An Evaluation of the Doctor-Interpretability of Generalized Additive Models with Interactions

Stefan Hegselmann* STEFAN.HEGSELMANN@UNI-MUENSTER.DE
Thomas Volkert† THOMAS.VOLKERT@UKMUENSTER.DE
Hendrik Ohlenburg† OHLENBURG@UNI-MUENSTER.DE
Antje Gottschalk† ANTJE.GOTTSCHALK@UKMUENSTER.DE
Martin Dugas* DUGAS@UNI-MUENSTER.DE
Christian Ertmer† CHRISTIAN.ERTMER@UKMUENSTER.DE

** Institute of Medical Informatics, University of Münster, Germany*
† Department of Anesthesiology, Intensive Care and Pain Medicine, University Hospital Münster, Germany

Abstract

Applying machine learning in healthcare can be problematic because predictions might be biased, can lack robustness, and are prone to overly rely on correlations. Interpretable machine learning can mitigate these issues by visualizing gaps in problem formalization and putting the responsibility to meet additional desiderata of machine learning systems on human practitioners. Generalized additive models with interactions are transparent, with modular one- and two-dimensional risk functions that can be reviewed and, if necessary, removed. The key objective of this study is to determine whether these models can be interpreted by doctors to safely deploy them in a clinical setting. To this end, we simulated the review process of eight risk functions trained on a clinical task with twelve clinicians and collected information about objective and subjective factors of interpretability. The ratio of correct answers for dichotomous statements covering important properties of risk functions was 0.83 ± 0.02 ($n = 360$) and the median of the participants' certainty to correctly understand them was *Certain* ($n = 96$) on a seven-level Likert scale (one = *Very Uncertain* to seven = *Very Certain*). These results suggest that doctors can correctly interpret risk functions of generalized additive models with interactions and also feel confident to do so. However, the evaluation also identified several interpretability issues and it showed that interpretability of generalized additive models depends on the complexity of risk functions.

1. Introduction

Healthcare is a sensitive domain to apply machine learning (ML) because unacceptable predictions can lead to significant consequences. Past studies show that ML might yield biased predictions (Bolukbasi et al., 2016), can lack robustness in the sense that small perturbations in the input cause misclassifications (Szegedy et al., 2013), and that ML is prone to overly rely on correlation rather than causation in big data (Lazer et al., 2014). As an example from healthcare, Caruana et al. (2015) present a model that, contrary to medical evidence, learned to associate a history of asthma with a lower risk of dying from pneumonia. While these issues are well known to the research community and are incorporated into current research practice, it is often impossible to completely eliminate

them. This can cause mistrust and uncertainty in healthcare professionals and stakeholders, which constitutes a barrier for deploying ML in patient care (Wiens et al., 2019).

Interpretable ML, which adds the ability to explain or present in terms understandable to a human (Doshi-Velez and Kim, 2017), can mitigate these issues. As Doshi-Velez and Kim (2017) state, the need for interpretability stems from an incomplete problem formulation that fails to formalize auxiliary criteria, such as lack of bias, robustness, and causality, as optimization criteria of the ML system. Interpretability can visualize these gaps in problem formalization and puts the responsibility for meeting these additional desiderata on the practitioner. Interpretable ML can be subdivided into *transparency* and *post-hoc interpretability* (Lipton, 2016). The latter approach enriches a prediction with additional information explaining a decision and usually allows for more complex models, such as artificial neural networks (Ribeiro et al., 2016). However, we believe that there are many clinical scenarios where healthcare professionals lack time to verify an explanation. In these cases transparent models might be favourable as they allow professionals to validate a model once and then use predictions without the necessity to verify each one separately.

Generalized additive models with interactions (GA^{2}Ms) are transparent models consisting of one- and two-dimensional risk functions that can be visualized and assessed by human practitioners. Due to this fact, the authors consider these models *intelligible*. Nevertheless, GA^{2}Ms clearly outperform logistic regression and are only slightly inferior to random forests (Lou et al., 2013). However, interpretability of these models has not been evaluated with human subjects, which we deem unacceptable for deployment in a clinical setting. The key objective of this study is to evaluate whether clinicians can correctly understand and interpret GA^{2}Ms to determine if they can validate them for clinical usage. To this end, we performed an application-grounded evaluation of doctors validating a GA^{2}M (Doshi-Velez and Kim, 2017). We trained the model on a Medical Information Mart for Intensive Care (MIMIC-III) benchmark task for in-hospital mortality (Harutyunyan et al., 2019), presented risk functions of varying complexity to clinicians, and asked them to decide which risk functions should be included in the final model. We measured their objective level of understanding risk functions with a self-developed questionnaire. Moreover, clinicians could state their confidence in understanding a risk function and could provide feedback about factors that support or hinder interpretability (subjective level of understanding).

Generalizable Insights about Machine Learning in the Context of Healthcare

- Our evaluation suggests that doctors are able to correctly interpret risk functions of a GA^{2}M and feel confident in doing so, but it shows disagreement on which functions are against medical knowledge and should be excluded from the model.
- Interpretability of GA^{2}Ms cannot be generalized for the whole model, but it depends on the complexity of the learned risk functions and, hence, on the specific application.
- Varying bin sizes, complex function shapes, weak signals, and outlier values impede interpretability of risk functions.
- Application-grounded evaluations with medical professionals are important to assess the interpretability of a ML system in the context of healthcare.

2. Methods

We defined the methods of this study a priori in a study protocol (available on request) to reduce bias during study execution and statistical analysis. This report contains more detailed descriptions of the methods and we state all deviation from the protocol.

2.1. Data Preprocessing

We used the MIMIC-III database (Johnson et al., 2016a,b) to train a GA^{2}M for our interpretability evaluation because it is publicly accessible, contains a large amount of data, and is well known in the research community. MIMIC-III contains rich critical care data covering 53,423 ICU stays of 38,597 distinct patients from the Beth Israel Deaconess Medical Center in Boston, USA, collected between 2001 and 2012. To ensure reproducibility of the data preprocessing and performance evaluation, we used an existing MIMIC-III benchmark task (Harutyunyan et al., 2019). We chose the task *in-hospital mortality* (the first 48 hours of a stay are used to predict mortality) since it is a binary classification problem that fits well into the GA^{2}M framework and it is intuitive for clinicians. We adopted the feature generation pipeline of the logistic regression model with some adjustments that we published in a Zenodo repository to ensure reproducible experiments.[1] Instead of generating six different statistics for seven subsequences of each time series, we only computed the mean and standard deviation over the last 48 hours for each of the 17 input variables. This resulted in 34 features for our study, in contrast to the 714 features used in Harutyunyan et al. (2019). Fewer features were used to reduce the number of risk functions learned by the GA^{2}M. The mean and standard deviation over the last 48 hours were chosen as both meaningful and intuitive features in discussions with clinicians. Moreover, we adjusted three aspects of the preprocessing pipeline: we replaced mean imputation with a constant value (-1) imputation to treat unknowns separately, we disabled normalization to obtain risk functions with a meaningful scaling of the axes, and we removed 157 implausible measurements that would have distorted the visualization of risk functions.[2] The final training, validation, and test sets contained 14,681, 3,222, and 3,236 ICU stays consisting of 34 features and a binary label indicating in-hospital mortality.

2.2. Training Generalized Additive Model with Interactions

To train a GA^{2}M and visualize its risk functions, we developed a simple web application based on the source code[3] from Lou et al. (2013). We used default parameters for the training procedure. Training included a discretization of input values into bins. The only slight modification to the original code, was that we required an extra bin for unknown values to handle them separately in the risk functions. We trained the GA^{2}M with 1,000 iterations for one- and two-dimensional functions on the training set and chose the model with the highest validation score. We stopped the training earlier when the model clearly overfitted. For the final model used in the interpretability evaluation, we selected 34 of 561 two-dimensional risk functions to reduce the set of functions for questionnaire development.

1. https://doi.org/10.5281/zenodo.3597992
2. The values are given in mimic3models/resources/plausible_values.json (see Zenodo repository)
3. https://github.com/yinlou/mltk

GA^{2}M training determined the importance (variance) of risk functions and we selected the 17 most important two-dimensional risk functions for two mean features and a mean and standard deviation feature. A single run on the test set was performed for each model and we used the evaluation script provided by Harutyunyan et al. (2019) to determine the area under the receiver operating characteristic (AUC-ROC) and the area under the precision-recall curve (AUC-PR). The JavaScript library D3.js was used to visualize the risk functions. We plotted the unknown bin separately, implemented a simple mouse-over functionality to read the plots, and aligned the risk axes for one- and two-dimensional risk functions. A diverging two-sided color map was used to visualize two-dimensional risk functions (Moreland, 2009).

2.3. Study Participants and Instruction Procedure

Twelve doctors from the Department of Anesthesiology, Intensive Care and Pain Medicine at University Hospital Münster, Germany, were included in the study. The participants had no prior knowledge of the evaluation. To study the effect of different backgrounds, participants were chosen from three different subgroups: doctors with a scientific background (SB), medical specialists (MS), and medical residents (MR; see Table 1). All clinicians at the department were notified about this evaluation and the included doctors participated voluntarily. The evaluation was performed within four weeks at a single computer in a controlled and standardized environment.

Before the participants performed the interpretability evaluation, they were instructed about the task and the goals of the evaluation. A ten-minute video was used for this purpose to standardize the procedure, reduce the workload, and to give clinicians full flexibility to perform the evaluation. This video included the following aspects: (1) basic explanation of ML to learn from past data, (2) introduction of our application for in-hospital mortality prediction based on MIMIC-III, (3) statement of the goal of the evaluation, (4) introduction of GA^{2}Ms and one- and two-dimensional risk functions, and (5) an example questionnaire and an explanation of its structure. The example questionnaire had the same structure as the final questionnaire and contained one one-dimensional (mean feature) and one two-dimensional (mean and standard deviation feature) risk function.

2.4. Questionnaire Development

The goal of the evaluation was to simulate a clinical validation of a GA^{2}M and to determine the objective and subjective level of understanding of its risk functions. We were unable to identify an existing and validated survey in the literature that suited this purpose. Hence, we were forced to use a self-designed questionnaire (Table 1 summarizes the dimensions of the evaluation). To increase the questionnaire's validity, it was developed by a multidisciplinary team consisting of a ML researcher, a medical specialist with a scientific background, and a doctor with additional expertise in medicine didactics. Question and questionnaire design was guided by the principles in Krosnick and Presser (2010). In addition to that, we performed cognitive interviews with two doctors (subgroups: SB and MS) to ensure correct understanding of the questions and to improve the instruction procedure and structure of the survey. The feedback was used to create the final version of the questionnaire and instruction procedure. All changes were approved by the multidisciplinary team.

Table 1: Dimensions of the interpretability evaluation of GA^{2}Ms. Values in brackets indicate the number of participants or questions from the respective group.

Participants	Risk Function	Function Selection	Questions
Scientific background (4) Medical specialist (4) Medical resident (4)	1D: mean 1D: sd 2D: mean x mean 2D: mean x sd	Follows medical knowledge Against medical knowledge	Determine risk for given value Determine values for given risk Important properties y/n (3-5) Include risk function y/n Confidence for interpretability What supports/hinders interpretability

The final GA^{2}M for the evaluation contained 34 one-dimensional and 34 two-dimensional risk functions. To keep the questionnaire at a reasonable size, we only used a selection of eight risk functions for the evaluation. Risk functions were chosen from four categories to evaluate the effect of different function complexities: one-dimensional for mean (1D: mean), one-dimensional for standard deviation (1D: sd), two-dimensional for mean vs mean (2D: mean x mean), and two-dimensional for mean vs standard deviation (2D: mean x sd). For each function type, one function that followed medical knowledge and one that did not were selected in order to study the effect of risk functions that disagree with the experience and knowledge of doctors. Hence, the evaluation included eight risk functions. The risk function selection was performed by the same multidisciplinary team mentioned above to incorporate different perspectives. The team discussed every risk function and picked a representative function for each type that contained relevant medical information. When several candidates were identified, functions that received a higher importance during GA^{2}M training were favored. The survey was paper based. During the evaluation, participants had access to an interactive version of the risk functions in the aforementioned web application.

2.4.1. Questions

For each risk function, there were seven questions, including three to five dichotomous statements covering important properties of a function (indicated by (3-5) in Table 1). Figure 1 shows an excerpt from the final questionnaire for a single risk function. The first part (questions one to three) considers the objective level of interpretability, question four simulates the clinical validation of the model, and the last part (questions five to seven) covers subjective factors of interpretability.

First, participants were asked to read the risk for a given input value (one-dimensional) or pair of input values (two-dimensional) from the risk function. The second task was to find an input value or pair of input values that result in a certain risk score; hence, the participants had to read the inverse of the function. The input values and risk values for these questions were picked randomly. Risk function selection and testing revealed that certain bins were too small to be visualized. Hence, only bins visible in the evaluation environment were included. We considered these questions as relatively simple, so their significance for the evaluation was low. The main motivation for these tasks was to familiarize the participants with a risk function. Contrary to the study protocol, we only included one question for each type because testing revealed that participants get bored otherwise.

Respiratory rate (mean over 48h) [breaths per minute]

1. A patient has a respiratory rate (mean over 48h) of 37.93 breaths per minute. What is the associated risk of in-hospital mortality? ☐

2. Which respiratory rate (mean over 48h) results in a risk of in-hospital mortality of -0.154 (log10-odds)? ☐

3. Which of the following statements about the given risk function is correct?

a)	A respiratory rate (mean over 48h) above 15 breaths per minute is associated with an increasing risk of in-hospital mortality.	☐ Yes ☐ No
b)	A respiratory rate (mean over 48h) around 20 breaths per minute is associated with the lowest risk of in-hospital mortality.	☐ Yes ☐ No
c)	A respiratory rate (mean over 48h) of 35 breaths per minute is associated with a lower risk of in-hospital mortality than 30 breaths per minute.	☐ Yes ☐ No
d)	An unknown value of respiratory rate is associated with the lowest risk of in-hospital mortality.	☐ Yes ☐ No

4. Would you include this risk function for in-hospital mortality prediction? ☐ Yes ☐ No

5. How certain are you that you can correctly understand and interpret the given risk function?

☐ Very Uncertain	☐ Uncertain	☐ Slightly Uncertain	☐ Neutral	☐ Slightly Certain	☐ Certain	☐ Very Certain

6. Which factors support interpretability?

7. Which factors hinder interpretability?

Figure 1: Excerpt from the evaluation questionnaire including all questions for the one-dimensional risk function of the variable *Respiratory rate (mean over 48h)*.

Next, there was a variable number of statements that covered important properties of each function. We decided to use dichotomous questions because this required only a single statement about each function, while a multiple choice question would have required five possibly correct statements. The same multidisciplinary team selecting risk functions determined properties relevant for correctly interpreting a given risk function and generated statements accordingly. The goal was to verify whether the participants understood all relevant aspects of a function. The team tried to generate approximately the same number of statements for both answers (yes/no). The order was randomized afterwards.

Question four simulated the clinical validation where participants must decide to include or exclude the risk function for predictions. We reckoned that this question offered much room for interpretation. Hence, we expected low inter-rater reliability for this quantity. Nevertheless, we thought this question was necessary to simulate the clinical validation that requires exactly this decision. Lastly, participants were asked for their confidence in correctly interpreting the given risk function and to give written feedback about factors that facilitated or hindered interpretability (they could write in German). These two questions were meant to determine the subjective level of interpretability. In contrast to the study protocol, we added a commentary field for each function because testing showed that this increased their usage. The final questionnaire is given in Appendix A.

2.5. Statistical Analysis

To interpret the evaluation, four statistical quantities were considered: first, the ratio of correct answers for the first two questions as a measure of the ability to read risk scores and input values from the risk functions. Second, the ratio of correct answers to dichotomous statements about important properties. Third, the median of the confidence of the participants to correctly understand a risk function (fifth question). For these three quantities, we present descriptive statistics for all participants and for each subgroup separately. Moreover, to show the effect of different types of risk functions and to contrast functions that follow medical knowledge or are against medical knowledge, we generated plots that visualize the quantities across these two dimensions. Fourth, the inter-rater reliability (Krippendorff's alpha) for inclusion and exclusion of risk functions across all participants and for each subgroup (Zapf et al., 2016). Moreover, similar notes about factors that facilitated or hindered interpretability were clustered and sorted according to their frequency. The study protocol contains an R script that we used for statistical analysis with slight modifications to account for changes of the questionnaire.

3. Results

3.1. Training Generalized Additive Model with Interactions

Table 2 summarizes the main performance results for this study. The columns refer to all modifications of the data preprocessing and the usage of a separate unknown bin as discussed in Section 2.1 and 2.2: imputation with a constant value (column Impute), disabling z-score normalization (column Normalize), removal of implausible measurements (column Filter), and enforcing a separate unknown bin during GA^{2}M training (column Unk. Bin). The first two entries are the logistic regression baseline (Logistic Regr.) and the best

Table 2: Performance results for *in-hospital mortality* prediction. First two entries are from Harutyunyan et al. (2019). The third row shows the GA^{2}M performance on the same features. The following experiments use the 34 features of the interpretability evaluation and GA^{2}Ms are trained with a modified data preprocessing (for details see Table 3). The final model is restricted to 34 two-dimensional functions (bold).

Model	# Features	Impute	Normalize	Filter	Unk. Bin	AUC-ROC	AUC-PR
Logistic Regr.	714	mean	z-score	no	-	0.848 (0.828, 0.868)	0.474 (0.419, 0.529)
MC LSTM	-	-	-	no	-	0.870 (0.852, 0.887)	0.533 (0.480, 0.584)
GA^{2}M	714	mean	z-score	no	no	0.872 (0.853, 0.889)	0.533 (0.478, 0.586)
Logistic Regr.	34 (mean, sd)	mean	z-score	no	-	0.794 (0.770, 0.817)	0.347 (0.301, 0.400)
GA^{2}M	34 (mean, sd)	mean	z-score	no	no	0.851 (0.832, 0.870)	0.465 (0.412, 0.521)
GA^{2}M	34 (mean, sd)	-1	no	yes	yes	0.852 (0.833, 0.871)	0.468 (0.413, 0.525)
GA^{2}M 34 2D	34 (mean, sd)	-1	no	yes	yes	0.850 (0.830, 0.868)	0.456 (0.402, 0.511)

performing model (multitask channel-wise long short-term memory neural network [MC LSTM]) reported by Harutyunyan et al. (2019). To compare these models with a GA^{2}M, we carried out an experiment with the original features and data preprocessing (row three). Note, however, that we only used ten training iterations for two-dimensional risk function due to high computational complexity. The first iteration already gave the lowest validation score, so it seems only little performance can be gained from the interaction terms. We can observe that the performance of a GA^{2}M on the full set of features is on par with the best neural network model from Harutyunyan et al. (2019). The following experiments were performed with the reduced set of 34 features consisting of the mean and standard deviation of input variables over the first 48 hours of a stay. The logistic regression model from Harutyunyan et al. (2019) and the GA^{2}Ms show a large drop in performance, however GA^{2}Ms still perform similar to the logistic regression model on the full set of features. The modified data processing has only a slight effect on the performance (row six). We also carried out experiments for all modifications of the original data preprocessing separately to study their effect (see Table 3 in Appendix B). None of the modifications changed the performance considerably. Selecting the most important two-dimensional risk functions for the final model used in the interpretability evaluation shows only a small drop of the AUC-PR score (row seven). The base risk of this model is -2.48, which corresponds to a probability of 7.71%. All risk functions of the final model are given in Appendix C. Figure 2 contains the functions selected for the evaluation.

3.2. Questionnaire Development

We share our experiences from questionnaire development to justify our function selection and to highlight possible problems of GA^{2}M interpretability. Using standard deviation features is not meaningful for constant variables, such as height. Moreover, even though a standard deviation feature can be determined for ordinal variables (e.g. Glasgow Coma Scale), it is mathematically undefined and, hence, leads to confusion for end users. We excluded all risk functions that contained a standard deviation feature of a nominal value from risk function selection. Feature binning performed during GA^{2}M training aims to create bins with approximately the same number of instances. Hence, in dense regions of a

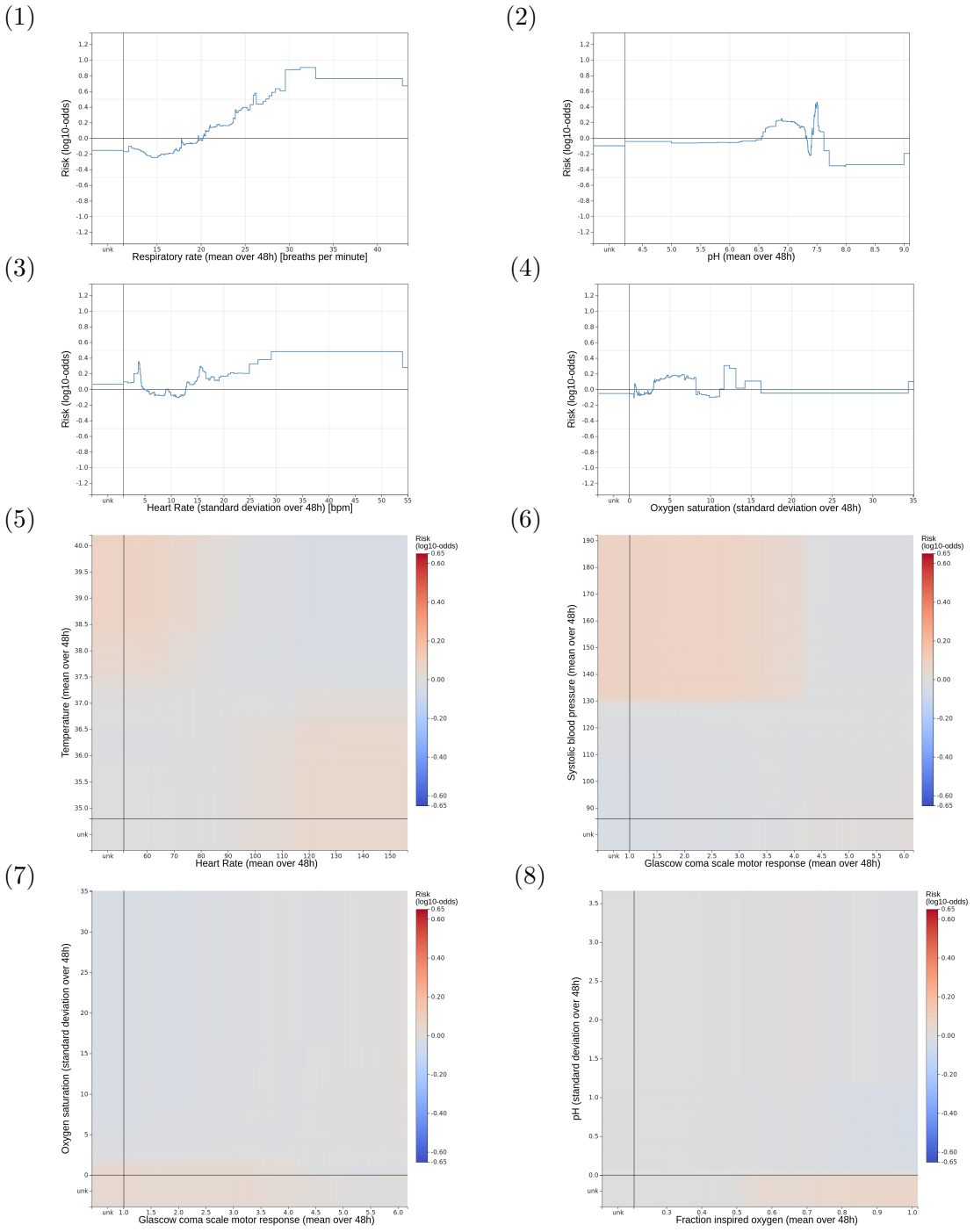

Figure 2: Eight risk functions selected for the interpretability evaluation from total of 68 functions of the final GA^{2}M (last entry in Table 2). Order according to Table 1.

feature very small bins were created that could not be visualized. This occurred for many standard deviation features that were often close to zero. Conversely, in sparse regions, large bins were created that proved problematic for outliers that received nonintuitive risk values. Risk functions learned by a GA^{2}M are not continuous, which can cause very complex function shapes. In our case, some one-dimensional risk functions especially exhibited spikes that might have decreased interpretability. Our impression from the selection procedure is that two-dimensional functions are often more difficult to interpret. Many two-dimensional functions only showed a weak signal and the importance was very low, with values between 0.0182% and 0.0078%. Note, however, that we aligned the risk axes for one- and two-dimensional functions, which might explain the weak signal and the importance values are with respect to all 595 risk functions learned during training. It was problematic to make out maxima or minima by sight and to determine a clear function behavior. Moreover, it proved difficult to decide whether a function followed or was against medical knowledge. The team agreed to declare a function as against medical knowledge if it showed clear behavior that contradicted medical knowledge in relevant regions of the function.

3.3. Statistical Analysis

The ratio and standard deviation of correct answers for reading risk scores and the inverse function (first and second questions) was 0.91 ± 0.02 ($n = 192$) for all participants. The subgroup performances were 0.89 ± 0.04 ($n = 64$; SB), 1.00 ± 0.00 ($n = 64$; MS), and 0.84 ± 0.05 ($n = 64$; MR). We rounded to two decimals and also accepted correct intervals. Most of the wrong answers were due to rounding issues or simple reading errors (e.g. the risk value of the neighboring bin was given). The first row of plots in Figure 3 indicates that one-dimensional risk functions with standard deviation inputs caused the biggest problems. There was no difference between risk functions following or against medical knowledge.

Yes/no statements about important properties of the risk functions (third question) were answered correctly with a ratio of 0.83 ± 0.02 ($n = 360$; SB: 0.82 ± 0.04 [$n = 120$], MS: 0.89 ± 0.03 [$n = 120$], MR: 0.78 ± 0.04 [$n = 120$]). Two questions were not answered and were considered as false. The order of the subgroup's performances are the same as for the first task, which might suggest that the ability to read a risk function is associated with interpreting the graph correctly. The second row in Figure 3 shows no large differences between one- and two-dimensional graphs and risk functions that follow or are against medical knowledge. There is a slight correspondence between subgroups and risk functions, which might indicate that certain functions were more difficult to interpret by all participants.

The median of the participants' certainty to correctly understand a risk function (fifth question) was *Certain* ($n = 96$). This subjective level of interpretability decreased with the level of specialization (SB: *Certain* [$n = 32$], MS: *Certain - Slightly Uncertain* [$n = 32$], MR: *Slightly Uncertain* [$n = 32$]). The last two plots in Figure 3 show that the certainty of medical residents for one-dimensional graphs with a mean feature was rather low. This could be due to the fact that these two graphs were presented first and the participants gained more experience and, hence, confidence in correctly interpreting a risk function. Inter-rater reliability for inclusion and exclusion of functions (fourth question) across all participants was very low $K_{alpha} = 0.01$, 95% CI −0.07-0.10 (SB: $K_{alpha} = 0.21$, 95% CI −0.21-0.52, MS: $K_{alpha} = -0.06$, 95% CI −0.22-0.08, MR: $K_{alpha} = -0.12$ 95% CI −0.29-0.02).

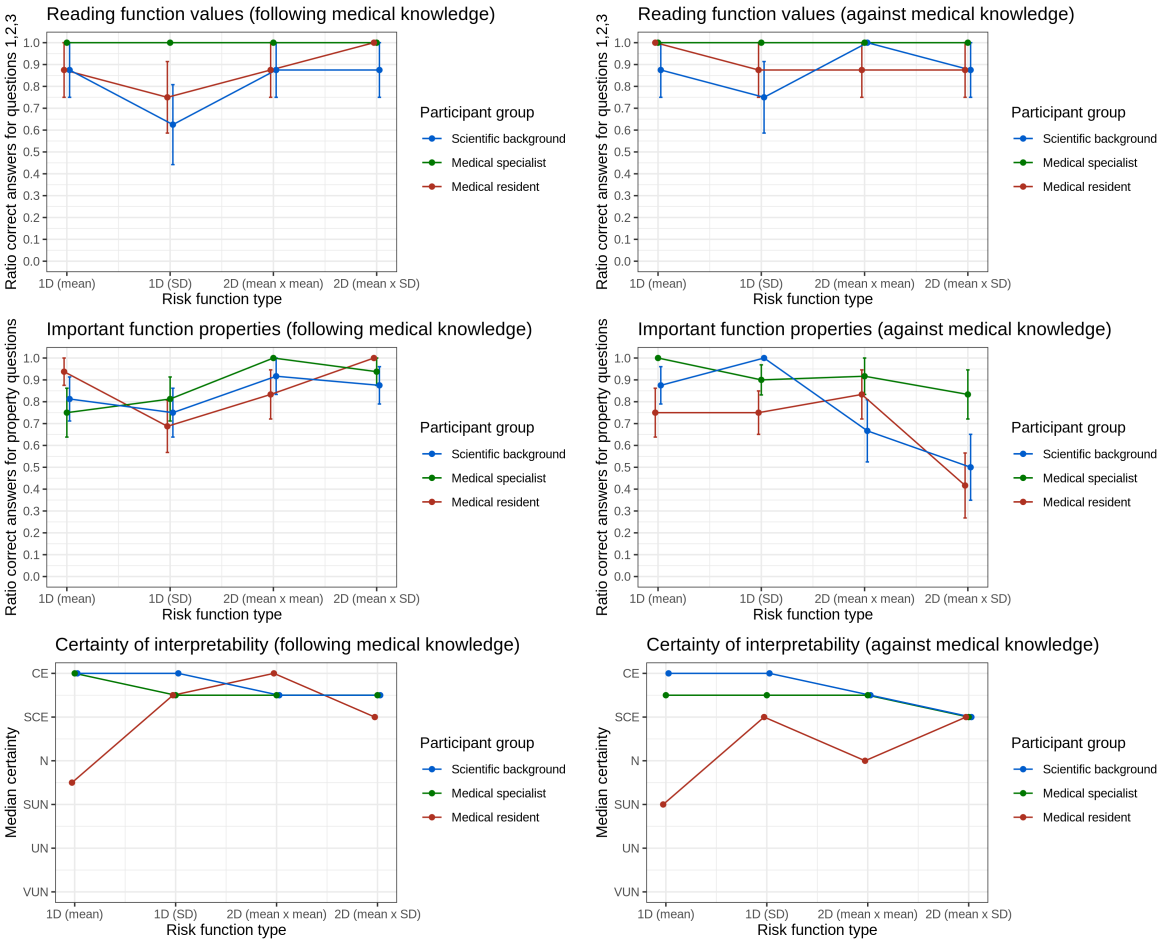

Figure 3: Ratio of correct answers for reading risk functions (first row), ratio of correct answers for dichotomous statements covering important function properties (second row), and median of participants' certainty to understand a function (third row) across different subgroups and function types. Bars show standard deviation.

All comments about factors that facilitated or hindered interpretability were, if necessary, translated to English and clustered (see Appendix D). There was one comment that was not meaningful. Positive factors that were mentioned most frequently are "Risk function visualization is easy to use and interpret" (9 times), "Linear relationships are easy to interpret" (5 times), "Color map for 2D risk functions is helpful" (5 times), and "The mouseover functionality to read the plots aids interpretability" (5 times). Comments about factors that hindered interpretability with the most occurrences were "2D risk functions show only a weak signal and no sharp borders" (7 times), "Fluctuating risks for small changes of the input variables are difficult to interpret (non-linear behavior)" (7 times), "Parts of the risk function are against medical knowledge" (4 times), "The binning for attributes is too small. It is difficult to distinguish very small bins and to find them with the cursor" (5 times), and "Nonintuitive risk function behavior for very high or very low input values" (5 times). Some participants wrote many comments, some wrote no comments at all.

4. Discussion

The key objective of this study is to evaluate whether clinicians can correctly understand and interpret risk functions of GA^{2}Ms to determine if they can validate these models for a clinical application. Our evaluation showed that reading risk scores and identifying input values that cause a certain risk were performed correctly with a very high ratio. However, in our opinion this is a very weak indicator for interpretability. Dichotomous questions about important risk function properties were answered correctly with a relatively high ratio of 0.83 ± 0.02 ($n = 360$). This shows that doctors without prior knowledge about GA^{2}Ms are able to grasp the concept of risk functions and can interpret the graphs to answer important statements about them. Moreover, doctors felt confident in understanding the functions correctly, and there were several positive comments about the interpretability of the graphs. However, a considerable number of questions were answered wrong, indicating that interpreting risk functions of GA^{2}Ms can be problematic for doctors and that there is a risk of overestimating the own ability to understand them. In addition, the self-developed statements about important function properties can only cover interpretability to a limited extent, so our conclusions must be considered with caution. Most of the positive comments mention simple function surfaces and linear relationships that are easy to grasp. Many negative notes include complaints about complex and fluctuating function surfaces, weak signals, especially of two-dimensional graphs, and bin sizes. This suggests that interpretability of GA^{2}Ms depends on the complexity of the risk functions and cannot be generalized for the whole model. As expected, inter-rater reliability for inclusion and exclusion of risk functions was very low. From our point of view, this is due to the fact that doctors demanded different levels of clinical validity to include a function. Some doctors tended to include many functions and would consider the model's output with caution, while other doctors only included graphs that reflected their medical knowledge. This shows that even though the risk functions might be interpretable, there will probably be disagreements about inclusion and exclusion of risk functions. Surprisingly, the results show no large differences between one- and two-dimensional graphs and functions that follow or are against medical knowledge (see Figure 3). We conclude that the results of this evaluation suggest that GA^{2}Ms are interpretable by doctors. However, interpretability of GA^{2}Ms depends on the complexity of the learned risk functions and, hence, on the specific task they are applied to.

We identified several factors from the clinicians' comments in the evaluation and during questionnaire development in the multidisciplinary team that could improve interpretability of GA^{2}M in a clinical setting. First, aligning the risk axis of the two-dimensional graphs led to large areas of the functions with a weak signal. Maybe it could prove beneficial to use different risk axes for these graphs. Second, the number of bins for discretizing continuous features during GA^{2}M training should be low enough to visualize all intervals properly. For our experiments, we used the default bin size of 256, which caused some very small bins that could not be displayed. Decreasing the number of bins could also help to flatten the functions' surfaces and prevent fluctuating behavior of the risk functions. Lastly, clinicians should assist during data preprocessing and feature generation to develop features that are clinically sensible and interpretable by doctors: cut-off points for valid input data are especially important when using GA^{2}Ms because outliers skew the axes of the risk functions.

Apart from the interpretability, there are further characteristics of GA^{2}Ms that, in our point of view, support an application in a clinical setting. Risk functions not only contain information about the predictions of the model but also offer insights into the input data. Following the advice of Caruana et al. (2015) we implemented histograms to visualize the input distribution (they were excluded from the evaluation for simplicity reasons). This allowed for an overview of the input data and could help to identify problems in the data preprocessing. For instance, we excluded 157 implausible cases from the benchmark task after analyzing the risk functions. Moreover, it is possible to add confidence intervals to risk functions to assess their reliability, which were also excluded from the evaluation. In addition to that, we found that visualizing a model and establishing a collaborative validation process could stimulate discussions to critically review a model. This could help building trust in healthcare professionals, which is important for a successful deployment into patient care (Wiens et al., 2019). Lastly, we think that it is a very useful property of GA^{2}Ms and transparent models in general that they can be validated once and then predictions can be used without the necessity to verify them.

However, we identified several issues related to GA^{2}Ms. First of all, we experienced a trade-off between performance and interpretability. The final model used in the interpretability evaluation contained 34 one-dimensional and 34 two-dimensional risk functions to allow validation through human practitioners in a reasonable amount of time. However, it performed much worse than a GA^{2}Ms on the full feature set (714 one-dimensional and 254,540 two-dimensional risk functions) and slightly worse than logistic regression model (714 logistic functions). However, validating those models would be more time consuming impeding interpretabilty. One possibility to alleviate this effect could be an automatic feature selection during GA^{2}M training instead of manually selecting features a priori. In addition to that, GA^{2}Ms cannot handle time-series data, which is very common in the medical domain. As a consequence, the need for data preprocessing increases and valuable structure from the input is removed. Features also have to be sensible for humans to make risk functions interpretable. For our experiments, we used the mean and standard deviation of an input variable over the first 48 hours of a stay, which constitutes a strong information reduction. Moreover, features that are correlated can lead to correlated risk functions that are very difficult to interpret because a group of functions must be interpreted together, contradicting the modularity of a GA^{2}M. We experienced this during questionnaire development with the Glasgow Coma Scale total (mean over 48h) feature and the separate Glasgow Coma Scales for eye opening, verbal response, and eye opening where each score had its own risk function, but also contributed risk with the total score.

This paper emphasizes the importance of application-grounded evaluations of interpretable ML (Doshi-Velez and Kim, 2017). While GA^{2}Ms were introduced as intelligible models (Lou et al., 2013), our study shows that interpretability is much more intricate than reading one- and two-dimensional risk functions. We had several discussions with clinicians about implementation details that would affect interpretability, especially during data preprocessing and development of the risk function visualization. For instance, which cut-off values to choose for input variables, how to integrate unknown values in the plots, and which color map to use for two-dimensional functions. In addition, our evaluation revealed many areas for improvement in the interpretability of GA^{2}Ms.

Our work has limitations. We performed an application-grounded evaluation by simulating the validation of a GA^{2}M in a clinical setting. However, we measured no performance quantity that is directly affected by this validation for instance, the clinical utility of the model or acceptance by the medical staff. Instead, we used a self-developed questionnaire to collect data on the objective and subjective level of interpretability. In addition, only twelve doctors participated in our study and our evaluation was limited to GA^{2}Ms and in-hospital mortality prediction based on MIMIC-III, so it remains open if doctors prefer different models and if our results hold for other application scenarios. Moreover, we had to rely on face validity of the survey and only performed two cognitive interviews to test and validate it. We also excluded some risk functions from the evaluations because standard deviation features were not meaningful in some cases. Lastly, the questionnaire is in English, but the evaluation was performed with German speaking participants, which could lead to misunderstandings.

Future work could evaluate interpretability of GA^{2}Ms in a real world setting instead of simulating the validation process. In addition, integration of GA^{2}Ms into the clinical workflow and the clinical utility of these models should be investigated. We reckon that a regular reiteration, for instance, in a monthly interval, of training based on new data with a subsequent model validation in a multidisciplinary team could be a useful approach. Future research could also consider possibilities to improve interpretability of GA^{2}Ms on a more technical level. This could include a refined binning mechanism preventing very small bins or constraints for risk function complexity. Lastly, a very useful extension would be the integration of time-series data and an automatic feature selection during GA^{2}M training.

5. Conclusion

Our evaluation suggests that doctors are able to interpret risk functions of a GA^{2}M for a clinical task and also feel confident in doing so. Interpretability of risk graphs depends on the complexity of the function surfaces. We identified several issues that could be improved to increase interpretability, and, in our opinion, developing a fully GA^{2}M remains a difficult task. The results underline that application-grounded evaluations are very important to reveal the quality of an interpretable ML system. We conclude that GA^{2}Ms can be useful in a clinical setting when there is not much experience with ML systems and it is necessary to build trust, there are high safety demands, and the input data is not very complex. We see GA^{2}Ms as a good candidate for a strong baseline model with a high level of control.

References

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 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. ACM, 2015.

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

Hrayr Harutyunyan, Hrant Khachatrian, David C. Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019. ISSN 2052-4463. doi: 10.1038/s41597-019-0103-9.

Alistair EW Johnson, Tom J Pollard, and Roger G Mark. Mimic-iii clinical database. *PhysioNet*, 2016a. doi: 10.13026/C2XW26. URL http://dx.doi.org/10.13026/C2XW26.

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. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016b.

Jon A Krosnick and Stanley Presser. Question and questionnaire design, second edition. In Peter V Marsden and James D Wright, editors, *Handbook of survey research*, chapter 9, pages 263–314. Emerald Group Publishing, 2010.

David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. The parable of google flu: traps in big data analysis. *Science*, 343(6176):1203–1205, 2014.

Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 623–631. ACM, 2013.

Kenneth Moreland. Diverging color maps for scientific visualization. In *International Symposium on Visual Computing*, pages 92–103. Springer, 2009.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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. ACM, 2016.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, pages 1–4, 2019.

Antonia Zapf, Stefanie Castell, Lars Morawietz, and André Karch. Measuring inter-rater reliability for nominal data–which coefficients and confidence intervals are appropriate? *BMC medical research methodology*, 16(1):93, 2016.