# The grammar of interactive explanatory model analysis*

Workshop on Trust and Reliance in AI-Assisted Tasks at CHI 2023

HUBERT BANIECKI†, Warsaw University of Technology, Poland and University of Warsaw, Poland
DARIUSZ PARZYCH, Warsaw University of Technology, Poland
PRZEMYSLAW BIECEK†, Warsaw University of Technology, Poland and University of Warsaw, Poland

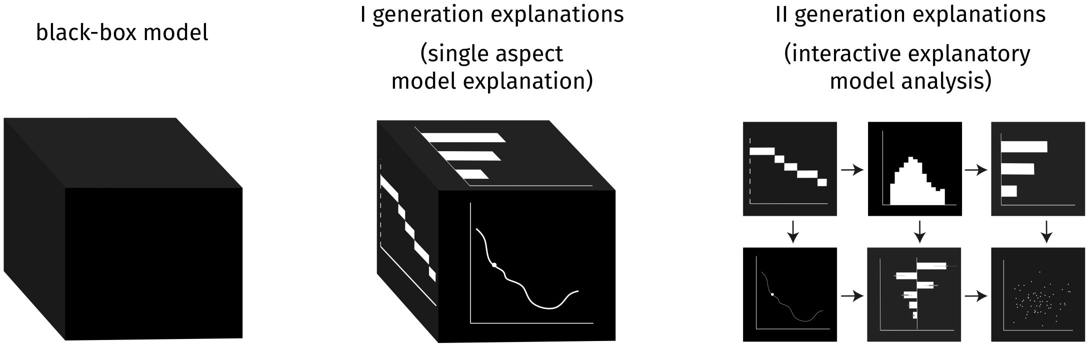


Fig. 1. Increasing computing power and the availability of automated machine learning tools resulted in complex models that are effectively black-boxes. The first generation of model explanations aims at exploring individual aspects of model behavior. The second generation of model explanation aims to integrate individual aspects into a vibrant and multi-threaded customizable story about the black-box that addresses the needs of various stakeholders. We call this process Interactive Explanatory Model Analysis (IEMA).

The growing need for in-depth analysis of predictive models leads to a series of new methods for explaining their local and global properties. Which of these methods is the best? It turns out that this is an ill-posed question. One cannot sufficiently explain a black-box machine learning model using a single method that gives only one perspective. Isolated explanations are prone to misunderstanding, leading to wrong or simplistic reasoning. This problem is known as the *Rashomon effect* and refers to diverse, even contradictory, interpretations of the same phenomenon. Surprisingly, most methods developed for explainable and responsible machine learning focus on a single-aspect of the model behavior. In contrast, we showcase the problem of explainability as an interactive and sequential analysis of a model. This paper proposes how different Explanatory Model Analysis (EMA) methods *complement* each other and discusses why it is essential to *juxtapose* them. The introduced process of Interactive EMA (IEMA) derives from the algorithmic side of explainable machine learning and aims to embrace ideas developed in cognitive sciences. We formalize the grammar of IEMA to describe human-model interaction. It is implemented in a widely used human-centered open-source software framework that adopts interactivity, customizability and automation as its main traits. We conduct a user study to evaluate the usefulness of IEMA, which indicates that an interactive sequential analysis of a model may increase the accuracy and confidence of human decision making.

CCS Concepts: • **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **User studies**.

Additional Key Words and Phrases: explainable AI, user study, decision making

[*] This is a shortened encore submission of the already published paper, which is licensed under a Creative Commons Attribution 4.0 International License http://creativecommons.org/licenses/by/4.0/; cite this article as: Hubert Baniecki, Dariusz Parzych, and Przemyslaw Biecek. **The grammar of interactive explanatory model analysis**. *Data Mining and Knowledge Discovery* (2023), 1–37. https://doi.org/10.1007/s10618-023-00924-w

[†] Correspondence to h.baniecki@uw.edu.pl and przemyslaw.biecek@pw.edu.pl.

## 1 INTRODUCTION

There are many technical discoveries in the field of explainable and interpretable machine learning (XIML) praised for their mathematical brilliance and software ingenuity [2, 3, 8, 13, 17]. However, in all this rapid development, we forgot about how important is the visual and interactive interface between human and model [7]. We agree with [12] that in practice, there are three main approaches to overcoming the opaqueness of black-box models: evading it and using algorithms interpretable by design [18], bias checking and applying mitigation techniques [11], or using post-hoc explainability methods [14]. Although the first two are precise, the last solution is of particular interest to ours in this paper. We base our contribution on the philosophies of Exploratory Data Analysis (EDA) [20], which presents tools for in-depth data analysis, Explanatory Model Analysis (EMA) [9], which presents tools for in-depth model analysis, and The Grammar of Graphics [21], which formalizes and unifies language for the visual description of data. The objective is set to bridge the research gap concerning opaque predictive models developed for *tabular data*, but the introduced concept can be generalized to other tasks, specifically in deep learning. We propose a new paradigm in XIML, which is to use a sequence of single-aspect explanations aiming to significantly extend our understanding of machine learning models (see Figure 1). Interactivity involves a sequence of operations; thus, explanatory model analysis can be seen as a cooperation between the operator and the explanatory interface. We adhere to the *Rashomon effect* [10] by juxtaposing complementary explanations, whereas conventionally it is used to denote analyzing diverging models.

*Contribution.* In [6], we formally define a language for human-model communication. The introduced grammar of Interactive Explanatory Model Analysis (IEMA) provides a multifaceted look at various possible explanations of the model's behavior. We validate its usefulness in three real-world machine learning use-cases: an approachable and illustrative example based on the FIFA-20 regression task, an external model audit based on the COVID-19 classification task, and a user study based on the Acute Kidney Injury prediction task. The paper introduces and validates a methodology for which we already implemented and contributed an open-source software framework [4], as well as prototyped its applicability [5]. This paper describes the results from a user study aiming to evaluate IEMA. We acknowledge work related to evaluating interactive explanations in user studies in Section 2.3 in [6].

## 2 EVALUATION WITH HUMAN SUBJECTS: A USER STUDY (SECTION 5 IN [6])

We conduct a user study on 30 human subjects to evaluate the usefulness and need for IEMA in a real-world setting. The goal is to assess if an interactive and sequential analysis of a model brings value to explaining black-box machine learning. In that, we aim to answer the main hypothesis of *"Juxtaposing complementary explanations increases the usefulness of explanations."* The *usefulness* can be measured in varied ways; in this case, we aim to check if juxtaposing complementary explanations *increases*:

- $H_1$: human *accuracy* in understanding the model,
- $H_2$: human *confidence* in understanding the model.

The latter can alternatively be viewed as increasing *trust* in machine learning models.

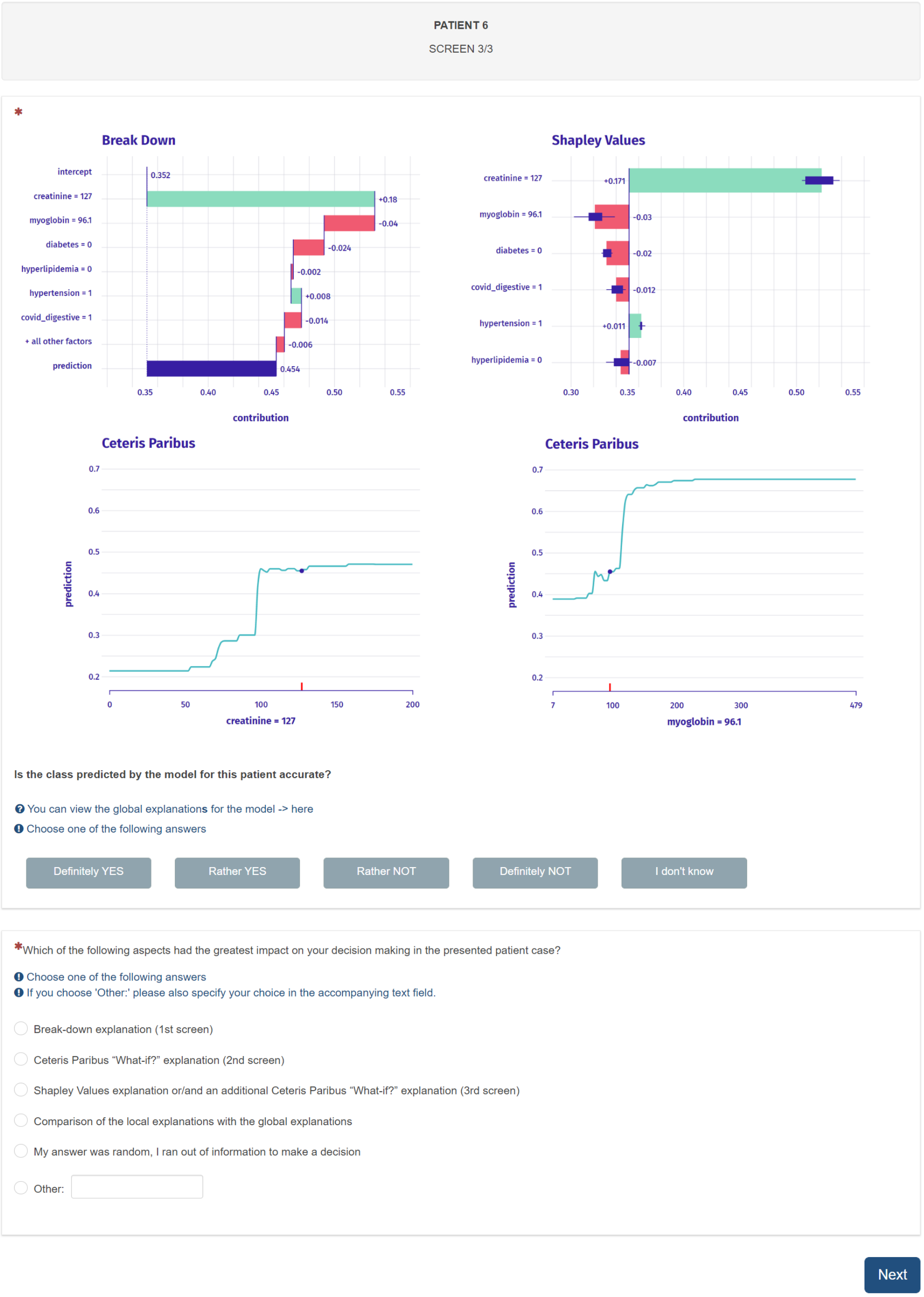
*Task description.* We chose a binary classification task from a medical domain for this study. It considers an authentic machine learning use case: predicting the occurrence of Acute Kidney Injury (AKI) in patients hospitalized with COVID-19. Physicians aim to estimate the probability of AKI based on the patient's blood test and medical history. Model engineers are tasked with developing and auditing a random forest algorithm for supporting such decisions. Overall, practitioners aim to use model explanations to allow for meaningful interpretation of its predictions. Let's consider a scenario in which, before deploying the model, a developer performs its audit by examining predictions with their explanations. Part of this audit is to look for wrong model behaviour based on abnormalities in either one. We aim to analyze how juxtaposing complementary explanations affect human accuracy and confidence in finding wrong model predictions.

*Experimental setting.* In this study, we rely on data of 390 patients from the clinical department of internal diseases in one of the Polish hospitals. The original values were altered slightly to maintain their anonymity. For each patient, we have information about 12 variables determined during the patient's admission: two quantitative variables that are biomarkers from a blood test: creatinine and myoglobin, five binary variables indicating chronic diseases: hypertension (among 62% of patients), diabetes (28%), cardiac atherosclerosis (19%), hyperlipidemia (32%), chronic kidney disease (5%); and five binary variables indicating symptoms related to COVID-19: fever (among 82% of patients), respiratory problems (90%), digestive problems (26%), neurological problems (8%), a critical condition requiring ventilator (6%). The classified target variable is a relatively rare binary variable: an occurrence of AKI during the patient's hospitalization (among 18% of patients). Overall, the above-described structure of the data was designed to be easily comprehended by the participants of our user study. There are two critical continuous variables, and the remaining binary ones can additionally affect the predicted outcome. Based on the data, we trained a random forest model with 100 trees and a tree depth of 3 for predicting AKI, which is treated as a black-box, later with an intention to deploy it in a hospital. To balance the training process, patients were weighted by the target outcome, therefore the model returns a rather uniformly-distributed probability of AKI (a number between 0 and 1). Assuming a classification threshold of 0.5, it achieved the following binary classification measures: Accuracy (0.896), AUC (0.946), F1 (0.739), Precision (0.644), Recall (0.866), which is more than needed for our user study.

*Questionnaire description.* We design a user study as an about 45-minute questionnaire, in which each participant is tasked with sequentially auditing the predictions with explanations for 12 patients. Specifically, a participant is asked to answer the question *"Is the class predicted by the model for this patient accurate?"* based on:

1. a single Break-down explanation ($Q_1$),
2. the same Break-down explanation with an additional Ceteris Paribus "What-if?" explanation of the most important variable based on the highest value in Break-down ($Q_2$),
3. the above-mentioned set of explanations with an additional Shapley Values explanation and a Ceteris Paribus "What-if?" explanation of an arbitrarily chosen variable ($Q_3$).

These three combinations of evidence were shown *sequentially* so that the participant could change their answer to the pivotal question of class prediction correctness. Note that the results obtained in $Q_1$ serve as a "control group" in our study since we aim to compare them with the results obtained after a sequence $Q_1$–$Q_3$ (see Section **??** for analogous studies comparing to other baselines). For answers, we chose a 5 point Likert scale consisting of "Definitely/Rather YES/NO" and "I don't know". On purpose, half of the presented observations were classified as wrong by the model (6/12). An example of such classification would be when the model predicts a probability of 0.6 while AKI did not occur for this patient. Figure 2 presents the 3rd screen from an exemplary audit process for a single patient. The participant was asked to answer an additional question on the third screen for each patient: *"Which of the following aspects had the greatest impact on the decision making in the presented case?"* ($Q_4$). The exemplary 1st and 2nd screens for this patient are presented in Appendix A.



PATIENT 6

SCREEN 3/3

*

**Is the class predicted by the model for this patient accurate?**

❓ You can view the global explanations for the model -> here

❗ Choose one of the following answers

*Which of the following aspects had the greatest impact on your decision making in the presented patient case?

❗ Choose one of the following answers

❗ If you choose 'Other:' please also specify your choice in the accompanying text field.

○ Break-down explanation (1st screen)

○ Ceteris Paribus "What-if?" explanation (2nd screen)

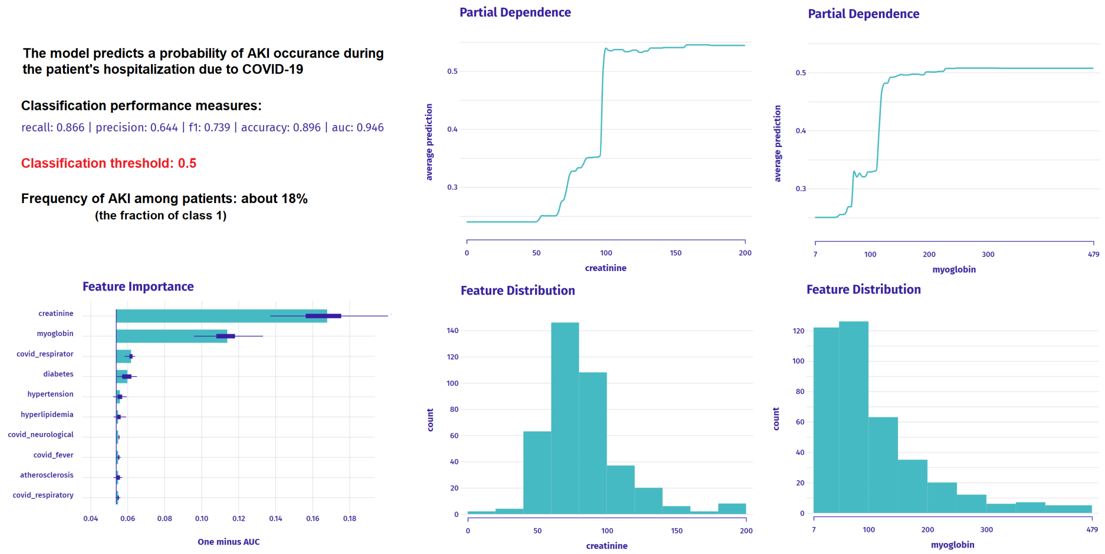
○ Shapley Values explanation or/and an additional Ceteris Paribus "What-if?" explanation (3rd screen)

○ Comparison of the local explanations with the global explanations

○ My answer was random, I ran out of information to make a decision

○ Other: ▭

Fig. 2. Screenshot from the user study's questionnaire showing the 3rd screen related to Patient 6 containing a set of four explanations: a Break-down explanations with an additional Ceteris Paribus explanations of the most important variable, and with additional Shapley Values and Ceteris Paribus explanations. At the end of each patient case, we asked for additional input on the most important factor affecting the participant's decision.



The model predicts a probability of AKI occurance during the patient's hospitalization due to COVID-19

Classification performance measures:

recall: 0.866 | precision: 0.644 | f1: 0.739 | accuracy: 0.896 | auc: 0.946

Classification threshold: 0.5

Frequency of AKI among patients: about 18%
(the fraction of class 1)

Fig. 3. Screenshot from the user study's questionnaire showing the explanatory context containing global explanations: Permutational Importance, and Partial Dependence explanations of the two most important variables with their distributions. These information were available at all times during the task.

To sum up the main task, each participant was tasked with answering 3 sequential questions ($Q_1$–$Q_3$ in Table 1), plus one additional indicating the participant's thought process ($Q_4$ in Table 2), about 12 patient cases each. Before the main task, the questionnaire made each participant familiar with a broad instruction, which discussed the task, data, model, and explanations, with a particular emphasis on data distributions and global model explanations (Figure 3). These visualizations were also available to each participant at all times during the questionnaire filling; hence, they are indicated as a possible answer in $Q_4$ for each patient case (shown in Figure 2). After the main task, there are some additional descriptive questions asked about the process, which allow us to qualitatively analyze the researched phenomenon (Section 2.2).

To make sure that the questionnaire is clear, we conducted a *pilot study* in person before the formal study, in which we validated our methodology with 3 participants and took their feedback into account. Additionally, the formal study contained a 13th patient case as a test case before the main task. The study was conducted as a computer-assisted web interview (CAWI) using a professional on-premise software with targeted invitations sent by email.

*Participants.* The target population in this study are data science practitioners with varied experience in machine learning and explainability, spanning from machine learning students to scientists researching XIML. Overall, there were 46 answers to our questionnaire, of which 31 were fully completed. Please note that we exclude one of the fully completed answers across reporting the results as it contains an answer of "I don't know" at each step of the questionnaire, which is rather redundant (see Figure 4). Thus, we rely on 30 answers in total. Crucially, this user study was anonymous with respect to the participants' identity, not their origin, as we aim to represent the target population correctly. The questionnaire was concluded with questions related to the participants' demographic data, e.g. about the participant's occupation and machine learning experience, which we report in Appendix B.

Table 1. Aggregated results from the user study validate our hypotheses. We report $mean_{\pm sd}$ across the participants' performance in 12 patient cases, and measure their difference between $Q_3$ and $Q_1$ marked as $\Delta Q_3 Q_1$. We validate each hypothesis with the t-test and Wilcoxon signed-rank test, hence two $p$ values. There is a significant increase in accuracy and confidence between the sequential questions. Additionally, the frequency of ambiguous answers decreases.

| **Hypothesis (number of cases = 12)** | $Q_1$ | $Q_3$ | $\Delta Q_3 Q_1$ | **$p$ values** |
|---|---|---|---|---|
| Accuracy increases between $Q_3$ and $Q_1$ | $52.2_{\pm 29.3}$ | $65.8_{\pm 24.2}$ | $13.6_{\pm 11.4}$ | 0.002; 0.004 |
| Confidence increases between $Q_3$ and $Q_1$ | $23.1_{\pm 13.7}$ | $35.3_{\pm 15.6}$ | $12.2_{\pm 11.8}$ | 0.004; 0.018 |
| "I don't know" *decreases* between $Q_3$ and $Q_1$ | $12.8_{\pm 9.8}$ | $5.2_{\pm 5.0}$ | $-7.5_{\pm 7.8}$ | 0.007; 0.007 |

*Expert validity phase to choose proper patient cases.* We conducted an expert validity study on 3 explainable machine learning experts before the described pilot and formal studies. The task was similar; it included answering the main question about the accuracy of model predictions based on information in all of the available explanations for 24 patients from the data (like in $Q_3$). We used the results to unambiguously pick the 12 patient cases where users of the highest expertise most agreed on answers concerning the information carried in explanations. This made the user study less biased with respect to our personal views.

### 2.1 Quantitative analysis

We first validate the two hypotheses by measuring the performance change between the sequential questions for each patient case using the following statistics:

- Accuracy: frequency of participants choosing "Definitely/Rather YES" when the prediction was accurate and "Definitely/Rather NO" when it was wrong.
- Confidence: frequency of participants choosing "Definitely YES/NO" as oppose to "Rather YES/NO" or "I don't know".

Additionally, we validate if the frequency of answers "I don't know" decreases over the course of questions, which corresponds to increasing human confidence and trust. Table 1 reports the aggregated quantitative results from the user study. We use the t-test and Wilcoxon signed-rank test to compare the differences $\Delta$ between $Q_3$ and $Q_1$, which serves as a baseline scenario in our study. We omit the analogous results for $Q_2$ where the difference is, as expected, smaller and report detailed numbers for each patient case in Figure 5 and Appendix C.

Since $\Delta Q_3 Q_1$ is positive, the overall conclusion is that the sequential analysis of a model $Q_1$–$Q_3$ with juxtaposing complementary explanations increases both human accuracy and confidence with respect to the single aspect-model explanation $Q_1$. Moreover, Table 2 presents the frequency of answers to $Q_4$ across all cases and participants. We observe an increasing relationship between the impact of consecutive explanations. Participants highlight that in about 19% of cases, juxtaposing global and local explanations had the greatest impact on their decision making, which we also view as a positive outcome towards our thesis.

### 2.2 Qualitative analysis

At the end of the user study, we asked our participants to share their thoughts on the user study. In the first question, we asked if they saw any positive aspects of presenting a greater number of explanations to the model. This optional question was answered by 19 participants, who most often pointed to the following positive aspects: the greater number of the presented explanations, the more information they obtain (n = 18; 95%), which allows a better understanding

Table 2. Frequency of answers for $Q_4$ averaged across 12 cases times 30 participants.

| *$Q_4$: Which of the following aspects had the greatest impact on your decision making in the presented patient case?* | |
|---|---|
| **Answer** | **Frequency** |
| Break-down explanation (1st screen) | 16.7% |
| Ceteris Paribus "What-if?" explanation (2nd screen) | 27.5% |
| Shapley Values explanation or/and an additional Ceteris Paribus "What-if?" explanation (3rd screen) | 35.3% |
| Comparison of the local explanations with the global explanations | 19.2% |
| My answer was random, I ran out of information to make a decision | 0.5% |
| Other (three descriptive answers in total: a Permutational Importance explanation, both Ceteris Paribus explanations, a high residual value) | 0.8% |

of the model (n = 13; 68%), and ultimately increases the certainty of the right decision making (n = 8; 42%) as well as minimizes the risk of making a mistake (n = 2; 11%). Additionally, we asked if the participants identified any potential problems, limitations, threats related to presenting additional model explanations? In 21 people answering this question, the most frequently given answers were: too many explanations require more analysis, which generates the risk of cognitive load (n = 15; 71%), and which may, in consequence, distract the focus on the most important factors (n = 7; 33%). Therefore, some participants highlighted the number of additional explanations as a potential limitation (n = 10; 48%). Moreover, the participants noticed that the explanations must be accompanied by clear instructions for a better understanding of the presented data, because otherwise they do not fulfill their function (n = 6; 29%), and may even introduce additional uncertainty to the assessment of the model (n = 4; 19%).

### 2.3 Detailed results

To analyze the results in detail, we deliver the following visualizations. Figure 4 presents specific answers given by the participants at each step of the questionnaire. Participants are clustered based on their answers with hierarchical clustering using the Manhattan distance with complete linkage, which is the best visual result obtained considering several clustering parameters. Note a single gray column corresponding to the removed participant. Overall, looking at the columns, we perceive more and less certain groups of participants, while in rows, we see blocks of three answers of similar color. Figure 5 aggregates the results presented in Figure 4 and divides them between wrong and accurate predictions. In this example, we better see the characteristic division into blue and red answers, as well as the change in the participants' certainty over $Q_1$–$Q_2$–$Q_3$. There were some hard cases in our study. Specifically, in case no. 10, participants were on average less accurate in $Q_2$ than in $Q_1$, and in case no. 12, participants were less accurate in $Q_3$ than in $Q_2$. Finally, the user study involved some follow-up questions asked at the end (Figure 6). The task was rather difficult for the users, from which we deduce that the created test in the form of a user study has high power in a statistical sense. Also, the participants think that presenting more explanations has the potential to increase the certainty and trust in the models.

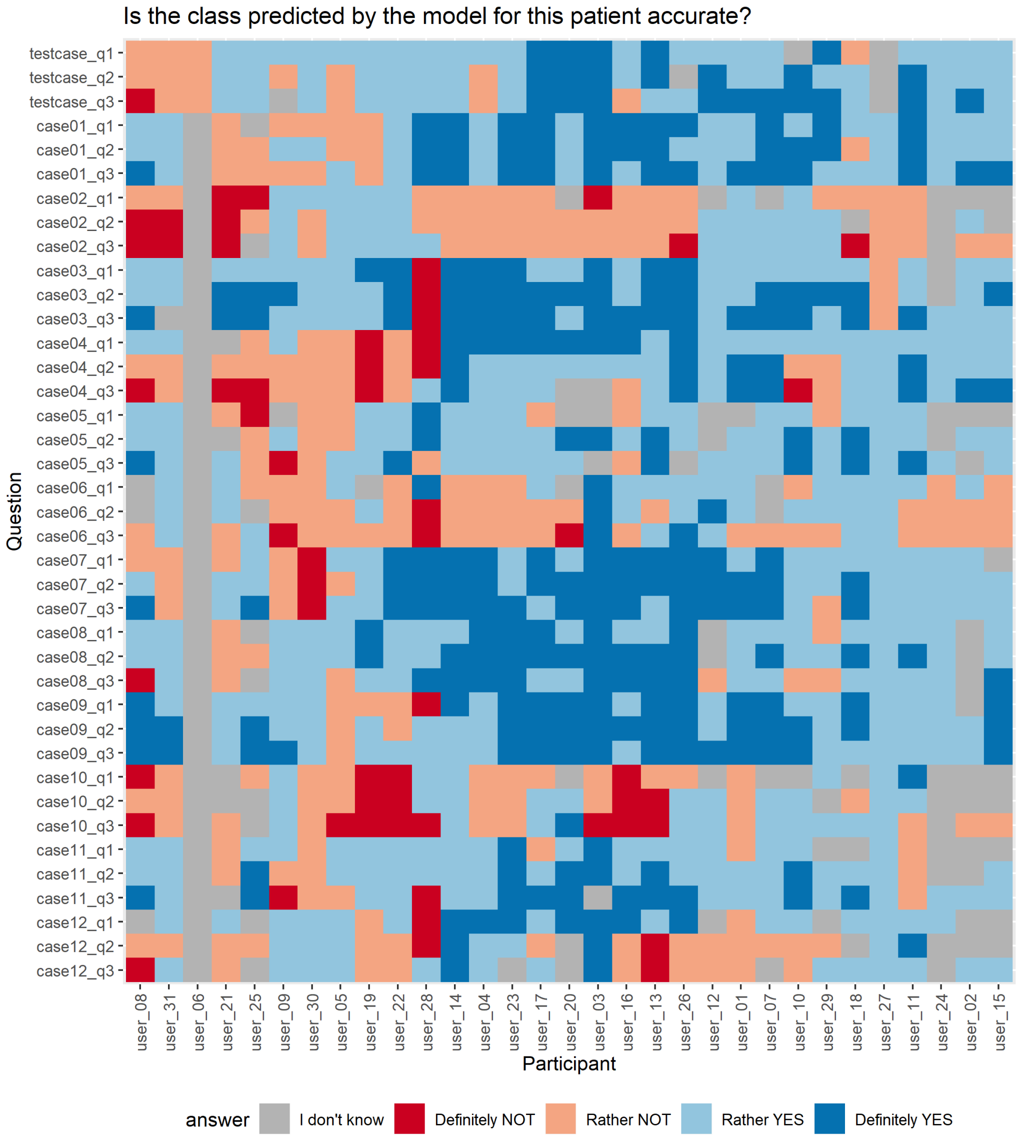
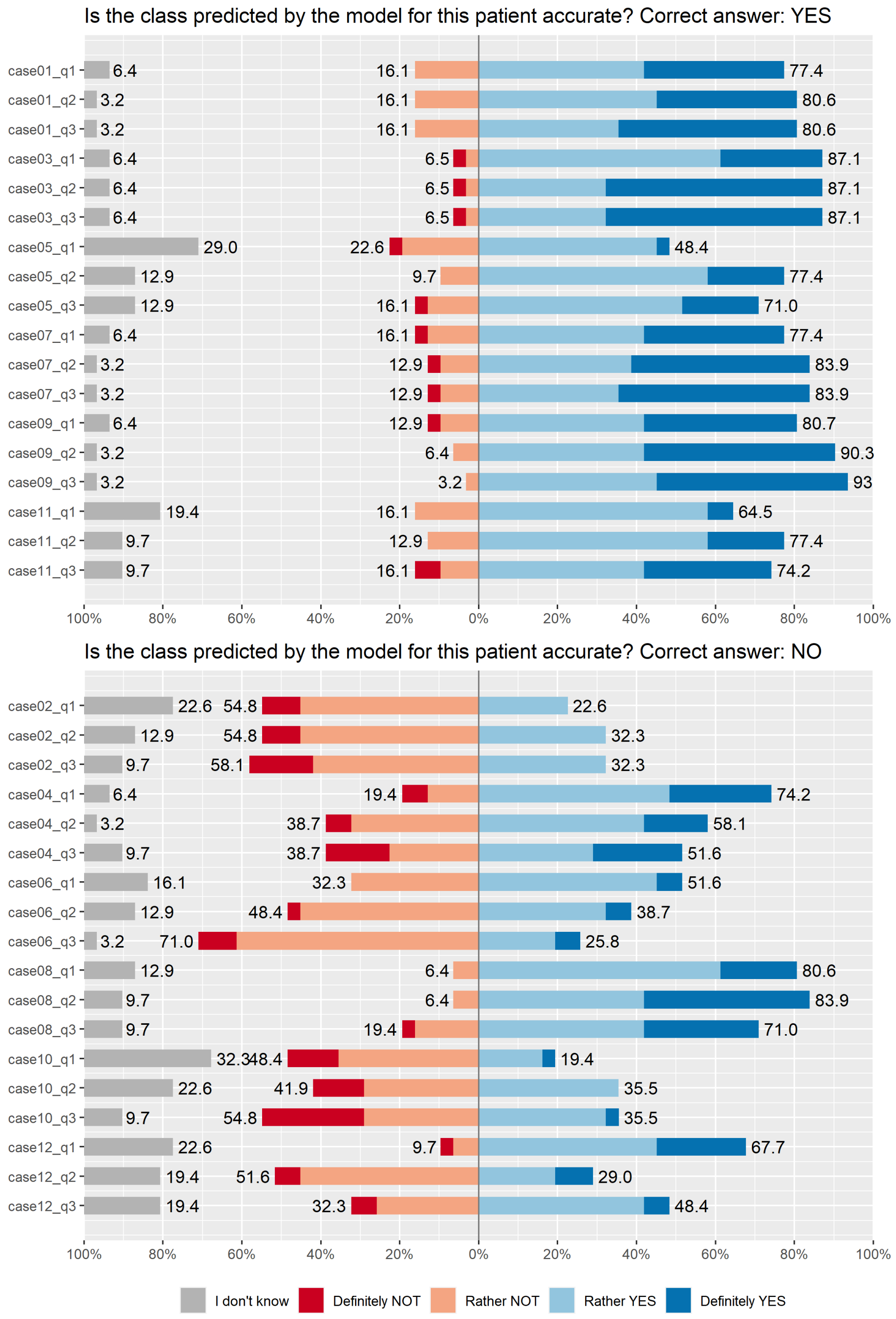


Fig. 4. Individual answers in the user study. Rows correspond to consecutive questions. Columns correspond to participants. Colors encode answers. Participants are clustered based on their similarity.

## 3 DISCUSSION

Our user study follows the One-Group Pretest-Posttest Design [16, 19], in which we compare the pretest observation, e.g. an answer to $Q_1$, with the posttest observation, e.g. an answer to $Q_2$. Crucially, this allows us to measure the change in human performance across time. Many threats to the internal validity of such a study are least plausible. For example, instrumentation, selection differences, and cyclical changes threats [16] are non-existent. Further, both the time and


Is the class predicted by the model for this patient accurate? Correct answer: YES

Is the class predicted by the model for this patient accurate? Correct answer: NO

Fig. 5. Summary of answers from the main part of the user study. Colors and questions correspond to these presented in Figure 4. Top panel corresponds to questions related to cases with correct predictions while the bottom panel corresponds to questions with incorrect predictions.

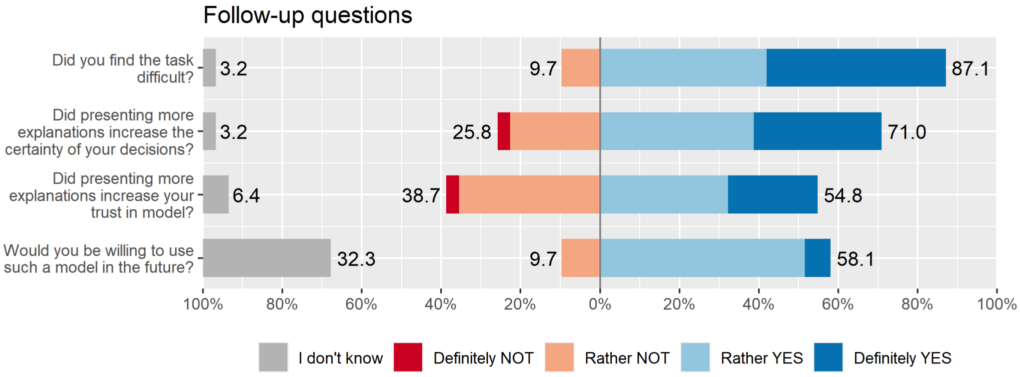


Fig. 6. Summary of answers from the follow-up part of the user study.

maturation effects are directly embedded into the principles of sequential model analysis. Nevertheless, the results like an increase in accuracy and confidence should be interpreted with respect to this experimental design assumption.

Many variables can affect the outcome of such a user study, yet most of them need to be fixed. First, we used only a specific predictive task, a real-world scenario of performing a model audit, and specific sets of explanations, which correspond to the available paths in the grammar of IEMA. To quantify the process and answers at different steps, we constructed a constrained questionnaire containing multiple views instead of allowing the users to interact with the dashboard themselves. In the future, it would be desirable to find ways of measuring a change in human performance when interacting in an open environment. To extend the results, we would like to perform a similar study on another group of stakeholders, e.g. physicians, in the case of predictive tasks concerning medicine. It could also involve other rules from the context-free grammar of IEMA, which correspond to alternative human-model interactions.

When choosing patient cases, we tried to account for a balanced representation of classes and balanced difficulties of predictions. When choosing participants, we aimed to gather answers from machine learning experts as opposed to crowd-sourced laypeople. Considering the above, we constrained the questionnaire to 12 patient cases aiming for about 45 minutes, which we believe allows for a reasonable inference. Overall, participants and results agree with evidence from previous work [1, 15] that finding wrong predictions based on explanations is a difficult task, which, in our view, makes it a more robust evaluation of our methodology. The experiment with human subjects confirmed our hypotheses that juxtaposing complementary explanations increases their usefulness. However, participants raised to attention the *information overload* problem [15] – the quantity of provided information needs to be carefully adjusted so as not to interfere with human decision making.

## 4 CONCLUSION

We conducted a user study to evaluate the usefulness of IEMA, which indicates that an interactive sequential analysis of a model may increase the accuracy and confidence of human decision making. The grammar of IEMA is founded on related work and our research neighbourhood's experiences in the explanatory analysis of black-box machine learning predictive models. The domain-specific observations might influence both practical and theoretical insight; thus, in the future, we would like to perform more human-centric experiments to study how possibly unidentified stakeholders analyze models.

## ACKNOWLEDGMENTS

This work was financially supported by the National Science Centre (Poland) grants number 2017/27/B/ST6/01307 and 2019/34/E/ST6/00052.

## REFERENCES

[1] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging Tests for Model Explanations. In *Conference on Neural Information Processing Systems (NeurIPS)*, Vol. 33. 700–712.

[2] Daniel W. Apley and Jingyu Zhu. 2020. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 82, 4 (2020), 1059–1086.

[3] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. 2010. How to Explain Individual Classification Decisions. *Journal of Machine Learning Research* 11, 61 (2010), 1803–1831.

[4] Hubert Baniecki and Przemyslaw Biecek. 2019. modelStudio: Interactive Studio with Explanations for ML Predictive Models. *Journal of Open Source Software* 4, 43 (2019), 1798. https://github.com/ModelOriented/modelStudio

[5] Hubert Baniecki and Przemyslaw Biecek. 2021. Responsible Prediction Making of COVID-19 Mortality (Student Abstract). *AAAI Conference on Artificial Intelligence (AAAI)* 35, 18 (2021), 15755–15756. https://doi.org/10.1609/aaai.v35i18.17874

[6] Hubert Baniecki, Dariusz Parzych, and Przemyslaw Biecek. 2023. The grammar of interactive explanatory model analysis. *Data Mining and Knowledge Discovery* (2023), 1–37. https://doi.org/10.1007/s10618-023-00924-w

[7] Gagan Bansal, Alison Marie Smith-Renner, Zana Buçinca, Tongshuang Wu, Kenneth Holstein, Jessica Hullman, and Simone Stumpf. 2022. Workshop on Trust and Reliance in AI-Human Teams (TRAIT). In *CHI Conference on Human Factors in Computing Systems (CHI)*.

[8] Przemyslaw Biecek. 2018. DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research* 19, 84 (2018), 1–5.

[9] Przemyslaw Biecek and Tomasz Burzykowski. 2021. *Explanatory Model Analysis*. Chapman and Hall/CRC.

[10] Leo Breiman. 2001. Statistical Modeling: The Two Cultures. *Statist. Sci.* 16, 3 (2001), 199–231.

[11] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 259—-268.

[12] Navdeep Gill, Patrick Hall, Kim Montgomery, and Nicholas Schmidt. 2020. A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing. *Information* 11, 3 (2020), 137.

[13] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Conference on Neural Information Processing Systems (NeurIPS)*, Vol. 30. 4765–4774.

[14] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38.

[15] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and Measuring Model Interpretability. In *CHI Conference on Human Factors in Computing Systems (CHI)*.

[16] Charles S Reichardt. 2019. *Quasi-experimentation: A guide to design and analysis*. Guilford Publications.

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 1135–1144.

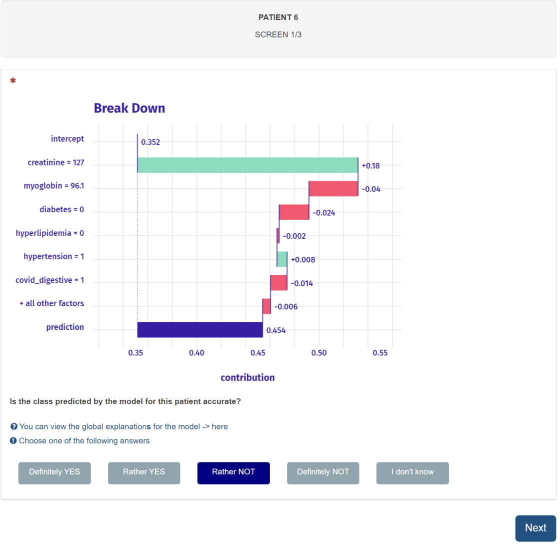
[18] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1 (2019), 206–215.

[19] William R Shadish, Thomas D Cook, and Donald T Campbell. 2002. *Experimental and quasi-experimental designs for generalized causal inference*. Houghton, Mifflin and Company.

[20] John W. Tukey. 1977. *Exploratory Data Analysis*. Addison-Wesley.

[21] Leland Wilkinson. 2005. *The Grammar of Graphics (Statistics and Computing)*. Springer-Verlag.

# A SCREENSHOTS OF THE QUESTIONNAIRE



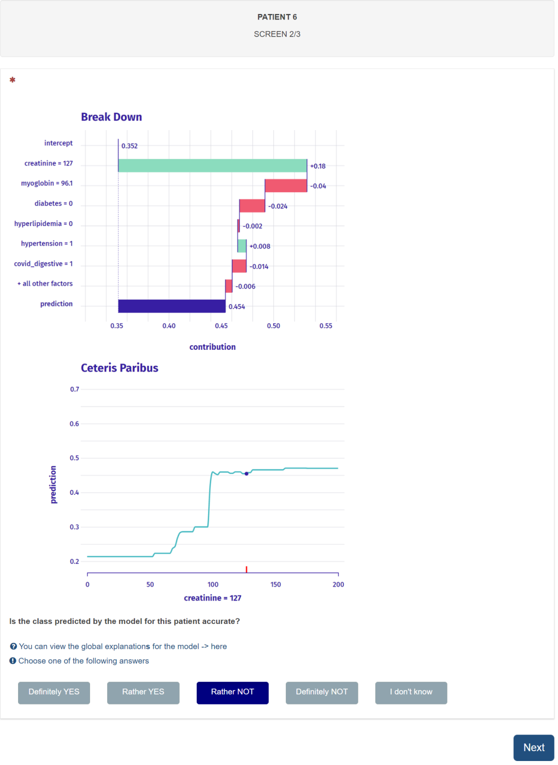
PATIENT 6

SCREEN 1/3

Is the class predicted by the model for this patient accurate?

❓ You can view the global explanations for the model -> here

❗ Choose one of the following answers



PATIENT 6

SCREEN 2/3

Is the class predicted by the model for this patient accurate?

❓ You can view the global explanations for the model -> here

❗ Choose one of the following answers

Fig. 7. **Left**: Screenshot from the user study's questionnaire showing the 1st screen containing a single Break-down explanation related to Patient 6. **Right**: Screenshot from the user study's questionnaire showing the 2nd screen containing a Break-down explanation with an additional Ceteris Paribus "What-if?" explanation of the most important variable.

## B DEMOGRAPHIC PROFILE OF PARTICIPANTS

Table 3. The demographic profile of 31 participants who fully answered the questionnaire in the user study.

| | Count | Frequency (%) |
|---|---|---|
| **Gender** | | |
| Man | 25 | 80.6 |
| Woman | 6 | 19.4 |
| I don’t want to answer | 0 | 0 |
| **What year of study are you in?** | | |
| 3rd year (BSc) | 17 | 54.8 |
| 4th year (BSc & MSc) | 7 | 22.6 |
| 5th year (MSc) | 3 | 9.7 |
| Other, e.g. PhD | 4 | 12.9 |
| **How much experience do you have in machine learning?** | | |
| 1 (No experience) | 0 | 0.0 |
| 2 | 4 | 12.9 |
| 3 | 8 | 25.8 |
| 4 | 7 | 22.6 |
| 5 | 7 | 22.6 |
| 6 (Extensive experience) | 5 | 16.1 |
| **How much experience do you have in explainable machine learning?** | | |
| 1 (No experience) | 4 | 12.9 |
| 2 | 6 | 19.4 |
| 3 | 10 | 32.3 |
| 4 | 6 | 19.4 |
| 5 | 3 | 9.7 |
| 6 (Extensive experience) | 2 | 6.5 |
| **How much experience do you have in using machine learning in medical applications?** | | |
| No experience | 9 | 29.0 |
| Participation in one project | 18 | 58.1 |
| Multiple projects and/or collaboration with medical staff | 4 | 12.9 |

## C DETAILED QUANTITATIVE RESULTS

Table 4. Accuracy for each patient case measured across the answers of 30 participants. $\Delta Q_3 Q_1$ indicates the difference in accuracy between $Q_3$ and $Q_1$. In **bold**, we highlight the results reported in Table 1.

| **Case no.** | $Q_1$ | $\Delta Q_2 Q_1$ | $Q_2$ | $\Delta Q_3 Q_2$ | $Q_3$ | $\Delta Q_3 Q_1$ |
|---|---|---|---|---|---|---|
| 01 | 80.0 | +3.3 | 83.3 | 0.0 | 83.3 | +3.3 |
| 02 | 56.7 | 0.0 | 56.7 | +3.3 | 60.0 | +3.3 |
| 03 | 90.0 | 0.0 | 90.0 | 0.0 | 90.0 | 0.0 |
| 04 | 20.0 | +20.0 | 40.0 | 0.0 | 40.0 | +20.0 |
| 05 | 50.0 | +30.0 | 80.0 | -6.7 | 73.3 | +23.3 |
| 06 | 33.3 | +16.7 | 50.0 | +23.3 | 73.3 | +40.0 |
| 07 | 80.0 | +6.7 | 86.7 | 0.0 | 86.7 | +6.7 |
| 08 | 6.7 | 0.0 | 6.7 | +13.3 | 20.0 | +13.3 |
| 09 | 83.3 | +10.0 | 93.3 | +3.3 | 96.7 | +13.3 |
| 10 | 50.0 | -6.7 | 43.3 | +13.3 | 56.7 | +6.7 |
| 11 | 66.7 | +13.3 | 80.0 | -3.3 | 76.7 | +10.0 |
| 12 | 10.0 | +43.3 | 53.3 | -20.0 | 33.3 | +23.3 |
| **Mean** | **52.2** | +11.4 | 63.6 | +2.2 | **65.8** | **+13.6** |
| **SD** | **29.3** | 14.4 | 26.3 | 10.9 | **24.2** | **11.4** |
| **Median** | 53.3 | +8.3 | 68.3 | 0.0 | 73.3 | +11.7 |

Table 5. Confidence for each patient case measured across the answers of 30 participants. $\Delta Q_3 Q_1$ indicates the difference in confidence between $Q_3$ and $Q_1$. In **bold**, we highlight the results reported in Table 1.

| **Case no.** | $Q_1$ | $\Delta Q_2 Q_1$ | $Q_2$ | $\Delta Q_3 Q_2$ | $Q_3$ | $\Delta Q_3 Q_1$ |
|---|---|---|---|---|---|---|
| 01 | 36.7 | 0.0 | 36.7 | +10.0 | 46.7 | +10.0 |
| 02 | 10.0 | 0.0 | 10.0 | +6.7 | 16.7 | +6.7 |
| 03 | 30.0 | +30.0 | 60.0 | 0.0 | 60.0 | +30.0 |
| 04 | 33.3 | -10.0 | 23.3 | +16.7 | 40.0 | +6.7 |
| 05 | 6.7 | +13.3 | 20.0 | +3.3 | 23.3 | +16.7 |
| 06 | 6.7 | +3.3 | 10.0 | +6.7 | 16.7 | +10.0 |
| 07 | 40.0 | +10.0 | 50.0 | +3.3 | 53.3 | +13.3 |
| 08 | 20.0 | +23.3 | 43.3 | -10.0 | 33.3 | +13.3 |
| 09 | 43.3 | +6.7 | 50.0 | 0.0 | 50.0 | +6.7 |
| 10 | 16.7 | -3.3 | 13.3 | +16.7 | 30.0 | +13.3 |
| 11 | 6.7 | +13.3 | 20.0 | +20.0 | 40.0 | +33.3 |
| 12 | 26.7 | -10.0 | 16.7 | -3.3 | 13.3 | -13.3 |
| **Mean** | **23.1** | +6.4 | 29.4 | +5.8 | **35.3** | **+12.2** |
| **SD** | **13.7** | 12.3 | 17.6 | 8.9 | **15.6** | **11.8** |
| **Median** | 23.3 | +5.0 | 21.7 | +5.0 | 36.7 | +11.7 |