Approaches to sample size calculation for clinical trials in rare diseases

Frank Miller[1] | Sarah Zohar[2] | Nigel Stallard[3] | Jason Madan[4] | Martin Posch[5] | Siew Wan Hee[3] | Michael Pearce[6] | Mårten Vågerö[7] | Simon Day[8]

[1]Department of Statistics, Stockholm University, Stockholm, Sweden
[2]INSERM, U1138, team22, Centre de Recherche des Cordeliers, Université Paris 5, Université Paris 6, Paris, France
[3]Statistics and Epidemiology, Division of Health Sciences, Warwick Medical School, University of Warwick, Coventry, UK
[4]Clinical Trials Unit, Warwick Medical School, University of Warwick, Coventry, UK
[5]Section for Medical Statistics, Center for Medical Statistics, Informatics, and Intelligent Systems, Medical University of Vienna, Austria
[6]Complexity Science, University of Warwick, Coventry, UK
[7]Swedish Orphan Biovitrum, Stockholm, Sweden
[8]Clinical Trials Consulting and Training Limited, Buckingham, UK

Correspondence
Frank Miller, Department of Statistics, Stockholm University, Stockholm, Sweden.
Email: frank.miller@stat.su.se

Funding information
European Union Seventh Framework Programme for research, technological development and demonstration, Grant Agreement Number: FP HEALTH 2013602144.

ABSTRACT

We discuss three alternative approaches to sample size calculation: Traditional sample size calculation based on power to show a statistically significant effect, sample size calculation based on assurance, and sample size based on a decision-theoretic approach. These approaches are compared head-to-head for clinical trial situations in rare diseases. Specifically, we consider three case studies of rare diseases (Lyell's disease, adult-onset Still's disease, and cystic fibrosis) with the aim to plan the sample size for an upcoming clinical trial. We outline in detail the reasonable choice of parameters for these approaches for each of the three case studies and calculate sample sizes. We stress that the influence of the input parameters needs to be investigated in all approaches and recommend investigating different sample size approaches before deciding finally on the trial size. Highly influencing for the sample size are choice of treatment effect parameter in all approaches and the parameter for the additional cost of the new treatment in the decision-theoretic approach. These should therefore be discussed extensively.

KEYWORDS

Assurance, clinical trial, decision theory, rare disease, sample size calculation

1 | INTRODUCTION

The choice of sample size for clinical trials is of key importance and should be made in a well-informed way. Sample size planning is especially critical if the target population for the investigation is difficult to study, e.g. a rare disease population where not enough patients exist to conduct a trial of traditional size or a setting where possibilities for inclusion in trials are restricted, such as paediatric populations.

There are different approaches to sample size calculation. In this paper, we will compare three approaches: 1) the traditional sample size calculation ensuring a prespecified power for showing a statistically significant effect versus control, 2) sample size calculation based on assurance where uncertainty about assumed treatment effects are modelled, 3) optimal sample size based on a decision-theoretic approach where we will distinguish between an acute and a chronic disease.

During study planning an appropriate sample size approach needs to be chosen. In some situations it can turn out that the traditional goal, power for a significance test, is neither a reasonable nor a desired goal for the study. However, an important step to sample size calculation is not only to choose this approach but also to specify the required parameters for these approaches in a reasonable way. The specification of parameters can be a difficult step even when the traditional way of sample size determination is used.[1] A non-traditional sample size approach is often more difficult since there is less familiarity with it in the scientific community and therefore less agreement how parameters should be specified. Further, we will see in this paper that we have to specify more parameters in advance for the decision-theoretic approach since it is more flexible.

The aim of this paper is to show how parameters in different sample size approaches can be chosen in rare disease settings. We will discuss good sample size praxis – a discussion which is needed in the literature especially for the decision-theoretic approach to increase the acceptance in clinical trials. Further we want to compare traditional and non-traditional sample size approaches head-to-head in rare disease case studies in this paper: Does the decision-theoretic approach lead to smaller sample sizes than the traditional approach? Which parameters drive the sample size?

We focus in this article on the sample size recommendation coming from a primary efficacy variable for the study. This is simplifying the real situation as other aspects, such as the need for adequate safety data can influence sample sizes as well. Other approaches or combinations of approaches for sample size than the three considered here exist as well and we will briefly discuss alternatives in the concluding section.

After a general description of the three considered approaches in Section 2, we discuss in Section 3 three case studies of rare diseases and compare the different approaches to sample size

calculation. In our first case study, a new cellular therapy (acute treatment) for Lyell's disease is planned. The sample size was finally based on the decision-theoretic approach, pre-study preparation work is currently ongoing and the study is expected to start in 2018. In the second case, the topic is a study for use of an interleukin I antagonist for the chronic treatment of adult-onset Still's disease. The third case is looking for the size of a cystic fibrosis trial investigating the chronic treatment with inhaled dry powder mannitol. We highlight in the case studies that it is good praxis for all sample size approaches to consider robustness of the sample size by checking the influence on sample size when some specifications are changed.

It turns out that specifications for the treatment effect are especially challenging for all three methods. Therefore, we discuss this choice in more detail in Section 4. We conclude this article by discussing some practical aspects important when the approaches (especially the decision-theoretic approach) are to be used for an upcoming study. We give thoughts about a treatment-licencing situation and the situation when a trial for an ultra-rare disease is planned.

2 | METHODS FOR SAMPLE SIZE DETERMINATION

2.1 | Traditional approach: Ensuring specific power for a target treatment effect

Traditionally, the determination or motivation of the sample size for a clinical trial in many cases uses the following approach: A specific treatment effect is targeted and possibly additional values for nuisance parameters are assumed. As the goal is to reject the null hypothesis of no treatment effect after the study, a significance level is chosen (often 5%) and the necessary sample size can be calculated to ensure a specific power to reject the null hypothesis if the treatment has the target effect (typically 80-90%). Statistical methods for this approach are discussed e.g. by Julious[2] and Julious and Campbell.[3] Nevertheless, of key importance for the clinical trial is the appropriate choice of the above mentioned parameters which are the basis for the computation (target treatment effect, assumed nuisance parameters, significance level, power). We will discuss the choice of these values in the context of the case studies in Section 3.

2.2 | Assurance approach: Handling uncertainty in assumed treatment effect

Power as calculated conventionally depends on a targeted effect which might be an assumed true treatment effect. If current information on the likely true effect can be characterised in the form of a prior distribution, the expected power of a proposed study can be calculated for a given sample size and this prior. This gives the expected probability of a successful trial (which shows a statistically significant treatment effect), also known as 'assurance'.[4]-[6] The sample size can then be chosen to ensure a specific assurance (80-90%). This approach is useful when previous studies of the treatment of interest exist and their results can form a prior (we will see this in two of our case studies). When no previous studies exist, assurance can be calculated assuming 'sceptical' or 'optimistic' priors, or by elicitation from suitable experts. Methods and examples for how to elicit information from experts about treatment effects are discussed e.g. by O'Hagan et al.[7], Zohar et

al.[8] and Kinnersley and Day.[9] In the same way as for the treatment effect, also uncertainties about nuisance parameters (like variability, success rate for control treatment) can be handled by using a prior distribution. We note that for priors with too large prior-probability for the new treatment being worse than control, the assurance can be below the target assurance even for infinitively large sample size. This issue occurring in Case Study 1 is well-known.[10]

2.3 | Decision-theoretic approach

The above approaches focus on the frequentist properties (type I error and power/assurance) since the intended analysis is to perform a significance test after the trial. As an alternative, decision-theoretic approaches have been proposed. These can be applied when the intention is not to base a treatment recommendation on a significance test with a certain type I error but instead a treatment recommendation is desired which maximizes an expected "gain" for the total population. The idea for designing the study including the choice of the sample size is to compare the expected gain which results from different decisions about the design (here: different choices about the sample size). The sample size which maximizes the expected gain is then chosen for the clinical study. Hee et al.[11] review such approaches for small trials and pilot studies. "Gain" is interpreted very broadly and can be defined from the patient, sponsor, regulatory, public health or society perspective – or from a combined perspective. The gain will depend on which perspective is taken. It accounts for the benefits from the treatment and is reduced by costs (e.g. monetary costs for development but also costs in terms of adverse effects). For each patient, gain functions are defined which depend on the received treatment (new or control), whether they receive it in or outside the trial, and on the treatment outcome for the patient. Like the assurance approach (Section 2.2), this approach assumes that current information about the new treatment and the control is characterized in the form of prior distributions.

2.3.1 Acute disease

We start with having a one-chance treatment in mind intended for an acute disease. In this situation, we assume that each patient in a target population of size N is treated exactly once: either in a trial where n_1 patients are allocated to the new treatment and n_2 patients are allocated to control, or after the trial where all N-n_1-n_2 remaining patients receive the treatment which is estimated to be the better based on the posterior distribution after the trial. Figure 1 shows this basic assumption for this approach assumed by Cheng et al.[12] and Stallard et al.[13]. Let δ_1 be the effect of the new treatment and δ_2 the effect of control. An overall expected gain is here defined to be

$$\begin{aligned} &G(n_1, n_2, \delta_1, \delta_2) = n_1 h_{New}(\delta_1) + n_2 h_{Control}(\delta_2) \\ &+(N - n_1 - n_2)[\mathbf{1}\{\text{New treatment chosen}\} g_{New}(\delta_1) + \mathbf{1}\{\text{Control chosen}\} g_{Control}(\delta_2)] \end{aligned} \quad (1)$$

where $h_{New}(\delta_1)$ and $h_{Control}(\delta_2)$ are expected gains for one patient when treated in the trial ("in-trial gain") with New Treatment or Control, respectively, g is the corresponding expected gain when

treated with a treatment outside of the trial ("out-trial gain") and $\mathbf{1}\{\}$ is the indicator function. Since there are no study specific costs for patients treated out-trial – it is assumed that the out-trial-gain of the superior treatment is at least as large as either of the in-trial-gains.

[Figure 1 around here]

Sample size determination then corresponds to the choice of n_1 and n_2 to maximize the overall expected gain. Here additional expectations are taken over the possible observations in and outside the trial and over the prior distribution of the parameters, see Stallard et al.[13] for more details. Further, it is assumed that the treatment decision after the trial is to choose the treatment maximizing the expected gain given the posterior distribution after the trial. One would like to have a sufficiently high sample size to have a good probability to determine the better treatment. On the other hand, a too large sample size will imply that more patients are treated with the inferior treatment in the trial and fewer patients have the possibility to gain from the treatment outside the trial. These conflicting arguments are the basis for searching for the optimal sample size.

Note that the described approach can be applied also when the trial is a single arm trial with n_1 patients on new treatment (formula (1) can be applied with $n_2=0$). This is especially important in rare disease settings when comparison with this "historical" information about control is desired. In Section 3.1, we will discuss a single arm case study.

2.3.2 Chronic disease

There are many rare diseases of chronic nature and the treatment must be taken continuously over time for benefitting from it. Patients included in the trial can benefit from one of the treatments for the trial duration and have an in-trial-gain. Let the treatment duration in the trial for each patient is be d. After being in the trial, the patients need further treatment. We assume that we are interested in a certain time horizon H for treating patients. Further we assume that the results from the trial can lead to a treatment policy recommending either new treatment or control from time S onwards which is after the end of the trial plus additional time e.g. for interactions between decision makers, drug production, advertisement, and marketing. We assume here that the N patients in the patient population will start the recommended treatment from a time-point with average at time S for the N patients. Since the time S depends on when the trial ends, it is a function of the recruitment speed and the trial's sample size n_1+n_2 and we write $S=S(n_1+n_2)$. In the case studies which consider chronic diseases, we will use functions S which depend linearly on the trial's sample size: $S= t_0 + t_1 (n_1+n_2)$ with a constant t_0 and a recruitment speed t_1 (time per patient recruited). Before time S, the control treatment is standard of care outside of the trial. Figure 2 shows these basic assumptions for a chronic disease graphically (where we illustrate for simplicity that all N patients start their out-trial treatment simultaneously at time S). We assume that the gain for a patient is proportional to the length of time a patient receives treatment. The

overall expected gain for a chronic disease case is

$$\begin{aligned} G(n_1, n_2, \delta_1, \delta_2) &= n_1 dh_{New}(\delta_1) + n_2 dh_{Control}(\delta_2) \\ &+ N(H - S(n_1 + n_2))[\mathbf{1}\{\text{New treatment chosen}\} g_{New}(\delta_1) + \mathbf{1}\{\text{Control chosen}\} g_{Control}(\delta_2)] \\ &+ (NS(n_1 + n_2) - (n_1 + n_2)d) g_{Control}(\delta_2) \end{aligned} \quad (2)$$

where the functions *h* and *g* are now expected gain per time unit. Again – since there are no study specific costs for patients treated out-trial – it is assumed that the out-trial-gain of the superior treatment is at least as large as either of the in-trial-gains. Like in the acute treatment situation, we determine the sample sizes n_1 and n_2 to maximize the overall expected gain, now given by (2) instead of (1).

[Figure 2 around here]

2.4 | Summary of the three approaches

Table 1 summarizes the specifications needed for the three sample size approaches discussed here. Irrespective which of the approaches is used, it is good practice to investigate the robustness of the calculated sample size with respect to changes in the specified values. To plot sample sizes versus a range of parameter values around the specified value helps understanding of the robustness and for discussions. If the result is that a certain parameter has big influence on the sample size, one can go back and critically reflect how likely certain values are.

[Table 1 around here]

3 | CASE STUDIES

3.1 | Case study 1: New cellular therapy for Lyell's disease

Stevens–Johnson syndrome and toxic epidermal necrolysis, also called Lyell's disease, are severe adverse drug reactions and are considered variants of the same pathologic process, differing only in severity.[14] The disease is characterized by necrolysis of the epithelium of skin and mucous membranes. Mortality is approximately 22% in Europe and estimated to be 9% in the French Referral Centre for Toxic Bullous Diseases where special disease management is applied.[15] The incidence is estimated at 2 per 1 million inhabitants in Europe.

A new cellular therapy is considered as treatment; however, treatment costs are high. A single arm Phase I/II trial is planned to investigate the effect of the cellular therapy. This disease is considered as acute: following successful therapy complete healing can be expected after an average of two weeks. The gain function specified by equation (1) is therefore used in the decision-theoretic approach and hence, we assume that every patient of a target population of size

N is treated either in the trial or afterwards outside the trial with the recommended treatment. The primary endpoint for the efficacy evaluation is complete healing at Day 7.

The costs per patient and gain for a successfully treated patient were discussed in repeated meetings between the responsible physician and two statisticians. The cellular treatment is particularly expensive such that the costs are currently 25 000 EUR per patients. In case of a successful trial, a larger scale production for the treatment is expected to reduce the cost considerably to 5 000 EUR per patient which we use for the out-trial costs. Fewer days in hospital with fewer medical exams and fewer co-morbidities for a successfully treated patient compared to a non-successfully treated patient lead to a positive gain for a successfully treated patient corresponding to 100 000 EUR since the daily costs in these intensive care units are particularly high.

Let p_1 be the response rate of the new therapy, p_2 the response rate of the current treatment. The in-trial gain for a patient, who in this single-arm trial will receive the new therapy, is then (in 1000 Euro): $h_{New}(p_1)=100\ p_1$-25. The out-trial gain is: $g_{New}(p_1)=100\ p_1$-5 or $g_{Control}(p_2)=100\ p_2$ depending on which treatment is recommended after the trial. This means that to achieve the same out-trial gain, the new cellular treatment needs to have a 5%-units higher response rate to compensate the higher costs. The total number of was considered to be N=500. This was based on the incidence of 2 per 1 million inhabitants and a judgement of the physicians how many of these patients could be reached with the new cellular therapy.

As there was uncertainty in judging the prior assumptions for the new therapy, the sample size was calculated for a set of different prior assumptions and for different target differences in the traditional approach. The expected success rate prior to the trial ranged from 0.5 (pessimistic, equal to current treatment) to 0.9 (optimistic hopes of the physician). For the assurance and decision-theoretic approach, we consider therefore the prior mean being between 0.5 and 0.9. Further, the weight of this prior trial information was assumed to correspond to between 2 and 20 patients, reflecting that there is only vague knowledge based on expert beliefs. Statistically, this prior information can be quantified using a beta distribution with two parameters a and b such that the mean is $a/(a+b)$ and the weight is $a+b$. We show densities of assumed priors in Figure 3 for 0.5, 0.55, 0.6, …, 0.9 and for weights corresponding to information from 2, 10, and 20 patients. Note that the case of prior mean = 0.5 and information = 2 patients corresponds to an uninformative prior which judges all treatment-response-rates equally likely (uniform distribution for all rates between 0 and 1). The response rate of the current treatment is judged to be $p_2=0.5$ based on experience according to the discussions between responsible physician and statisticians. In the actual study planning, it was agreed not to include an uncertainty for this proportion $p_2=0.5$ in the model – however an alternative would have been to consider a prior distribution for p_2 based on previous data with mean 0.5 and higher weight than the prior distribution for p_1 since the physicians have experience with the current treatment and there is less uncertainty than for

the new treatment. The parameters needed for the sample size approaches are summarized in Table 2.

[Table 2 and 3 around here]

Table 3 shows the sample size for the traditional, the assurance approach, and the decision-theoretic approach; for the first two approaches we specified a type I error of α=5% (two-sided) and require a power of 80%. We see that the sample sizes are in some cases unreasonably high for the first two approaches in relation to the population size $N=500$. From a theoretical perspective, the traditional approach can be viewed as assurance approach with infinitively large prior weight, therefore we can view the first four lines in Table 3 as assurance approach where the prior weight is infinity in the first row. When we look at these first four rows in Table 3 for fixed assumed mean response rate for new treatment (looking at each column separately), we see that the sample size is increasing when the prior weight decreases from infinity down to 2. Hence, with increasing prior uncertainty, larger sample sizes are needed.

The optimal sample sizes for the decision-theoretic approach in the last three rows of Table 3 range up to 17. When the prior belief is that the new treatment is equal or only a little better than the standard treatment (prior mean for p_1 slightly above 0.5), the highest sample sizes are necessary for this decision-theoretic approach. In this situation more evidence for the treatment recommendation to the N-n patients outside the trial is needed. For large prior means (e.g. prior mean > 0.7 for prior weight = 10), even no trial can be best when it is optimal to recommend the new therapy based on the prior belief only. However, before really choosing to conduct no trial, one would need to ensure that all stakeholders agree with such a prior belief which led to sample size=0. Looking at all sample size calculations for the considered priors, a sample size of 15 was chosen for the trial which seems appropriate for all priors as it is close to the maximum optimal sample size.

We see here a large difference between the recommended sample sizes from traditional or assurance approach versus the decision-theoretic approach. Especially if it is assumed that the new cellular therapy is only a little better than the previous treatment, a lot of trial-information is necessary to statistically demonstrate this small difference which increases the traditional and assurance sample size. However, in our situation for treatment of Lyell's disease, it makes no sense to include a large part of the total population in the trial before making a recommendation for future patients. A justified treatment recommendation has the benefit for the society to reduce future treatment costs. Therefore, it was decided to apply a decision-theoretic approach based on specification of trial costs and – importantly – costs for future treatment and with this aim for optimizing the total treatment gain. Using this approach, it was accepted that we cannot necessarily hope to prove that one treatment is better if the treatment effects are very similar. The use of decision-theoretic sample size calculation for this clinical trial was accepted by ethical

committee and the regulatory agency.

[Figure 3 around here]

3.2 | Case study 2: Treatment of adult-onset Still's disease with Interleukin I (IL-1) antagonists

Adult-onset Still's disease is a rare chronic symptomatic disease. According to Gerfaud-Valentin et al.,[16] "the reported prevalence rates range from 1 to 34 cases per 1 million persons" in Japan and Europe and they refer to published values of 1.6 and 4 (per 1 million) in France and Norway, respectively. If we assume approximately an incidence of 2 per 1 million in the EU and a total EU population of roughly 500 million, the population size with the disease is around $N=1000$ in EU.

When the work with this case study was done, no pharmaceutical treatment was licensed for the disease. A mechanistic justification has been proposed for treatment of adult-onset Still's disease with IL-1 antagonists and a few observational studies, as well as one randomised controlled study, have been conducted.[17] Given the information from these previous studies, we consider here the sample size determination for a potentially upcoming trial comparing the IL-1 antagonist anakinra with control when patients are randomized with 1:1-allocation. The treatment duration in the upcoming trial is going to be 6 months and the intended primary outcome is remission rate. After the planning work with this case study, canakinumab, a fully human monoclonal anti-human IL-1 beta antibody, has been approved for the treatment of adult-onset Still's disease.

Hong et al.[17] report in their meta-analysis that 36 of 47 patients treated with anakinra and 33 of 68 patients treated with control experienced remission. Therefore, we assume for the traditional approach, described in Section 2.1 above, a remission rate of 0.766 and 0.485 for treatment and control, respectively. For the assurance and decision-theoretic approach we assume that the remission rate has a beta prior distribution. Appropriate choices for treatment and control are Beta(36,11) and Beta(33,35), respectively corresponding to the data from Hong et al.[17] reported above, see Figure 4. For the traditional and assurance approach, we specify a type I error of α=5% (two-sided) and require a power (assurance) of 80% (β=20%).

[Figure 4 around here]

As this is a chronic condition, for the decision-theoretic approach we use the formulation of the gain function given by equation (2). We assume that after the trial, a treatment recommendation is made for either the new drug or control. Until the patients can be treated according to this recommendation, they are treated with control (if they are not in the trial). In total, our time horizon H is 10 years. After these 10 years, we expect other improved

treatments to take over. However, since there is uncertainty around this time horizon specification, we will below check how the sample size depends on changes in this parameter. The treatment recommendation is assumed to be made at time $S = S(n_1 + n_2) = t_0 + t_1(n_1 + n_2)$ years with t_0=2 (6 months treatment duration and an assumed 1.5 years from the end of the trial until the average start of recommended treatment) and recruitment speed of t_1=1/40 years per patient (we expect that recruitment of each 40 patients takes approximately one year). This recruitment speed is more optimistic than the experience from Nordström et al.[18] who could include 22 patients in 2 years, as it is hoped for more centres contributing to recruitment.

Let the unknown probability of remission be p_1 after anakinra treatment and p_2 after control treatment. We define the benefit of being in remission for 1 year as 1 unit (considering the whole H=10 year period where we want to optimize treatment). The assumed cost for being in the trial was judged to be c_P=0.05, i.e. the costs for 20 patients included in the trial was judged to correspond to the gain of one more patient experiencing remission. Outside of the trial, there are fewer costs; one more remission justifies treatment costs for 100 patients (c_O=0.01). Consequently, the in-trial gain is $h_{New}(p)=h_{Control}(p)=p-0.05$ per year and the out-trial gain is $g_{New}(p)=g_{Control}(p)=p-0.01$ per year, where $p=p_1$ for anakinra, $p=p_2$ for control treatment. For simplicity, we assume here that remission or non-remission is directly achieved and constant over the in-trial or out-trial treatment period. As we have restricted the possible allocation in the trial to 1:1, we have $n_1=n_2=n$. Table 4 summarizes the chosen parameters.

[Table 4 around here]

With these specifications, we can calculate recommended sample sizes according to the three approaches. Sample size for the traditional approach is $n_1=n_2=46$ to achieve 80% power based on the assumed remission rates of 0.766 and 0.485. To ensure 80% assurance, a sample size of $n_1=n_2=56$ is necessary based on the above specified beta prior distribution.

The decision-theoretic approach which recommends the treatment with the higher posterior expected gain for the future, recommends in this situation not to run any trial at all but to rely fully on the prior distribution. The new treatment is recommended directly. It is not beneficial to postpone the start of out-trial treatment S to collect more data on the treatment-control comparison when we have the overall gain function as specified. One reason is that the prior distributions here (Figure 4) are very distinct, implying that the prior probability is 99.9% for anakinra being better than control. If we would have prior distributions which are much closer to each other than those here (i.e. if the prior mean for anakinra would be between 0.53 and the control mean 0.485 leaving anything else unchanged, implying a prior probability of 68.2% of anakinra being better than control), this decision-theoretic approach leads to optimal sample sizes >0. Another reason is that we have assumed that there are no specific costs for anakinra

compared to control. However, one should reflect both additional monetary costs for the anakinra treatment and in general reflect concerns about safety differences: how much larger does the remission rate for anakinra need to be to compensate for additional monetary costs and for a potential increased safety burden? Even if we assume that 15 percentage points better remission rate are needed for anakinra compared to control (which is judged as a conservative assumption for this case study), we still get the decision-theoretic sample size recommended to be 0 confirming the support for not doing any trial and recommending anakinra directly based on the decision-theoretic approach (the prior probability is 93.3% that anakinra is at least 0.15 better than control).

For illustrative purposes, we assume now that the additional anakinra economic and safety costs correspond to c_T=0.3, i.e. 30 percentage points higher remission needed for anakinra compared to control, $h_{New}(p_1)=p_1-c_P-c_T$ and $g_{New}(p_1)=p_1-c_O-c_T$. In this case, we obtain the optimal sample size n*=45. The left panel in Figure 5 shows how the decision-theoretic sample size depends on the additional anakinra costs c_T. For c_T<0.25, no trial is recommended and anakinra should be the treatment recommendation for all out-trial patients. For c_T between 0.25 and 0.37 a trial is recommended with roughly increasing sample size which is at most n=61. For c_T>0.37, no trial is recommended and the treatment recommendation is to use control treatment. Using c_T=0.3, the middle panel in Figure 5 shows how the sample size depends on the population size N: If N is smaller than the assumed N=1000, smaller sample sizes are necessary. At N=1000, a plateau is (almost) reached (for very higher population sizes, n*=47 would be the optimal sample size). Finally, using again c_T=0.3 and N=1000, the right panel in Figure 5 shows the dependency on the time horizon H being in the interval from 8 to 15 years. The longer the time horizon, the larger is the sample size which is justified, but the dependency is not too drastic.

[Figure 5 around here]

In this case study (as well as in the following case study), we determined sample sizes for trials with 1:1-allocation. The traditional and assurance approach can easily be modified by computing power/assurance for another prespecified allocation ratio. In the same way, we can apply the decision-theoretic approach for any prespecified allocation ratio. However, when using the decision-theoretic approach, we have even the possibility to optimize the allocation ratio by searching the optimal pair (n_1, n_2), i.e. by dropping the restriction $n=n_1=n_2$ which we made in the case studies.

3.3 | Case study 3: Treatment of cystic fibrosis with inhaled dry powder mannitol (Bronchitol)

Cystic fibrosis is a progressive, genetic disease that causes persistent lung infections and limits the ability to breathe over time.[19] The prevalence was determined to be around 0.7 per 10 000 in average in the EU-countries and around 0.8 per 10 000 in the US.[20] No cure exists; treatment is currently symptomatic and chronic. In two clinical Phase III trials, called CF301 and CF302,

inhaled dry powder mannitol was investigated as treatment. The primary endpoint was lung function measured as forced expiratory volume in 1 s (FEV_1). While results of CF301 showed a statistical significant advantage for mannitol,[21] in the primary analysis of the second pivotal study "statistical significance was narrowly missed" (p=0.058).[22] FDA required an additional Phase III trial in an adult cystic fibrosis population, called CF303, before potentially licencing on the US market. The trial is currently ongoing. We pretend here that we aim to calculate the sample size for an upcoming trial given the results of CF301 and CF302, with the US market in mind.

To form prior distributions, we combine the results from the two available trials. For the mannitol treatment, the average change in FEV_1 over the 26 weeks double-blind phase was 111 ml (based on 177 patients in CF301 and 184 patients in CF302). For the control treatment, the average was 42 ml (based on 118 patients in CF301 and 121 patients in CF302). Given the confidence intervals in study CF302 for mannitol, the standard deviation was assumed to be 295 ml. Therefore we assumed a $N(111,16^2)$ distribution for mannitol and a $N(42,19^2)$ distribution for control (the variance in the mannitol prior is e.g. calculated as $295/\sqrt{177+184} \approx 16$). The prior for the treatment difference δ is then $N(\delta_0,\sigma_0^2)$ with δ_0=111-42=69 and $\sigma_0^2 = 16^2 + 19^2 \approx 25^2$. For the traditional sample size calculation, we can assume the treatment difference 69 ml and a standard deviation of 295 ml. In the traditional and assurance approach, we use α=0.05 (two-sided) and an (expected) power requirement of 1-β=0.8.

Again, we use for the decision-theoretic approach the chronic disease formulation of the gain function given by equation (2). The population size on the US market is around N = 325 000 000 x 0.8 / 10 000 = 26 000. According to the timelines planned for recruitment in trial CF303,[23] the sponsor anticipated to need around two years for 440 patients. Therefore, we assume a recruitment speed of 20 patients/month (240/year) and that the recommended treatment is available after $S = S(n_1 + n_2) = t_0 + t_1(n_1 + n_2)$ years with t_0=2 and recruitment speed of t_1=1/240 (fraction of a year for recruiting one patient). We assume a treatment horizon H=10 years for the treatment and will again as in the second case study investigate also values in the interval 8-15 years later. As a 26 week treatment period is desired, we use d=0.5 years.

In this case study, we use the base case (default treatment policy) to treat each of the N patients during the entire time of H years with the control treatment outside of any trial. We set the expected gain of this base case to 0 implying $g_{Control}(\delta_2)$=0 and are interested in the change of the gain when using another treatment policy.

For specifying the gain functions, we translate clinical results (FEV_1 improvement) into an economic value. According to Table 5 from DeWitt et al.[24], the mean total annual health care costs (without medication) decrease by 1700 USD when the variable percent predicted FEV_1 (FEV_1%) is increased with 1 %-unit. In trials CF301 and CF302, the average treatment effect of mannitol was FEV_1 = 69 ml and, measured in the variable FEV_1%, was 3.5%, meaning that a

change of FEV_1%=1% corresponded roughly to a change of FEV_1=20 ml. Therefore, we justify a decrease of 1700 USD in costs when FEV_1 is increased by 20 ml or in other words an increase of 1 ml in FEV_1 is expected to reduce health care costs by c_U=85 USD per year. We therefore use the gain function $\delta\, c_U$ (where δ is the mean difference in FEV_1 between treatment and control in ml) and deduce c_P=5000 USD for a patient in the trial and further c_T=6000 USD per year treatment costs for a patients' treatment with mannitol (not for the control treatment). Using the functions h and g introduced in (2), we can write $h_{New}(\delta)= \delta c_U - c_T - c_P$, $h_{Control}=-c_P$, $g_{New}(\delta)= \delta c_U - c_T$, $g_{Control}=0$.

As in the previous case study, we restrict the possible allocation in the trial to 1:1 and have $n_1=n_2=n$. A traditional sample size calculation for ensuring 80% power when the treatment difference is 69 ml with an assumed standard deviation of 295 ml for a significance test having α=5% (two-sided) yields a sample size of n=288 patients per group (for 1:1-allocation). Based on the prior distribution for the difference between mannitol and control, $N(69,25^2)$, we compute the sample size ensuring 80% assurance to be n=390.

[Table 5 around here]

In this situation with a normally distributed prior and observations, the gain function can be computed to have a quite simple form. We show this in the appendix following the computation by Willan[25] and Pearce et al.[26] in a similar situation but for an acute treatment. Optimizing this computed gain function over the sample size, we obtain the decision-theoretic sample size as n^*=221.

When applying the decision-theoretic approach it is essential to check how changes in assumed parameters change the situation. Some of the above assumptions were well justified by evidence from literature, others are more rough assumptions. To identify the parameters with important influence on the sample size helps to focus on those when critically reflecting the assumptions. In Figure 7, we show the influence of 9 parameters on the optimal sample size n^*. We keep the other 8 parameters constant with values as described above and vary one parameter at a time.

In the upper-left panel of Figure 7, we see that when changing the prior mean δ_0, the optimal sample size is 0 for small δ_0<35 ml where the control treatment is recommended and no attempt would be made to get the new treatment approved for this pessimistic prior. The sample size is between 300 and 360 for δ_0 between 35 ml and 59 ml. The optimal sample size is then decreasing for increasing prior mean until δ_0=82 ml. For larger prior means, the recommendation would be for mannitol without further trial (n=0). However in a licencing situation, decision for active treatment without trial would require that the authority agrees with the prior mean as well. Figure 6 shows the probability to choose the mannitol treatment after the study depending on the prior mean δ_0 when the optimal n^* is chosen. The optimal sample size is decreasing with increasing prior variance σ_0^2 and increasing with increasing variance of the observations σ^2 (second and third panel in top row of Figure 7), but the influence is not too strong.

Looking at the cost parameters (panels in the middle row in Figure 7), we observe basically no influence of the trial cost per patient c_P. In contrast, the annual treatment costs c_T and gain per increased FEV_1-unit c_U are critical for the sample size. For small gains per increased FEV_1-unit, c_U<59 USD, no trial is recommended and the control treatment should be taken. The highest optimal sample size is then n=346 for c_U=64 USD and is then decreasing with increasing gain. For c_U = 105 USD or more, mannitol should be recommended directly (n*=0). In this case, the gain outweighs the treatment costs based on the prior for the treatment difference. From the formula for the gain function we see that the results depend on c_U and c_T through c_T/c_U only. Therefore, the dependence on c_T when c_U is fixed is related. As noted in the appendix, the ratio c_T/c_U has the following interpretation: When the posterior mean for the treatment difference is at least c_T/c_U, mannitol should be recommended; when it is smaller, control should be recommended.

When the population size is at least N=10 000, the optimal sample size is basically constant – for very large population sizes, the optimal sample size tends to n*=227. If the population size would decrease below N=10 000, the optimal sample size will also decrease. We see some but not too drastic influence of the recruitment speed $1/t_1$ (the faster the recruitment, the higher sample size is justified) and of the treatment horizon H (the longer the treatment horizon, the higher the optimal sample size).

Overall, we conclude from these figures that the treatment difference in the prior δ_0 and the gain/cost-ratio c_T/c_U are the most critical parameters for the sample size. These should therefore be considered carefully in this situation before finally deciding on the sample size.

Comparing the recommended sample sizes from the traditional or assurance approach with the optimal decision-theoretic sample size, we see here only a relatively small discrepancy, especially if we use some more conservative assumptions based on Figure 7 and use e.g. the maximum sample size recommended in these sensitivity calculations shown in the figure.

[Figure 6 and Figure 7 around here]

4 | SPECIFICATION OF THE TREATMENT EFFECT

In the traditional approach, we need a specification for the targeted treatment effect. There are different ways of interpreting this target treatment effect: one possibility is to justify a “minimal clinically important difference” (MCID) and to require the desired power for this difference. While this concept is broadly applied, Burman and Carlberg[27] question the existence of a MCID by discussing that even very small differences in effect are important given two drugs with otherwise identical profile in e.g. safety, price, and dosing schedule. Another way of interpreting the target effect is to anticipate the likely effect difference which might be based on earlier experience. This interpretation is also often used when choosing the target effect for the

traditional sample size determination. With this interpretation and when prior distributions are used with very small variability around the likely values in the assurance approach, i.e. when the uncertainty becomes smaller and smaller, then the assurance approach converges to the traditional approach.

We have seen in the case studies that the assurance approach led to higher sample sizes compared to the traditional approach with the prior mean equal to the target effect. This can be explained since for the sample size leading to a traditional power of 80-90%, the assurance is lower since the power function is concave around the target effect so that the average of the power is lower than the power at the average assumed effect size.

For the assurance and decision-theoretic approach, a prior distribution for the treatment effect needs to be specified, which can be challenging in applications. In our first case study, there was only vague information elicitated from experts. We investigated therefore the optimal sample sizes for a large set of different priors and have chosen a sample size close to the maximum of all recommended sample sizes. In the other two case studies, results from earlier studies were available and formed beta and normal priors, respectively. The underlying justification was: When using uninformative priors before the earlier, published studies, the posterior after the earlier studies is the likelihood which is then used as prior for the study we are planning. However, there are limitations with this way of specifying the priors: Firstly, the earlier results may have a risk for bias, e.g. if they are from early phase, from unblinded or non-randomized studies. This could be especially a risk in Case Study 2. Secondly, starting with an uninformative prior before the earlier results might be challenged as well since e.g. for a disease difficult to treat where several other compounds failed, an uninformative prior could be too optimistic and more pessimistic priors might be justified.

5 | DISCUSSION AND CONCLUSION

In this manuscript, we have compared the traditional, assurance and decision-theoretic sample size approaches. These are not the only possible approaches. In some situations the interest is not in hypothesis testing but in estimating e.g. a treatment difference with good precision. Then the sample size can be motivated to ensure a certain confidence interval length for the treatment difference. When the intention is to analyse the data with Bayesian methods but a decision-theoretic specification of gain functions is not desired, Bayesian sample size determination as described by Adcock[28] can be applied. Several criteria to ensure limited length of posterior credibility intervals can be defined to determine a Bayesian sample size; see Joseph and Bélisle[29] for normal means and their differences and M'Lan, Joseph and Wolfson[30] for the binomial case. Moreover, aspects from different approaches might be combined: E.g. a significance test to decide upon treatment recommendation could be incorporated into the decision-theoretic frame[26],[31],[32]. Uncertainties about cost- or recruitment-parameters might be handled using prior distributions for these parameters as well.

We recommend not to use a single approach in isolation, but to compare results of several ways to determine sample size. An informed choice of sample size is ideally made after challenging the specifications made using several of the considered approaches. When very different recommendations result from different approaches the reasons should be understood before a choice is made.

We considered a single efficacy endpoint in this work but there are situations where several endpoints are important for sample size determination. A sample size chosen based on considerations for efficacy might be too small to ensure a sufficiently large safety database and then the minimum requirement for safety would define the sample size. In other situations when specific important safety endpoints are identified, these could be integrated in the decision-theoretic framework by using a utility score as a primary endpoint (for an example of a utility score incorporating efficacy and safety see Ouellet et al.[33]) Alternatively, the decision-theoretic framework could be used with more complex models with several endpoints, specifying priors, and gain functions based on them.[34-35]

An important and often difficult step in sample size determination is to specify the required parameters for these methods. It is good practice for all sample size approaches to consider robustness of the sample size by checking the influence on sample size when some of these specifications are changed. The parameters which are especially influential can be critically reflected, maybe re-discussed with subject-matter-experts and if necessary revised. In the investigated case studies, the parameter for the additional treatment cost of the new treatment and the assumed difference in prior means were parameters highly influencing the sample size for the decision-theoretic approach. The importance of the treatment cost parameter for sample size and treatment decision making was highlighted by Pearce et al.[26]

If the intention of the trial is to support drug licensing by regulatory agencies, the question arises if a decision-theoretic approach is acceptable. A critical parameter in the planning of the decision-theoretic design is in our view the treatment cost parameter: this parameter can reflect safety costs and can then be interpreted as safety penalty. The new treatment is only recommended for future treatment of patients when the data suggests that the effect is sufficiently better than control's effect to justify these safety costs. Therefore, we think that especially this parameter, in addition to the prior distribution for the effect, needs to be carefully discussed with regulatory agencies. If this is done, the decision-theoretic approach is appealing for a licencing situation. While the traditional and assurance approach build on an underlying significance test which aims to demonstrate that the new treatment is better than control, a positive demonstration says nothing about how much better the new treatment is and if it justifies the additional safety or monetary costs for the new treatment. In contrast, the decision-theoretic approach uses gain functions which connect directly to societal and regulatory considerations. By this, the approach defines the level of evidence required to make a treatment decision. Given the study with optimized sample size and treatment

decision rule, the corresponding type I and II errors can be calculated (see e.g. Stallard et al.[13] and Pearce et al.[26]).

For the decision-theoretic approach, the optimal sample size is increasing with increasing population size. Cheng et al.[12] and Stallard et al.[13] have shown for the acute case that the optimal sample size is proportional to the square root of N and is therefore unbounded. In contrast, we have seen for the approach with chronic diseases in Case study 2 and 3, that this optimal sample size reaches a plateau for large populations, i.e. is bounded. We explain this by following considerations: If N is relatively large, the trial covers only a small area of the total area of treatment need in Figure 2. It might then be reasonable to approximate the overall gain by ignoring the gain in the trial. The overall gain becomes

$$G(n_1, n_2, \delta_1, \delta_2) = N\{(H - S(n_1 + n_2))[\mathbf{1}\{\text{New treatment chosen}\}g_{New}(\delta_1) + \mathbf{1}\{\text{Control chosen}\}g_{Control}(\delta_2)] + S(n_1 + n_2)g_{Control}(\delta_2)\}. \quad (3)$$

The difference between (2) and the approximation (3) is that the trial part (dark blue and red in Figure 2) is handled as if these patients had been treated outside the trial with control. Since the overall gain in (3) is proportional to the population size, the sample size which maximises the gain depends no longer on the population size N. This means that if population size N increases, the optimal sample size according to (2) converges to a constant sample size (which can be computed based on (3)). These arguments showing that the optimal sample size has an upper bound were obtained assuming a recruitment speed independent of the population size. For smaller populations, it is reasonable that the recruitment depends heavily on the size of the population; for larger populations, we think it is meaningful that the speed cannot exceed a certain level even if the population is very big. The independence of recruitment speed on population size would be true for very big populations.

We have considered three case studies for rare disease with population sizes 500, 1000, or 26 000. There are also even smaller populations, “ultra-rare” diseases where e.g. around 100 or fewer patients exist worldwide. E.g. the Hutchinson–Gilford progeria syndrome which is a chronic disease of segmental premature aging syndrome and fatal by teenage years, had 54 known cases worldwide in 2009 and 146 in 2016.[36] In studies with enrolment during 2005-2006[37] and during 2007,[38] the authors state that they succeeded to include at least half of cases known at enrolment. A currently ongoing trial with objective to study survival when treated aims to include 80 patients.[36] The approach to sample size in the last study was to include as many as possible from the children with disease. A requirement for this approach is the existence of a good patient registry which is available here. In the case of an ultra-rare disease, the approach to collect as much information as possible instead of applying a sample size approach as considered in this paper seems to be the best alternative.

ACKNOWLEDGMENTS

We thank the principal investigator of the Lyell's disease study for discussing the studies background. This work was conducted as part of the InSPiRe (Innovative methodology for small populations research) project funded by the European Union Seventh Framework Programme for research, technological development and demonstration under grant agreement number FP HEALTH 2013602144.

REFERENCES

[1] R. V. Lenth. Some Practical Guidelines for Effective Sample Size Determination. *The American Statistician* **2001**, *55*, 187-193.

[2] S. A. Julious. Sample sizes for clinical trials with normal data. *Stat. Med.* **2004,** *23*, 1921-1986.

[3] S. A. Julious, M. Campbell. Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data. *Stat. Med.* **2012,** *31*, 2904-2936.

[4] A. O'Hagan, J. Stevens, M. Campbell. Assurance in clinical trial design. *Pharm. Stat.* **2005**, *4*, 187-201.

[5] C. Chuang-Stein. Sample size and the probability of a successful trial. *Pharm. Stat.* **2006**, *5*, 305–309.

[6] N. Stallard, M. Posch, T. Friede, F. Koenig, W. Brannath. Optimal choice of the number of treatments to be included in a clinical trial. *Stat. Med.* **2009**, *28*, 1321-1338.

[7] A. O'Hagan, C. Buck, A. Daneshkhah, J. Eiser, P. Garthwaite, D. Jenkinson, J. Oakley, T. Rakow. *Uncertain Judgements: Eliciting Experts' Probabilities.* Wiley, Chichester **2006**.

[8] S. Zohar, I. Baldi, G. Forni, F. Merletti, G. Masucci, D. Gregori. Planning a Bayesian early-phase phase I/II study for human vaccines in HER2 carcinomas. *Pharm. Stat.* **2011**, *10*, 218–226.

[9] N. Kinnersley, S. Day. Structured approach to the elicitation of expert beliefs for a Bayesian-designed clinical trial: a case study. *Pharm. Stat.* **2013**, *12*, 104-113.

[10] K. K. G. Lan, J. T. Wittes. Some thoughts on sample size: A Bayesian-frequentist hybrid approach. *Clin. Trials* **2012**, *9*, 561-569.

[11] S. W. Hee, T. Hamborg, S. Day, J. Madan, F. Miller, M. Posch, S. Zohar, N. Stallard. Decision theoretic designs for small trials and pilot studies: a review. *Stat. Methods Med. Res.* **2016**, *25*, 1022-1038.

[12] Y. Cheng, S. Fusheng, D. A. Berry. Choosing sample size for a clinical trial using decision analysis. *Biometrika* **2003**, *90*, 923-936.

[13] N. Stallard, F. Miller, S. Day, S. W. Hee, J. Madan, S. Zohar, M. Posch. Determination of the optimal sample size for a clinical trial accounting for the population size. *Biom. J.* **2017**, *59*, 609-625.

[14] M. Lissia, P. Mulas, A. Bulla, C. Rubino. Toxic epidermal necrolysis (Lyell's disease). *Burns* **2010,** *36*, 152-163.

[15] L. Valeyrie-Allanore, S. Ingen-Housz-Oro, O. Chosidow, P. Wolkenstein. French referral center management of Stevens-Johnson syndrome/toxic epidermal necrolysis. *Dermatologica Sinica* **2013**, *31*, 191-195.

[16] M. Gerfaud-Valentin, Y. Jamilloux, J. Iwaz, P. Sève. Adult-onset Still's disease. *Autoimmun Rev.* **2014**, *13*, 708–722.

[17] D. Hong, Z. Yang, S. Han, X. Liang, K. Ma, X. Zhang. Interleukin I inhibition with anakinra in adult-onset Still disease: a meta-analysis of its efficacy and safety. *Drug Design, Development and Therapy* **2014**, *8*, 2345-2357.

[18] D. Nordström, A. Knight, R. Luukkainen, R. van Vollenhoven, V. Rantalaiho, A. Kajalainen, J. G. Brun, A. Proven, L. Ljung, H. Kautiainen, T. Petterson. Beneficial effect of interleukin I inhibition with anakinra in adult-onset Still's disease. An open, randomized, multicenter study. *The Journal of Rheumatology* **2012**, *39*, 2008-2011.

[19] Cystic Fibrosis Foundation. About Cystic Fibrosis, https://www.cff.org/What-is-CF/About-Cystic-Fibrosis/ (accessed: December 2016).

[20] P. M. Farrell. The prevalence of cystic fibrosis in the European Union. *Journal of cystic fibrosis* **2008**, *7*, 450-453.

[21] D. Bilton, P. Robinson, P. Cooper, C. G. Gallagher, J. Kolbe, H. Fox, A. Jaques, B. Charlton, for the CF301 Study Investigators. Inhaled dry powder mannitol in cystic fibrosis: an efficacy and safety study. *Eur. Respir. J.* **2011**, *38*, 1071–1080.

[22] M. L. Aitken, G. Bellon, K. De Boeck, P. A. Flume, H. G. Fox, D. E. Geller, E. G. Haarman, H. U. Hebestreit, A. Lapey, I. M. Schou, J. B. Zuckerman, B. Charlton, for the CF302 Investigators. Long-Term Inhaled Dry Powder Mannitol in Cystic Fibrosis: An International Randomized Study. *Am. J. Respir. Crit. Care Med.* **2012**, *185*, 645–652.

[23] U.S. National Institutes of Health. A Safety and Efficacy Trial of Inhaled Mannitol in Adult Cystic Fibrosis Subjects, https://www.clinicaltrials.gov/ct2/show/NCT02134353 (accessed: December 2016).

[24] E. M. DeWitt, C. A. Grussemeyer, J. Y. Friedman, M. A. Dinan, L. Lin, K. A. Schulman, S. D. Reed. Resource Use, Costs, and Utility Estimates for Patients with Cystic Fibrosis with Mild Impairment in Lung Function: Analysis of Data Collected Alongside a 48-Week Multicenter Clinical Trial. *Value in health* **2012**, *15*, 277-283.

[25] A. R. Willan. Optimal sample size determinations from an industry perspective based on the expected value of information. *Clin. Trials* **2008**, *5*, 587–594.

[26] M. Pearce, S. W. Hee, J. Madan, M. Posch, S. Day, F. Miller, S. Zohar, N. Stallard. Value of information methods to design a clinical trial in a small population to optimise a health economic utility function. **2017** Manuscript.

[27] C. F. Burman, A. Carlberg. Future challenges in the design and ethics of clinical trials. In SC Gad (ed.), *Clinical Trials Handbook*, Wiley **2009**, pp. 1173–1200.

[28] C. J. Adcock. Sample Size Determination: A Review. *J. Royal Stat. Soc.: Series D (The Statistician)* **1997**, *46*, 261-283.

[29] L. Joseph, P. Bélisle. Bayesian sample size determination for normal means and differences between normal means. *J. Royal Stat. Soc.: Series D (The Statistician)* **1997**, *46*, 209-226.

[30] C. E. M'Lan, L. Joseph, D. B. Wolfson. Bayesian sample size determination for binomial proportions. *Bayesian Analysis* **2008**, *3*, 269-296.

[31] T. Ondra, S. Jobjoernsson, R. Beckman, C. Burman, F. Koenig, N. Stallard, M. Posch. Optimizing Trial Designs for Targeted Therapies. *PLoS ONE* **2016**, *11*, e0163726.

[32] A. Graf, M. Posch,, F. Koenig. Adaptive designs for subpopulation analysis optimizing utility functions. *Biom. J.* **2015**, *57*, 76-89.

[33] D. Ouellet, J. Werth, N. Parekh, D. Feltner, B. McCarthy, R. L. Lalonde. The use of a clinical utility index to compare insomnia compounds: A quantitative basis for benefit-risk assessment. *Clin. Pharmacol. Ther.* **2009**, *85*, 277-282.

[34] N. Stallard, P. F. Thall, J. Whitehead. Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics* **1999**, *55*, 971–977.

[35] T. Kikuchi, J. Gittins. A behavioral Bayes method to determine the sample size of a clinical trial considering efficacy and safety. *Stat. Med.* **2009**, *28*, 2293–2306.

[36] Progeria Research Foundation, www.progeriaresearch.org (accessed: March 2017).

[37] M. A. Merideth, L. B. Gordon, S. Clauss, V. Sachdev, A. C. M. Smith, M. B. Perry *et al*. Phenotype and course of Hutchinson–Gilford Progeria Syndrome. *New England J. Med.* **2008**, *358*, 592-604.

[38] L. B. Gordon, M. E. Kleinman, D. T. Miller, D. S. Neuberg, A. Giobbie-Hurder, M. Gerhard-Herman, L. B. Smoot, C. M. Gordon, R. Cleveland, B. D. Snyder, B. Fligor, W. R. Bishop, P. Statkevich, A. Regen, A. Sonis, S. Riley, C. Ploski, A. Correia, N. Quinn, N. J. Ullrich, A. Nazarian, M. G. Liang, S. Y. Huh, A. Schwartzman, M. W. Kieran. Clinical trial of a farnesyltransferase inhibitor in children with Hutchinson-Gilford progeria syndrome. *Proc. Natl. Acad. Sci.* **2012**, *109*, 16666-16671.

Tables

Table 1 Required specifications for sample size calculation

<table>
<tr><th>Traditional approach</th><th>Assurance approach</th><th>Decision-theoretic approach</th></tr>
<tr><td>Target / assumed effect of treatment and control</td><td>Prior distribution for effect of treatment and control</td><td>Prior distributions for effect of treatment and control</td></tr>
<tr><td>Nuisance parameters</td><td>Nuisance parameters or prior distribution for them</td><td>Nuisance parameters or prior distribution for them</td></tr>
<tr><td colspan="2">Type I error allowed (α)</td><td rowspan="2">Gain functions for patients treated with new treatment and treated with control (for in-trial patients and out-trial patients) including cost parameters</td></tr>
<tr><td rowspan="2">Power required (1-β)</td><td rowspan="2">Expected power (assurance) required (1-β)</td></tr>
<tr><td>Size of the population; for the chronic case even recruitment speed and time horizon</td></tr>
</table>

Table 2 Parameter specifications for Lyell's disease case study

Parameter	Specification
Significance level (*α*)	0.05 (two-sided)
Required power (1-*β*) or assurance	0.80
Prior mean proportion, new treatment	Uncertain, range 0.5 to 0.9 considered
Prior information weight, new treatment	Uncertain, range 2 to 20 considered
Mean proportion, current treatment	0.5
Gain for successfully treated patient	100 000 EUR
Costs for one patient in trial (cellular treatment)	25 000 EUR
Costs for one patient outside trial with cellular treatment	5 000 EUR
Population size *N*	500

Table 3 Sample size for Lyell's disease trial for traditional and assurance approach

Target / assumed mean response rate for new treatment (control response rate = 0.5)	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90
Traditional approach	>500	197	88	50	32	22	17	13
Assurance, prior weight=20	*	>500	283	88	44	27	18	13
Assurance, prior weight=10	*	*	>500	158	59	31	20	14
Assurance, prior weight= 2	*	*	*	*	>500	79	27	18
Decision-theoretic, prior weight=20	17	14	0	0	0	0	0	0
Decision-theoretic, prior weight=10	17	16	14	9	0	0	0	0
Decision-theoretic, prior weight= 2	11	12	12	11	11	10	9	8

*for these cases, the assurance would be < 80% even for infinitively large sample size since there is >20% prior-probability for the new treatment's response rate to be worse than previous treatment

Table 4 Parameter specifications for adult-onset Still's disease case study

Parameter	Specification
Significance level (α)	0.05 (two-sided)
Required power (1-β) or assurance	0.80
Prior mean proportion, treatment	0.766
Prior information weight (in patients), treatment	47
Prior mean proportion, control	0.485
Prior information weight (in patients), control	68
Gain for successfully treated patient	1 unit
Costs for one patient in trial	0.05 units
Costs for one patient outside trial	0.01 units
Recruitment function $S(n)$ (in years)	$2+n/40$
Time horizon H and treatment time d (in years)	H=10, d=0.5
Population size N	1000

Table 5 Parameter specifications for cystic fibrosis case study

Parameter	Specification
Significance level (α)	0.05 (two-sided)
Required power (1-β) or assurance	0.80
Prior mean difference (in ml), treatment-control	69
Prior information standard deviation for treatment difference	25
Standard deviation of observations (ml)	295
Gain c_U from 1 ml increase in FEV_1 (in USD)	85
Costs c_P for one patient in trial (in USD)	5000
Costs c_T for new treatment per year (in USD)	6000
Recruitment function $S(n)$ (in years)	$2+n/240$
Time horizon H and treatment time d (in years)	H=10, d=0.5
Population size N	26 000

Figures

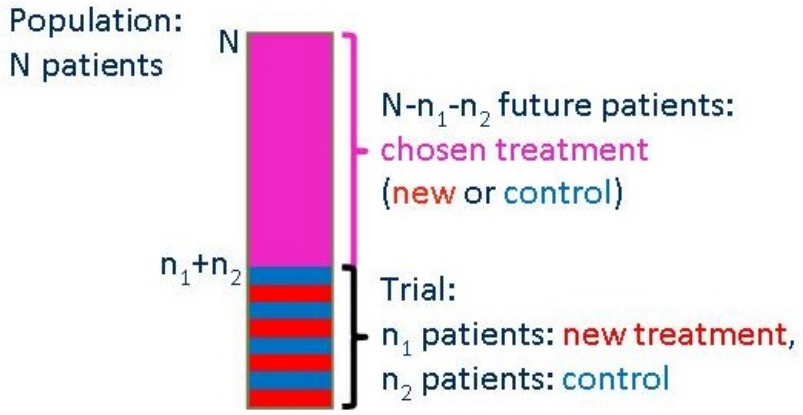

Figure 1 Basic assumption for decision-theoretic approach

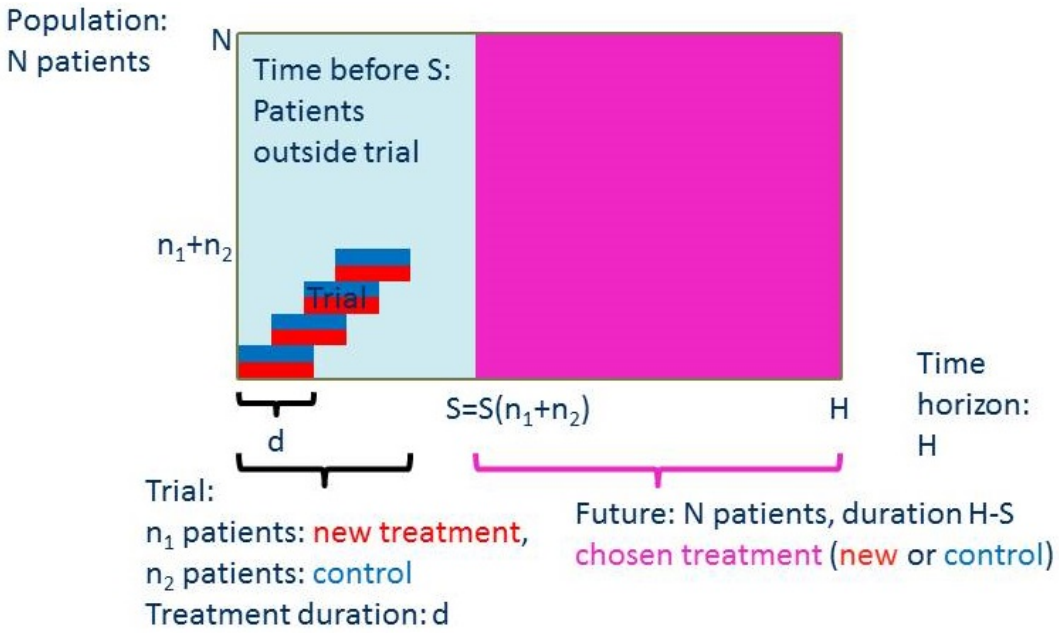

Figure 2 Basic assumption for decision-theoretic approach in chronic diseases

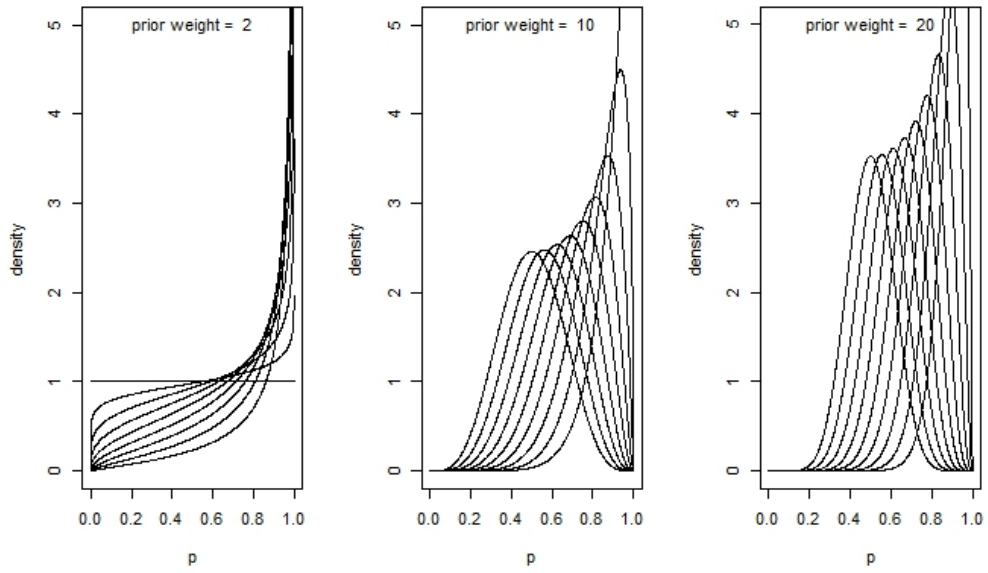

Figure 3 Prior distributions considered for Lyell's disease case study: Beta distributions with weight = 2 (left panel), 10 (middle panel), 20 (right panel), and mean = 0.5, 0.55, 0.6, …, 0.9 (from left to right in each panel)

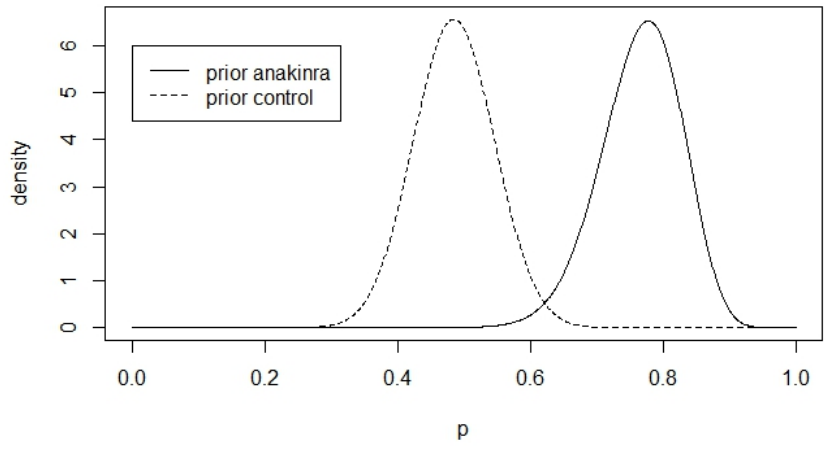

Figure 4 Prior distributions for case study on adult-onset Still's disease: Anakinra has Beta(36,11)-prior and control has Beta(33,35)-prior

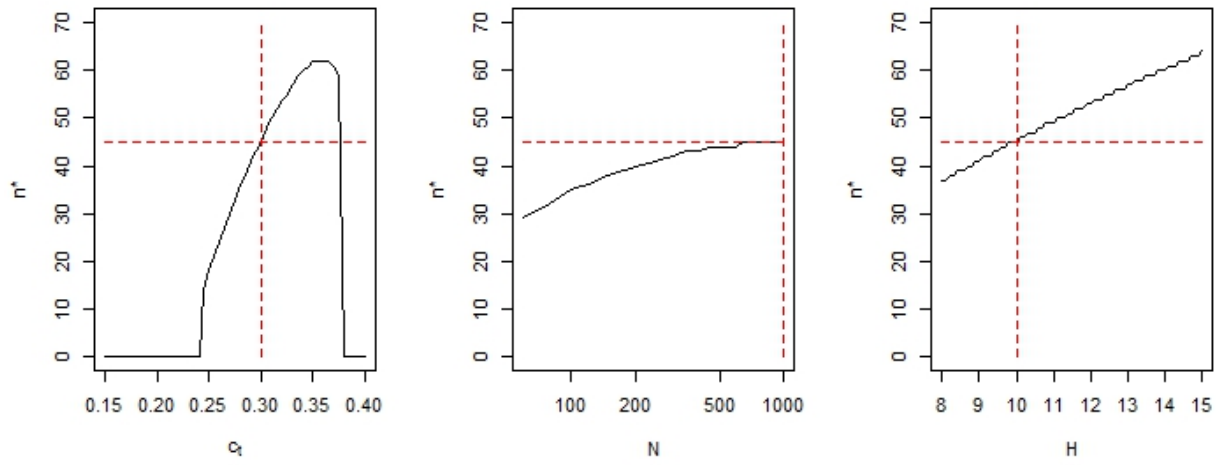

Figure 5 Anakinra case study: Dependence of optimal decision-theoretic sample size n^* on treatment costs c_T, on population size N, and on the time horizon H. Vertical dashed lines mark the values c_T=0.3, N=1000, and H=10 discussed in the text which give the optimal sample size n^*=45 (horizontal dashed lines).

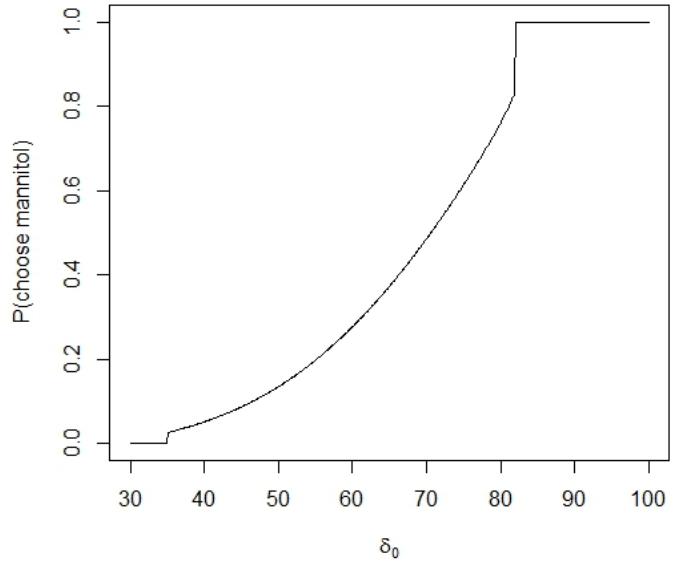

Figure 6 Probability to choose the mannitol treatment when the optimal decision-theoretic design is used depending on the prior mean δ_0

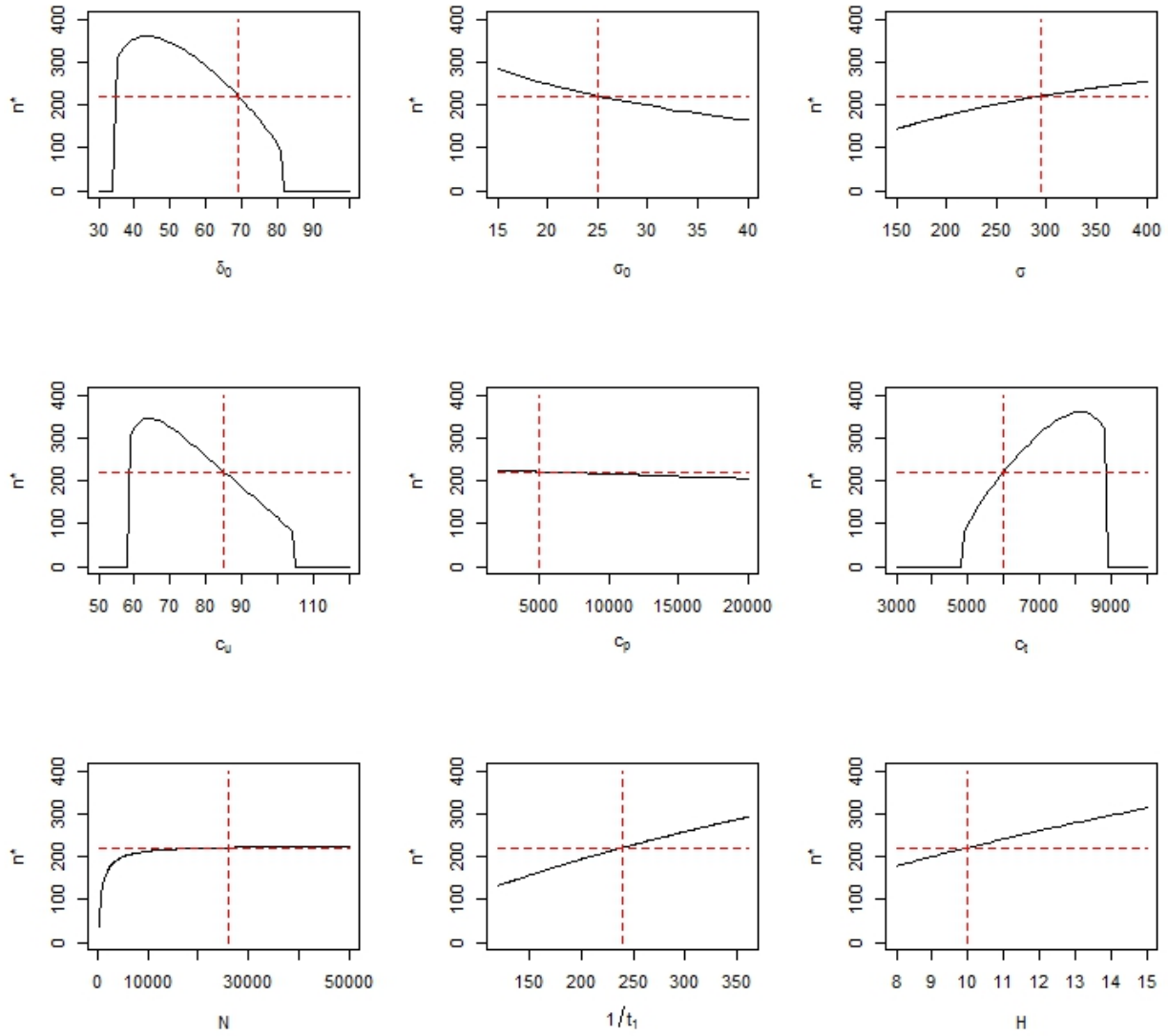

Figure 7 Influence of parameter changes on optimal decision-theoretic sample size *n for cystic fibrosis case study. Vertical dashed lines mark the motivated parameter specifications which give the optimal sample size *n**=221 (horizontal dashed lines).**

APPENDIX | The expected gain for decision-theoretic approach in Case Study 3

When everyone in the population would be treated with control over the whole treatment horizon H outside of a clinical trial (n=0), the gain would be $N * H * g_{Control}(\delta_2)$. Our interest is how we can change this gain by using the new treatment (having effect δ_1 but additional costs of c_T per year) instead of control (having effect δ_2). The gain function (or more exactly the additional gain compared to treating everyone with control outside of the trial) is:

$$\begin{aligned} G(n_1, n_2, \delta) &= n_1 dh_{New}(\delta) + n_2 dh_{Control} \\ &+ N(H - S(n_1 + n_2))\mathbf{1}\{\text{New treatment chosen}\} g_{New}(\delta) \end{aligned} \tag{4}$$

which is for $n_1=n_2=n$ and the functions h and g from Case Study 3:

$$G(n,\delta) = n\, d\, (\delta c_U - 2c_P - c_T) + \mathbf{1}\{\text{New treatment chosen}\}\, N\, (H - S(2n))\, (\delta c_U - c_T).$$

The factor $(\delta c_U - c_T)$ is positive if $\delta > c_T/c_U$ and therefore it is optimal for the overall expected gain to recommend the new treatment if and only if the posterior mean for the treatment effect δ is at least c_T/c_U.

For the prior $N(\delta_0,\sigma_0^2)$ for δ and n observations made in two groups with mean difference δ and known variance σ^2 having observed mean difference $\bar{y}$, the posterior mean is

$$\frac{\frac{\delta_0}{\sigma_0^2} + \frac{n\bar{y}}{2\sigma^2}}{\frac{1}{\sigma_0^2} + \frac{n}{2\sigma^2}}$$

In a similar way as by Willan[25] and Pearce et al.,[26] we derive the expected gain by integrating over observed treatment difference and the parameter δ. Here, for the chronic disease case, we obtain

$$G(n)/c_U = \{(\delta_0 - c_T/c_U)\Phi(-z) + \sigma_0^2/\sigma_x \phi(z)\} N\{H - S(2n)\} + 2nd\{(\delta_0 - c_T/c_U)/2 - c_P/c_U\},$$

where $z = \frac{\sigma_x}{\sigma_0^2}(c_T/c_U - \delta_0)$ and $\sigma_x^2 = \sigma_0^2 + \frac{2\sigma^2}{n}$. We maximize this expression with respect to n.