Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission

Rich Caruana
Microsoft Research
rcaruana@microsoft.com

Yin Lou
LinkedIn Corporation
ylou@linkedin.com

Johannes Gehrke
Microsoft
johannes@microsoft.com

Paul Koch
Microsoft Research
paulkoch@microsoft.com

Marc Sturm
NewYork-Presbyterian Hospital
mas9161@nyp.org

Noémie Elhadad
Columbia University
noemie.elhadad@columbia.edu

ABSTRACT

In machine learning often a tradeoff must be made between accuracy and intelligibility. More accurate models such as boosted trees, random forests, and neural nets usually are not intelligible, but more intelligible models such as logistic regression, naive-Bayes, and single decision trees often have significantly worse accuracy. This tradeoff sometimes limits the accuracy of models that can be applied in mission-critical applications such as healthcare where being able to understand, validate, edit, and trust a learned model is important. We present two case studies where high-performance generalized additive models with pairwise interactions (GA^2Ms) are applied to real healthcare problems yielding intelligible models with state-of-the-art accuracy. In the pneumonia risk prediction case study, the intelligible model uncovers surprising patterns in the data that previously had prevented complex learned models from being fielded in this domain, but because it is intelligible and modular allows these patterns to be recognized and removed. In the 30-day hospital readmission case study, we show that the same methods scale to large datasets containing hundreds of thousands of patients and thousands of attributes while remaining intelligible and providing accuracy comparable to the best (unintelligible) machine learning methods.

Categories and Subject Descriptors

I.2.6 [**Computing Methodologies**]: Learning—*Induction*

Keywords

intelligibility; classification; interaction detection; additive models; logistic regression; healthcare; risk prediction

1. MOTIVATION

In the mid 90's, a large multi-institutional project was funded by Cost-Effective HealthCare (CEHC) to evaluate the application of machine learning to important problems in healthcare such as predicting pneumonia risk. In the study, the goal was to predict the probability of death (POD) for patients with pneumonia so that high-risk patients could be admitted to the hospital while low-risk patients were treated as outpatients. In the study [3, 2], the most accurate models that could be trained were multitask neural nets.[1] On one dataset the neural nets outperformed traditional methods such as logistic regression by wide margin (the neural net had AUC=0.86 compared to 0.77 for logistic regression), and on the other dataset used in this paper outperformed logistic regression by about 0.02 (see Table 2). Although the neural nets were the most accurate models, after careful consideration they were considered too risky for use on real patients and logistic resgression was used instead. Why?

One of the methods being evaluated was rule-based learning [1]. Although models based on rules were not as accurate as the neural net models, they were *intelligible*, i.e., interpretable by humans. On one of the pneumonia datasets, the rule-based system learned the rule "HasAsthama(x) ⇒ LowerRisk(x)", i.e., that patients with pneumonia who have a history of asthma have lower risk of dying from pneumonia than the general population. Needless to say, this rule is counterintuitive. But it reflected a true pattern in the training data: patients with a history of asthma who presented with pneumonia usually were admitted not only to the hospital but directly to the ICU (Intensive Care Unit). The good news is that the aggressive care received by asthmatic pneumonia patients was so effective that it lowered their risk of dying from pneumonia compared to the general population. The bad news is that because the prognosis for these patients is better than average, models trained on the data incorrectly learn that asthma lowers risk, when in fact asthmatics have much higher risk (if not hospitalized).

One of the goals of the study was to perform a clinical trial to determine if machine learning could be used to predict risk prior to hospitalization so that a more informed decision about hospitalization could to be made. The ultimate goal was to reduce healthcare cost by reducing hospital admissions, while maintaining (or even improving) outcomes by more accurately identifying patients that need hospitalization. As the most accurate models, neural nets were a strong candidate for clinical trial. Deploying neural net models that could not be understood, however, was deemed too risky —

[1]SVMs and boosted trees were not in common use yet, and Random Forests had not yet been invented.

KDD'15, August 10-13, 2015, Sydney, NSW, Australia.

ACM 978-1-4503-3664-2/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2783258.2788613 .

if the rule-based system had learned that asthma lowers risk, certainly the neural nets had learned it, too. The rule-based system was intelligible and modular, making it easy to recognize and remove dangerous rules like the asthma rule. While there are methods for *repairing* the neural nets so they do not incorrectly predict that asthmatics are at lower risk and thus less likely to need hospitalization, e.g., re-train without asthmatics in the population, remove the asthma feature, modify the targets for asthmatics to "1" in the data to reflect the care they received (unfortunately confounding care with death), the decision was made to not use the neural nets not because the asthma problem could not be solved, but because the lack of intelligibility made it difficult to know what *other* problems might also need fixing. Because the neural nets were more accurate than the rules, it was possible that the neural nets had learned other patterns that could put some kinds of patients at risk if used in a clinical trial. For example, perhaps pregnant women with pneumonia also receive aggressive treatment that lowers their risk compared to the general population. The neural net might learn that pregnancy lowers risk, and thus recommend not admitting pregnant women, thus putting them at increased risk. In an effort to "do no harm", the decision was made to go forward only with models that were intelligible such as logistic regression, even if they had lower AUC than other unintelligible models. The logistic regression model also learned that having asthma lowered risk, but this could easily be corrected by changing the weight on the asthma feature from negative to positive (or to zero).

Jumping two decades forward to the present, we now have a number of new learning methods that are very accurate, but unfortunately also relatively unintelligible such as boosted trees, random forests, bagged trees, kernelized-SVMs, neural nets, deep neural nets, and ensembles of these methods. Applying any of these methods to mission-critical problems such as healthcare remains problematic, in part because usually it is not ethical to modify (or randomize) the care delivered to patients to collect data sets that will not suffer from the kinds of bias described above. Learning must be done with the data that is available, not the data one would want. But it is critical that models trained on real-world data be validated prior to use lest some patients be put at risk, which makes using the most accurate learning methods challenging.

In this paper we describe the application of a learning method based on high-performance generalized additive models [5, 6] to the pneumonia problem described above, and to a modern, much larger problem predicting 30-day hospital readmission. On both of these problems our GA^{2}M models yield state-of-the-art accuracy while remaining intelligible, modular, and editable. We believe this class of models represents a significant step forward in training models with high accuracy that are also intelligible. The main contributions of this paper are that it: shows that GA^{2}Ms yield competitive accuracy on real problems; demonstrates that the learned models are intelligible; demonstrates that the predictions made by the model for individual cases (patients) also are intelligible, and demonstrates how, because the models are modular, they can be edited by experts.

The remainder of the paper is organized as follows. Section 2 provides a brief introduction to GAM and GA^{2}M. Sections 3 and 4 present our case studies of training intelligible GA^{2}M model on the pneumonia and the 30-day readmission data, respectively. Section 5 discusses a wide range of issues that arise when learning with intelligible models and our general lessons for the research community.

2. INTELLIGIBLE MODELS

Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_1^N$ denote a training dataset of size N, where $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})$ is a feature vector with p features and y_i is the target (response). We use x_j to denote the jth variable in the feature space.

Generalized additive models (GAMs) are the gold standard for intelligibility when low-dimensional terms are considered [4, 5, 6]. Standard GAMs have the form

$$g(E[y]) = \beta_0 + \sum f_j(x_j), \tag{1}$$

where g is the link function and for each term f_j, $E[f_j] = 0$. Generalized linear models (GLMs), such as logistic regression, are a special form of GAMs where each f_j is restricted to be linear. Since the contribution of a single feature to the final prediction can be easily understood by examining f_j, such models are considered intelligible.

To improve accuracy, pairwise interactions can be added to standard GAMs, leading to a model called GA^{2}Ms [6]:

$$g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j). \tag{2}$$

Note that pairwise interactions are intelligible because they can be visualized as a heat map. GA^{2}M builds the best GAM first and then detects and ranks all possible pairs of interactions in the residuals. The top k pairs are then included in the model (k is determined by cross-validation).

There are various methods to train GAMs and GA^{2}Ms. Each component can be represented using splines, leading to an optimization problem which balances the smoothness of splines and empirical error [7]. Other representations include regression trees on a single or a pair of features. Empirical study showed gradient boosting with bagging of shallow regression trees yields as components very good accuracy [5]. Interested readers are referred to [5, 6] for details.[2]

3. CASE STUDY: PNEUMONIA RISK

In this case study we use one of the pneumonia datasets discussed earlier in the motivation [3]. This dataset has 14,199 pneumonia patients. To facilitate comparison with prior work, we use the same train and test set folds from the earlier study: the train set contains 9847 patients and the test set has 4352 patients (a 70:30 train:test split). There are 46 features describing each patient. These range from history features such as age and gender, to simple measurements taken at a routine physical such as heart rate, blood pressure, and respiration rate, to lab tests such as White Blood Cell count (WBC) and Blood Urea Nitrogen (BUN), to features read from a chest x-ray such as lung collapse or pleural effusion. See Table 1 for a complete list.

As discussed earlier, the goal is to predict probability of death (POD) so that patients at high risk can be admitted to the hospital, while patients at low risk are treated as outpatients.[3] 10.86% of the patients in the dataset (1542 patients) died from pneumonia. The GAM/GA^{2}M models are

[2]Code is available at `https://github.com/yinlou/mltk`.

[3]Hospitals are dangerous places, particularly for patients with impaired immune systems. Treating low-risk patients as outpatients not only saves money, but is actually safer.

Patient-history findings			
chronic lung disease	-	age	C
re-admission to hospital	-	gender	-
admitted through ER	-	diabetes mellitus	-
admitted from nursing home	-	asthma	-
congestive heart failure	-	cancer	-
ischemic heart disease	-	number of diseases	C
cerebrovascular disease	-	history of seizures	-
chronic liver disease	-	renal failure	-
history of chest pain	-		
Physical examination findings			
diastolic blood pressure	C	wheezing	-
gastrointestinal bleeding	-	stridor	-
respiration rate	C	heart murmur	-
altered mental status	-	temperature	C
heart rate	C		
Laboratory findings			
liver function tests	-	BUN level	C
glucose level	C	creatinine level	C
potassium level	C	albumin level	C
hematocrit	C	WBC count	C
percentage bands	C	pH	C
pO2	C	pCO2	C
sodium level	C		
Chest X-ray findings			
positive chest x-ray	-	lung infiltrate	-
pleural effusion	-	pneumothorax	-
cavitation/empyema	-	chest mass	-
lobe or lung collapse	-		

Table 1: Pneumonia attributes, grouped by type. Continuous features that will be shaped by GAM/GA^{2}M models are marked with a "C".

trained on this data using 100 rounds of bagging. Bagging is done to reduce overfitting, and to provide pseudo-confidence intervals for the graphs in the intelligible model.

The AUC area for different models trained on this data are shown in Table 2. On this dataset logistic regression achieves AUC = 0.843, Random Forests achieves 0.846, LogitBoost 0.849, GAM 0.854, and GA^{2}M is best with AUC = 0.857.[4] The difference in AUC between the methods is not huge (less than 0.02), but it is reassuring to see the GAM/GA^{2}M methods achieve the best accuracy on this problem. The important question is if the GAM/GA^{2}M models are able to achieve this accuracy while remaining intelligible?

Figure 1 shows 28 of the 56 terms in the GA^{2}M model for pneumonia. Unfortunately, the compact representation necessary for the paper reduces intelligibility. For small models like this with fewer than 100 terms we would prefer to present all terms, possibly sorted by their importance to the model. In the actual deployment, for each term we would also show a histogram of data density for different values of the feature, descriptive statistics about the feature, several different measures of term importance in the model, and links to online resources that provide information about the term, e.g., links to a hospital database, or Wikipedia or WebMD pages that describe features, how they are measured, what the normal ranges are, and what abnormal values indicate. Because of space limitations we have suppressed all of this auxiliary information (including some axis labels!) and just present shape plots for some of the more interesting terms. Presenting the terms in multicolumn format without the auxiliary information further hinders intelligibility — the models are more readable when

[4]The GA^{2}M model uses 10 of the $46 * 45/2 = 1035$ possible pairwise interaction terms (k chosen by cross-validation).

Model	Pneumonia	Readmission
Logistic Regression	0.8432	0.7523
GAM	0.8542	0.7795
GA^{2}M	0.8576	0.7833
Random Forests	0.8460	0.7671
LogitBoost	0.8493	0.7835

Table 2: AUC for different learning methods on the pneumonia and 30-day readmission tasks.

presented in sorted order as a scrollable list of graphs plus auxiliary information.

The 1st term in the model is for age. Age (in years) on the x-axis ranges from 18-106 years old (the pneumonia dataset contains only adults). The vertical axis is the risk score predicted by the model for patients as a function of age. The risk score for this term varies from -0.25 for patients with age less than 50, to a high of about 0.35 for patients age 85 and above. The green errorbars are pseudo-errorbars of the risk score predicted for each age: each errorbar is ± 1 standard deviation of the variation in the risk score measured by 100 rounds of bagging. We use ± 1 standard deviation instead of the standard error of the mean because it is well known that bagging underestimates the variance of predictions from complex models. We believe it is safer to be conservative than to present unrealistically narrow confidence intervals. (See the top of Figure 3(a) for an enlarged version of this graph, and the discussion in Section 5.5 for more detailed analysis of the age feature.)

The 2nd term in the model, asthma, is the one that caused trouble in the CEHC study in the mid-90's and prevented clinical trials with the very accurate neural net model. The GA^{2}M model has found the same pattern discovered back then: that having asthma lowers the risk of dying from pneumonia. As with the logistic regression and rule-based models trained then, but unlike with the neural net models, this term is easy to recognize and fix in the GA^{2}M model. We can "repair" the model by eliminating this term (effectively setting the weight on this graph to zero), or by using human expertise to redraw the graph so that the risk score for asthma=1 is positive, not negative. Because asthma is boolean, it is not necessary to use a graph, and we could present a weight and offset (RiskScore = w*hasAsthma + b) instead. We prefer to use graphs for boolean terms like asthma for three reasons: 1) it is necessary to show graphs for features with multiple or continuous values such as age as well as for interactions between features, and it is awkward for the user to jump from terms presented as graphs to terms presented as equations; 2) we find graphs provide an intuitive display of risk where up implies higher risk, down implies lower risk, and the magnitude of the change in risk is captured by the distance moved; and 3) some users are not as comfortable with numbers as they are with graphs, and it is important that the model is intelligible to real users, whatever their background.

The 3rd term in the model is BUN (Blood Urea Nitrogen) level. Most patients have BUN=0 because, as in many medical datasets, if the variable is not measured or assumed normal it is coded as 0. The model says risk is reduced for patients where BUN was not measured, suggesting that this test typically is not ordered for patients who appear to be healthy. BUN levels below 30 appear to be low risk,

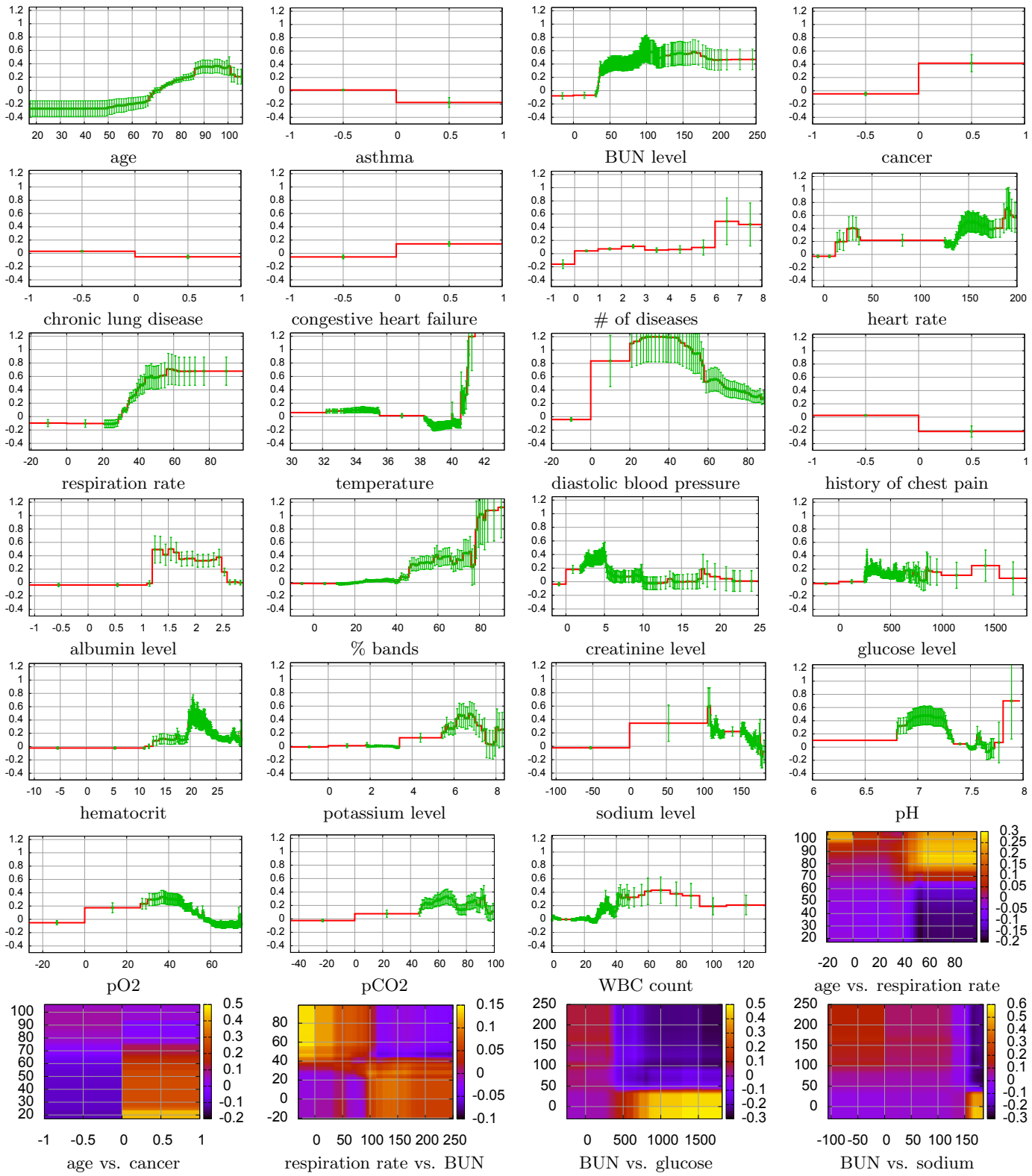


Figure 1: 28 (of 56 total) components for the GA^{2}M model trained on the pneumonia data. The line graphs are terms that contain single features. The heat maps at the bottom are pairwise interaction terms. The vertical scale on all line graphs are the same to facilitate rapid scanning of the relative contribution of each term. The green errorbars are pseudo-errorbars from bagging. Boolean features such as asthma are presented as graphs because this aids interpretation among other features that must be presented as graphs.

while levels from 50-200 indicate higher risk. This is consistent with medical knowledge which suggests that normal, healthy BUN is 10-20, and that elevated levels above 30 may indicate kidney damage, congestive heart failure, or bleeding in the gastrintestinal tract.

The cancer term in the model is clear: having cancer significantly increases the risk of dying from pneumonia, probably because it explains why the patient has pneumonia (either from lung cancer, from immuno suppressive drugs used to treat cancer, or from hospitalization associated with cancer) and/or because it explains the stage of cancer (terminal stages of cancer frequently lead to failing health and being bed-ridden, both of which can lead to pneumonia).

The next term in the model, chronic lung disease, and the history of chest pain term that occurs later, are interesting because the model suggests that chronic lung disease and a history of chest pain both *decrease* POD. We suspect that this may be a similar problem as asthma: patients with lung disease and chest pain may receive care earlier, and may receive more aggressive care. If further investigation suggests this to be the case, both terms would be removed from the model, or edited, similar to the asthma term.

The # of diseases (# of comorbid conditions) is a general measure of illness. The graph suggests that having no diseases other than pneumonia lowers risk, that risk increases slowly as the number of comorbid conditions increases from 1-3, then is flat or decreases until it rises dramatically above 6, but the errorbars are large enough to be consistent with risk being somewhat flat for 3-8 comorbidities.

Heart rate is an unusual looking graph. 91% patients have rate=0, indicating it was not measured or assumed normal. Risk is high for very low heart rates (10-30), and for very high rates (125-200), but the model does not appear to discriminate between patients with heart rates 40-120. On further inspection, there are no patients with heart rate recorded between 40-120! Apparently all patients in this range were considered "normal" and coded as 0. (Normal heart rate in adults is about 60-100, 40-60 for athletes, and somewhat higher than 100 for patients with "White Coat" Syndrome). The unusual shape of the graph for heart rate has lead us to discover a surprising aspect of the data, though it is not clear what risk we would want to model to predict for rates between 40-120 where there is no data?

The respiration rate term is very clear: rate=0 (missing) or 20-28 is low risk, and risk rises rapidly as breathing rate climbs from 28-60. Normal respiration rate for adults is 16-20. In the body temperature term, temps from 36°C-40°C are low risk (normal is 37°C), risk is somewhat elevated at low body temps (32°C-36°C), and greatly elevated for temps above 40.5°C (fever this high often is a sign of serious infection). Having a fever above 41.5°C increases the risk score by a full point or more.[5] Diastolic blood pressure also can dramatically increase risk: low diastolic in the range 20-50 (normal is 60-80) increase risk as much as a full point. % bands is also a strong term (bands in a blood test are a sign of bacterial infection—bacterial pneumonia is more deadly than viral pneumonia): bands above 40% indicate elevated risk, with bands above 80% indicating very elevated risk.

Before leaving pneumonia, let us examine one of the interaction terms. In the age vs. cancer term, we see that risk is highest for the youngest patients (probably cancers acquired in childhood but not cured when the patient reaches age 18), and declines for patients who acquire cancer later in life, but for patients without cancer risk rises as expected with age. This is a classic interaction effect that likely results from the difference between childhood and adult cancers.

Space prevents us from discussing each term individually, or from discussing terms in great detail. See Section 5.5 for a deeper dive on the age term. To summarize this section, the GA^{2}M model discovered the same asthma pattern that created problems in the CEHC study, provides a simple mechanism to correct this problem, and uncovered other similar problems (chronic lung disease and history of chest pain) that were not recognized in the CEHC study but which warrant further investigation and probably repair. As hoped, the GA^{2}M model is accurate, intelligible, and repairable.

[5]An increase in risk of 1 point more than doubles the odds of dying. See Section 5.1.

4. CASE STUDY: 30-DAY READMISSION

In this section we apply GA^{2}M to a modern and much larger dataset for 30-day hospital readmission. The data comes from a collaboration with a large hospital. There are 195,901 patients in the train set (2011-2012), 100,823 patients in the test set (2013), and 3,956 features for each patient. Features include lab test results, summaries of doctor notes, and details of previous hospitalizations. In this problem, the goal is to predict which patients are likely to be readmitted to the hospital within 30 days after being released from the hospital. All patients in this dataset have already been hospitalized at least once, and the goal is to predict if they will need to return to the hospital unusually quickly (within 30 days). Hospitals with abnormally high 30-day readmission rates are penalized financially because a high rate suggests the hospital did not provide adequate care on the earlier admission, or may have released the patient prematurely, or did not provide adequate instructions to the patient when they were released, or did not perform adequate follow-up after release. In the data 8.91% of patients are readmitted within 30 days. For this problem we use 10 iterations of bagging. Training the 10 models takes 2-3 days on a small cluster of 10 general purpose computers. Table 2 shows the AUC for different models on this data.

In Section 3 we examined the GA^{2}M model for the pneumonia problem. Unfortunately, the readmission dataset contains almost 100 times as many features. Instead of trying to examine the full model, we instead examine the predictions made by the model for three patients. Two of these patients have very high predicted probability of readmission (p=0.9326 and p=0.9264), and one of the patients has a typical readmission probability (p=0.0873). This allows us to demonstrate that the models are intelligible not only taken as a whole, but that the predictions GA^{2}M models make for individual patients also are intelligible.

In Figure 2, each of the three columns is a patient, and each row is a term in the model. Terms are sorted for each patient (in each column) by the risk they contribute to that patient for 30-day readmission. Space limits us to showing the top 6 terms for each patient that contributed most to risk. Patient #1 has a very high probability of readmission within 30 days: p=0.9326. The four terms that contribute most to their high probability of readmission are: their total number of visits to the hospital is 40, they have been an inpatient in the hospital 19 times in the last 12 months, they have been in the hospital 10 times in the last 6 months, and 4 times in the last 3 months. This is not unusual: the most

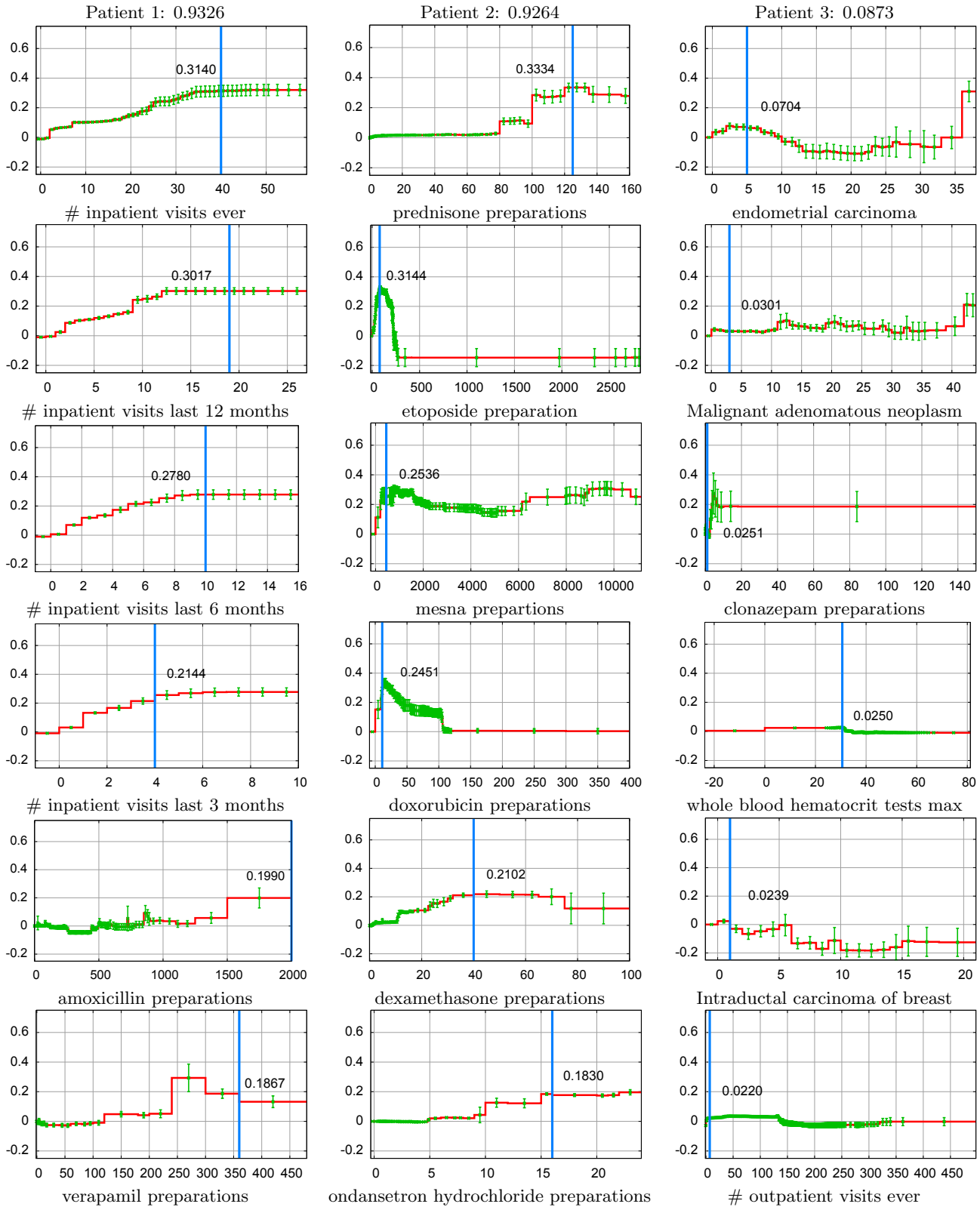


Figure 2: Top 6 terms (of 4456) in the GA^{2}M for three patients. The patients on the left have high risk of readmission. The patient on the right has moderate risk. Terms are sorted by their contribution to risk. Blue lines highlight feature values and corresponding risk scores. Six terms cannot tell the full story for these patients, but even these few terms provide insight into the patients and their risk of readmission.

predictive terms in the 30-day readmission model measure the number of visits patients have made in the last 12 month, 6 months, and 3 months to the ER, as an outpatient, and as an inpatient. As we see with this patient, a large number of recent inpatient visits (admissions) is associated with a high probability of readmission.[6] The next two terms suggest why patient #1 may have been in the hospital often: this patient has received large doses of amoxicillin (an antibiotic used to treat infections like strep and pneumonia) and verapamil (a treatment for hypertension and angina), i.e., they have an ongoing infection that may not be responding to antibiotics, and also probably have heart disease. The main reason this patient is predicted to be likely to return is because they have been in the hospital often in the last year, but the first few terms in the model also give us a hint of the medical conditions that put them at elevated risk.

The terms that are most important for patient #2 (also high risk: p=0.9364) are different from the terms that were important for patient #1. The most important 6 terms are preparations that the patient received during their last visit: prednisone is a corticosteroid used as an imummosuppressant, etoposide in an anticancer drug, mesna is a cancer chemotherapy drug, doxorubicin is a treatment for blood and skin cancers, dexamethosone is another immunosuppressant steroid, and ondansetron is a drug used to treat nausea from chemotherapy. Patient #2 has received doses of each of these preparations that suggest cancer may not be responding well to treatment and that they are receiving aggressive chemotherapy. The contribution to risk from these 6 terms alone is greater than +1.5, i.e., these 6 terms alone triple the odds of their being readmitted within 30-days.

Patient #3 has moderate risk: p=0.0873 (baseline rate is 8.91%). This 6 terms that increase this patient's readmission risk the most are: 1) the patient has endrometrial carcinoma (a cancer common in post-menopausal women that often can be treated effectively by hysterectomy without radiation- or chemo-therapy); 2) a benign abdominal tumor (malignant adenomatous neplasm =3); 3) a relaxant typically prescribed to calm patients or reduce spasms; 4) a fairly typical (i.e. low risk) hematocrit test result; 5) a precancerous non-invasive lesion in the breast; and 6) a small number of outpatient visits suggesting they have been receiving treatment as an outpatient without needing to be hospitalized (the inpatient and ER risk factors for this patient are all low). Patient #3 is a typical patient as far as 30-day readmission is considered. They are post-menopausal, have cancers that respond well to treatment if caught early, the treatments themselves are relatively low-risk, and they have not needed unusual medications or to be hospitalized often in the last year.

The patients above provide a small glimpse of what the GA^{2}M model learned from a 200,000 patient train set with 4,000 features: we have only been able to examine three patients, and have only looked at the top 6 terms for each of these patients. To a medical expert, the sorted terms for each patient present a comprehensive picture of the risk factors that contribute to the probability of readmission predicted for a patient. The model is not causal — it does not say that because the patient has X, they received treatments A, B, and C, and we can see from the amount of A, B, and C they received that they are not responding well. Instead, it learns that high doses of A, B, and C are associated with high risk or readmission, and it is up to the human experts to infer the underlying causal reasons for the feature values and the risk they predict. Nevertheless, compared to an unintelligible model such as an ensemble of 1000 boosted trees or a complex neural net, the model is fairly transparent, and the predictions it makes can be fully "understood", both at the per-patient level, and at the macro-model level.

[6]A large number of visits to the ER also is associated with increased chance of readmission, but outpatient visits are more interesting: a small number of recent outpatient visits increases risk of readmission, but a very large number of outpatient visits (100-200 in the last year) indicates lower risk of readmission because the patient is receiving primary care as an outpatient—many of these patients are dialysis patients who visit the hospital 1-2 times per week.

RiskScore	Probability	RiskScore	Probability
-5.0	0.0067	+5.0	0.9933
-4.0	0.0180	+4.0	0.9820
-3.0	0.0474	+3.0	0.9526
-2.0	0.1192	+2.0	0.8808
-1.0	0.2689	+1.0	0.7311
0.0	0.5000		

Table 3: Risk scores (log odds) and the corresponding probabilities.

5. DISCUSSION

5.1 How To Interpret Risk Scores

Each term in the intelligible model returns a risk score (log odds) that is added to the patient's aggregate predicted risk. Terms with risk scores above zero increase risk; terms with scores below zero decrease risk. The term risk scores are added to a baseline risk, and the sum converted to a probability. Both penumonia and 30-day readmission have baseline rates near 0.1, which corresponds to TotalRiskScore = -2.197. So patients with aggregate risk scores above -2.2 have higher than average risk, and patients with total risk scores below -2.2 have lower than average risk scores. A patient with TotalRiskScore = 0 (including the baseline offset) has quite high risk: $p = 1/(1+\exp(-1*TotalRiskSccore) = 1/(1+\exp(0)) = 0.5$. Table 3 shows a sample of total risk scores and the corresponding probabilities.

5.2 Modularity

In the intelligible models discussed in this paper, the average risk score for each graph (i.e., each term: each feature or pair of features) averaged across the training set is set to zero by subtracting the mean score. A single bias term is then added to the model so that the average predicted probability across all patients equals the observed baseline rate. This is done to make models identifiable and modular. Because of this property, each graph can be removed from the model (zeroed out) without introducing bias to the predictions. If all terms were removed from the model, the only remaining term would be the bias term, and the probability predicted for all patients would be the observed baseline rate in the training set. Adding terms (graphs) to the model increases the model's discriminativeness without altering the prior. This is important because it increases modularity and makes it easier to interpret the contribution of each term:

negative scores decrease risk, and positive scores increase risk compared to the baseline risk.

5.3 Sorting Terms by Importance

If a model contains a modest number of terms (e.g., less than 50), it is best to show terms in the model to experts in the order they are most familiar with. Because experts are often used to seeing features in logical groupings, interpretation is aided by preserving these groupings when the model is presented. However, when the number of terms grows large, it becomes infeasible for experts to examine all terms carefully. Term importance often follows a power-law distribution, with a few terms being very important, a modest number of terms being somewhat important, and many terms being of little importance. When this is the case, intelligibility can be improved by sorting terms by a measure of importance such as the drop in AUC when the term is removed, or the skill of the term measured in isolation, or the maximum contribution (positive or negative) that the term can make for any patient. No one measure is correct or best, and we find that a sort that reflects a combination of these metrics seems to work well.

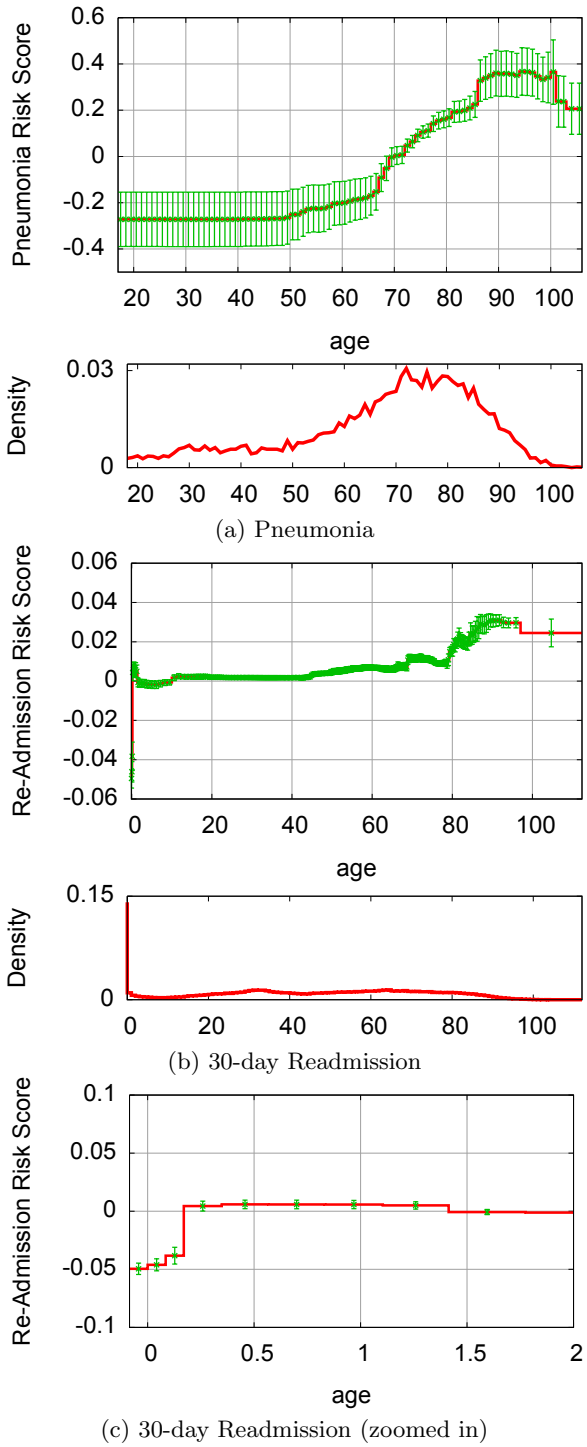
It is much easier to sort terms by importance when making prediction for a single patient: because each term yields a single risk score for each patient at the point where that patient's feature value lies on the term graph, it is possible to sort terms by how much they increase or decrease risk for the patient. This provides a well-defined ordering of the terms for a patient from terms that increase risk most to terms that decrease risk most. Often this ordering quickly identifies the key patient characteristics that best explain the model's prediction, and which help experts quickly understand the patient's condition. This is the method we used to describe the predictions made by the 30-day readmission model — although that model contains more than 4000 terms, the number of terms that are relevant for each patient are, in practice, often quite small (e.g., less than 100).

5.4 Feature Shaping vs. Expert Discretization

Significant effort was made in the CEHC pneumonia study to train accurate models with logistic regression and other methods that could not handle continuous attributes. Medical experts carefully discretized each continuous attribute into clinically meaningful ranges used to define boolean variables. For example, the intervals for age were 18-39, 40-54, 55-63, 64-68, 69-72, 73-75, 76-79, 80-83, 84-88, and 89+. We used these expert-defined intervals for the logistic regression model reported in Table 2. We also trained a GA^{2}M model with these discretized features, and observed a drop in AUC of about 0.01 on the test set compared to the GA^{2}M trained with the continuous features, suggesting that the GA^{2}M model gains some of its accuracy by shaping continuous features more accurately than expert discretization.

5.5 Deep Dive: Risk as a Function of Age

In this section we drill down on how the feature "Age" is shaped by the pneumonia and 30-day readmission models. Age is present in both data sets and measured in years. But the relevance of age to the two prediction tasks is very different. In pneumonia, age is a critical factor that can explain why a patient has acquired pneumonia, and what the outcome is likely to be.[7] In 30-day all-cause readmission,



(a) Pneumonia

(b) 30-day Readmission

(c) 30-day Readmission (zoomed in)

Figure 3: Risk as a function of Age for the Pneumonia and 30-day Readmission problems.

[7]Pneumonia is sometimes called "The Old Man's Best Friend", not because pneumonia is good for elderly patients, but because it often results in rapid death for patients that otherwise could linger for months or years before their primary illness causes death.

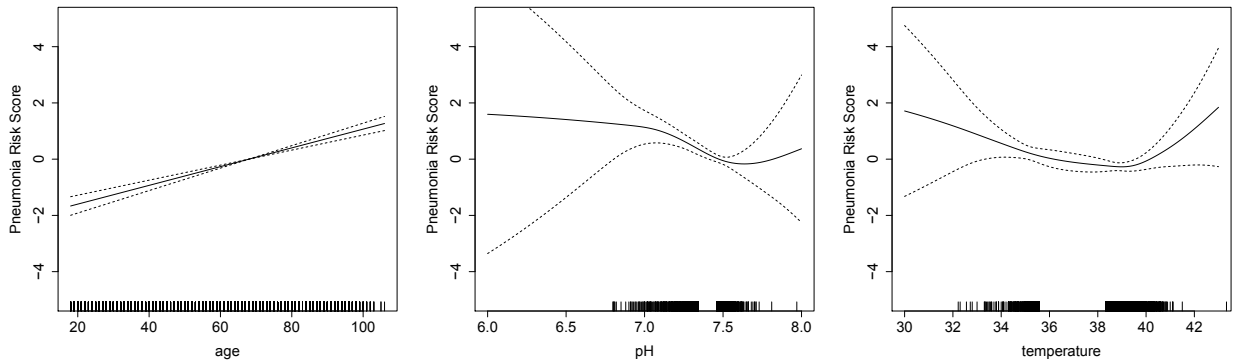


Figure 4: Selected splines in pneumonia dataset.

however, age is just one of thousands of factors that affect a patient's health and course of illness. Moreover, because the prediction task is hospital readmission, not probability of death, age represents a weaker, more generic characterization of patient health and their likeliness to need additional hospitalization within 30 days. If the patient is elderly, but just had a successful hip replacement or kidney stone removed, they are not likely to need to return to the hospital within 30 days for this condition. Similarly, an elderly patient who was admitted to the hospital because of pneumonia, but who is now being released because they responded to treatment, is unlikely to need further care for pneumonia within 30 days if they take proper medications. All-cause readmission is very different from probability of death for a specific condition such as pneumonia.

Figure 3(a) shows the risk profile for age in the pneumonia model, and the distribution of age in the pneumonia data. The majority of pneumonia patients are age 60-90. Qualitatively, the risk of dying from pneumonia is low and constant from age 18-50, rises slowly from age 50-66, then rises quickly from age 66-90, and then levels off at very high risk above age 90. The low-risk region to the left of age 50 is remarkably flat, suggesting that the underlying trees rarely if ever found it useful to split this region into subregions. Note that the risk score for this region is -0.27, suggesting that being young significantly reduces the risk of dying from pneumonia. But risk slowly increases as age increases above 50, though the contribution to risk does not become positive until about 70 years. Beyond 70 years old, the contribution to risk rises rapidly from 0.0 at 70 to +0.20 at age 82 and +0.35 at age 86. According to the model, the increase in risk of going from 70 to 86 is larger than the decrease is risk of going from 70 down to 50 or less.

Beyond the risk vs. age profile described above, there are intriguing details in the graph. 1) There is a small jump in risk at age 67, and again at age 86. The error bars are reasonably tight around age 65-70, suggesting that the jump in risk at 67 may be real. One possible explanation for this is that in a dataset from the 90's, many patients would have retired at around age 65, and that this may yield differences in activity levels, health insurance, and willingness to get access healthcare early enough to improve outcomes — pneumonia responds well to treatment with antibiotics, but can be life threatening if not treated. The 2nd jump in risk around age 86 is harder to explain. It may be that practitioners, either consciously or subconsciously, treat patients older than 85 differently, and that this ultimately increases their POD. Or the jump at 86 may be an artifact of the model — the error bars at age 86 and above are larger. One approach to investigating this issue further would be to train on another sample sample of data (or on different subsamples) to see if the rise at age 86 persists.[8] 2) There is an apparent drop in risk above age 100. We suspect that this drop probably is not real and may be due to mild overfitting — there are very few patients age 95 and older, and the error bars from age 90 to 106 are large and consistent with risk being constant in this region.[9] 3) Surprisingly, there is no evidence that risk, although very high, increases above age 85. Either medical treatments are equally effective for patients older than 85, or other medical conditions are more likely to be responsible for death at this age than pneumonia, or risk does increase above 85 and the model has failed to learn it.

Figure 3(b) shows the age term and density for 30-day readmission. One of the key differences between the pneumonia and 30-day readmission datasets is that pneumonia dataset contains only adult patients age 18 and older, but the readmission dataset contains patients of all age, including newborn infants. The importance of age to 30-day readmission is *very* different. Age has little effect on readmission between age 2 and 50, risk slowly increases from age 50 to 80, and then increases a little more above age 80. The largest increase in score is +0.03 at age 90 and above. There are many reasons why age is less important for readmission than for pneumonia: most patients independent of age would not be released if the hospital thought they were likely to need to be readmitted in less than a month, in this dataset there are thousands of other more specific variables that can better explain variance in the risk of readmission (the model is more illness specific) than age, and some patients who are very elderly will die at home (either unexpectedly or by choice) and thus will not be readmitted.

[8]It is only because the model is so intelligible that we are able to recognize and question such fine detail in the risk vs. age profile. We assume that similar jumps in predicted risk occur in other accurate models such as boosted trees as well, but because those models are less intelligible the jumps are not recognized or investigated.

[9]Or it might be due to *successful agers*, a rare but genetically identifiable class of people with traits that better enable them to survive into old age

An interesting feature of the model for 30-day readmission is highlighted in Figure 3(c) where the x axis has been expanded to show age 0-2. In this dataset newborn infants are born into the hospital, and thus will be treated as readmitted if they need to be hospitalized within 30 days after going home. In part because newborns would not be released if they were at risk, the risk score for newborns age 0-2 months is -0.04—this is a larger negative risk score than the increase in risk for elderly patients. This suggests that most newborns tend to be healthy when they are released from the hospital and are less likely to need to be readmitted within 30 days. But this reduction in risk from being newborn diminishes after 2-3 months, and the model suggests that infants age 3-15 months have slightly higher positive risk of being readmitted to the hospital. Thus infants age 3-15 months have higher risk of readmission than infants that are younger or older, and it is not until age 45 that the risk of readmission rises to this level again.

5.6 Shaping with Splines

Generalized additive models are often fit with splines [7]. Splines allow GAMs to be trained with careful control over regularization and provide more principled error bars. Unfortunately, the spline methods tend to over regularize, yield less accuracy than GA^{2}M models, and yield risk profiles that sometimes miss detail discovered by GA^{2}M models. Figure 4 shows three terms from a spline GAM model trained on the pneumonia data. The 1st term is age, the 2nd is pH, and the 3rd is temperature. Although the splines capture the basic trends (e.g., risk increases with age, pH risk is least around 7.6, and fever risk rises above 40°C), the splines miss detail learned by GA^{2}M. For example, the GA^{2}M model for age is much more nuanced, and the spline model may not properly model temperature in the normal range 36°C-38°C. The spline GAM model has accuracy closer to logistic regression than GA^{2}M, so the extra detail learned by GA^{2}M increases accuracy and probably reflects genuine structure.

5.7 Correlation Does Not Imply Causation

Because the models in this paper are intelligible, it is tempting to interpret them causally. Although the models accurately explain the predictions they make, they are still based on correlation. If features were added to or subtracted and the model retrained, the graphs for some terms that had remained in the model would change because of correlation with the features added or subtracted. Although details of some of the shape plots are suggestive (e.g., does pneumonia risk truly jump as age increases above 65, and again above 85?), it is not (yet) clear if some details like this are due to a) overfitting; b) correlation with other variables; c) interaction with other variables; d) correlation or interaction with unmeasured variables; or e) due to true underlying phenomena such as retirement and change in insurance provider.

Perhaps the strongest statement we can make right now is that the models are intelligible enough to provide a window into the data and prediction problem that is missing with many other learning methods, and that this window allows questions to be raised that will require investigation and further data analysis to answer. In future versions of these models we hope to automate some of these analyses so that it is clearer what features in the intelligible model are "real" or due to random factors such as overfitting and spurious correlation. Adding causal analysis to the models would be tremendously useful, but is, of course, difficult.

6. CONCLUSIONS

We present two case studies on real medical data where GA^{2}Ms achieve state-of-the-art accuracy while remaining intelligible. On the pneumonia case study the GA^{2}M model learns patterns that previously prevented complex machine learning models from being deployed, but because GA^{2}M is intelligible and modular it is possible to edit the model to reduce deployment risk. On the larger, more complex 30-day hospital readmission task the GA^{2}M model achieves excellent accuracy while yielding a manageable, surprisingly intelligible model despite incorporating over 4000 terms. Using this problem we demonstrate how GA^{2}Ms can be used to explain the predictions the model makes for individual patients in a concise way that places focus on the most important/relevant terms for each patient. We believe GA^{2}Ms represent a significant step forward in the tradeoff between model accuracy and intelligibility that should make it easier to deploy high-accuracy learned models in applications such as healthcare where model verification and debuggability are as important as accuracy.

Acknowledgements. We thank Michael Fine, MD, University of Pittsburgh School of Medicine, and Greg Cooper, MD, PhD, University of Pittburgh for help with the Pneumonia data and model. We thank Eric Horvitz, MD, PhD, Microsoft Research Redmond for help with the 30-day hospital readmission data and model. The 30-day hospital readmission experiment was reviewed and approved by the institutional review board at Columbia University Medical Center.

7. REFERENCES

[1] R. Ambrosino, B. Buchanan, G. Cooper, and M. Fine. The use of misclassification costs to learn rule-based decision support models for cost-effective hospital admission strategies. In *Proceedings of the Annual Symp. on Comp. Application in Medical Care*, 1995.

[2] G. Cooper, V. Abraham, C. Aliferis, J. Aronis, B. Buchanan, R. Caruana, M. Fine, J. Janosky, G. Livingston, T. Mitchell, S. Montik, and P. Spirtes. Predicting dire outcomes of patients with community acquired pneumonia. *Journal of Biomedical Informatics*, 38(5):347–366, 2005.

[3] G. Cooper, C. Aliferis, R. Ambrosino, J. Aronis, B. Buchanan, R. Caruana, M. Fine, C. Glymour, G. Gordon, B. Hanusa, J. Janosky, C. Meek, T. Mitchell, T. Richardson, and P. Spirtes. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial Intelligence in Medicine*, 9(2):107–138, 1997.

[4] T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman & Hall/CRC, 1990.

[5] Y. Lou, R. Caruana, and J. Gehrke. Intelligible models for classification and regression. In *KDD*, 2012.

[6] Y. Lou, R. Caruana, J. Gehrke, and G. Hooker. Accurate intelligible models with pairwise interactions. In *KDD*, 2013.

[7] S. Wood. *Generalized additive models: an introduction with R*. CRC Press, 2006.