Towards Generalizable Methods for Automating Risk Score Calculation

Jennifer J. Liang[1,3,⋆], Eric Lehman[2,3,#], Ananya S. Iyengar[4,§], Diwakar Mahajan[1,3,†], Preethi Raghavan[2,5,‡], Cindy Y. Chang[6,*] Peter Szolovits[2,3,◇]

[1]IBM Research, [2]MIT CSAIL, [3]MIT-IBM Watson AI Lab, [4]Northeastern University, [5]Fidelity Investments, [6]Brigham and Women's Hospital, Harvard Medical School

{⋆jjliang,†dmahaja}@us.ibm.com, {#lehmer16,◇psz}@mit.edu, §iyengar.a@northeastern.edu, ‡preethi.raghavan@fmr.com, *cchang@bwh.harvard.edu

Abstract

Clinical risk scores enable clinicians to tabulate a set of patient data into simple scores to stratify patients into risk categories. Although risk scores are widely used to inform decision-making at the point-of-care, collecting the information necessary to calculate such scores requires considerable time and effort. Previous studies have focused on specific risk scores and involved manual curation of relevant terms or codes and heuristics for each data element of a risk score. To support more generalizable methods for risk score calculation, we annotate 100 patients in MIMIC-III with elements of CHA_2DS_2-VASc and PERC scores, and explore using question answering (QA) and off-the-shelf tools. We show that QA models can achieve comparable or better performance for certain risk score elements as compared to heuristic-based methods, and demonstrate the potential for more scalable risk score automation without the need for expert-curated heuristics. Our annotated dataset will be released to the community to encourage efforts in generalizable methods for automating risk scores.

1 Introduction

Clinical risk scores are standardized metrics to estimate the risk of a particular future outcome based on available clinical parameters and are commonly used at the point-of-care to inform decision-making around diagnosis and treatment (Steyerberg et al., 2019). An example of this is the CHA_2DS_2-VASc score (Lip et al., 2010), which uses 7 patient data elements to estimate the risk of stroke in patients with non-valvular atrial fibrillation and thus guide strategies around stroke prevention. It has successfully demonstrated clinical impact and is referenced in the practice guidelines for management of atrial fibrillation released by the American Heart Association, American College of Cardiology, and Heart Rhythm Society in 2014 (January et al., 2014). In general, data elements that contribute to a risk score may include information about the patient's age, gender, medical history, presenting symptoms, medication use, etc. While risk scores are generally designed for use at the point-of-care, calculating them can require considerable time and effort, as each data element must be manually gathered, often from multiple locations within the electronic health record (EHR). A previous study investigating the feasibility of automating clinical score calculation identified 534 unique patient data elements from 168 externally validated clinical scores, with each score requiring anywhere from 3 to 31 elements (Aakre et al., 2017). Automating extraction of clinical data elements necessary to calculate risk scores could save clinicians time and help them more effectively leverage risk scores to improve care at the bedside (Aakre et al., 2017).

Prior efforts to automate data extraction for risk score calculations have targeted specific risk scores. Some of these efforts focused only on leveraging information from structured EHR data. Koziatek et al. (2018) developed and automated a structured-data-only version of the Wells and revised Geneva risk scores for estimating pulmonary embolism (PE) risk. Similarly, in automating the Padua Prediction Score for venous thromboembolism risk, Pavon et al. (2018) either operationalized variables to rely only on structured data or omitted them entirely. Other efforts have also incorporated unstructured EHR data into their work. Jonnagaddala et al. (2015) used a rule-based text mining system to extract elements of the Framingham risk score for coronary artery disease. Mark et al. (2018) and Zhang et al. (2022) used text string searches on a set of custom-built keywords/search phrases to automate coronary risk scores and Wells score for PE, respectively. Bean et al. (2019), Grouin et al. (2011), and Elkin et al. (2021) explored the use of named entity recognition (NER) tagging combined with heuristics to automatically calculate

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 426–431

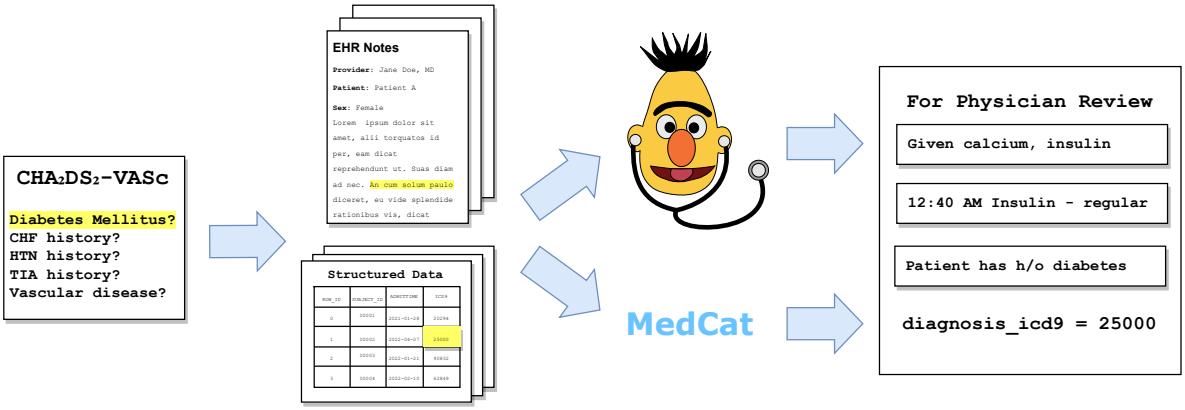

Figure 1: Demonstration of how our proposed QA-based risk score analysis system would work in conjunction with a physician.

the CHA_2DS_2-VASc score over unstructured EHR data. While these efforts found strong agreement with expert human evaluators, heuristic-based approaches are often rigid and struggle to generalize to other problems. Thus, we propose using off-the-shelf tools and pretrained language models to extract evidence from both structured and unstructured EHR data, without the need for manually-curated rules.

In this study, we explore two commonly used risk scores – the CHA_2DS_2-VASc score for atrial fibrillation stroke risk (Lip et al., 2010) and the Pulmonary Embolism Rule-out Criteria (PERC) rule (Kline et al., 2004) – to demonstrate our approach. We use a transformer-based model trained on emrQA (Pampari et al., 2018) and an off-the-shelf biomedical ontology linker paired with a SQL query component to extract evidence from unstructured and structured EHR data, respectively, for each element of the two risk scores (Figure 1). The main contributions of this work are:

- the first community-shared dataset based on MIMIC-III for automating risk scores,
- a demonstration of the potential for off-the-shelf tools and QA models to automate risk scores over heuristics and rules,
- the need for better negation/hypothetical detection and clinical knowledge embeddings.

2 Dataset

To evaluate our models, we randomly sample 100 patients from the Medical Information Mart for Intensive Care III (MIMIC-III) dataset (Johnson et al., 2016) to annotate with elements of CHA_2DS_2-VASc and PERC. CHA_2DS_2-VASc uses 7 patient data elements to estimate the risk of stroke in patients with non-valvular atrial fibrillation: congestive heart failure (CHF), hypertension, age, diabetes mellitus, stroke/transient ischemic attack (TIA)/thromboembolism (TE), vascular disease (prior myocardial infarction, peripheral artery disease, or aortic plaque), and sex. PERC uses 8 elements to evaluate the risk of PE in low-risk patients: age, heart rate, oxygen saturation, unilateral leg swelling, hemoptysis, recent surgery or trauma, prior PE or deep venous thrombosis (DVT), and hormone use.

We frame our scenario as a new patient being seen in the emergency department (ED) requiring calculation of CHA_2DS_2-VASc or PERC because of suspected atrial fibrillation or PE, respectively, and the data in MIMIC-III is the available past medical history for this patient. Therefore, we limited our dataset to non-expired patients at least 18 years of age at time of last discharge with at least one discharge summary. Since PERC only rules out PE when none of the criteria are met, one of which is age ≥ 50, we further adjust our sampling such that at least half of the patients selected are under 50 years of age at time of last discharge to ensure non-trivial calculation of PERC.

The dataset was annotated by two independent annotators, with a 20% overlap for inter-annotator agreement ($\kappa = 0.800$), and then reviewed by a physician. Annotators reviewed the entire EHR data provided in MIMIC-III, including both structured and unstructured sources, and annotated evidence relevant to each risk score element. Evidence in structured data include coded diagnoses, procedures, and past medical history. Evidence in unstructured data consist of text snippets from

Patient	Risk Score	Element	Evidence Source	Evidence Text	Answer
1234	CHA_2DS_2-VASc	CHF	noteevents.row_id = xxxx noteevents.row_id = xxxx diagnoses_icd.icd9_code = 4280	88 yo M with h/o dCHF Findings compatible with moderate congestive heart failure, with interval worsening since [**2157-8-30**] Congestive heart failure, unspecified	yes
1234	CHA_2DS_2-VASc	Hypertension	noteevents.row_id = xxxx diagnoses_icd.icd9_code = 4019	Hypertension Unspecified essential hypertension	yes
1234	CHA_2DS_2-VASc	Stroke/TIA/TE	NA		no data
1234	CHA_2DS_2-VASc	Vascular disease	chartevents.value = CAD	CAD	yes
1234	PERC	Hemoptysis	NA		no data
1234	PERC	Recent surgery/trauma	noteevents.row_id = xxxx noteevents.row_id = xxxx	s/p Pedestrian struck by auto presented to an outside hospital after reportedly being struck by a car traveling at 35mph	yes

Table 1: Example of annotated dataset. Under Evidence Source, NA indicates not applicable because no evidence found, noteevents indicates unstructured EHR data (xxxx indicates elided data), and all other sources are considered structured EHR data.

discharge summaries, admission notes, progress notes, and their addenda. Patients in our subset had an average of 44 notes with average length of 289 tokens.

Since we frame our scenario as a new patient being seen in the ED, vital signs (e.g., heart rate, oxygen saturation) as recorded in their history (i.e., MIMIC-III in our scenario) would not be relevant and are therefore excluded from annotation. For other elements in PERC that may also be time-sensitive, since the exact time frame is not always apparent from the given documentation, for the purposes of this study, we annotate all instances of unilateral leg swelling, hemoptysis, surgery and trauma as evidence for their respective elements regardless of when they occurred. In addition to the evidence, annotators also provided an overall answer for each risk score element: `"Yes"`, `"No"`, `"Unclear"` (evidence present but conflicting or inconclusive), or `"No data"`. A sample of the annotated dataset is presented in Table 1.

3 Task Setup

To extract information relevant to the specified risk score, we query the system with risk score elements expressed as short natural language phrases containing the entities (e.g., "hypertension"). Elements containing multiple concepts are split into multiple phrases, each containing a single concept. For the purposes of evaluation, `"Yes"` and `"Unclear"` in the ground truth are considered to be equivalent because both provide some positive evidence, while `"No"` and `"No data"` are considered to be equivalent because in practice, lack of data would be presumed to be negative.

For unstructured data, a system is tasked with predicting the presence or absence of the given risk score element. The system must also provide the sentence it selected to make its decision. Predictions considered true positives when compared to the ground truth are further reviewed by a physician to ensure that the sentence used for prediction can reasonably be used to determine if the patient has the given condition; if the sentence used for prediction cannot be used to logically determine whether the patient has the given condition, the prediction is marked as a false positive.

For structured data, the model is tasked with retrieving a `Yes/No` answer along with the relevant billing code (when present) for each risk score element. We evaluate the system by matching the retrieved `Yes/No` answer with the ground truth, and calculating the precision, recall, and F1-score.

4 Models

4.1 Structured Data Information Retrieval

To extract answers from structured EHR data, we employ a two step process. We (1) use MedCAT[1] (Kraljevic et al., 2019), an off-the-shelf biomedical ontology linker, to curate a set of Concept Unique Identifiers (CUIs) for each risk score element, which are then mapped to institution-specific billing codes (here, ICD9 for MIMIC-III) using the Unified Medical Language System (UMLS) APIs[2] (Bodenreider, 2004), and then (2) use these element-specific code-sets to form SQL queries (derived from emrKBQA (Raghavan et al., 2021)) to retrieve answers, i.e., `Yes/No` marked by the presence/absence of element-specific codes for a patient in the structured data. We evaluate our output only against risk score elements with answers from structured data (i.e., Evidence Source $\neq$ noteevents). Results are presented in Table 2.

[1] https://github.com/CogStack/MedCAT

[2] https://documentation.uts.nlm.nih.gov/rest/home.html

Risk Score	Element	Count	R	P	F1
CHA_2DS_2-VASc	CHF	16	1.0	0.94	0.97
CHA_2DS_2-VASc	Hypertension	43	0.97	0.81	0.89
CHA_2DS_2-VASc	Stroke/TIA/TE	17	1.0	0.12	0.21
CHA_2DS_2-VASc	Vascular disease	27	0.92	0.44	0.60
CHA_2DS_2-VASc	Diabetes mellitus	18	0.87	0.72	0.79
CHA_2DS_2-VASc	Overall	121	0.95	0.64	0.76
PERC	Unilateral leg swelling	5	0	0	0
PERC	Hemoptysis	1	0	0	0
PERC	Recent surgery/trauma	79	0	0	0
PERC	Prior PE/DVT	8	1.0	0.38	0.55
PERC	Hormone use	0	NA	NA	NA
PERC	Overall	93	0.84	0.17	0.29

Table 2: Performance of the structured data information retrieval component. We only calculate performance on risk score elements with structured data answers in the ground truth.

4.2 Baseline Model

To ground the results of our QA model, we implement a NER-based approach based on Bean et al. (2019). We use MedCAT to tag CUIs in the notes. We then return the top sentence that contains relevant affirmed CUIs based on the MedCAT negation detection system. Bean et al. (2019) defines a set of CUIs with respect to CHA_2DS_2-VASc. However, there is no such definition for PERC. We thus find relevant CUIs for the main categories of PERC (e.g., hormone use, surgery, etc.) and use all possible descendants of the selected CUIs based on the UMLS hierarchy. Results are shown in Table 3.

4.3 Unstructured QA Model

To retrieve relevant information from unstructured EHR data, we use ClinicalBERT (Alsentzer et al., 2019; Devlin et al., 2019), a transformer-based model pretrained on MIMIC-III. We sample 5% of the data[3] from the medication, relations, and risk subsections and train on emrQA (Pampari et al., 2018). Due to the vast number of notes likely containing irrelevant information, we additionally negative sample (1:1 ratio) unanswerable questions from other notes in emrQA during training. Further, due to the vague elements often used in risk scores (e.g., recent surgery or trauma), we augment 20% of existing emrQA questions containing a clinical entity to instead contain its parent MeSH[4] hierarchy entity. Similar to Bean et al. (2019), we select model predicted relevant spans and use MedCAT's negation detector to determine whether or not the patient has the given risk score element.

We additionally show how performance improves when unstructured data predictions are paired with structured data ones. To combine unstructured and structured data, we use the structured data prediction if it predicts that the patient has the given condition, otherwise we default to the selected unstructured data answer. All results are presented in Table 3.

5 Discussion

We make a few key observations. We find that the structured data model is able to achieve extremely high performance in a number categories, but unable to find any relevant information for the rest. We hypothesize that this is due to chronic conditions (e.g., CHF, hypertension) being more consistently recorded in the structured data, while acute events (e.g., PE/DVT, stroke/TIA/TE) are coded only in the limited time frame when such conditions are being actively managed. Also, structured data, in the form of billing codes, would not be expected to capture symptoms without a formal diagnosis (e.g., unilateral leg swelling). We additionally find that the QA model on unstructured data alone is able to improve on the results of Bean et al. (2019) on a number of categories, without the need for expert-crafted heuristics. However, we find that the QA model struggles due to a lack of clinical knowledge and ability to distinguish hypothetical mentions versus true affirmations of the given condition.

With respect to vascular disease, an error analysis of the QA-based model showed that 69% of the false positives were due to a lack of clinical understanding, as the model considered a much broader definition of vascular disease than the one specified in the CHA_2DS_2-VASc score. Similarly, with respect to stroke/TIA/TE, we find that 93% of the false positives can be attributed to imprecise understanding of medical terminology and the model's inability to use contextual clues to differentiate between stroke and other conditions. We additionally see extremely low precision for PE/DVT. This can largely be attributed to faulty negation detection, as MedCAT often fails to distinguish between affirming and hypothetical/negated mentions in over 70% of the false positives.

One issue we found when implementing Bean et al. (2019)'s approach is that it is nontrivial to determine which CUIs to select, specifically for general categories like surgery and trauma. Using all UMLS descendants of surgery and trauma results in 3,413,446 unique CUIs, which will clearly result in an enormous number of false positive re-

[3] Yue et al. (2020) found that sampling 5% of the data was equivalent to training on the entire dataset.

[4] https://www.nlm.nih.gov/mesh/meshhome.html

		Model	Bean et al. (2019)			QA					
		Data	Unstructured			Unstructured			Structured + Unstructured		
Risk Score	Element	Support	R	P	F1	R	P	F1	R	P	F1
CHA_2DS_2-VASc	CHF	16	0.385	0.294	0.333	0.615	0.533	0.571	0.938	0.789	0.857
CHA_2DS_2-VASc	Hypertension	43	0.929	0.736	0.821	0.883	0.864	0.874	0.977	0.875	0.923
CHA_2DS_2-VASc	Stroke/TIA/TE	17	0.588	0.303	0.400	0.385	0.263	0.312	0.538	0.333	0.412
CHA_2DS_2-VASc	Vascular disease	27	0.423	0.846	0.564	0.810	0.250	0.382	0.870	0.290	0.435
CHA_2DS_2-VASc	Diabetes mellitus	18	0.818	0.167	0.277	0.667	0.667	0.667	0.833	0.652	0.732
CHA_2DS_2-VASc	Overall	121	0.679	0.435	0.530	0.741	0.488	0.588	0.876	0.550	0.676
PERC	Unilateral leg swelling	5	0.200	1.000	0.333	0.500	0.375	0.429	0.500	0.375	0.429
PERC	Hempoptysis	1	1.000	0.250	0.400	1.000	0.118	0.211	1.000	0.118	0.211
PERC	Recent surgery/trauma	79	0.750	0.030	0.058	0.397	0.610	0.481	0.397	0.610	0.481
PERC	Prior PE/DVT	8	0.714	0.161	0.263	0.750	0.064	0.118	0.833	0.106	0.189
PERC	Hormone use	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
PERC	Overall	93	0.611	0.078	0.138	0.440	0.270	0.335	0.455	0.287	0.352

Table 3: Performance of Bean et al. (2019) heuristics, QA model, and a combination structured and QA model predictions.

sults when selecting sentences, as seen in Table 3. We find that the QA-based approach significantly outperforms the Bean et al. (2019)-based approach with respect to identifying surgery/trauma. This suggests that QA may offer a solution for these more general categories.

6 Conclusion

We explore risk score automation using QA and off-the-shelf ontology entity linkers without the need for expert-curated rules, and demonstrate its potential for easy adaptation to unexplored risk scores. We find that QA models can achieve comparable or better performance for certain risk score elements as compared to heuristic-based methods, and demonstrate the potential for more scalable risk score automation without the need for expert-curated heuristics. Our annotated dataset will be released to the community to encourage efforts in generalizable methods for automating risk scores.

References

Christopher Aakre, Mikhail Dziadzko, Mark T Keegan, and Vitaly Herasevich. 2017. Automating clinical score calculation within the electronic health record. a feasibility assessment. *Appl. Clin. Inform.*, 8(2):369–380.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Dan Bean, James Teo, Honghan Wu, Ricardo Oliveira, Raj Patel, Rebecca Bendayan, Ajay Shah, Richard Dobson, and Paul Scott. 2019. Semantic computational analysis of anticoagulation use in atrial fibrillation from real world data. *PLOS ONE*, 14:e0225625.

Olivier Bodenreider. 2004. The unified medical language system (umls): Integrating biomedical terminology. *Nucleic acids research*, 32:D267–70.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter L Elkin, Sarah Mullin, Jack Mardekian, Christopher Crowner, Sylvester Sakilay, Shyamashree Sinha, Gary Brady, Marcia Wright, Kimberly Nolen, Joann Trainer, Ross Koppel, Daniel Schlegel, Sashank Kaushik, Jane Zhao, Buer Song, and Edwin Anand. 2021. Using artificial intelligence with natural language processing to combine electronic health record's structured and free text data to identify nonvalvular atrial fibrillation to decrease strokes and death: Evaluation and case-control study. *J. Med. Internet Res.*, 23(11):e28946.

Cyril Grouin, Louise Deléger, Arnaud Rosier, Lynda Temal, Olivier Dameron, Pascal van Hille, Anita Burgun, and Pierre Zweigenbaum. 2011. Automatic computation of cha2ds2-vasc score: Information extraction from clinical texts for thromboembolism risk assessment. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:501–10.

Craig T. January, L. Samuel Wann, Joseph S. Alpert, Hugh Calkins, Joaquin E. Cigarroa, Joseph C. Cleveland, Jamie B. Conti, Patrick T. Ellinor, Michael D. Ezekowitz, Michael E. Field, Katherine T. Murray, Ralph L. Sacco, William G. Stevenson, Patrick J. Tchou, Cynthia M. Tracy, and Clyde W. Yancy. 2014. 2014 aha/acc/hrs guideline for the management of patients with atrial fibrillation: A report of the american college of cardiology/american heart association task force on practice guidelines and the heart rhythm society. *Journal of the American College of Cardiology*, 64(21):e1–e76.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Sci. Data*, 3(1):160035.

Jitendra Jonnagaddala, Siaw-Teng Liaw, Pradeep Ray, Manish Kumar, Nai-Wen Chang, and Hong-Jie Dai. 2015. Coronary artery disease risk assessment from unstructured electronic health records using text mining. *Journal of Biomedical Informatics*, 58:S203–S210. Supplement: Proceedings of the 2014 i2b2/UTHealth Shared-Tasks and Workshop on Challenges in Natural Language Processing for Clinical Data.

J A Kline, A M Mitchell, C Kabrhel, P B Richman, and D M Courtney. 2004. Clinical criteria to prevent unnecessary diagnostic testing in emergency department patients with suspected pulmonary embolism. *J. Thromb. Haemost.*, 2(8):1247–1255.

Christian A Koziatek, Emma Simon, Leora I Horwitz, Danil V Makarov, Silas W Smith, Simon Jones, Soterios Gyftopoulos, and Jordan L Swartz. 2018. Automated pulmonary embolism risk classification and guideline adherence for computed tomography pulmonary angiography ordering. *Acad. Emerg. Med.*, 25(9):1053–1061.

Zeljko Kraljevic, Daniel M Bean, Aurelie Mascio, Lukasz Roguski, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, and Richard J. B. Dobson. 2019. Medcat - medical concept annotation tool. *ArXiv*, abs/1912.10166.

Gregory Y H Lip, Robby Nieuwlaat, Ron Pisters, Deirdre A Lane, and Harry J G M Crijns. 2010. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, 137(2):263–272.

Dustin G. Mark, Jie Huang, Uli Chettipally, Mamata V. Kene, Megan L. Anderson, Erik P. Hess, Dustin W. Ballard, David R. Vinson, and Mary E. Reed. 2018. Performance of coronary risk scores among patients with chest pain in the emergency department. *Journal of the American College of Cardiology*, 71(6):606–616.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Juliessa M Pavon, Richard J Sloane, Carl F Pieper, Cathleen S Colón-Emeric, Harvey J Cohen, David Gallagher, Miriam C Morey, Midori McCarty, Thomas L Ortel, and Susan N Hastings. 2018. Automated versus manual data extraction of the padua prediction score for venous thromboembolism risk in hospitalized older adults. *Appl. Clin. Inform.*, 9(3):743–751.

Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, Rachita Chandra, and Peter Szolovits. 2021. emrKBQA: A clinical knowledge-base question answering dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73, Online. Association for Computational Linguistics.

Ewout W Steyerberg et al. 2019. *Clinical prediction models*. Springer.

Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: A thorough analysis of the emrqa dataset. *ArXiv*, abs/2005.00574.

Nasen Jonathan Zhang, Philippe Rameau, Marsophia Julemis, Yan Liu, Jeffrey Solomon, Sundas Khan, Thomas McGinn, and Safiya Richardson. 2022. Automated pulmonary embolism risk assessment using the wells criteria: Validation study. *JMIR Form Res*, 6(2):e32230.