Bayesian Inference Clinical Trials and Nonparametric Models

Non-Stat team

Part 1: Introduction and Conclusion

Determine the Maximum Tolerated Dose (MTD) in Phase I dose-finding clinical trials

- **Why frequency is not suitable for phase I dose-finding trials?**
- **Why traditional 3+3 not good enough?**
- **Why continual reassessment method (CRM) better?**

- **Introduction for our research question and finding.**
- **Why our project important?**
- **A brief conclusion for our project**

We will make deeper discussion in:

- **What our finding in Literature Review**
- **What our finding in Model Reproduction & Simulation**

Why not frequency?

Phase I dose-finding clinical trials

- Determine the MTD.
- Not fully understood the actions and side effects of new drugs in the human body.
- Experimenting at low doses level may delay patients' chances of treatment.
- I dose-finding trials often have a small sample size (around 3 each dose-level cohort).

Why frequency is not suitable for phase I dose-finding trials?

- Very small sample size, high likelihood of type II error in hypothesis test.
- Can not make good use of prior information (e.g. animal experiment results).
- Even reject the null hypothesis, can't say anything for the alternative hypothesis.

Traditional 3+3 method

Currently, traditional 3+3 method is the mainstream method for determining MTD in phase I dose-finding trials (by Tongtong's finding in literature review part).

- Clear results, easy to operate, widely accepted.
- Good in safety

Disadvantages:

- Cannot efficiently utilize all the information in the trials.
- Not flexible. May expose more patients to ineffective low doses level treatment.
- TTR (Target toxicity rate) is fixed at 33%.
- Unable to establish a dose-response curve.

Continual reassessment method (CRM)

Compared with the traditional 3+3 method, CRM has the following advantages:

- CRM makes better use of all the data in the trials.
- Can easily adjust the subsequent experiment plan during the trials.
- Therefore, CRM can reduce the sample size and time wasted during the low-dose level.
- Can analyze the complex relationship between dose and toxicity in different skeleton models.
- Can better reflect the relationship between toxicity and dose by dose-response curve.
- In the literature review part, many studies have shown that the CRM have better performance than 3+3.

Continual reassessment method (CRM)

CRM still has some potential limitations:

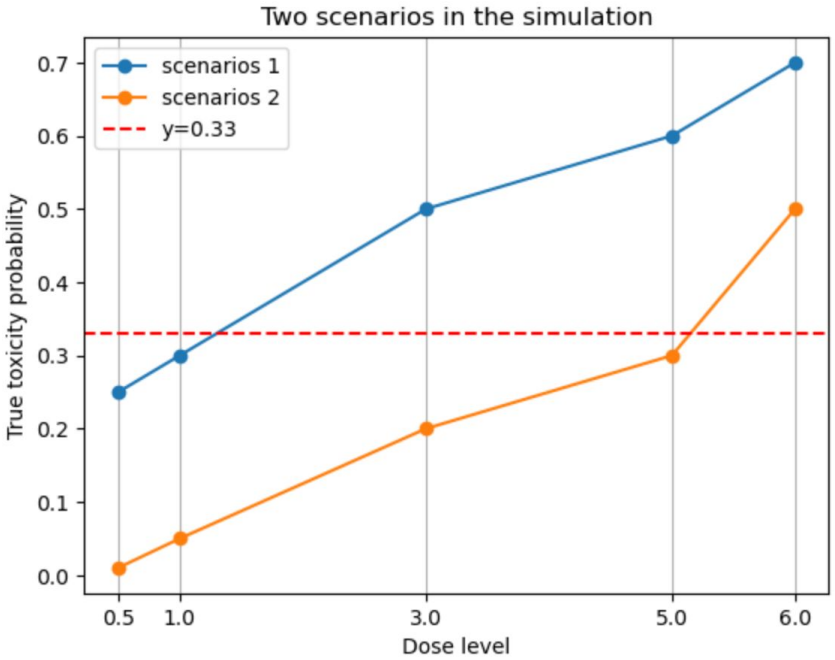
- Relatively complicated, no mature SOP, needs a statistical expert for inference and decision-making.
- Some (single-parameter) models may have a rapid dose level increase. May need ad-hoc rules.
- Should choose appropriate model and parameters for implementation.
- Might be intentionally chosen inappropriate models due to vested interests.

Why our project important?

Although CRM has been shown to be better performance than traditional 3+3, it is a complicated and flexibility model with no mature SOP. It is important to know how to choose suitable model for the various relationship between dose and toxicity. This will help the clinical team estimate MTD more accurately and helps in making the mature SOP to increasing the use of the CRM.

Introduction for Final Project Research

Analysis the performance of CRM method with different true toxicity probability drugs



True toxicity probability

Scenarios 1: (0.25, **0.3**, 0.5, 0.6, 0.7)

Scenarios 2: (0.01, 0.05, 0.2, **0.3**, 0.5)

Target toxicity level = 0.33

Probability in finding true MTD

- Scenarios 1 = 47.3%, Scenarios 2 = 69.2%
- 0.25 is closer to 0.33 than 0.20
- The start toxicity probability in Scenarios 1 is much higher than Scenarios 2

Which one is the reason for the difference?

0.20 and 0.25

- Scenarios 3: (0.25, 0.3, 0.5, 0.6, 0.7), Scenarios 4: (0.20, 0.3, 0.5, 0.6, 0.7)
- Scenarios 3 = 39.0%, Scenarios 4 = 48.8%
- The difference is around 10%.
- Hard to avoid. We do not know the true toxicity probability in the real world. And larger difference in does level will reduce accuracy and may increase the risk of DLT.

Start toxicity probability

- Scenarios 5: (0.2, 0.3, 0.4, 0.5, 0.6), Scenarios 6: (0.01, 0.1, 0.2, 0.3, 0.4)
- Scenarios 5 = 36.2%, Scenarios 6 = 48.4%
- The difference is around 12%.

Why start toxicity probability impact the performance of CRM?

Note in the simulation, we only have the sample size = 36, and in each cohort, we have 3 sample. The number of DLTs obeys a Binomial distribution.

True toxicity probability = 0.1

DLT	Prob
0	0.729
1	0.243
2	0.027
3	0.001

0.729 → 0.33 ← 0.500

True toxicity probability = 0.5

DLT	Prob
0	0.125
1	0.375
2	0.375
3	0.125

The variance is n*p*(1-p). We are facing higher uncertainty in the right case.

How can the difference be mitigated?

We try to do the following modifies:

- Make penalization for underdosing.
- Force the selection to go to dose 3 instead of dose 6 if first three patients do not get sick.
- Limit the simulations which are not just stick to one dose.
- Adding extra lower dose levels if the distribution is too skewed to the right.

Finally, the accuracy for scenario 1 increase from 48% to 62%.
(We will discuss more in the following)

A brief conclusion

- Rule-based designs like the 3+3 are still more prevailingly used now, but there many studies show that the model-based designs like CRM are performance better and can make more accurate MTD.

- The starting point of the dose-level is important for the performance of the CRM. In classic CRM designs, it may be better to set a conservative starting point.

- We can modify our model to mitigate the performance difference caused by the start point. In the next step, we can test the robustness of our new model.

Literature Review: Phase 1 trials (MTD)

- Goal of Phase 1 trials: identify maximum tolerated dose (MTD).
- MTD : the dose expected to produce some degree of medically unacceptable dose limiting toxicity (DLT) in a specified proportion θ of patients.

  Namely, Prob(DLT|Dose = MTD) = θ, where the proportion θ is also defined as the target toxicity level (TTL) (Babb and Rogatko 2004; Cater 1972).

- Dose-escalation trial design
  - Popular 3+3 design
  - Bayesian model-based designs
  - Bayesian curve-free methods
  - (Jaki, Clive, and Weir 2013)

Literature Review: Current Patterns

- Rule-based designs like the 3+3 design are still more prevailingly used.
- CRM or model based
  - 1.6% trials for cancer phase 1 trial between 1991 to 2006
  - 5.4% used for phase 1 oncology between 2008 to 2015
  - (Rogatko et al. 2007; Chiuzan et al. 2017).
  - 7% or 6 trials when reviewing 84 phase I cancer clinical trials (Le Tourneau et al. 2012)

Literature Review: Benefits of Model-based designs

- Use of all toxicity information accumulated during the trial
- Perform better in rapid information accumulation
- Reduce excessive exposure to subtherapeutic doses. (Le Tourneau, Lee, and Siu 2009).
- CRM is more likely to recommend the correct MTD and dose more trial patients close to the MTD.
- CRM and SM are comparable in terms of how fast they reach the MTD. As the number of dose levels was increased, CRM reached the MTD in fewer patients when used with a fixed sample of 20 patients (Iasonos et al. 2008).
- Flexible

Literature Review: How to implement CRM

- The continual reassessment method for dose-finding studies (Garrett-Mayer 2006)
- Handbook of Methods for Designing, Monitoring, and Analyzing Dose-Finding Trials
- Dose Finding by the Continual Reassessment Method

Literature Review: Other developments

- FDA guidance indicates Bayesian methods can be used in clinical trial for Drugs and Biologics.
- Preset levels using CRM is more acceptable to clinicians (Onar, Kocak, and Boyett 2009)
- Rolling-6 may be preferable over the CRM if very few or no toxicity is expected with the agent under study and if the dose finding period is long. (Onar-Thomas and Xiong 2010)

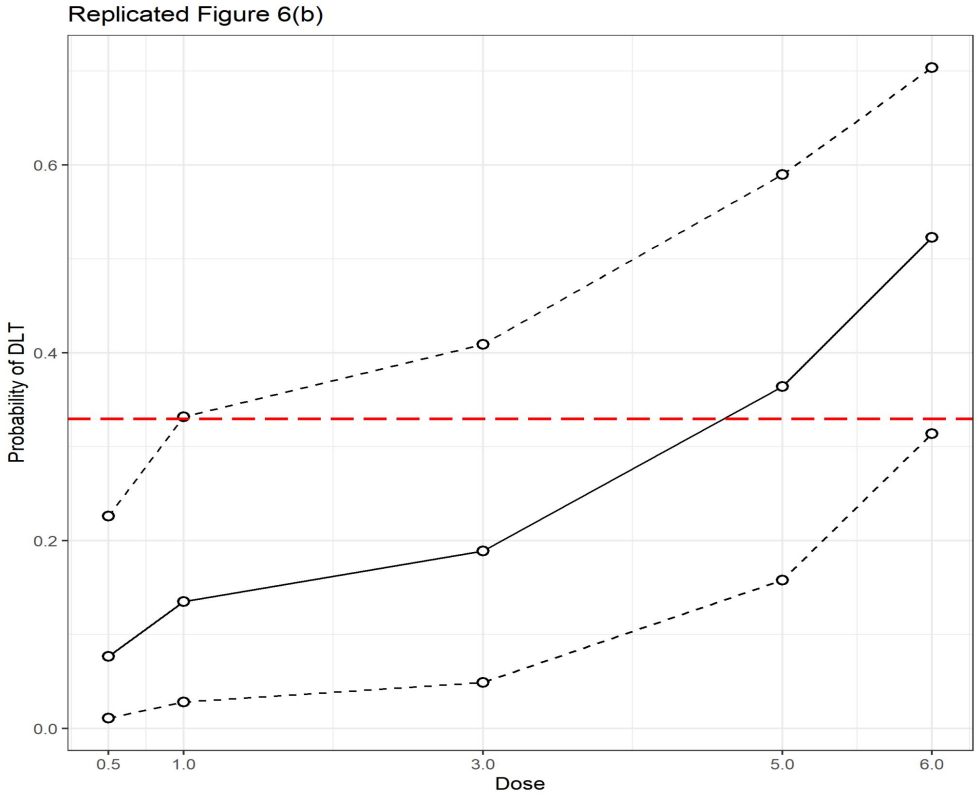
The set-up: original study

- The model: one-parameter logit;

$$\frac{exp(3+ exp\ (\beta)\ d)}{1+ exp(3+ exp\ (\beta)\ d)}$$

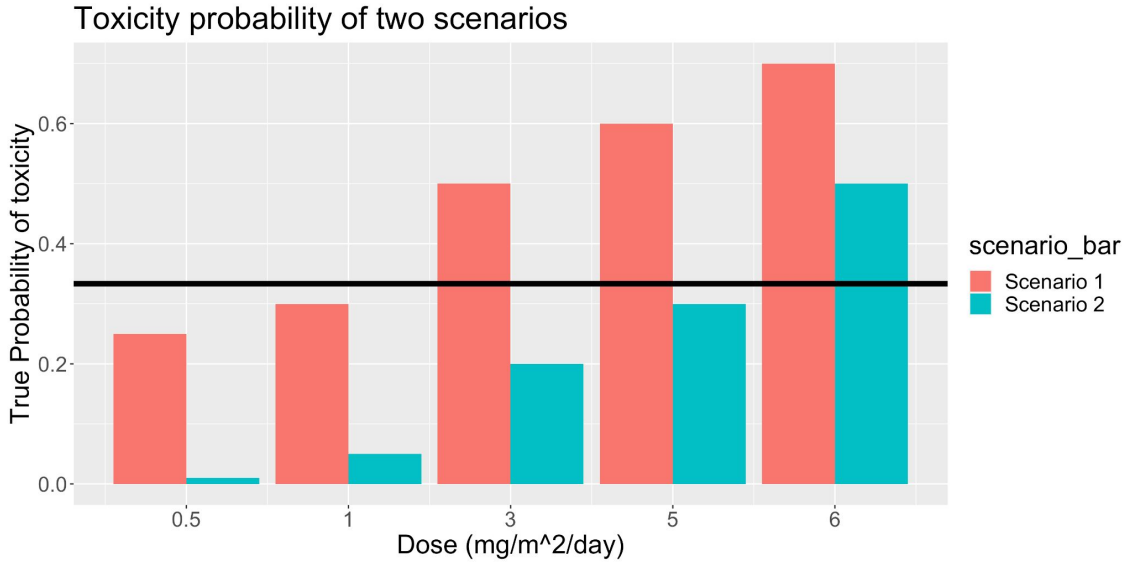
- Prior distribution of param: exponential with mean 0
- TTL: 0.33
- Drug-specific doses: (0.5, 1, 3, 5, 6)
- Skeleton probabilities: (0.05, 0.10, 0.15, 0.33, and 0.50)
- Escalation did not strictly conform to what the posteriors recommended

The set-up: replication of Figure 6



Q1: Simulation under two scenarios

- Target toxicity level: 33%
- Sample size: 36
- Cohort size: 3



Two-parameter logistic regression

Dose-toxicity model

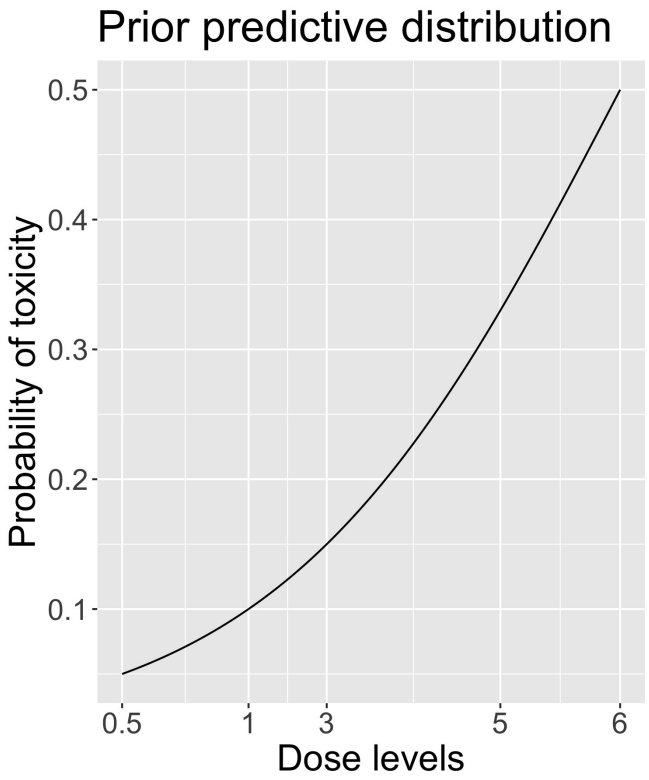
$$\frac{exp(\beta_1 + exp\ (\beta_2)\ d)}{1 + exp(\beta_1 + exp\ (\beta_2)\ d)}$$

d as in the general form of dose labels:

$$\frac{\ln\ (\frac{p_i}{1-p_i}) - \beta_1}{exp(\beta_2)}$$

Prior intercept mean as 0,

Prior slope mean as 1.



Simulation results

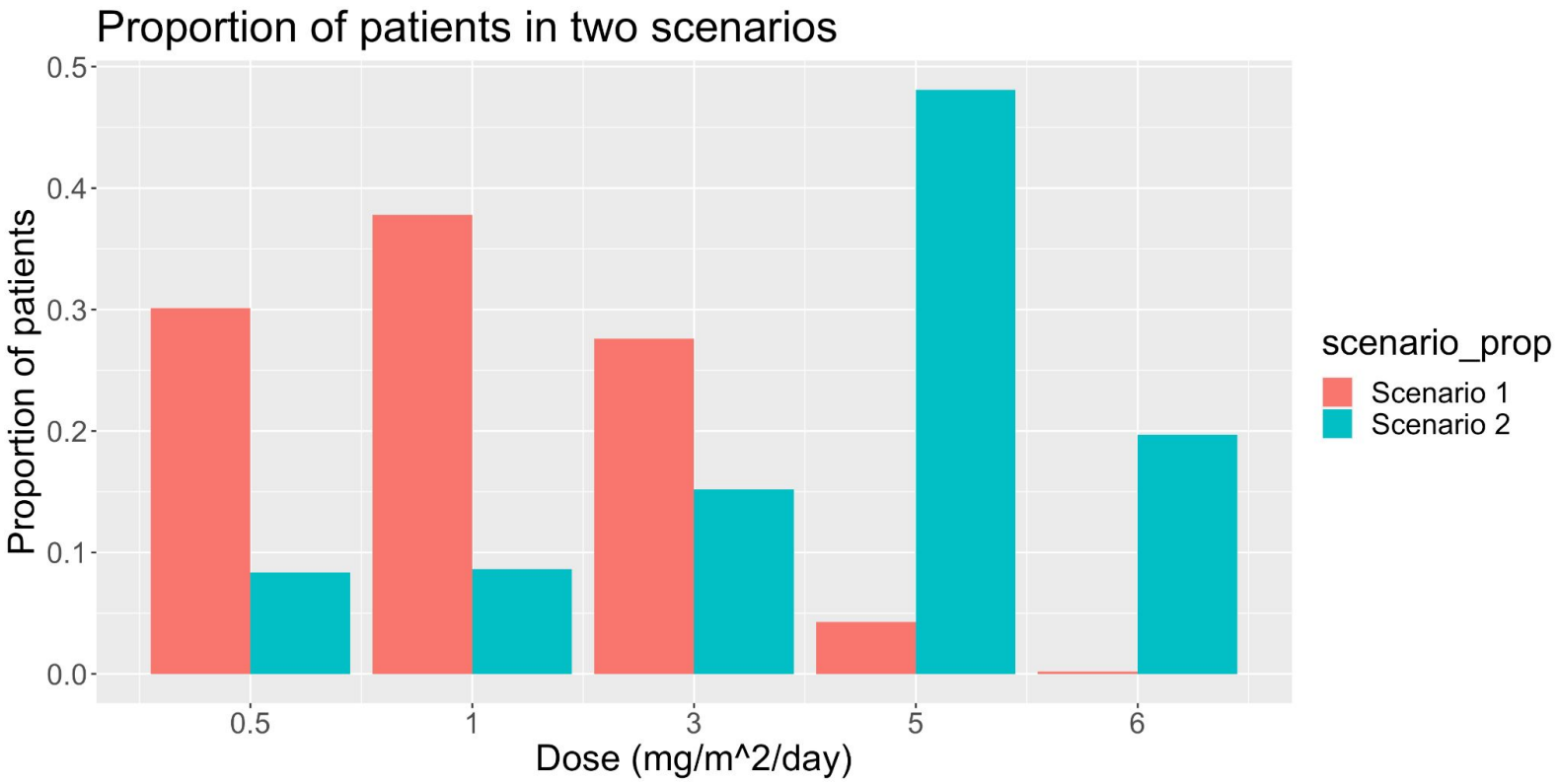
We calculate the posterior toxicity probability of each dose.

MTD is selected as the dose with the closest posterior probability to TTL (33%).

Proportion of each dose selected as MTD under the two scenarios:

Dose level (mg/m^2/day)	0.5	1	3	5	6
Scenario 1	0.241	**0.473**	0.275	0.011	0
Scenario 2	0	0	0.069	**0.692**	0.239

Simulation results – proportion of patients at each dose level



Q2 Why is there a difference?

1. Scenario 1 has two doses close to each other.

2. The simulation setup is favored toward Scenario 2 given the starting values.

Bonus Q. One potential solution

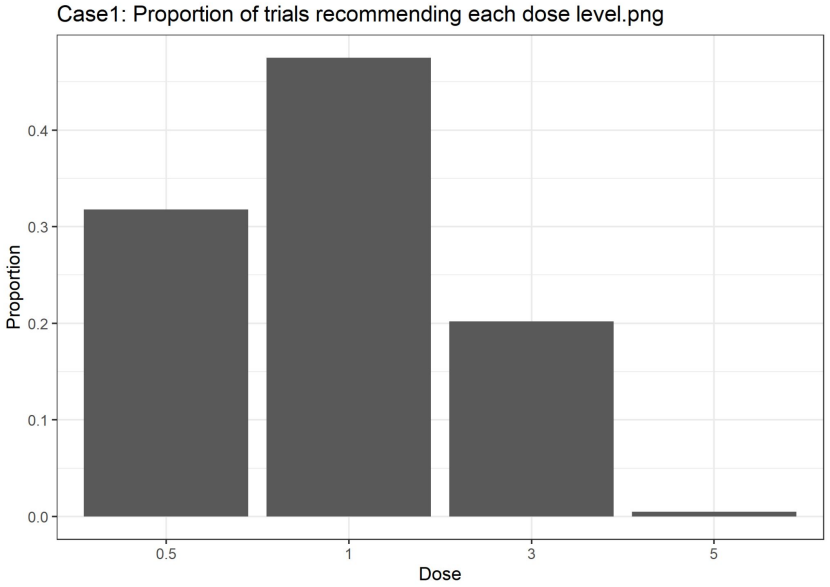
- Some penalization for underdosing (Neuenschwander, Branson, and Gsponer, 2008)

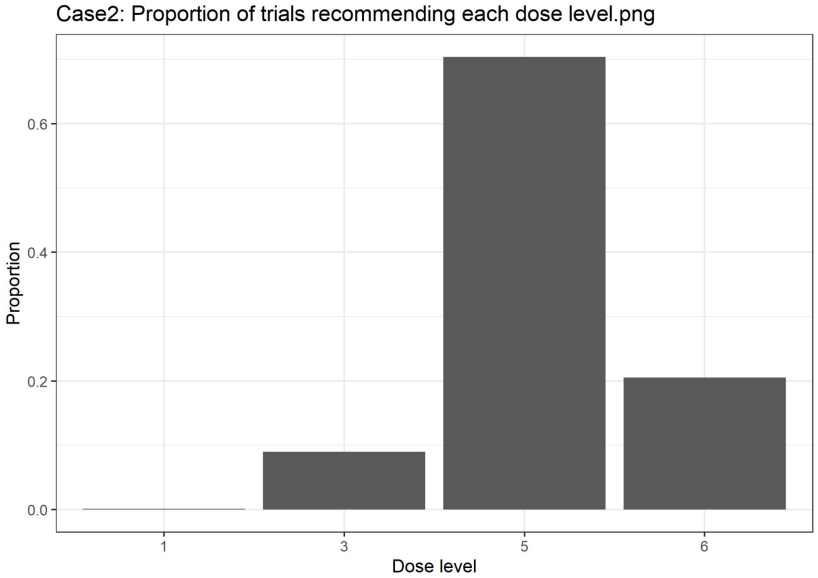
$$L(\theta, d) = \begin{cases} \ell_1 = 1 & \text{if } \pi_\theta(d) \in (0, 0.31] \\ \ell_2 = 0 & \text{if } \pi_\theta(d) \in (0.31, 0.35] \\ \ell_3 = 1 & \text{if } \pi_\theta(d) \in (0.35, 0.6] \\ \ell_4 = 2 & \text{if } \pi_\theta(d) \in (0.6, 1] \end{cases}$$

- Bayes risk:

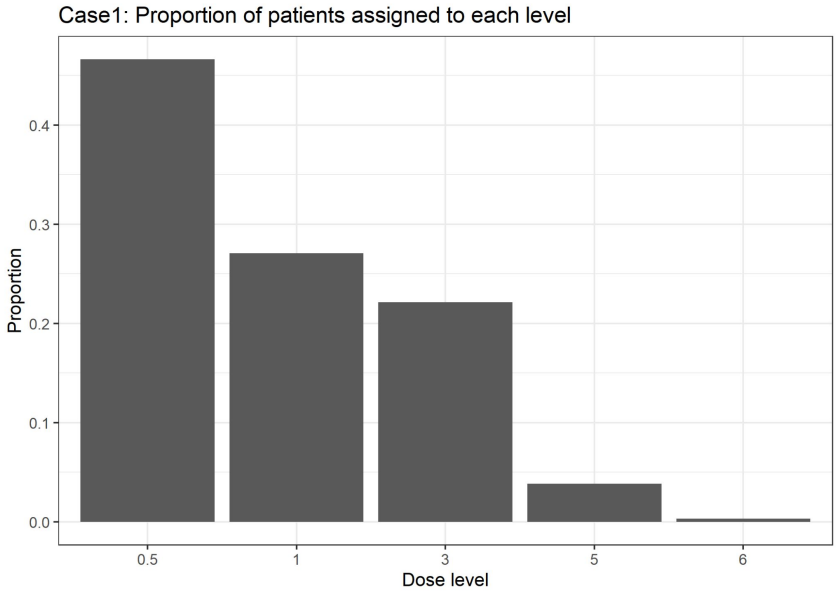
$$(\ell_1 P\{\pi_\theta(d) \in (0, 0.31]\} + \ell_2 P\{\pi_\theta(d) \in (0.31, 0.35]\} + \ell_3 P\{\pi_\theta(d) \in (0.35, 0.6]\} + \ell_4 P\{\pi_\theta(d) \in (0.6, 1]\})$$

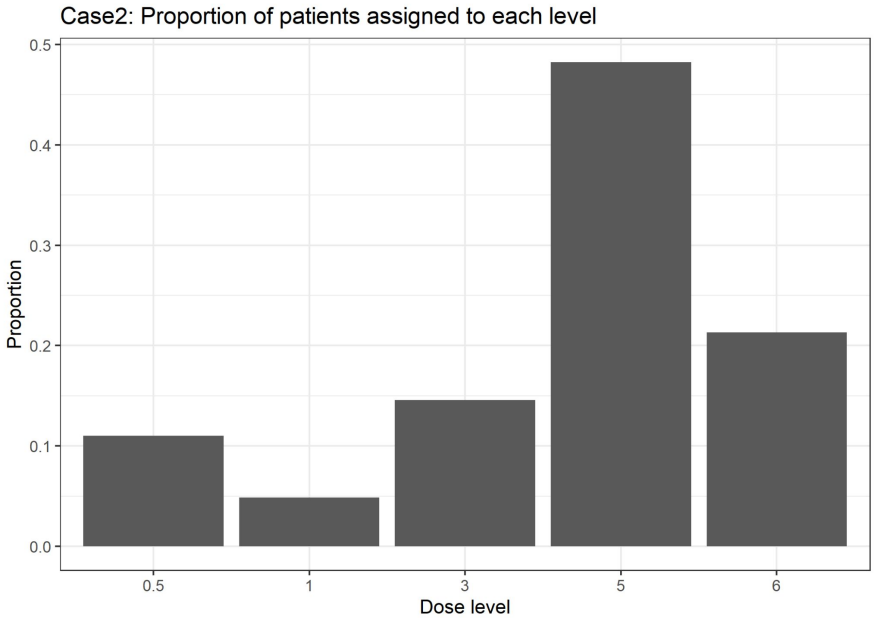
Results


Case1: Proportion of trials recommending each dose level.png
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
1
3
5
Dose


Case2: Proportion of trials recommending each dose level.png
Proportion
0.0
0.2
0.4
0.6
1
3
5
6
Dose level

Results


Case1: Proportion of patients assigned to each level
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
1
3
5
6
Dose level


Case2: Proportion of patients assigned to each level
Proportion
0.0
0.1
0.2
0.3
0.4
0.5
1
3
5
6
Dose level

Bonus Q. Can we reduce the difference?

- Main challenge is doing this in a non-data-dependent way.

- Jumping from 0 to 6 is a good place to start (they avoid doing this in the study).

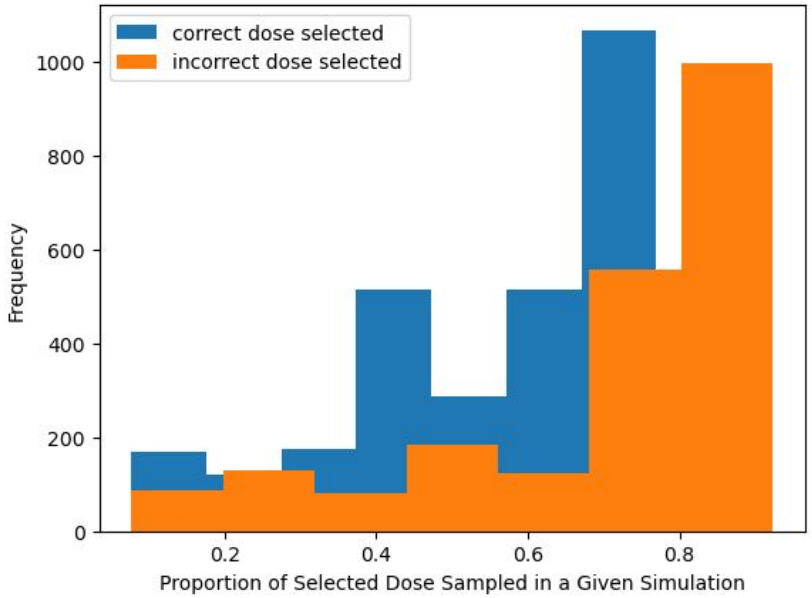
- The starting point matters a lot with low sample sizes.

Where does it go wrong?

Because the starting point is worse for scenario 1, it gets stuck sampling one dose more often.

```
selected_dose                                                    0
doses_sampled                    [0, 4, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]


selected_dose                                                    2
doses_sampled                    [0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]
proportion_selected_dose_sampled                          0.923077
```



Solution:

1. Force the selection to go to dose 3 instead of dose 6 if first three patients do not get sick.

2. Limit the analysis to simulations were the sampling does not just stick to one dose (< 90% cutoff).

Accuracy for scenario 1: 48.74% -> 62.4%

Accuracy for scenario 2: 62.4% -> 60.8%

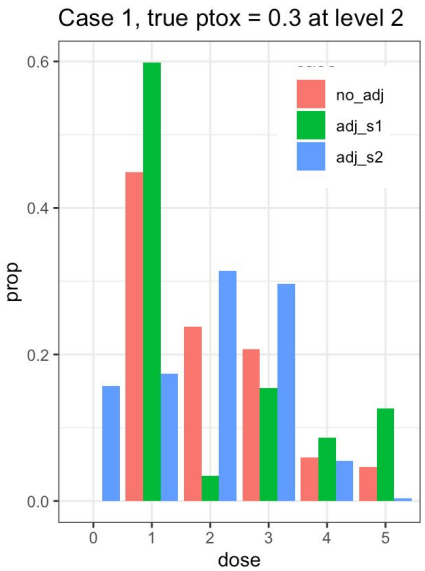
Bonus Q. One potential solution

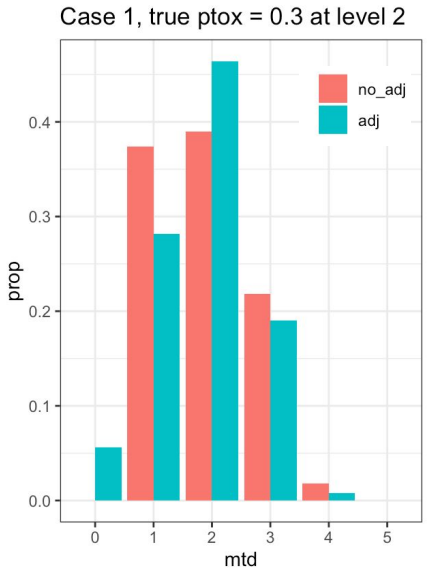
Idea: increase estimate accuracy by expanding the dose range.

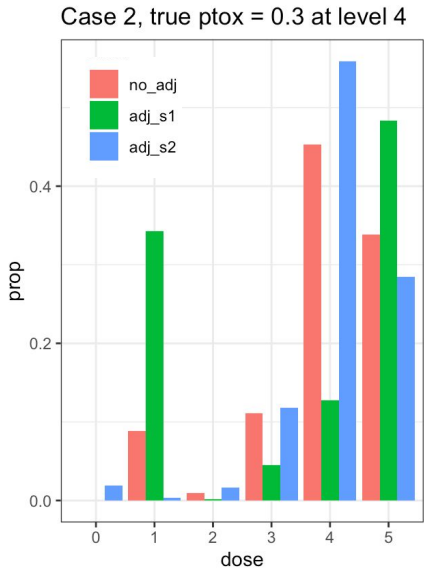
Motivation: in a linear regression $Y = a+bX+e$, $Var(b\_hat) = \sigma^2 (X'X)^{-1}$.

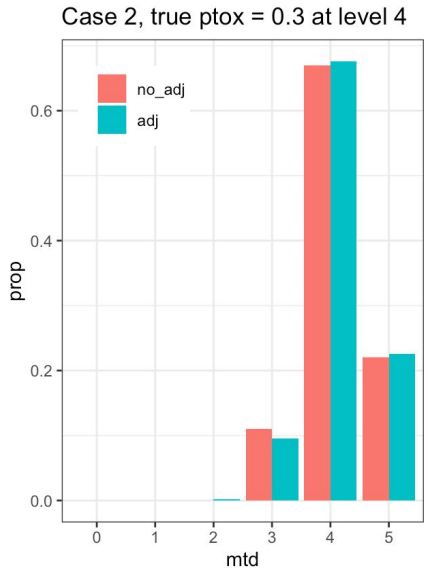
Design: At the 1/3 point of the trail, evaluate the dose visit frequency. If the distribution is too skewed to the right (most patients are assigned to dose level 1), an extra lower level 0 will be added for the remaining 2/3 patients. Otherwise, continue the previous design.

Bonus Q. Discrepancy mitigation and accuracy improvement









For scenario 1 and 2, level 0 was added to 79% and 17% simulated trials after evaluation, respectively.