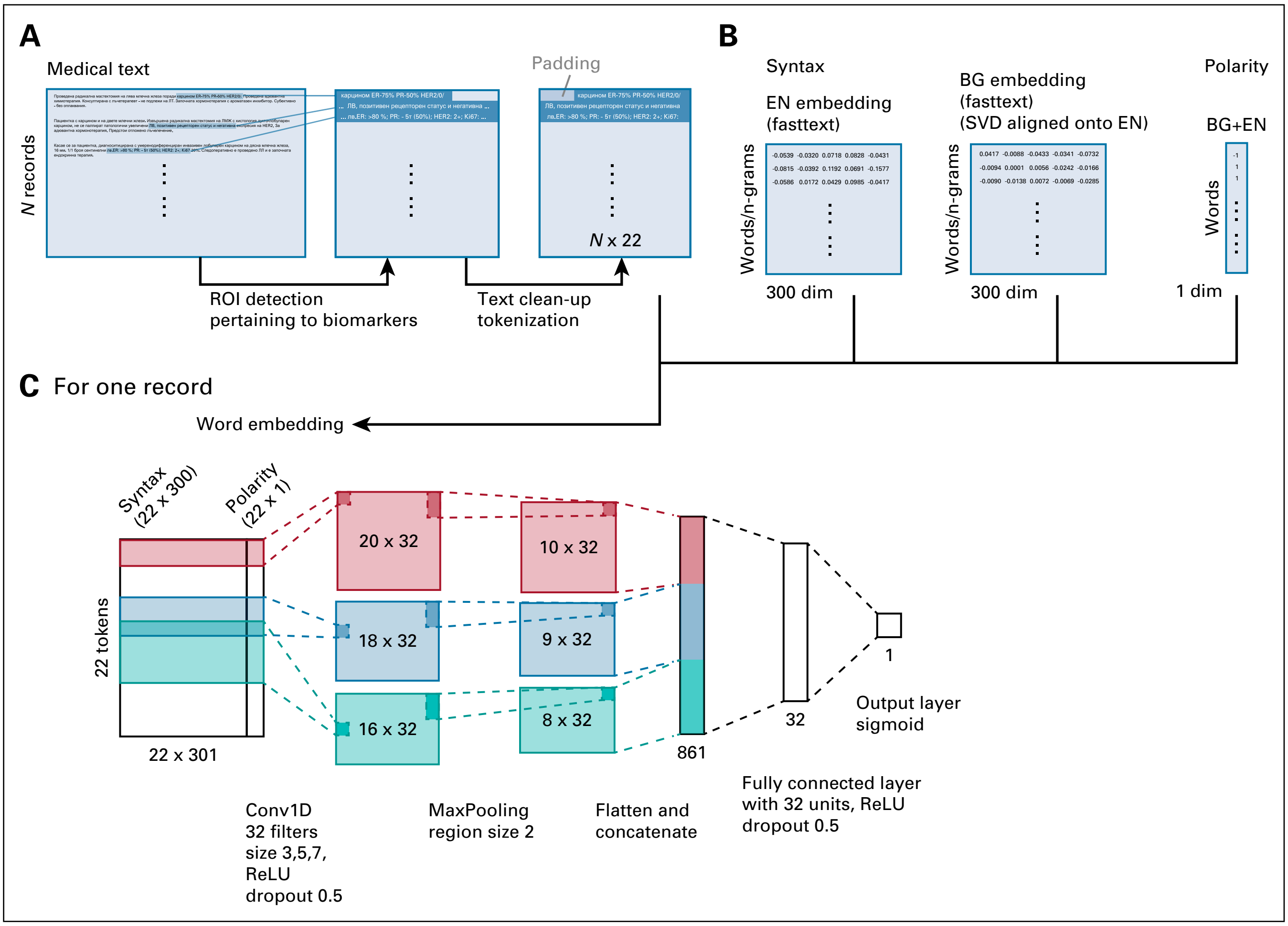

FIG 1. Schematic of the data extraction and normalization pipeline. (A) Unstructured medical text was parsed, and text segments pertaining to biomarker information (region of interest [ROI]) were extracted. The text segment was subsequently cleaned, tokenized, and padded. The medical text in the hematic was modified from the actual patient record for the purposes of illustration. (B) Word embeddings used were fastText embeddings pretrained on Wikipedia and Common Crawl data for English (EN) and Bulgarian (BG). Bulgarian embedding was aligned onto the English embedding vector space using singular value decomposition (SVD). A polarity vector was also used, which contained a known dictionary of words with +1 for positive sentiment and –1 for negative sentiment. (C) The embedded input was a 22 × 301 matrix that was passed onto a convolutional neural network for predicting the positive and negative values. Conv1D, one-dimensional convolutional layer; dim, dimension; ReLU, rectified linear unit.

activation function and dropout (probability, 0.5). The outputs of the convolutional layer were pooled using max pooling and were concatenated. This was followed by a fully connected hidden layer with 32 neurons and ReLU activation, a dropout layer (probability, 0.5), and an output layer with sigmoid activation function. The activities of the hidden and output layer were also regularized with L1 and L2 regularization ($\lambda = 0.001$).

Bidirectional Gated Recurrent Unit

We also examined model architectures containing recurrent units (Appendix Fig A1). Here the input embedding (with padding masked) was passed onto a bidirectional gated recurrent unit,[22] followed by a fully connected hidden layer with 32 neurons, ReLU activation, and an output layer with sigmoid activation function. Dropouts were used following both the gated recurrent unit and fully connected hidden layer, with a dropout rate of 0.5.

Model Training and Evaluation

The data set was split 85%:15% into training and validation sets. The model was trained for 200 epochs, using Adam optimizer and binary cross entropy as the loss function. The loss function results were also weighted according to the proportion of samples with positive and negative samples. The model was evaluated using F_1 score, calculated as,

$$F_1 = 2\frac{\text{precision}\cdot\text{recall}}{\text{precision} + \text{recall}}.$$

The F_1 scores for each label were average weighted by support. Receiver operating characteristic curves (true-positive rate *v* false-positive rate) and AUC values were also calculated.

RESULTS

We evaluated the models using the CNN architecture for the biomarkers PR, ER, and HER2. The model was trained with no indications of overfitting (Fig 2A) and achieved high AUC values (≥ 0.98 for all three biomarkers; Fig 2B) and F_1 scores (≥ 0.98 for all three biomarkers; Table 1). In addition to models with CNN architecture, we also evaluated models with recurrent neural networks and observed comparable performance results (Table 1 "Validation"). In contrast, a baseline classifier on the basis of an SVM was not highly predictive of biomarker status, with F_1 scores of 0.52 to 0.67 and AUC values of 0.53 to 0.64 (Appendix Table A2). We also evaluated our final models on a completely independent test data set (different time periods and including one additional hospital never used in training or validation). The models continued to perform well with F_1 scores of 0.90 or higher for all three biomarkers (Table 1 "Test").

We reran the model on the entire original data set and investigated the reasons for the erroneous predictions. A majority of the results had agreement between actual biomarker label and predicted label (Fig 2C). For those with a mismatch between actual and predicted, we observed that many were actually a result of human error, for which the model predicted with high confidence what the correct biomarker status was to be. In the few cases in which the model was incorrect, the confidence level was also low (Fig 2C).

Next, we examined whether the models were able to extract from out-of-vocabulary (OOV) values. We tried first to see if the model could distinguish potentially conflicting symbols or numeric values (a positive symbol followed by a negative

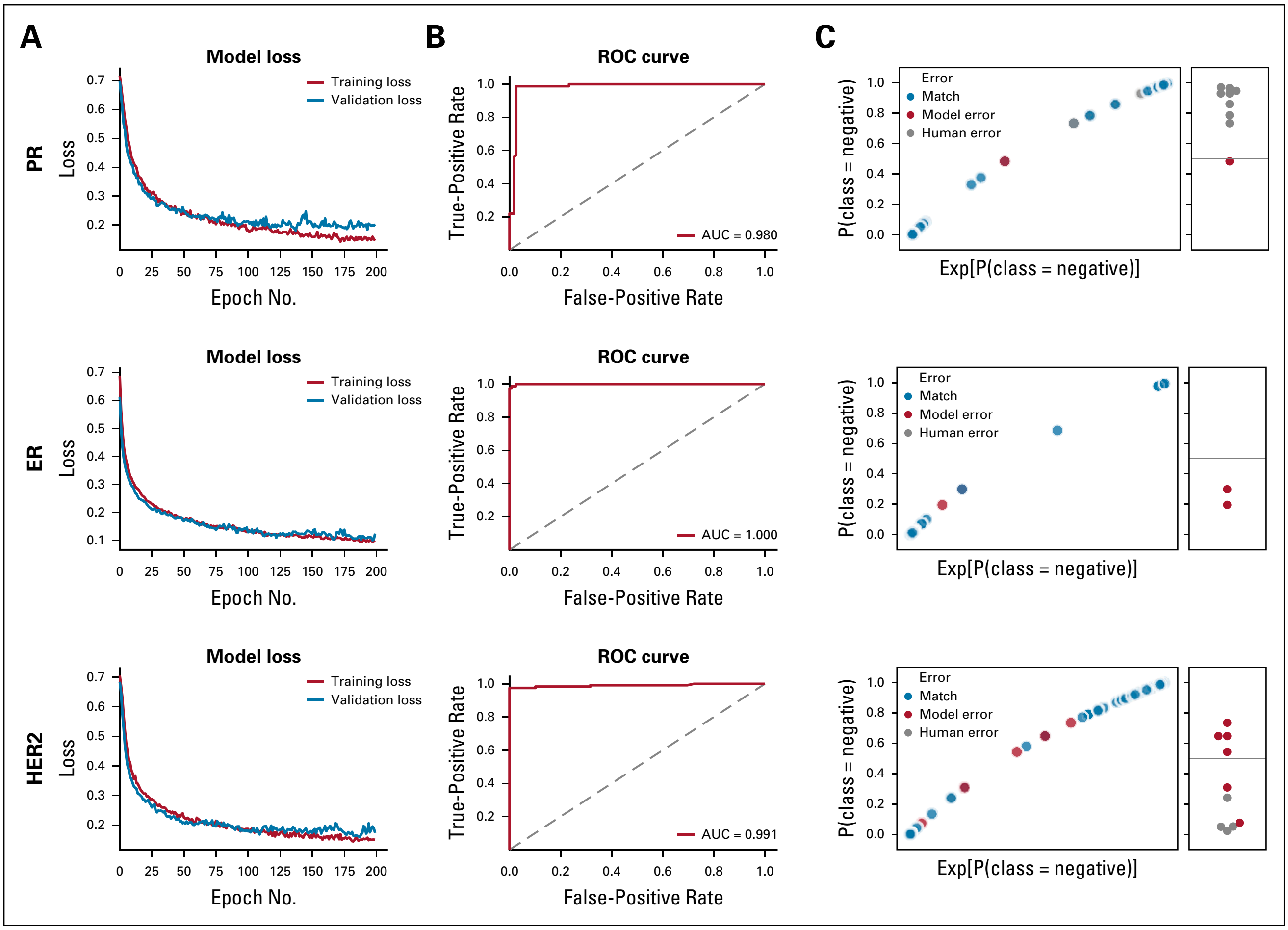

FIG 2. Convolutional neural network (CNN) model evaluation for biomarkers progesterone receptor (PR), estrogen receptor (ER), and human epidermal growth factor receptor 2 (HER2). (A) Training history over 200 epochs. *Y* values are shown as the binary cross entropy loss. (B) Receiver operating characteristic (ROC) curves for the CNN models for the validation data set. (C) Comparison between predicted and actual labels for each biomarker by running the model on the entire original data set. Blue represents a match between predicted and actual labels, red represents discrepancy between predicted and actual and the difference was the result of a model error, and gray represents that such difference was the result of human error. AUC, area under the curve.

TABLE 1. Model Performance for PR, ER, and HER2 as Evaluated by F_1 Score and AUC on the Validation and Test Data Sets

	Validation		Test
Model	**AUC**	**F_1 Score**	**F_1 Score**
PR			
3xConv1D-dense	0.980	0.98	0.95
BiGRU-dense	0.989	0.97	0.93
ER			
3xConv1D-dense	1.000	0.98	1.00
BiGRU-dense	0.998	0.99	0.99
HER2			
3xConv1D-dense	0.991	0.99	0.92
BiGRU-dense	0.994	0.97	0.90

NOTE. Only F_1 scores were reported for the test data set because it was highly imbalanced.

Abbreviations: 3xConv1D, convolutional neural network-based model; AUC, area under the curve; BiGRU, bidirectional gated recurrent unit; ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; PR, progesterone receptor.

numeric value [eg, +0%] or a negative symbol followed by a positive numeric value [eg, –81%]). By using a subset of the data sets containing PR biomarker status, we inserted a small set of OOV test values (ie, +0%, +0, –51%, –81%) in lieu of the original PR value and used that value as input to the models. We observed that the models maintained high AUC values of more than 0.98 (Table 2).

To examine OOV words, we included a set of Bulgarian and English words that mean positive or negative that were never part of the original training set. Although these words are unlikely to occur, we show that it is hypothetically possible for the model to also make correct predictions on the label of the biomarker on the basis of the syntax and polarity embeddings of words never seen during training, achieving 0.84 and 0.94 in AUCs for Bulgarian and English vocabulary, respectively (Table 2). In contrast, models with no polarity embeddings were not substantially better than guessing for these OOV words.

TABLE 2. Model Performance for OOV PR Values Evaluated by F_1 Score and AUC

OOV Input	**Embedding**	**Model**	**F_1 Score**	**AUC**
OOV (symbols)				
Symbols	Syntax + polarity	3xConv1D-dense	0.92	0.988
		BiGRU-dense	0.84	0.990
OOV (word)				
EN	Syntax + polarity	3xConv1D-dense	0.83	0.943
		BiGRU-dense	0.72	0.687
EN	Syntax only	3xConv1D-dense	0.55	0.516
		BiGRU-dense	0.66	0.618
BG	Syntax + polarity	3xConv1D-dense	0.80	0.840
		BiGRU-dense	0.80	0.871
BG	Syntax only	3xConv1D-dense	0.40	0.493
		BiGRU-dense	0.55	0.530

Abbreviations: 3xConv1D, convolutional neural network-based model; AUC, area under the curve; BG, Bulgarian; BiGRU, bidirectional gated recurrent unit; EN, English; OOV, out-of-vocabulary [values]; PR, progesterone receptor.

DISCUSSION

One of the major barriers in the adaptation of artificial intelligence (AI) into medicine is the tremendous volume of unstructured text that exists in EHRs. We demonstrate here that using a deep-learning–based NLP algorithm, we can achieve high-performance (AUC $\geq$ 0.98) for the extraction and normalization of biomarker values in EHRs containing mixed languages and with substantial heterogeneity in the target parameter positions and values.

In addition, we found that incorrect high-confidence predictions were mainly a result of human errors, and those with model errors were restricted to those with lower-confidence scores. The ability to achieve high accuracy despite human errors in the training data set highlights the robustness of the approach. It is worth noting that, aside from manual annotation errors, other errors are also known to occur in medical records and registries,[23] with error rates that can be as high as 27%. Here, the final models can be incorporated into an overall pipeline to facilitate an automated extraction and normalization of biomarker values. Low-confidence predictions can be marked for subsequent manual validation.

The approach described here is also applicable to other mixed-language medical text data (eg, EHRs in Russian, Ukrainian, or Serbian). The corresponding embedding for those languages can be trained independently and can be aligned to the English embedding vector space. We found that the broad pretrained embeddings were sufficient for the purposes of biomarker classification, but the embeddings can be improved for other tasks. The pretraining of embeddings can be continued with specific medical text for the incorporation of medical syntactic information. Embeddings for polarity can also be extended to include additional words in the lexicon. It is worth noting that state-of-the-art language models have recently moved toward context-specific embeddings (eg, ELMo) and novel self-attention–based neural network architectures (eg, transformers, openAI transformer, BERT). Although these models are starting to be used in tasks such as language translation and named entity recognition, we found that our proposed approach was sufficient for the extraction and normalization described. For parameters with more challenging ROI identifications, algorithms on the basis of character or word embedding and bidirectional recurrent neural networks commonly used in named entity recognition may be used for ROI proposals.

Data extraction from an unstructured data source is critical for clinical research, for identifying eligible patients for clinical trial enrollments, and for monitoring treatment outcomes for value-based care not only for oncology, but across all fields of medicine. Undoubtedly, there are many other challenging tasks, such as deriving drug treatment durations and therapy effectiveness. Approaches such as the one presented here are critical in automating data extraction and improving the accuracy of multilingual EHRs.

AFFILIATION

[1]Sqilline Health, Boston, MA

CORRESPONDING AUTHOR

Boyang Zhao, PhD, Sqilline Health, 50 Milk St, 16th Floor, Boston, MA 02109; Twitter: @boyangzhao; e-mail: boyang.zhao@sqillinehealth.com.

AUTHOR'S DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.

Boyang Zhao
Employment: Sqilline Health
Leadership: Sqilline Health
Stock and Other Ownership Interests: Sqilline Health

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

Thanks to Chavdar Grudev and the R&D team for data preparation and helpful comments on the manuscript.

REFERENCES

1. Lee J, Song HJ, Yoon E, et al: Automated extraction of biomarker information from pathology reports. BMC Med Inform Decis Mak 18:29, 2018
2. Buckley JM, Coopey SB, Sharko J, et al: The feasibility of using natural language processing to extract clinical information from breast pathology reports. J Pathol Inform 3:23, 2012
3. Breischneider C, Zillner S, Hammon M, et al: Automatic extraction of breast cancer information from clinical reports. IEEE 30th International Symposium on Computer-Based Medical Systems, Thessaloniki, Greece, June 22-24, 2017
4. Tang R, Ouyang L, Li C, et al: Machine learning to parse breast pathology reports in Chinese. Breast Cancer Res Treat 169:243-250, 2018
5. Yala A, Barzilay R, Salama L, et al: Using machine learning to parse breast pathology reports. Breast Cancer Res Treat 161:203-211, 2017
6. Wieneke AE, Bowles EJ, Cronkite D, et al: Validation of natural language processing to extract breast cancer pathology procedures and results. J Pathol Inform 6:38, 2015
7. Weegar R, Dalianis H: Creating a rule based system for text mining of Norwegian breast cancer pathology reports. Proc Sixth International Workshop on Health Text Mining and Information Analysis, Lisbon, Portugal, September 17, 2015, Association for Computational Linguistics, 73-78
8. Boytcheva S, Angelova G, Angelov Z, et al: Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. Cybern Inf Technol 15:58-77, 2015
9. Boytcheva S, Nikolova I, Paskaleva E, et al: Obtaining status descriptions via automatic analysis of hospital patient records. Informatica 34:269-278, 2010
10. Boytcheva S: Shallow medication extraction from hospital patient records. Stud Health Technol Inform 166:119-128, 2011
11. Qiu JX, Yoon HJ, Fearn PA, et al: Deep learning for automated extraction of primary sites from cancer pathology reports. IEEE J Biomed Health Inform 22:244-251, 2018
12. Bojanowski P, Grave E, Joulin A, et al: Enriching word vectors with subword information. Trans Assoc Comput Linguist 5:135-146, 2017
13. Grave E, Bojanowski P, Gupta P, et al: Learning word vectors for 157 languages. Proc 11th Language Resource Evaluation Conference, Miyazaki, Japan, May 9-11, 2018, Association for Computational Linguistics https://arxiv.org/pdf/1802.06893.pdf
14. Mikolov T, Chen K, Corrado G, et al: Efficient estimation of word representations in vector space. arXiv 1-12, 2013 https://arxiv.org/pdf/1301.3781.pdf
15. Pennington J, Socher R, Manning C: Glove: Global vectors for word representation. Proc 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, October 25-29, 2014, Association for Computational Linguistics, 1532-1543 https://www.aclweb.org/anthology/D14-1162
16. Smith SL, Turban DHP, Hamblin S, et al: Offline bilingual word vectors, orthogonal transformations and the inverted softmax. ICLR 1-10, 2017 https://arxiv.org/pdf/1702.03859.pdf
17. Labutov I, Lipson H: Re-embedding words. Proc 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, August 4-9, 2013, Association for Computational Linguistics, 489-493, 2013 https://igorlabutov.com/static/papers/re.pdf
18. Maas AL, Daly RE, Pham PT, et al: Learning word vectors for sentiment analysis. Proc 49th Annual Meeting of the Association for Computational Linguistics, Portland, OR, June 19-24, 2011, Association for Computational Linguistics, 142-150, https://www.aclweb.org/anthology/P11-1015
19. Rezaeinia SM, Rahmani R, Ghodsi A, et al: Sentiment analysis based on improved pre-trained word embeddings. Expert Syst Appl 117:139-147, 2019
20. Yu LC, Wang J, Lai KR, et al: Refining word embeddings for sentiment analysis. Proc 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, September 7-11, 2017, Association for Computational Linguistics, 534-539 https://aclweb.org/anthology/D17-1056
21. Hu M, Liu B: Mining and summarizing customer reviews. Proc ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, WA, August 22-25, 2004 https://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf
22. Cho K, van Merrienboer B, Gulcehre C, et al: Learning phrase representations using RNN encoder-decoder for statistical machine translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, October 25-29, 2014
23. Goldberg SI, Niemierko A, Turchin A: Analysis of data errors in clinical research databases. AMIA Annu Symp Proc 242-246, 2008

APPENDIX

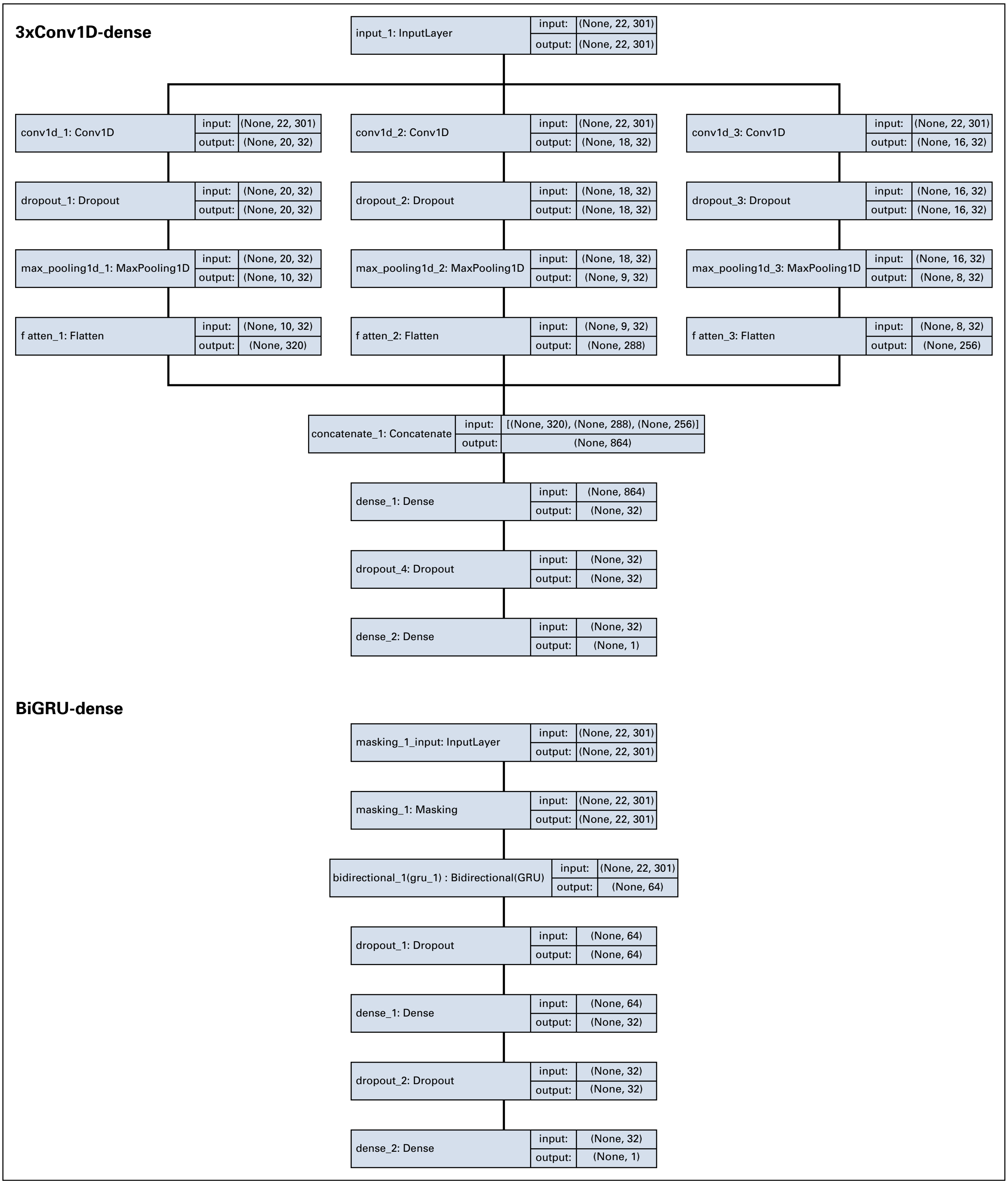

FIG A1. Model architecture for convolutional neural network (CNN)–based model (3xConv1D-dense) and GRU-based model (bidirectional gated recurrent unit [BiGRU]-dense); Conv1D, one-dimensional convolutional layer.

TABLE A1. Model Performance for PR Training, Validation, and Test Data Sets Evaluated by F_1 Score and AUC

Model	F_1 Score	AUC
Training		
3xConv1D-dense	0.99	1.000
BiGRU-dense	0.99	1.000
Validation		
3xConv1D-dense	0.98	0.977
BiGRU-dense	0.96	0.979
Test		
3xConv1D-dense	0.97	0.999
BiGRU-dense	0.96	0.997

Abbreviations: 3xConv1D, convolutional neural network-based model; AUC, area under the curve; BiGRU, bidirectional gated recurrent unit; PR, progesterone receptor.

TABLE A2. Model Performance for Baseline SVM Classifier for PR, ER, and HER2 Evaluated by F_1 Score and AUC on Training and Validation Data Sets

Marker	Model	F_1 Score (training)	F_1 Score (validation)	AUC (validation)
ER	SVM	0.67	0.52	0.525
PR	SVM	0.66	0.53	0.550
HER2	SVM	0.65	0.67	0.641

Abbreviations: AUC, area under the curve; ER, estrogen receptor; HER2, human epidermal growth factor receptor 2; PR, progesterone receptor; SVM, support vector machine.