Therapeutic Innovation & Regulatory Science
https://doi.org/10.1007/s43441-024-00689-4

DIA

ORIGINAL RESEARCH

On the Application of Artificial Intelligence/Machine Learning (AI/ML) in Late-Stage Clinical Development

Karl Köchert[1] · Tim Friede[2,3] · Michael Kunz[1] · Herbert Pang[4] · Yijie Zhou[5] · Elena Rantou[6]

Received: 3 January 2024 / Accepted: 18 July 2024

Abstract

Whereas AI/ML methods were considered experimental tools in clinical development for some time, nowadays they are widely available. However, stakeholders in the health care industry still need to answer the question which role these methods can realistically play and what standards should be adhered to. Clinical research in late-stage clinical development has particular requirements in terms of robustness, transparency and traceability. These standards should also be adhered to when applying AI/ML methods. Currently there is some formal regulatory guidance available, but this is more directed at settings where a device or medical software is investigated. Here we focus on the application of AI/ML methods in late-stage clinical drug development, i.e. in a setting where currently less guidance is available. This is done via first summarizing available regulatory guidance and work done by regulatory statisticians followed by the presentation of an industry application where the influence of extensive sets of baseline characteristics on the treatment effect can be investigated by applying ML-methods in a standardized manner with intuitive graphical displays leveraging explainable AI methods. The paper aims at stimulating discussions on the role such analyses can play in general rather than advocating for a particular AI/ML-method or indication where such methods could be meaningful.

Keywords Regulatory Decision Making · Machine Learning · Phase-III Trial · Exploratory Data Analysis

Introduction

In modern health care Artificial Intelligence/Machine Learning – (AI/ML) methods continue to gain popularity in several clinical fields of application. An overview of applications in cardiovascular medicine can be found in [1] and a specific use case algorithm that predicts circulatory failure in the intensive care unit using machine learning methods is provided in [2]. Another example is the application MyIUS (https://www.bayoocare.com/en/myius/ and [3]) that was developed for the prediction of vaginal bleeding patterns during usage of a particular birth control device, more specifically a so called Intra Uterine System (IUS). This application is available to patients as an app from both the Google Play and Apple App Store. For an easily accessible non-technical introduction into the field of ML we refer to [4].

AI/ML methods are far away from being a new phenomenon and date back to even the 1950s and 60s. [5, 6] may serve as early references in the literature on this topic. However, as represented by the above-mentioned recent applications, it can be seen that AI/ML can now have a direct impact for the individual patient or may even become a part of routine submissions of pharmaceutical companies to regulatory agencies like the US Food and Drug Agency (FDA). The latter aspect is of particular interest for us. This new setting for AI/ML requires interactions between all relevant stakeholders including regulators and industry for a clear mutual understanding of reporting requirements.

✉ Michael Kunz
michael.kunz@bayer.com

1 Bayer AG, Berlin, Germany

2 Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

3 DZHK (German Center for Cardiovascular Research), partner site Göttingen, Göttingen, Germany

4 Genentech, South San Francisco, CA, USA

5 Vertex Pharmaceuticals, Boston, MA, USA

6 Office of Biostatistics, FDA/CDER/OTS, Silver Spring, MD, USA

Published online: 21 August 2024

In recent years some regulatory guidance has become available, to a considerable degree driven by regulators overseeing medical devices. The FDA's Center for Devices and Radiologic Health (CDRH) released an action plan for AI or ML-based software as medical device [7], also compare a recent discussion paper on the use of AI/ML in the development of drugs and biological products [8] and a paper where several centers at FDA describe how they are working together on this topic [9]. European Medicines Agency (EMA) also just released a draft version of a guidance document on The use of Artificial Intelligence in the medical product lifecycle [10]. A further example is the creation of a sort of score card for Good Machine Learning Practice resulting from a collaboration between the FDA's CDRH, Health Canada and the UK regulatory agency MHRA [11].

We would like to note that current applications that are discussed in clinical research cover e.g., recruitment of participants, selection of trial participants coupled with the goal to decrease variability in the data, adherence to study intervention, and retention in the trial. Apart from these applications in clinical research the application of AI/ML is discussed across the whole value chain, i.e. also in drug discovery, non-clinical research, post market safety surveillance and pharmaceutical manufacturing.

A comprehensive overview on fields of applications and challenges in connection with safety topics can be found in [12]. The key examples for our considerations are phase-III trials that are analyzed using AI/ML methods after a thorough statistical planning of the whole trial was done.

In the following we reflect on the use of AI/ML in the reporting phase of a clinical study.

Despite these examples cited above, the application of AI/ML methodology in the analysis phase of late-stage clinical development for new drugs and biologics is limited. Statistical models typically applied in such settings model the outcome based on treatment and a usually very small amount of baseline covariates; see for instance the recently released FDA guidance "Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products" [13]. In a time-to-event setting an approach termed 5-step Stratified Testing and Amalgamation Routine (5-STAR) was introduced that overcomes this limitation via identifying strata where risks are more homogeneous than in the overall population [14]. The authors propose ML-methods like elastic net Cox-regression to overcome "noise" in the covariates defining "strata", hence providing a prognostic factor for the outcome of interest, which might then also turns out to be a predictive factor for the treatment of interest. In statistical terms identifying a prognostic factor means identifying a factor that is associated with an outcome of interest irrespective of treatment while identifying a predictive factor corresponds to identifying a treatment by covariate interaction.

These considerations on the use of baseline covariates also relate to what is now known as "precision medicine" or "personalized medicine", which defines a paradigm shift in the development of biopharmaceuticals, see also a presidential address by Barack Obama [15]. This paradigm shift started to penetrate into public perception as well as into other fields of science and industry. Most importantly, it impacted and shaped new approaches in drug development which can capitalize on the ever-increasing amount of data being measured to gauge the precise state of a biological system or a particular disease as a system deviating from "normal average biology", see [16, 17] in connection with Alzheimer's disease.

One shortcoming of current practice in applying statistical models is that higher-order interactions, non-linear, or non-monotonic functional associations are often not appropriately reflected in the models applied. Although it would be possible to explicitly pre-specify and include those in subsequent confirmatory hypothesis testing, this would require in-depth a priory knowledge to ensure unbiased modelling.

In contrast, AI/ML-based analysis methods are able to accomplish these tasks rather automatically considering the complete hypothesis space including implicitly interactions as well, i.e. all measured data and variables at once, thus deducing a holistic mathematical model of a disease and the respective treatment effect in a completely data-driven fashion; see [18] for a related discussion. As such, AI/ML methodology is key in enabling true data driven decision making and a prerequisite for precision medicine since it complements the currently predominantly applied low dimensional approaches to enable as thorough as technically possible insights generation.

For decision making during the drug development process, it is of importance that analyses are provided in a timely fashion and with high quality. Looking back at the authors' own experiences these analyses were more in the style of prototypes, but with respect to the aspects time and quality progress was made recently via standardization and the introduction of agreed reporting and documentation standards within organizations.

It is beyond the scope of this paper to give a complete overview of the available literature on the topic of AI/ML in clinical development. It is also not the aim of this paper to advocate a particular method. We feel that the time has come to consider examples like the one presented here for inclusion into formal regulatory submissions. This would constitute a big step forward because these analyses would by nature not necessarily be reported in forms of tables and figures that can be easily reproduced. There might also

be overlaps to other initiatives to create interactive study reports. It is the aim of this paper to present some experiences made so far, embed this into work and documents available from regulators and encourage others to think into a similar direction.

In summary the applications of AI/ML in a particular clinical study can be allocated to the stages of feasibility, study set-up, study monitoring, study quality assessment or study reporting. Our focus in the use case is on the reporting phase under special consideration which principles should be adhered to and which benefits could potentially be provided to regulators.

The remainder of this paper is organized as follows. Section "AI/ ML in the FDA Regulatory Environment" provides a short overview on available regulatory guidance and on activities where FDA statisticians apply AI/ML methods. Moreover, Section "Adoption of AI/ML Methodology for Late-Stage Clinical Development by Industry: A Case Study" provides a specific efficient framework to apply AI/ML-methods in the setting of a pivotal study in late-stage clinical development. Of note this example refers to the reporting phase of a clinical trial. As can be seen other fields of application also exist, but are not further discussed throughout this paper. The paper avoids technical language wherever possible to enable a discussion among all stakeholders. We conclude this paper with a discussion and recommendations.

AI/ML in the FDA Regulatory Environment

Modern drug development programs tend to create increasingly larger volumes of patient data. Using appropriate methods to analyze such data has the potential to reshape biopharmaceutical development and might even have subsequently an impact on regulatory decision making. This section highlights how AI/ML methods have been used in a regulatory environment, namely at the US (FDA Center for Drug Evaluation and Research (CDER) and Center for Devices and Radiological Health (CDRH)). The use of such methods can be generally decomposed into:

a. AI/ML-related submissions to FDA.
b. AI/ML application in the FDA review and research work.

Background – Guiding Principles

In a common effort to promote safe and effective medical devices that use AI/ML, FDA's CDRH along with Health Canada and UK's Regulatory Agency (MHRA), have jointly identified 10 guiding principles that can inform the development of Good Machine Learning Practice (GMLP) [11]. Although they are meant for the development of medical devices and medical software, we are citing these principles as we feel they are general enough to also guide the application of AI/ML-methods in biopharmaceutical development. These guiding principles may be used to adopt good practices that have been proven in other sectors and to tailor practices from other industry sectors, so they are applicable to medical technology and the health care sector. This list of principles can be considered as a first step and a drive to adjust and expand such practices in the sector of biopharmaceutical development.

Some of these principles address the technical aspect of the process and are current and applicable to all ML applications, such as the independence of the training and test sets, the performance monitoring and management of the re-training risks, and good software engineering and security practices.

With respect to the application of the 10 guiding principles we would especially like to highlight the following four:

a. Multi-disciplinary expertise is leveraged throughout the total product life cycle.
b. Emphasis is put on the human interpretability of the ML model, rather than the model performance in isolation.
c. Statistically sound test plans are developed and executed to generate clinically relevant drug performance information independently of the training data set. Such considerations include the intended patient population, important subgroups, and potential confounding factors among others.
d. Information is established and is available to patients regarding the model performance for appropriate subgroups, characteristics of the test and training data, acceptable inputs, known limitations, user interface interpretation, and clinical workflow integration of the model. Additionally, the drug users are aware of any updates from real-world performance monitoring and able to communicate product concerns to the developer.

The complete list of principles is included as an appendix to this paper.

Overview of AI/ML in Regulatory Submission Processes

In recent years there is an increasing trend of AI/ML–related regulatory submissions received by the FDA's Center for Drug Evaluation and Research (CDER). For an overview

of recent applications of AI/ML in such submissions see [19]. This work presents the most common analysis types and objectives of these submissions. A rapidly increasing number of submissions including AI/ML was received by CDER through Investigational New Drug (IND), New Drug Application (NDA) and Abbreviated New Drug Application (ANDA) (with the vast majority being IND) applications during the period from 2016 to 2021. The leading therapeutic areas of these submissions were oncology (27%), psychiatry (15%), gastroenterology (12%) and neurology (11%). The most common objectives of these regulatory submissions were disease prognosis and treatment response prediction in efficacy and safety studies, covariate selection and confounding adjustment, pharmacometric modeling, imaging/video/voice analysis, drug discovery, drug toxicity prediction, enrichment design, dose selection and optimization, adherence to drug regimen, generating synthetic controls, endpoint and biomarker assessment and post-marketing surveillance. The most common AI/ML analysis approaches of these regulatory submissions were decision tree-based models and deep learning (analysis of imaging data). Further details on the different types of ML-algorithms were reported in [20].

AI/ML in FDA Review and Research work

Besides the use of AI/ML in regulatory submissions received by CDER, FDA reviewers/researchers have been applying AI/ML algorithms for various applications in their everyday work and research. Such applications are applied to the areas of determination of the optimal dosage range, prediction of drug toxicity, analysis of the FDA adverse event reporting system (FAERS), prediction of clinical sites that will qualify for FDA inspection, regulatory drug safety evaluation and hierarchical clustering of pharmacokinetic (PK) curves.

In the case of drug safety evaluation at FDA, ML has been applied in both prediction and causal inference problems. A recently published paper summarizes the different ML applications and discusses the challenges and considerations when using ML methods with RWD to generate Real World Evidence (RWE; [12]). It provides a summary of cases where ML was employed to impute missing data, improve algorithms to identify health outcomes in the Sentinel system (FDA's national electronic system used by researchers to monitor the safety of FDA-regulated medical products and devices [21]), reduce the risk of potentially inappropriate opioid prescriptions and estimate causal effects/association measures for RWD. Various ML tools are being discussed in reference to the above issues. Such tools include unsupervised learning for clustering and discovering patterns for drug utilization as well as supervised learning techniques where typically a relationship between a clinical endpoint and potential explanatory factors shall be established. The same paper emphasizes that caution is required when ML methods are used in safety evaluation mainly because of their data driven nature and degree of algorithm complexity. Such caution should be maximized when RWD is used to generate RWE. More specifically, special attention should be paid to reproducibility, data transparency maintaining patient confidentiality, sparsity of rare events in the training data set and interpretability of study findings.

The use of AI/ML in clinical site inspection selection is considered in [22, 23]. These contributions compare several methods and combinations of methods. In a clinical trial setting, data reliability can be jeopardized by poorly collected, processed, or reported data or even by fraudulent data. The substantial increase in both the number and complexity of clinical trials makes it difficult for regulators to choose clinical sites for inspection however, due to limited resources such that only less than 1% of the sites can be inspected annually. It is therefore crucial to select the appropriate clinical sites for inspection.

Site inspection results can be classified into NAI (No Action Indicated), VAI (Voluntary Action Indicated) and OAI (Official Action Indicated). One of the main challenges in using AI/ML methods for site inspection selection comes from the imbalanced outcomes since OAI classification is considered a rare event (approximately 1% of all cases).

Lastly but not least, another area FDA researchers use AI/ML is on clustering pharmacokinetic (PK) concentration curves, as presented by [24]. The authors conclude that hierarchical clustering for grouping patients by PK concentration curve shape has the potential to identify signals and draw conclusions that would otherwise be hidden when applying standard analyses of the Area Under the (concentration-time-) Curve (AUC) or the Maximum concentration (Cmax) and to identify outliers. This technique would have to be tested and its clinical value assessed in a setting where clinical outcome or safety data are available.

The key papers are summarized in Table 1.

Adoption of AI/ML Methodology for Late-Stage Clinical Development by Industry: A Case Study

With all of the considerations above in mind Bayer introduced AI/ML-based analyses in the reporting phase of late phase trials, i.e. Phase II and Phase III trials, on a regular basis to detect complex efficacy and/or safety signals that may not be detected using low-dimensional statistical models. Whilst at this point these analyses are done in an

Table 1 Key papers recently authored by FDA-statisticians on the application of AI/ML-methodology in the drug development process

Author(s)	Year	Journal	Title	Objective	Paper type
Liu et al.	2022	Clinical Pharmacology and Therapeutics	Landscape Analysis of the Application of Artificial Intelligence and Machine Learning in Regulatory Submissions for Drug Development From 2016 to 2021.	Summarizes AI/ML in regulatory submissions	Review paper
Liu et al.	2020	Clinical Pharmacology and Therapeutics	Application of Machine Learning in Drug Development and Regulation: Current Status and Future Potential.	Discusses AI/ML in regulatory submissions	Review paper
Zhang et al.	2022	Statistics in Biopharmaceutical Research	The Use of Machine Learning in Regulatory Drug Safety Evaluation.	Reviews AI/ML work in drug safety evaluation	Review paper
Hein et al.	2019	Journal of biopharmaceutical Statistics	Comparing Methods for Clinical Investigator Site Inspection Selection: A Comparison of Site Selection Methods of Investigators in Clinical Trials.	Compares methods of clinical site inspection selection	Original Research
Lautier et al.	2022	Initially presented at JSM 2022, Washington DC	Applications of Machine Learning in Pharmacogenomics: Clustering Pharmacokinetic Concentration Curves,	Explores the use of hierarchical clustering in pharmacogenomics	Original Research

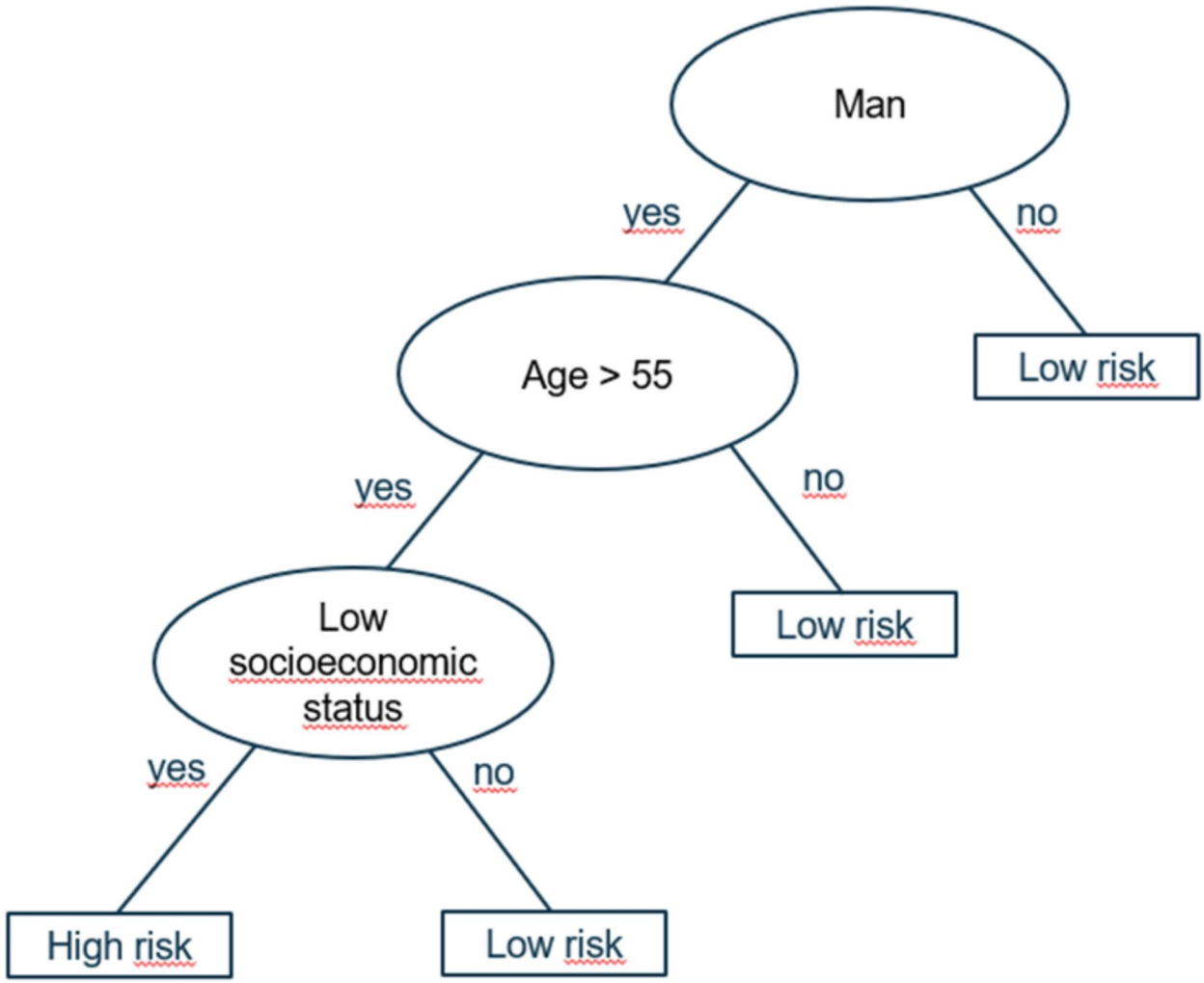


Fig. 1 Example of a decision tree to classify patients into "High risk" or "Low risk" patients

exploratory and post-hoc fashion, they are anticipated to become part of prospective planning in the future. Results from these analyses are expected to be included in clinical study reports which are the basis for regulatory approval decisions. To the best of our knowledge analyses such as those presented here have not been included in submission packages yet, but we consider this a realistic option and therefore would like to contribute to a discussion on the appropriateness of these analyses in a regulatory setting. However there are other applications of particular AI/ML-methodology that we have included into the Statistical Analysis Plan of a phase-IIb-study.

From Artisan AI/ML Prototyping to Standardized and Efficient Software Solutions

Across the pharmaceutical industry AI/ML methodology has been applied for many years in the space of high-dimensional biomarker development and analysis (such as gene expression data). However, these biomarker projects are mostly research driven, in an exploratory nature, and have different needs and expectations attached. In our experience, typically an AI/ML solution is tailored specifically for each biomarker project; as a result, each project consumes a considerable amount of resources and time but only a small part of it, for example a small part of the software code that is written, can be re-used in later projects. Therefore, when applying AI/ML in late phase and especially pivotal clinical trials to enable data-driven decision making this execution approach is inefficient and not fit-for-purpose. We hence set out to streamline the AI/ML capabilities, creating highly flexible yet standardized, reproducible and validated AI/ML software modules following usual terminology in bioinformatics called pipeline to efficiently compute and report holistic AI/ML models for all typical Phase II /III clinical endpoints based on all measured data, i.e. all Clinical Data Interchange Standards Consortium (CDISC) domains from late phase interventional trials.

The software is built on the principles of software engineering best practices; it is developed in *R* and is highly modularized, leveraging efficient coding and computation concepts from *tidyverse*, *tidymodels* [25] and *mlr3* [26]. The modularization allows for seamless plug-in of different AI/ML engines and explainable AI components needed to interpret complex models in a way amenable to clinicians. With this set up we are striking a balance between the flexibility needed to address trial specific peculiarities whilst ensuring as much standardization as possible for efficiency.

The software modules were developed over the last 4 years. Bayer is currently in the process of making these modules publicly available on github.com/Bayer-Group, including extensive documentation and usage examples. For now code can be requested from the authors.

Quality Control and best Practices

When leveraging AI/ML data analyses in late-stage clinical trials it is key to have the same level of quality and robustness just like any other analysis performed and potentially brought forward to and discussed with regulators. While generating new insights, AI/ML technology and its application is also more complex in terms of methods, data preprocessing, software engineering and documentation. In this vein, ensuring reproducibility and reusability are imperative as mentioned in [27], specifically when using AI/ML approaches with human data in interventional clinical trials, a highly regulated environment.

To address this a framework of best practices was developed. Teaming up with colleagues from several different statistics departments within Bayer we have developed an internal AI/ML best practice document that entails guidance on what needs to be decided, done and documented when planning and executing AI/ML analyses within a standardized analysis framework. This guidance, at its core, is built on the "data, optimization, model and evaluation" (DOME) concept published in [27] as well as the Good Machine Learning Practice previously mentioned [11]. To accommodate the clinical trial setting at a pharmaceutical company, we have extended the DOME concept specifically in terms of (i) exact data traceability and (ii) software validation in line with Bayer's standard operating procedures and programming best practices. The final document has been rolled out, accompanied by trainings, to our statistical staff to ensure that planning, execution, and interpretation of AI/ML analysis will be based on common quality and robustness standards.

A slightly redacted version of the final document (without company internal cross references) is added as supplementary material to this paper. For all analyses conducted in a regulatory setting it is essential that data derivations, objectives, (statistical) algorithms, and validation steps are addressed. The document addresses the particularities that especially occur in the setting of computational intensive AI/ML-methods.

Leveraging Value of AI/ML Analyses for Decision Making within Clinical Teams

One of the most important aspects of complementing conventional statistical analysis methods with AI/ML approaches in the setting of late phase clinical statistics is the interpretation of highly complex results. They must be made amenable to stakeholders of diverse backgrounds

including clinicians and regulatory scientists. At Bayer, we have developed a framework for presenting key AI/ML results with this audience in mind. It consists of a standardized set of tables, listings and figures (TLFs), all of which are derived from explainable AI concepts. Most of these TLFs can be derived from any of the standard ML methods typically used for structured data. For the purpose of this paper however, we will focus on one particular ML method, Random Forests (RF) introduced in [28]. A notable body of evidence (see e.g. [29, 30]) (and company internal benchmarking over many years and differing projects) shows that RF provides good balance among ML model performance, interpretability and ability to deduce complex holistic models. Additionally it features the flexibility needed to model common endpoint types in Phase II/ III settings. In the following sections, we will introduce the components of the aforementioned framework of results presentation. Note that the purpose here is not to describe in great depth the technical details but rather to show what we believe are the fundamental building blocks of such a framework in a clinical development organization of a pharmaceutical company.

As random forests are the technique used to generate the output presented in this paper we provide some further illustration so that all readers irrespective of their background get an idea of the basic principle of this technique:

Consider the situation where the risk status for a certain disease (high/ low) of a subject shall be classified based on available (health) data – and based on this assessment further diagnostics may be considered necessary. Using a decision tree that takes suitable information into account can be a natural choice. In the admittedly somewhat artificial setting where it is known that only men above the age of 55 years with a low socioeconomic status have a high risk then a possible decision tree could look as follows:

Intuitively, if many variables (in this setting usually called features) are available then the classification derived from an available data set can be close to perfect. However such a classification usually does not turn out to be too useful as the decision tree will usually not yield good results in an independent sample of subjects.

Random forests overcome this challenge via not deriving only one decision tree, but many decision trees, hence the name random "forest". First several data sets $D_1,\ldots, D_n$ are derived from the original dataset via randomly drawing from the available dataset D with replacement. "Randomly" refers to both choosing subjects randomly from all available subjects and to choosing not all, but a subset of features to derive a classification rule for each randomly selected data set D_i. An overall classification rule is then obtained via using the majority vote of the single classification rules.

Information Content: how well do we Understand the Biological System Given the data?

One of the key components of any AI/ML model fed with complex data sets is the model performance on independent data not used to build the model, i.e. its ability to generalize and correctly predict a clinical outcome in new patients. Estimating this performance in an unbiased manner is by itself a complex task and details on how this can be done can be found for example in [31]. In our framework and the corresponding analysis practice, appropriate re-sampling methods are applied, potentially at various steps when computing an AI/ML model, ultimately allowing to obtain a robust estimate of model performance. The key information we can derive is twofold:

- Firstly the performance estimate, which is conceptually related to the idea of "information content" described already in the year 1948 in [32]. It allows us to gauge how well we are able to understand and predict the behavior of a complex information processing system using a number between 0 and 1. Depending on the (clinical) outcome type that is modelled (e.g. continuous normally distributed, or binary, or time-to-event), well known performance measures such as R^2, Receiver Operating Characteristics – Area Under the Curve (ROC-AUC) or Harrel's C can be used. For the remainder of this paper we focus on Harrel's C. Harrel's C, also known as concordance index, is a measure of a model's ability to discriminate between lower and higher risk patients. It yields values close to 1 if subjects with a higher risk score tend to have the event of interest later than subjects with a lower risk score. Values close to 0.5 indicate that the risk score does not yield more information than flipping a coin.
- Secondly, comparing the AI/ML model performance to the performance of the conventional statistical modeling which is typically pre-specified in a statistical analysis plan helps to quantify how much information content is not captured with the latter model (see Fig. 2). It is possible that the AI/ML model does not have more information than the relatively straightforward conventional statistical modeling, which then may suggest that information to build a meaningful knowledge basis of the disease and potential treatment effect may be incomplete. Such finding may inform scientific (what else ought to be measured and when to measure, to get a better information content) and strategic discussions (what is the probability of success when moving forward with the development program).

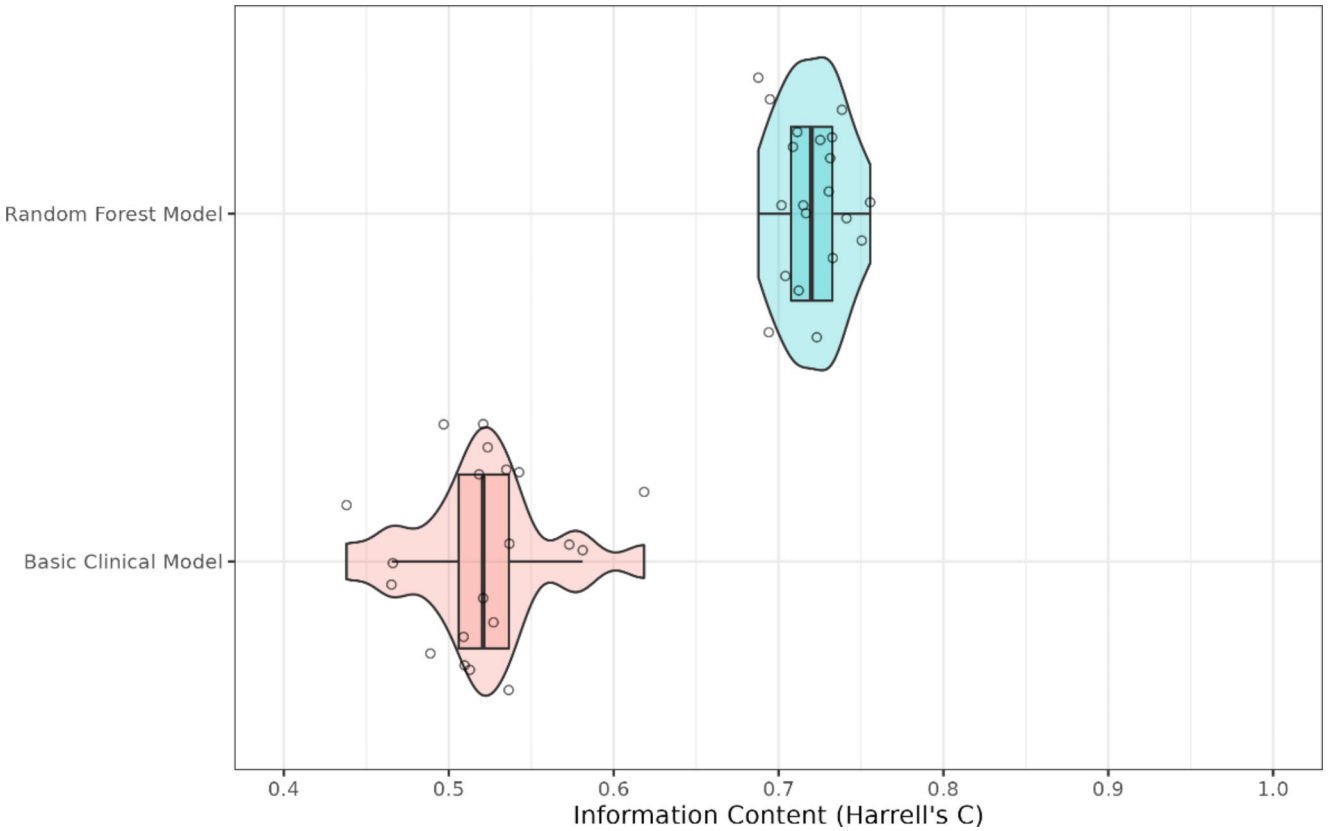


Fig. 2 Information content (approximated by Harrell's C in this case) quantification of a time to event endpoint example case with 112 simulated variables including a binary treatment label. 0.5 is equivalent to no information captured by the model given the data and 1.0 is equivalent to complete information and a perfect model to explain the clinical endpoint. The basic clinical model is the typical one defined for primary analysis, i.e. treatment and strata terms are the only independent variables used for modelling the endpoint. In contrast, the Random Forest model was built with all 112 available clinical covariates using a logrank splitrule to accommodate for the time to event endpoint. This type of Random Forest is also known as Random Survival Forest, see [35]. It captures substantially more information content than the basic clinical model. Information content quantification was based on 4 × 5-fold cross validation

As can be seen from this figure the Random Forests model provides a better prediction compared to the basic clinical model. However the Random Forests model still leads to a model which is far away from being perfect.

What are the Major Drivers of the Disease and Treatment Effect and what is Their Functional Relationship to the Clinical Outcome?

Conventional statistical modelling approaches usually avoid modelling all possible interactions, although in principle this is possible. For the AI/ML-models we have in mind this is more or less a by-product, but clinical teams will need to understand and interpret these complex types of models to derive actionable insights. In our framework we focus on three types of tasks to enable interpretability that build on the concepts of explainable AI.

Variable Importance

In Random Forest, a concept to estimate the impact of each assessed variable on the clinical outcome of interest is called variable importance. Various types and derivatives of this variable importance exist and each can serve slightly different purposes. However, all of them allow us to gain an understanding of the impact of each variable in relation to all other variables as the variable importance integrates the main effect of a variable with all potential interaction effects with other variables. For example, with a continuous outcome assumed, it is straightforward to understand how much of the outcome variance is actually explained by treatment, adjusting for all other covariates measured in the trial (see Table 2), which is scientifically as well as strategically a very important piece of information.

The table shows that treatment only ranks 6th among all variables in the model. However, it has to be considered that many often occurring factors like age either cannot be modified at all or need high efforts to modify, while for treatment

Table 2 *Example of a* variable importance table representing the complete hypothesis space. Permutation importance was used here and the value is equivalent to the information content (Harrell's C in this case) of the respective variable. It is important to note, that variable importance is not a univariate measure but integrates the main effect of the respective variable with all potential (higher order) interaction effects with any of all the other variables that are part of the Random Forest model

Importance Rank	Variable	RF Variable Importance
1	X1	0.014
2	X2	0.006
3	X3	0.006
4	X4	0.006
5	X5	0.005
6	treatment (TRT)	0.005
…	…	…
111	X111	0
112	X112	0

this is not the case and in many situations can even rather easily be applied.

Detecting Complex treatment-effect Subgroups

One of the most important features of AI/ML, specifically of tree-based methods such as Random Forests, is the property of implicit inference of interactions among all measured variables of essentially any order and complexity. To identify and reliably quantify such interaction effects from AI/ML models, various approaches have been reported in the literature. For example [33–35] are under the concept of generally detecting interactions between any pair of covariates based on Random Forests models. As another example with a different concept, [36] focuses on detecting interactions of covariates with treatment which can be used to define patient subgroups with an e.g. enhanced treatment effect using virtual twins. Clinical teams will have to decide on a case-by-case basis which of those to apply and it is important to note that, as with any subgroup analysis, this is merely hypothesis generation in need of confirmation by other means. Also, as part of the robustness checks when analyzing subgroups in a specific trial, various approaches exist and ought to be included in AI/ML subgroup analyses, see [37, 38].

With a hypothesis space of potential interactions deduced from clinical trial data, one can visualize a virtual biological system in the form of a network graph in which variables are nodes sized proportionally by their variable importance and edges between nodes indicate interaction between the variables (see Figs. 3 and 4). For example, Fig. 4 shows that there should be an interaction between treatment and e.g. X_1, X_2, X_3 resp.

When detection of interactions is based on tree models such as Random Forests, edges can optionally also be given a direction by deriving the interactions based on the mean minimal depth between a given root node covariate (the tail, i.e. start of the edge) and a subsequent child node covariate (the head, i.e. end of the edge). The idea here is that a root node covariate is normally more important than a child node covariate, which may be important information for correct interpretation of a potential interaction, specifically within a network graph.

Interpreting Complex Functional Relationships between Variables and a Clinical Outcome

To quantify, visualize and interpret the complex functional relationships between a clinical outcome and measured variables modelled by an AI/ML method, clinical teams can decide to do this at population level or patient level. In the Bayer framework, typically these analyses are performed on the population level using Partial Dependency Plots (PDP) and Accumulated Local Effects (ALE) plots. By doing so we can visualize the ML-inferred functional relationship of a variable with outcome whilst adjusting for all other covariates (PDP) and cross-correlation among them (ALE), see Fig. 5. With PDP and ALE plots, the complexity of a functional relationship becomes visible to clinical teams. Eg In the ALE plot of Fig. 5a the umbrella type shape of the curve central values of X_3 i.e. values between 75 and 125 lead to values clearly above 0, while values outside this central area lead to values below 0. Clinically this would mean that subjects inside this central range would have a survival benefit while this would not or to a lesser degree apply to subjects outside this central range. The non-monotonic non-linear functions provided by such visualizations are amenable to interpretation and potential decision making, because it allows to determine cut-off ranges of a variable's value at which clinical benefit may turn into detrimental effects. This can be an important step when deciding of treatment populations moving forward in a drug development program. A very interesting feature of these explainable AI analyses is that interaction effects can be incorporated and visualized, especially for any two-way interaction of a variable with treatment to determine whether the treatment effect is potentially modified as function of the level of any other variable.

For a technical description of all these methods and references to software implementations of them please see [39]. Currently Bayer is in the process of making available the software package that was used to generate the output provided here.

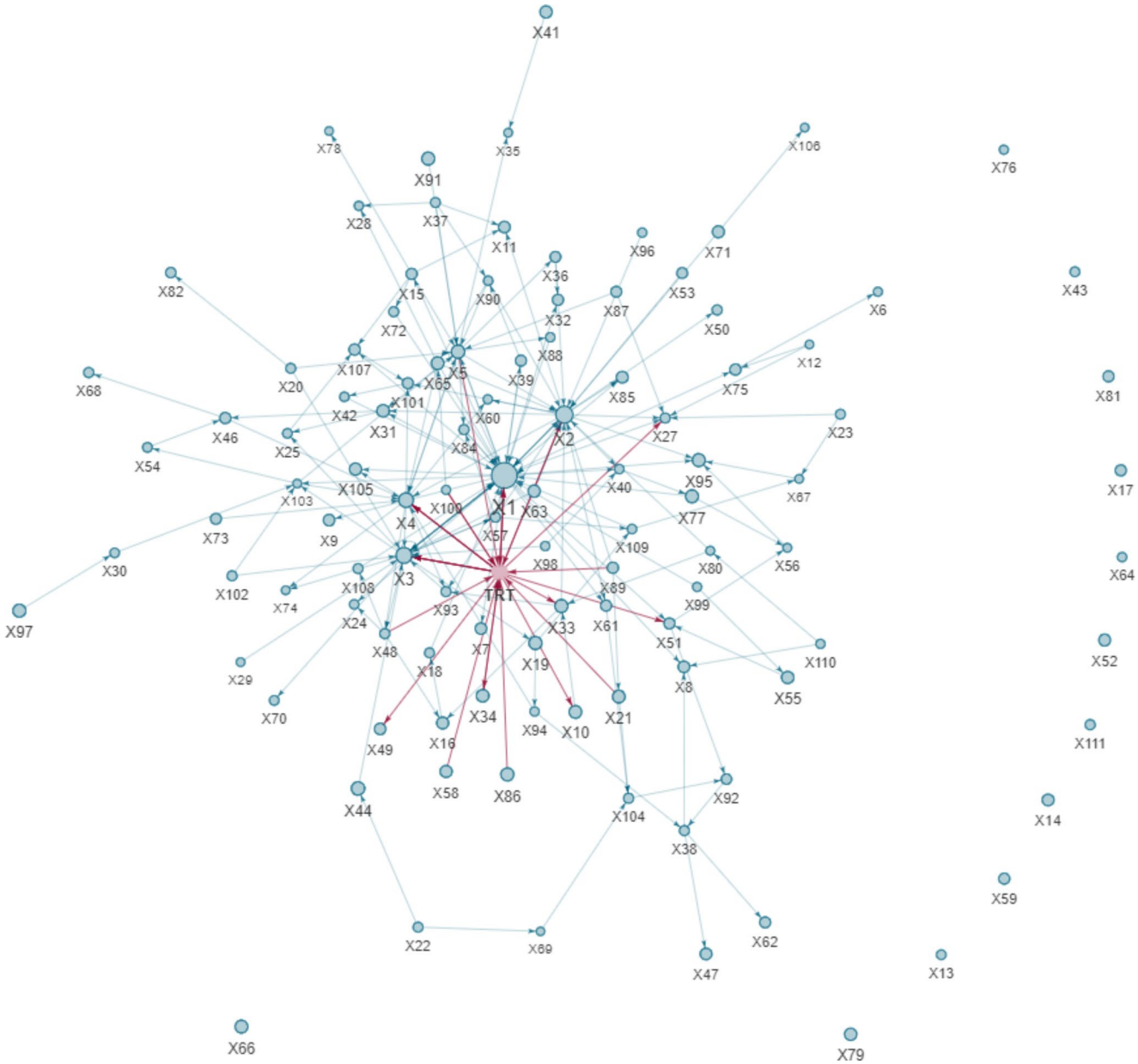


Fig. 3 System network graph deduced from the holistic RF model example case describing how variables interact to determine the patient's system journey on the trajectory to a clinical event. It is based on all detected pairwise interactions between variables. Such networks represent the best hypothesis space given the data and can be interrogated interactively to check specific hypothesis or be informed about previously unknown relationships between variables

Discussion and Recommendations

AI/ML methods follow a different concept than the conventional statistical modelling that is currently applied in late-stage clinical development and usually forms the basis for the formal decision from a statistical point of view whether a drug shall be approved or not. Despite clearly specified success criteria, however, submissions to regulatory agencies usually consist of a large amount of additional analyses that contribute to the overall body of evidence. We believe that an analysis using AI/ML methodology as described here has the potential to deepen the understanding of a compound. Our considerations show that AI/ML methods can indeed be informative for several stakeholders in the drug development process. The present paper illustrates that both regulatory and industry are heavily investing in understanding the role these concepts can play in the future, and in particular, we have depicted a use case in detail.

These methods have a lot to offer mainly in settings where a holistic analysis of all available baseline data on

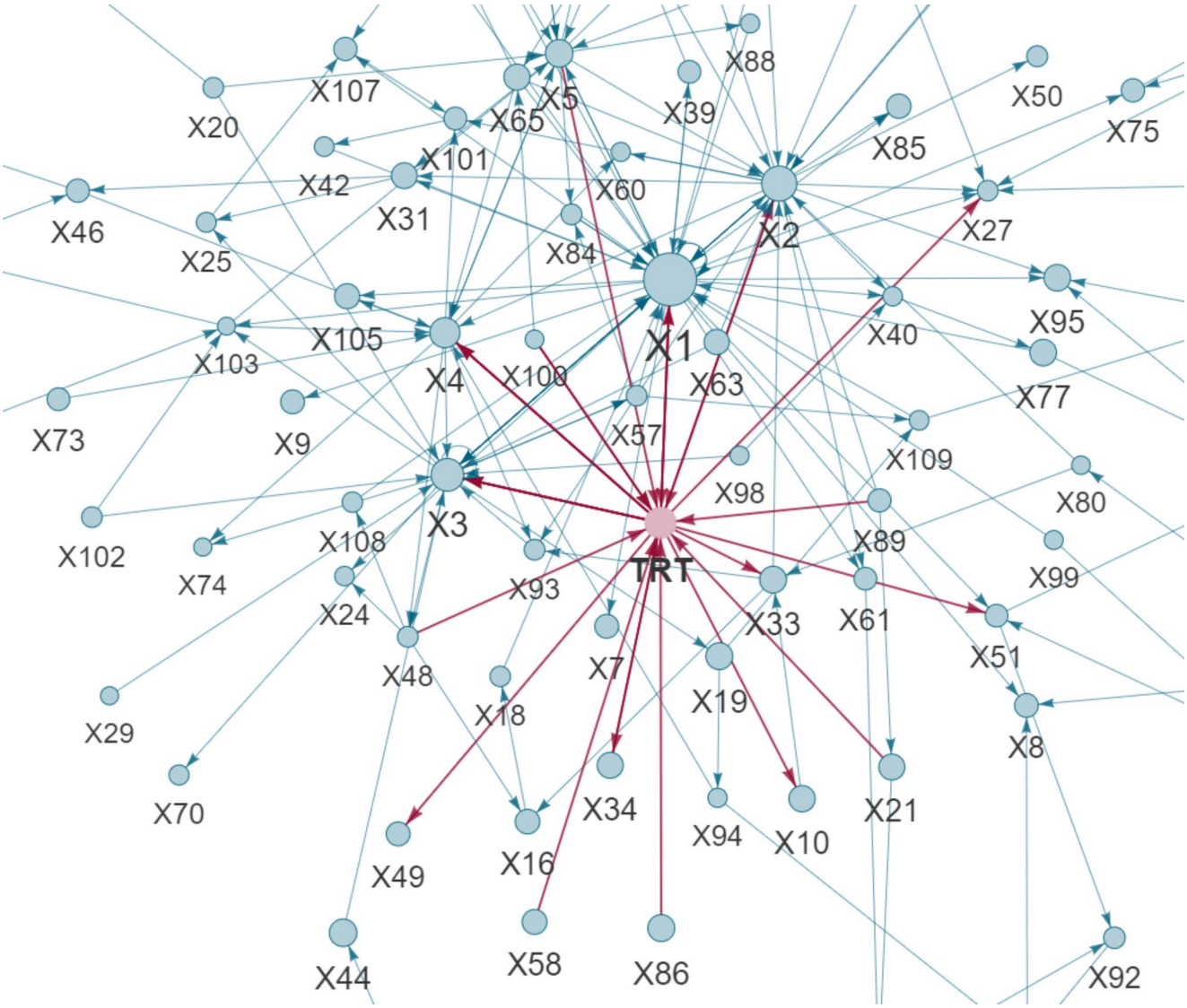
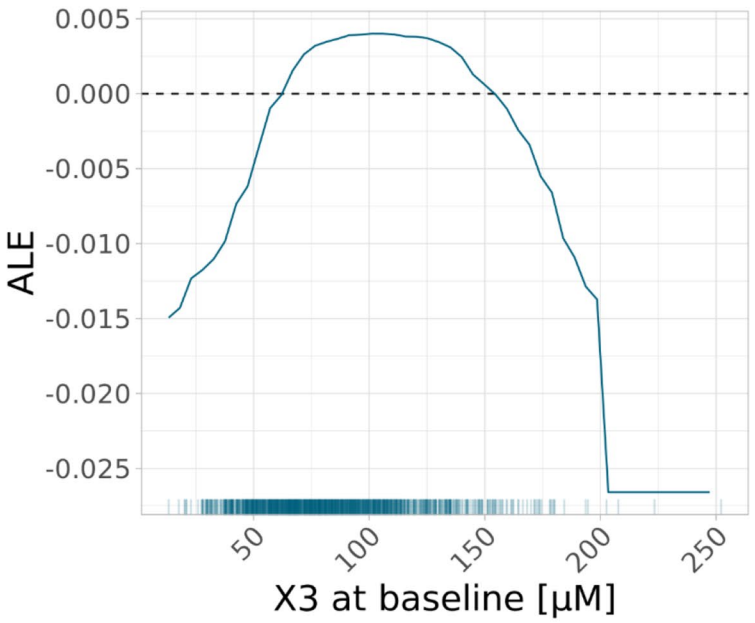


Fig. 4 Zooming into the network graph to specifically scrutinize variables that interact with treatment and may thus modify the treatment effect

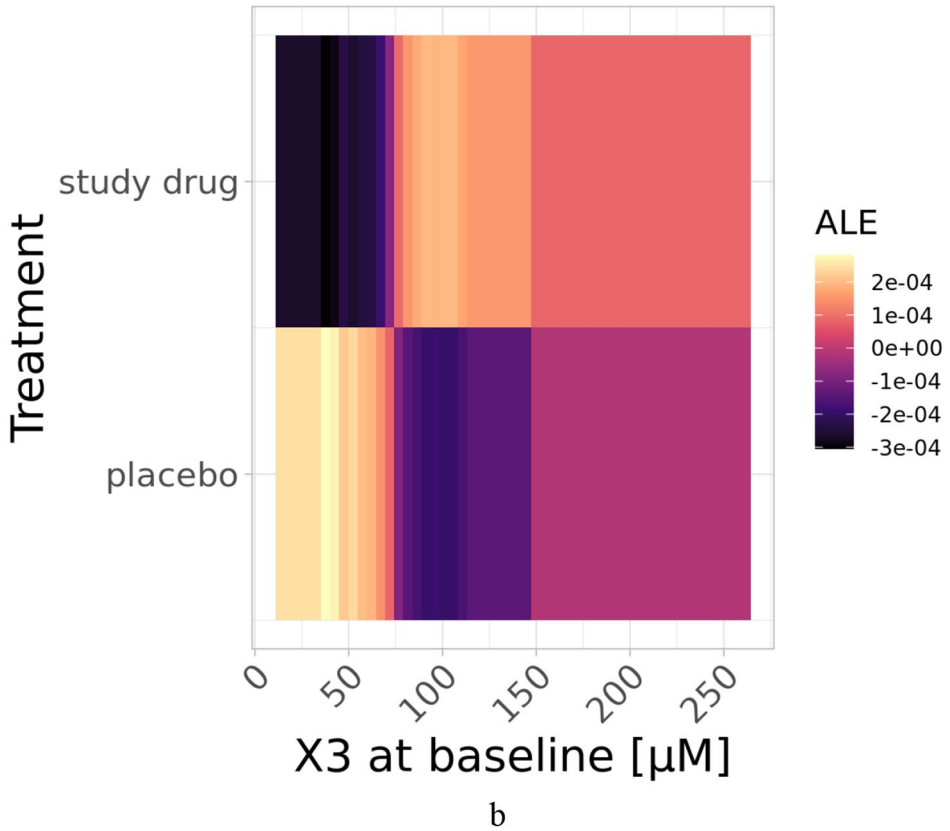
an outcome–efficacy or safety – is of particular importance. The focus of this paper was on data collected in randomized clinical trials, which by construction allows to assess the influence of an investigational treatment on an outcome. Often individual trials might be too small to investigate for instance interactions of baseline characteristics with treatment effects. Hence, data from several trials might be used. More experience is needed on how to deal with potential between-trial heterogeneity in treatment effects when combining data across trials for more complex AI/ML driven exploration [36]. Whereas in randomized controlled trials the internal validity is given by design, AI/ML concepts might be of even more importance when RWD shall form the basis of regulatory decision making or if precision medicine concepts shall be implemented in regulatory decision making or more general in settings where the external validity cannot be taken as a given.

When applying AI/ML – methods in the framework of pivotal clinical trials we especially recommend the following based on our experiences:

- **Reproducibility**: Make sure that the results obtained are reproducible.
- **Education**: Make sure that the AI/ML models used and the quantities presented are understood by all parties involved and that all parties involved have received an appropriate training in advance.
- **Standardization**:
  - Make sure that appropriate best practices are in place within the company.



a



b

Fig. 5 **a** and **b**: ALE analysis of complex functional associations of X3 (an example covariate) with a time to event clinical outcome at a specific time point, without (**a**) and with (**b**) dependence on treatment. An ALE value of 0 is equivalent to the average event probability and deviations depict the relative increase or decrease of the event probability. ALE analyses are, by virtue of the method, adjusted for all other co-variates and any potential correlation between them

  - Make sure that standards as lined out in [7, 8], and [11] are adhered to.

- **Specialization**: Consider on an organizational level to let these analyses be conducted by statisticians who are specialized on AI/ML methods and who have the necessary technical skills.

For AI/ML analyses to become a valuable part of late phase trial data analysis, considerations range from very technical aspects (e.g. software engineering and software development) in the statistical or data science area (i.e., which method to choose, how to ensure robustness etc.), to aspects of integrating AI/ML results into decision making in cross-functional development teams, and ultimately regulatory acceptance of such results as supportive evidence. To touch on all the aspects in detail was beyond the scope of this paper. In this paper, we focused on the key aspects that have turned out to be particularly important in order to establish the AI-ML concept in a current clinical development function of a pharmaceutical company. We would also like to note that we focused here on supervised AI/ML-methods as the focus of the framework we presented was on a supervised setting.

The hurdles for AI/ML-based analyses in terms of documentation and traceability are higher than for statistical methodology currently implemented in pivotal trials, which in most cases can be easily reproduced by drug regulatory bodies. Therefore, a deep understanding of the applied methodology both in terms of interpretability and transparency on both sides – regulatory and sponsor, and last but not least in some cases patients, is required. First steps on requirements for documentation and principles to apply when using these methods have been made both from regulators and as presented here also from single pharmaceutical companies.

The use case presented has not been submitted to any regulatory agency, hence we cannot definitely answer the question whether the analyses presented would improve and extend the evidence provided by the submission package but we are convinced that they add value also for regulators.

Nevertheless, concrete experiences on the regulatory side when being confronted with these analyses in a submission might help with the willingness to accept these approaches on a larger scale. A sort of mock submission might also be an option to gain experiences by both regulators and sponsors.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43441-024-00689-4.

Acknowledgements The authors were all involved in a session termed "Application of AI/ML in Late-Stage Clinical Development" at the 2022 American Statistical Association's Regulatory Industry Statistics Workshop and decided afterwards to write a paper on the topic. The authors would like to express their gratitude for the opportunity to cover the topic at this occasion.
The authors would like to acknowledge the input that further colleagues from Bayer have provided when compiling the Best Practice Document, especially Silke Janitza and Bohdana Ratich.
All authors are employed by their respective institutions and may hold stock or stock options of the pharmaceutical companies they are working for.

References

1. Friedrich S, Groß S, König, et al. „Application of Artificial Intelligence/ Machine Learning approaches in Cardiovascular Medicine: a systematic review with recommendations. Eur Heart J– Digit Health. 2021;2:424–36. https://doi.org/10.1093/ehjdh/ztab054.
2. Hyland S, Faltys M, Hüser et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nat Med 2020.
3. Frenz AK, Ahlers C, Beckert V, et al. Predicting menstrual bleeding patterns with levonorgestrel-releasing intrauterine systems. Eur J Contracept Reproductive Health Care. 2021;26:48–57. https://doi.org/10.1080/13625187.2020.1843015.
4. Rajkomar A, Dean J, Kohane I. Machine learning in Medicine. N Engl J Med. 2019;380:1347–58. https://doi.org/10.1056/nejmra1814259. https://www.nejm.org/doi/full/.
5. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65(6):386–408. https://doi.org/10.1037/h0042519.
6. Kelley H. Gradient theory of optimal flight paths. ARS J. 1960;30(10):947–54. https://doi.org/10.2514/8.5282.
7. FDA/ CDRH Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan 2021; https://www.fda.gov/media/145022/download. Last accessed December 21, 2023.
8. FDA. Using Artificial Intelligence & Machine Learning in the Development of Drug & Biological Products 2023; https://www.fda.gov/media/167973/download. Last accessed December 21, 2023.
9. FDA Artificial Intelligence & Medical Products: How CBER, CDER, CDRH, and FDA are Working Together" 2024, Last accessed May 27, 2024. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device
10. EMA „The use of Artificial. Intelligence (AI) in the medicinal product lifecycle draft reflection paper 2023, https://www.ema.europa.eu/en/use-artificial-intelligence-ai-medicinal-product-lifecycle. Last accessed May 31, 2024.
11. FDA, Health Canada, and MHRA. Good Machine Learning Practice for Medical Device Development: Guiding Principles 2021, https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles. Last accessed December 21, 2023.
12. Zhang D, Song J, Dharmarajan S et al. (2022, The Use of Machine Learning in Regulatory Drug Safety Evaluations, Statistics in Biopharmaceutical Research 2022; https://www.tandfonline.com/doi/full/https://doi.org/10.1080/19466315.2022.2108135
13. FDA. Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products 2023; https://www.fda.gov/regulatory-information/search-fda-guidance-documents/adjusting-covariates-randomized-clinical-trials-drugs-and-biological-products

14. Mehrotra DV, West RM. Survival analysis using a 5-step stratified testing and amalgamation routine (5-STAR) in randomized clinical trials. Stat Med. 2020. https://doi.org/10.1002/sim.8750.
15. Obama B. FACT SHEET: President Obama's Precision Medicine Initiative 2015;available at FACT SHEET: President Obama's Precision Medicine Initiative| whitehouse.gov (archives.gov). Last Accessed Dec 21, 2023.
16. Craig J. Complex diseases: Research and applications. Nat Educ. 2008;1(1):184.
17. Ferrari C, Sorbi S. The complexity of Alzheimer's disease: an evolving puzzle. Physiol Rev. 2021. https://doi.org/10.1152/physrev.00015.2020.
18. Myszczynska MA, Ojamies PN, Lacoste AMB, et al. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nat Reviews Neurol. 2020. https://doi.org/10.1038/s41582-020-0377-8.
19. Liu Q, Huang R, Hsieh J, et al. Landscape Analysis of the Application of Artificial Intelligence and Machine Learning in Regulatory submissions for Drug Development from 2016 to 2021. Clin Pharmacol Ther. 2022. https://doi.org/10.1002/cpt.2668. https://ascpt.onlinelibrary.wiley.com/doi/.
20. Liu Q, Zhu H, Liu C, et al. Application of machine learning in drug development and regulation: current status and future potential. Clin Pharmacol Ther. 2020. https://doi.org/10.1002/cpt.1771.
21. FDA's Sentinel Initiative. https://www.fda.gov/safety/fdas-sentinel-initiative. Last Accessed Dec 21, 2023.
22. Hein N, Rantou E, Schuette P. Comparing methods for clinical investigator site inspection selection: a comparison of site selection methods of investigators in clinical trials. J Biopharm Stat. 2019;29(5). https://doi.org/10.1080/10543406.2019.1657134.
23. Tang M, Rantou E, Schuette P. Performance of Data Mining Methods in an Example with Ordinal and Imbalanced Data Conference of Statistical Practice-American Statistical Association, February 2017, Jacksonville, FL.
24. Lautier J, Grosser S, Kim J et al. Applications of Machine Learning in Pharmacogenomics: Clustering Pharmacokinetic Concentration Curves. 2022; available under https://arxiv.org/abs/2210.13310. Last accessed December 21, 2023.
25. Wickham H, Averick M, Bryan J, et al. Welcome to the tidyverse. J Open Source Softw. 2019;4(43):1686. https://doi.org/10.21105/joss.01686.
26. Lang M, Binder M, Richter J et al. mlr3: A modern object-oriented machine learning framework in R. Journal of Open Source Software. 2019; https://joss.theoj.org/papers/10.21105/joss.01903. Last accessed December 21, 2023.
27. Walsh I, Fishman D, Garcia-Gasulla D, et al. DOME: recommendations for supervised machine learning validation in biology. Nat Methods. 2021. https://www.nature.com/articles/s41592-021-01205-4. Last accessed December 21, 2023.
28. Breiman L. Random forests. Machine Learning; 2001.
29. Douglas PK, Harris S, Yuille A, et al. Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. J Neuroimage. 2011. https://doi.org/10.1016/j.neuroimage.2010.11.002.
30. Uddin S, Khan A, Hossain E, et al. Comparing different supervised machine learning algorithms for disease prediction. Med Inf Decis Mak. 2019. https://doi.org/10.1186/s12911-019-1004-8.
31. Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer; 2009.
32. Shannon CE. A mathematical theory of communication, in The Bell System Technical Journal 1948; vol. 27, no. 3, pp. 379–423, July 1948, https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
33. Ishwaran H. Variable importance in binary regression trees and forests, Electronic J. Statist 2007; 1 519–537, 2007. https://doi.org/10.1214/07-EJS039
34. Ishwaran H, Kogalur UB, Gorodeski EZ, et al. High-dimensional variable selection for survival data. J Amer Statist Assoc. 2012. https://doi.org/10.1198/jasa.2009.tm08622.
35. Ishwaran H, Kogalur UB, Chen X, et al. Random survival forests for highdimensional data. Stat Anal Data Min. 2011. https://doi.org/10.1002/sam.10103.
36. Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. Stat Med. 2011. https://doi.org/10.1002/sim.4322.
37. Dane A, Spencer A, Rosenkranz G, Lipkovich I, Parke T, PSI/EFSPI Working Group on Subgroup Analysis. Subgroup analysis and interpretation for phase 3 confirmatory trials: white paper of the EFSPI/PSI working group on subgroup analysis. Pharm Stat. 2019. https://doi.org/10.1002/pst.1919.
38. Huber C, Benda N, Friede T. Subgroup identification in individual participant data meta-analysis using model-based recursive partitioning, Advances in Data Analysis and Classification 2022; https://link.springer.com/article/10.1007/s11634-021-00458-3
39. Molnar C. Interpretable Machine Learning - A Guide for Making Black Box Models Explainable 2022; christophm.github.io/interpretable-ml-book/