Model Distillation for Faithful Explanations of Medical Code Predictions

Zach Wood-Doughty[1,2] *, **Isabel Cachola**[1]*, **Mark Dredze**[1]
[1] Johns Hopkins University, Baltimore, MD 21211
[2] Northwestern University, Evanston, IL 60208
zach@northwestern.edu, icachola@cs.jhu.edu, mdredze@cs.jhu.edu

Abstract

Machine learning models that offer excellent predictive performance often lack the interpretability necessary to support integrated human machine decision-making. In clinical or other high-risk settings, domain experts may be unwilling to trust model predictions without explanations. Work in explainable AI must balance competing objectives along two different axes: 1) Models should ideally be both *accurate* and *simple*. 2) Explanations must balance *faithfulness* to the model's decision-making with their *plausibility* to a domain expert. We propose to use knowledge distillation, or training a student model that mimics the behavior of a trained teacher model, as a technique to generate faithful and plausible explanations. We evaluate our approach on the task of assigning ICD codes to clinical notes to demonstrate that the student model is faithful to the teacher model's behavior and produces quality natural language explanations.

1 Introduction

Machine learning (ML) methods have demonstrated predictive success in medical settings, leading to discussions of how ML systems can augment clinical decision-making (Caruana et al., 2015). However, a prerequisite to clinical integration is the ability for healthcare professionals to understand the justifications for model decisions. Clinicians often disagree on the proper course of care, and share their justifications as a means of agreeing on a treatment plan. Explainable Artificial Intelligence (AI) can enable models to provide the explanations needed for them to be integrated into this process (Lundberg et al., 2018; Caruana et al., 2015). However, modern AI models that often rely on complex deep neural networks with millions or billions of parameters pose challenges to creating explanations that satisfy clinician's demands (Feng et al., 2018).

Similar concerns over model explanations across domains have inspired a whole field of interpretable ML. Work in this area typically considers two goals: faithfulness (explanations that accurately convey the decision-making process of the model) and plausibility (explanations that make sense to domain experts) (Jacovi and Goldberg, 2020). These two goals may be in conflict: faithful explanations that accurately convey the reasoning of complex AI systems may be implausible to a domain expert, and vice versa (Kumar and Talukdar, 2020; Wiegreffe et al., 2021). Models must also balance performance against transparency. The methods that perform best on a task may be unable to provide explanations (Rudin, 2019).

We propose to disentangle these competing goals by using knowledge distillation. We train a bag-of-words linear student model to predict the *predictions* of the teacher model, so that the behavior of the student model mimics the teacher model's behavior, rather than independently modeling the target task. We then rely on the interpretable student to create explanations without changing the original teacher model. We evaluate the student's faithfulness to the teacher model and the plausibility of the student's explanations.

We demonstrate our approach on the task of medical code prediction. While ML methods have achieved predictive success on various versions of International Classification of Diseases (ICD) clinical code assignment, the best-performing methods have been neural networks that are notoriously difficult to interpret. We train student models for three teacher models: (1) DR-CAML, a method designed to produce explainable predictions which outperformed several baselines when evaluated by a clinical expert (Mullenbach et al., 2018); (2) Hierarchical Attention Networks, a Bi-GRU document classifier first introduced by Yang et al. (2016) and adapted to ICD code prediction by Dong et al. (2021); and (3) TransICD, a transformer-based

* Equal contribution

Proceedings of the BioNLP 2022 workshop, Dublin, Ireland, pages 412–425
May 26, 2022.

method (Biswas et al., 2021). We show that our student models are faithful to the teacher models and can generate natural language explanations that are comparably plausible. We also show that our student model outperforms a logistic regression baseline in comparison to the true ICD-9 labels, despite being of equal complexity. We release the code under an MIT license for both our method and for reproducing Mullenbach et al. (2018).[1]

2 Background

2.1 Interpretable ML

Interpretable machine learning falls within the growing field of Explainable AI (Doshi-Velez, 2017). We present an overview of major themes in the literature, and direct the reader to recent surveys for more details (Doshi-Velez, 2017; Guidotti et al., 2018; Gilpin et al., 2018).

Past work distinguishes between "transparent" or "inherently interpretable" models that offer their own explanations, and "post-hoc" methods that produce explanations for a separately-trained model. Linear methods such as logistic regression are often considered transparent, while deep neural networks are generally not and rely on post-hoc methods for explainability (Guidotti et al., 2018; Feng et al., 2018). However, even simple models can prove difficult to interpret, e.g., when the model's features are complex (Lipton, 2018). LIME and SHAP are commonly used post-hoc methods (Ribeiro et al., 2016; Lundberg and Lee, 2017); given a trained model of arbitrary complexity they produce explanations for individual predictions by sampling perturbed inputs. Unlike LIME and SHAP, our method produces global explanations, and the student model can be used for predictions on future input. Prior work has shown that such methods can produce contrasts which are misleading or unintuitive (Mittelstadt et al., 2019) and that LIME or SHAP can be fooled into providing innocuous explanations for models that demonstrate racist or sexist behavior (Slack et al., 2020). These methods' feature importance scores are difficult to aggregate across a dataset and do not provide global faithfulness (van der Linden et al., 2019; Lakkaraju et al., 2017).

Lipton (2018) argues that interpretability is never "inherent" and must satisfy several criteria. These include simulatability, or whether a human can reasonably work through each step of the model's calculations; decomposibility, or whether each parameter of the model can be understood on its own; and algorithmic transparency, or whether the model belongs to a class with known theoretical behaviors. Lou et al. (2012) highlights linear and additive models as particularly decomposible (or intelligible) classes of models, because "users can understand the contribution of individual features in the model." Our proposed approach uses a linear bag-of-words model to provide a simulatible, decomposible, and transparent method.

Interpretability methods are also distinguished by the form and quality of the explanations they produce. We follow Jacovi and Goldberg (2020) in recognizing two primary desiderata for post-hoc explanations of ML systems: "faithfulness" and "plausibility."[2] A faithful explanation accurately represents the original model, by closely approximating its behavior or exposing its internal state (Yeh et al., 2019; Lakkaraju et al., 2020). A plausible model produces explanations that can be interpreted by a human expert (Jacovi and Goldberg, 2020; Ehsan et al., 2019). Prior work has explored methods such as forcing a faithful classifier to make predictions from a limited set of (plausible) rationales (Jain et al., 2020), or focusing on extracting rationales to constrain predictors to be inherently interpretable (Lei et al., 2016; Bastings et al., 2019). Methods should attempt to achieve both goals, but there is a trade-off between the two; explanations typically cannot be both concise and perfectly descriptive. Plausibility, unlike faithfulness, necessarily requires an evaluation based on human perception (Herman, 2017; Jain et al., 2020). A strength of our proposed method is that it is designed for plausibility and transparency, but optimized for faithfulness.

2.2 Knowledge Distillation

Knowledge distillation is a technique in which a simpler "student" model is trained to behave like a high performing, but more complex "teacher" model (Hinton et al., 2015). This approach has been widely studied under a variety of other names such as model approximation or compression (Buciluǎ et al., 2006), or simply 'copying' (Unceta et al., 2020). In many of these threads of research, the goal is to produce a student model that is

[1] https://github.com/isabelcachola/mimic-proxy

[2] Faithfulness is also referred to as fidelity, validity or completeness; plausibility is alternatively referred to as persuasiveness (Herman, 2017). See Jacovi and Goldberg (2020) for a longer discussion of alternate terminology.

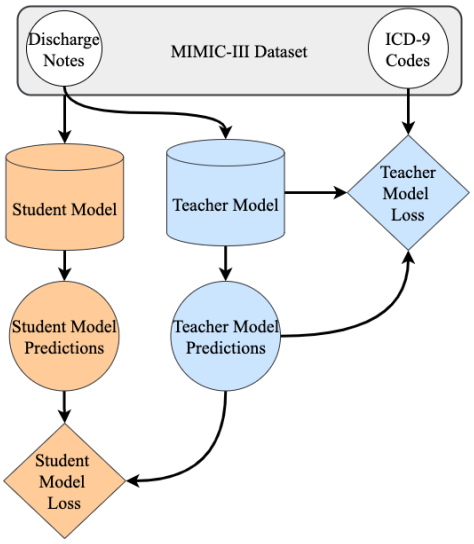

Figure 1: Relationship between the teacher and student models. The student model is trained to predict the teacher's outputs, rather than the true ICD-9 codes. This optimizes the student model for faithfulness.

smaller or faster than the teacher, while achieving high accuracy. Our work most closely follows approaches such as Lakkaraju et al. (2017); Asadulaev et al. (2019) that have sought to produce an *interpretable* student. The experimental and theoretical properties of knowledge distillation are well studied (Tan et al., 2018b; Phuong and Lampert, 2019).

Knowledge distillation has been applied to a variety of domains for the purposes of interpretability, such as crime and lending data (Tan et al., 2018a) and image classification (Asadulaev et al., 2019). Furthermore, a wide variety of model architectures have been shown to be effective as the student model, including decision trees (Elshawi et al., 2019), elastic nets (Guo et al., 2017), decision sets (Lakkaraju et al., 2017). We apply knowledge distillation to the task of medical code prediction, for the purposes of interpretability. We show that knowledge distillation is both an effective technique for training a faithful student model and can be used to generate plausible natural language explanations.

2.3 Explainable prediction in the medical domain

Explainability techniques have been applied to a variety of tasks in the medical domain, such as pneumonia and hospital readmission risk (Caruana et al., 2015) or real-time hypoxaemia prediction (Lundberg et al., 2018). Our work considers the task of predicting medical codes from hospital discharge notes. This task has been widely studied, and we use three published models on which we evaluate our approach: DR-CAML (Mullenbach et al., 2018), HAN (Yang et al., 2016), and TransICD (Biswas et al., 2021). As all three models contain millions of parameters, they are not simulatible or decomposible. However, DR-CAML and TransICD seek to produce their own explanations using a per-label attention mechanism that highlights regions in the input text that were correlated with the model's predictions. HAN was not designed with the goal of interpretability.

The use of attention to produce explanations has sparked discussion. Jain and Wallace (2019) showed that attention mechanisms can provide misleading explanations, whereas Wiegreffe and Pinter (2019) argued that attention-based explanations are often plausible, even when unfaithful. More recent work has explored when and how attention mechanisms can be either useful or deceptive (Zhong et al., 2019; Grimsley et al., 2020; Jain et al., 2020; Pruthi et al., 2020). As researchers continue to use this domain to explore methods for explainability and document classification (Kim and Ganapathi, 2021; Vu et al., 2020), we should strive to produce models that are both faithful and plausible.

3 Methods

Our proposed method is post-hoc and seeks to balance faithfulness and plausibility. We assume that we have a trained teacher model with good predictive performance but low interpretability. We train a student model that takes the same input from the dataset, but uses the teacher model's predictions as its labels. Figure 1 gives a visual representation of our model distillation setup.

The MIMIC-III dataset contains anonymized English-language ICU patient records, including physiological measurements and clinical notes (Johnson et al., 2016). Following Mullenbach et al. (2018), we focus on discharge summaries which describe a patient's visit and are annotated with ICD-9 codes. There are 8,922 different ICD-9 codes that describe procedures and diagnoses that occurred during a patient's stay. The manual assignment of these codes to patient records are required by most U.S. healthcare payers (Topaz et al., 2013).

To train the teacher models, we duplicate the experimental setup of Mullenbach et al. (2018) and Dong et al. (2021), which use the text of the discharge summaries as input to the DR-CAML and

HAN models, respectively, which then are trained to predict all ICD-9 codes associated with that document. After applying their pre-processing code to tokenize the text, the dataset contains 47,724 discharge summaries divided into training, dev, and test splits. We also duplicate the experimental setup of Biswas et al. (2021), which has a similar experimental setup but only predicts the top 50 most common ICD-9 codes.

We apply DR-CAML and HAN to the texts in MIMIC-III and save its continuous-valued probabilities as the labels for our student model. We similarly apply TransICD to MIMIC-III-50, which contains the top 50 most frequent labels in MIMIC-III and save the continuous-valued probabilities as the labels.[3] For all three models, we use the code released by the authors.[4] Training the student model on predictions from the existing teacher model optimizes for faithfulness.

We want the student model to produce plausible explanations and fulfill the criteria from Lipton (2018): simulatibility, decomposibility, and algorithmic transparency. To fulfill these desiderata, each student is a linear regression trained on bag-of-words representation of the clinical text. The fundamental trade-off here is that if we overly restrict our model class, the student will be unfaithful and unable to mimic the behavior of the teacher model. But if we allow for a student model that is too complex, it may not provide plausible or otherwise desirable explanations. These trade-offs may be domain-specific based, for example, on the target audience of the explanations. If the student model demonstrates sufficient empirical performance, a domain expert may even prefer to use it in place of the teacher model, an option unsupported by LIME or SHAP models.

We train a student model for each medical code independently, and we refer to the student model trained on X model's predictions as "Student-X" (e.g. Student-DRCAML). Each student uses only 50k parameters, allowing us to train each model on a single CPU in a matter of minutes. We implement our method using the linear `SGDRegressor` model from `sklearn` (Pedregosa et al., 2011), and apply a log transform to the model's probability outputs and train the student to minimize squared loss. After a brief[5] grid search on the validation set, we use L1 regularization with $\alpha = 0.0001$ for the DR-CAML student and $\alpha = 0.01$ for HAN and TransICD proxies.

To extract a rationale, we take the feature importance weights of the student model and average over a sliding window of n-grams from the discharge summary. We extract the n-gram with the highest average feature importance weight. Future work could use extracted rationales to train a student model that remains faithful to a black-box model.

In the next two sections, we introduce our evaluation for the student model's faithfulness to each model and the plausibility of its explanations.

4 Faithfulness evaluation

To establish that this collection of linear regressions is faithful to the trained models, we want to show that it makes similar predictions across all ICD-9 codes on held-out data. Recall from Figure 1 that the student is trained not to predict the true ICD-9 codes but to output the same label probabilities as the teacher model. In fact, the student model never sees the true ICD-9 codes. We evaluate faithfulness by comparing the outputs of the student and teacher models on the held-out test set. If the two systems produced identical outputs on held-out data, we would say that the student was perfectly faithful. We make this comparison in three different ways – first with regression metrics for the continuous outputs of the two models, then using classification metrics with binarized teacher predictions, and finally comparing student outputs as predictions for the true ICD-9 codes. For all these comparisons, we use a logistic regression baseline that is trained to directly predict the ICD-9 codes, independent of any black-box model. While we would expect the logistic baseline's predictions to roughly correlate with those of other models, we would not expect it to be faithful.

Similar to Tan et al. (2018a), our first evaluation uses regression metrics that assess the correlation between the student's predictions and the original teacher model's predicted probabilities. We use Spearman and Pearson correlation coefficients and

[3] Training on the full label set was prohibitively computationally expensive to reproduce and the authors did not release the trained model weights. In Table 1, TransICD and its student only use these 50 codes. These codes do not include those used in Tables 3 and 7, so TransICD models are omitted.

[4] Mullenbach et al. (2018) released their code under an MIT license, while Yang et al. (2016) and Biswas et al. (2021) did not specify a license.

[5] We considered L1, L2, and elastic net regularization with α from 0.1 to 10^{-7}. For HAN, which was not trained using the published dev set, we simply adopted α=0.01.

	Regression			Classification			
				AUC		F1	
Model	Spearman	Pearson	Kendall	Macro	Micro	Macro	Micro
Logistic to...							
DRCAML	0.036	-0.195	-0.135	0.734	0.936	0.012	0.353
HAN	0.204	0.036	-0.139	0.885	0.994	0.017	0.511
TransICD	0.587	0.662	0.419	0.894	0.927	0.476	0.580
Student-X							
-DRCAML	0.794	0.498	0.608	0.980	0.995	0.052	0.416
-HAN	0.736	0.519	0.543	0.975	0.997	0.014	0.454
-TransICD	0.838	0.539	0.650	0.960	0.960	0.507	0.592

Table 1: Comparison of the logistic baseline and the student model to the DR-CAML, HAN, and TransICD predictions. For the F1 evaluation, we threshold the student outputs at 0.5. The logistic model was trained to predict the ICD codes; the student model to predict DR-CAML's, HAN's, or TransICD's predictions, respectively. The student model dramatically outperforms the logistic baseline in terms of faithfulness to the DR-CAML and TransICD models. On classification metrics, the baseline is a surprisingly excellent student for the HAN model.

		DR CAML		HAN		Trans ICD	
	Logistic	Student	Orig	Student	Orig	Student	Orig
Macro AUC	0.561	0.901	0.906	0.870	0.884	0.883	0.897
Micro AUC	0.937	0.967	0.972	0.962	0.967	0.907	0.924
Macro F1	0.011	0.142	0.224	0.026	0.077	0.426	0.586
Micro F1	0.271	0.326	0.536	0.251	0.390	0.478	0.640
Prec @ 8	0.541	0.483	0.701	0.519	0.599	0.479	0.502
Prec @ 15	0.412	0.407	0.548	0.406	0.455	0.333	0.343

Table 2: Comparison of DR-CAML, HAN, and TransICD and their respective student models to the true ICD labels. Although the logistic regression baseline was trained to directly predict ICD codes and our student models were not, the Student-DRCAML and Student-TransICD models outperform the baseline in AUC and F1.

the non-parametric Kendall Tau rank correlation. These metrics range from -1 to 1 with 1 indicating perfect faithfulness. Regression results are on the left side of Table 1.

Our second evaluation treats the teacher model's predictions as binary labels to compute F1, AUC, and precision scores. We then evaluate the faithfulness of our student model by treating its outputs as probabilities and using classification metrics such as F1 score. Prec @ n is the fraction of the n highest scored labels that are present in the ground truth. These metrics range from 0 to 1, where perfectly faithful predictions would have 1.0 AUC and F1 scores. The student model is considered faithful if it correctly predicts whether the teacher model will make a binary prediction. We again use the logistic regression baseline. Classification results are on the right side of Table 1.

Finally, we use the student model's predictions to predict the ground-truth ICD code labels and compare its predictive performance against that of the teacher model's in Table 2. While the student model was not trained using these labels, we can use its predictions as probabilities for these codes. By comparing against the logistic regression baseline (a linear model of equal complexity), we can see whether our training setup allows the student model to learn a better predictor.

Our results show that our proxies are quite faithful to the teacher models. Table 1 shows that the Student-DRCAML and Student-HAN models are dramatically more faithful to their corresponding black-box models than the logistic regression baseline. Interestingly, the baseline is in fact quite faithful to the TransICD model. Comparing the classification metrics of Table 1 to the results in Table 2, we see that on AUC metrics, all three proxies are more faithful to the their target models than

934.1: "Foreign body in main bronchus"

Mullenbach et al. (2018)		
CAML	(HI)	*... line placed bronchoscopy performed showing* **large mucus plug on** *the left on transfer to ...*
Cosine		*... also needed medication to help* **your body maintain your** *blood pressure after receiving iv ...*
CNN		*... found to have a large* **lll lingular pneumonia on** *chest x ray he was ...*
Logistic		*... impression confluent consolidation involving nearly* **the entire left lung** *with either bronchocentric or vascular ...*
Ours		
DR-CAML	0.38	*... line placed bronchoscopy performed showing* **large mucus plug on** *the left on transfer to ...*
Logistic	0.28	*... tube down your throat to* **help you breathe you** *also needed medication to help ...*
Student-DRCAML	0.38	*... a line placed bronchoscopy performed* **showing large mucus plug** *on the left on transfer ...*
Student-HAN	0.39	*... line and r radial a* **line placed bronchoscopy performed** *showing large mucus plug on ...*

Table 3: Comparison of the clinical evaluation from Mullenbach et al. (2018) with our plausibility evaluation. The example above contains the explanations produced by eight systems. The first four systems for each example are directly copied from Table 1 of Mullenbach et al. (2018). The (HI) and (I) labels in the second column indicate whether the clinician labeled those explanations as Highly Informative or Informative. The four systems below the dotted line are from our evaluation, for which the second column indicates the probability output of our plausibility classifier. Here, Student-DRCAML and DR-CAML produce almost identical explanations. The Student-HAN explanation highlights that our student method can generate explanations for black-box models which cannot explain themselves. Additional comparisons are shown in Tables 5 and 7.

those black-box models were to the original ICD codes. In Table 2, we hypothesize that the relatively low precision scores result from our student regressions being fit for each ICD code independently, which prevents the combined model from encoding relative frequency information.

Rudin (2019) critiques post-hoc methods in general, arguing that "if we cannot know for certain whether our [post-hoc] explanation is [faithful], we cannot know whether to trust either the explanation or the original model." Because no post-hoc method can ever be perfectly faithful to an original model, our explicit measurement of faithfulness provides a useful approach for understanding whether the student is "faithful enough" for a given application. It also allows for a prediction-specific analysis – if we wish to use the student model to explain a high-stakes prediction made by a black-box model, we can first check whether both agree upon that specific prediction.

In applications where explainability is essential, our student model could be used as a more interpretable replacement for a high-performing black-box model. In such a case, a domain expert might care less about the evaluation of faithfulness in Table 1 and more about the ground-truth predictive performance evaluated in Table 2. We leave for future work the challenge of whether a student model produced by our method could be fine-tuned to better predict ground-truth ICD codes.

5 Plausibility Evaluation

Explanations are considered plausible if they can be reasoned about by a person (Wiegreffe and Pinter, 2019). Evaluating plausibility is thus typically more difficult than faithfulness, because it requires input from annotators (Herman, 2017; Arora et al., 2021). Furthermore, an explanation that is plausible to a domain expert may not be plausible to a layperson. Mullenbach et al. (2018) evaluated the plausibility of their models' explanations by collecting annotations from a clinical expert. For 100 notes, each of four models produced an explanation in the form of a 14-token subsequence taken from the discharge summary. The clinician read the four (anonymized) explanations and the corresponding ICD code and subjectively rated each explanation as "informative"[6]. Across the 100 examples, the clinician rated CAML as slightly more informative than the logistic regression and CNN baselines. Table 3 shows explanations produced by our and Mullenbach et al. (2018)'s models.

The format of Mullenbach et al. (2018)'s plausibility evaluation does not easily lend itself to replication. While the authors shared their annotations with us, missing metadata (see Appendix A.2) prevented a direct reproduction of their analysis. Additionally, since the clinical annotator considered explanations in a comparative setting, we cannot easily add our student model as another method us-

[6]The annotator was told to mark as informative all explanations that "adequately explain[ed] the presence of the given ICD code" (Mullenbach et al., 2018).

Model	Score	Interval	Best
Logistic	35	(31, 49)	7%
Cosine	38	(32, 51)	13%
CNN	42	(33, 52)	14%
CAML	44	(33, 52)	16%
DR-CAML	48	(34, 53)	22%
Student-DRCAML	52	(34, 54)	19%
Student-HAN	47	(33, 52)	10%

Table 4: Binary plausibility evaluation using classifier annotations. We collapse the Highly Informative and Informative labels from Mullenbach et al. (2018) to a single positive class. The Score column is out of 99; we use a binary threshold of 0.45 so that the proportion of predicted plausible explanations matches the data. To highlight the uncertainty of this evaluation, we bootstrap sample 1000 informative labels for each method's explanations. The Interval column shows the 95% interval of informative scores across those 1000 samples. The Best column shows the percentage of samples in which each method scored highest.

296.20: "Major depressive affective disorder, single episode, unspecified"

DR-CAML	*... diagnosis overdose of medications narcotics* **benzodiazepine suicide attempt chronic** *migraine headaches depression stage iv ...*
Student-DRCAML	*... up from the medications you* **were evaluated by psychiatry** *and will be transferred to ...*

Table 5: Examples of differing explanations between DR-CAML and its student. Our informative classifier gives the DR-CAML and student explanations scores of 0.47 and 0.33, respectively. Additional examples are shown in Table 8.

ing the same annotations. Therefore, we replicate this evaluation as best as possible by using a classifier to predict synthetic labels as to whether the clinical domain expert *would have* labeled our models' explanations as plausible. Using BioWordVec embeddings released by Zhang et al. (2019), the text of the ICD-9 code description, and the 14-gram explanation produced by each model from Mullenbach et al. (2018), we train a classifier that predicts whether an explanation would have been rated as informative. This annotation classifier achieves a binary classification accuracy of 67.2% and an AUC score of 0.726 when evaluated with leave-one-out validation. This relatively low accuracy and our model training details are discussed in Appendix A.3.

To conduct our plausibility evaluation, we first use or reproduce the baseline methods from Mullenbach et al. (2018) and Biswas et al. (2021). Each model, including the student, produces a 14-token explanation from the discharge summary by first finding the 4-gram with the largest *average feature importance* and then including five tokens on either side of the 4-gram. The logistic regression baseline is the same as in § 4, where feature importance is computed using the coefficients of the logistic model. The student model's explanations are computed in the same manner, finding the 4-gram with the largest average coefficient weights. For CAML, DR-CAML, and the CNN models, we use the code released by Mullenbach et al. (2018) to extract explanations. The CNN baseline primarily differs from CAML in that it does not use an attention mechanism. Finally, we reimplement their Cosine baseline which picks the 4-gram with the highest cosine similarity to the ICD-9 code description text.

We extract the model's explanations for the same[7] discharge summaries as were evaluated by Mullenbach et al. (2018). For each explanation, we use the annotation classifier described above to predict the probability that each explanation would have been labeled as informative. If we set the classifier threshold such that 45% of explanations are rated as informative (matching the proportion from the original annotations), we get the results in the Score column of Table 4. The student model produces the largest number of informative explanations according to our classifier; however, the classifier's inaccuracy can introduce substantial uncertainty. Rather than thresholding the outputs of the annotation classifier, we can use its probability outputs to sample a set of informative labels for each explanation. If we sample 1000 such sets of labels and report the 95% confidence interval for each model's score in the Interval column of Table 4, the interval overlap makes the methods essentially indistinguishable. The Best column in this table shows the percentage of samples in which each method scored highest. While the Interval column highlights the inherent limitation of evaluating plausibility on this small fixed dataset of human evaluations, the Best and Score columns combined with the qualitative comparisons in Table 3 suggest that our student model explanations are at least comparably plausible to those of DR-CAML.

Table 3 shows that DR-CAML and Student-DRCAML produce qualitatively similar explana-

[7] Using the 99 (of 100) discharge summaries that could be uniquely identified. See Appendix A for details.

tions. The other two examples presented in Mullenbach et al. (2018) are in Appendix A.4. The similarity is perhaps surprising because DR-CAML extracts explanations using its attention mechanism, whereas the student model uses unigram feature importance values that do not vary between examples. For this example, it appears that the student is faithful both in the predictions it makes and how it makes those predictions. We additionally include the explanations for Student-HAN. As HAN cannot produce its own explanations, this highlights that our method can also be applied to models that are not interpretable by design. Table 5 shows an example where the student and DR-CAML diverge the most. We include two additional examples in Appendix A.4. These cases highlight two benefits of the student model. First, its feature importance weights are *global* across all predictions, providing an aggregate representation of the student's behavior. Second, the approach for extracting student explanation n-grams is transparent and simulatible; it is just the average of n feature weights. These factors may be particularly appealing in cases where explainability is paramount.

6 Discussion

We have introduced a method that uses knowledge distillation to generate post-hoc explanations and is designed to be interpretable and plausible while maintaining faithfulness to the trained model. By constraining the student to a class of models that is decomposible, simulatible, and algorithmically transparent, our optimization for faithfulness gives us a clear way to evaluate several dimensions of interpretability. We evaluated our method on the task of clinical code prediction. A key benefit of our method is its simplicity and wide applicability. Even for a proprietary trained model for which the learned parameters are unknown, a student can be trained as long as we have a dataset that includes the trained model's predictions. Our approach has the additional benefit of producing a standalone student model that can provide *global* feature explanations. If the student has sufficient predictive performance, a skeptic of post-hoc methods (e.g. Rudin (2019)) might prefer to use the inherently-interpretable student.

The present work has several limitations that are left for future work. Though the task of medical code prediction has important implications and has been widely studied in interpretability research, we only consider this single task on a single English-language dataset. While we have shown our student approach works for three different black-box models, it requires additional study in new domains and tasks. There may be black-box models for which no linear student is faithful. Our evaluation is also limited to only a single form of explanation: n-grams extracted via importance or attention weights. Counterfactual explanations (i.e., an alternative input that would have been classified differently) might be harder or easier for our student method to generate (Barocas et al., 2020). Our plausibility evaluations rely on a small set of annotations from which we extrapolate. Future work should collect new annotations that consider metrics such as sufficiency and simulatibility that require human evaluations (Jain et al., 2020; Hase and Bansal, 2020; Arora et al., 2021).

As the ML community continues to explore new directions for interpretable methods, new desiderata may arise based on the domain experts who turn to ML methods for decision support. Interpretable ML methods should clearly define how they expect to satisfy criteria such as faithfulness or plausibility; by designing for plausibility and transparency and optimizing for faithfulness, our proposed method is broadly applicable. We release our code to enable future work.

7 Ethics and Broader Impacts

This paper is situated in a broader field of clinical applications of machine learning. While our work does not raise new ethical issues within this domain, there are general concerns that also apply to this work. ML methods should not be deployed in real-world settings without extensive validation (Wiens et al., 2019). In the clinical domain, particular attention must be paid to the possibility of perpetuating disparities that have been encoded in the training data (Rajkomar et al., 2018). While MIMIC-III provides a useful benchmark for developing and evaluating methods, it is not representative of the the enormous variety of clinical and linguistic data. Domain experts and those most likely to be affected by new ML systems should be given oversight of potential deployments.

Acknowledgements

We acknowledge support provided by the Johns Hopkins Institute for Assured Autonomy. We thank Sarah Wiegreffe and Jacob Eisenstein for their help and plausibility annotations.

References

Siddhant Arora, Danish Pruthi, Norman Sadeh, William W Cohen, Zachary C Lipton, and Graham Neubig. 2021. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. *arXiv preprint arXiv:2112.09669*.

Arip Asadulaev, Igor Kuznetsov, and Andrey Filchenkov. 2019. Interpretable few-shot learning via linear distillation. *arXiv preprint arXiv:1906.05431*.

Solon Barocas, Andrew D Selbst, and Manish Raghavan. 2020. The hidden assumptions behind counterfactual explanations and principal reasons. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 80–89.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977.

Biplob Biswas, Thai-Hoang Pham, and Ping Zhang. 2021. Transicd: Transformer based code-wise attention model for explainable icd coding. In *AIME*.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. 2006. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541.

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*, pages 1721–1730.

Hang Dong, V'ictor Su'arez-Paniagua, William Whiteley, and Honghan Wu. 2021. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *Journal of biomedical informatics*, page 103728.

Been Doshi-Velez, Finale; Kim. 2017. Towards a rigorous science of interpretable machine learning. In *eprint arXiv:1702.08608*.

Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O Riedl. 2019. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 263–274.

Radwa Elshawi, Mouaz Al-mallah, and Sherif Sakr. 2019. On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making*, 19.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE.

Christopher Grimsley, Elijah Mayfield, and Julia RS Bursten. 2020. Why attention is not explanation: Surgical intervention and causal reasoning about neural models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1780–1790.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Wenbo Guo, Kaixuan Zhang, Lin Lin, Sui Huang, and Xinyu Xing. 2017. Towards interrogating discriminative machine learning models. *ArXiv*, abs/1705.08564.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552.

Bernease Herman. 2017. he promise and peril of human evaluation for model interpretability. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS 2017)*.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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 3543–3556.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

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.

Byung-Hak Kim and Varun Ganapathi. 2021. Read, attend, and code: Pushing the limits of medical codes prediction from clinical notes by machines. *ArXiv*, abs/2107.10650.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural language inference with faithful natural language explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Himabindu Lakkaraju, Nino Arsov, and Osbert Bastani. 2020. Robust and stable black box explanations. In *International Conference on Machine Learning*, pages 5628–5638. PMLR.

Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2017. Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Yin Lou, Rich Caruana, and Johannes Gehrke. 2012. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 150–158.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st international conference on neural information processing systems*, pages 4768–4777.

Scott M. Lundberg, Bala G. Nair, Monica S Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor L. Adams, David Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry H. Kim, and Su-In Lee. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2:749 – 760.

Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in ai. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 279–288.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Mary Phuong and Christoph H. Lampert. 2019. Towards understanding knowledge distillation. *ArXiv*, abs/2105.13093.

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793, Online. Association for Computational Linguistics.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186.

S. Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018a. Distill-and-compare: Auditing black-box models using transparent model distillation. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.

S. Tan, Giles Hooker, Paul Koch, Albert Gordo, and Rich Caruana. 2018b. Considerations when learning additive explanations for black-box models.

Maxim Topaz, Leah Shafran-Topaz, and Kathryn H Bowles. 2013. Icd-9 to icd-10: evolution, revolution, and current debates in the united states. *Perspectives in health information management/AHIMA, American Health Information Management Association*, 10(Spring).

Irene Unceta, Jordi Nin, and Oriol Pujol. 2020. Copying machine learning classifiers. *IEEE Access*, 8:160268–160284.

Ilse van der Linden, Hinda Haned, and Evangelos Kanoulas. 2019. Global aggregations of local explanations for black box models. In *SIGIR Workshop on FACTS-IR*.

Thanh Vu, Dat Quoc Nguyen, and Anthony N. Nguyen. 2020. A label attention model for icd coding from clinical text. In *IJCAI*.

Sarah Wiegreffe, Edward Choi, Sherry Yan, Jimeng Sun, and Jacob Eisenstein. 2019. Clinical concept extraction for document-level coding. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 261–272.

Sarah Wiegreffe, Ana Marasović, and Noah A Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10266–10284.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard H. Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL*.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. 2019. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32:10967–10978.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. Biowordvec, improving biomedical word embeddings with subword information and mesh. *Scientific data*, 6(1):1–9.

Ruiqi Zhong, Steven Shao, and Kathleen McKeown. 2019. Fine-grained sentiment analysis with faithful attention. *arXiv preprint arXiv:1908.06870*.

A (Re-)implementation details

A.1 Reproducing CAML predictive performance

The trained DR-CAML model released by Mullenbach et al. (2018) produced predictions that matched the published F1 and ROC scores. We were unable to precisely replicate the outputs of the CAML model. Table 6 shows the scores published by Mullenbach et al. (2018) as well as those for a CAML reimplementation done by Wiegreffe et al. (2019). We include the scores we observe using the model weights released on GitHub as well as the scores for a model we retrained from scratch. We use the released model instead of the retrained model as its performance is much closer to the published numbers.

A.2 Reproducing plausibility scores

The clinical plausibility annotations provided to us by the authors of Mullenbach et al. (2018) contains the text explanations and their corresponding annotations, but was missing the crucial metadata of which models produced which explanations. The metadata also did not indicate from which specific discharge summary the texts were derived; while the text explanations were uniquely identifying for all but one of the 100 examples. For that one example, because some patients had multiple documents sometimes containing duplicated segments of text, there were three discharge summaries from which the explanations could have been drawn. We thus excluded this example from our analyses. To replicate their analysis the best we could, we retrained or reimplemented their logistic regression, vanilla CNN, and cosine similarity methods. We then looked at the attention or feature importance weights for each trained model and the text explanations that had been annotated, and assigned each model the text explanation for which it provided the highest weight. This assignment did not perfectly align with past work: there were six cases (out of 99) where a text explanation was "chosen" by more models than times it appeared as an option. Ignoring that issue and then simply aggregating the Informative and Highly Informative clinician annotations, we obtained the plausibility scores in the Ours column of Table 9. The Theirs column shows the published numbers from Mullenbach et al. (2018). While the numbers change substantially, the ordering is relatively stable with only two swaps: CAML and Cosine, and Logistic and CNN. The other columns of the table are described below.

A.3 Plausibility annotation classifier

To evaluate the plausibility of our student model's explanations, we trained a classifier to predict whether an explanation would have been labeled as plausible by the clinical domain expert. We treat this as a binary classification task by grouping the "Informative" and "Highly Informative" annotations as a single "plausible" label. Conscious of the fact that we have only 99 examples with four text explanations each, we use two approaches with which to train and evaluate our classifier. The first used leave-one-out cross validation at the example level, such that the classifier was trained on 98 examples at a time and then evaluated on the remaining one. We refer to this evaluation as "E1" in Table 9. The second also used leave-on-out cross validation but at the explanation level; we held out a single text explanation, trained on all other explanations across all examples, and then evaluated on the held-out explanation. When an explanation appeared more than once in a single example, we made sure to remove its duplicates from the training data for predicting that explanation. We refer to this evaluation as "E2" in Table 9.

The trained model is a simple logistic regression classifier trained on a fastText embedding of both the explanation and the target ICD-9 code description. Using the BioWordVec embeddings released by Zhang et al. (2019), we embed each both the explanation and code description into a 200-dimensional vector, concatenate the two vectors, and pass it to the logistic regression. In the E1 evaluation, the model achieves an accuracy of 60.6% and an ROC AUC score of .640. In the E2 evaluation, that increases to an accuracy of 67.2% and an AUC score of .726, indicating that the additional within-example explanations substantially help the classifier.

When using these classifiers to label the explanations generated by each model instead of the plausibility scores derived in A.2, we get the results shown in columns E1 and E2 of Table 9.

Finally, we retrain our final classifier on all the explanations, leaving none held out. Rather than using our classifier to evaluate the explanations that were actually shown to the clinician, we instead use our (re-)implementation of the four models to extract an explanation from each of the 99 discharge summaries. These explanations thus may or may

	AUC		F1		P@n	
	Macro	Micro	Macro	Micro	8	15
Mullenbach et al. (2018)	0.895	0.986	0.088	0.539	0.709	0.561
Wiegreffe et al. (2019)	0.889	0.985	0.080	0.542	0.712	0.562
Ours (using released weights)	0.892	0.978	0.090	0.298	0.636	0.471
Ours (retrained)	0.628	0.884	0.001	0.024	0.042	0.027

Table 6: Published predictive performance of CAML and our replicated results. Our experiments throughout the paper use the model with the released weights, which is closest to the published numbers (despite Micro F1).

442.84: "Aneurysm of other visceral artery"

Mullenbach et al. (2018)		
CAML	(I)	... *and gelfoam embolization of right* **hepatic artery branch pseudoaneurysm** *coil embolization of the gastroduodenal* ...
Cosine		... *coil embolization of the gastroduodenal* **artery history of present** *illness the pt is a* ...
CNN		... *foley for hemodynamic monitoring and* **serial hematocrits angio was** *performed and his gda was* ...
Logistic	(I)	... *and gelfoam embolization of right* **hepatic artery branch pseudoaneurysm** *coil embolization of the gastroduodenal* ...
Ours		
DR-CAML	0.55	... *gelfoam embolization of right hepatic* **artery branch pseudoaneurysm coil** *embolization of the gastroduodenal artery* ...
Logistic	0.57	... *biliary stents hx cbd r* **colonic fistula r colectomy** *partial l nephrectomy for renal* ...
Student-DRCAML	0.55	... *embolization of right hepatic artery* **branch pseudoaneurysm coil embolization** *of the gastroduodenal artery history* ...
Student-HAN	0.55	... *embolization of right hepatic artery* **branch pseudoaneurysm coil embolization** *of the gastroduodenal artery history* ...

428.20: "Systolic heart failure, unspecified"

Mullenbach et al. (2018)		
CAML		... *no mitral valve prolapse moderate* **to severe mitral regurgitation** *is seen the tricuspid valve* ...
Cosine		... *is seen the estimated pulmonary* **artery systolic pressure is** *normal there is no pericardial* ...
CNN		... *and suggested starting hydralazine imdur* **continue aspirin arg admitted** *at baseline cr appears patient* ...
Logistic	(HI)	... *anticoagulation monitored on tele pump* **systolic dysfunction with ef** *of seen on recent echo* ...
Ours		
DR-CAML	0.39	... *anticoagulation monitored on tele pump* **systolic dysfunction with ef** *of seen on recent echo* ...
Logistic	0.37	... *seen the mitral valve leaflets* **are mildly thickened there** *is nomitral valve prolapse* ...
Student-DRCAML	0.39	... *anticoagulation monitored on tele pump* **systolic dysfunction with ef** *of seen on recent echo* ...
Student-HAN	0.36	... *blood cultures obtained repeated cxr* **echocardiogram showed an ef** *of and therefore zestril was* ...

Table 7: Comparison of the clinical evaluation from Mullenbach et al. (2018) with our plausibility evaluation. There are two examples above, each which contains the explanations produced by eight systems. The first four systems for each example are directly copied from Table 1 of Mullenbach et al. (2018). The (HI) and (I) labels in the second column indicate whether the clinician labeled those explanations as Highly Informative or Informative. The four systems below the dotted line are from our evaluation, for which the second column indicates the probability output of our plausibility classifier.

not appear in the training data for the classifier. For the Full evaluation we are not worried about the classifier overfitting, as the classifier functions as a direct replacement for the clinician who produced the training data. The results of this analysis are the numbers shown in Table 4 in § 5, reproduced in Table 9 in the "Full" column. The Logistic model does much worse on the Full evaluation than in either E1 or E2. This may be because the explanations selected by the trained model were worse than those selected by the model which was used for the original clinical evaluation.

455.0: "Internal hemorrhoids without mention of complication"

DR-CAML	0.38	*... and she then underwent a* **colonoscopy with gi that** *also did not detect evidence ...*
Student-DRCAML	0.52	*... past medical history diverticular disease* **diverticulitis sbo anxiety hemorrhoids** *past surgical history s p ...*

592.0 : "Calculus of kidney"

DR-CAML	0.30	*... if you develop any of* **these symtpoms please call** *the office or go to ...*
Student-DRCAML	0.46	*... the colon gastroesophageal reflux asthma* **irritable bowel syndrome gastroparesis** *osteoporosis anxiety and or depression ...*

Table 8: Additional differing explanations and classifier scores between DR-CAML and the student.

Model	Theirs	Ours	E1	E2	Full
Logistic	41	43	47	49	35
Cosine	48	48	41	40	38
CNN	36	46	51	47	42
CAML	46	54	47	43	44
DR-CAML	–	–	45	44	48

Table 9: Plausibility evaluations and comparison to Mullenbach et al. (2018). The Theirs column shows the published numbers; Ours shows our best attempt at matching the clinical evaluation to the trained models. While the numbers change dramatically, the ordering only changes by two swaps. The clinical evaluation did not include DR-CAML. E1 and E2 show the results with predicted plausibility labels under the two evaluation settings described in A.3. Full duplicates the results from Table 4 for comparison.

A.4 Additional Examples

We provide two additional examples of eight different models' explanations in Table 7. These are the same examples shown in (Mullenbach et al., 2018). We include the four explanations as published in Mullenbach et al. (2018), our reproduction of DR-CAML, the logistic regression baseline, and the explanations from two student models, Student-DRCAML and Student-HAN. As we can see from the examples, Student-DRCAML produces similar explanations to DR-CAML. Student-HAN shows that our method is able to produce explanations for models not originally designed to do so. We also include two additonal examples in which DR-CAML and Student-DRCAML diverge the most in Table 8.