Towards Reducing Diagnostic Errors with Interpretable Risk Prediction

Anonymous ACL submission

Abstract

Many diagnostic errors occur because clinicians cannot easily access relevant information in patient Electronic Health Records (EHRs). In this work we propose a method to use LLMs to identify pieces of evidence in patient EHR data that indicate increased or decreased risk of specific diagnoses; our ultimate aim is to increase access to evidence and reduce diagnostic errors. In particular, we propose a Neural Additive Model to make predictions backed by evidence with individualized risk estimates at time-points where clinicians are still uncertain, aiming to specifically mitigate delays in diagnosis and errors stemming from an incomplete differential. To train such a model, it is necessary to infer temporally fine-grained retrospective labels of eventual "true" diagnoses. We do so with LLMs, to ensure that the input text is from *before* a confident diagnosis can be made. We use an LLM to retrieve an initial pool of evidence, but then refine this set of evidence according to correlations learned by the model. We conduct an in-depth evaluation of the usefulness of our approach by simulating how it might be used by a clinician to decide between a pre-defined list of differential diagnoses.

1 Introduction

A major source of poor patient outcomes and unnecessary costs in healthcare are missed or delayed diagnoses. A recent report estimated that diagnostic errors result in around 795,000 serious harms annually (Newman-Toker et al., 2023). Furthermore, many diagnostic errors result from information transfer problems (Zwaan et al., 2010). This is unsurprising given "note bloat", i.e., the widespread problem of information overload in EHR notes, often due to copied or irrelevant information which obfuscates relevant information. All of this motivates the potential of providing more efficient mechanisms to access relevant information in EHRs as a means to reduce these errors.

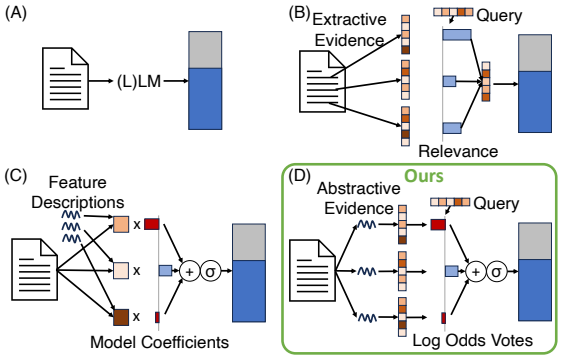

Figure 1: **Inherently "interpretable" approaches to prediction**. Typically, 'interpretable' models trade off between the expressiveness of intermediate representations and the faithfulness of the resulting interpretability to the models' true mechanisms. Our approach (D) manages to use very expressive intermediate representations in the form of abstractive natural language evidence while still maintaining true transparency during aggregation of this evidence. See Table 1 for more details.

One approach to helping practitioners make use of EHR is to train NLP models on free-text notes to provide predictions about patient risk for various illnesses (Rasmy et al., 2021; Li et al., 2021; Yang et al., 2023), but these systems are often lack transparency. Even when systems have high accuracy, clinicians may still prefer simple linear models as clinical decision support tools (Goldstein et al., 2016). Prior work has focused on developing inherently interpretable[1] models with minimal tradeoff in predictive performance, e.g., in the general domain with Neural Additive Models (Agarwal et al., 2020) and in healthcare with GA^{2}Ms (Caruana et al., 2015). Recently, zero-shot instruction-tuned LLMs have been shown capable of extracting information from clinical text (Agrawal et al., 2022), which in turn facilitates interpretable predictions (McInerney et al., 2023; Alsentzer et al., 2023).

[1]Interpretability is a famously ambiguous term; we are focused on having explicit measure of the contribution of individual pieces of evidence to an output.

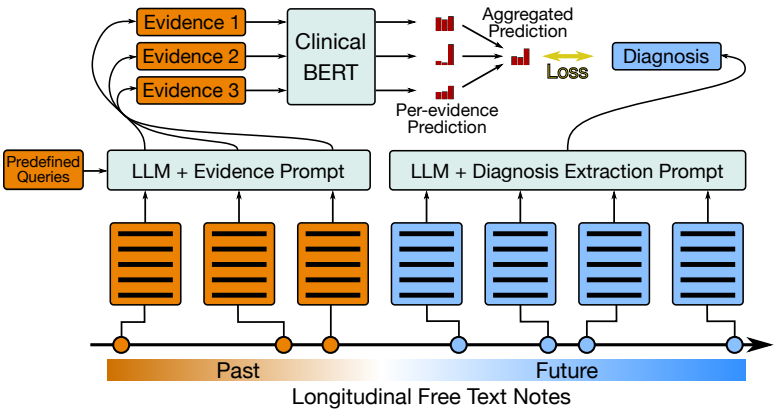

Figure 2: **Explainable Risk Prediction and Training.** An overview of our approach. Left: We retrieve evidence snippets from past notes with an LLM for predefined queries posed by a clinician. Then we use our risk prediction model to estimate risk of various diagnoses given each piece of evidence individually, and aggregate these scores. Right: We automatically extract diagnosis 'labels' from future reports with an LLM to use to train the risk predictor.

In this work, we combine the power and flexibility of zero-shot instruction-tuned LLMs with the transparency and modeling ability of Neural Additive Models (NAMs) to train a risk-prediction model that can also surface evidence to support predictions. We use an LLM (FLAN-T5-XXL; Chung et al. 2022) to generate abstractive "evidence" from EHR, which is then processed by a simpler model (Clinical BERT; Alsentzer et al. 2019) to produce features for a Neural Additive Model (Figure 2). This provides flexibility—the model can make inferences and condense information into fluent text snippets—but brings risk of "hallucinations".

We view this approach as "interpretable" in that it produces "evidence" in the form of human-understandable intermediate variables: abstractive text with associated risks, providing insights into factors important to a given prediction. Relative to other approaches to inherent interpretability (Figure 1), like those that use a relevance weights to weight and combine information from different sentences (B) and those that use large language model prompts to infer feature values (C), our approach permits greater flexibility in comparison to (C), while maintaining a more faithful interpretability in comparison to (B); see Table 1 for contrasts.

One complication is that we would like fine-grained, accurate labels to train our predictor (see section 4.1); ICD codes do not meet these criteria (Searle et al., 2020). Instead of ICD codes, which are noisy and temporally coarse (observed at the end of an encounter with discharge summaries), we propose to synthetically extract diagnosis labels from each report using an LLM. In some cases, this has been shown to be more aligned with true diagnoses (Alsentzer et al., 2023).

We focus our evaluation on how this system impacts clinical decision-making. Specifically, we examine settings where risk of misdiagnosis is high and the consequences severe. Our methods work within the confines of data present in electronic health record, which allows the model to be trained on any EHR. LLMs can be run locally and are only used for inference, so privacy and compute resources are not an issue.

Our contributions are summarized as follows:

Interpretable Risk Prediction with LLMs. We propose an approach to risk prediction that offers a particular form of interpretability in that it can expose faithful relationships between specific pieces of retrieved evidence and an output prediction.

Extracting Future Targets with LLMs. We present a method to extract target diagnoses for use in training from the unstructured text in the future of a patient's medical record that are more granular than ICD codes in the time dimension, and we validate with clinician annotations that the extracted labels are accurate.

In-depth Annotation of Usefulness. We validate how much evidence-wise interpretability can positively impact a clinician's expert judgement in high-impact settings which feature the greatest risk of misdiagnosis.

2 Dataset

We use MIMIC-III (Johnson et al., 2016a,b), an open-source dataset of EHRs from ICU patients. The ICU is one of the hospital settings (along with, e.g., the ER and Radiology) where misdiagnosis or

Modeling Approach	Intermediate representation(s)	Aggregation	Interpretability
(A) Direct **Black-box** Prediction (e.g., zero/few-shot, fine-tuned LLM)	None	CLS or last token embedding + classification or LM head	No inherent interpretability
(B) Aggregating chunked input with **relevance weights**	*Extractive* text snippets	Weighted avg. of CLS embeddings + class. head	Positive, real-valued relevance scores per query
(C) **Logistic regression** with LLM-inferred features	*Inferred*, real-valued numbers relating to predefined natural language queries	Logistic regression	Negative and positive real-valued static model coefficients
(D) **Log odds voting** with LLM-inferred text snippets (ours)	*Inferred/abstractive* text snippets relating to predefined natural language queries	Neural Additive Model (conditioned on the query/condition vector)	Negative and positive real-valued dynamic impact scores

Table 1: Types of interpretability afforded by the different modeling approaches for EHR data visualized in Figure 1. Red and green denote negative and positive aspects of each model.

delayed diagnosis are often caused by incomplete information, since clinicians typically do not have enough time to fully examine a patient's EHR.

In healthcare, cancer, infunction, and vascular dysfunction (termed the "big three") account for about 75% of all mis-diagnosis-related harms (Newman-Toker et al., 2023). Within the ICU, the latter two categories mostly manifest as pneumonia, and pulmonary edema (which in this paper we treat as interchangeable with congestive heart failure). For this reason, we will focus on predicting the risk of ICU patients for cancer, pneumonia, and pulmonary edema. These are also conditions for which clinical correlation with notes from the past EHR is important for diagnosis. We use all patients in the MIMIC dataset so that we have both negative and positive examples of the conditions.

3 An Interpretable Risk Prediction Model

We propose a multi-stage approach to risk prediction, capitalizing on a modern LLM, FLAN-T5-XXL (Chung et al., 2022; Wei et al., 2022) in this case, to implement each of the following steps.

Retrieval (Section 3.1). We generate abstractive evidence from free text notes by prompting an LLM with appropriate queries. The evidence snippets provide a form of interpretability, in that they can be inspected directly to verify predictions.

Risk Prediction (Section 3.2). We input the evidence into the risk predictor, which models relationships between the evidence and each of the potential diagnoses and outputs multi-label classification probabilities, i.e. the predicted risk that the patient will be diagnosed with each condition.

Evidence Re-ranking (Section 3.3). The retrieved evidence may still be too large a pool to review given the time constraints of the clinician. Therefore, we re-rank the evidence so as to only show that which promotes risk predictions that most deviate from the baseline risks of each condition.

To train risk prediction models we use use synthetic labels extracted from *future* notes in a patient's record (Section 4). Figure 2 provides an overview of our model and training approach.

3.1 Evidence Retrieval

Following prior work (Ahsan et al., 2023), we use a sequential prompting strategy to retrieve evidence that is relevant to a queried diagnosis or a risk factor. Specifically, we first ask the LLM for a binary response as to whether evidence for a condition exists; if the answer is affirmative, we then issue a second prompt tasking the LLM to generate supporting evidence. Formally, we define the evidence retrieved for report n and query q_i as follows:

$$e_{n,q_i} = \begin{cases} \text{GetEvidence}(r_n, q_i) \\ \quad \text{if EvidenceExists}(r_n, q_i) = \text{“yes”} \\ \texttt{null} \quad \text{otherwise} \end{cases} \tag{1}$$

where "GetEvidence" and "EvidenceExists" represent the corresponding prompt functions.

This approach does have limitations. For example, it cannot produce more than one snippet of evidence per report/query pair. Retrieved evidence may also be abstractive rather than extractive, which introduces the risk of model "hallucinations", but permits flexibility and interpretability (Ahsan et al., 2023). It also significantly reduces the amount of text (therefore requiring a relatively small context window) by going from all reports to sentence-length snippets for some reports. The resulting "summarization" in the form of evidence snippets is also controllable through the querying process and works zero-shot, i.e., it requires no

specialized or in-domain training. Queries, written by a clinician co-author, are in appendix Figure 9.

3.2 Risk Prediction

Because a patient can have more than one diagnosis, we treat risk prediction as a multi-label classification problem where each label corresponds to a diagnosis. To realize interpretability, we use a Neural Additive Model (Agarwal et al., 2020). Specifically, we do not model *interactions* between evidence snippets. Instead, we predict scores individually for each piece of evidence, and average these[2] to obtain a logit for risk prediction:

$$p(\hat{y}_i = 1|e_{1:E}) = \sigma(b_i + w_i \cdot (\frac{1}{E}\sum_{j=1}^{E} f_\theta^{\text{BERT}}(e_j))) \quad (2)$$

where $w_i \in \mathbb{R}^d$ is the embedding of diagnosis i, $e_{1:E}$ is the flattened list of evidence snippets with `null` evidence omitted, f_θ^{BERT} is the ClinicalBERT (Alsentzer et al., 2019) `[CLS]` embedding function (which yields a d-dimensional vector), and $b_i \in \mathbb{R}$ is the bias for diagnosis i. The prior over conditions can be defined as the same equation excluding the evidence term: $p(\hat{y}_i) = \sigma(b_i)$, and the **relative risk** follows as $p(\hat{y}_i|e_{1:E})/p(\hat{y}_i)$.

While the bias could be learned, we instead simply set it to the inverse sigmoid of the observed prevalence of the disease in the training sample distribution: $b_i = \sigma^{-1}(\text{prevalence}_i^{\text{train}})$. This means that if we wanted to transfer the model to a new population, where the prevalence differed but the contributions of different evidence were assumed to remain, we could simply update the b_i term.

Excluding interactions between evidence snippets is a sacrifice in model complexity, but it also allows us to compute an interpretable "vote" for any individual piece of evidence as

$$p(\hat{y}_i|e_j) = \sigma(b_i + w_i \cdot f_\theta^{\text{BERT}}(e_j)) \quad (3)$$

and compute an individualized relative risk for each piece of evidence using this value.

Conveniently, forcing the bias term to be the inverse sigmoid of the training prevalence, by definition, also means we can interpret the evidence term in Equations 2 and 3 as the **log odds ratio**, i.e., the difference between the logits when conditioning vs. not conditioning on the evidence. The model is effectively estimating this log odds ratio directly. This variable's expected value does not change if we sample conditions for training with a frequency different from the the natural prevalence of the conditions (Simon, 2001). Because of this, we can estimate the likelihood and the relative risk during inference on a differently sampled population by simply changing the bias term in the prior and in equations 2 and 3 to reflect the estimate of the natural prevalence of the conditions (Zhang and Kai, 1998), which we can get from the training set before sampling: $b'_i = \sigma^{-1}(\text{prevalence}_i^{\text{train}})$.

3.3 Evidence Re-ranking

Because of the simplicity of the risk prediction, we can use the internal variables it exposes to re-rank evidence. The intuition behind the re-ranking is that the most important evidence will be that which most changes our risk assessment from the prior over the diagnoses, and we would like the chosen metric to capture this across all of the potential diagnoses. We use Mean Squared Error (MSE) of the predicted logits with the logits of the prior $p(y)$. This makes the formulation of the MSE metric simple as the mean (over Q conditions) of the squares of the log odds ratio for a piece of evidence:

$$\text{MSE}(\sigma^{-1}p(\hat{y}|e_j), \sigma^{-1}p(\hat{y})) = \frac{1}{Q}\sum_{i=1}^{Q}(w_i \cdot f_\theta^{\text{BERT}}(e_j))^2. \quad (4)$$

It is necessary to use the *log odds* ratio term in this score function because we care not only about increasing but also about decreasing the probability of a condition, so it makes most sense to compare and sum these two different effects in log space. The reason to choose MSE over other scores (e.g. the absolute distance) comes from the intuition that it is more important to see the evidence that is "very opinionated" about one condition rather than to see evidence that is "slightly opinionated" about many. Therefore, it is necessary to square this log odds ratio before averaging across conditions to reflect this idea when sorting evidence.

4 Certain Diagnosis Extraction

We make an assumption about the EHR of patients that eventually receive a diagnosis that there is some period of time in the record where a diagnosis is "uncertain" before it becomes "certain", and the eventual "certain" diagnosis is correct. Of course

[2]Neural Additive Models typically use a sum instead of an average, but we found that given varying amount of evidence retrieved, it worked better to use an average.

just because a diagnosis is definitive as noted by clinician in the record does not necessarily mean that it is correct—sometimes clinicians are wrong.

However, it is hard to detect such cases, so here we focus on reducing delayed diagnosis errors where we assume some evidence in the medical record from that "uncertain" period could have influenced a clinician to make a diagnosis or order a certain kind of test sooner than they did, or keep a diagnosis in the running list of differentials for longer. If notes are incorporated into the input where the diagnosis is already certain, the prediction problem becomes too easy, which is why a time-wise fine-grained label is necessary—such a label could more accurately weed out all of this obvious evidence. To extract these certain diagnoses with an LLM, we use three sequential prompts and a normalization step.

4.1 3-Stage Extraction with LLMs

In this section we describe the prompts for certain diagnosis extraction, which are shown in full in the appendix (Section B). Following prior work (Ahsan et al., 2023), we first prompt the LLM with a binary question asking if there exists a confident diagnosis for a patient. If the answer is "yes", we then ask the model for the diagnoses. Unfortunately, creating a list of diagnosis terms from the answer to this prompt is not just a matter of parsing because we found that the model will often return extended phrases that are not easily mapped to diagnoses. Therefore, we issue one more prompt that only takes in the output of the previous prompt to create a structured list of diagnostic terms. We then parse this final output of the LLM into a list of strings.

4.2 Normalization

To normalize produced diagnostic terms, we take a two-step apporach. First we use string matching heuristics to handle easy cases. Then we embed sentences with `SentenceTransformers` (Wang et al. 2020; Reimers and Gurevych 2019; specifically, `all-MiniLM-L6-v2`) and calculate cosine similarities, matching a term in the parsed list to the most similar term (with similarity $>.85$) in the predefined set ("cancer", "pneumonia", and "pulmonary edema"). We ignore terms with no match.

5 Evaluation

Because our targets are synthetically generated using an LM, we first evaluate how well our labels align with the "ground truth" (Section 5.1). Next, we aim to evaluate how well the model can realistically help with risk prediction. Though it is straightforward to assess the accuracy of the risk prediction itself—we use the standard metrics of precision, recall, F1 and AUROC scores to compare to various uninterpretable baselines—it is not as easy to assess what we really care about: How helpful is the interpretability offered by the proposed model to clinicians (section 5.2)? For this we resort to manual evaluation by our clinical coauthors and develop bespoke interfaces to facilitate annotation.

5.1 Future Target Extraction

To evaluate how well the LLM extracts targets in the form of "confident" diagnoses, we enlist our clinical collaborators to annotate the precision with which the LLM infers "confident" diagnoses. In particular, for every report where one of the three diagnoses—cancer, pneumonia, and pulmonary edema—was automatically extracted, an ICU clinician is first tasked with answering the question "Is [diagnosis] a confident diagnosis of the patient according to the report?". If the answer is "yes", they are asked: "Is it likely that this confident diagnosis could be identified in earlier reports?".

5.2 Risk Prediction Interpretability

To assess the viability of clinicians using this model in practice, we collect in-depth annotations intended to simulate the real-world use of this technology. We evaluate a number of baseline models and model ablations to assess the relative benefits of different model components.

Interface and Annotations To conduct annotations, we develop an interface that simulates as closely as possible the envisioned use case: A clinician is seeing an ICU patient's chart for the first time and trying diagnose the patient or determine what they are at risk of. The clinician may not have much time to spend with the patient's chart, so we ask clinician annotators to work quickly—specifically, to try and keep annotation time to a few minutes—and we record the amount of time they take to review the patient's record. When they are done, the annotation process starts, and though they are allowed to access the patient's notes, they are encouraged not to.

We first ask if a diagnosis is noted explicitly in the patient's record. Given that we are aiming to

evaluated records where the diagnosis is not yet clear, we skip the rest of the annotations on the instance if a diagnosis is explicit. If not, we ask for estimates of the likelihood ("unlikely", "somewhat likely", or "very likely") of each of the possible conditions. Note that we explicitly do not show any model predictions until after this question, to avoid bias. Then, we show the annotator the model predictions and ask if the predicted risk for the conditions aligns with intuition.

Moving onto the evidence (appendix Figure 13), we allow the annotator to look at the sorted evidence one snippet at a time along with the individualized risk prediction only based on that snippet. The annotator notes the usefulness of the evidence with respect to each condition. If the evidence is useful, they are asked whether or not the impact of this evidence on the risk scoring (for the particular condition) aligns with intuition, and whether the annotator remembers seeing this piece of evidence during their initial review of the patient's notes. After two pieces of evidence, if the annotator feels like more evidence is needed to form a reasonable opinion of the patient's risk, they can request more evidence snippets (up to a maximum of 10), annotating each as they go. Finally, the annotator is asked if any of the evidence presented impacted their original assessment of likelihood.

Ablations While the task of risk prediction is standard, there is less work on the the task of surfacing relevant evidence (abstracted or extracted) to support such predictions. Consequently, there is not a large set of baselines to serve as natural comparators to our approach. Therefore, in our analysis we focus on showing the importance of each component of our model through ablations. We can decompose our approach into two evidence retrieval components, generating the evidence, which we refer to as "**LLM Evidence**" and reranking it, which we refer to as "**Log Odds Sorting**". The following ablations show the importance of both of these components in identifying useful evidence.

We use prior work (Ahsan et al., 2023) as a starting point for generating the evidence, so it is natural to ask what that component can do by itself without re-ranking using the risk prediction scores for each piece of evidence. A natural comparison is to present the same evidence retrieved but in a *random* or *reverse chronological* order (as recency is probably important). But we can also use the model certainty in evidence, given that this has been shown to correlate with the utility of snippets (Ahsan et al., 2023). We adopt this approach for comparison and call it "**Confidence Sorting**".

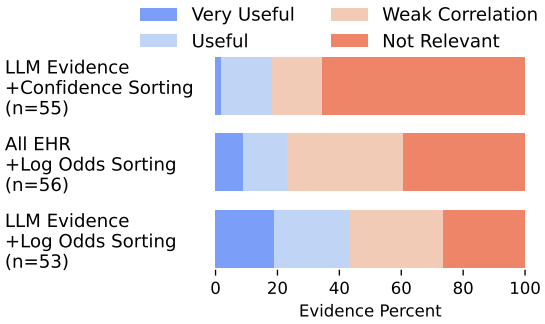

Figure 3: **Evidence Usefulness** (the maximum score across conditions) for our approach and two ablations. "LLM Evidence+Confidence Sorting" uses model evidence, but sorts by (length-normalized) log probability instead of the log odds. "All EHR+Log Odds Sorting" does not use LLM evidence and instead takes the last 1000 sentences in the record as evidence.

It is also natural to question the importance of using the language model to abstractively generate evidence at all. We might instead simply use every sentence in the report as evidence and train our prediction model with this retrieved evidence, reranking it in the normal way ("Log Odds Sorting") with the prediction model's scores. We call this the "**All EHR**" model.

6 Results and Discussion

The majority of our results are based on annotations from 4 annotators on 24 instances and 3 models. Each instance has a maximum of 3 annotators, each annotating different models (assigned randomly). Table 2 reports detailed statistics.

Our main goal is to understand if our approach can retrieve better evidence. To this end, we plot the percentage of evidence annotated in each category of usefulness for each model in Figure 3. Though we record usefulness for each condition individually, here we combine these annotations by taking the maximum score across the conditions for each piece of evidence. The results highlight the necessity of both the "**LLM Evidence**" retrieval component and the "**Log Odds Sorting**" method, as both other variants retrieve significantly less "Useful" and "Very Useful" evidence and more "Weakly Correlated" and "Not Relevant" evidence.

How much of the relevant retrieved evidence is

Annotator	LLM Evidence+Confidence Sorting				All EHR+Log Odds Sorting				LLM Evidence+Log Odds Sorting			
	Inst.	Evid.	Rep.	Percent Useful	Inst.	Evid.	Rep.	Percent Useful	Inst.	Evid.	Rep.	Percent Useful
1	8	20	195	5.0	5	14	81	7.1	6	13	154	30.8
2	2	6	26	50.0	2	5	72	40.0	5	14	162	50.0
3	4	13	105	23.1	6	17	224	35.3	5	14	119	50.0
4	5	16	132	18.8	6	20	127	20.0	4	12	85	41.7
Aggregated	19	55	458	24.2	19	56	504	25.6	20	53	520	43.1

Table 2: Annotations. We report the statistics for the number instances annotated, the amount of evidence snippets annotated, the total number of reports in the annotated instances, and the percent of evidence annotated as "Useful" and "Very Useful". Aggregated statistics are computed by summing over the annotators except in the case of "Percent Useful", where scores are macro-averaged over annotators. (This is slightly different from Figure 3 where percentages are macro-averaged, i.e., we combine all annotated evidence).

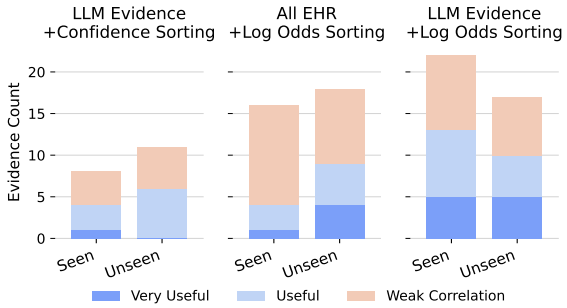

Figure 4: **Seen vs. unseen** evidence counts for all evidence that at least weakly correlates with a condition.

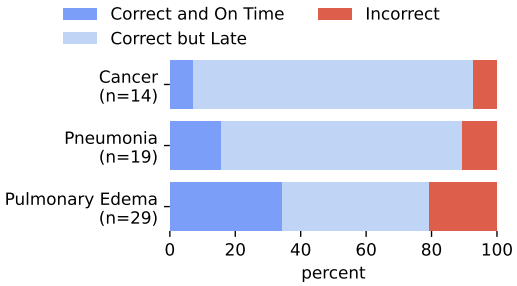

Figure 5: Synthetic label precision. For each confident diagnosis label extracted by the system, annotators check whether the diagnosis actually appears in the report (and is definitive), and subsequently if subjectively they believe that report is likely the *first* time the diagnosis was definitive based on the report language.

redundant with the information already uncovered during the annotator's initial review of the patient? We plot evidence counts separately for seen vs unseen evidence in Figure 4 and find that there is a significant amount of unseen evidence that is useful and very useful in all models.

The rated usefulness of evidence does not necessarily matter if it does not affect the clinician's decision. An example of how these models might work in practice is when our LLM Evidence model with Confidence Sorting surfaced the following: "Atrial fibrillation with rapid ventricular response. Compared to the previous tracing atrial fibrillation is seen. Other findings are similar. The patient is at risk of pulmonary edema." In this case the annotator changed their estimate of the likelihood of pulmonary edema from unlikely to somewhat likely, and it turns out that pulmonary edema did appear in a future report.

We show all 7 instances where annotators changed their mind after viewing evidence in Appendix Table 5. Of these we find 2 instances (including the example above) where annotators' increased their likelihood of conditions that were extracted from future records, and 5 where condition(s) other than the synthetically labeled condition(s) were affected (mostly by increasing the annotators' risk assessments). Though more data should be collected, this indicates the model might improve annotator recall (though at some cost in precision); recall is arguably more important here.

Given that we are using synthetic labels of future diagnoses for both training and evaluation for risk prediction (discussed next), it is important to evaluate how well our labels align with ground truth. Given that ICD codes are not fine-grained enough and are not always accurate, we turn to manual annotations of precision for this evaluation. In Figure 5, we report the precision of these labels for being correct or for being "correct and on time". This second category is a stronger correctness in which the annotator also noted that the note where the label was detected subjectively seems to be the first note where that label should have been given as judged using the phrasing in the note.[3]

We see reasonable precision when using automatic labeling with the LLM pipeline (about 80 percent and above for all conditions). We also

[3]It would be time-consuming to annotate this directly because it involves looking at a lot of prior notes.

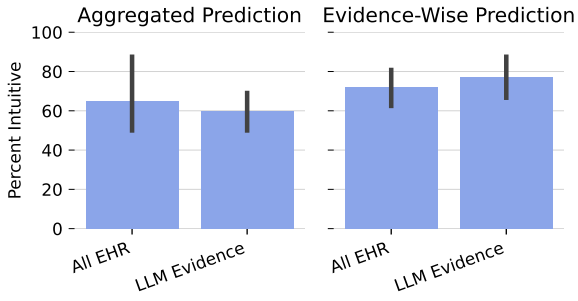

Figure 6: Intuitiveness of predictions macro-averaged across annotators.

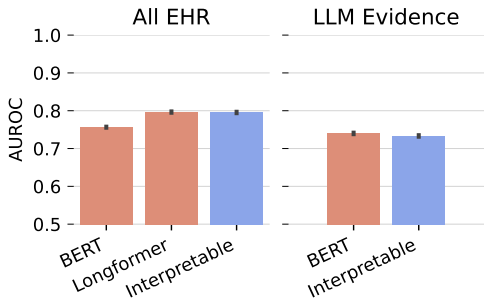

Figure 7: Risk Prediction Performance evaluated on synthetic labels and averaged over 5 random seeds for choosing the which time-point in the EHR to use prior to the diagnosis label. Error bars represent standard deviation of the random seeds. Here, BERT and Longformer refer to Clinical BERT and Clinical Longformer.

compute inter-annotator agreement for these annotations of precision across the 4 annotators by enforcing that 8 annotated predictions overlap for all the annotators. The Fleiss' Kappa score for these synthetic label annotations was .68 for the 3-category classification shown in Figure 5 and .86 for the 2-category classification obtained by simplifying the labels into just "Correct" or "Incorrect".

We would also like to assess how well our models' risk estimates aligns with the intuitions of clinicians with respect to the aggregated and individual predictions. Though for the aggregated prediction for an instance, we ask annotators to take the magnitude of the risk, not just the direction (i.e. increased compared to baseline or decreased compared to baseline) into account, for evidence-level predictions, we ask annotators to take the magnitude with a grain of salt and mostly judge based on the direction. This is because the magnitudes appeared to be somewhat artificially inflated potentially either due to the strong evidence trying to "componsate" for the evidence that does not actively contribute to the log odds (see Figure 12) or because of the sorting method.[4] Figure 6 shows that both models do reasonably well with respect to the aggregated prediction and the evidence-wise predictions, and both do slightly better on evidence-wise predictions than aggregated predictions.

Finally, it is important to evaluate the actual prediction performance of our models on our synthetic labels. Here we also compare against baseline models that are not interpretable: BERT and Longformer. These black-box models are trained on both the All EHR and the concatenated retrieved LLM evidence. Figure 7 shows that including all evidence usually helps prediction performance, but using the blackbox vs interpretable models on the same input does not effect performance.

[4]Future work might investigate how to bring make this *magnitude* more interpretable.

7 Conclusions

Clinicians should have access to all the pertinent information to make well-grounded decisions for diagnosing a patient, but currently they are inundated with (unstructured) information from the EHR. This is exacerbated by the time constraints faced by practitioners. We have proposed an approach that aims to facilitate efficient access to potentially important data within EHR; our method capitalizes on the capabilities of LLMs to produce digestable, abstractively generated text evidence, which is then consumed by a Neural Additive Model (NAM) to yield a prediction.

We find that using NAMs does not sacrifice predictive quality, but does enable models to surface useful evidence to clinicians. Using the LLM to create the starting set of evidence to feed into the NAM does sacrifice some performance, but it also significantly increases the usefulness of the evidence in comparison with using the raw sentences from EHR notes as evidence.

Further, we find that in some cases the surfaced evidence is able to change a clinician's mind, increasing the clinician's recall though decreasing precision, which warrants future work to improve on this system. One major concern is that this type of system could increase clinician's workload rather than decrease it. Future work should assess exactly how and when it might be beneficial to show snippets to clinicians.

8 Limitations

Though the proposed approach of combining abstractive LLM evidence with Neural Additive Models shows promise, there are still many concerns that need to be addressed in future work. One of the biggest concerns is about the use of abstractive "evidence" produced by LLMs. Hallucinations in this evidence could at best negatively impact trust of clinicians in the system and at worst mislead clinicians and negatively affect patient outcomes. Our work does not directly study this given the substantial extra annotator time needed to check for hallucinations. We also did not experiment much with different prompts or models for producing this evidence given that our main focus was on validating the system-level approach rather than individual components.

Another limitation concerns the lack of a significant number of baseline models. Though not many baselines exist for a task that involves retrieving evidence supporting predictions in EHR, there are still potential baselines that use relevance weights or cosine similarity with clinical BERT that we could have included. However, due to the extensive amount of time needed for just one annotation on one model, we chose to focus on ablating over the LLM evidence retrieval and sorting method components of the model.

Finally, our analysis mostly relies on a relatively small amount of annotations from one dataset. This again stems from the time cost of annotations. Each annotator must first look through a whole patient's record to get a sense of the patient before even getting to any annotations. On average, this took almost 3 minutes, which is all before annotators even see any of the questions. Then, because the study focuses on just the top evidence presented for each instance, each annotator only annotates 3.2 evidence snippets on average per instance. This time-consuming process did limit the number of annotations we could obtain.

References

Rishabh Agarwal, Nicholas Frosst, Xuezhou Zhang, Rich Caruana, and Geoffrey E. Hinton. 2020. Neural additive models: Interpretable machine learning with neural nets. *ArXiv*, abs/2004.13912.

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hiba Ahsan, Denis Jered McInerney, Jisoo Kim, Christopher Potter, Geoffrey Young, Silvio Amir, and Byron C. Wallace. 2023. Retrieving evidence from ehrs with llms: Possibilities and challenges.

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.

Emily Alsentzer, Mary-Jette Rasmussen, Raíssa Schmitt Fontoura, Andrew Cull, Brett K. Beaulieu-Jones, Kathryn J. Gray, D. Bates, and Vesela P. Kovacheva. 2023. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. *NPJ Digital Medicine*, 6.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY, USA. Association for Computing Machinery.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John P A Ioannidis. 2016. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24(1):198–208.

Alistair E. W. Johnson, Tom J. Pollard, and Roger G. Mark. 2016a. MIMIC-III clinical database (version 1.4).

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. 2016b. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. 2021. Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE journal of biomedical and health informatics*, 27:1106 – 1117.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Denis McInerney, Geoffrey Young, Jan-Willem van de Meent, and Byron Wallace. 2023. CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8477–8494, Singapore. Association for Computational Linguistics.

David E Newman-Toker, Najlla Nassery, Adam C Schaffer, Chihwen Winnie Yu-Moe, Gwendolyn D Clemens, Zheyu Wang, Yuxin Zhu, Ali S. Saber Tehrani, Mehdi Fanai, Ahmed Hassoon, and Dana Siegal. 2023. Burden of serious harms from diagnostic error in the usa. *BMJ Quality & Safety*.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Thomas Searle, Zina Ibrahim, and Richard JB Dobson. 2020. Experimental evaluation and development of a silver-standard for the mimic-iii clinical coding dataset. *arXiv preprint arXiv:2006.07332*.

Stephen D Simon. 2001. Understanding the odds ratio and the relative risk. *Journal of andrology*, 22(4):533–536.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers.

Jason Wei, Maarten Paul Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew Mingbo Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Zhichao Yang, Avijit Mitra, Weisong Liu, Dan Berlowitz, and Hong Yu. 2023. Transformehr: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nature Communications*, 14.

Jun Zhang and F Yu Kai. 1998. What's the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes. *Jama*, 280(19):1690–1691.

Laura Zwaan, Martine de Bruijne, Cordula Wagner, Abel Thijs, Marleen Smits, Gerrit van der Wal, and Daniëlle R. M. Timmermans. 2010. Patient Record Review of the Incidence, Consequences, and Causes of Diagnostic Adverse Events. *Archives of Internal Medicine*, 170(12):1015–1021.

A Description of terms.

Table 3 shows all of the terms used to describe different models and settings.

B Certain Diagnosis Extraction Prompts

Prompt 1:

> Read the following report:
>
> <input>
>
> Question: Is there a confident diagnosis of the patient's condition? Choice: -Yes -No
> Answer:

Prompt 2:

> Read the following report:
>
> <input>
>
> Answer step by step: What is the correct diagnosis of the patient's condition?
> Answer:

We use Chain of Thought (CoT) prompting here because—similar to the evidence retrieval step—we want the model first to extract the parts of the report that refer to a diagnosis, as this seems to work better than going straight to the list of diagnoses. In initial experiments, using the CoT prompt appeared to more easily elicit these verbose extractions.

Prompt 3:

> Here is a diagnosis of a patient:
>
> <confident diagnosis>
>
> Question: Provide a list of diagnostic terms or write none.
> Answer:

C Prompting Problems

In our 3-stage prompting process, we initially had some problems with false positives in scenarios where pneumonia was negated (Figure 8). We discovered that this was because our 3nd prompt was originally:

> Here is a diagnosis of a patient:
>
> <confident diagnosis>
>
> Question: Based on this diagnosis, provide a list of diagnostic terms.
> Answer:

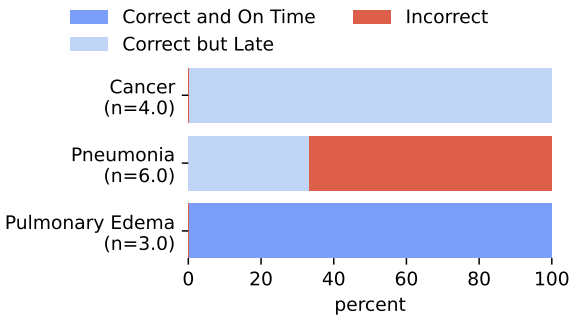

Figure 8: Synthetic labels on validation examples before correcting the prompting problem.

This particular prompt sometimes produced positive synthetic labels for pneumonia when pneumonia was actually negated in the confident diagnosis generated by the previous prompt. We realized this when starting to annotate validation examples, so we changed our prompt (see section 4.1). All of the test annotations reported in the main paper do not include or overlap patients with these annotated validation examples.

D Experiments

We use Clinical BERT for the NAM prediction model. For all models, we train for up to 10 epochs on one Quadro RTX 8000 GPU and pick the best checkpoint (where checkpoints occur every 5 percent of an epoch). For the LLM for both evidence retrieval and synthetic label extraction we use FLAN-T5-XXL (Chung et al., 2022; Wei et al., 2022). In the case of All EHR used as input to the NAM, we split sentences with NLTK. We will make code open-source on acceptance.

E Usefulness of Queries

Unlike (Ahsan et al., 2023), we do not directly evaluate how relevant the retrieved evidence is to the query used to retrieve it; we instead focus on how relevant the evidence is to the risk predictions. However, we would like to examine which queries produce useful evidence. Figure 9 shows counts of evidence in each category separated across which query was used to retrieve that evidence. It seems as though the most useful evidence came from the three queries that directly ask about the condition for which we are predicting risk (the three left-most queries), but a few additional queries sometimes did prove useful.

LLM Evidence	Models that use the evidence retrieved with an LLM.
All EHR	Models that use the all of the text in the EHR. For Interpretable Neural Additive Model, this text is split at the sentence level.
BERT or Longformer	Blackbox models that take either All EHR or LLM Evidence (concatenated) as input. BERT refers to Clinical BERT (Alsentzer et al., 2019) and Longformer refers to Clinical-Longformer (Li et al., 2023).
Interpretable	The proposed Interpretable Neural Additive Model, which can operate either on LLM Evidence or All EHR inputs.
Confidence Sorting	Sorting LLM Evidence by the length-normalized log-likelihood of the evidence under the LLM.
Log Odds Sorting	Sorting either LLM Evidence or All EHR inputs by the mean squared error of the predicted log odds (equation 4).

Table 3: Description of terms.

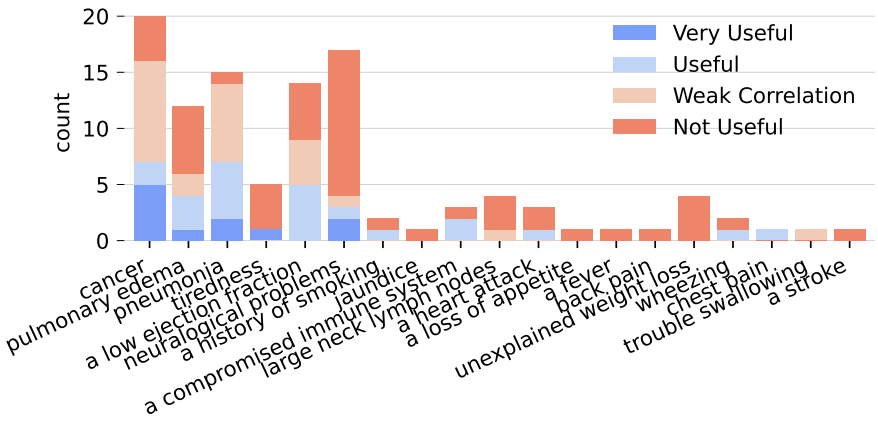

Figure 9: Usefulness per Query

F Full Prediction Performance

We report the full prediction performance in Table 4.

G Annotators Changing Their Minds

Table 5 presents all the occurrences of annotators changing their mind.

H Ablation over amount of evidence used

Figure 10 shows performance if we limit to a set number of evidence that can be used in the Neural Additive Model's final aggregated score.

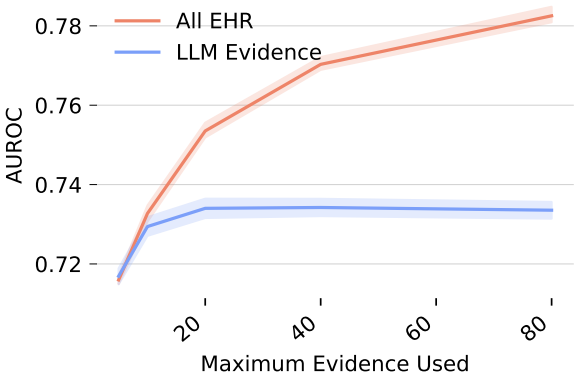

Figure 10: Ablation over amount of evidence used to make a risk prediction.

I Evidence Histograms

Figure 11 shows a histogram of the amount of evidence per each instance, and Figure 12 shows what the distribution over the log odds votes looks like.

J Annotation Interface

Figure 13 shows a screenshot of what the part of the interface dedicated to annotating evidence looks like.

	AUROC	Precision	Recall	F1
BERT (All EHR)	75.6 ± .19	65.6 ± 1.38	16.8 ± .38	26.8 ± .43
Longformer (All EHR)	79.6 ± .22	55.5 ± .32	28.8 ± .43	37.9 ± .38
Interpretable (All EHR)	79.5 ± .23	56.5 ± .57	20.5 ± .58	30.1 ± .60
BERT (LLM Evidence)	74.0 ± .27	51.6 ± 1.32	22.7 ± .27	31.5 ± .42
Interpretable (LLM Evidence)	73.3 ± .27	53.6 ± 1.09	15.0 ± .36	23.4 ± .48

Table 4: **Risk Prediction Performance** on the synthetic labels averaged over 5 different random seeds used for choosing the time-point in each patient that separates the past from the future.

Annotator	Model	Sorting	Changes	Best Evidence	Usefulness	Synthetic Label
2	LLM Evidence	Confidence Sorting	Pneumonia: Unlikely → Somewhat likely	There is a small right pneumothorax. There is extensive consolidation of the right upper lobe. Consolidation in the right lower lobe is mostly located in the superior segment. The left lung is grossly clear. There. Signs: There is extensive consolidation of the right upper lobe. Consolidation in the right lower lobe is mostly located in the superior segment. The left lung is grossly clear. There is no left pleural effusion. There is	Useful for **Pneumonia**	Pneumonia
4	LLM Evidence	Confidence Sorting	Pulmonary Edema: Unlikely → Somewhat likely	Atrial fibrillation with rapid ventricular response. Compared to the previous tracing atrial fibrillation is seen. Other findings are similar. The patient is at risk of pulmonary edema.	Useful for **Pulmonary Edema**	Pulmonary Edema
3	All EHR	Log Odds Sorting	Cancer: Unlikely → Very likely	Basal cell skin ca. [**27**].	Useful for **Cancer**	Pulmonary Edema
4	All EHR	Log Odds Sorting	Cancer: Unlikely → Somewhat likely	o.b.resident to see pt., pt.waiting for a "biopsy".	Useful for **Cancer**	Pulmonary Edema
4	All EHR	Log Odds Sorting	Pulmonary Edema: Somewhat likely → Unlikely, Pneumonia: Somewhat likely → Very likely	There is increased opacity in the. retrocardiac left lower lobe, as well as the right lower lobe, which could be. due to atelectasis, aspiration, or possibly pneumonia.	Very Useful for **Pneumonia**	
1	LLM Evidence	Log Odds Sorting	Pneumonia: Somewhat likely → Very likely	CXR showed L middle/lower lobe PNA, prob asp PNA.	Very Useful for **Pneumonia**	
4	LLM Evidence	Log Odds Sorting	Cancer: Unlikely → Very likely	CLL. Signs: id: pmh of CLL	Very Useful for **Cancer**	

Table 5: Examples of the 5 instances where annotators changed their mind based on evidence shown.

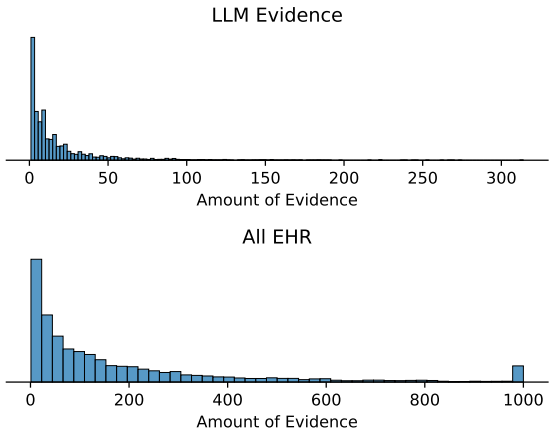

Figure 11: Histogram of the number of text snippets for each instance.

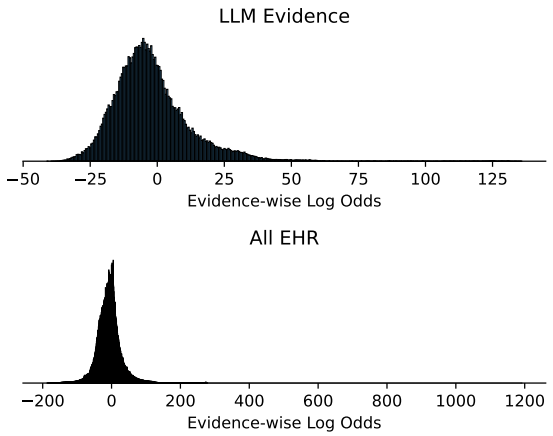

Figure 12: Histogram of the log odds of each individual piece of evidence.

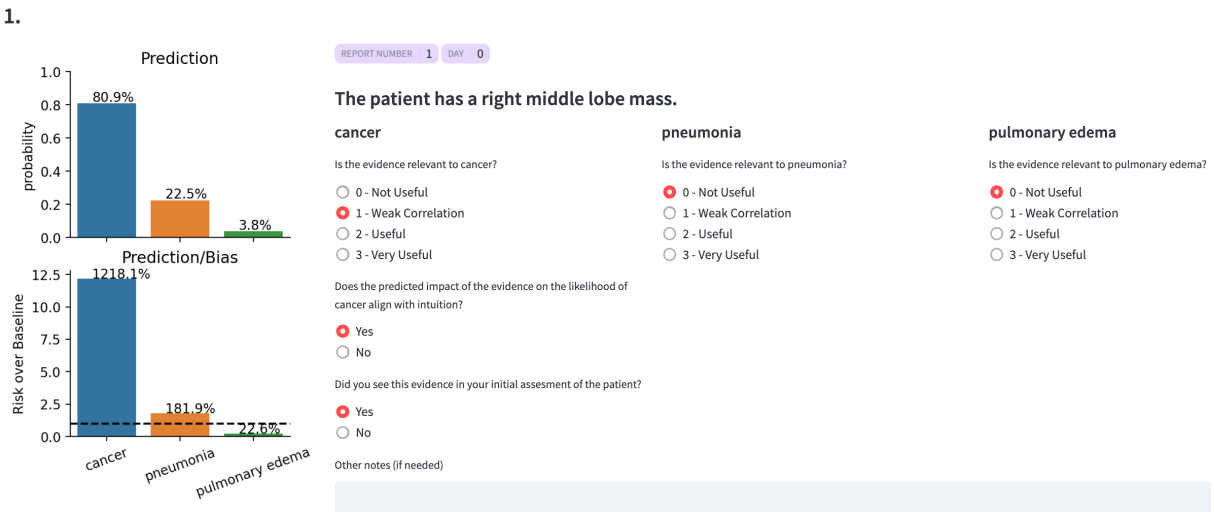

Figure 13: An example part of the **evidence** annotation interface. The plots on the left indicate the predicted likelihood (top) and the odds ratio (bottom).