Faithful and Plausible Explanations of Medical Code Predictions

Anonymous EMNLP submission

Abstract

Machine learning models that offer excellent predictive performance often lack the interpretability necessary to support integrated human machine decision-making. In clinical medicine and 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 train a proxy model that mimics the behavior of a trained model and provides control over these trade-offs. We evaluate our approach on the task of assigning ICD codes to clinical notes to demonstrate that the proxy model is faithful to the trained model's behavior and produces quality 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. 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.

Similar concerns over model explanations across domains have inspired a whole field of interpretable ML. Work in this area considers two goals: faithfulness (explanations that accurately convey the decision-making process of the model) and plausibility (explanations that make sense to domain experts). Balancing these goals can be challenging; faithful explanations that accurately convey the reasoning of complex AI systems may be implausible to a domain expert, and vice versa. Models must also balance sophistication against transparency. The sophisticated methods may yield the best performance on a task, but be least able to provide explanations.

We propose to disentangle these competing goals by introducing a *proxy model*. We assume a trained model exists that makes accurate predictions on a dataset but that may not be interpretable. We train a fundamentally-interpretable linear model on the *predictions* of the trained ML model, so that the behavior of the proxy model mimics the trained model's behavior, rather than independently modeling the target task. We then rely on the interpretable proxy model to create explanations, allowing the trained model to use sophisticated methods to achieve high accuracy. We pose two questions to validate our approach: 1) Is the proxy faithful to the workings of the trained model? and 2) Are the produced explanations of high quality to domain experts?

We demonstrate our approach on the task of medical code prediction. While ML methods have achieved predictive success on various versions of ICD clinical code assignment, the best-performing methods have been neural networks that are notoriously difficult to interpret. Mullenbach et al. (2018) introduced DR-CAML, a method designed to produce explainable predictions, which outperformed several baselines when evaluated by a clinical expert.

We reproduce this work and compare to our proxy model.We use a linear logistic regression proxy model that learns to mimic the behavior of the trained DR-CAML model. We show that the proxy model is faithful to the original model

and produces plausible explanations, as measured on clinician annotations of generated explanations. We release the code both for our method and for reproducing Mullenbach et al. (2018).

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. Methods such as logistic regression are often considered transparent or inherently interpretable, because their simplicity allows a domain expert to understand how a change in input would produce a different output (Guidotti et al., 2018). However, even simple models can prove difficult to interpret in certain settings, such as when the model's features are complex (Lipton, 2018). LIME is an example of a post-hoc method (Ribeiro et al., 2016); given a trained model of arbitrary complexity it produces explanations for individual predictions. The trade-off in the different methods is that inherently-interpretable methods are often limited in model complexity. Deep neural networks, for example, often demonstrate better performance but are not inherently interpretable (Feng et al., 2018), and typically rely upon post-hoc methods to derive explanations (Guidotti et al., 2018).

Lipton (2018) critiques the idea of "inherent" interpretability and argues that methods that are intended to be transparently understood should pursue several traits. These include simulatability, or whether a human can reasonably work through each step of the model's calculations to understand how a prediction is made; decomposibility, or whether each parameter of the model can be intuitively 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 will use a linear model trained on bag-of-word features to provide a simulatible, decomposible, and transparent method.

Interpretability methods are also distinguished by the form and quality of the explanations they produce. Two primary desiderata for explanations of ML systems are "faithfulness" and "plausibility."[1] A faithful method accurately describes the true machinery of the model's prediction, while a plausible model produces explanations that can be interpreted by a human expert (Jacovi and Goldberg, 2020). A method could be faithful but not plausible, if it accurately explains a model's predictions but does so in terms of high-dimensional feature vectors that a human cannot interpret. Similarly, a method could be plausible but not faithful, if it produces concise natural language summaries that are unrelated to the calculations that produce the model's predictions. 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). A strength of our proposed method is that it is designed for plausibility and transparency, but optimized for faithfulness.

2.2 Explainable prediction of medical codes

Our work closely follows that of Mullenbach et al. (2018). We use the same dataset of clinical texts and associated medical codes (described in § 4) and compare against their method: Description-Regularized Convolutional Attention for Multi-Label classification (DR-CAML). DR-CAML is a neural model that seeks to produce its own faithful explanations using a per-label attention mechanism that highlights n-grams in the input text that were correlated with the model's predictions. Because DR-CAML has over six million learned parameters, it does not fulfill simulatability or decomposability; a single parameter cannot be understood in any intuitive way. However, the attention mechanism allows for some insight into the model's decision-making, as it indicates which regions of the input text were given more weight in the prediction.

DR-CAML's use of attention to produce explanations has sparked discussion. Jain and Wallace (2019) showed that attention mechanisms can pro-

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

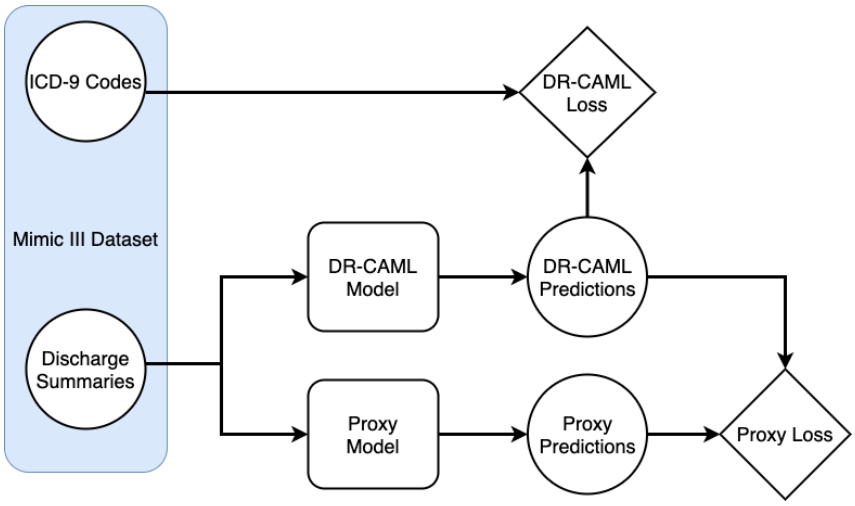

Figure 1: Relationship between trained DR-CAML model and proxy model. The proxy model is trained to predict DR-CAML's outputs, rather than the true ICD-9 codes. This optimizes the proxy model for faithfulness.

vide misleading explanations that are not faithful to the model's true reasoning. Wiegreffe and Pinter (2019) argued that while the explanations produced by attention may not always be faithful, they are often plausible. This discussion has continued in the interpretable ML literature, with methods demonstrating how attention mechanisms can be useful or deceptive (Zhong et al., 2019; Grimsley et al., 2020; Jain et al., 2020; Pruthi et al., 2020). Creating models that are both faithful and plausible remains a challenge.

3 Methods

Our proposed method is post-hoc and seeks to balance faithfulness and plausibility. We assume that we have a trained model with good predictive performance but low interpretability. Given this trained model and a dataset on which it can be applied, we train a *proxy model* that takes the same input from the dataset, but uses the trained model's predictions as its labels. In other words, given the dataset's input, the proxy model predicts the outputs of the uninterpretable model. Figure 1 gives a visual representation of the proxy model setup. For the medical code classification task, the original model (DR-CAML) is trained on the text of discharge summaries and produces a probability for each of the 8,922 possible medical codes. We apply DR-CAML to the texts in MIMIC III (Johnson et al., 2016) and save its continuous-valued probabilities as the labels for our proxy model. Training the proxy model on predictions from the existing model optimizes for faithfulness by design.

We also want the proxy model to produce plausible explanations and fulfill the criteria from Lipton (2018): simulatibility, decomposibility, and algorithmic transparency. To do so, we restrict our proxy model to a class of models that fulfills these desiderata. The fundamental trade-off here is that if we restrict our model class too much, the proxy will be unfaithful and unable to mimic the behavior of the trained model. But if we allow for a proxy model that is too complex, it may not provide plausible or otherwise desirable explanations. The choice of proxy model requires some consideration of the particular domain, as feature preprocessing and similar details may affect its behavior and explanations.

For the task of medical code prediction, we use a linear regression model trained on a bag-of-words representation of the clinical texts. We train 8,922 proxy models, one for each medical code in the dataset's labels. 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 proxy to minimize squared loss. We include release the code for training and evaluating our method as an Appendix.

Our approach is similar to LIME (Ribeiro et al., 2016) in that it learns a simple (linear) model to explain a pretrained model. However, whereas LIME learns a linear model to post-hoc explain a single prediction, our linear model is trained to predict and explain the entire dataset of predictions. This has several consequences. Unlike LIME, we do not require sampling perturbed inputs that do not exist in the training data, which can produce con-

trasts which are misleading or unintuitive (Mittelstadt et al., 2019). Slack et al. (2020) showed that LIME can be fooled into providing innocuous explanations for models that demonstrate racist or sexist behavior by exploiting its reliance on perturbations. It also means that our proxy model is given a more difficult task than a LIME model – it may be that a given proxy model is insufficiently flexible to model the complexity of the pretrained model, in which case we can measure this failure in terms of our faithfulness evaluation (see § 4). Because LIME trains a model linear only in the neighborhood of a given instance, its feature importance scores are difficult to aggregate across a dataset, making extrapolation difficult (van der Linden et al., 2019). When our proxy model is faithful to the trained model, our approach gives us explanations that we can expect to apply to future predictions. If the proxy model demonstrates sufficient empirical performance, a domain expert may even prefer to use it in place of the original trained model, an option unsupported by LIME models.

By applying our proxy model method to the DR-CAML model from Mullenbach et al. (2018), we enable an evaluation of both faithfulness and plausibility. We evaluate whether our model is faithful by seeing how closely its outputs match the predictions of DR-CAML. Because DR-CAML was designed to be interpretable using its attention mechanism, we can compare its explanations against those produced by our proxy. In the next two sections, we introduce our evaluation for the proxy model's faithfulness to the DR-CAML model and the plausibility of its explanations.

4 Faithfulness evaluation

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). We duplicate the experimental setup of Mullenbach et al. (2018) which uses the text of the discharge summaries as input to the DR-CAML model, which then is 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, validation, and test splits.

Our proxy model is the combination of 8,922 linear regression models trained to predict DR-CAML's log probability for each ICD-9 code. After a brief grid search on the validation set, we chose to apply L1 regularization with $\alpha = 0.0001$ for each regression. To establish that this collection of linear regressions is faithful to the trained DR-CAML model, we want to show that it makes similar predictions across all ICD-9 codes on held-out data. Recall from Figure 1 that the proxy is trained not to predict the true ICD-9 codes but to output the same label probabilities as DR-CAML. In fact, the proxy model never sees the true ICD-9 codes. We evaluate faithfulness by comparing the outputs of DR-CAML and the proxy model on the held-out test set. If the two systems produced identical outputs on held-out data, we would say that the proxy was perfectly faithful. We make this comparison in three different ways – first using regression metrics that compare the continuous outputs of the two models, then using classification metrics with binarized DR-CAML predictions, and finally by using the proxy model's 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 without knowledge of DR-CAML's predictions. While we would expect the logistic baseline's predictions to be somewhat correlated with those of DR-CAML, we would not expect the baseline to be faithful.

Our first evaluation uses regression metrics that assess the correlation between the proxy's predictions and DR-CAML's predicted probabilities. We use Spearman and Pearson correlation coefficients and 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 DR-CAML's predictions as binary labels based on whether they exceed the threshold used by Mullenbach et al. (2018) to compute F1 scores. We then evaluate the faithfulness of our proxy model by treating its outputs as unnormalized probabilities and using classification metrics such as F1 score. These metrics range from 0 to 1, where perfectly faithful predictions would have 1.0 AUC and F1 scores. The proxy

	Regression			Classification			
				AUC		F1	
Model	Spearman	Pearson	Kendall	Macro	Micro	Macro	Micro
Logistic	0.036	-0.195	-0.135	0.734	0.936	0.012	0.353
Proxy	0.498	0.794	0.608	0.980	0.995	0.052	0.416

Table 1: Comparison of the logistic baseline and the proxy model to the DR-CAML predictions. For the F1 evaluation, we threshold the unnormalized proxy outputs at 0.5. The logistic model was trained to predict the ICD codes; the proxy model to predict DR-CAML's predictions. As expected, the proxy model dramatically outperforms the logistic baseline in terms of faithfulness to the DR-CAML model.

	Logistic	Proxy	DR-CAML
Macro AUC	0.596	0.901	0.906
Micro AUC	0.889	0.967	0.972
Macro F1	0.033	0.142	0.224
Micro F1	0.278	0.326	0.536
Prec @ 8	0.547	0.483	0.701
Prec @ 15	0.413	0.407	0.548

Table 2: Comparison of the logistic baseline, the proxy model, and DR-CAML to true ICD labels. Although the logistic model was trained for this specific task and the proxy model was not, the proxy model outperforms the baseline in terms of AUC and F1. The proxy model's outputs are unnormalized, which partially explains the gap between its F1 scores, which are computed with a threshold of 0.5, and its AUC scores, which are invariant to normalization. This lack of normalization may also explain the proxy model's low precision scores, as each code is predicted independently of the others.

model is considered faithful if it correctly predicts whether DR-CAML 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 proxy model's predictions to predict the ground-truth ICD code labels and compare its predictive performance against that of DR-CAML in Table 2. While the proxy model was not trained using these labels, we can use its predictions as unnormalized 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 proxy model to learn a better predictor.

Our results show that the proxy model is quite faithful to the DR-CAML model. Compared to the logistic regression baseline, the proxy model is dramatically better on all metrics in Table 1. Combining the results from Tables 1 and 2 we can see that on AUC metrics, the proxy model is closer to the DR-CAML predictions than DR-CAML is to the ground-truth labels. The proxy model also outperforms the logistic regression baseline in the classification metrics (AUC and F1), indicating that the proxy model is more faithful to the DR-CAML predictions. In Table 2, we see a large gap between its performance on the AUC metrics and the F1 and precision metrics. This is likely because the outputs of the proxy model are not normalized to be valid probabilities and AUC is invariant to normalization, unlike F1 and precision.

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, we believe our approach to explicitly measuring faithfulness provides a useful approach for understanding whether the proxy is "faithful enough" for a given application. It also allows for a prediction-specific analysis – if we wish to use the proxy model to explain a high-stakes prediction made by DR-CAML, we can first check to see whether the two models agree upon that specific prediction.

In applications where explainability is essential, our proxy 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 proxy model produced by our method could be fine-tuned to improve its performance at predicting ground-truth ICD codes.

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 ...*
Proxy	0.38	*... a line placed bronchoscopy performed* **showing large mucus plug** *on the left on transfer ...*

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 ...*
Proxy	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	0.38	*... anticoagulation monitored on tele pump* **systolic dysfunction with ef** *of seen on recent echo ...*
Logistic	0.36	*... seen the mitral valve leaflets* **are mildly thickened there** *is nomitral valve prolapse ...*
Proxy	0.38	*... anticoagulation monitored on tele pump* **systolic dysfunction with ef** *of seen on recent echo ...*

Table 3: Comparison of the clinical evaluation from Mullenbach et al. (2018) with our plausibility evaluation. There are three examples above, each which contains the explanations produced by seven 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 three systems below the dotted line are from our evaluation, for which the second column indicates the probability output of our plausibility classifier. Here, the proxy and DR-CAML produce almost identical explanations; additional comparisons between DR-CAML and the proxy are shown in Table 4.

5 Plausibility Evaluation

Explanations are considered plausible if they can be reasoned about by a person. Thus, evaluating plausibility is typically more difficult than faithfulness, because it requires input from annotators (Herman, 2017). 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 CAML's explanations by collecting annotations from a clinician. Wiegreffe and Pinter (2019) argued that the attention mechanism of CAML and DR-CAML generally provide plausible explanations, even if they at times are not faithful to the model's internal decision-making. For each model they considered, they extracted an explanation in the form of a 14-token subsequence taken from the discharge summary. The clinician read all (anonymized) four explanations and the corresponding ICD code and rated each explanation as either "informative" or not. CAML was rated slightly more informative than logistic regression and CNN baselines. Table 3 shows explanations produced by Mullenbach et al. (2018)'s methods as well as the ones we consider in this work.

The format of Mullenbach et al. (2018)'s plausi-

296.20: "Major depressive affective disorder, single episode, unspecified"		
DR-CAML	0.47	... *diagnosis overdose of medications narcotics* **benzodiazepine suicide attempt chronic** *migraine headaches depression stage iv* ...
Proxy	0.33	... *up from the medications you* **were evaluated by psychiatry** *and will be transferred to* ...
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* ...
Proxy	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* ...
Proxy	0.46	... *the colon gastroesophageal reflux asthma* **irritable bowel syndrome gastroparesis** *osteoporosis anxiety and or depression* ...

Table 4: Differing explanations and classifier scores between DR-CAML and the proxy.

Model	Score	Interval
Logistic	35	(31, 49)
Cosine	38	(32, 51)
CNN	42	(33, 52)
CAML	44	(33, 52)
DR-CAML	48	(34, 53)
Proxy	52	(34, 54)

Table 5: 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 same total proportion of explanations are deemed plausible. The Interval column shows a 95% bootstrap interval from sampling 1000 labels from the classifier probabilities.

bility evaluation does not easily lend itself to replication. While the authors shared their annotations with us, missing metadata prevented a direct reproduction of their analysis. Additionally, since the clinical annotator considered explanations in a comparative setting, we cannot easily add our proxy model as another method using the same annotations. Therefore, we replicate this evaluation 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.[2] This annotation classifier achieves an accuracy of 67.2% and an AUC score of 0.726 on held-out explanations, indicating it is a useful but noisy stand-in for the clinician. Additional training details are in Appendix A.3.

To conduct our plausibility evaluation, we first use or reproduce the baseline methods from Mullenbach et al. (2018). Each model, including the proxy, 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 proxy 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[3] 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 5. The proxy model produces the largest number of informative explanations according to our classifier; however, the clas-

[2] We collapse the "informative" and "highly informative" labels into a single positive class.

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

sifier's inaccuracy introduces 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. We sample 1000 such sets of labels and report the 95% confidence interval for each model's score in the Interval column of Table 5. Accounting for this uncertainty dramatically reduces the differences between the methods. Because 95% of all classified plausibility probabilities are between 24.1% and 58.1%, these intervals skew towards lower scores. Despite the inherent uncertainty involved in extrapolating plausible scores from a fixed set of clinical annotations, our evaluations suggest that the proxy model produces explanations that are at least as plausible as those of DR-CAML.

Table 3 shows that for the three examples considered in Mullenbach et al. (2018), DR-CAML and our proxy model produce very similar explanations. This is perhaps surprising because DR-CAML extracts explanations using its attention mechanism, whereas the proxy model uses unigram feature importance values that do not vary between examples. For these examples, it appears that the proxy is faithful both in the predictions it makes and how it makes those predictions. Table 4 shows three examples where the proxy and DR-CAML diverge the most. These rare cases highlight two benefits of the proxy model. First, its feature importance weights are *global* across all predictions, providing an aggregate representation of the proxy's behavior. Second, the approach for extracting proxy 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 for post-hoc explanations that is designed to be interpretable and plausible while maintaining faithfulness to the trained model. By constraining the proxy 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. Furthermore, our proxy model has only 50K parameters, compared to CAML's 6 million. 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 proxy 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 proxy model that can provide *global* feature explanations. Depending on the gap in predictive performance between the proxy and original model, a skeptic of post-hoc methods (e.g. Rudin (2019)) might prefer to discard the original model altogether and simply use the proxy's predictions, for which its explanations are faithful by design.

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. We believe this proxy model approach is generally applicable as a post-hoc interpretability method for arbitrary models, but this must be further studied on new datasets and different trained models. It is possible that in some domains, trained models might be more difficult to mimic than DR-CAML. If so, the application may require a trade-off between a less restrictive proxy model class and a less faithful proxy.

Our evaluation is also limited in that it only considers a single form of explanation: n-grams extracted via feature importances or attention weights. Recent work has explored alternate formulations for a quality explanation (Barocas et al., 2020); some formulations may be more or less accommodating of our proxy model method. Our plausibility evaluations rely heavily on a single set of expert annotations from which we extrapolate using a classifier. To demonstrate that our method can reliably provide both plausible and faithful explanations, additional evaluations must collect new plausibility annotations or build off of existing resources (DeYoung et al., 2020).

As the ML community continues to explore new directions for interpretable methods, definitions of desiderata may continue to evolve. Such criteria will always depend on the domain experts who turn to an ML method for decision support. Interpretable ML methods should clearly define how they expect to satisfy a criterion 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.

References

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.

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.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.

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

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.

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)*.

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.

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*.

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.

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.

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.

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.

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).

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 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.

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 7. 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 proxy 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 7. 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 7.

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 7.

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).

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 7: 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 5 for comparison.

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 5 in § 5, reproduced in Table 7 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.