# PLOS ONE

# Designing an exploratory phase 2b platform trial in NASH with correlated, co-primary binary endpoints

**Elias Laurin Meyer**[1], **Peter Mesenbrink**[2], **Nicholas A. Di Prospero**[3], **Juan M. Pericàs**[4,5], **Ekkehard Glimm**[6,7], **Vlad Ratziu**[8], **Elena Sena**[4], **Franz König**[1] *, on behalf of the EU-PEARL NASH Investigators**¶**

1 Center for Medical Data Science, Medical University of Vienna, Vienna, Austria, 2 Novartis Pharmaceuticals Corporation, One Health Plaza, East Hanover, NJ, United States of America, 3 Janssen Research and Development, Raritan, NJ, United States of America, 4 Liver Unit, Internal Medicine Department, Vall d'Hebron University Hospital, Vall d'Hebron Institute for Research (VHIR), Barcelona, Spain, 5 Centros de Investigación Biomédica en Red Enfermedades Hepáticas y Digestivas (CIBERehd), ISCIII, Madrid, Spain, 6 Novartis Pharma AG, Basel, Switzerland, 7 Institute of Biometry and Medical Informatics, University of Magdeburg, Magdeburg, Germany, 8 Assistance Publique-Hôpitaux de Paris, Hôpital Pitie-Salpetriere, University of Paris, Paris, France

¶ Membership of The EU-PEARL NASH Investigators is provided in the Acknowledgments.
* franz.koenig@meduniwien.ac.at

## Abstract

Non-alcoholic steatohepatitis (NASH) is the progressive form of nonalcoholic fatty liver disease (NAFLD) and a disease with high unmet medical need. Platform trials provide great benefits for sponsors and trial participants in terms of accelerating drug development programs. In this article, we describe some of the activities of the EU-PEARL consortium (EU Patient-cEntric clinicAl tRial pLatforms) regarding the use of platform trials in NASH, in particular the proposed trial design, decision rules and simulation results. For a set of assumptions, we present the results of a simulation study recently discussed with two health authorities and the learnings from these meetings from a trial design perspective. Since the proposed design uses co-primary binary endpoints, we furthermore discuss the different options and practical considerations for simulating correlated binary endpoints.

## 1 Introduction

The recent years have seen unprecedented challenges for many branches of modern medical research. The desire to accelerate development and approval of new treatments has called into question some long-standing drug development paradigms, such as the strict succession of phase 1, 2 and 3 trials and the insistence on separate trials for every experimental compound [1]. Consequently, substantial effort has been made into the development of master protocol trials and in particular platform trials [2–5]. These types of trials allow evaluation of many investigational treatments in parallel and hence their implementation has increased over the last years. The interest in platform trials has increased further with the emergence of the global

pandemic due to the SARS-CoV-2 virus [6–11]. However, many operational, logistical and statistical challenges around platform trials remain.

The definition of platform trials used in this article is that they are clinical trials which investigate multiple treatments or treatment combinations in the context of a single disease, possibly within several sub-studies for different disease sub-types or targeting different trial participant populations. In a platform trial, both drugs or drug combinations within existing sub-studies, as well as new sub-studies, may enter or leave the trial over time, allowing the trial to run infinitely, in principle. Within each sub-study, many adaptive and innovative design elements may be combined that clearly separate platform trials from more classical trial designs [4]. For a more detailed introduction, we refer to [2], where we conducted a comprehensive systematic search to review current literature on master protocol trials from a design and analysis perspective. A compact glossary of common terms related to platform trials can be found in Table 1, while a more detailed list of terms and explanations can be found online [12].

Platform trials can leverage their main strengths such as adaptive design elements, testing multiple hypotheses in a single trial framework, reduced time to make decisions, ease of

**Table 1. Glossary for important terms related to platform trials, taken partly from ICH E9 [25], partly EU-PEARL D2.1 [12].**

| Term | Description |
|---|---|
| Adaptive Design | An adaptive design allows the pre-specification of flexible components to the major aspects of the trial, like the treatment arms used (dose, frequency, duration, combinations, etc.), the allocation to the different treatment arms, the eligible patient population, and the sample size. An adaptive design can learn from the accruing data what the most therapeutic doses or arms are, allowing for example, the design to home in on the best arms. |
| Integrated Research Platform | An Integrated Research Platform (IRP) is a novel clinical development concept centered on a master trial protocol which can accommodate multi-sourced interventions using the existing infrastructure of hospitals and federated patient data in design, planning and execution, while an optimized regulatory pathway for these novel treatments has been assured. |
| Master Protocol | The term "master protocol" refers to a single overarching design developed to evaluate multiple hypothesis, and the general goals are to improve efficiency and establish uniformity through standardization of procedures in the development and evaluation of different interventions. Under a common infrastructure, the master protocol may be differentiated into multiple parallel sub-studies to include standardized trial operational structures, patient recruitment and selection, data collection, analysis, and management. In a platform trial the protocol will have the infrastructure to drop interventions and allow new interventions or combinations of interventions to enter the study based on decision rules in the master protocol. |
| Platform Trials | Clinical trials which investigate multiple treatments or treatment combinations in the context of a single disease, possibly within several sub-studies for different disease sub-types or targeting different trial participant populations. For more information, see section 1. |
| Multi-center Trial | A clinical trial conducted according to a single protocol but at more than one site, and therefore, carried out by more than one investigator. |
| Frequentist Methods | Statistical methods, such as significance tests and confidence intervals, which can be interpreted in terms of the frequency of certain outcomes occurring in hypothetical repeated realisations of the same experimental situation. |
| Bayesian Methods | Approaches to data analysis that provide a posterior probability distribution for some parameter (e.g. treatment effect), derived from the observed data and a prior probability distribution for the parameter. The posterior distribution is then used as the basis for statistical inference. |
| Interim Analysis | Any analysis intended to compare treatment arms with respect to efficacy or safety at any time prior to the formal completion of a trial. |

https://doi.org/10.1371/journal.pone.0281674.t001

incorporating new investigational treatments into the ongoing trial and possibilities for collaboration between different consortia/sponsors. In 2018, the Innovative Medicines Initiative (IMI) put forth a call for proposals for the development of integrated research platforms to conduct platform trials to enable more patient-centric drug development. A consortium of 36 private and public partners have come together in a strategic partnership to deliver on the IMI proposal goals; the project is called EU Patient-cEntric clinicAl tRial pLatforms (EU-PEARL) [13]. Among the expected outputs of the initiative are publicly available master protocol templates for platform trials and four disease-specific master protocols for platform trials ready to operate in disease areas still facing high unmet clinical need; one of those diseases being non-alcoholic steatohepatitis (NASH).

NASH is a more progressive form of non-alcoholic fatty liver disease (NAFLD) and is estimated to affect approximately 5% of the world population. The disease is characterized by the accumulation of fat in the liver in the absence of significant alcohol intake or other secondary causes of hepatic steatosis [14, 15]. Over time, chronic inflammation and liver cell injury lead to fibrosis and eventually cirrhosis including complications of end-stage liver disease and hepatocellular carcinoma. Indeed, NASH complications are rapidly becoming the leading indication for liver transplantation. In addition, NASH is associated with higher risks of developing cardiovascular diseases, which is the primary cause of death for most people affected. Currently, there are no approved treatments for NASH in the US and EU and in recent years several compounds failed to meet their phase 3 primary endpoint(s) [16, 17]. However, developing treatments for NASH is a very active area of clinical research with dozens of industry-sponsored interventional studies active or recruiting trial participants across phases 1 through 3 with the vast majority in phase 1 or 2 according to ClinialTrials.gov and the EU clinical trials register (https://www.clinicaltrialsregister.eu).

To facilitate and accelerate the identification of the most effective and promising novel treatment options for trial participants with NASH, multiple potential novel therapies, as well as combinations of novel mechanisms of action, will need to be evaluated in well-designed early clinical studies before advancing to pivotal phase 3 programs. From a platform study perspective, Phase 2b is often the preferred trial design as it generally offers a robust pipeline for most indications and the ability to make decisions more rapidly before committing to longer, more costly development. This is particularly true for NASH where there is an abundance of compounds in early development and phase 3 programs tend to run over several years. Importantly, there are broadly common design elements, study populations, procedures, and endpoints for NASH phase 2b clinical studies which are aligned with Health Authority (HA) guidance.

Both the United States Food and Drug Administration (FDA) and the European Medicines Agency (EMA) have put forward advice for developing drugs for patients with non-cirrhotic NASH [18–20]. Both HAs note that the risk of progression to clinical outcomes (i.e., both liver-related and non liver-related morbidity and mortality) is mainly related to fibrosis stage. Therefore, the non-cirrhotic NASH population that should be studied are individuals with either fibrosis stage 2 (F2) or stage 3 (F3) since they are at increased risk of progression relative to those with little (F1) or no (F0) liver fibrosis [21–23]. In addition, the recognition by the HAs that the length of time necessary to observe a sufficient number of clinical events to assess drug efficacy may hamper drug development has led the HAs to recommend improvement in liver histology as clinical trial endpoints (i.e., resolution of steatohepatitis and no worsening of liver fibrosis, improvement in liver fibrosis greater than or equal to one stage with no worsening of steatohepatitis), which can be used as surrogates for approval in Phase 3 according to the accelerated approval pathways. Therefore, the FDA guidance advises that phase 2b studies demonstrate efficacy on a histological endpoint after at least 12–18 months of treatment, given

that histological change takes an extended period of time to occur using a range of doses to support phase 3 dose selection. Therefore, members of EU-PEARL are currently developing a master protocol (see Table 1) to support a phase 2b platform trial in NASH and this paper, as well as a previously published simulation study [24], describe the initiative's efforts to simulate the performance of the parameters used to make decisions on whether or not the treatment being evaluated is effective.

## 2 Materials and methods

### 2.1 Platform design

An overview of the proposed platform trial design can be found in Fig 1. Generally, it is assumed that after an initial inclusion of a certain number of cohorts each consisting of treatment and matching control, further cohorts will enter over time while some of the existing cohorts might be discontinued for efficacy or futility. Trial participants entering the platform will be allocated between open cohorts. Within open cohorts, trial participants will be equally allocated between control and treatment arm using a block randomization of length two. Finally, the platform ends when all cohorts have finished their analyses. If the inclusion and exclusion criteria of the different cohorts are similar, it might be preferable to share the accumulating information on the control treatments, at least for concurrently enrolling trial participants. While there is a lot of controversy regarding the use of non-concurrent controls [26], sharing only information on trial participants that could have been randomized to the arm under investigation seems uncontroversial (note that this requires data to be concurrent). As noted before, platform trials can run perpetually without limiting the number of drugs going into the trial. Any potentially successful compound in a NASH phase 2b trial would have to show either resolution of NASH without worsening of fibrosis (binary endpoint 1) and/or 1-stage fibrosis improvement without worsening of NASH (binary endpoint 2).

Endpoints 1 and 2 are correlated binary endpoints and clinical studies have demonstrated a strong link between histologic resolution of steatohepatitis with improvement in fibrosis [27,

## Platform Design
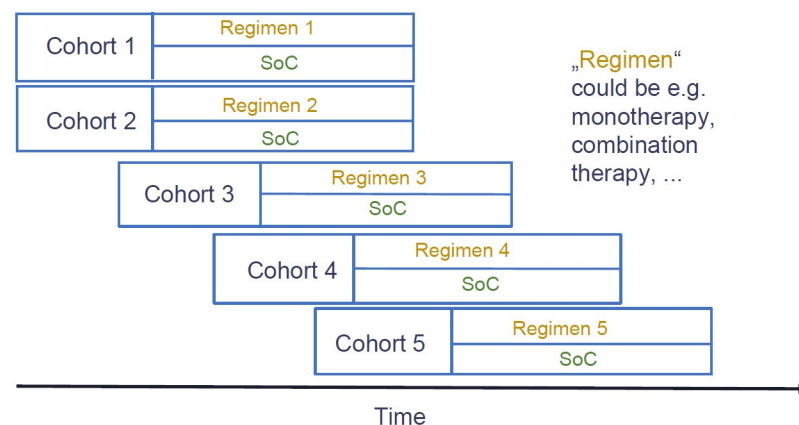


**Fig 1. Phase 2b platform trial design in non-alcoholic steatohepatitis (NASH).** After an initial inclusion of two cohorts consisting of control (usually the standard-of-care, "SOC") and "regimen" arm (which could be a monotherapy or a combination therapy), more cohorts of the same structure are entering the trial over time. Within each cohort, several interim and a final analysis are conducted using the co-primary binary endpoints "NASH resolution without worsening of fibrosis" and "Fibrosis improvement without worsening of NASH". The platform trial ends when all cohorts have been evaluated.

28], therefore, improvement in endpoint 1 could lead to improvement in endpoint 2 but not necessarily the converse and not necessarily during the same time frame. The current regulatory guidance is that the FDA recommends demonstrating endpoint 1 OR endpoint 2 and the EMA recommends demonstrating endpoint 1 AND endpoint 2 [21–23]. For this simulation study, we decided to follow FDA endpoint recommendations. Within EU-PEARL, several possible phase 2b platform trial designs for NASH were considered—treatment (one dose; could be monotherapy or combination therapy) versus control, treatment (multiple doses; could be monotherapy or combination therapy) versus control, combination therapy versus monotherapies versus control, etc. Furthermore, it was considered whether the final endpoints (which are observed after roughly 48–52 weeks) should be used for interim decision making or whether a short-term surrogate endpoint should be used. Based on the proposed design, comprehensive simulations were run for two scenarios: monotherapy (one dose) versus control and combination therapy versus monotherapies versus control. We will present results of the former in this paper and results of the latter can be found in [24, 29].

## 2.2 Decision rules

Decisions on whether or not to promote treatments to the next stage of development can be based on different principles such as fixed thresholds for treatment effect estimates, the p-values of statistical frequentist tests for treatment efficacy, conditional or predictive probabilities of final trial success. Many readers might be familiar with group-sequential trials where early stopping for futility or efficacy is based on the p-values from statistical tests which are adjusted for repeated looks into the data, such as the O'Brien-Fleming test [30, 31]. In some simple situations (e.g. if stopping the entire clinical trial for efficacy or futility is the only permitted interim decision option), it is possible to convert such decision rules into each other [32] (in the sense that a decision rule given by a threshold on conditional power can equivalently be stated by a correspondingly recalculated threshold on the estimated treatment effect, say). In platform trials, however, the decision space is usually more complicated and comprises interdependent decisions such as stopping arms without stopping the entire trial or selecting treatments if they are sufficiently superior to other treatments. In such situations, there is no simple 1-to-1 correspondence between decision rules formulated on different scales (e.g. a decision rule which is influenced by the treatment effect estimates from several treatments cannot be converted into a fixed threshold for one specific treatment). It is also very difficult to provide decision rules on un-standardized measures such as treatment effect estimates, since these would have to be derived anew for every concrete application. For these reasons, we focus on Bayesian posterior probabilities [5] as the main vehicle for making decisions in this paper. The benefit of using Bayesian decision rules is their flexibility regarding extensions to several criteria and interim analyses. To illustrate the basic mechanics, we introduce the concept for comparing the response rate of a new treatment ($\pi_E$) with the response rate of the Standard-of-Care (SoC) ($\pi_S$) in a clinical trial. For an analysis after observing data $D$, we are conducting a Bayesian analysis with the aim of deciding whether there is enough evidence to declare the treatment effective. First, we will introduce the concept for a Bayesian decision rule testing a single endpoint using a parsimonious notation for illustrative purposes. Later and in the Appendix, we will show how the parameters could be specified for the endpoints at hand in NASH. Different levels of evidence will be introduced depending on the parameterization of the Bayesian decision rule. Typically, a Bayesian decision rule of the following sort could be used for comparing the new treatment to the control SoC (the priors on $\pi_S$ and $\pi_E$ are omitted for better readability):

$$\text{Declare Efficacy, if } P(\pi_E > \pi_S + \delta | D) > \gamma \tag{1}$$

with some pre-specified probability threshold $\gamma$ and pre-defined margin $\delta$ for the targeted treatment effect of interest. Such a decision rule based on a posterior distribution can, but does not have to, correspond to a particular null hypothesis (e.g. $H_0$: $\pi_E > \pi_S + \delta$). For example, it is sometimes appropriate to use so-called "shrinkage estimators" where the single treatment effect estimates in a platform trial are "shrunk" towards a common average effect. This is appropriate if drugs share a common mechanism of action and it is therefore a priori plausible that they may have similar effects. In such situations, the decision on a single drug is influenced by the performance of the entire class of drugs. For better readability, the dependence on the data $D$ is omitted in following sections.

Decision rules should provide a high level of confidence that graduating compounds are competitive with respect to the current landscape of compounds in development with publicly accessible phase 2/3 studies published. In particular, semaglutide demonstrated a 42 percentage point response rate increase in NASH resolution (endpoint 1) as compared to placebo [33] and lanifibranor demonstrated a 19 percentage point response rate increase in fibrosis improvement (endpoint 2) as compared to placebo [34]. In the process of eliciting which exact decision rules to use, we first conducted a review of studies in NASH. Several structured discussions were held between statisticians and clinical experts to define endpoints and targeted effect sizes. Finally, the experts provided confidence intervals based on which they would accept graduation of compounds from the platform trial. Confidence intervals are generally centered around the observed response rate and the width of the confidence interval describes the remaining uncertainty, which is directly linked to the sample size, i.e. larger sample sizes lead to narrower confidence intervals. The width of the confidence intervals the experts were presented with corresponds to a sample size of 75, which is the lowest treatment group size investigated in this simulation study and corresponds to a treatment group size usually used in NASH phase 2b trials. In frequentist decision making, one might tailor the decision rules such that the confidence interval does not include a certain lower bound of efficacy. The multi-component Bayesian decision rules we propose will allow for refined specification of evidence on the efficacy of a new compound required in order to graduate, while at the same time controlling basic type 1 error with one of its decision rule components. It should also be noted that our efficacy decision rules are based on checking the criterion that there is sufficient confidence that the effect size (i.e. the difference in response rate between the experimental and control treatment) exceeds (a) certain margin(s) with certain confidence(s), i.e., the posterior probabilities for decision rules as defined in Eq 1. If the margin is selected close to the true (but unknown) effect size, there are limits for the achievable confidence, which in some situations seems counter-intuitive. As an example, if the true success rate is 0.5 and assuming a weakly informative prior, we will never be able to achieve a confidence greater than 50% that the true success rate is 0.5 or larger (for large sample sizes). The reason is that the posterior distribution will be centered around the true success rate of 0.5, i.e., resulting in a probability of maximum 50% that the value will be equal or larger then 0.5. This is equivalent to achieving 50% power to detect a success rate of 0.5 if we require a confidence greater or equal to 50%. If indeed we wanted to detect a success rate of 50% with a larger power, we need to either reduce the required confidence or the targeted success rate in our Bayesian decision rules. This is illustrated further in Fig 8 in Section 5.1.3 in Appendix, where the resulting posterior distribution for a theoretical success rate of 0.5 is shown if a weakly informative Beta (1,1) prior and a sample size of 75 participants per group are chosen and the observed success rate equals the assumed response rate. Finally, communication of the chosen decision rules to clinicians and general audiences is not always straightforward, e.g. while graduating a compound if there is 20% confidence that the effect is sufficiently large (say $\Delta$) is equivalent to dropping the compound if there is more than 80% confidence that the effect is smaller than $\Delta$, the latter is much

**Table 2. Different levels of evidence required to graduate treatment for efficacy.** The ordering is hierarchical in nature, i.e. requiring two levels of evidence means level 1 and level 2 need to be simultaneously fulfilled. E1 and E2 refer to endpoint 1 (resolution of NASH without worsening of fibrosis) and endpoint 2 (1-stage fibrosis improvement without worsening of NASH) respectively.

| Level of Evidence $l$ | Margin $\delta$ for targeted difference | Confidence $\gamma$ required | Description |
|---|---|---|---|
| 1 | 0 (both endpoints) | 95% | First level of efficacy evidence required serving as a threshold to establish sufficient confidence in any treatment effect larger than 0, i.e. superiority of a treatment to control. Note that when using non-informative priors, the one-sided frequentist type I error will be about $1 - \gamma$ for a single endpoint, i.e. 5% in this example. |
| 2 | 0.30 (E1) 0.175 (E2) | 85% | Second level of efficacy evidence required serving as a threshold to establish a high confidence that the true effect is larger than moderate treatment effects. In the example a higher margin for the first endpoint is required compared to the second endpoint. |
| 3 | 0.40 (E1) 0.25 (E2) | 60% | Third level of efficacy evidence required serving as a threshold to establish sufficient confidence in large treatment effects. |

more generally understood. For the lack of a comparable frequentist design, no direct comparison of operating characteristics was conducted.

We propose a Bayesian framework for a multi-level efficacy decision rule which incorporates different levels of evidence, ranging from information whether the treatment is simply superior to the control up to information on how likely larger effect sizes of interest are. A specification for such multi-level efficacy decision rules using three levels of evidence can be found for both endpoints of interest in NASH in Table 2 and Fig 2. At level 1, the main target is to show whether the experimental treatment is superior to the control by setting the margin $\delta_1 = 0$. To ensure sufficient type 1 error control (assuming weakly informative priors on the success rates) the required confidence is set to $\gamma_1 = 0.95$. At level 2, there should be sufficient
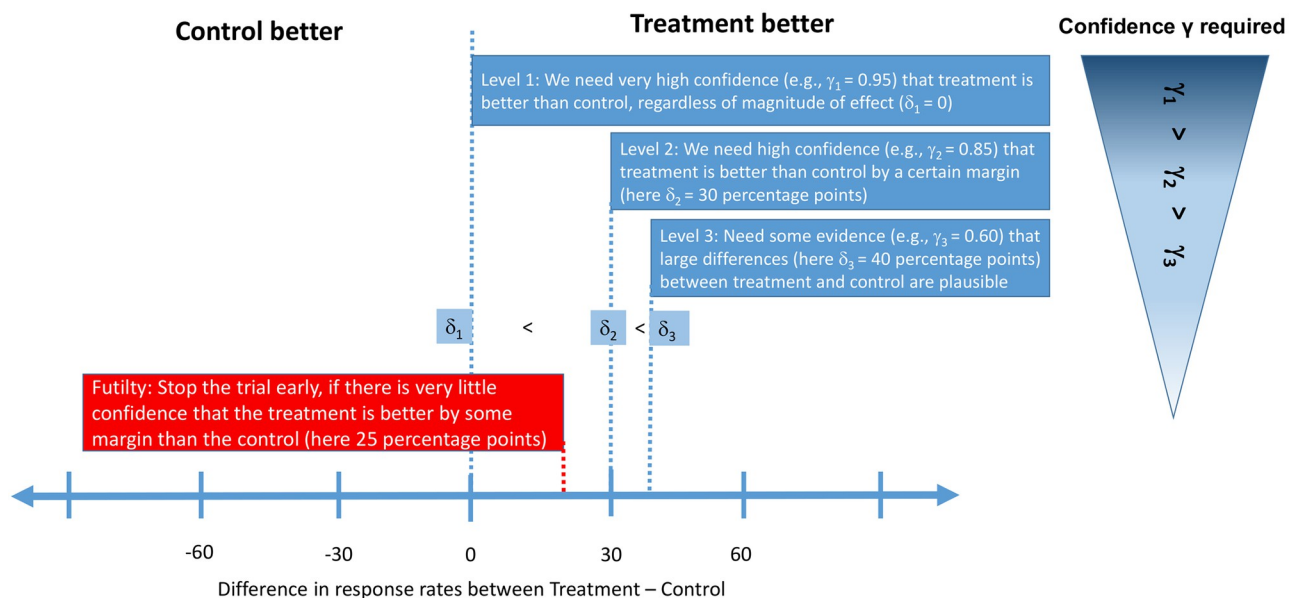


**Fig 2. Schematic overview of decision rules used.** On the x-axis, the difference in response rates between the control and treatment group (i.e. treatment effect) in percentage points is shown. At the two interim analyses, cohorts can be stopped early for futility, if there is very little evidence (interim analysis 1: less than 20%, interim analysis 2: less than 30%) that the treatment is better than control by 25 percentage points or more (red box). At all analysis time points, the same efficacy decision rules are used (blue boxes). Depending on the aim of the study, all or only certain levels of evidence could be required (see also Table 2). The treatment effects ($\delta$s) presented in this figure correspond to the decision rules used for endpoint 1— for endpoint 2, we used $\delta_1 = 0$, $\delta_2 = 0.175$, $\delta_3 = 0.25$, as well as a futility margin of 10 percentage points.

evidence provided that the true effect size is larger than moderate differences with a certain level of confidence. For example, we set $\delta_2$ for each endpoint to an effect size which we elicited should be close to the lower end of any 95% confidence interval based on which a treatment would be graduated and the required confidence, $\gamma_2$, to see an effect size at least as large as this to 85% (in accordance with our considerations regarding the confidence interval). Level 3 requires that there is also sufficient confidence in observing larger effects. For example we set $\delta_3$ for each endpoint to an effect size which we elicited should be slightly below the center of any 95% confidence interval based on which a treatment would be graduated and the required confidence, $\gamma_3$, to see an effect size at least as large is this to 60% (again in accordance with our considerations regarding the confidence interval). Of course any such ordering of $\delta$s and $\gamma$s should fulfill the conditions $\delta_1 < \delta_2 < \delta_3$ and $\gamma_1 > \gamma_2 > \gamma_3$ to be meaningful. To motivate the multi-level decision rules, we simulated all scenarios using the different levels of required evidence. Please note that while $\delta_1 < \delta_2 < \delta_3$ and $\gamma_1 > \gamma_2 > \gamma_3$, this does not mean that level 3 decision rules are generally "stricter" than level 2 or level 1 decision rules. In fact, if the posterior was extremely flat, the level 1 requirement would be the strongest.

In addition to graduating a treatment based on (possibly multi-level) Bayesian decision rules, one might also be interested in dropping a treatment at an interim analysis for futility based on posterior probabilities. In this case, the Bayesian decision rule for a single endpoint can be expanded introducing margins $\delta_E^T$ and $\delta_F^T$ with thresholds $\gamma_E^T$ and $\gamma_F^T$ for graduating and dropping, respectively. At a given point in time $T$, after observing data $D$, we are conducting an analysis with the aim of deciding whether we have enough evidence to declare the treatment efficacious or futile.

$$\begin{aligned}
&\text{Declare Efficacy, if } P(\pi_E > \pi_S + \delta_E^T | D) > \gamma_E^T \\
&\text{Declare Futility, if } P(\pi_E > \pi_S + \delta_F^T | D) < \gamma_F^T \\
&\text{Continue Trial, otherwise}
\end{aligned} \qquad (2)$$

with some pre-specified probability thresholds $\gamma_E^T$ and $\gamma_F^T$ and required treatment effects $\delta_E^T$ and $\delta_F^T$. Since we decided to follow FDA requirements, a treatment will be graduated if the efficacy decision rules are met for at least one of the two endpoints (this will be referred to as the "OR" decision rule), while it will only be dropped at interim for futility, if the futility decision rules are met for both of the endpoints. The specification of the efficacy decision rules for both endpoints is given in Table 2, whereby the same margins $\delta$ and confidence $\gamma$ are used for the interim and final analyses. The trial will be stopped for futility, if there is only a small likelihood that the response rates of the treatment arm exceeds the control arm by at least 25 and 10 percentage points for endpoint 1 and 2, respectively. For endpoint 1, both efficacy and futility decision rules are illustrated in Fig 2. For more information, including a verbal description of the decision rules and a more formal definition, please refer to Section 5.1 in Appendix.

## 3 Simulations

### 3.1 Simulation setup

For classical randomized-controlled trials (RCTs) or even some multi-arm, multi-stage trials, the required sample size to achieve a certain power might be a deterministic function of several assumptions and design parameters such as treatment effects, significance level and chosen test procedure. Due to the complex designs, platform trials usually require simulations to be run in order to calculate operating characteristics such as power and average trial duration. In order to simulate the EU-PEARL NASH phase 2b platform trial, we chose a set of assumptions and design choices that were fixed and a set of assumptions and design choices that were

**Table 3. Specification of important simulation parameters.** Values are either fixed or varied in different simulation scenarios. For different simulation parameters, we differentiate between parameters that are considered a design choice ("D") and parameters that are considered an assumption ("A") regarding the future course of the platform trial or treatment effects (see second column "Type").

| Name | Type | Investigated Values | Description |
|---|---|---|---|
| Timing of new cohorts | A | 24 | Number of weeks after which a new treatment enters the platform trial. |
| Accrual rate | A | 6 | Number of participants entering the trial per week (approximation based on the number of participating centers and trial participants per center). |
| SoC responder rates | A | 10% (E1) 20% (E2) | Success rates for endpoint 1 (E1) and endpoint 2 (E2) in the standard-of-care (SoC) arm. |
| Endpoint 1 Responder Rate | A | Range from 0.10 to 0.55 | Assumed responder rate of the investigational treatments for endpoint 1 (NASH resolution without worsening of fibrosis). |
| Endpoint 2 Responder Rate | A | Range from 0.20 to 0.55 | Assumed responder rate of the investigational treatments for endpoint 2 (fibrosis improvement without worsening of NASH). |
| Time trend | A | 0 | Assumed drift in the outcome responder rates over time. Simulations were conducted assuming no such drift. |
| Correlation between endpoint 1 and endpoint 2 | A | -0.3, 0, 0.3, 0.7 | Assumed correlation between the latent continuous analogues to endpoint 1 and endpoint 2 (see section 5.2.4 in Appendix for more details). While we assume the true correlation to be positive, a negative value was added in the simulation study in order to investigate a larger range of values. |
| Initial cohorts | D | 2 | Number of cohorts the platform trial is initiated with. |
| Cohort limit | D | 5 | Maximum number of cohorts that can enter the platform trial over time. |
| Outcome observation time | D | 52 | Number of weeks after enrollment at which the primary outcome is observed. |
| Interim Analyses Timing | D | 50% (IA1) 75% (IA2) | Timings of interim analyses relative to final planned sample size (counting observed outcomes). |
| Final Cohort Sample Size | D | 150, 250 | Number of trial participants after which final analysis in a cohort is conducted. These numbers correspond to sample sizes usually used in NASH phase 2b trials (i.e. 75/125 per arm). |
| Data Sharing | D | concurrent, cohort | Different methods of data sharing used at analyses, either using concurrent data ("concurrent") or not sharing at all ("cohort"). |
| Decision Rule | D | Specific rules | See section 2.2, as well as Table 2 and Fig 2 for more details on the decision rules used. For a formal definition, see Section 5.1.2 in Appendix. |
| Evidence Level | D | 1,2,3 | Different levels of evidence required in the Bayesian decision rules (see section 2.2). By default, the highest level of evidence is required (i.e. level 3). |

https://doi.org/10.1371/journal.pone.0281674.t003

varied (see Table 3). In general, for every trial participant we observe two correlated binary outcomes, NASH resolution (Endpoint 1) and fibrosis improvement (Endpoint 2). Outcomes are simulated using the approach described in section 5.2.4 in Appendix, such that in the simulations we can fix the success rates of endpoint 1 and endpoint 2 and their latent variable correlation $\rho$ (see Fig 3). The correlation between the two endpoints can be interpreted in such a way that if there is a positive correlation between the two endpoints, then there is an increased likelihood that either events are observed in both endpoints jointly or not at all. If there is a negative correlation, it means if an event is observed in one endpoint, then there is a larger likelihood that no event is observed for the other one. If the two endpoints are uncorrelated ($\rho = 0$), knowing if an event was observed for one endpoint gives no information as to whether or not an event is observed for the second one. Sample sizes reflect number of trial participants with complete observations (i.e. paired biopsies) by the time of final analysis. After this number of trial participants were enrolled in a given treatment arm, enrollment to this treatment arm stops.

Simulations of this trial design were performed using the **cats** package, which is downloadable on Github (https://github.com/el-meyer/cats) and CRAN (https://cloud.r-project.org/web/packages/cats/index.html) and validated using the **simple** package (https://github.com/el-meyer/simple). For each of the distinct combinations of simulation parameters the platform trial was simulated 10000 times. Results of those 10000 simulated platform trial trajectories
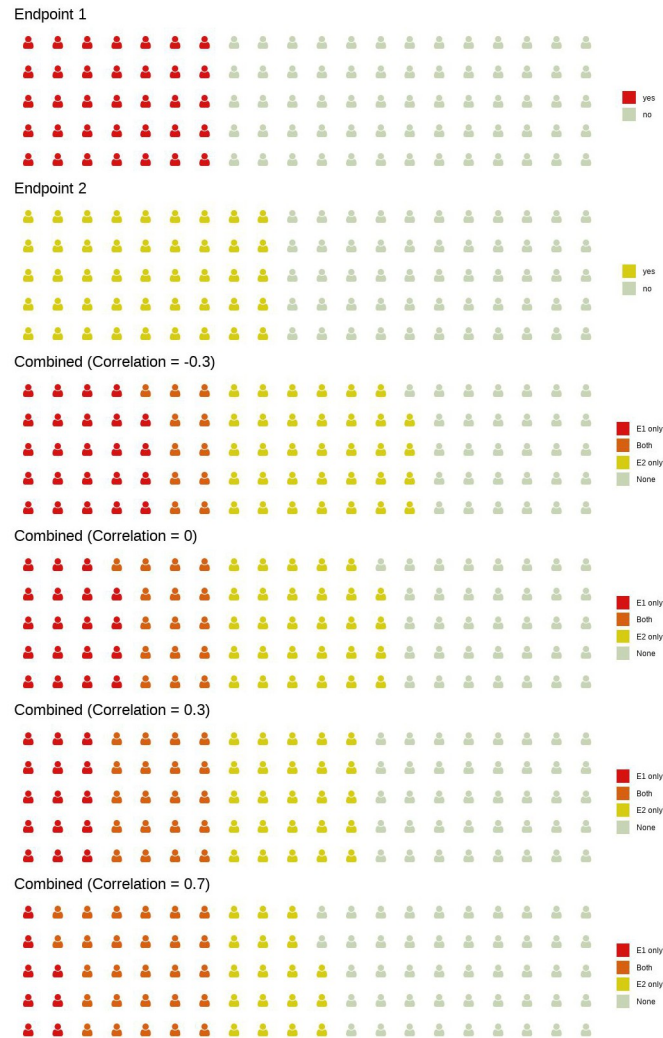
**Fig 3. Impact of different levels of correlation between endpoints 1 and 2 on the expected number of responders.** Assuming a response rate of 30% for endpoint 1 and 40% for endpoint 2, we expect 30/100 patients to reach endpoint 1 (in red) and 40/100 patients to reach endpoint 2 (in yellow), regardless of the correlation. Depending on different levels of the correlation, the number of responders that reach both endpoints simultaneously (in orange) varies; it increases with increasing correlation. In contrast, the expected number of trial participants that reach at least on of the two endpoints decreases with increasing correlation (in this example, this number is 69 when the correlation is -0.3 and 53 when the correlation is 0.7).

https://doi.org/10.1371/journal.pone.0281674.g003

were summarized for each of the sets of simulation parameters and visualized using lattice plots [35]. In particular, we present the success probability (i.e. the probability for a particular drug to be declared superior to control; this is equivalent to type 1 error when the drug is in truth futile and power when the drug is in truth efficacious).

## 3.2 Main simulation results

Trial success probabilities with respect to the chosen simulation parameters are shown in Fig 4. It becomes apparent that when the drug under investigation is not effective for either of the endpoints (i.e. responder rate of 10% for endpoint 1 and 25% for endpoint 2), the success
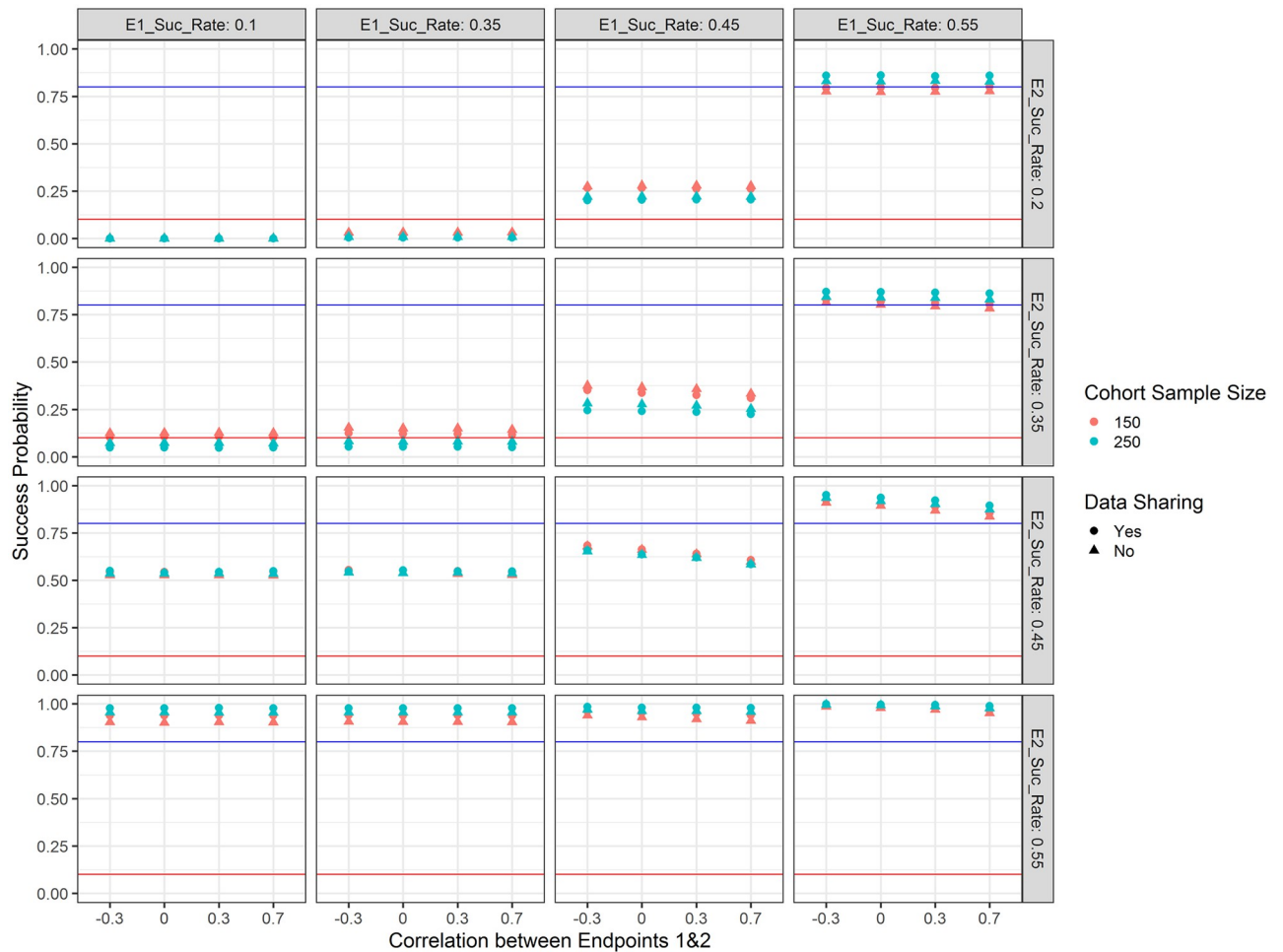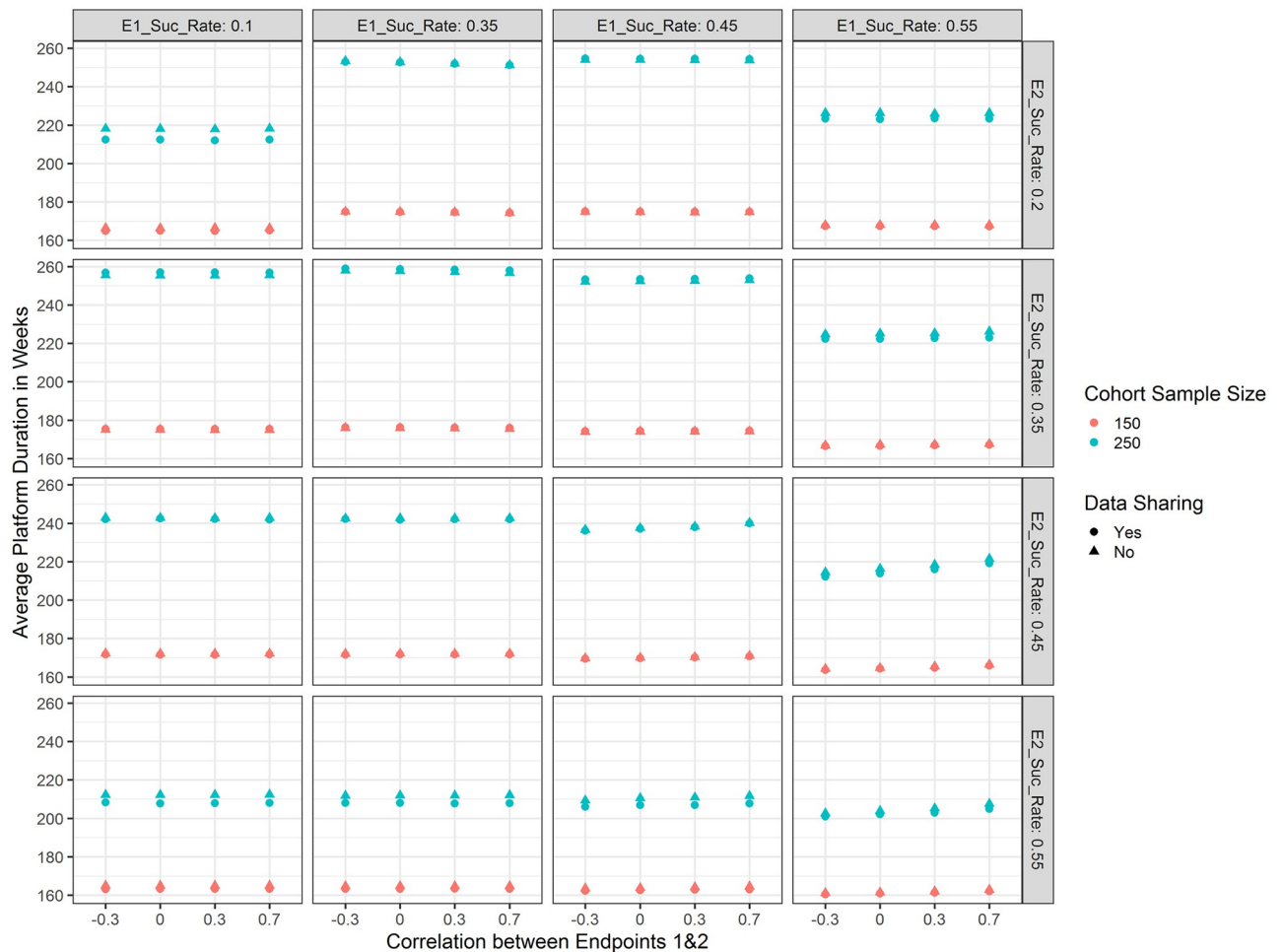
**Fig 4. Success probabilities for the treatment arm with respect to the response rate for endpoint 1 (E1_Suc_Rate; columns), the response rate for endpoint 2 (E2_Suc_Rate; rows), the correlation between the two endpoints (x-axis), the type of data sharing used (point shape) and the planned cohort sample size per arm (colour).** The blue horizontal line marks 80% as a common target for the power and the red horizontal line marks 10% as a common target for type 1 error in early phase clinical trials. When the drug is truly effective, success probabilities correspond to power; when the drug is not effective, success probabilities correspond to type 1 error.

https://doi.org/10.1371/journal.pone.0281674.g004

probability (i.e. in this case type 1 error) is negligible (about 0.1%) regardless of the sample size. On the other hand, when the drug is highly efficacious for either or both of the endpoints (i.e. responder rate of 55% for both endpoints), the success probability (i.e. in this case power) is close to 1. When the treatment under investigation exhibits a responder rate of 35% for one or both endpoints, the success probability is below 20% and in case of a larger sample size (125 per arm) below 10%. When the treatment under investigation is promisingly efficacious on both endpoints (i.e. responder rate of 45%), the success probability is between 60–70%, depending on the sample size and correlation. In terms of data sharing we observe the same pattern as with increased sample size—when the treatment is highly efficacious, the success probability increases, otherwise it decreases—this is a feature of the Bayesian decision rules (in frequentist analyses we would expect the type 1 error to be the same). As an example, when the sample size is 125 per arm, there is no correlation between the two endpoints and the success rate is 35% for both endpoints, sharing data reduces the success probability (in this case corresponding to type 1 error) to 5%, while the success probability is 8% when not sharing data. In

general, we see a difference in success probabilities when only one of the two endpoints is reached—the success probability is higher if the treatment is efficacious on the Fibrosis endpoint. This is intended and consistent with our assumption that NASH resolution leads to Fibrosis improvement. Higher correlations between the two endpoints lead to reduced success probabilities—this is explained via the "OR" decision rule and the fact that outcomes are simulated using a latent bivariate normal distribution. This effect is most pronounced when the treatment is moderately efficacious for one or both of the endpoints (i.e. responder rate of 45%).

Average platform trial durations (i.e. time until a decision is made for the last investigational treatment) are shown in Fig 5. It becomes apparent that when the treatment is weakly efficacious for either or both of the endpoints (i.e. 35% responder rate), trial duration is the longest (since it is unlikely that the treatment will be stopped for efficacy or futility at any of the interim analyses). For similar reasons, trial duration is decreased when the treatment is efficacious for neither of the endpoints and shortest when the treatment is highly efficacious for both of the endpoints. Generally, sharing concurrent data can lead to savings in trial duration of at most 2 weeks (sample size per arm 75; bottom left panel in Fig 5; 163 vs 165 weeks)
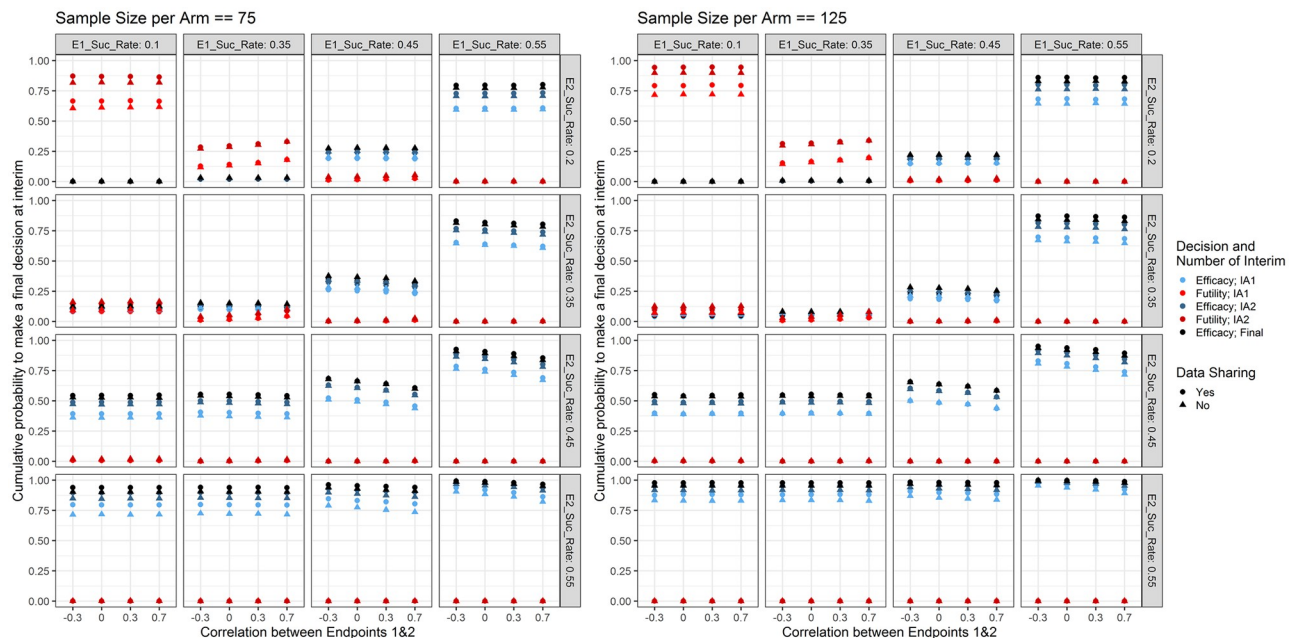


**Fig 5. Average platform trial duration in weeks with respect to the response rate for endpoint 1 (E1_Suc_Rate; columns), the response rate for endpoint 2 (E2_Suc_Rate; rows), the correlation between the two endpoints (x-axis), the type of data sharing used (point shape) and the planned cohort sample size per arm (colour).**

or 6 weeks (sample size per arm 150; top left panel in Fig 5; 212 vs 218 weeks) compared to not sharing data (please note these numbers are also influenced by our assumption that new treatments would enter the platform every 24 weeks and the platform would necessarily run until five treatments are evaluated). Please note that since there is a lag of observing the final endpoints of 52 weeks, even if a decision is made early, this might not translate to savings in trial participants. In case the sample size per arm is 75, we observed no savings in terms of trial participants enrolled (i.e. always 750 trial participants are enrolled in the course of the trial). This is due to the assumed recruitment rate, which would lead to full recruitment in the time frame before the first interim analysis (a slower recruitment rate or more treatment arms investigated simultaneously might lead to savings). When the sample size per arm is 150, analogously to average trial duration, we observed savings in trial participants when treatments are overwhelmingly efficacious or futile (in the most extreme case of overwhelming efficacy on both endpoints and sharing data, approximately 382 out of 1500 trial participants were saved, compared to if no early stopping rules had been in place).

Probabilities of stopping early with respect to treatment efficacy are shown in Fig 6. In case the treatment is not better than placebo, the futility rules eliminate approximately 60% of treatments at the first interim analysis and approximately 80% by the second interim analysis. This is true when the sample size per treatment arm is 75 and the probabilities increase further with increased sample size, i.e. 125. When the treatment is highly efficacious on both endpoints, the efficacy decision rules graduate most of the treatments at the first or second interim analysis. We also see that the futility interim decision rules eliminate more treatments that are only weakly effective on endpoint 1 and ineffective on endpoint 2 than if the reverse was true (this is a result of identical futility stopping rules while the standard-of-care response rate is lower for endpoint 1 than for endpoint 2).



**Fig 6. Cumulative probabilities to make a decision early** (i.e. making an early efficacy or futility decision either at the first or second interim analysis) with respect to the response rate for endpoint 1 (E1_Suc_Rate; columns), the response rate for endpoint 2 (E2_Suc_Rate; rows), the correlation between the two endpoints (x-axis), the type of data sharing used (point shape) and the planned sample size per cohort (left panel 150 and right panel 250).

https://doi.org/10.1371/journal.pone.0281674.g006
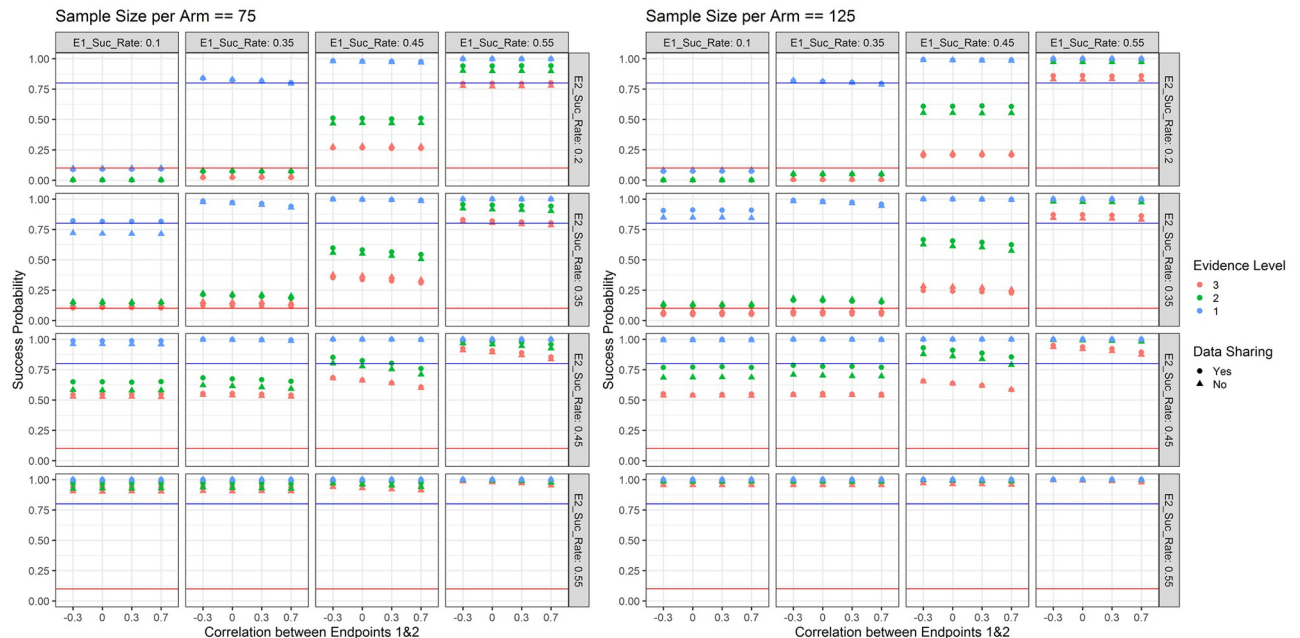
**Fig 7. Success probabilities for the treatment arm with respect to the response rate for endpoint 1 (E1_Suc_Rate; columns), the response rate for endpoint 2 (E2_Suc_Rate; rows), the correlation between the two endpoints (x-axis), the type of data sharing used (point shape), the level of evidence required (colour) and the planned sample size per cohort (left panel 150 and right panel 250).** The blue horizontal line marks 80% as a common target for the power and the red horizontal line marks 10% as a common target for type 1 error in early phase clinical trials. Level of evidence required refers to how many of the Bayesian efficacy rules specified in section 2.2 need to simultaneously hold for a treatment to be declared efficacious.

## 3.3 Impact of Bayesian multi-level decision rules

Trial success probabilities with respect to the chosen simulation parameters as well as the level of evidence required in the Bayesian decision rules (see section 2.2) are shown in Fig 7. It becomes apparent that when requiring only the lowest level of evidence (evidence level 1), type 1 error is controlled and for all investigated effect sizes there is a large success probability (which—depending on the target product profile—might be desired or not). When requiring a second level of evidence, success probabilities for effect sizes identified as insufficiently promising in our decision rule drop significantly, while success probabilities for large effect sizes stay large. When requiring all three levels of evidence, only treatments with a very large effect size are advanced with a high probability.

## 4 Discussion

For this exploratory phase 2b platform trial in NASH, a Bayesian framework has been chosen to incorporate the information from two endpoints (resolution of NASH without worsening of fibrosis (endpoint 1) and/or 1-stage fibrosis improvement without worsening of NASH (endpoint 2)) in the Bayesian decision rule. Based on the regulatory requirements, it has been decided that for the success criteria it is sufficient to demonstrate efficacy in either of the two endpoints. However, based on emerging phase 2b and phase 3 data for compounds under development, it is not sufficient to simply show superiority, instead there should be sufficient evidence that the effect sizes are large enough to show differentiation in order to graduate a treatment from phase 2b to phase 3. Therefore, the Bayesian decision rules have been extended to allow for different levels of evidence. Firstly, a high confidence is needed that the experimental treatment is better than control by any margin. Secondly, high confidence is required

that the true effect is at least of moderate effect. Finally, some evidence is required that the true effect is relatively large and competitive with respect to the current landscape of compounds in the development pipeline.

In frequentist trials, a study is powered to show superiority. The assessment whether the observed effect is relevant would be deferred to the lower bound of the confidence interval of the observed effect. Such a strategy could also be translated into a shifted hypothesis test using the duality between confidence intervals and frequentist tests. The proposed Bayesian framework allows a convenient way of combining the evidence from both endpoints and the extension to futility stopping rules in interim analyses. In this simulation study, we proposed that in order to declare success it is sufficient to demonstrate efficacy in either of the two endpoints, whereas both endpoints have to show insufficient efficacy to declare futility. In a frequentist trial, further multiplicity adjustments would have been needed for both repeated significance testing and testing two endpoints. We also investigated different ways in which two correlated binary endpoints can be simulated. The investigation of different correlations is critical for the trial's probability of success. Due to using an "OR" criterion (i.e. the treatment is graduated if either of the endpoints are met), a higher correlation will lead to decreased probabilities. Therefore, for planning purposes and determining sample sizes, it might be preferable to assume larger correlations, but ultimately this also depends on the chosen method to generate correlated binary endpoints.

A draft of the proposed platform trial protocol including some simulation results were discussed with FDA in a critical path innovation (CPIM) meeting in January 2022 [36]. There were no objections to using Bayesian decision rules and it was acknowledged that in this phase 2b setting there is no need to correct for multiplicity on a platform level resulting from testing several compounds in the same trial. Regarding sharing of data, FDA supported the idea of using concurrent control data (defined as data of trial participants who were randomized to control arm in another cohort while randomization is ongoing in the cohort of interest who meet the inclusion/exclusion for all cohorts), but was opposed to using non-concurrent control data. Similar responses were received when discussing this NASH platform trial design with EMA in an Innovation Task Force (ITF) meeting in November 2022. Within EU-PEARL, these results currently serve as a discussion foundation on how to choose the sample size and decision rules for the platform trial protocol as it seems that there is still some latitude with respect to type 1 error, especially considering that this is planned as a phase 2b design for decision making and not for registration purposes.

One of the main advantages of platform trials is that they reduce the time and number of trial participants required to make a decision [2]. This is usually achieved both by operational and statistical efficiencies, such as multiple interim analyses and sharing data concurrently/non-concurrently across investigational cohorts. The impact of the interim analyses in reducing the duration of study and/or the number of participants will be lessened when the recruitment rate is fast relative to the time needed to observe the final endpoints (for example, if all of the participants needed for analysis in an investigational cohort are recruited in 3–6 months and the interim and final analyses are conducted at 12 months, there will be no savings in the number of participants entering the platform trial). This could lead to a large number of participants who have been randomized into a cohort in the trial, but have not yet had their primary endpoint observed when a decision is made to stop the cohort either due to superior or futile efficacy [37]. This problem becomes most evident in the trial design investigated here, when the sample size per treatmen arm is 75 with an accrual rate of 6 participants/week, even if a decision is made at one of the interim analyses to stop the cohort for superior efficacy, potentially there will be regulatory interest in seeing the treatment effect of the full cohort to provide further demonstration of the robustness of the efficacy observed at the interim

analysis. If futility was demonstrated, then ethically you could prevent those randomized who have not reached week 52 from having an additional invasive liver biopsy. Therefore, initiatives to establish validated short-term endpoints in NASH based on biomarkers are critical. Savings in trial participants might be more pronounced when the recruitment rate is slower and/or more treatments are evaluated at the same time and or the sample size is larger. When the effect size is large, we observed time savings of around six weeks and up to 20–25% reduction in the number of required participants when assuming 125 participants per treatment group and overwhelmingly superior or futile efficacy is observed along with the use of a concurrent control, indicating that even under the simple trial simulation assumptions studied the savings can be achieved under the ideal conditions. Further work is needed to determine how a Phase 2b NASH platform trial can achieve greater efficiency in the number of participants that need to be evaluated while making the best decisions in advancing those investigational treatments that have the potential of demonstrating transformative efficacy.

Many extensions of this simulation study can be considered. First, more simulation parameters could be investigated—especially with respect to a range of accrual rates we would expect the savings in trial duration and participants to show meaningful change. The time between new treatments entering was set to 24 weeks and the maximum number of cohorts to 5, with two cohorts starting initially. If these assumptions were changed such that either more or less cohorts would be enrolling concurrently, differences in trial duration and success probabilities might be more pronounced for a concurrent control versus using cohort control groups only. Also, the use of non-concurrent controls could be investigated. We observed no significant changes in success probabilities when the sample size was increased beyond 75 trial participants per arm—therefore we believe no larger sample sizes are warranted based on the simulation parameters that have been evaluated to date. So far, it is assumed that trial participants are equally randomized between open cohorts and within cohorts between treatment arms. The use of response-adaptive-randomization might allow effective treatments to graduate faster— we did not investigate its use further, because we assumed identical treatment effects for all treatments within one platform trial. It was assumed that it is enough to show efficacy on one of the endpoints (i.e. "OR" decision rule). Future research could aim to show efficacy on both endpoints (i.e. "AND" decision rule). Both co-primary endpoints proposed for the this phase 2b platform trial have been discussed and agreed on by regulatory agencies, i.e. at an ITF meeting with EMA in November 2022 and a CPIM meeting with FDA in January 2022. While the resolution of steatosis or fibrosis improvement are clearly important endpoints, it may be scientifically interesting to evaluate disease progression to NASH cirrhosis. This might disclose that treatment regimens are unable to improve baseline condition, but possibly able to prevent disease progression. The only progression endpoint that is of regulatory interest is the one showing a progression from F2/F3 to F4. However, this is usually part of the clinical outcome composite endpoint evaluated in most NASH phase 3 trials and it would require longer follow-up than 12–18 months that is used in most Phase 2b NASH clinical trials. So when simulating phase 3 designs for NASH it should also incorporate the cumulative incidence of important clinical outcomes [38, 39] as the baseline risk and include clinically relevant outcomes such as prevention of cirrhosis. However, the design of a phase III platform trial goes beyond the scope of this paper. The proposed platform trial is designed to incorporate phase 2b trials, but not phase 3 trials.

To conclude, we have found, based on our assumptions, that a NASH phase 2b platform study design employing Bayesian decision rules can demonstrate some time efficiencies when the effect sizes are large for either primary (histologic) endpoint, which is consistent with accelerating the development of transformational therapies. These time efficiencies would be in addition to those offered by a platform study in terms of accelerating start-up activities

thereby creating a favorable proposition for drug developers, especially small biotechs. It is possible that once short-term (i.e., 12–24 week) biomarkers (alone or in combination) have sufficient data to predict efficacy for histological and/or clinical outcomes, the use of a platform design may become even more powerful using a phase 2a/2b seamless design, which could offer not just time savings but reduced sample size as well. However, for the moment based on current knowledge in the NASH field, the proposed design offers the benefits of potentially creating more opportunities for participants and overall reduced trial conduct time for developers leading to the main goal of EU-PEARL: providing tools for accelerating drug development with increased efficiency in a cross-sponsor approach that will ultimately benefit patients.

# 5 Appendix

## 5.1 Detailed description of decision rules

**5.1.1 Verbal description.** In the following, a verbal description of the decision rules presented in section 2.2 with respect to the analysis time point is given.

**Interim analysis 1** Declare efficacy if either of the following two conditions is true:

1. • Posterior probability of at least 95% that success rate of NASH resolution in treatment arm is by any margin larger than in SOC arm AND

   • Posterior probability of at least 85% that success rate of NASH resolution in treatment arm is by at least 30 percentage points larger than in SOC arm AND

   • Posterior probability of at least 60% that success rate of NASH resolution in treatment arm is by at least 40 percentage points larger than in SOC arm

2. • Posterior probability of at least 95% that success rate of fibrosis improvement in treatment arm is by any margin larger than in SOC arm AND

   • Posterior probability of at least 85% that success rate of fibrosis improvement in treatment arm is by at least 17.5 percentage points larger than in SOC arm AND

   • Posterior probability of at least 60% that success rate of fibrosis improvement in treatment arm is by at least 25 percentage points larger than in SOC arm

Declare futility if both of the following are true:

1. Posterior probability of less than 20% that success rate of NASH resolution in treatment arm is by at least 25 percentage points larger than in SOC arm

2. Posterior probability of less than 20% that success rate of fibrosis improvement in treatment arm is by at least 10 percentage points larger than in SOC arm

If neither efficacy nor futility is declared, continue the trial.

**Interim analysis 2** Declare efficacy if either of the following two conditions is true:

1. • Posterior probability of at least 95% that success rate of NASH resolution in treatment arm is by any margin larger than in SOC arm AND

   • Posterior probability of at least 85% that success rate of NASH resolution in treatment arm is by at least 30 percentage points larger than in SOC arm AND

   • Posterior probability of at least 60% that success rate of NASH resolution in treatment arm is by at least 40 percentage points larger than in SOC arm

2. • Posterior probability of at least 95% that success rate of fibrosis improvement in treatment arm is by any margin larger than in SOC arm AND

   • Posterior probability of at least 85% that success rate of fibrosis improvement in treatment arm is by at least 17.5 percentage points larger than in SOC arm AND

   • Posterior probability of at least 60% that success rate of fibrosis improvement in treatment arm is by at least 25 percentage points larger than in SOC arm

   Declare futility if both of the following are true:

1. Posterior probability of less than 30% that success rate of NASH resolution in treatment arm is by at least 25 percentage points larger than in SOC arm

2. Posterior probability of less than 30% that success rate of fibrosis improvement in treatment arm is by at least 10 percentage points larger than in SOC arm

   If neither efficacy nor futility is declared, continue the trial.
   **Final analysis** Declare efficacy if either of the following two conditions is true:

1. • Posterior probability of at least 95% that success rate of NASH resolution in treatment arm is larger than in SOC arm AND

   • Posterior probability of at least 85% that success rate of NASH resolution in treatment arm is by at least 30 percentage points larger than in SOC arm AND

   • Posterior probability of at least 60% that success rate of NASH resolution in treatment arm is by at least 40 percentage points larger than in SOC arm

2. • Posterior probability of at least 95% that success rate of fibrosis improvement in treatment arm is larger than in SOC arm AND

   • Posterior probability of at least 85% that success rate of fibrosis improvement in treatment arm is by at least 17.5 percentage points larger than in SOC arm AND

   • Posterior probability of at least 60% that success rate of fibrosis improvement in treatment arm is by at least 25 percentage points larger than in SOC arm

**5.1.2 Formal definition.** Based on the decision rules given in Eq 2 in section 2.2, the proposed multi-component decision rules for several endpoints and interim analyses can be generalized as follows:

$$
\begin{aligned}
\text{GO, if } & (P(\pi_E > \pi_S + \delta_{k,1}^{G,T}|Data) > \gamma_{k,1}^{G,T}) \ \wedge \\
& (P(\pi_E > \pi_S + \delta_{k,2}^{G,T}|Data) > \gamma_{k,2}^{G,T}) \ \wedge \\
& \ldots \qquad\qquad\qquad\qquad\qquad\qquad \wedge \\
& (P(\pi_E > \pi_S + \delta_{k,l}^{G,T}|Data) > \gamma_{k,l}^{G,T}) \\
\\
\text{STOP, if } & (P(\pi_E > \pi_S + \delta_{k,1}^{F,T}|Data) < \gamma_{k,1}^{F,T}) \ \vee \\
& (P(\pi_E > \pi_S + \delta_{k,2}^{F,T}|Data) < \gamma_{k,2}^{F,T}) \ \vee \\
& \ldots \qquad\qquad\qquad\qquad\qquad\qquad \vee \\
& (P(\pi_E > \pi_S + \delta_{k,m}^{F,T}|Data) < \gamma_{k,m}^{F,T})
\end{aligned}
\tag{3}
$$

whereby $\pi_S$ denotes the response rate in the standard-of-care arm, $\pi_E$ denotes the response rate in the experimental treatment arm, $T \in 1, 2, \ldots, N$ denotes the analysis time point, subscript $k$ denotes the endpoint ($k \in \{1, \ldots, K\}$) and subscripts $l$ and $m$ denote the possibility to have multiple decision rules at any given point in time. At interim ($T \in \{1, 2, ..N-1\}$), if neither a decision for early efficacy or futility is made, the cohort continues unchanged. At final ($T = N$), if the efficacy boundaries are not met, the cohort automatically stops for futility. The initial letters E or F in the superscript of the thresholds $\delta$ and $\gamma$ indicate if this boundary is used to stop for efficacy (G) or futility (F). Choosing, for example, $\gamma_{k,1}^{E,u} = 1 \; \forall u \in \{1, \ldots, k\}$ corresponds to not allowing early stopping for efficacy for endpoint $k$ at interim 1. If at any point in time both stopping for early efficacy and futility is allowed, parameters need to be chosen carefully such that GO and STOP and decisions are not simultaneously possible. Please note that the requirements in Eq 3 refer to the evidence needed to declare efficacy or futility of a single endpoint, i.e. while we might have multiple efficacy requirements which need to be simultaneously fulfilled to declare a single endpoint efficacious, we advance the treatment if it is found to be efficacious in either of the two endpoints (analogously for futility decisions).

In order to achieve the multi-level decision rules described in Table 2 and generalized in Eq 3, we set the following parameters for efficacy ($l = 3$) and futility ($m = 1$):

- $\delta_{k,1}^{G,T} = 0, \gamma_{k,1}^{G,T} = 0.95, \gamma_{k,2}^{G,T} = 0.85, \gamma_{k,3}^{G,T} = 0.60, \; \forall T \in \{1,2,3\} \; \forall k \in \{1,2\}$

- $\delta_{1,2}^{G,T} = 0.30, \delta_{1,3}^{G,T} = 0.40, \delta_{2,2}^{G,T} = 0.175, \delta_{2,3}^{G,T} = 0.25, \; \forall T \in \{1,2,3\}$

- $\gamma_{k,1}^{F,1} = 0.20, \gamma_{k,1}^{F,2} = 0.30, \gamma_{k,1}^{G,3} = 0, \; \forall k \in \{1,2\}$

- $\delta_{1,1}^{F,T} = 0.25, \delta_{2,1}^{F,T} = 0.10 \; \forall T \in \{1,2,3\}$

**5.1.3 Visualization of Bayesian decision making based on Beta posteriors.** The properties of Bayesian decision making using posterior probabilities as described in more detail in section 2.2 is highlighted in Fig 8.

## 5.2 Sampling correlated binary endpoints

As mentioned previously, a successful clinical trial in NASH would investigate two correlated binary endpoints. This appendix is dedicated to establish the theory behind the sampling of correlated binary endpoints, which is used in our simulations in section 3.1. The proposed options are not meant to reflect the most efficient approaches (from a statistical or computational perspective), but rather the most realistic in terms of an interdisciplinary collaboration between clinicians and statisticians, whereby the role of the latter is to translate the received information as accurately as possible into simulation programs.

Let us assume during the course of a clinical trial investigating one or more treatments, we observe data on two correlated binary outcomes for every trial participant, $S$ and $L$. This could be the case if e.g. $L$ is a long-term endpoint and $S$ is a short-term endpoint which might be used as a surrogate for $L$. Another example would be if $S$ and $L$ are co-primary endpoints, as is the case in the trial design under investigation where S is NASH resolution (i.e. endpoint 1) and L is fibrosis improvement (i.e. endpoint 2). For the remainder of this section however, for reasons of better readability, we will refer to the two endpoints as "short-term" (S) and "long-term" (L) endpoint. For any given trial participant $i$, the four possible events which we could observe and their respective probabilities are $P(S = 0, L = 0) = p_{00}^i, P(S = 1, L = 0) = p_{10}^i, P(S = 0, L = 1) = p_{01}^i$ and $P(S = 1, L = 1) = p_{11}^i$. We assume that $p_{sl}^i$ is fully characterized by trial participant $i$'s covariates $x_i$. If treatment is the only covariate, $p_{sl}^i = p_{sl}^{k_i}$ where $k_i = k$
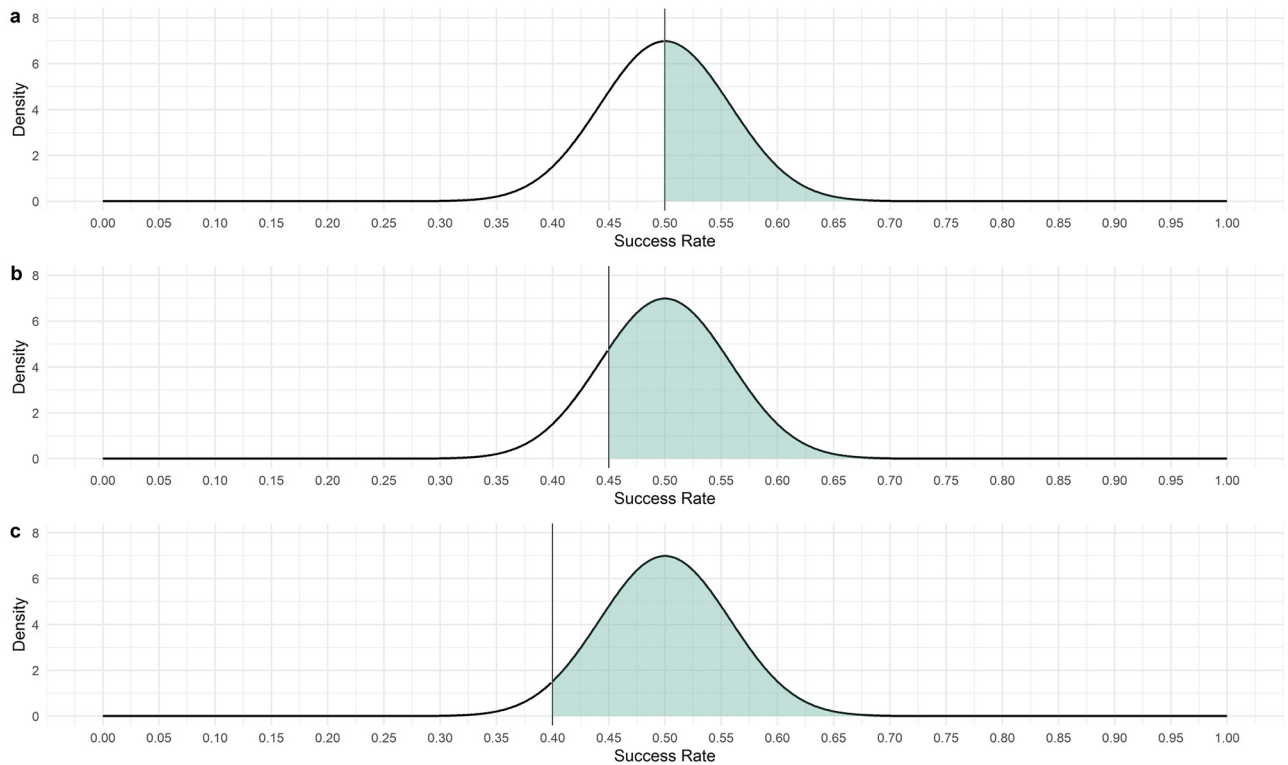
**Fig 8. Beta distributions corresponding to the posterior we would observe if a Beta(1,1) prior and a sample size of 75 was used and the observed response rate would equal 0.5.** Panel a: The posterior probability for a success rate greater or equal to 0.5 is 50%. If in our Bayesian decision rules we set a target of 0.5 and require a confidence of 50%, for large sample sizes we will graduate compounds with true responder rate of 0.5 in 50% of the cases, i.e. achieve a power of 50%. Panel b: The posterior probability for a success rate greater or equal to 0.45 is 81%. If in our Bayesian decision rules we set a target of 0.45 and require a confidence of 81%, for large sample sizes we will graduate compounds with true responder rate of 0.5 in 50% of the cases, i.e. achieve a power of 50%. Panel c: The posterior probability for a success rate greater or equal to 0.40 is 96%. If in our Bayesian decision rules we set a target of 0.40 and require a confidence of 96%, for large sample sizes we will graduate compounds with true responder rate of 0.5 in 50% of the cases, i.e. achieve a power of 50%. In order to achieve larger power values, required confidences and target success rates need to be adapted.

denotes treatment. Therefore, for the remainder of this section we drop the superscript $i$ in favor of a superscript $k$ on these probabilities, such that e.g. $p_{00}^k$ denotes the probability to have no success on either endpoint for all trial participants receiving treatment $k$. Hence, $(S_i, L_i)$ follows a bivariate Bernoulli distribution with expected value $\mathbf{p_i^k}$, covariance $p_{11}^k * p_{00}^k - p_{01}^k * p_{10}^k$ and correlation (Pearson $\phi$):

$$\phi^k = \frac{p_{11}^k * p_{00}^k - p_{01}^k * p_{10}^k}{\sqrt{p_{0\cdot}^k * p_{1\cdot}^k * p_{\cdot0}^k * p_{\cdot1}^k}} = \frac{p_{11}^k - p_{1\cdot}^k * p_{\cdot1}^k}{\sqrt{p_{0\cdot}^k * p_{1\cdot}^k * p_{\cdot0}^k * p_{\cdot1}^k}}$$

where $\mathbf{p_i^k}$ denotes vector of marginal success probabilities for $S$ and $L$ for trial participant $i$ (i.e. we only consider two different such vectors $\mathbf{p_i^k}$ if treatment is the only covariate) [40]. For an overview see Table 4. For a recent article discussing incorporation of both short-term and long-term binary data at interim, see [37], where simulations were based on the **R** packages **mvtnorm** and **psych**. Generalizations of this problem for more than two dimensions are discussed in [41].

Different diagnostic and predictive properties that arise are:

- Sensitivity of the short-term endpoint in predicting the long-term endpoint ($sens_{SL}^k$): $\frac{p_{11}^k}{p_{1\cdot}^k}$

**Table 4. 2x2 crosstable of possible short-term (S) and long-term (L) outcomes and their respective probabilities, as well as the marginal probabilities, with respect to treatment (denoted by superscript *k*).**

| S/L | 0 | 1 | |
|---|---|---|---|
| 0 | $p_{00}^{k}$ | $p_{01}^{k}$ | $p_{0\cdot}^{k}$ |
| 1 | $p_{10}^{k}$ | $p_{11}^{k}$ | $p_{1\cdot}^{k}$ |
| | $p_{\cdot 0}^{k}$ | $p_{\cdot 1}^{k}$ | 1 |

- Specificity of the short-term endpoint in predicting the long-term endpoint ($spec_{SL}^{k}$): $\frac{p_{00}^{k}}{p_{0\cdot}^{k}}$

- Sensitivity of the long-term endpoint in predicting the short-term endpoint ($sens_{LS}^{k}$): $\frac{p_{11}^{k}}{p_{\cdot 1}^{k}}$

- Specificity of the long-term endpoint in predicting the short-term endpoint ($spec_{LS}^{k}$): $\frac{p_{00}^{k}}{p_{\cdot 0}^{k}}$

There are multiple ways to specify the required probabilities in Table 4. We will now explore a few options in more detail which as mentioned previously focus on practical applicability in a setting of interdisciplinary collaboration and not necessarily on statistical or mathematical efficiency. Except for the method in section 5.2.4 in Appendix, all of the explored methods are reparametrizations of $\mathbf{p} = (p_{00}, p_{10}, p_{01}, p_{11})$ with the constraints $p_{ij} \in [0, 1]$ and $\Sigma_{ij} p_{ij} = 1$ (e.g. bijective functions which map $(p_{00}, p_{01}, p_{10})$ onto another set of three parameters).

**5.2.1 Direct specification.** The easiest and most straightforward way to specify the joint distribution is by fixing the four probabilities $p_{00}^{k}, p_{10}^{k}, p_{10}^{k}$ and $p_{11}^{k} = 1 - (p_{00}^{k} + p_{10}^{k} + p_{01}^{k})$. We can then calculate the correlation and diagnostic and predictive properties as described in the previous section. A drawback of this method is that likely neither any diagnostic/predictive properties nor the correlation are the same for both treatments unless the same probabilities are specified, which makes this approach rather unintuitive and results difficult to communicate.

**5.2.2 Implicit specification via sensitivity & specificity or PPV & NPV.** Next, let us reconsider Table 4 and note that there are various different ways of picking at least 3 of the unknown variables to fully characterize the bivariate Bernoulli distribution. In the previous section, we chose three unknown parameters from the "inside" of the Table (the fourth directly followed), but in this subsection we explore the option of specifying one marginal probability (i.e. either $p_{1\cdot}^{k}$ or $p_{\cdot 1}^{k}$) and the diagnostic and predictive properties of this outcome in predicting the other outcome (i.e. in case of specifying $p_{1\cdot}^{k}$, we additionally specify $sens_{SL}^{k}$ and $spec_{SL}^{k}$). This is possible without any constraints on the chosen probabilities. This approach might make sense if, for example, $S$ is a surrogate endpoint for $L$. Note that specifying one marginal probability and the diagnostic and predictive properties of the other outcome in predicting this outcome (i.e. in case of specifying $p_{1\cdot}^{k}$, additionally specifying $sens_{LS}^{k}$ and $spec_{LS}^{k}$) is possible only under heavy constraints on the chosen triplet of values and most combinations of values are not attainable (for more information see Fig 9). Another drawback is that in most clinical examples, it might not make sense to require $p_{1\cdot}^{k}$ and $sens_{SL}^{k}$ and $spec_{SL}^{k}$ as input parameters for a simulation study, but rather $p_{\cdot 1}^{k}$ and $sens_{SL}^{k}$ and $spec_{SL}^{k}$, which, as discussed before, leads to invalid probabilities for most combinations of these three parameters.

The required transformations are reported exemplary for specifying $p_{1\cdot}^{k}$, $sens_{SL}^{k}$ and $spec_{SL}^{k}$. From $sens_{SL}^{k} = \frac{p_{11}^{k}}{p_{1\cdot}^{k}}$ and $spec_{SL}^{k} = \frac{p_{00}^{k}}{p_{0\cdot}^{k}}$ it follows that $p_{11}^{k} = sens_{SL}^{k} * p_{1\cdot}^{k}$ and $p_{00}^{k} = spec_{SL}^{k} * (1 - p_{1\cdot}^{k})$. Subsequently, $p_{10}^{k} = p_{1\cdot}^{k} - p_{11}^{k}$ and $p_{01}^{k} = (1 - p_{1\cdot}^{k}) - p_{00}^{k}$. Another drawback of this method is
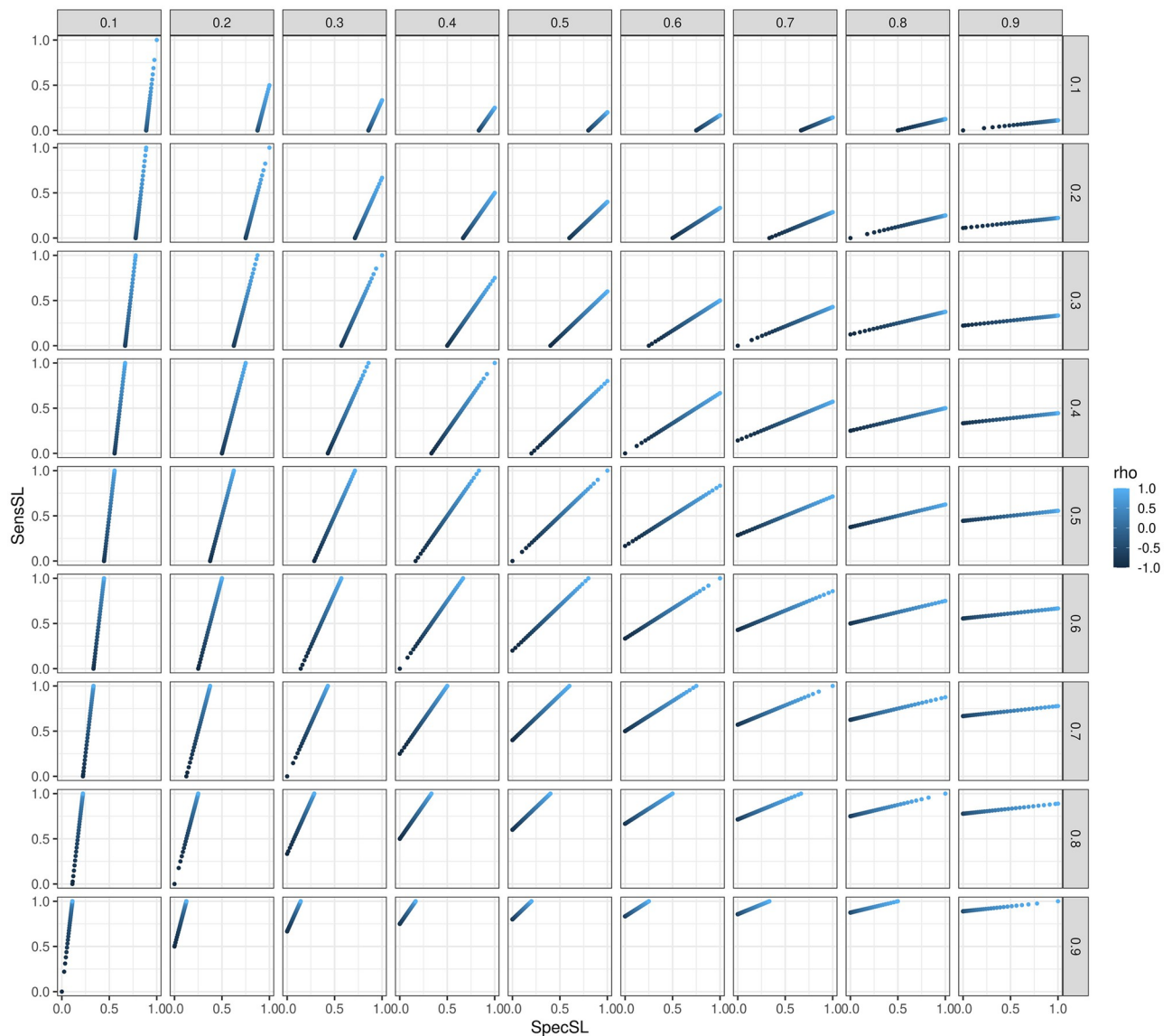
**Fig 9. $sens_{SL}$ and $spec_{SL}$ with respect to $p^k_{1\cdot}$ (columns), $p^k_{\cdot 1}$ (rows) and $\rho$ (shade of color).** For scenarios $p^k_{1\cdot} < p^k_{\cdot 1}$, there is an upper $u_{spec} < 1$ bound for $spec_{SL}$. For scenarios $p^k_{1\cdot} > p^k_{\cdot 1}$, there is an upper $u_{sens} < 1$ bound for $sens_{SL}$. For scenarios $p^k_{1\cdot} = p^k_{\cdot 1}$ either $sens_{SL}$ ($p^k_{1\cdot} = p^k_{\cdot 1} \leq 0.5$) or $spec_{SL}$ ($p^k_{1\cdot} = p^k_{\cdot 1} \geq 0.5$) can take any values between 0 and 1. If the matrix of figures was transposed, we would see $sens_{LS}$ and $spec_{LS}$ instead of $sens_{SL}$ and $spec_{SL}$. Please note that in the Figure the label "rho" is used for $\rho$.

https://doi.org/10.1371/journal.pone.0281674.g009

that the correlation and diagnostic and predictive properties of $S$ in predicting $L$ will likely differ between treatments.

**5.2.3 Implicit specification via correlation.** In this subsection we explore the option of specifying $p^k_{1\cdot}, p^k_{\cdot 1}$ and $\phi^k$. From the second formula for $\phi^k$ provided in the previous section, it follows that

$$p^k_{11} = \phi^k * \sqrt{p^k_{0\cdot} * p^k_{1\cdot} * p^k_{\cdot 0} * p^k_{\cdot 1}} + p^k_{1\cdot} * p^k_{\cdot 1}$$

As other authors have noted [42], after fixing $p^k_{\cdot 1}$ and $p^k_{1\cdot}$, $\phi^k$ cannot be chosen freely anymore

and is bounded below by

$$max\left(-\sqrt{\frac{p^k_{1.} * p^k_{.1}}{(1 - p^k_{1.})(1 - p^k_{.1})}}, -\sqrt{\frac{(1 - p^k_{1.})(1 - p^k_{.1})}{p^k_{1.} * p^k_{.1}}}\right)$$

and above by

$$min\left(\sqrt{\frac{p^k_{1.} * (1 - p^k_{.1})}{(1 - p^k_{1.}) * p^k_{.1}}}, \sqrt{\frac{(1 - p^k_{1.}) * p^k_{.1}}{p^k_{1.} * (1 - p^k_{.1})}}\right)$$

Finally, $p^k_{01} = p^k_{.1} - p^k_{11}, p^k_{10} = p^k_{1.} - p^k_{11}$ and $p^k_{00} = 1 - (p^k_{11} + p^k_{01} + p^k_{10})$ can simply be derived.

The constraints on the triplet $(p^k_{1.}, p^k_{.1}, \phi^k)$ have two major impacts in the scenario of simulating correlated binary outcomes for different treatments: 1) If we assume the correlation to be the same across all treatments, it might happen that for some treatments we derive valid probabilities and for some treatments we do not derive valid probabilities. 2) For any chosen pair $(p^k_{1.}, p^k_{.1})$, we are unable to investigate the whole range of correlations. Frequently statisticians will receive the following information, based on which the simulations should be performed: "We know $(p^k_{1.}$ and $p^k_{.1})$ and the correlation". While this method would technically be the correct approach to take, it is very hard to keep track of the restrictions on the parameters (since one of the parameters has a support which depends on the values of the other two) and they are also difficult to communicate.

**5.2.4 Specification via bivariate normal distribution.** Finally, the required probabilities can be derived using a latent variable approach from a bivariate normal distribution. In this approach, we would usually like to fix $p^k_{.1}$ (and thereby $p^k_{.0}$) and $p^k_{1.}$ (and thereby $p^k_{0.}$), i.e. the short-term and long-term response rates of treatment $k$. Furthermore, we would like to specify a naive correlation $\rho$ for the two outcomes, i.e. we need to specify a triplet $(p^k_{.1}, p^k_{1.}, \rho)$. We then define the joint probabilities as follows:

$$p^k_{00} = \int_{-\infty}^{p^k_{0.}} \int_{-\infty}^{p^k_{.0}} f(\vec{x}, \vec{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}) dx_1 dx_2$$

$$p^k_{10} = \int_{p^k_{0.}}^{\infty} \int_{-\infty}^{p^k_{.0}} f(\vec{x}, \vec{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}) dx_1 dx_2$$

$$p^k_{01} = \int_{-\infty}^{p^k_{0.}} \int_{p^k_{.0}}^{\infty} f(\vec{x}, \vec{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}) dx_1 dx_2$$

$$p^k_{11} = \int_{p^k_{0.}}^{\infty} \int_{p^k_{.0}}^{\infty} f(\vec{x}, \vec{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}) dx_1 dx_2$$

whereby $f(\vec{x}, \vec{\mu}, \Sigma)$ is the probability density function of the bivariate normal distribution with mean $\mu$, covariance matrix $\Sigma$ and $\vec{x} = (x_1, x_2)$. It should be obvious that these four probabilities sum up to 1. This basically corresponds to splitting the bivariate normal distribution into 4 new quadrants and setting the probabilities equal to the probability mass in each of these four quadrants.

The most obvious drawback of this methods is that not all combinations of $\mathbf{p_i}$ are attainable this way (e.g. (0.33, 0.34, 0.33, 0) is impossible). Another drawback of allowing the specification

of triplets $(p_{.1}^k, p_{1.}^k, \rho)$ is that the diagnostic and predictive properties differ for different treatments (if they have different response rates and equal correlations). In fact, for a given pair of response rates $(p_{1.}^1, p_{.1}^1)$, the sensitivity and specificity are a function of the correlation $\rho$. Therefore, there are constraints regarding achievable sensitivity and specificity for a given set of response rates $(p_{.1}^k, p_{1.}^k)$. Fig 9 shows the achieved sensitivity and specificity for different ranges of $p_{.1}^k$ and $p_{1.}^k$. Another drawback of this method is the non-linear relationship between the specified $\rho$ and the actual correlation of the binary endpoints, $\phi$, which also depends on $p_{.1}^k$ and $p_{1.}^k$. In practise, this means that in most cases (as in our simulation study) we will think of $\rho$ as the correlation, when in fact the actual correlation $\phi$ between the binary endpoints might differ substantially. See Fig 10 in the Appendix for more details.
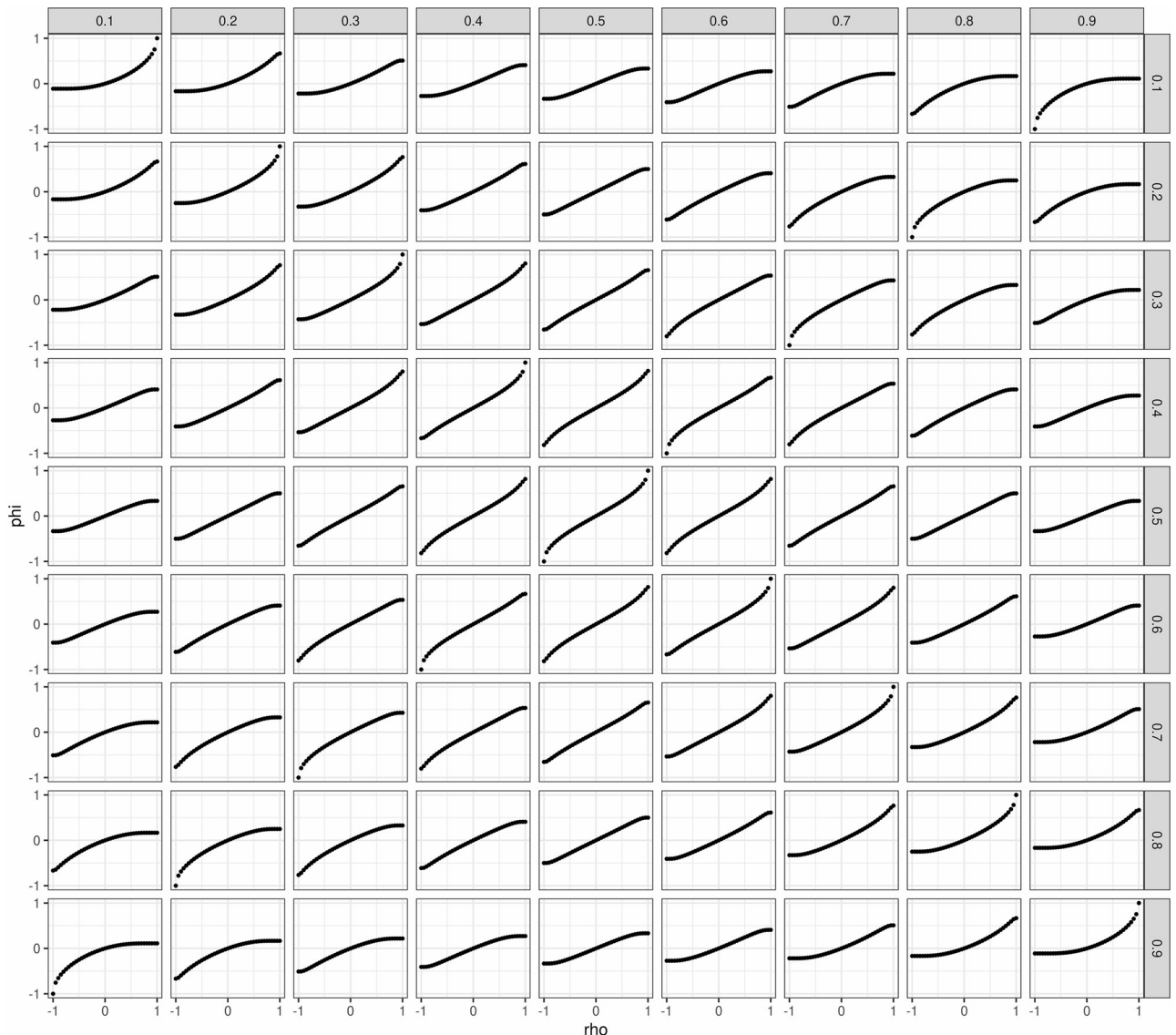


**Fig 10. Relationship between specified correlation of the two continuous endpoints $\rho^k$ prior to dichotomization and achieved correlation $\phi^k$ of the two binary endpoints after dichotomization.** Only when $p_{1.}^k = p_{.1}^k = 0.5$ can $\phi^k \in [0, 1]$ be achieved, otherwise it is bounded above and/or below (see paragraph "Correlation" in section "Implicit specification" for more details on the bounds). The more $p_{1.}^k$ and $p_{.1}^k$ differ, the closer either the upper or lower bound of $\phi^k$ is to 0. Please note that in the Figure the labels "phi" and "rho" are used for $\phi$ and $\rho$.

https://doi.org/10.1371/journal.pone.0281674.g010

As demonstrated in the previous sections, no single specification option comes without limitations. For this simulation study, we chose to sample trial participants' correlated binary endpoints via the bivariate normal distribution, because requiring marginal success probabilities and a "naive" correlation seemed like the best trade-off between being unproblematic in terms of constraints, being easy to implement in the software and finally—and most importantly—easy to communicate to clinical teams.

## Supporting information

**S1 File. Simulated data used for results.**
(XLSX)

## Acknowledgments

## Author Contributions

**Conceptualization:** Elias Laurin Meyer, Peter Mesenbrink, Nicholas A. Di Prospero, Juan M. Pericàs, Franz König.

**Methodology:** Elias Laurin Meyer, Peter Mesenbrink, Nicholas A. Di Prospero, Juan M. Pericàs, Ekkehard Glimm, Franz König.

**Project administration:** Elena Sena.

**Software:** Elias Laurin Meyer.

**Supervision:** Franz König.

**Validation:** Franz König.

**Visualization:** Elias Laurin Meyer.

**Writing – original draft:** Elias Laurin Meyer, Franz König.

**Writing – review & editing:** Elias Laurin Meyer, Peter Mesenbrink, Nicholas A. Di Prospero, Juan M. Pericàs, Ekkehard Glimm, Vlad Ratziu, Elena Sena, Franz König.

## References

1. Redman MW, Allegra CJ. The master protocol concept. Seminars in Oncology. 2015; 42(5):724–730. https://doi.org/10.1053/j.seminoncol.2015.07.009 PMID: 26433553

2. Meyer EL, Mesenbrink P, Dunger-Baldauf C, Fülle HJ, Glimm E, Li Y, et al. The Evolution of Master Protocol Clinical Trial Designs: A Systematic Literature Review. Clinical Therapeutics. 2020; 42(7):1330–1360. https://doi.org/10.1016/j.clinthera.2020.05.010 PMID: 32622783

3. Meyer EL, Mesenbrink P, Mielke T, Parke T, Evans D, König F. Systematic review of available software for multi-arm multi-stage and platform clinical trial design. Trials. 2021; 22(1):1–14. https://doi.org/10.1186/s13063-021-05130-x PMID: 33663579

4. Woodcock J, LaVange LM. Master Protocols to Study Multiple Therapies, Multiple Diseases, or Both. New England Journal of Medicine. 2017; 377(1):62–70. https://doi.org/10.1056/NEJMra1510062 PMID: 28679092

5. Meyer EL. Designing exploratory platform trials. Medical University of Vienna; 2022.

6. Kunz CU, Jörgens S, Bretz F, Stallard N, Lancker KV, Xi D, et al. Clinical trials impacted by the COVID-19 pandemic: Adaptive designs to the rescue? Statistics in Biopharmaceutical Research. 2020; 0(ja):1–41. https://doi.org/10.1080/19466315.2020.1799857 PMID: 34191979

7. Stallard N, Hampson L, Benda N, Brannath W, Burnett T, Friede T, et al. Efficient Adaptive Designs for Clinical Trials of Interventions for COVID-19. Statistics in Biopharmaceutical Research. 2020; 0(0):1–15. https://doi.org/10.1080/19466315.2020.1790415 PMID: 34191981

8. Dodd LE, Follmann D, Wang J, Koenig F, Korn LL, Schoergenhofer C, et al. Endpoints for randomized controlled clinical trials for COVID-19 treatments. Clinical Trials. 2020; 17(5):472–482. https://doi.org/10.1177/1740774520939938 PMID: 32674594

9. Horby PW, Mafham M, Bell JL, Linsell L, Staplin N, Emberson J, et al. Lopinavir–ritonavir in patients admitted to hospital with COVID-19 (RECOVERY): a randomised, controlled, open-label, platform trial. The Lancet. 2020; 396(10259):1345–1352. https://doi.org/10.1016/S0140-6736(20)32013-4

10. Angus DC, Derde L, Al-Beidh F, Annane D, Arabi Y, Beane A, et al. Effect of hydrocortisone on mortality and organ support in patients with severe COVID-19: the REMAP-CAP COVID-19 corticosteroid domain randomized clinical trial. Jama. 2020; 324(13):1317–1329. https://doi.org/10.1001/jama.2020.17022 PMID: 32876697

11. Macleod J, Norrie J. PRINCIPLE: a community-based COVID-19 platform trial. The Lancet Respiratory Medicine. 2021; 9(9):943–945. https://doi.org/10.1016/S2213-2600(21)00360-X PMID: 34331866

12. EU-PEARL. D2.1. Report on Terminology, References and Scenarios for Platform Trials and Master Protocols; 2020. Available from: https://eu-pearl.eu/wp-content/uploads/2020/06/EU-PEARL_D2.1_Report-on-Terminology-and-Scenarios-for-Platform-Trials-and-Masterprotocols.pdf.

13. EU Patient-Centric Clinical Trial Platforms; 2021.

14. Chalasani N, Younossi Z, Lavine JE, Charlton M, Cusi K, Rinella M, et al. The diagnosis and management of nonalcoholic fatty liver disease: practice guidance from the American Association for the Study of Liver Diseases. Hepatology. 2018; 67(1):328–357. https://doi.org/10.1002/hep.29367 PMID: 28714183

15. Powell EE, Wong VWS, Rinella M. Non-alcoholic fatty liver disease. The Lancet. 2021; 397 (10290):2212–2224. https://doi.org/10.1016/S0140-6736(20)32511-3

16. Ratziu V, Friedman SL. Why do so many NASH trials fail? Gastroenterology. 2020. https://doi.org/10.1053/j.gastro.2020.05.046 PMID: 32439497

17. Ratziu V, Francque S, Sanyal A. Breakthroughs in therapies for NASH and remaining challenges. Journal of Hepatology. 2022; 76(6):1263–1278. https://doi.org/10.1016/j.jhep.2022.04.002 PMID: 35589249

18. FDA/CDER. Noncirrhotic Nonalcoholic Steatohepatitis With Liver Fibrosis: Developing Drugs for Treatment—DRAFT GUIDANCE; 2018. Available from: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/noncirrhotic-nonalcoholic-steatohepatitis-liver-fibrosis-developing-drugs-treatment.

19. European Medicines Agency. Reflection paper on regulatory requirements for the 5 development of medicinal products for chronic non6 infectious liver diseases (PBC, PSC, NASH) (draft); 2018. Available from: https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-regulatory-requirements-development-medicinal-products-chronic-non-infectious-liver_en.pdf.

20. Anania FA, Dimick-Santos L, Mehta R, Toerner J, Beitz J. Nonalcoholic Steatohepatitis: Current Thinking From the Division of Hepatology and Nutrition at the Food and Drug Administration. Hepatology (Baltimore, Md). 2020.

21. Hagström H, Nasr P, Ekstedt M, Hammar U, Stål P, Hultcrantz R, et al. Fibrosis stage but not NASH predicts mortality and time to development of severe liver disease in biopsy-proven NAFLD. Journal of hepatology. 2017; 67(6):1265–1273. https://doi.org/10.1016/j.jhep.2017.07.027 PMID: 28803953

22. Dulai PS, Singh S, Patel J, Soni M, Prokop LJ, Younossi Z, et al. Increased risk of mortality by fibrosis stage in nonalcoholic fatty liver disease: systematic review and meta-analysis. Hepatology. 2017; 65 (5):1557–1565. https://doi.org/10.1002/hep.29085 PMID: 28130788

23. Ratziu V. A critical review of endpoints for non-cirrhotic NASH therapeutic trials. Journal of Hepatology. 2018; 68(2):353–361. https://doi.org/10.1016/j.jhep.2017.12.001 PMID: 29223369

24. Meyer EL, Mesenbrink P, Dunger-Baldauf C, Glimm E, Li Y, König F. Decision rules for identifying combination therapies in open-entry, randomized controlled platform trials. Pharmaceutical Statistics. 2022; 21(3):671–690. https://doi.org/10.1002/pst.2194 PMID: 35102685

25. ICH E9 Expert Working Group. Statistical principles for clinical trials. Statistics in Medicine. 1999; 18:1905–1942.

26. Roig MB, Krotka P, Burman CF, Glimm E, Gold SM, Hees K, et al. On model-based time trend adjustments in platform trials with non-concurrent controls. BMC medical research methodology. 2022; 22 (1):1–16. https://doi.org/10.1186/s12874-022-01683-w PMID: 35971069

27. Brunt EM, Kleiner DE, Wilson LA, Sanyal AJ, Neuschwander-Tetri BA, Network NSCR. Improvements in histologic features and diagnosis associated with improvement in fibrosis in nonalcoholic steatohepatitis: results from the nonalcoholic steatohepatitis clinical research network treatment trials. Hepatology. 2019; 70(2):522–531. https://doi.org/10.1002/hep.30418 PMID: 30549292

28. Kleiner DE, Brunt EM, Wilson LA, Behling C, Guy C, Contos M, et al. Association of histologic disease activity with progression of nonalcoholic fatty liver disease. JAMA network open. 2019; 2(10): e1912565–e1912565. https://doi.org/10.1001/jamanetworkopen.2019.12565 PMID: 31584681

29. Meyer EL, Mesenbrink P, Dunger-Baldauf C, Glimm E, König F. CohortPlat: Simulation of cohort platform trials investigating combination therapies. Final Draft. 2022.

30. Jennison C, Turnbull BW. Group sequential methods with applications to clinical trials. CRC Press; 1999.

31. Zehetmayer S, Posch M, Koenig F. Online control of the False Discovery Rate in group-sequential platform trials. Statistical Methods in Medical Research. 2022; 31(12):2470–2485. https://doi.org/10.1177/09622802221129051 PMID: 36189481

32. Gallo P, Mao L, Shih VH. Alternative views on setting clinical trial futility criteria. Journal of biopharmaceutical statistics. 2014; 24(5):976–993. https://doi.org/10.1080/10543406.2014.932285 PMID: 24933121

33. Newsome PN, Buchholtz K, Cusi K, Linder M, Okanoue T, Ratziu V, et al. A placebo-controlled trial of subcutaneous semaglutide in nonalcoholic steatohepatitis. New England Journal of Medicine. 2021; 384(12):1113–1124. https://doi.org/10.1056/NEJMoa2028395 PMID: 33185364

34. Francque SM, Bedossa P, Ratziu V, Anstee QM, Bugianesi E, Sanyal AJ, et al. A randomized, controlled trial of the pan-PPAR agonist lanifibranor in NASH. New England Journal of Medicine. 2021; 385 (17):1547–1558. https://doi.org/10.1056/NEJMoa2036205 PMID: 34670042

35. Meyer EL, Kumaus C, Majka M, König F. An interactive R-Shiny app for quickly visualizing a tidy, long dataset with multiple dimensions with an application in clinical trial simulations for platform trials. Long Dataset with Multiple Dimensions with an Application in Clinical Trial Simulations for Platform Trials (October 12, 2022). 2022.

36. FDA. Critical Path Innovation Meetings (CPIM); 2022.

37. Niewczas J, Kunz CU, König F. Interim analysis incorporating short-and long-term binary endpoints. Biometrical Journal. 2019; 61(3):665–687. https://doi.org/10.1002/bimj.201700281 PMID: 30694566

38. Allen AM, Therneau TM, Ahmed OT, Gidener T, Mara KC, Larson JJ, et al. Clinical course of non-alcoholic fatty liver disease and the implications for clinical trial design. Journal of hepatology. 2022; 77 (5):1237–1245. https://doi.org/10.1016/j.jhep.2022.07.004 PMID: 35843374

39. Sanyal AJ, Van Natta ML, Clark J, Neuschwander-Tetri BA, Diehl A, Dasarathy S, et al. Prospective study of outcomes in adults with nonalcoholic fatty liver disease. New England Journal of Medicine. 2021; 385(17):1559–1569. https://doi.org/10.1056/NEJMoa2029349 PMID: 34670043

40. Marshall AW, Olkin I. A family of bivariate distributions generated by the bivariate Bernoulli distribution. Journal of the American Statistical Association. 1985; 80(390):332–338. https://doi.org/10.1080/01621459.1985.10478116

41. Jiang W, Song S, Hou L, Zhao H. A set of efficient methods to generate high-dimensional binary data with specified correlation structures. The American Statistician. 2021; 75(3):310–322. https://doi.org/10.1080/00031305.2020.1816213

42. Sozu T, Sugimoto T, Hamasaki T. Sample size determination in clinical trials with multiple co-primary binary endpoints. Statistics in medicine. 2010; 29(21):2169–2179. https://doi.org/10.1002/sim.3972 PMID: 20687162