Literature-Augmented Clinical Outcome Prediction

Aakanksha Naik [1*] **Sravanthi Parasa**[2] **Sergey Feldman**[3] **Lucy Lu Wang**[3] **Tom Hope**[3,4]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Swedish Medical Group [3]Allen Institute for Artificial Intelligence
[4] Paul G. Allen School for Computer Science & Engineering, University of Washington
anaik@andrew.cmu.edu {tomh,lucyw,sergey}@allenai.org

Abstract

We present BEEP (Biomedical Evidence-Enhanced Predictions), a novel approach for clinical outcome prediction that retrieves patient-specific medical literature and incorporates it into predictive models.[1] Based on each individual patient's clinical notes, we train language models (LMs) to find relevant papers and fuse them with information from notes to predict outcomes such as in-hospital mortality. We develop methods to retrieve literature based on noisy, information-dense patient notes, and to augment existing outcome prediction models with retrieved papers in a manner that maximizes predictive accuracy. Our approach boosts predictive performance on three important clinical tasks in comparison to strong recent LM baselines, increasing F1 by up to 5 points and precision@Top-K by a large margin of over 25%.

1 Introduction

Predicting the medical outcomes of hospitalized patients holds the promise of enhancing clinical decision making. With the advent of electronic health records (EHRs), more clinical data has become available to train AI models for outcome prediction (Rajkomar et al., 2018; Hashir and Sawhney, 2020). In particular, language models pretrained on biomedical and/or clinical text are demonstrating increasing proficiency when fine-tuned for the task of predicting outcomes such as in-hospital mortality or length of stay (van Aken et al., 2021).

In this work, we explore a novel approach for improving clinical outcome prediction by dynamically retrieving relevant medical literature for each patient, and incorporating this literature into language models (LMs) trained for outcome prediction from clinical notes. This is in contrast to existing outcome prediction work that uses *only* clinical notes (Boag et al., 2018; Hashir and Sawhney, 2020). Recent LM-based approaches van Aken et al. (2021) have designed pretraining schemes over corpora of clinical notes and *general* biomedical literature. This is in contrast to our work, where we directly incorporate a literature retrieval mechanism into our outcome prediction model, by finding papers relevant to *specific* patient cases. Our approach, named BEEP (Biomedical Evidence-Enhanced Predictions), is broadly inspired by Evidence Based Medicine (EBM)—a leading paradigm in modern medical practice which calls for finding the "current best evidence" to support optimal clinical decisions for each *individual* patient (Sackett et al., 1996).

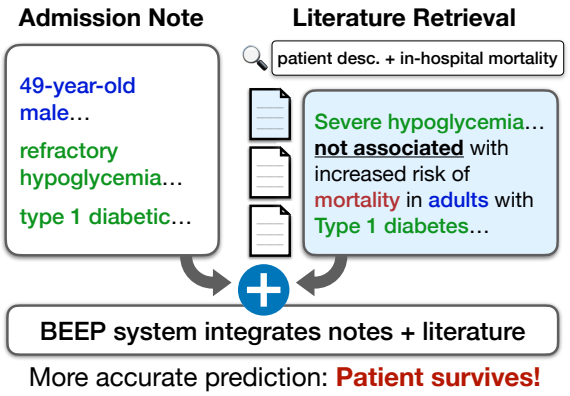


Figure 1: Overview of BEEP. We retrieve literature relevant to the patient description and an outcome of interest, in-hospital mortality in this example. We combine both sources of information to train a model to predict the outcome with better accuracy.

Our setting presents unique challenges. First, our approach requires retrieving literature based on noisy EHR notes containing multitudes of information (e.g., medical history, ongoing treatments), unlike orthogonal efforts on extracting and summarizing scholarly information related to well-formed questions (e.g., the efficacy of ACE inhibitors in adult patients with type-2 diabetes) (Wallace, 2019; Lehman et al., 2019; DeYoung et al., 2020, 2021).

*Work done during internship at AI2.
[1]Our code is available at https://github.com/allenai/BEEP.

In addition, as our end task is predicting patient outcomes, another challenge lies in aggregating the retrieved literature in a way that maximizes prediction accuracy. Toward these challenges, we make the following key contributions:

- **Literature-Augmented Model.** As illustrated in Figure 1, for each ICU patient and each target outcome to be predicted (e.g., mortality), our model retrieves papers from PubMed, encoded and fused together with the ICU admission note for making a final prediction. We present several architectures for retrieving papers and for aggregating and combining them with clinical notes. We make our code, cohort selection, paper identifiers and models publicly available.
- **Adding Literature Boosts Results.** For evaluation, we measure both overall performance and precision/recall@Top-K, to account for the real-world scenario where "alarms" are only raised for high-confidence predictions to avoid alarm fatigue (Sendelbach and Funk, 2013). BEEP provides substantial improvements over baselines, with strong gains in overall classification performance and precision@Top-K. For example, we improve F1 by up to 5 points and precision@Top-K by a large margin of over 25%.
- **Exploring Patient-Specific Retrieval.** We explore a range of sparse and dense retrieval approaches, including language models, for the complex and underexplored task of retrieving relevant literature based on a patient's noisy, information-dense clinical note. Our final retrieval module employs a retrieve-rerank approach that effectively retrieves helpful literature, as shown in our analysis (section 5).

We hope our work opens new research directions for automatically scanning literature for patient-specific evidence, and combining it with EHR information to boost accuracy of medical predictive models. Finally, our work raises the more general prospect of building predictive models that can dynamically *learn to retrieve literature* for optimizing task accuracy, in medicine and other related areas.

2 Related Work

Patient-Specific Literature Retrieval. Since 2014, the Text REtrieval Conference (TREC) has organized a series of challenges to advance research in this area. The TREC Clinical Decision Support (CDS) tracks focused on evaluating systems on the task of retrieving biomedical articles relevant for answering generic clinical questions about patient medical records (e.g., identifying potential diagnoses, treatments, and tests) (Simpson et al., 2014; Roberts et al., 2015, 2016). TREC CDS 2014 and 2015 used short case reports as idealized representations of medical records due to the lack of available de-identified records. TREC 2016 shifted to using real-world medical records from the Medical Information Mart for Intensive Care (MIMIC) database (Johnson et al., 2016).[2] In our work, our focus is on *predicting clinical outcomes* using ICU admission notes and patient-specific retrieved literature.

Ueda et al. (2021) use contextualized representations on more structured retrieval tasks not involving clinical notes (Voorhees et al., 2021), leaving open the question of how large pretrained language models (LMs) would fare on long, noisy EHR text. We explore this by experimenting with LMs for retrieval based on EHR text.

Clinical Outcome Prediction. The idea of using automated outcome prediction for assisting clinical triage, workflow optimization, and hospital resource management has received much interest recently, especially given the conditions of the COVID-19 pandemic (Li et al., 2020). Predictive models based on structured (e.g., lab results) and unstructured (e.g., nursing notes) information have been built for key clinical outcomes including mortality (Jain et al., 2019; Feng et al., 2020), length of hospital stay (van Aken et al., 2021), readmission (Jain et al., 2019), sepsis (Feng et al., 2020), prolonged mechanical ventilation (Huang et al., 2020), and diagnostic coding (Jain et al., 2019; van Aken et al., 2021). Increasingly, models have leveraged unstructured text from notes since they can contain key information for outcome prediction (Boag et al., 2018; Jin et al., 2018). Most recently, van Aken et al. (2021) attempted this using large pretrained LMs. Our work compares the performance of a broader range of state-of-the-art pretrained language models on outcome prediction tasks.

3 BEEP: Literature-Enhanced Clinical Predictive System

Task & Approach Overview. Our goal is to improve models for clinical outcome prediction

[2] Since 2017, the focus has switched to TREC-PM (precision medicine) tracks where articles are retrieved based on short structured queries with attributes such as patient condition and demographics, a less realistic scenario.

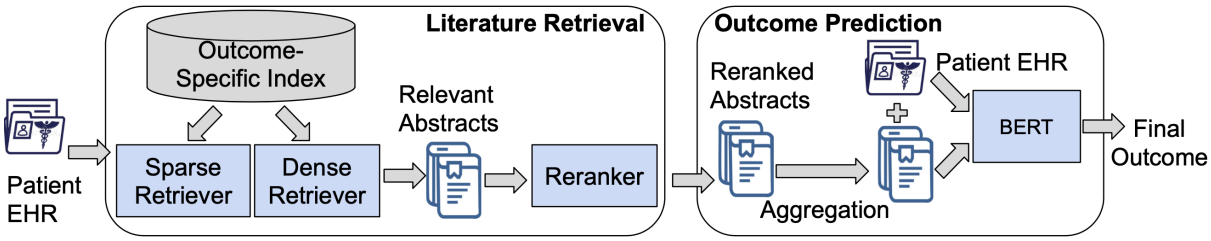


Figure 2: Complete system pipeline, unpacking the high-level overview seen in Figure 1. For a given patient ICU admission note, the literature retrieval module first retrieves relevant biomedical abstracts from a clinical outcome-specific index, then reranks a top-ranked subset of abstracts. The outcome prediction module aggregates information from these reranked abstracts and fuses it with the admission note to make the final prediction

from EHR notes by augmenting them with relevant biomedical literature. BEEP consists of two main stages: (i) literature retrieval, and (ii) outcome prediction. We also briefly experiment with a formulation that trains both jointly (details in section 4). Given a patient EHR note Q and a clinical outcome of interest y, the first stage is to identify a set of biomedical abstracts $Docs(Q) = \{D_1, ..., D_n\}$ from PubMed[3] that may be helpful in assessing the likelihood of the patient having that outcome. The next stage is to augment the input to an EHR-based outcome prediction model with these retrieved abstracts ($Q \cup Docs(Q)$) and predict the final outcome. Figure 1 provides a high-level illustration of BEEP, and Figure 2 unpacks it with more detail. Next, we describe our system's main components.

3.1 Literature Retrieval Module

Our literature retrieval module consists of three components: (i) an index of biomedical abstracts pertaining to the outcome of interest, (ii) a *retriever* that retrieves a ranked list of abstracts relevant to the patient note from the index, and (iii) a *reranker* that reranks retrieved abstracts using a stronger document similarity computation model. For the retriever, we experiment with both sparse and dense models. We follow the standard retrieve-rerank approach, which has been shown to achieve good balance between efficiency and retrieval performance (Dang et al., 2013), and has recently also proved useful for large-scale biomedical literature search (Wang et al., 2021). In the retrieval step, we prioritize efficiency, using models that scale well to large document collections but are not as accurate, to return a set of top documents. In the reranker step, we prioritize retrieval performance by running a computationally expensive but more accurate model on the smaller set of retrieved documents.

3.1.1 Outcome-Specific Index Construction

Since we are interested in identifying information related to a *specific outcome* for a patient, we begin by constructing an index of all abstracts from PubMed relevant to that outcome to limit search scope. To gather all abstracts relevant to a clinical outcome, we first identify MeSH (Medical Subject Heading) terms associated with the outcome by performing MeSH linking on the outcome descriptions using scispaCy (Neumann et al., 2019). These associated MeSH terms are then used as queries to retrieve abstracts.[4] For some MeSH terms that are too broad (e.g., "mortality"), we include additional qualifiers (e.g., "human") to make sure we do not gather articles that are not relevant to our overall patient cohort. Appendix A lists the final set of queries used for all clinical outcomes considered in this work. Abstracts retrieved via this process are used to construct the outcome-specific index.

3.1.2 Sparse Retrieval Model

The sparse retrieval model returns top-ranked abstracts based on cosine similarity between TF-IDF vectors of MeSH terms for the query (clinical note) and the documents (outcome-specific abstracts). MeSH terms from abstracts are extracted by running scispaCy MeSH linking over the abstract text. PubMed MeSH tagging is done only at the abstract level, and does not reflect actual term frequency in the text, requiring our extraction step. However, extracting MeSH terms from clinical notes requires a more elaborate pipeline, due to two major issues:

- **Entity type and boundary issues:** Off-the-shelf entity extractors like scispaCy and cTAKES (Savova et al., 2010) extract some entity types that are uninformative for relevant literature retrieval, e.g., hospital names, references

[3] https://pubmed.ncbi.nlm.nih.gov

[4] https://www.ncbi.nlm.nih.gov/books/NBK25499/

to family members, etc. They also have a tendency to ignore important qualifiers. For example, given a sentence containing the entity "right lower extremity pain", both extractors returned "extremity" and "pain" as separate entities.

- **Negated entities:** Clinical notes have a high density of negated entities (up to 50% of (Chapman et al., 2001)). These entities must be identified and discarded prior to literature retrieval to avoid retrieving articles about symptoms and conditions that are *not* exhibited by the patient.

To handle these issues, we train an entity extraction model that focuses on problems, tests, and treatments with empirically good coverage of important qualifiers (Uzuner et al., 2011). We then filter negated entities with negation detection (Harkema et al., 2009) and perform entity linking to MeSH terms. For more information and implementation details see Appendix B.

3.1.3 Dense Retrieval Model

We add a dense retrieval model to complement the sparse retriever, an approach that has shown promise in recent work (Gao et al., 2021). Our dense retrieval model maps clinical notes (queries) and biomedical abstracts (documents) to a shared dense low-dimensional embedding space. Computing similarity between these encoded vectors allows for softer matching beyond surface form. For dense retrieval, we use a BERT-based bi-encoder model. We use a bi-encoder to support scaling to large document collections, as opposed to cross-encoder models which are much slower (e.g., (Gu et al., 2021)). We use PubmedBERT (Gu et al., 2021) as the encoder and train our bi-encoder using the dataset from the TREC 2016 clinical decision support task (Roberts et al., 2016). For more details, see Appendix B. Our bi-encoder achieves mean precision@10 score of 45.67 on TREC 2016 data in 5-fold cross-validation, comparable to state-of-the-art results (Das et al., 2020).

3.1.4 Reranker Model

The reranker model takes a subset of top-ranked documents from both the sparse and dense retrieval models and rescores them. We use a BERT-based cross-encoder model for reranking, prioritizing ranking performance over efficiency on this smaller subset. Given a query clinical note Q and an abstract document D_i, we run a PubmedBERT-based encoder over the concatenation of both (`[CLS]` Q `[SEP]` D_i `[SEP]`) to compute an embedding E_{QD_i}. This embedding is run through a linear layer to produce a relevance score, trained using cross-entropy loss with respect to document relevance labels from the TREC 2016 dataset. Our cross-encoder achieves a mean precision@10 score of 48.33 on TREC 2016 in 5-fold cross-validation, which is also comparable to state-of-the-art performance on TREC CDS 2016 (Das et al., 2020).

From the top-ranked documents returned by the reranker, the top k are selected[5] to be passed alongside the patient clinical note to the outcome prediction module, which we describe next.

3.2 Outcome Prediction Module

The goal of this module is to compute an aggregate representation from the set of top k abstracts relevant to the clinical note, and then predict the outcome of interest using this aggregate representation and the note representation.

3.2.1 Aggregation Strategies

Let $Docs(Q) = D_1, ..., D_k$ be the set of relevant abstracts retrieved for clinical note Q and $\text{BERT}(X)$ be the encoder function that returns an embedding E_X given a document X. We experiment with four different strategies to compute an aggregate literature representation for $Docs(Q)$, which we denote by $LR(Q)$.

Averaging. Averaging encoder representations:

$$LR(Q) = \frac{1}{k}\sum_{i=1}^{k} \text{BERT}(D_i) \tag{1}$$

Weighted Averaging. Weighted average of encoder representations:

$$LR(Q) = \frac{1}{\sum_{i=1}^{k} w_i}\sum_{i=1}^{k} w_i \cdot \text{BERT}(D_i) \tag{2}$$

where weights w_i are the relevance scores computed by the reranker. The final outcome is computed by concatenating note representation $\text{BERT}(Q)$ with $LR(Q)$ and running this through a linear layer.

We also concatenate the note embedding with each abstract ($E_{QD_i} = [\text{BERT}(Q); \text{BERT}(D_i)]$), run outcome prediction and aggregate output probabilities as follows.

Soft Voting. Averaging per-class probabilities from k outcome prediction runs:

$$p(y = c) = \frac{1}{k}\sum_{i=1}^{k} p(y = c|E_{QD_i}) \tag{3}$$

[5] We treat k as a hyperparameter, see appendix C.

Outcome	0	1	2	3
PMV	3,776	3,335	-	-
MOR	43,609	5,136	-	-
LOS	5,596	16,134	13,391	8,488

(a) Class distribution for all outcomes. For PMV, classes 0 and 1 refer to cases that don't/do require prolonged ventilation. For MOR, classes 0 and 1 refer to patients that don't/do die in admission. For LOS, classes 0-3 refer to stay lengths of <3 days, 3-7 days, 1-2 weeks, and >2 weeks respectively.

Outcome	Train	Dev	Test	#Articles
PMV	5,691	712	708	81,311
MOR	33,997	4,918	9,830	90,125
LOS	30,421	4,391	8,797	93,594

(b) Training, development and test splits, and total number of PubMed articles in our outcome-specific index for each clinical outcome.

Table 1: Data statistics per outcome

Weighted Voting. Weighted average of per-class probabilities from k outcome predictions runs:

$$p(y = c) = \frac{1}{\sum_{i=1}^{k} w_i} \sum_{i=1}^{k} w_i \cdot p(y = c|E_{QD_i}) \tag{4}$$

4 Experiments & Results

We test our system on the task of predicting clinical outcomes from patient admission notes. Predicting outcomes from admission notes can help with early identification of at-risk patients and assist hospitals in resource planning by indicating how long patients may require hospital/ICU beds, ventilators etc. (van Aken et al., 2021).

4.1 Clinical Outcomes

We evaluate our system on three clinical outcomes:

- **PMV:** Prolonged mechanical ventilation prediction, identifying whether a patient will require ventilation for >7 days (Huang et al., 2020).
- **MOR:** In-hospital mortality prediction, identifying whether a patient will survive their current admission (van Aken et al., 2021).
- **LOS:** Length of stay prediction is the task of identifying how long a patient will need to stay in the hospital. We follow van Aken et al. (2021) and group patients into four major categories based on clinician recommendations: <3 days, 3-7 days, 1-2 weeks, and >2 weeks.

PMV and MOR are binary classification tasks, while LOS is a multi-class classification task. We predict these outcomes from patient admission notes extracted from the MIMIC III v1.4 database (Johnson et al., 2016), which contains de-identified EHR data including clinical notes in English from the Intensive Care Unit (ICU) of the Beth Israel Deaconess Medical Center in Massachusetts between 2001 and 2012. Admission notes are constructed by filtering discharge summary documents from MIMIC to only retain the following sections typically known at admission: Chief complaint, (History of) Present illness, Medical history, Admission medications, Allergies, Physical exam, Family history and Social history. Notes that do not contain any of these sections are excluded. For PMV, we follow the cohort selection process from Huang et al. (2020), and include all patients who were above 18 years of age and were on mechanical ventilation for at least 2 days with more than 6 hours each day. Patients transferred from other hospitals, organ donors, and patients with neuromuscular disease, head and neck cancer, and extensive burns, which always lead to PMV and may act as confounds, were excluded. For MOR and LOS, we follow the same cohort selection process as van Aken et al. (2021), and include all patients except newborns and remove duplicate admissions. Following these cohort selection processes results in the data splits shown in Table 1b. Table 1b also shows the numbers of relevant PubMed articles for all three clinical outcomes.

4.2 Selecting the Encoder Language Model

Since the encoder used for outcome prediction needs to produce representations for both clinical notes and relevant abstracts, we choose language models that have been pretrained on *both* biomedical and clinical text. We evaluate the following models on outcome prediction (without literature augmentation) to choose a suitable encoder:

- **ClinicalBERT** (Alsentzer et al., 2019): ClinicalBERT further pretrains BioBERT (Lee et al., 2020), a biomedical language model, on EHR notes from MIMIC III. We evaluate both versions: one trained on discharge summary notes only, and one trained on both discharge summaries and nursing notes.
- **CORe** (van Aken et al., 2021): CORe further pretrains BioBERT with a next sentence prediction objective on sentences describing admissions and outcomes. CORe jointly trains on EHR notes and biomedical articles.

Model	PMV			MOR			LOS		
	AUROC	Micro F1	Macro F1	AUROC	Micro F1	Macro F1	AUROC	Micro F1	Macro F1
BLUEBERT	54.27	53.25	51.64	81.49	89.11	**62.69**	**73.22**	45.66	**44.18**
+Avg	57.21	54.66	52.32	83.90	90.52	61.62	71.66	45.22	40.66
+SVote	**58.16**	**56.07**	52.63	84.21	**90.60**	61.00	72.54	**46.02**	42.46
+WVote	57.71	57.91	**56.67**	84.00	90.45	61.02	71.49	44.82	39.55
+WAvg	57.59	55.65	52.21	**84.26**	90.44	60.49	72.58	45.90	42.39
UMLSBERT	56.44	56.07	**54.97**	83.34	87.93	66.93	72.19	43.12	42.20
+Avg	58.36	**56.50**	54.62	84.02	90.41	60.28	72.25	45.61	41.58
+SVote	55.92	54.66	50.94	83.30	84.82	**67.23**	72.14	45.55	42.12
+WVote	**59.43**	56.07	54.26	**84.65**	**90.62**	62.93	**72.71**	**46.44**	**42.71**
+WAvg	59.30	**56.50**	53.70	83.59	90.35	59.61	71.02	44.58	39.95

Table 2: Performance of baseline and literature-augmented outcome prediction models on all clinical outcomes. We note that LOS is a multiclass target; we observe substantial gains in 2/4 of the classes (Table 10 in the Appendix).

- **BLUEBERT** (Peng et al., 2019): BLUEBERT further pretrains BERT (Devlin et al., 2019) jointly on EHR notes and PubMed abstracts.
- **UMLSBERT** (Michalopoulos et al., 2021): UMLSBERT further pretrains ClinicalBERT on EHR notes from MIMIC, with tweaks to the architecture and pretraining objective to incorporate conceptual knowledge from the Unified Medical Language System (UMLS) Metathesaurus (Schuyler et al., 1993).

Note that in this experiment, we predict clinical outcomes from patient admission notes only, without incorporating literature. We also use weighted cross-entropy loss to manage class imbalance (see Appendix B). Table 5 in the Appendix shows the performance of the above language models on the validation sets for all clinical outcomes. We select the top-performing language models BLUEBERT and UMLSBERT for our remaining experiments.[6]

4.3 Literature Augmentation Results

We provide two sets of results: for overall performance, and for high-confidence predictions.

Overall Performance. Table 2 shows the overall performance of our literature-augmented outcome prediction system on all three clinical outcomes. We test our system using both UMLSBERT and BLUEBERT as encoders, as well as all four literature aggregation strategies. We report three metrics for each setting: (i) area under the receiver operating characteristic (AUROC), (ii) micro-averaged F1 score, and (iii) macro-averaged F1 score. From Table 2, we observe that incorporating literature leads to performance improvements on two of three clinical outcomes, PMV and mortality. On LOS prediction, results are more mixed, with minor improvements on micro F1 but no improvements on other metrics. Comparing BLUEBERT and UMLSBERT, variants that use UMLSBERT do slightly better on PMV and mortality, while results on LOS are more mixed. Comparing across literature aggregation strategies, there is no clear winner, though voting-based strategies seem to have a slight advantage, especially on UMLSBERT.

[6]We also experiment with CORe but observe consistently lower scores (Table 8 in Appendix F).

Model	No PMV		PMV	
	Prec@10	Rec@10	Prec@10	Rec@10
BLUEBERT	52.86	9.95	55.71	11.61
+Avg	**64.29**	**12.1**	60.0	12.5
+SVote	61.43	11.56	**64.29**	**13.39**
+WVote	62.86	11.83	52.86	11.01
+WAvg	58.57	11.02	52.86	11.01
UMLSBERT	58.57	11.02	57.14	11.90
+Avg	67.14	12.63	**64.29**	**13.39**
+SVote	61.43	11.56	62.86	13.1
+WVote	64.29	12.1	**64.29**	**13.39**
+WAvg	**68.57**	**12.9**	62.86	13.1

(a) For PMV

Model	No MOR		MOR	
	Prec@10	Rec@10	Prec@10	Rec@10
BLUEBERT	**99.8**	**11.15**	46.39	**23.62**
+Avg	99.59	11.13	68.91	17.81
+SVote	99.69	11.14	**73.39**	16.55
+WVote	99.59	11.13	68.36	16.94
+WAvg	**99.8**	**11.15**	69.46	16.07
UMLSBERT	99.8	11.15	42.06	**39.21**
+Avg	99.59	11.13	**69.07**	15.78
+SVote	99.8	11.15	40.69	38.72
+WVote	99.49	11.12	68.44	19.94
+WAvg	**100.0**	**11.17**	68.92	14.81

(b) For MOR

Table 3: Precision and recall scores for top 10% high-confidence predictions per class.

Evaluating High-Confidence Predictions. In addition to standard evaluation, we evaluate the

top 10% high-confidence predictions per class for all models (precision/recall@TOP-K), informative for two key reasons. First, when using automated outcome prediction systems in a clinical setting, it is reasonable to only consider raising alarms for high-confidence positive predictions to avoid alarm fatigue (Sendelbach and Funk, 2013). Second, high-confidence predictions for both positive and negative classes can be used to reliably assist with hospital resource management (e.g., predicting future ventilation and hospital bed needs).

Tables 3a and 10 show the precision/recall@TOP-K scores for all models on prolonged mechanical ventilation, mortality, and length of stay prediction. In Table 3a, we see that our literature-augmented models achieve much higher precision scores than the baseline (∼9-12 points higher in most cases) for the PMV negative class. We also see higher precision scores than the baseline for the positive class (∼5-9 points higher in most cases). This is a strong indicator that our literature-augmented pipeline might offer more utility for PMV detection in a clinical setting than using EHR notes only. Table 3b shows similarly encouraging trends for mortality prediction. The mortality prediction dataset is the most skewed of the three datasets, and therefore we do not see much performance difference across models on the negative class. However, on the positive class, our literature-augmented models show dramatic increase in precision. In particular, BLUEBERT-based literature models show an increase in precision of ∼22-27 points, at the expense of only ∼6-7 point drop in recall relatively to non-literature models.[7] This also indicates that literature-augmented mortality prediction might be more precise and reliable in a clinical setting than using clinical notes alone. From Table 10 (Appendix H), we can see that for LOS prediction, our models show clear gains (∼2-5 points) on classes 1 and 2 (i.e., 3-7 days and 1-2 weeks), and minor gains for some variants on class 3 (>2 weeks). We also perform an alternate evaluation in which we only score predictions from our literature-augmented models that show a relative confidence increase of at least 10% over the baseline prediction, presented in Appendix H.

Learning To Retrieve Using Outcomes. BEEP trains separate models for literature retrieval and outcome prediction. Inspired by Lee et al. (2019), we develop a learning-to-retrieve (L2R) formulation that trains both jointly to ensure that the retriever can learn from outcome feedback. However, our L2R model does not improve performance over BEEP (results in Table 7 in Appendix E). We provide discussion for potential reasons in Appendix E. This is an interesting direction for future work.

5 Analysis and Discussion

Given BEEP's improved performance, we further assess the utility of retrieved literature and cases where adding literature is particularly helpful.

Diversity of retrieved literature. As a preliminary analysis, we evaluate the diversity of the abstracts retrieved for admission notes in our datasets, as a proxy for the degree to which literature is personalized to specific patient cases. For the 100 most frequently retrieved abstracts for each clinical outcome, Figures 4a, 4b, and 4c in Appendix H show proportions of patient notes for which these abstracts are judged as relevant by our retriever-rerank pipeline. From these histograms, we see a stark difference for LOS which is much less diverse than both PMV and MOR, indicating that the literature retrieved for length of stay prediction may be less personalized to patient cases than the literature retrieved for other outcomes. We leave to future work exploration of diversifying retrieved papers across patients and examining the effect on outcome prediction performance.[8]

Qualitative examination of retrieved literature. We qualitatively examine literature retrieved for cases in which our model shows large confidence increases over the baseline to determine its utility in making the right prediction. We study increases in both directions, i.e. cases in which adding literature resulted in a confidence increase in either the *correct* outcome label (good) or *incorrect* outcome label (bad). For each clinical outcome, a bio-NLP expert looked at the top 5 cases from each category based on the magnitude of confidence increase (total 10 cases per outcome). For each case, the expert looks at the top 5 abstracts retrieved for the case (total 50 abstracts per outcome) and assigns each abstract to one of 8 categories we define for categorizing degree of relevance and type of evidence provided, including retrievals considered *helpful*

[7] Note that since the MOR class is rare, a larger recall drop could still translate to a small number of incorrect cases only

[8] We perform an ablation in which we use only the retrieved literature for prediction, showing quantitative evidence for the utility of retrieved literature (see Appendix G).

Patient EHR	Retrieved Abstract	Evidence Type	Outcome
CHIEF COMPLAINT: liver tranplant **PRESENT ILLNESS:** ...s/p liver tranplant...Dx: ESLD secondary to alcoholic **cirrhosis**. **MEDICAL HISTORY**: EtOH Cirrhosis	Retrospective review of data of 73 consecutive patients with **cirrhosis requiring MV**...majority of patients, 51/64 **(79.7%), dying in the first 48 hours of intubation**...	Patient condition and outcome directly related	No PMV
CHIEF COMPLAINT: Aortic dissection **PRESENT ILLNESS:** ...72-year-old woman...chest pain...had **type A aortic dissection**...an intramural hematoma...proceed with **surgery**... **MEDICAL HISTORY:** HTN Renal failure	Acute **type A aortic dissection** presents a formidable challenge...the most important variables associated with **in-hospital mortality** in patients undergoing **surgery** for this condition...suggests that **CPB time, diabetes mellitus and postoperative bleeding** are the main determinants of in-hospital death.	Known outcome indicators **not** present in patient	No MOR
CHIEF COMPLAINT: Dyspnea, fever **PRESENT ILLNESS:** 58F w/ HCV **cirrhosis**...requiring BiPAP, ultimately urgent **intubation**... extubated ... short of breath... **MEDICAL HISTORY:** HCV cirrhosis	...study identifies specific predictors of increased mortality and resource utilization in **cirrhotic** patients...**Increased LOS** in the MICU was **associated with mechanical ventilation**...	Ongoing treatment and outcome related	LOS >2 weeks

Table 4: Qualitative examples of retrieved literature that is helpful for increasing prediction confidence of the correct outcome. Case 1 shows an example of retrieved literature that strongly matches patient condition and provides direct evidence linking it to the outcome of interest. Case 2 shows an example with indirect evidence, in which retrieved literature lists outcome indicators not present in the patient. Case 3 shows an example of retrieved literature describing a link between patient's ongoing treatment and outcome of interest. **green**: patient characteristics; **blue**: outcome of interest; **red**: known indicators of the outcome measure not present in the patient.

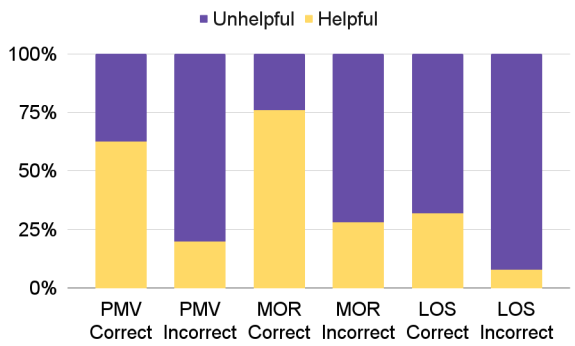


Figure 3: Literature categorization for both correct and incorrect outcome cases. For PMV and MOR, retrieved literature for correct cases is more often categorized as helpful, and unhelpful literature dominates for incorrect cases. For LOS, literature for both categories is more often categorized as unhelpful.

and *unhelpful*. For example, see Table 4 (evidence type column; more in Appendix).

As seen in Table 4, for *helpful* categories, retrieved literature matches patient characteristics (especially current condition) and includes evidential links between outcome of interest and patient conditions/treatment. In the first case, the retrieved abstract provides evidence that patients with cirrhosis have high mortality in the first 48 hours of intubation, entails the patient might not undergo prolonged ventilation. In the second case, the abstract lists comorbidities associated with in-hospital mortality (outcome of interest), but none are present in the patient under consideration, which can be taken as weak indication that the patient may survive. Similarly, for the third case, the retrieved abstract mentions that cirrhotic patients may have longer hospital stays if they are on mechanical ventilation. This matches our patient's treatment history since she has cirrhosis and was briefly intubated and extubated, before experiencing shortness of breath again. Given this, the patient might have a longer length of stay. Conversely, *unhelpful* retrieved literature often does not match patient characteristics or may not contain evidence relevant to the outcome. See more example explanations in Appendix I.

Figure 3 presents the distribution of helpful and unhelpful categories for both kinds of cases for all outcomes. We can see that for correct outcome cases from both PMV and mortality, retrieved literature is more frequently assigned to one of the helpful categories, while for incorrect outcome cases, retrieved literature is more frequently assigned to one of the unhelpful categories. For LOS, unhelpful categories dominate both types of cases, especially prevalent in incorrect outcomes.

6 Conclusion

In this paper, we introduced BEEP, a system that automatically retrieves *patient-specific literature* based on intensive care (ICU) EHR notes and uses the literature to enhance clinical outcome prediction. On three challenging tasks, we obtain substantial improvements over strong recent baselines, seeing dramatic gains in top-10% precision for mortality prediction with a boost of over 25%.

Our hope is that this work will open new research directions into bridging the gap between AI-based clinical models and the Evidence Based Medicine (EBM) paradigm in which medical decisions are based on explicit evidence from the literature. An interesting direction is to incorporate evidence identification and inference (Wallace, 2019; DeYoung et al., 2020) directly into our retrieval and predictive models. Another important question to explore relates to the implications our approach has on increasing the interpretability of clinical AI models.

Acknowledgements

This work was supported in part by NSF Convergence Accelerator Award #2132318. The authors would like to thank the members of the Semantic Scholar team, and the anonymous reviewers for their helpful feedback on this work.

Ethical Concerns

Incorporating outcome prediction models into a medical decision-making pipeline effectively will require these technologies to adhere to standards set by the core principles of medical ethics: beneficence, non-maleficence, autonomy, and justice (Beauchamp et al., 2001). These requirements may raise the following concerns when deploying outcome prediction models in clinical settings:

- **Out-of-Cohort Generalization:** The extent to which outcome prediction models generalize to patient cohorts that may not have been present in their training data is unclear. If model accuracy is significantly lower on "out-of-cohort" patients, using inaccurate/uncertain predictions during decision making may violate the requirement that any application of technology must be *beneficent* and *non-maleficent* to individual patients. Our proposed technique can partly mitigate the generalization issue by identifying additional supporting evidence from literature, which may be better tailored to individual patient characteristics, instead of using only cohort-level evidence. However, biomedical literature can also have blind spots, with certain cohorts and disease combinations being under-studied, and even literature-augmented prediction may not be sufficiently accurate.
- **Algorithmic Biases:** Since outcome prediction models are trained on historical health data, existing inequities in healthcare access may translate into models continuing to perpetuate unintentional discrimination against patients from under-served demographics. For example, models might predict poorer outcomes (e.g., high mortality, poor response to treatment, etc.) for specific demographics that have historically had worse outcomes due to poor access to care. Such issues are a clear violation of the *justice* requirement, and must be tackled before deployment.
- **Informed Consent:** Lastly, if outcome prediction models are used in clinical settings, patients and their caregivers must be made aware of their use, since the principle of *autonomy* emphasizes that patients must be provided all relevant medical information to support autonomous decision making. The black-box nature of these models raises another issue: how can we help patients/caregivers understand and interpret outcome predictions to further support their autonomy in decision making? We hope that literature-augmented prediction techniques can partly ease this by using evidence snippets from literature that contributed to the model's prediction as explanations.

References

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.

Tom L Beauchamp, James F Childress, et al. 2001. *Principles of biomedical ethics*. Oxford University Press, USA.

Willie Boag, Dustin Doss, Tristan Naumann, and Peter Szolovits. 2018. What's in a note? unpacking predictive value in clinical note representations. *AMIA Summits on Translational Science Proceedings*, 2018:26.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001.

Evaluation of negation phrases in narrative clinical reports. In *Proceedings of the AMIA Symposium*, page 105. American Medical Informatics Association.

Van Dang, Michael Bendersky, and W Bruce Croft. 2013. Two-stage learning to rank for information retrieval. In *European Conference on Information Retrieval*, pages 423–434. Springer.

Manirupa Das, Juanxi Li, Eric Fosler-Lussier, Simon Lin, Steve Rust, Yungui Huang, and Rajiv Ramnath. 2020. Sequence-to-set semantic tagging for complex query reformulation and automated text categorization in biomedical IR using self-attention. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 14–27, Online. Association for Computational Linguistics.

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.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Wang. 2021. MSˆ2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132.

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. 2021. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings 2021*.

Jinyue Feng, Chantal Shaib, and Frank Rudzicz. 2020. Explainable clinical decision support from text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1478–1489, Online. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Tongfei Chen, Zhen Fan, Benjamin Van Durme, and Jamie Callan. 2021. Complement lexical retrieval model with semantic residual embeddings. In *European Conference on Information Retrieval*, pages 146–160. Springer.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. 2009. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851.

Mohammad Hashir and Rapinder Sawhney. 2020. Towards unstructured mortality prediction with free-text clinical notes. *Journal of Biomedical Informatics*, 108:103489.

Kexin Huang, Abhishek Singh, Sitong Chen, Edward Moseley, Chih-Ying Deng, Naomi George, and Charolotta Lindvall. 2020. Clinical XLNet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 94–100, Online. Association for Computational Linguistics.

Sarthak Jain, Ramin Mohammadi, and Byron C Wallace. 2019. An analysis of attention over clinical notes for predictive tasks. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 15–21.

Mengqi Jin, Mohammad Taha Bahadori, Aaron Colak, Parminder Bhatia, Busra Celikkaya, Ram Bhakta, Selvan Senthivel, Mohammed Khalilia, Daniel Navarro, Borui Zhang, et al. 2018. Improving hospital mortality prediction with medical named entities and multimodal learning. *arXiv preprint arXiv:1811.12276*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. 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 3705–3717.

Matthew D Li, Nishanth Thumbavanam Arun, Mishka Gidwani, Ken Chang, Francis Deng, Brent P Little, Dexter P Mendoza, Min Lang, Susanna I Lee, Aileen O'Shea, et al. 2020. Automated assessment and tracking of covid-19 pulmonary disease severity on chest radiographs using convolutional siamese neural networks. *Radiology: Artificial Intelligence*, 2(4):e200079.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):1–10.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, and William R. Hersh. 2016. Overview of the TREC 2016 clinical decision support track. In *Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016*, volume 500-321 of *NIST Special Publication*. National Institute of Standards and Technology (NIST).

Kirk Roberts, Matthew S Simpson, Ellen M Voorhees, and William R Hersh. 2015. Overview of the trec 2015 clinical decision support track. In *TREC*.

David L Sackett, William MC Rosenberg, JA Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217.

Sue Sendelbach and Marjorie Funk. 2013. Alarm fatigue: a patient safety concern. *AACN advanced critical care*, 24(4):378–386.

Matthew S Simpson, Ellen M Voorhees, and William Hersh. 2014. Overview of the trec 2014 clinical decision support track. Technical report, LISTER HILL NATIONAL CENTER FOR BIOMEDICAL COMMUNICATIONS BETHESDA MD.

Alberto Ueda, Rodrygo L. T. Santos, Craig Macdonald, and Iadh Ounis. 2021. Structured fine-tuning of contextual embeddings for effective biomedical retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2031–2035, New York, NY, USA. Association for Computing Machinery.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix Gers, and Alexander Loeser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 881–893, Online. Association for Computational Linguistics.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Byron C Wallace. 2019. What does the evidence say? models to help make sense of the biomedical literature. In *IJCAI: proceedings of the conference*, volume 2019, page 6416. NIH Public Access.

Yu Wang, Jinchao Li, Tristan Naumann, Chenyan Xiong, Hao Cheng, Robert Tinn, Cliff Wong, Naoto Usuyama, Richard Rogahn, Zhihong Shen, et al. 2021. Domain-specific pretraining for vertical search: Case study on biomedical literature. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3717–3725.

A PubMed Queries Per Outcome

Following are the MeSH terms that we use to retrieve literature from PubMed to construct the

outcome-specific index for each clinical outcome under consideration:

- **Prolonged Mechanical Ventilation (PMV):** "Respiration, Artificial". We also query using the terms "Ventilation, Mechanical" and "Ventilator Weaning" but do not find any new results.
- **In-Hospital Mortality (MOR):** "Hospital Mortality", "Mortality+Humans+Risk Factors". Note that the "+" operator is interpreted as AND by PubMed search.
- **Length of Stay (LOS):** "Length of Stay". All other MeSH terms from the tagger are aliases of this term.

B Implementation Details

Entity Extraction. First, we extract entities from clinical notes using a model trained on the i2b2 2010 concept extraction dataset (Uzuner et al., 2011). This dataset consists of clinical notes annotated with three types of entities: problems, tests, and treatments. These entity types cover the pertinent medical information that can be used to retrieve abstracts relevant to a clinical note. Moreover, the i2b2 guidelines require annotators to include all qualifiers within an entity span, so training a model on these annotations should bias it towards including pertinent entity qualifiers. Our entity extraction model uses a BERT-based language model to compute token representations, followed by a linear layer to predict entity labels.

We use ClinicalBERT (Alsentzer et al., 2019) as the the language model to train our i2b2 entity extractor. Table 6 shows the performance of our model on the i2b2 2010 test set. These numbers are close to the exact F1 scores reported by Alsentzer et al. (2019) on i2b2 2010 (87.8).

Entity Filtering. After extracting entities, we filter out all negated entities. Negated entities are detected using the ConText algorithm for negation detection from clinical text (Harkema et al., 2009). We use the implementation of ConText negated entity detection algorithm provided by medspaCy (Eyre et al., 2021).

MeSH Linking. Finally, the set of filtered entities is linked to MeSH terms using scispaCy. Entities not linked to MeSH terms are discarded. MeSH terms linked in clinical notes and abstracts are used to compute TF-IDF vectors for the sparse retrieval model.

Bi-Encoder Given a query clinical note Q and an abstract document D_i, a BERT-based encoder is used to compute dense embedding representations E_Q and E_{D_i}. A scoring function S is defined as the Euclidean distance between query and document embeddings:

$$S(Q, D_i) = \|E_Q - E_{D_i}\|_2 \tag{5}$$

Documents closest to the query vector in the embedding space are returned as top-ranked results. The bi-encoder is trained using a triplet loss function defined as follows:

$$L(Q, D_i^+, D_i^-) = \max(S(Q, D_i^+) - S(Q, D_i^-) + m, 0) \tag{6}$$

Here D_i^+ is an abstract more relevant to the clinical note Q than D_i^- and m is a margin value. We use PubmedBERT (Gu et al., 2021) as the encoder and train our bi-encoder using the dataset from the TREC 2016 clinical decision support task (Roberts et al., 2016).[9] This dataset consists of 30 de-identified EHR notes, along with ~1000 PubMed abstracts per note marked for relevance. We select relevant abstracts per note as positive candidates (D_i^+), and irrelevant abstracts for the same note as negative candidates (D_i^-).

Outcome prediction module training. We use a weighted cross-entropy loss function to handle class imbalance. Given a dataset with N total examples, c classes and n_i examples in class i, class weights are computed as follows:

$$w_i = \frac{N}{c \cdot n_i} \tag{7}$$

We use Adam optimizer, treating initial learning rate as a hyperparameter. All models are implemented in PyTorch, and we use Huggingface implementations for all pretrained language models.

C Hyperparameter Tuning

We do a grid search over the following hyperparameter values for each aggregation:

Learning Rate (LR): [5e-4, 1e-5, 5e-5, 1e-6,5e-6]

Number of top abstracts (k): [1, 5, 10]

[9] We do not use data from TREC 2014 and 2015 since they use idealized case reports instead of actual EHR notes. Combining all three datasets degraded performance, likely due to differences in language between case reports and EHRs.

	PMV		MOR		LOS	
LM	**AUROC**	**Micro F1**	**AUROC**	**Micro F1**	**AUROC**	**Micro F1**
ClinicalBERT (Full)	54.66	53.93	81.78	86.34	70.94	40.00
ClinicalBERT (Disc.)	54.91	54.21	81.78	86.34	71.44	40.36
CORe	54.98	54.35	81.58	84.85	69.15	37.94
BLUEBERT	56.60	55.34	82.40	84.75	**71.87**	**41.93**
UMLSBERT	**57.42**	**55.48**	**83.31**	**87.29**	71.60	41.84

Table 5: Performance of various language models trained on clinical and biomedical text on all clinical outcomes. For ClinicalBERT, Disc. and Full refer respectively to variants trained on discharge summaries only and both discharge summaries and nursing notes.

Category	**Exact F1**
Overall	86.66
Test	87.48
Problem	86.53
Treatment	86.03

Table 6: Entity extraction model performance on i2b2 2010 test set

Gradient accumulation steps (GA): [10, 20] This hyperparameter grid stays consistent across all outcome prediction experiments. For all experiments, we currently report the outcome of a single run.

D Computing Infrastructure

Our experiments were carried out on 2 AWS p3.16xlarge instances, which are 8-GPU machines with 16 GB RAM per GPU. All our experiments can be run on a single 16 GB GPU.

E Results from Learning To Retrieve Model

Given a note Q, we first obtain a set of top 100 relevant abstracts ($Docs(Q) = \{D_1, ..., D_{100}\}$) from the BEEP retrieve-rerank pipeline. The retriever component is then defined as follows:

$$E_Q = BERT_Q(Q) \quad (8)$$

$$E_{D_i} = BERT_D(D_i) \quad (9)$$

$$S_{retr}(Q, D_i) = cosine(E_Q, E_{D_i}) \quad (10)$$

$BERT_Q(X)$ and $BERT_D(X)$ are the query and document encoder functions. Based on retriever scores S_{retr}, we select the top k abstracts and perform outcome prediction using the same structure as the BEEP outcome prediction module. We also add the following early update loss term to the outcome loss for the retriever component:

$$P_{early}(D_i|Q) = \frac{exp(S_{retr}(Q, D_i))}{\sum_{D_j \in Docs(Q)} exp(S_{retr}(Q, D_j))} \quad (11)$$

$$L_{early} = -\log \sum_{D_j \in Docs(Q)} y_j P_{early}(D_j|Q) \quad (12)$$

where y_j is set to 1 if using document D_j alongside Q results in a confidence increase in the correct outcome (as per BEEP) and 0 otherwise. Our L2R model does not improve performance over BEEP (results in Table 7). We speculate that this may partly be due to the fact that the heuristic we use to assign y_j values in early update loss is not as accurate as the one used by Lee et al. (2019) (directly checking for presence of the answer in a document, for the reading comprehension task).

Table 7 presents results for the learning-to-retrieve model on all clinical outcomes using UMLSBERT as the encoder. From the table, we can see that while L2R improves performance over a notes-only baseline, its performance is comparable to BEEP. As mentioned earlier, we speculate that this may partly be attributed to the fact that the heuristic we use to assign y_j values in early update loss is not as accurate as the one used by Lee et al. (2019) (directly checking for presence of answer in document, for the reading comprehension task). We believe that experimenting with other sources of supervision to generate y_j values and weighting mechanisms to better combine outcome and early update losses might lead to larger improvements, but we leave those to future work.

F Literature-Augmented Outcome Prediction with CORe

Table 8 shows the overall performance of our literature-augmented outcome prediction system

	PMV			MOR			LOS		
Model	**AUROC**	**Micro F1**	**Macro F1**	**AUROC**	**Micro F1**	**Macro F1**	**AUROC**	**Micro F1**	**Macro F1**
UMLSBERT	56.44	56.07	54.97	83.34	87.93	66.93	72.19	43.12	42.20
+Avg	54.17	53.53	41.51	84.54	90.47	60.53	71.90	44.88	41.26
+SVote	54.29	52.82	39.93	84.50	90.51	61.10	72.17	45.56	41.68
+WVote	57.60	**56.50**	**55.93**	83.92	90.54	61.20	72.72	46.46	42.17
+WAvg	**58.65**	55.79	53.68	84.68	90.59	62.78	72.16	45.04	40.87

Table 7: Performance of learning to retrieve (L2R) model on all clinical outcomes using the UMLSBERT language model

	PMV			MOR			LOS		
Model	**AUROC**	**Micro F1**	**Macro F1**	**AUROC**	**Micro F1**	**Macro F1**	**AUROC**	**Micro F1**	**Macro F1**
CORe	55.91	53.96	53.71	79.96	78.92	62.46	71.52	42.59	**42.33**
+Avg	**58.76**	55.51	55.43	82.41	84.67	**66.06**	**71.99**	40.54	40.39
+SVote	58.40	**58.62**	55.23	81.90	**89.90**	55.76	71.35	**45.07**	40.16
+WVote	58.03	56.92	53.14	**82.81**	**89.87**	53.16	70.96	44.74	39.73
+WAvg	57.53	55.51	**55.49**	81.98	81.86	64.63	71.17	39.48	39.67

Table 8: Performance of baseline and literature-augmented outcome prediction models on all clinical outcomes using the CORe language model

	PMV			MOR			LOS		
Model	**AUROC**	**Micro F1**	**Macro F1**	**AUROC**	**Micro F1**	**Macro F1**	**AUROC**	**Micro F1**	**Macro F1**
BLUEBERT	–	–	–	–	–	–	–	–	–
+Avg	55.72	54.38	46.95	68.72	89.49	47.23	63.40	39.40	29.15
+SVote	57.11	56.50	52.21	71.04	89.49	48.73	63.46	39.41	28.90
+WVote	55.83	53.25	43.43	71.00	89.50	48.73	63.40	39.56	27.52
+WAvg	56.99	55.65	47.97	71.39	89.48	49.26	63.46	39.34	27.99
UMLSBERT	–	–	–	–	–	–	–	–	–
+Avg	59.15	55.37	50.79	71.22	89.49	48.54	63.84	39.49	30.30
+SVote	56.53	55.09	51.76	69.31	89.50	47.71	63.14	38.95	27.12
+WVote	57.06	54.38	53.77	70.54	89.46	49.34	63.46	39.40	27.55
+WAvg	56.99	54.94	54.29	70.04	89.46	49.16	63.51	39.51	28.32

Table 9: Performance of models that only use retrieved literature for outcome prediction on all clinical outcomes

	<3 days		**>=3 and <=7 days**		**>7 and <=14 days**		**>14 days**	
Model	**Prec@10**	**Rec@10**	**Prec@10**	**Rec@10**	**Prec@10**	**Rec@10**	**Prec@10**	**Rec@10**
BLUEBERT	47.6	37.11	61.09	16.14	44.98	14.15	50.74	26.93
+Avg	54.23	24.0	60.64	16.02	45.45	14.49	49.48	25.66
+SVote	54.48	27.12	62.12	16.41	46.38	14.97	51.33	26.87
+WVote	55.73	21.68	61.66	16.29	46.68	12.78	47.99	25.18
+WAvg	52.48	28.28	60.75	16.05	47.33	15.12	51.03	26.99
UMLSBERT	47.33	37.11	59.95	15.84	44.83	13.04	48.92	25.97
+Avg	53.08	26.14	60.41	15.96	48.3	15.27	49.6	26.03
+SVote	52.37	28.55	59.5	15.72	44.38	14.38	49.36	25.72
+WVote	57.22	27.21	64.28	16.98	45.43	14.78	50.4	26.33
+WAvg	52.86	20.61	59.84	15.81	44.9	14.38	48.44	25.24

Table 10: Precision and recall scores for top 10% high-confidence predictions per class (precision/recall@TOP-K) for LOS.

Model	No PMV	PMV
BLUEBERT+Avg	55.47	57.48
BLUEBERT+SVote	56.82	55.56
BLUEBERT+WVote	62.50	62.67
BLUEBERT+WAvg	56.34	61.29
UMLSBERT+Avg	63.71	60.71
UMLSBERT+SVote	50.39	65.62
UMLSBERT+WVote	61.83	59.09
UMLSBERT+WAvg	57.80	63.33

(a) Precision on PMV, when considering cases for which literature-augmented models achieve >10% increase in prediction confidence over baseline.

Model	No MOR	MOR
BLUEBERT+Avg	87.91	69.77
BLUEBERT+SVote	87.49	75.00
BLUEBERT+WVote	86.99	76.09
BLUEBERT+WAvg	87.29	77.68
UMLSBERT+Avg	85.33	83.33
UMLSBERT+SVote	90.33	31.01
UMLSBERT+WVote	86.66	52.17
UMLSBERT+WAvg	85.29	60.00

(b) Precision on MOR, when considering cases for which literature-augmented models achieve >10% increase in prediction confidence over baseline.

on all three clinical outcomes when the CORe language model is used as an encoder. From this table, we can see that adding literature improves performance in this setting as well (with the exception of macro F1 on length of stay). However the overall scores are lower than the settings in which UMLSBERT and BLUEBERT are used as encoders (Table 2).

G Literature-Only Outcome Prediction

To quantitatively test the quality of the retrieved literature, we run an ablation study in which we predict the clinical outcome using only the literature retrieved for a specific patient case, without incorporating any information from the patient clinical note. Table 9 shows the results for this ablation study, using both BLUEBERT and UMLSBERT encoders. From this table, we can see that while removing the clinical note leads to performance drops, especially on mortality and length of stay, the retrieved literature does have some predictive ability. We take this as indication that the retrieved literature contains some clinical indicators associated with the outcome, that are also present in the patient's clinical note.

H Analyzing High Confidence Increases Over Baseline

Finally, we also examine an alternate way of using high-confidence predictions made by our models. We run both baseline and literature-augmented systems, and only consider predictions from the literature-augmented system that show a high increase in confidence, such as $> 10\%$ increase relative to the baseline predictions for the same cases. Tables 11a and 11b show the precision scores of all models on prolonged mechanical ventilation and mortality in this setting. We can see that precision scores in this setting are fairly high, especially for the negative class in mortality prediction. Most averaging variants also do well on the positive class in mortality prediction.

I Examples of Literature For Incorrect Outcome Cases

We categorize examples into the following:

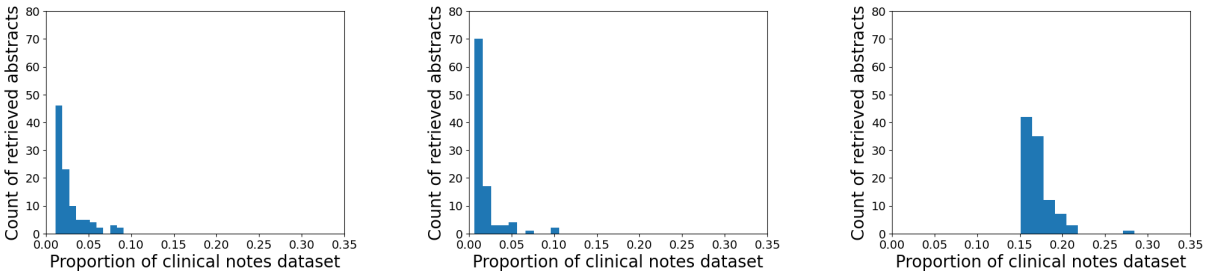
1. Patient condition and outcome directly related
2. Patient history and outcome related
3. Known outcome indicators not present in patient
4. Ongoing treatment and outcome related
5. No cohort match
6. No/weak condition match
7. Condition-outcome pair not studied
8. No evidence for outcome/Weak evidence for direct relationship between patient condition and outcome

From table 12, we can see that retrieved literature from *unhelpful* categories often does not match patient characteristics. The first case discusses a patient who has had an ICD firing incident, but the retrieved literature discusses ICD implantation therapy. While related, there is no discussion of the impact of ICD firing on various clinical outcomes.

For the second case, we see that the retrieved article discusses strokes in general, without matching any of the patient's indications or demographic characteristics. Moreover, the outcome of interest (mortality) is mentioned briefly, but links between the outcome and patient conditions are not studied. Finally, the third case provides an example of a common phenomenon we observe. There are a fair number of review articles retrieved that do not have strong evidential statements in the abstract. For the third case, the retrieved abstract discusses the need for early triage/transfer (which could lead to low length of stay), but then do not provide any conclusive evidence.

Patient EHR	Retrieved Abstract	Evidence Type	Outcome
CHIEF COMPLAINT: ICD firing **PRESENT ILLNESS:** 57 yo M presenting s/p ICD discharges...shocks preceded by prodrome of dizziness,...and was shocked once...Has not had ICD firing prior to these events since implant **MEDICAL HISTORY:** Heart failure...	...assess if selected clinical markers of organ dysfunction were associated with increased 1-year mortality despite ICD therapy...Clinical markers of liver dysfunction, recent mechanical ventilation, and renal impairment were independently associated with increased 1 year mortality...	Weak condition match, condition-outcome pair not studied	PMV
CHIEF COMPLAINT: acute onset right hemiplegia and aphasia **PRESENT ILLNESS:** 84yo M...acute onset of inability to speak and right hemiplegia...head CT showed dense L MCA and hypodensities in left inferior frontal lobe and left corona radiata. **MEDICAL HISTORY:** HTN Afib, off coumadin...	Stroke is indicated by an abrupt manifestation of neurologic deficits secondary to an ischemic or hemorrhagic insult to a region of the brain...ranked as the third leading cause of death in the United States...report shows that despite the use of antithrombotic and/or antiplatelet aggregating drugs, the key to stroke management is primary prevention.	No cohort match, condition-outcome pair not studied	MOR
CHIEF COMPLAINT: Substernal chest pain **PRESENT ILLNESS:** ...62 yo M... no prior cardiac history... substernal CP... mild SOB, nausea, diaphoresis and numbness in left arm... **MEDICAL HISTORY:** foot surgery 2 weeks ago ?COPD ?gastritis?	..rising health care costs have created pressures to increase efficiency of coronary care units. Possible strategies seek to decrease resource use by identifying low-risk patients for initial triage or early transfer to lower levels of care...	No cohort match, no evidence for outcome	LOS <3 days

Table 12: Qualitative examples of retrieved literature that is categorized as unhelpful for cases where adding literature increases confidence in incorrect outcome. Case 1 shows an example of retrieved literature that has a weak match with patient condition, but no evidence linking condition to outcome. Case 2 shows an example in which retrieved literature does not match patient case or contain evidence for outcome. Case 3 shows an example of a review article that again does not match patient case or provide outcome evidence.



(a) PMV (b) MOR (c) LOS

Figure 4: Proportion of admission notes associated with the 100 most highly retrieved abstracts for each clinical outcome. From these graphs, we can see that frequently-retrieved abstracts for LOS are associated with a larger proportion of cases from the dataset, than frequently retrieved abstracts for PMV and MOR (indicative of lower literature diversity in LOS).