Communicating Performance of Regression Models Using Visualization in Pharmacovigilance

Ashley Suh* Tufts University
Gabriel Appleby† Tufts University
Erik W. Anderson‡ Data Science and AI Novartis Pharmaceuticals
Luca Finelli§ Data Science and AI Novartis Pharmaceuticals
Dylan Cashman¶ Data Science and AI Novartis Pharmaceuticals

ABSTRACT

Statistical regression methods can help pharmaceutical organizations improve the quality of their pharmacovigilance by predicting the expected quantity of adverse events during a trial. However, the use of statistical techniques also changes the risk profile of any downstream tasks, due to bias and noise in the model's predictions. That risk profile must be clearly understood, documented, and communicated across many different stakeholders in a highly regulated environment. Aggregated performance metrics such as explained variance or mean average error fail to tell the whole story, making it difficult for subject matter experts to feel confident in deciding to use a model. In this work, we describe guidelines for communicating regression model performance for models deployed in predicting adverse events. First, we describe an interview study in which both data scientists and subject matter experts within a pharmaceutical organization describe their challenges in communicating and understanding regression performance. Based on the responses in this study, we develop guidelines for which visualizations to use to communicate performance, and use a publicly available trial safety database to demonstrate their use.

Keywords: Visual Communication, Regression Models, Pharmacovigilance

1 INTRODUCTION

Advanced analytics and statistical methods have seen increasing use by pharmaceutical organizations to improve the quality of their pharmacovigilance [1, 11]. Notably, regression models have been shown to accurately predict and detect the under-reporting of adverse events (AEs) in clinical trials [16, 20, 21] – a persistent and recurrent issue raised by the FDA and GCP [33]. By identifying outliers in AE reporting through advanced regression methods, pharmaceutical companies can improve the early-detection of data collection or processing issues at trial sites, prevent delays in the required approval process, and ultimately improve patient safety [15].

Although regression methods can augment traditional pharmacovigilance approaches, a new challenge emerges: how can a diverse set of stakeholders and subject matter experts (SMEs) assess a model's risk profile in a highly regulated space, such as clinical safety? Stakeholders and SMEs often rely on model builders themselves to translate the reliability and limitations of predictive models, under the assumption that these translations will be accessible by the audience [14]. However, empirical studies suggest that even SMEs can be overwhelmed and disappointed during presentations by data scientists, particularly when metrics alone are presented as an assessment of the model's performance [14, 23, 31]. Without a more careful consideration for the communication of model performance between data scientists and SMEs, pharmacovigilance groups can fail to capitalize on the predictive power of modern machine learning and artificial intelligence techniques. It has been suggested that similar conditions limited the use of machine learning in other applications and domains, such as cybersecurity [30].

In this workshop paper, we present a design study on effective communication of regression models deployed for pharmacovigilance. First, we describe a preliminary interview study within a pharmaceutical corporation conducted with two participant groups: data scientists who build regression models, and SMEs who make decisions based on a regression model's performance and outputs. From our interviews, we outline common challenges that occur when communicating and assessing regression models, and offer guidelines for visual communication methods for a regression model's performance. Lastly, we demonstrate the use of our guidelines with a pharmacovigilance use case, illustrating the performance of three different regression models trained on a publicly available trial safety database for predicting AEs in clinical trials.

*e-mail: ashley.suh@tufts.edu
†e-mail: gabriel.appleby@tufts.edu
‡e-mail: erik.anderson@novartis.com
§e-mail: luca.finelli@novartis.com
¶e-mail: dylan.cashman@novartis.com

2 RELATED WORK

The interpretability and transparency of machine learning models is an active field being tackled at every level of industry and academia. We take a more targeted approach, focusing on the specific needs of users within a pharmaceutical corporation. However, we still review related works to identify previous examples of the use of statistical methods in pharmacovigilance. We then offer a high-level summary of relevant work in model communication to understand if any current solutions address our use case.

2.1 Predictive Models in Pharmacovigilance

Advanced analytics have been in use in pharmacovigilance for decades to investigate both the risks and benefits of medicines [1, 8, 11]. For example, Ménard et al. developed a predictive model that enables the oversight of AE reporting in clinical trials at the program, study, site, and patient levels [20, 21]. The authors describe that the deployment of these predictive models can lessen the labor-intensive load of manual investigations by pharmaceutical sponsors [33], however, the authors do not detail whether challenges occurred in the model's actual adoption by end-users. Previous studies suggest that these techniques are rarely deployed in practice at healthcare and pharmaceutical organizations, regardless of their ability to improve pharmacovigilance and QA practices [4, 7, 29].

Seneviratne et al. call to bridge the implementation gap of machine learning in healthcare by merging ML algorithms into the 'socio-technical' milieu of the organization [27]. Shah et al. suggests that the utility of ML algorithms could be better demonstrated in practice if stakeholders and healthcare patients could better assess the performance of a predictive model without relying on standard performance metrics [28]. In this work, we intentionally study how the performance of a regression model can be effectively communicated to SMEs and decision-makers, with the goal of improving the accessibility and use of predictive models in pharmacovigilance.

978-1-6654-2067-9/21/$31.00 ©2021 IEEE
DOI 10.1109/VAHC53616.2021.00006

2.2 Model Communication

Effective presentation of a predictive model's performance to domain scientists, SMEs, and other stakeholders is of ongoing study in literature. Researchers in explainable AI seek to help users interpret and explain the inferences of AI models by visualizing the internal workings of those models [6, 19, 22, 34]. Metrics and principles are posed for explainable AI [12, 25], guidelines for defining *interpretability* are suggested [5, 36], and visual analytic tools enhance machine learning and AI transparency [2, 13, 17, 26]. While much of this research is relevant to the use case in this paper, they target explainability at too low of a level – the proposed solutions are typically complex and often require training. From our interview study, we found that there was a need for better solutions at a higher level to facilitate communication between data scientists and SMEs.

Our work is closer to the user-centered approaches that interview and observe builders and consumers of AI models for improved ML workflows. For example, previous work has examined the workflow for machine learning practitioners to characterize common challenges faced by those in industry settings [14, 23]. Similarly, Suresh et al. suggest improving ML workflows by characterizing stakeholders by their personal knowledge and expertise outside of ML [31]. Our work attempts to consider both 'expert' and 'non-expert' roles, and examine the bridge of communication between them within a small scope of regression models for pharmacovigilance.

While most related work at least touches on how visualization can be used as a communication method for the interpretability of ML models, to our knowledge, no previous work aims to understand the communication gap between data scientists and SMEs who must make decisions based on a regression model's performance. To address this, we identify what data scientists who build models and SMEs that use their predictions find most valuable in the interpretation of a regression model's outcome. Our interviews with members of both groups, described in Section 3, lead us to create guidelines (Section 3.3) that can be used broadly by the community for communicating regression model performance to SMEs.

3 Interview Study

To identify visualization techniques that are most effective in communicating a regression model's performance, we conducted an interview study within a pharmaceutical company with two participant groups: data scientists who regularly build regression models, and SMEs who make decisions with regression models in their daily work. In this section, we describe our study design, interview protocol, and identify our participants' priorities when communicating and interpreting the capabilities of a regression model.

3.1 Study Design

Participants: In total, 6 data scientists and 6 subject matter experts were recruited via email. During our email exchange, potential participants were informed that the purpose of the interview was to discuss their experiences interpreting and communicating a regression model's performance. When recruiting SMEs, we specifically targeted those without direct expertise in statistics, but who have worked with or seen a regression model in the past. For data scientists, we targeted those who have developed or assessed regression models at some point in their daily work. Demographics for our participants, including their area of expertise and level of familiarity with regression models, can be seen in Table 2.

Procedure: All of our interviews were semi-structured and took 45-60 minutes to complete. Each interview was conducted virtually on Microsoft Teams with audio only. Shortly before each interview, participants were given a copy of the consent form which contained information about the study, its design, and their rights as participants. Each participant verbally consented to the study over a recording and was given an anonymous demographics survey to complete. At the start of each interview, participants were given a refresher on regression models which included: (1) a definition, (2) an example of a regression model being used at a weather station to predict daily temperature, (3) the semantic difference between regression and classification models.

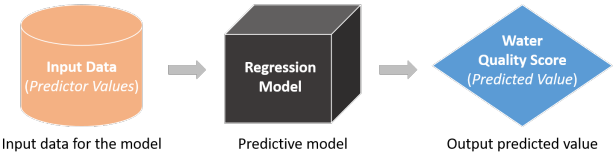


Figure 1: Visual representation provided to participants for the regression model they are tasked with assessing in the first part of our interview study (Section 3). Participants were shown 10 rows of the input data, with labeled attributes for each predictor variable.

Participants were shown the same set of prepared slides to walk through three different scenarios of assessing and communicating the performance of a regression model, one question at a time. Regardless of the participant's organizational role (data scientist or SME), the same questions were given during the interview. For time sensitivity, we did not ask all data scientists and SMEs every question that was included in our slide deck. Our slides and interview questions are included as supplementary material.

Interviews: While previous work has investigated common machine learning interpretability challenges faced by data scientists [23], machine learning practitioners [14] and stakeholders [31], our work focuses directly on how communication can be improved between data scientists and subject matter experts when the end-goal is to use a regression model in their workflow. Our interviews consider three major topics, presented as individual scenarios we stepped through with participants during the interview:

1. **What would you need to know about a regression model to recommend its use?** Participants were described a theoretical scenario in which a new regression model was being presented at their workplace. In our slides, we illustrated the regression model as one that predicts a water potability score based on a set of input predictor values (e.g., pH, sulfate). A subset of the input data was shown to participants, and the output of the model displayed a predicted water potability score as a numerical value. The visual representation of the scenario shown to participants is available in Figure 1. Participants described what they would need to know about the model to recommend and trust its use.
2. **How have you assessed and communicated a regression model at work previously?** We asked participants to think of a time when a regression model was introduced to their daily work. In our slides, we presented examples of regression models commonly deployed for pharmacovigilance. Once the participant had a particular model in mind, they described how its performance was assessed, communicated, and scrutinized.
3. **How could communicating a regression model's performance with *data scientists* and *subject matter experts* be improved?** We asked participants to describe what they would like the opposite role to better communicate or give feedback on when assessing the performance of a regression model together. In other words, SMEs were asked what they would like data scientists to communicate to them regarding a model's performance, and data scientists were asked what they would like SMEs to communicate to them regarding a model's performance.

For the first scenario of our interview study, we asked participants how they would assess a regression model that they had never seen before. Our goal was to engage participants in a broad discussion on factors that influence their trust and interpretation of a regression model when being presented on its performance. In the second sce-

nario, a number of data scientists and SMEs elaborated on projects related to modeling adverse events in clinical trials, which we consider for our use case in Section 4. For the remainder of the section, we summarize results for all three interview scenarios.

3.2 Study Results

A preliminary analysis of our interviews revealed several high-level differences and similarities between the experiences and concerns of data scientists and SMEs. In this workshop paper, we highlight the findings that were most relevant to our use case, and present guidelines based on these findings. We hope to provide further analysis of our interview results in future work.

To recommend a model's use, we found data scientists strongly valued performance metrics, the distribution and correlation of features, and overall data quality when assessing regression models. On the other hand, SMEs were primarily concerned with the limitations of the regression model, particularly where the model might fail. One SME detailed the risks that he must consider when using a regression model to predict under-reporting of AEs in clinical trials: "*Under-reporting is very critical for clinical trials because it's safety information... which needs to be notified from the scientists, to the teams, to the company... There can be a risk to the patient's safety, [if] we have not addressed a safety issue which has come up on time, that may have an impact on changing the safety profile of the drug. So, it [the model's performance] has quite an impact.*"

Another SME working in pharmacovigilance was strongly concerned about the reliability of the model's outputs, particularly when it affects patient safety. She was the only interviewee of ours to adamantly question how the predicted water potability score, described in the first scenario of our interview, would be used in the real world. In particular, she noted that she would need to exhaustively know the limitations and risks of the model before recommending its use: "*If we're telling people that water is potable, that means they're going to use it for washing, cooking, cleaning, drinking... there are consequences to that decision. You know, there is a human being at the end of it. That's the reason why I'd want as much information about the model as possible. So if I'm going to make a decision about applying this in the real world, then at least I know exactly what its limitations are before making a decision on something.*"

During the second scenario of our interview, all participants mentioned charts, graphs, or interfaces that would be or have been useful to them in assessing a model's performance. Overwhelmingly, SMEs requested the ability to see how certain inputs (or features) affected the output of a model to test out its performance. This offers evidence that there is an appetite for the interactive explainable systems that are seen in literature, e.g., the What-if tool [35] for black box models, and similar tools for classification models [24, 37].

When discussing how communication could be improved, data scientists told us they expected to spend time explaining the technical details of their models' performance to SMEs. They did not expect SMEs to already know regression metrics by name, noting that slides are prepared in advance to cover questions about quantifying performance. However, SMEs stated they did not always feel comfortable asking questions during presentations, often due to the pacing of the explanation: "*Data scientists show a regression curve and it's so normal for them... they don't always realize that people don't understand some of the visuals for the models and what they really mean. Sometimes it just goes over your head, and I think the end-user a good chunk of the time would be too embarrassed to say - I don't get what you're talking about*" (SME).

Similar feedback from interview participants suggests that there can be a mismatch in the interpretation of the conversation between data scientists and SMEs, where visualization could more effectively act as an explanatory bridge. Further, it suggests that commonly used charts for visualizing regression performance may not be as easily interpretable or recognizable to SMEs as data scientists perceive.

3.3 Guidelines

From the analysis of our interviews, we derive a set of guidelines for communicating the performance of a regression model to SMEs.

The first two guidelines address a lack of context and comfort identified by SMEs: "*You have to make the end-user feel comfortable both in the data scientist's language, and also that if they don't understand something they can easily ask, what is this?*" (SME).

G1: When articulating results, start slow and offer to speed up. All SMEs we interviewed suggested that data scientists could spend more time highlighting aspects of their presentation that could be considered "obvious", in order to establish a common baseline for the language spoken and understood. For example, data scientists could define common performance metrics or potentially nuanced visual encodings before detailing their results.

G2: Tie in use cases for the model by illustrating real-life, objective-driven examples. Across all of our interviews, when asked what they would want to know about a model's performance to recommend its use, a common request made by SMEs was to understand how a regression model's performance relates to their end-goals for the model. This request is especially critical when using regression models that affect patient and/or public safety.

The next three guidelines relate to the choice of visualization style when communicating and presenting regression model performance to SMEs: "*Some people don't have experience with visualization outside of BBC infographics. I do realize it can be hard for me to remove my data scientist hat and put myself into the role of somebody who's not looking at a log plot every day*" (data scientist).

G3: Provide context for performance by annotating plots with stories. Each annotated story serves to decode the intended message of the visualization, beyond the visualized data and provided legend. By guiding the audience through sensible conclusions on a provided visualization, an SME could more quickly arrive at new conclusions with the same visualization

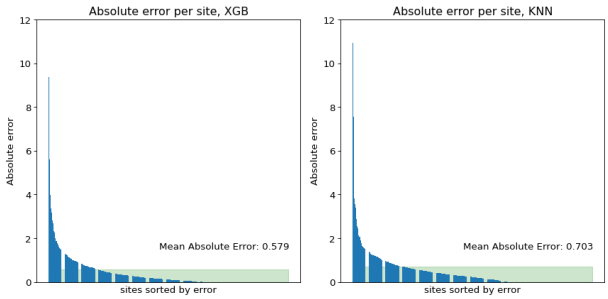
G4: For any chart that communicates a model's performance, provide a range of comparisons. SMEs found that assessing the results of a regression model's performance is easier if it is compared against their current practices, an interpretable naive baseline model, and if possible, an oracle or perfect model.

G5: Visually explain significance of global metrics. Global metrics such as explained variance or mean absolute error can seem abstract and removed from the use case. Showing metrics in visual context can help ground them; for example, visualizing the enveloping ellipse in a correlation scatterplot can give a proxy for the correlation between predicted and actual values.

The final three guidelines address concerns by both data scientists and SMEs in understanding the caveats, edge cases, outliers, and limitations of the model: "*If data scientists said, 'when you run these models, here is the area where we think you're going to have the most problems, or the most risk. And here's the explanation for why we think that's happening.'... I think upfront and transparent communication about why we should expect those issues is a very big way for us to build trust and confidence in the model*" (SME).

G6: Point to outliers in the model's performance with known or plausible explanations. The source of outliers and anomalies is often dependent on the scenario, therefore, data scientists should point SMEs to known or potential outliers, and include at least reasonable speculations behind their anomalous behavior.

G7: Be descriptive about the data used for training and testing a model, and provide examples. The distribution, weighting, correlation, and availability of the data used in the modeling process were notable concerns from both SMEs and data scientists. Many data scientists agreed that SMEs provide essential context



(a) XGB model (b) KNN model

Figure 2: Absolute error plots for two models predicting adverse event rate. To explain the mean absolute error global metric, the average error is shown as a green rectangle compared with the sorted errors of individual sites, showing that approximately 80% of sites have better than average error with the XGB model. Residual plots for the remaining models can be seen in Figure 5 of the appendix.

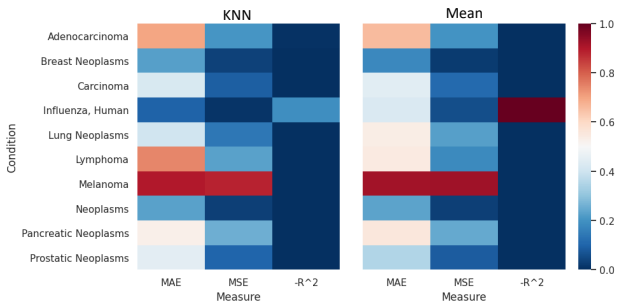


Figure 3: Heat map showing the apportionment of error of a KNN model vs. a baseline (mean), split by treatment in each row. This visualization explains whether the model is biased towards certain subgroups in the data. An annotation could be added to point out that the KNN model has a better r2 for Influenza trials than the mean, but a worse mean absolute error for Lymphoma trials. Heat maps for remaining models can be seen in Figure 4 of the appendix.

for the data domain, ultimately leading to improvements in model performance and transparent communication.

G8: Explicitly demonstrate the error, limitations, and weaknesses of the model, not just the strengths. From our interviews, SMEs want transparent information regarding the limitations of a model with both qualitative and quantitative assessments of those errors or weaknesses. Both SMEs and data scientists noted that they were able to help each other improve a model's performance once the limitation of the model was fully understood.

4 Use Case: Modeling Adverse Events

To demonstrate how our guidelines for communicating regression model performance can be used, we present a use case on modeling AEs in clinical trials. We train three different models and two baselines, then provide example visualizations that could be used in explaining the performance of these models. A write-up of our modeling process is included in our appendix, and error metrics for the regressors trained are available in Table 1. A subset of visualizations for our models are provided in the main text, however, examples for each can be found in the appendix (Figures 4, 5, 6).

4.1 Communicating Model Performance

For the remainder of this section, we provide examples of visualizations and explanations that follow the proposed guidelines in Section 3.3 for the five models described above.

Adding context and comfort: To address **G1** and **G2**, we suggest starting with an explanation of the model's basic functions. For example, if a KNN model was used, it could be explained that the model looks at the rates of similar historical trials based on multiple aspects: the program, patient, site, and study phases. Walking through a single example of inference, or a simplified illustration, can establish a level of comfort with SMEs and better tie the model to the use case at hand. For context on the data used in the modeling process (**G7**), a description of the data can be provided, as in Table 3.

Provide annotated visualizations explaining global metrics: For the use case of predicting the rate of AEs, we propose using a bar chart showing the residuals between predicted and actual, sorted by size of the residual. The shape of this plot shows how error is apportioned globally (**G5**). Explanatory annotations (**G3**) can show the mean absolute error and a comparison against a baseline (**G4**) can show the different shapes of error, as seen in Figure 2. A containing ellipse can also be shown to provide a visual of the r2 of the regression model (e.g., Figure 6) [10].

Caveats, edge cases, outliers, and limitations:. Examples of the input data should be provided, with special focus given to trials where the model has greatest error **(G6)**. SMEs also want to know if there are segments of the data that the model has high error on; we recommend looking at the most important features of the model and using a heat map to show which categories have high error (**G7**), as seen in Figure 3. This heat map can show multiple error metrics across all categories. However, different error metrics have different scales, so some normalization must be applied to make outliers visually salient. The most important limitations of a model can be communicated textually or visually (**G8**). For example, if a KNN model is used, it should be explained that they are highly sensitive to noisy or junk data, while it can be communicated that an XGB model might be slower to train.

5 Conclusion and Future Work

In this work, we present guidelines for communicating regression model performance within a pharmaceutical organization. Based on interviews with both data scientists and subject matter experts, we identify common gaps in communication and suggest broadly applicable solutions for data scientists to use in communicating their results to SMEs. Lastly, we demonstrate how our guidelines could be used in practice by illustrating a pharmacovigilance use case.

We hope to have future work in several directions. First, we would like to quantitatively analyze our interview data to better understand mismatches in language between data scientists and SMEs. Characterizing these gaps could lead to more pointed recommendations about common language to use, or a visual language to facilitate translation. We would also like to evaluate commonly used regression visualizations, including those suggested in this workshop paper, to evaluate if SMEs indeed find them helpful when interpreting a model's performance. Finally, empirical studies can comparatively analyze the efficacy of our suggested guidelines.

Acknowledgments

We thank our collaborators at Novartis for their time and participation in our study, as well as the reviewers for their helpful feedback.

References

[1] J. S. Almenoff. Innovations for the future of pharmacovigilance. *Drug safety*, 30(7):631–633, 2007.

[2] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. The state of the art in enhancing trust in machine learning models with the use of visualizations. In *Computer Graphics Forum*, vol. 39, pp. 713–756. Wiley Online Library, 2020.

[3] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794. ACM, New York, NY, USA, 2016. doi: 10.1145/2939672.2939785

[4] K. Cresswell and A. Sheikh. Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. *International journal of medical informatics*, 82(5):e73–e86, 2013.

[5] B. Davis, M. Glenski, W. Sealy, and D. Arendt. Measure utility, gain trust: Practical advice for xai researchers. In *2020 IEEE Workshop on TRust and EXpertise in Visual Analytics (TREX)*, pp. 1–8. IEEE, 2020.

[6] F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[7] E. J. Emanuel and R. M. Wachter. Artificial intelligence in health care: will the value match the hype? *Jama*, 321(23):2281–2282, 2019.

[8] S. J. Evans. Pharmacovigilance: a science or fielding emergencies? *Statistics in medicine*, 19(23):3199–3209, 2000.

[9] E. Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *International Statistical Review / Revue Internationale de Statistique*, 57(3):238–247, 2021/07/09/ 1989. Full publication date: Dec., 1989. doi: 10.2307/1403797

[10] M. Friendly, G. Monette, and J. Fox. Elliptical insights: understanding statistical methods through elliptical geometry. *Statistical Science*, 28(1):1–39, 2013.

[11] M. Hauben and X. Zhou. Quantitative methods in pharmacovigilance. *Drug safety*, 26(3):159–186, 2003.

[12] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*, 2018.

[13] F. Hohman, A. Srinivasan, and S. M. Drucker. Telegam: Combining visualization and verbalization for interpretable machine learning. In *2019 IEEE Visualization Conference (VIS)*, pp. 151–155. IEEE, 2019.

[14] S. R. Hong, J. Hullman, and E. Bertini. Human factors in model interpretability: Industry practices, challenges, and needs. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–26, 2020.

[15] ICH. E26(r2) guideline for good clinical practices. *International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use.*, 2:1–60, 2016.

[16] B. Koneswarakantha, T. Ménard, D. Rolo, Y. Barmaz, and R. Bowling. Harnessing the power of quality assurance data: can we use statistical modeling for quality risk assessment of clinical trials? *Therapeutic innovation & regulatory science*, 54(5):1227–1235, 2020.

[17] J. Krause, A. Dasgupta, J. Swartz, Y. Aphinyanaphongs, and E. Bertini. A workflow for visual diagnostics of binary classifiers using instance-level explanations. In *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pp. 162–172. IEEE, 2017.

[18] A. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. Nineteenth Century Collections Online (NCCO): Science, Technology, and Medicine: 1780-1925. F. Didot, 1805.

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

[20] T. Ménard, Y. Barmaz, B. Koneswarakantha, R. Bowling, and L. Popko. Enabling data-driven clinical quality assurance: predicting adverse event reporting in clinical trials using machine learning. *Drug safety*, 42(9):1045–1053, 2019.

[21] T. Ménard, B. Koneswarakantha, D. Rolo, Y. Barmaz, L. Popko, and R. Bowling. Follow-up on the use of machine learning in clinical quality assurance: can we detect adverse event under-reporting in oncology trials? *Drug safety*, 43(3):295–296, 2020.

[22] S. Mohseni, N. Zarei, and E. D. Ragan. A survey of evaluation methods and measures for interpretable machine learning. *arXiv preprint arXiv:1811.11839*, 1, 2018.

[23] S. Passi and S. J. Jackson. Trust in data science: Collaboration, translation, and accountability in corporate data science projects. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–28, 2018.

[24] D. Ren, S. Amershi, B. Lee, J. Suh, and J. D. Williams. Squares: Supporting interactive performance analysis for multiclass classifiers. *IEEE transactions on visualization and computer graphics*, 23(1):61–70, 2016.

[25] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251*, 2021.

[26] D. Sacha, M. Kraus, D. A. Keim, and M. Chen. Vis4ml: An ontology for visual analytics assisted machine learning. *IEEE transactions on visualization and computer graphics*, 25(1):385–395, 2018.

[27] M. G. Seneviratne, N. H. Shah, and L. Chu. Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations*, 6(2), 2020.

[28] N. H. Shah, A. Milstein, and S. C. Bagley. Making machine learning models clinically useful. *Jama*, 322(14):1351–1352, 2019.

[29] S. Shilo, H. Rossman, and E. Segal. Axes of a revolution: challenges and promises of big data in healthcare. *Nature medicine*, 26(1):29–38, 2020.

[30] R. Sommer and V. Paxson. Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE symposium on security and privacy*, pp. 305–316. IEEE, 2010.

[31] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021. doi: 10.1145/3411764.3445088

[32] A. Tasneem, L. Aberle, H. Ananth, S. Chakraborty, K. Chiswell, B. J. McCourt, and R. Pietrobon. The database for aggregate analysis of clinicaltrials.gov (aact) and subsequent regrouping by clinical specialty. *PloS one*, 7(3):e33677–e33677, 2012. 22438982[pmid]. doi: 10.1371/journal.pone.0033677

[33] F. R. Varallo, S. d. O. P. Guimarães, S. A. R. Abjaude, and P. d. C. Mastroianni. Causes for the underreporting of adverse drug events by health professionals: a systematic review. *Revista da Escola de Enfermagem da USP*, 48(4):739–747, 2014.

[34] H. J. Weerts, W. van Ipenburg, and M. Pechenizkiy. A human-grounded evaluation of shap for alert processing. *arXiv preprint arXiv:1907.03324*, 2019.

[35] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

[36] F. Yang, Z. Huang, J. Scholtz, and D. L. Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pp. 189–201, 2020.

[37] J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics*, 25(1):364–373, 2018.

A Appendix

A.1 Modeling Process

The data used is a small subset of the ClinicalTrials.gov data accessed through the AACT database [32], which contains a large quantity of basic trial summary and results information. We use this information to predict the number of AEs per enrolled person.

The dataset was filtered by requiring all values to be filled, in addition to the study being a completed interventional study that lasted one or more years. We also filtered the dataset by removing any studies that did not have a MeSH[1] term for the condition, or an intervention in the 40 most popular within the current subset. Finally, all categorical features were transformed into a one-hot encoding. The resulting dataset totals 2572 instances with 96 derived features, 88 of which are one-hot encodings of categorical features.

Three basic regression models were chosen: Linear Regression [18] (OLS), K-Nearest Neighbors Regression [9] (KNN), and Gradient Boosting Tree Regression [3] (XGB). Two 'dummy' regression models were included as interpretable baselines: one that always predicts the mean, and one that always predicts the median.

Some hyperparameter tuning was done for two of the regression models, KNN and XGB. We split the dataset into a train and test set of 75% and 25%. The training set was then used to perform three fold cross validation grid search to find the best hyperparameters based on mean squared error. For the KNN we looked for the number of nearest neighbors, and for XGB we examined different max depths. We found the best value for the nearest neighbors to be 25, and the best value for max depth to be 3.

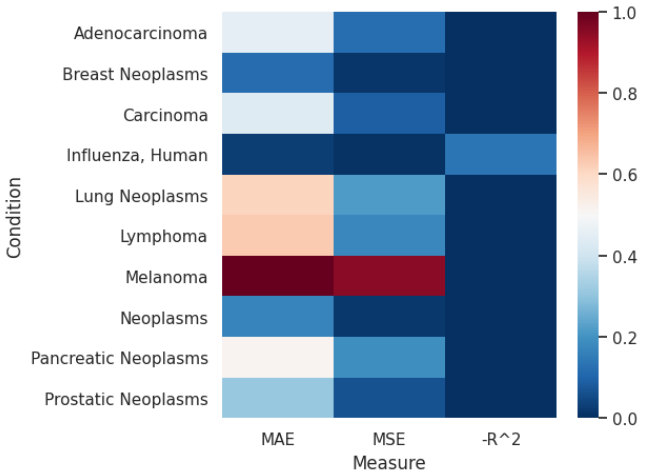
Each of the regression models were then trained on the full training set, using the best hyperparameter values found with the grid search where appropriate. Finally, the mean squared error (MSE), mean absolute error (MAE), and coefficient of determination (R^2) on the test set was recorded. Error metrics for all regressors trained are available in Table 1.

A.2 Additional Tables and Figures

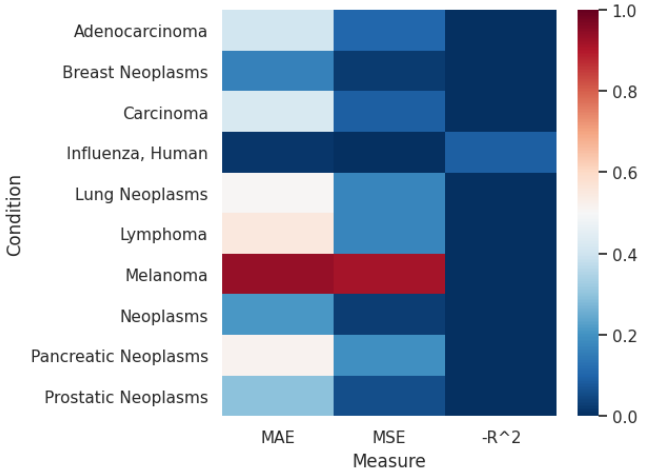
Regressor	MAE	MSE	R^2
OLS	0.564	1.064	0.283
XGB	0.579	1.090	0.266
KNN	0.703	1.364	0.082
Mean	0.788	1.486	-0.001
Median	0.736	1.621	-0.091

Table 1: Error metrics across the three regressors, and two heuristic methods used as baselines. In order of lowest mean squared error, regressors were Linear Regression (OLS), Gradient Boosting Trees (XGB), and K-Nearest Neighbors Regression (KNN). The two baseline methods were Mean and Median, which predicted the Mean and Median of the training set respectively.

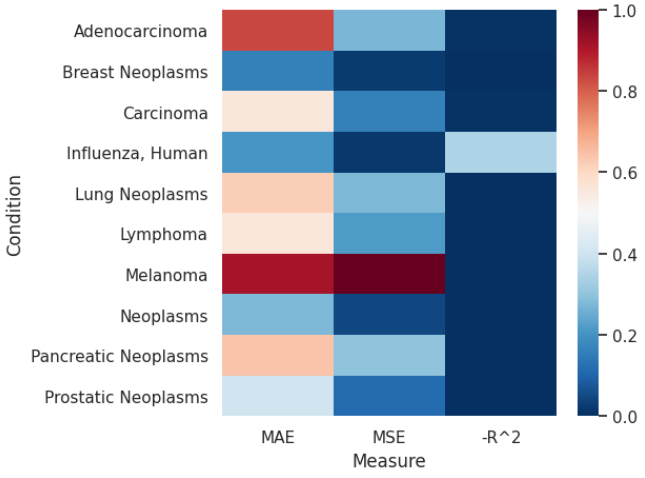
[1] https://www.ncbi.nlm.nih.gov/mesh/



(a) XGB model



(b) OLS model



(c) Median (baseline)

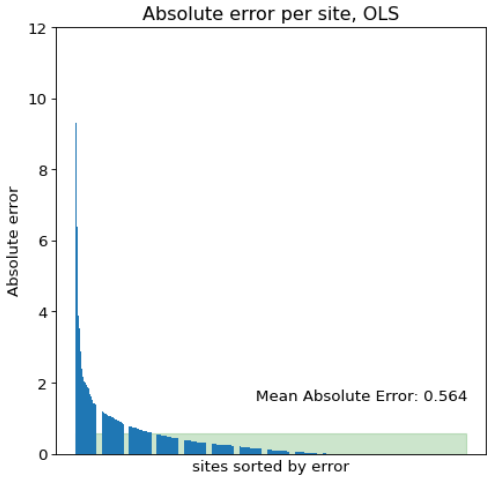
Figure 4: Heat maps for the remaining three models described in Section 4.

Measure	Count
N	12
Age	18-29: 1, 30-39: 5, 40-49: 5, 50-59: 1, 60-69: 0
Gender	Female: 5, Male: 7, Non-binary: 0
Education	Associates: 0, Bachelors: 1, Masters: 7, Doctorate: 4
Role	Data scientist: 6, Subject matter expert: 6
Data science experience	No experience: 0, Somewhat familiar: 1, Familiar: 5, Very familiar: 3, Expert: 5
Frequency using data tools	Never: 0, 1-3x/month: 0, 1-3x/week: 4, 1-3x/day: 2, All day: 6
Frequency using regression	Never: 1, 1-3x/month: 4, 1-3x/week: 1, 1-3x/day: 5, All day: 1
Expertise (SMEs only):	Finance 1, Commercial: 1, Pharmacovigilance: 2, Quality Assurance: 2

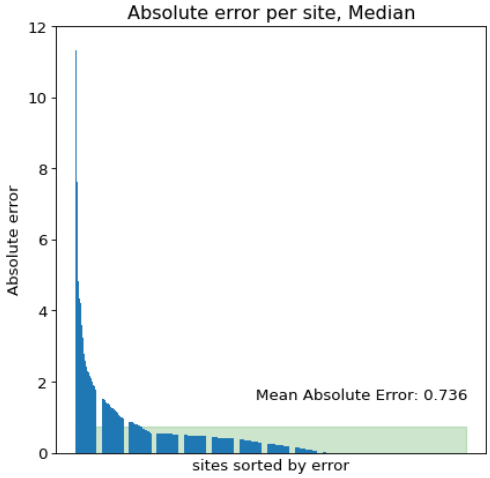
Table 2: Demographics table for the interview study described in Section 3.

Feature	Description
Phase	For a clinical trial of a drug product (including a biological product), the numerical phase of such clinical trial, consistent with terminology in 21 CFR 312.21 and in 21 CFR 312.85 for phase 4 studies.
Enrollment	The estimated total number of participants to be enrolled (target number) or the actual total number of participants that are enrolled.
Number of Arms	The number of arms in the clinical trial. For a trial with multiple periods or phases that have different numbers of arms, the maximum number of arms during all periods or phases.
Has Expanded Access	Whether there is expanded access to the investigational product for patients who do not qualify for enrollment in a clinical trial. Expanded Access for investigational drug products (including biological products) includes all expanded access types under section 561 of the Federal Food, Drug, and Cosmetic Act: (1) for individual participants, including emergency use; (2) for intermediate-size participant populations; and (3) under a treatment IND or treatment protocol.
Number of Facilities	The number of participating facility in a clinical study.
Actual Duration	Number of months between the start date and primary completion date. Start date: the estimated date on which the clinical study will be open for recruitment of participants, or the actual date on which the first participant was enrolled. Primary completion date: the date that the final participant was examined or received an intervention for the purposes of final collection of data for the primary outcome, whether the clinical study concluded according to the pre-specified protocol or was terminated. In the case of clinical studies with more than one primary outcome measure with different completion dates, this term refers to the date on which data collection is completed for all of the primary outcomes.
Months to Report Results	Number of months between primary completion date and first received results date.
Minimum Age	The numerical value, if any, for the min. age a potential participant must meet to be eligible for the clinical study. (Years only for us)
Number of Primary Outcomes	"Primary outcome measure" means the outcome measure(s) of greatest importance specified in the protocol, usually the one(s) used in the power calculation. Most clinical studies have one primary outcome measure, but a clinical study may have more than one.
Number of Secondary Outcomes	"Secondary outcome measure" means an outcome measure that is of lesser importance than a primary outcome measure, but is part of a pre-specified analysis plan for evaluating the effects of the intervention(s) or interventions under investigation in a clinical study. , and is not specified as an exploratory or other measure. A clinical study may have more than one secondary outcome measure.
Condition Mesh Term	Condition MeSH terms generated by NLM algorithm
Intervention Mesh Term	Intervention MeSH terms generated by NLM algorithm

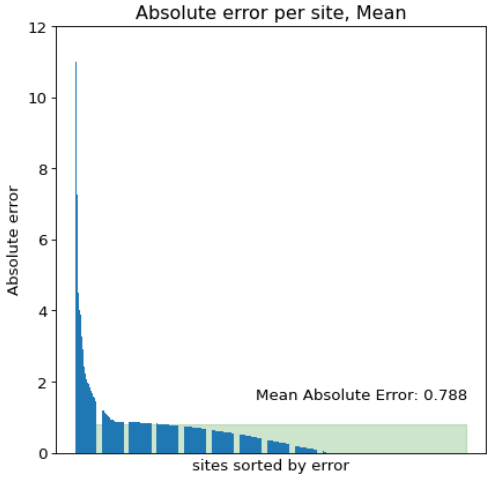
Table 3: Descriptions of the data used for training in Section 4.



(a) OLS model

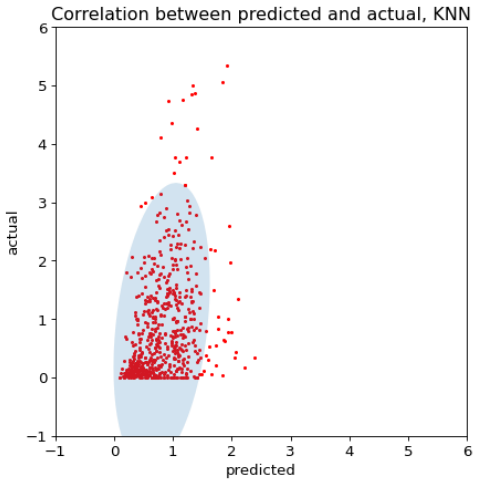


(b) Median (baseline)

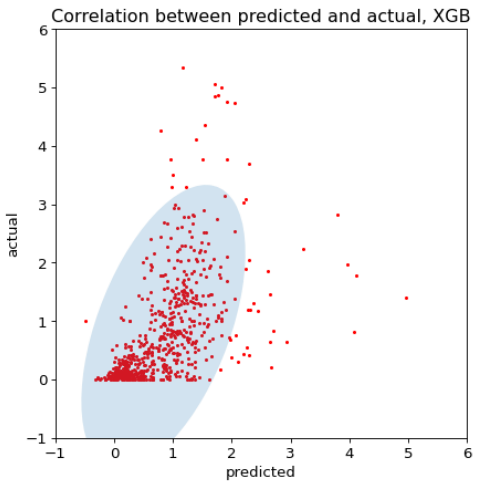


(c) Mean (baseline)

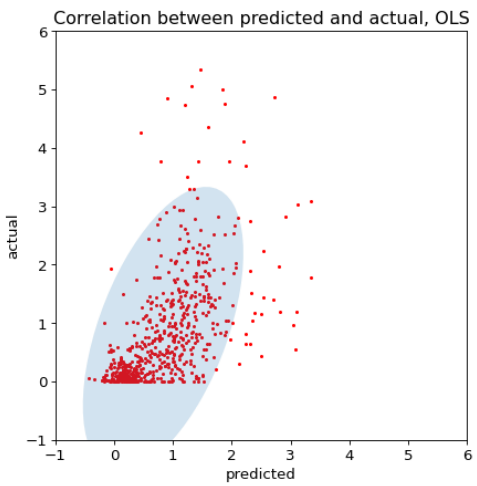
Figure 5: Absolute error plots for the three remaining models described in Section 4.



(a) KNN model



(b) XGB model



(c) OLS model

Figure 6: Correlation scatter plots for three of the models described in Section 4.