A Simulation Evaluation of the Effectiveness and Usability of the 3+3 Rules-based Design for Phase I Clinical Trials

Jonathan Luu
Master of Science (Biostatistics)
Keck School of Medicine
University of Southern California

Degree Conferral Date: May 10, 2019

Table of Contents

1 Introduction

Clinical trials constitute a research methodology for the search and establishment of better treatment of human diseases. These trials are a form of a “planned experiment which involves patients and is designed to elucidate the most appropriate treatment for future patients with a given medical condition” (Pocock, 1984). The first of four phases in traditional clinical trials, phase I trials test the usability and safety of new agents. Many designs have been proposed to optimize phase I trials, but the traditional 3+3 rule-based design remains the most popular (Hansen et al, 2014). This study examines different attributes of the 3+3 design to determine its strengths and weaknesses under various conditions, as well as discusses the popularity of the 3+3 design among investigators despite it being heavily criticized.

2 Background

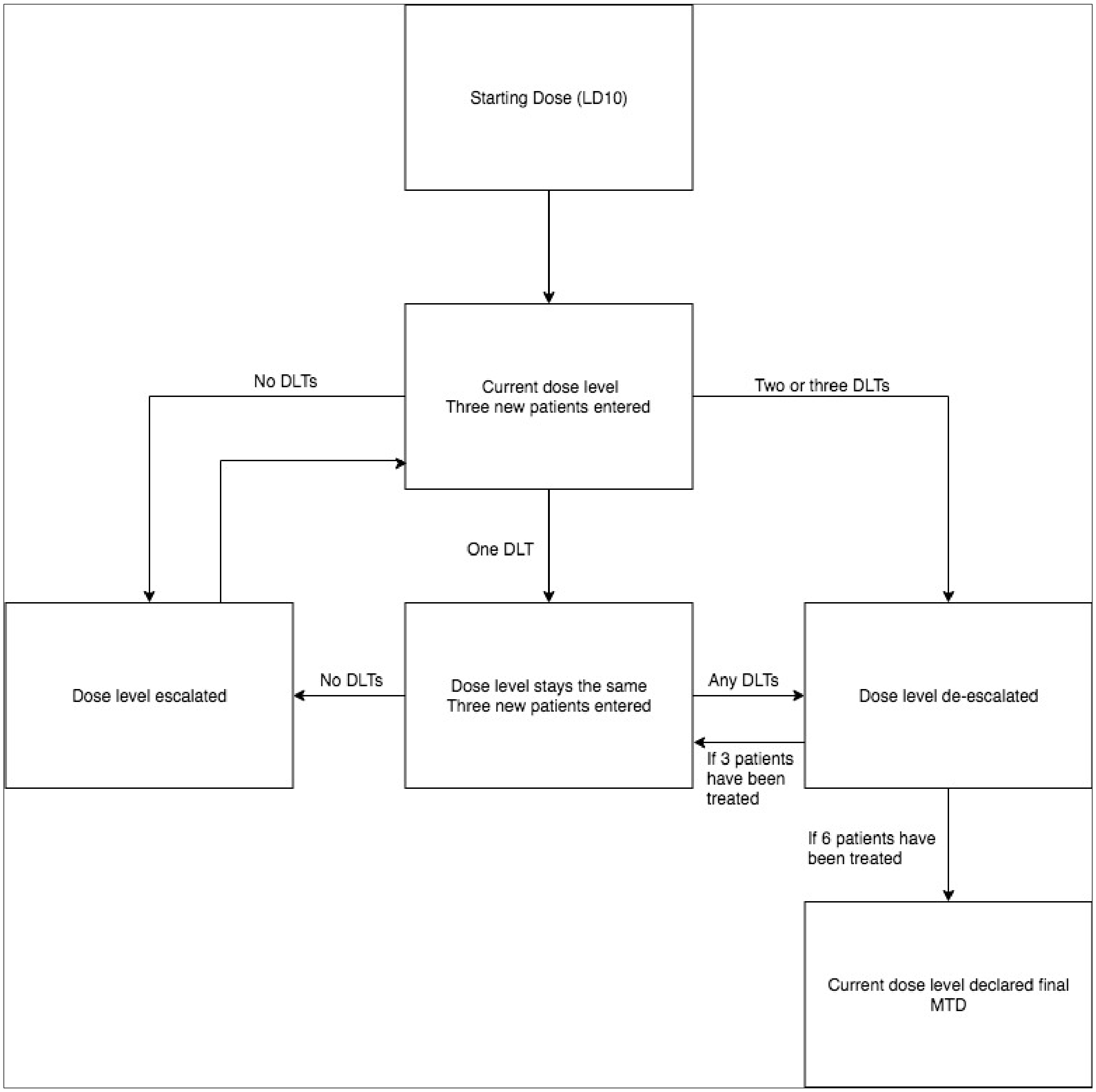
Phase I trials are the first stage of administered treatment on humans. Typically done on a small number of subjects for whom there are no effective therapies, phase I trials attempt to find the maximum tolerated dose (MTD) of a new agent. The MTD is the highest dose tested with acceptable levels of toxicity outlined by the trial protocol and is usually passed on as the recommended dose for phase II trials. However, to determine what is an acceptable level of toxicity, investigators must specify their agent’s dose limiting toxicities (DLT) as these vary from trial to trial. A DLT for one trial could be a serious, life-threatening toxicity, while a DLT for another trial could be a long-lasting toxicity that is unable to resolve at a reasonable interval. Because of this variation, DLTs must be precisely defined *a-priori* based on the NCI Common Terminology Criteria for Adverse Events.

After establishing an agent’s criteria for a DLT, investigators experiment with a variety of dose levels until an MTD is found. However, finding the MTD is a difficult process because these new agents have very limited usage on humans. In an ideal phase I trial, an investigator could set up a relatively large range of doses to search for the MTD, randomly assign subjects to each dose level, and estimate the MTD based on their responses. Although this design is efficient and estimates the MTD with minimal bias, it is not feasible as it is unethical to give patients extremely high or low doses without previous knowledge of the drug. Extremely high doses of a drug could potentially be lethal, while extremely low doses could have no effect on a patient. Furthermore, giving patients subtherapeutic doses is ethically questionable as these drugs could possibly be these patients' last chance for survival. Because of the issues inherent in this "ideal" design, investigators must utilize more practical designs to determine the MTD of a drug.

Numerous dose-finding designs have been proposed over the last three decades to find this ideal design, manipulating factors such as patient selection, starting dosage, escalation rules, and dose scheduling. For example, the Storer method manipulates patient selection by proposing that only one patient is evaluated at a time until a DLT occurs, after which accrual to dose levels is done with three patients. Using this method, simulations have shown improved estimates of the MTD and a reduced proportion of patients treated at low dose levels without greatly increasing the proportion of patients treated at unacceptably high dose levels (Storer, 1989). However, both the 3+3 and Storer designs are often criticized for their inflexibility due to being unable to specify an unacceptable probability of DLT. Another popular method, the continual reassessment method (CRM), addresses this concern by requiring the specification of unacceptable DLT probability (Quigley et al, 1990). It utilizes a prior distribution of the MTD for the starting dose and a long-memory model which recalculates the MTD estimate after treating each patient. As this method utilizes previous results to determine the next dose, the MTD estimates are improved over the traditional design. However, many investigators critique this method as it is based on *a-priori* assumptions about the dose-toxicity relationship regarding the starting dose, leading to dose escalations that skip several dose levels which could lengthen the trial and endanger patients if the dosage gets too high. Despite strong support from statisticians because of their superior statistical properties, none of these new designs have overtaken the 3+3 as the most commonly used design (Jaki et al, 2013).

In the 3+3 design, escalation is determined by the number of DLTs observed at each level. Initially, three patients are entered at the first dose level. In human phase I studies, the first dose is determined by using animal models, often taking 10% of the mouse LD10 (the dose at which 10% of mice die during preclinical tests) or some fraction of a dose that has no adverse events in a larger animal model (Morgensztern et al, 2015). If no patients experience a DLT at this initial dose, three patients are entered at the next dose level and the process repeats. If two or three patients experience a DLT at the initial dose, dose escalation is discontinued, and the dose is potentially de-escalated. However, if one patient experiences a DLT at the initial dose, three additional patients are added and treated at this dose level. If only one patient out of the six experiences a DLT, the dose is escalated, and the process repeats; otherwise, the dose is concluded to be above the MTD and is de-escalated. If the dose is above the MTD, the next lower dose is declared the MTD if six patients have already been treated at that dose; otherwise, three additional patients are treated at the next lower dose. If 0 or 1 have DLTs, this is declared the MTD; if two or more have DLTs, it is further de-escalated following this protocol until the MTD is declared. Figure 1 below illustrates the 3+3 design as a flowchart.

Figure 1 - 3+3 design flowchart


Starting Dose (LD10)
Current dose level
Three new patients entered
No DLTs
Two or three DLTs
One DLT
Dose level escalated
No DLTs
Dose level stays the same
Three new patients entered
Any DLTs
Dose level de-escalated
If 3 patients have been treated
If 6 patients have been treated
Current dose level declared final MTD

The 3+3 method, which is known as a “rules based” design, is a simple design with an easily followed set of rules that can be taught to investigators. Its simplicity is one of its biggest strengths, requiring limited statistical knowledge to understand its execution. Furthermore, the 3+3 has been used extensively over the past few decades, so research teams understand its intricacies and know what to expect due to familiarity. This familiarity allows trial protocols to pass institutional review boards and biostatistics reviews more easily and increases the throughput of potential drug trials (Ji et al, 2013). The sample size required by the 3+3 is typically much smaller compared to model-based designs as well (Ji et al, 2013). While there are technically very few statistical advantages, the 3+3 method has been found to pick doses with a relatively low probability of DLT, a desired property of phase I trials (Hansen et al, 2014). Other designs can be altered to pick lower dose probability levels as well, but this property is intrinsic of the 3+3. It can be speculated that the 3+3 has other properties that satisfy investigators because of its continued usage, but the main reasons for its popularity are its simplicity, flexibility, and its history of producing desirable results.

However, there are several issues with the 3+3 rule-based design. First, it does not converge to the true MTD but instead provides an estimate. This is because it does not allow for dose oscillation (switching between dose escalation and de-escalation repeatedly) or intermediate doses without amending the protocol (Nie et al, 2016). Furthermore, confidence intervals perform poorly, as nominal 80% confidence intervals do not include the correct value 80% of the time, and 95% intervals often cover the full range of possibilities [0,1] which provides no information (Nie et al, 2016). It is sensitive to both the starting dose and the dose-toxicity relationship, and subjects may also experience different levels of toxicity which the design does not account for. Finally, the most common criticism of this design is that many patients may be treated at sub-therapeutic doses if the initial dose is much lower than the MTD. Due to the poor statistical properties of this design and the potential ethical and financial issues caused by incorrect starting doses which may lengthen the trial, the 3+3 design has been heavily criticized by investigators who suggest that it be replaced.

To determine the effectiveness of the 3+3 design, this simulation study evaluates its performance over a range of potential scenarios. From low dosage, low dose escalation drugs to high dosage, high dose escalation drugs, it is important to consider how various outcomes change in these different scenarios. Different drugs have different toxicity curves, so outlining scenarios where the 3+3 excels may assist researchers when deciding on a trial design. Many previous studies have evaluated the 3+3 rule-based design; however, these studies chose very specific scenarios that limit their outcomes and are therefore less applicable to the average study. Furthermore, these studies typically are done in a comparative fashion to examine how the 3+3 compares to another method like the CRM. The focus of this study is to examine scenarios where the 3+3 is effective so researchers can justify their usage of the 3+3, allowing the results to be applied more broadly. The main outcome being considered in this study is the probability of DLT at the selected MTD and its distribution over several dose levels. This outcome determines how a DLT probability curve under different scenarios can affect the MTD. Other outcomes such as the number of DLTs patients experience in a trial, the highest dose level tested, the average number of patients needed in a trial, and whether the MTD selected is missing (the MTD value selected is below the lowest dose level) will also be considered.

3 Methods

3.1 Simulations

The methods for generating phase I trial data are split up into the following five sections: simulation of phase I toxicity data, the application of the 3+3 rules to simulated data, validating the 3+3 algorithm, number of simulations, and collection of data for evaluation of the performance of the 3+3.

3.1.1: Simulation of phase I toxicity data

To create realistic and varied DLT probability arrays, the program utilized the DLT probability function of the form (Sposto et al, 2010):

Figure 2 - DLT probability function

$$\boxed{P(D) = e^{\beta D(1 \ + \ \gamma D)} - 1}$$

where D is the maximum dose level, P(D) is the maximum DLT probability generated using uniform distribution U(0,1), beta is the slope of the DLT probability curve estimated using uniform distribution U(0, P(D)/D), and gamma which is calculated by plugging in the previous variables. Previous DLT probabilities P(d) for d=1,…D-1 are calculated using the gamma value. This creates a DLT probability set that assigns a DLT probability to each dose level ranging from 1 to D.

Using this DLT probability function, a variety of DLT curves can be created. As beta dictates the slope of the DLT probability curve, we can change it to increase how quickly a drug becomes toxic. Similarly, changing the maximum DLT probability, P(D), also changes the shape of the DLT curve by limiting its height. These two parameters, in addition to the maximum number of dose levels, D, will be used to evaluate the effectiveness of the 3+3 method. Both P(D) and beta were categorized into three levels and defined in further detail below (Table 1). There are nine different combinations of P(D) and beta, giving nine potential DLT curves. This study examines four different maximum dose levels (D=3, 5, 7 and 9) to understand the effect of incrementing dosage, creating a total of 36 (nine DLT curves * 4 different dose levels) different scenarios.

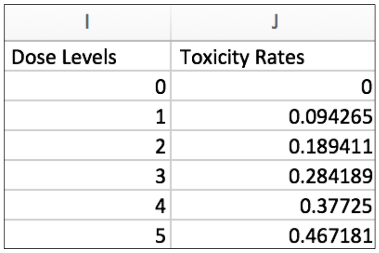
Table 1 - All experimental conditions

Experimental Condition	Definition
P(D)	The probability of a dose limiting toxicity occurring at the highest dose level in a simulation. In other words, the maximum DLT probability generated in a simulation using the uniform distribution U(0,1).
Low P(D)	Probability ranges from 10% to 30%
Medium P(D)	Probability ranges from 30% to 50%
High P(D)	Probability ranges from 50% to 80%
Beta	How steep the initial slope of the dose limiting toxicity probability curve is – the larger the beta, the steeper the slope
Low Beta	Beta value ranges from 0 to 1/3(P(D)/D)
Medium Beta	Beta value ranges from 1/3(P(D)/D) to 2/3(P(D)/D)
High Beta	Beta value ranges from 2/3(P(D)/D) to P(D)/D
Dose levels	The total number of dose levels in a simulation. The highest dose level in a simulation is indicated by D. Four different dose levels will be tested: D=3, 5, 7, and 9.

3.1.2: Application of the 3+3 rules to simulated data

The data analyzed in this study were created with a simulation program written in C++. The simulation applies the 3+3 rules by taking in two initial parameters: a starting dose level and an array of DLT probabilities that indicate the probability that a patient will experience a DLT at each dose level. An example of a DLT probability array is shown in Figure 3 below. In this given example, if a patient was given the drug at dose level 3, there would be a 28.4% chance he or she would experience a DLT.

Figure 3 - DLT probability array



I	J
Dose Levels	Toxicity Rates
0	0
1	0.094265
2	0.189411
3	0.284189
4	0.37725
5	0.467181

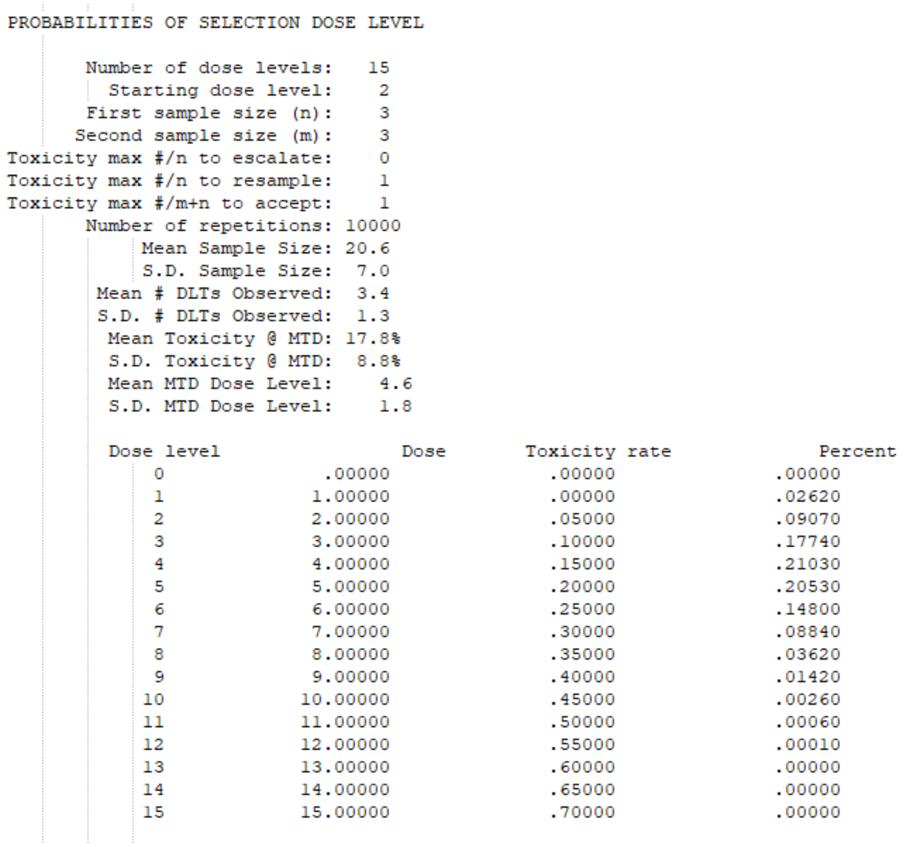
To simulate patients experiencing a DLT at the current dose level, the program performs an inverse transform of the uniform distribution using the given DLT probability array. The inverse transform works by using a random number generator with uniform distribution that generates a number between 0 and 1 -- if the generated number is less than the DLT probability for the current dose level, the patient experiences a DLT; otherwise, the patient does not experience a DLT.

In our simulations, the starting dose level for all simulations is set at dose level 1 for this experiment, giving the program the potential to de-escalate the dosage by one level if necessary. Furthermore, the full dose range consists of levels 0 to D (e.g. D=9 contains [0,9] dose levels). The program begins by setting the current dose level as the starting dose level. Based on the number of DLTs at the current dose level, the dosage is either escalated or de-escalated based on the rules of the 3+3 design. The current dose level is updated, and the entire process is repeated until the MTD is found.

3.1.3: Validating the 3+3 algorithm

Initially, the DLT probability array was created using set intervals of 5%, ranging from 0% to 70% for 15 dose levels. Although this probability array is unrealistic, it allowed for initial verification of the program's 3+3 algorithm as it could be compared to another third-party simulator. Using a third-party simulator written in FORTRAN, 150 simulations were run and its MTD proportions were used as the "gold-standard" (Figure 4). This program was written by Dr. Richard Sposto who kindly permitted the use of his program to verify the performance of the algorithm written in C++ for this thesis.

Figure 4 - Gold standard MTD proportions



```
PROBABILITIES OF SELECTION DOSE LEVEL

       Number of dose levels:   15
         Starting dose level:    2
       First sample size (n):    3
      Second sample size (m):    3
Toxicity max #/n to escalate:    0
Toxicity max #/n to resample:    1
Toxicity max #/m+n to accept:    1
       Number of repetitions: 10000
            Mean Sample Size: 20.6
            S.D. Sample Size:  7.0
        Mean # DLTs Observed:  3.4
        S.D. # DLTs Observed:  1.3
         Mean Toxicity @ MTD: 17.8%
         S.D. Toxicity @ MTD:  8.8%
         Mean MTD Dose Level:    4.6
         S.D. MTD Dose Level:    1.8

         Dose level                Dose       Toxicity rate            Percent
             0               .00000            .00000             .00000
             1              1.00000            .00000             .02620
             2              2.00000            .05000             .09070
             3              3.00000            .10000             .17740
             4              4.00000            .15000             .21030
             5              5.00000            .20000             .20530
             6              6.00000            .25000             .14800
             7              7.00000            .30000             .08840
             8              8.00000            .35000             .03620
             9              9.00000            .40000             .01420
            10             10.00000            .45000             .00260
            11             11.00000            .50000             .00060
            12             12.00000            .55000             .00010
            13             13.00000            .60000             .00000
            14             14.00000            .65000             .00000
            15             15.00000            .70000             .00000
```

Using the same DLT probability array of 5% increments, 150 simulations were then generated with the C++ program for comparison with the following result:

Dose Level	1	2	3	4	5	6	7	8	9	10
Frequency	7	14	24	27	28	25	15	6	2	2
Proportion	0.047	0.093	0.160	0.180	0.187	0.167	0.100	0.040	0.013	0.013

R was used to compare the observed count of MTD values in the C++ program to the “gold standard” proportions of the FORTRAN program using a chi-squared goodness of fit test, and both programs had very similar results (Chi2=10.98, df=11, p=0.445), verifying the correctness of the C++ algorithm.

3.1.4: Number of simulations

At each maximum dose level, D, for each of the nine (P(D) x beta) categories, 150 simulations were run - a total of 1350 simulations per value of D. A sample size of 150 per scenario was calculated using nQUERY. A two-way ANOVA was used to calculate the variance in means for the two parameter variables, beta and P(D), using a 3 by 3 table (Figure 5). These numbers were chosen based on a standard deviation of 1, with the values in row 3 column 3 being a half standard deviation higher to account for potential interaction between the two variables. With these calculated variances in means, a significance value of 0.05, and a power of 90%, nQUERY gave an estimated sample size of 147 which was rounded up to 150.

Figure 2 – Variability among the means – used for sample size determination

Standardized Mean Probability of DLT at the MTD (differences are relative to a common standard deviation of 1.0)			
	Low P(D)	Medium P(D)	High P(D)
Low Beta	0.0	0.5	1.0
Medium Beta	0.5	1.0	1.5
High Beta	1.0	1.5	2.5

3.1.5: Collection of data for evaluation of the performance of the 3+3

Within each simulation, five different endpoints (Table 2) were recorded and exported to Excel (Figure 6-7) to evaluate the properties of the 3+3 design. The main endpoint is the probability of DLT at the MTD, or the probability of a dose limiting toxicity occurring when using the dosage indicated at the maximum tolerated dose. This endpoint is considered the most important because it is the main goal of a phase I clinical trial – finding a dosage to recommend to investigators conducting phase II trials if the drug has potential. However, phase I clinical trials also focus on several other factors – one being the safety of the patients. One outcome related to safety is the number of DLTs, or the total number of dose-limiting toxicities that occur in a single simulation. The number of DLTs may indicate how safe a drug may be, as more DLTs given the same set of parameters is undesired. The highest dose level tested in a single simulation may also be of interest because at least three patients will be administered this dosage.

Determining if the highest dose level tested is too high, which could cause unnecessary toxicity, or if the highest dose level is too low, which leads to increased trial costs and patients treated at sub-therapeutic doses, may be useful to researchers. The total number of patients enrolled in a single trial is also important, as having more patients increases cost and could indicate an issue with the initial dosage. Additionally, the 3+3 design may not always select an MTD if the starting dosage is too high, so simulations where the MTD is not found will be considered.

Table 2 - All endpoints

Endpoint	Definition
Probability of DLT at the MTD	The probability of a dose limiting toxicity occurring when using the dose selected as at the maximum tolerated dose based on the 3+3 rules
Number of DLTs	The total number of dose limiting toxicities that occur in a single simulation
Highest dose level tested	The highest dose level tested in a single simulation, regardless if it is selected as the MTD
Total number of patients	The total number of patients enrolled in a single simulation
Missing MTD	Whether the probability of DLT at the MTD selected is missing or not found, as the dose level selected is below 1

Figure 3 – Endpoint results format for 4 simulated trials

Simulation #	Probability of DLT at the final MTD	Number of DLTs	Highest dose level tested	Total number of patients
1	0.46	3	5	15
2	0.34	3	5	18
3	0.20	4	4	15
4	0.55	6	8	30

Figure 4 – The probability of DLT, P(d), at each of the 9 possible dose levels for 4 simulated trials

Simulation #	P(1)	P(2)	P(3)	P(4)	P(5)	P(6)	P(7)	P(8)	P(D)
1	0.12	0.24	0.35	0.46	0.56	0.64	0.71	0.76	0.79
2	0.09	0.18	0.26	0.34	0.41	0.47	0.52	0.56	0.59
3	0.07	0.13	0.20	0.26	0.32	0.37	0.42	0.47	0.51
4	0.10	0.19	0.28	0.36	0.44	0.50	0.55	0.59	0.62

3.2 Analysis

After running all the simulations and exporting their results, R was used to calculate their descriptive statistics. Mean, median, min, max, IQR, and standard deviation were calculated for each endpoint. The ggplot2 data visualization package in R was used to create DLT probability graphs using the given DLT probabilities (Figure 7) in each simulation, as well as histograms of the various endpoints. These descriptive statistics and graphs were gathered to better understand the distribution of each endpoint in each scenario. Understanding the average probability of DLT and its range may help researchers decide whether to use the 3+3 design if the profile of their drug fits one of these scenarios.

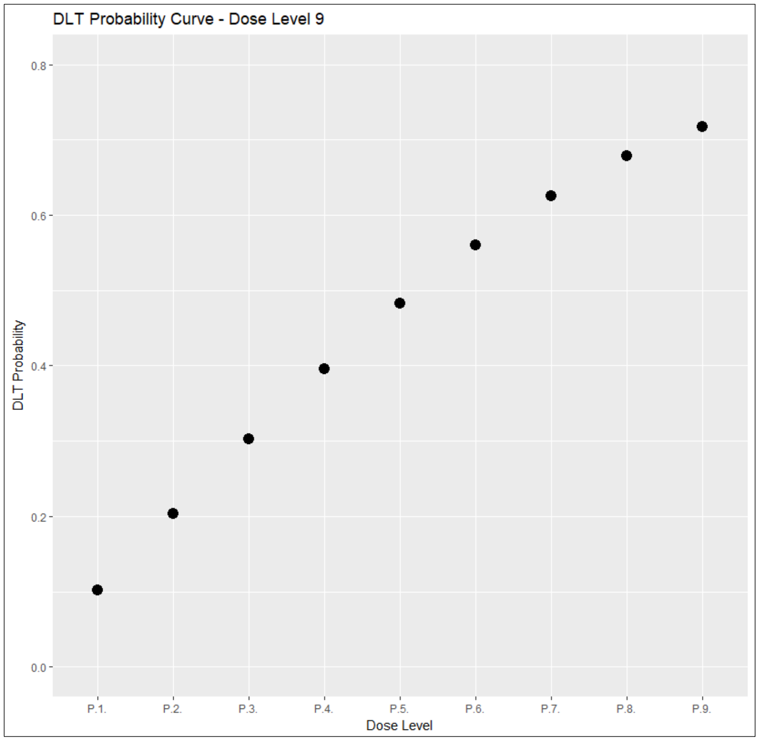
In addition to the descriptive statistics, SAS was used to analyze how the parameters P(D) and beta affected the main endpoint, the MTD. Understanding how changing the maximum dosage of a trial and dose escalation speed may affect the MTD is important because it assists researchers in determining an effective dose escalation strategy. P(D) and beta were read in as continuous variables. Categorical versions (catBeta and catP(D)) using the cutoff points in Table 1 were created of both variables to check for association, as the categories may contain as much information as their continuous counterparts. Interaction terms between P(D) and beta -- P(D)*catBeta, catP(D)*beta, and catP(D)*catBeta -- were also considered. As the outcome, MTD, is continuous, linear regression was run to check for these associations, and the predicted and residual values were checked to see if they met all assumptions (normality, homoscedasticity, linearity, and independence). Linear regression was also used to predict the best model, where the full model included P(D), beta, catP(D), catBeta, and the interaction terms listed above. Four prediction methods were used to determine what would be the best model in terms of a high adjusted R^2 and number of variables: forward selection, backwards elimination, stepwise, and all-subsets. The prediction tests used a significance level of entry/exit of 0.10. Their type I and III sums of squares as well as unadjusted R^2 values were also examined.

4 Results

4.1 Simulation Results

For each simulation, a random DLT probability array is generated using the DLT probability function in Figure 3. To see the general shape of the probability curves, R was used to graph each of these probability arrays. For example, Figure 8 illustrates a single simulation's DLT probability curve for D=9. As the dose level increases from 1 to 9, the probability that a patient will experience a DLT also increases.

Figure 5 - DLT probability curve of a single simulation



However, the DLT curve from a single simulation is not very informative due to the variation within each of the nine categories. Therefore, to illustrate the overall range of shapes for the nine categories, all DLT probability curves (n=150 for each category) were combined into overall graphs, categorized by dose level, with jitter for visibility (Figure 9).

*Figure 6 - DLT probability curve for each category at **D=9** with jitter, n=1350 (150 per category)*

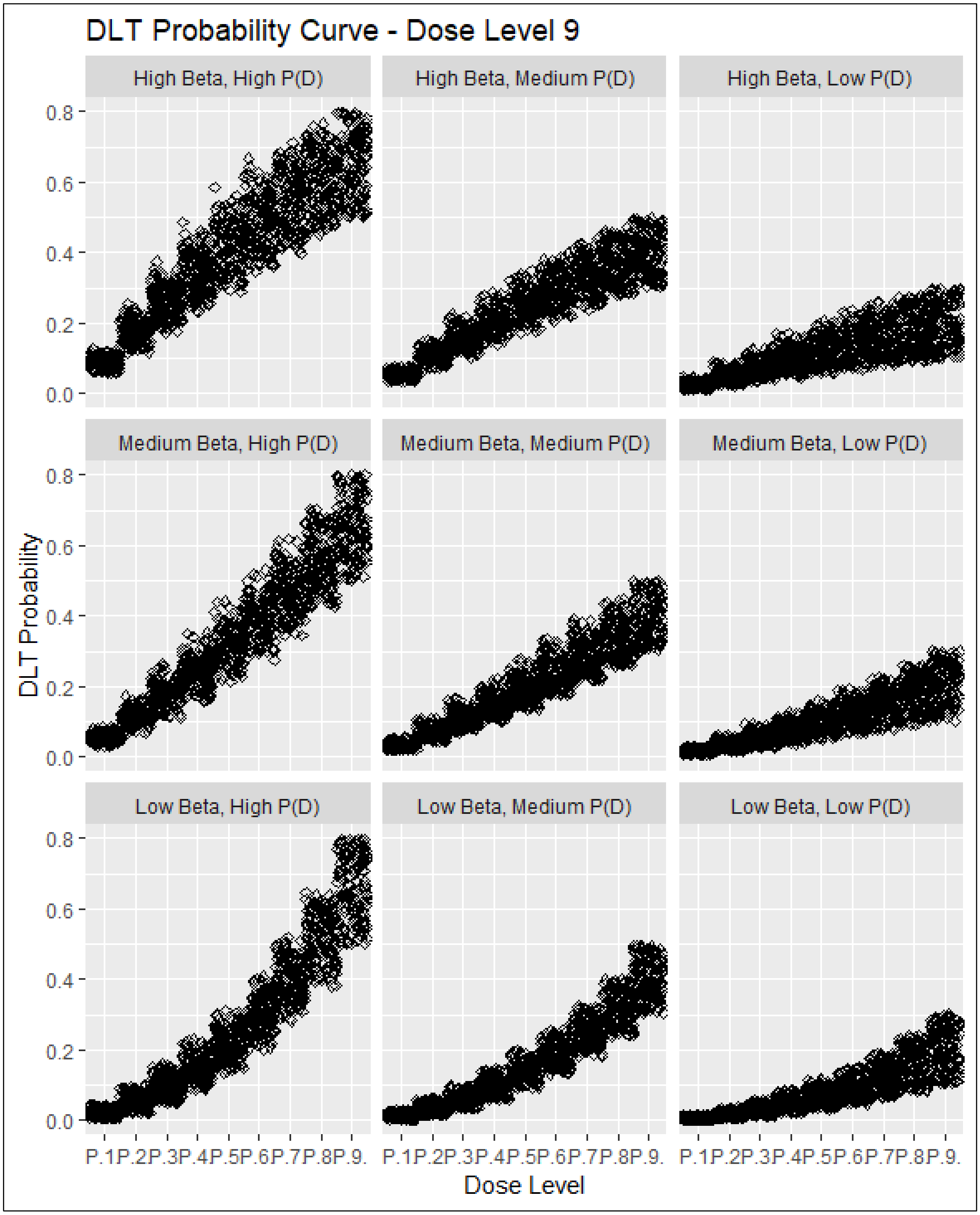


Figure 9 illustrates the DLT probability curve shapes for D=9, while D=3, 5, and 7 can be found in the appendix (Appendix 1-3). As beta controls the slope of the graph, lower beta values create a concave-up parabolic shape which simulates a drug that is less toxic at lower doses but ramps up in toxicity as the dose level increases. A medium beta value creates a positively-sloped linear DLT curve that simulates a drug with toxicity that increases approximately linearly with the dosage. A high beta value creates a concave-down parabolic shape which simulates a drug that increases in toxicity more quickly initially but begins to level out as the dosage increases. P(D) determines the overall height of the graph as it sets the maximum DLT probability a simulation can reach. Higher P(D) values have graphs ending between 0.5 and 0.8, medium P(D) values have graphs ending between 0.3 and 0.5, and low P(D) values have graphs ending between 0.1 and 0.3. These P(D) values simulate a variety of drug toxicities, with lower P(D) values simulating less toxic drugs and higher P(D) values simulating very toxic drugs. Overall, each category represents a different scenario with a different height or different shape, allowing this experiment to evaluate how the 3+3 performs under a variety of conditions.

4.2 Descriptive Statistics

4.2.1 Endpoint: Probability of DLT at the MTD

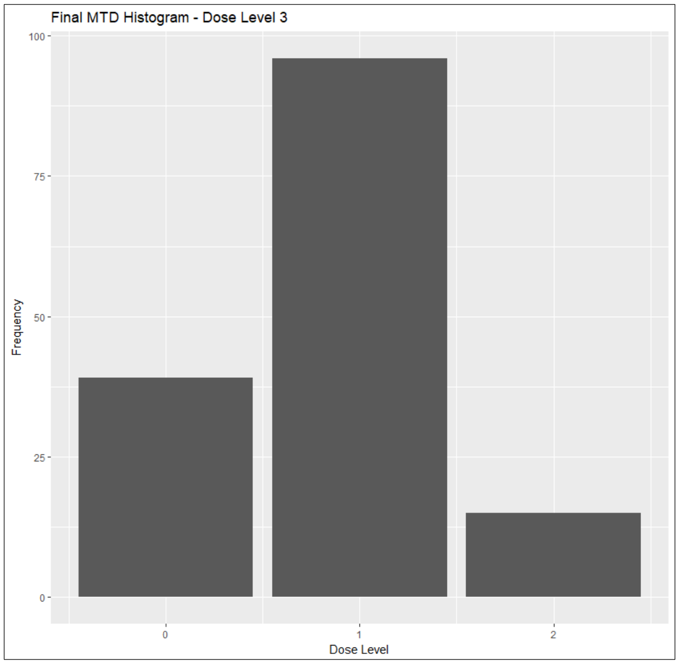
For each number of doses (D=3, 5, 7, and 9) and for each of the 9 P(D) and beta combinations, 150 simulations were run using their respective DLT probability curve and their results were recorded. Descriptive statistics were run on each endpoint for each category and the four maximum dose levels, D. Table 3 shows the descriptive statistics for the endpoint of probability of DLT at the MTD for D=9, as estimated by the simulations. Holding beta constant, the higher the P(D) category, the higher the mean value is for this endpoint because the drug is more toxic and has a higher limit for what the maximum probability of DLT will be at D=9. Holding P(D) constant, beta makes a smaller difference in the mean values for each category. Low and medium beta categories have very similar mean values, while the high beta category has marginally higher mean values. Median values for all categories have similar findings. Minimum and maximum values for low and medium beta, while holding P(D) constant, are very similar as well, while high beta has a higher variance. When comparing dose levels and holding P(D) and beta constant, D=9 had the lowest mean values of DLT probabilities, while D=3 had the highest mean values of DLT probabilities (Appendix 7). This is expected because the 3+3 method does not allow for intermediate dose levels. Therefore, D=3 has lower precision compared to D=9 as the intervals between each dose level are larger, leading to overestimation of the endpoint. Similarly, the maximum and minimum values for D=3 are more extreme compared to D=9 because of the same intermediate dose level reason.

Table 3 Probability of DLT at MTD for ***D=9*** *with n=150 simulations in each column*

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
mean	0.2	0.14	0.19	0.17	0.14	0.17	0.18	0.14	0.17
sd	0.1	0.053	0.086	0.083	0.061	0.093	0.086	0.054	0.082
median	0.19	0.12	0.19	0.16	0.13	0.15	0.17	0.14	0.18
q1	0.11	0.1	0.11	0.1	0.1	0.11	0.12	0.11	0.11
q3	0.26	0.17	0.25	0.22	0.18	0.22	0.24	0.17	0.23
min	0	0.025	0	0.025	0.019	0.018	0	0.017	0.024
max	0.55	0.26	0.37	0.4	0.29	0.4	0.42	0.3	0.4

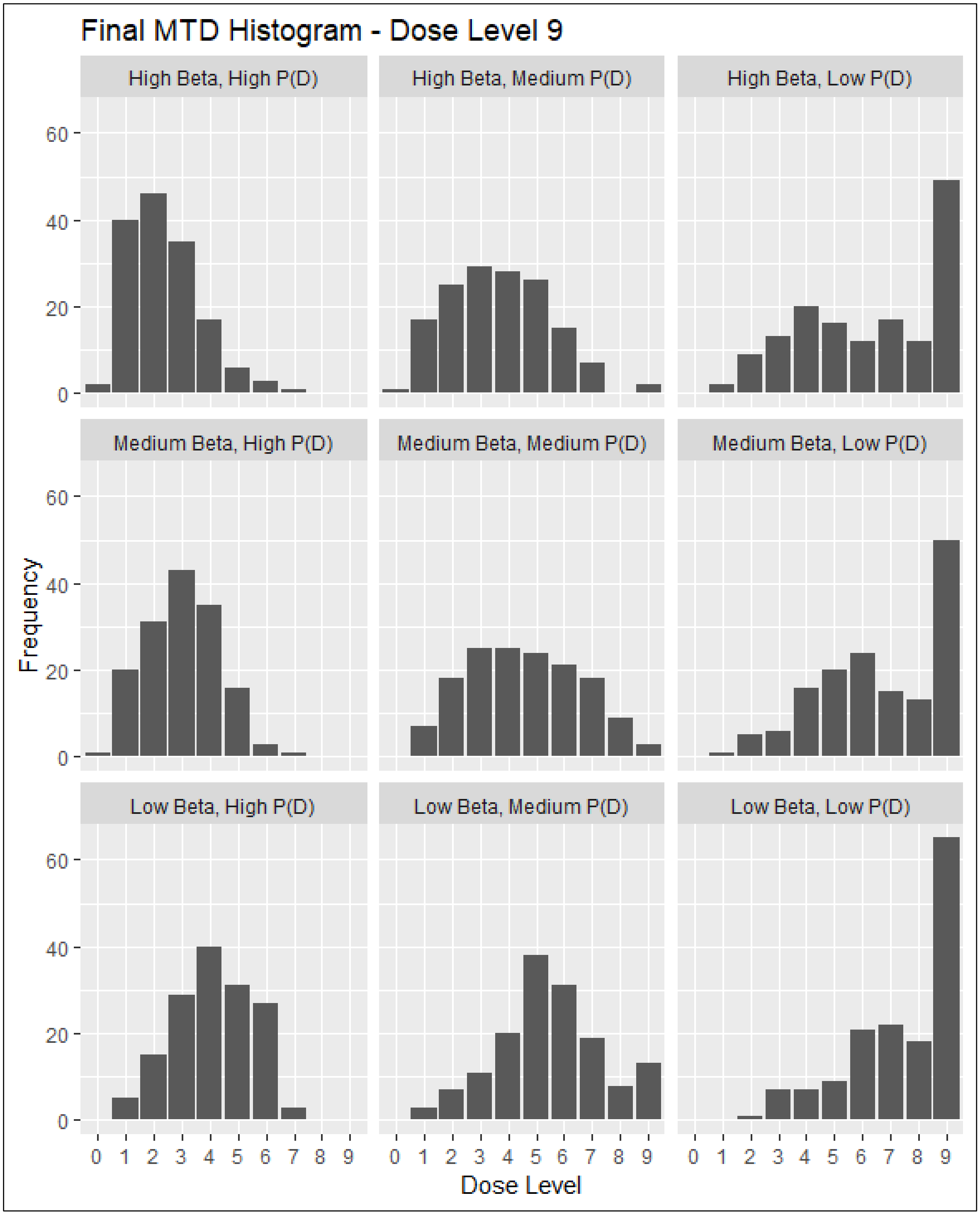
To better visualize the distributions of the probability of DLT at the selected MTD, histograms of the frequency count of the MTD dose level were created for each category using the results from their respective 150 simulations. Figure 10 illustrates the histogram for one category (high beta, high P(D), D=3). Out of its 150 simulations, around 40 were not found (dose level 0), around 90 ended on dose level 1, around 12 ended on dose level 2, and 0 ended on dose level 3.

Figure 7 - MTD Histogram for one category (High Beta, High P(D))



A graph combining all nine categories within a dose level can be found in Figure 11. By categorizing these graphs, the effects of beta and P(D) become more apparent. Figure 11 illustrates these categories for D=9, while D=3,5,7 can be found in the appendix (Appendix 4-6). When looking at just the P(D) categories, high P(D) and medium P(D) plots have relatively normal distributions with some slight left and right skewness. However, low P(D) plots have a non-normal, bimodal distribution with the largest peak at dose level 9. Although the maximum probability of DLT is 30% for the low P(D) category, the each simulation's P(D) value is selected using a uniform random distribution U(0.1, 0.3), so a majority of P(D) values are much lower than the maximum of 30%. Because of these low probability values for all nine dose levels, the chances of it reaching the maximum dose value is much higher. Beta, which controls the initial slope of the DLT probability curve, also influences these histograms. Histograms for low beta categories' are more left skewed than high beta categories, indicating that low beta MTD dose levels are higher. With a lower initial slope, the initial dose levels have a lower probability of DLT as well, increasing the chance that fewer DLTs will be experienced at these lower dose levels. For high beta values, the initial dose levels have a higher probability of DLT while later dose levels begin to level off, increasing the chance that more DLTs will be experienced at these lower dose levels. Medium beta values are the most linear, so their histograms fall in between low and high beta values. For D=9, missing MTD values (dose level 0) are less of a concern due to the increased precision of higher dose levels. However, when looking at the D=3 graphs (Appendix 4), there are significantly more missing values, especially for the high P(D) categories where there are around 40, 20, and 5 missing values for high beta, medium beta, and low beta, respectively. Similarly, for D=5 (Appendix 5), there are still missing values but not as severe as D=3, with a downward trend for missing values as D increases. The histograms for D=3,5, and 7 have relatively similar distributions to D=9, where beta and P(D) determine the skewness and normality of each histogram.

*Figure 8 - MTD Histogram for all categories at **D=9***



4.2.2 Endpoint: Number of patients who experienced a DLT

Descriptive statistics were run on each of the secondary endpoints as well. Analyzing the total number of DLTs is important, as a high number of DLTs is undesired in a trial. Table 4 shows the descriptive statistics for the endpoint of total number of DLTs for D=9, as estimated by the simulations. Holding beta constant, high P(D) has a very similar mean number of DLTs compared to medium P(D), while low P(D) has a lower mean count. Standard deviations are very similar throughout all nine categories, ranging from 1 to 1.5. Minimum values are very similar throughout the nine categories, ranging from 0 to 2. However, maximum values have more variation ranging from 6 to 11, although there is also no clear trend amongst the categories. Holding P(D) constant, there is no clear trend on its effect on the mean value for each of the beta categories. When comparing D=9 to D=3 simulations (Appendix 13), D=3 has larger variation because the interval between each dose level is larger; however, DLT count per simulation remained very similar. Median values were also more varied, ranging from 1 to 4 for D=3 compared to D=9 which ranged from 2 to 3. Like the histogram for the MTD frequency above, Figure 12 graphs the frequency of the number of DLTs, categorized by the MTD level. All simulations where there are 0 or 1 DLTs occur in the low P(D) and medium P(D) categories. Similarly, all simulations where there are 0 or 1 DLTs have a MTD value of 9 (pink in the graph below), indicating that it has reached the highest dose level. Low and medium P(D) categories also have more normal-looking distributions compared to high P(D). High P(D) has a maximum MTD level at the 7th dose level, never reaching the 9th dose level. A majority of MTD levels in the high P(D) + high beta category is dose levels 1 and 2, but this becomes more dispersed as beta decreases. Histograms for D=3,5,7 can be found in Appendix 22-24. D=7 has a similar distribution to D=9. D=5 also has a similar distribution to D=9, although the graph is slightly more left skewed due to more simulations reaching the highest dose level. D=3 has a very different distribution from D=9, with several simulations reaching dose level D even in the high P(D) category. There are a lot more simulations that have a total DLT count of 0 or 1, and the mean values for the low P(D) and medium P(D) because of this.

Figure 9 - Histogram of Number of DLTs categorized by MTD level for D=9

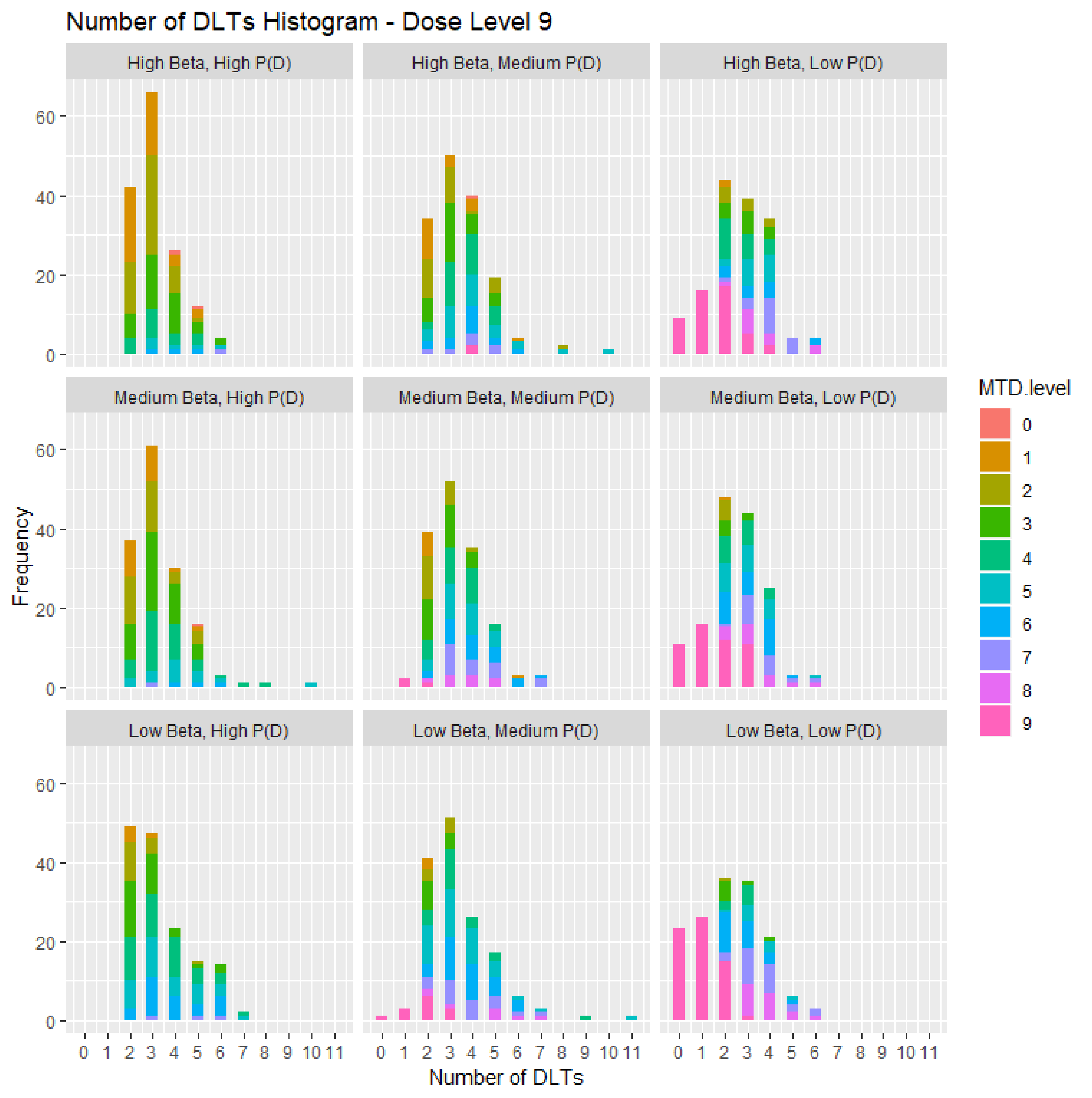


*Table 4 – Number of DLTs for **D=9***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
mean	3.1	2.7	3.5	3.4	2.2	3.4	3.3	2.5	3.3
sd	1	1.3	1.3	1.3	1.5	1.5	1.2	1.3	1.2
median	3	3	3	3	2	3	3	2.5	3
q1	2	2	3	2	1	2	3	2	2
q3	4	4	4	4	3	4	4	3	4
min	2	0	2	2	0	0	2	0	1
max	6	6	10	7	6	11	10	6	7

4.2.3 Endpoint: Number of dose levels tested

Table 5 shows the descriptive statistics for the endpoint of highest dose level tested for D=9, as estimated by the simulations. Median values range from 3 to 9 between the nine categories. Interquartile ranges for the high P(D) categories had the smallest intervals with a value of 2, while lower P(D) categories had larger intervals with values ranging from 2 to 4. Minimum and maximum values for all categories were similar ranging from 2 to 3 for minimum and 8 to 9 for maximum. As P(D) increased, the mean dose level count went down because as the probability of DLT increases more rapidly, the more likely 2 or more patients will experience a DLT and end the simulation. Similarly, as beta increases, the mean dose levels also go down. The steeper the slope is, the quicker the probability of DLT increases which leads to lower dose levels tested. It is difficult to compare the values of D=9 to D=3 (Appendix 10), as the lower dose level simulations cap off the number of potential dose levels. Therefore, it makes sense that lower dose level simulations have lower mean and median values, but little can be interpreted from these results. As the number of patients is directly correlated with the highest dose level tested, determining the number of doses tested is important when considering ethics and cost. Furthermore, when looking at Figure 13 which graphs the frequency of the highest dose level tested categorized by the simulation's MTD level, the highest dose level tested must be one dose level below the MTD except when it the MTD is equal to D. As expected, low P(D) categories have a higher dose level tested average as the probability of DLT is lower. Medium P(D) categories have a more normal distribution compared to the other two P(D) categories, while the high P(D) categories are more left skewed. High beta categories are also more left skewed due to having a steeper slope, causing the probability of DLT to increase more rapidly. Appendix 25-27 shows the histograms for D=3,5,7. D=5 and 7 have relatively similar distributions to D=9; however, D=3 is very different due to having only two values for highest dose level tested, d=2 and 3. For D=3, both low P(D) and beta categories decrease the number of d=2 values while increasing d=3, while high P(D) and beta categories increase the d=2 while decreasing d=3. The histogram also illustrates the possibility of de-escalation when comparing the highest dose level tested to the MTD level. For example, in Figure 13 low beta, medium P(D), d=9 has MTD level values of 9, 8, 7 and 6, indicating that some trials ended at d=9 while some trials de-escalated several dose levels due to toxicity.

Figure 10 - Histogram of highest dose level tested categorized by MTD level for D=9

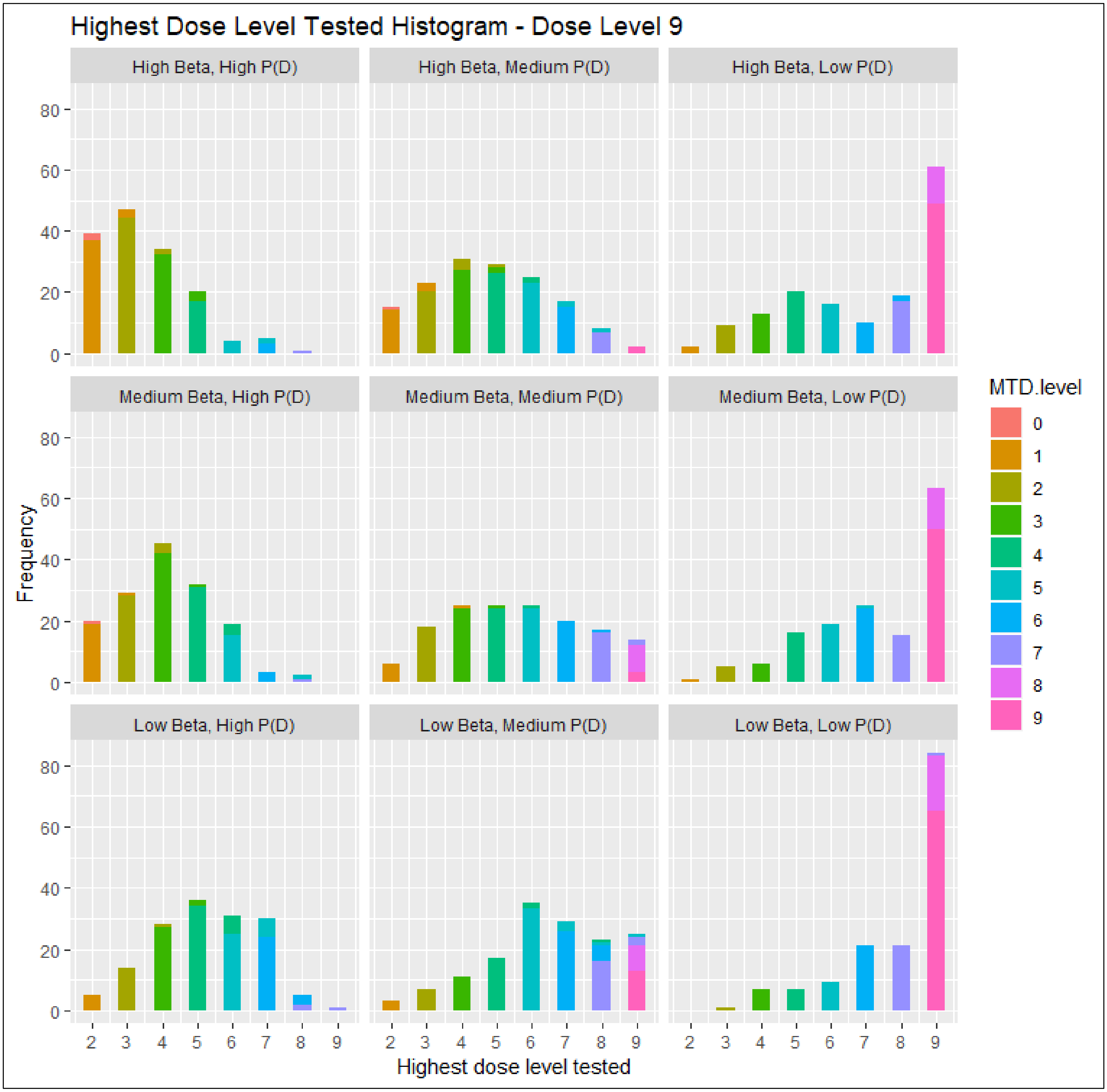


Table 5 – Highest dose level tested for D=9

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
mean	3.5	7	4.8	5.3	7.9	6.5	4.1	7.3	5.6
sd	1.3	2.1	1.7	1.5	1.5	1.8	1.4	1.8	2
median	3	8	5	5	9	7	4	8	6
q1	2	5	3.2	4	7	5.2	3	6	4
q3	4	9	6	6	9	8	5	9	7
min	2	2	2	2	3	2	2	2	2
max	8	9	9	9	9	9	8	9	9

4.2.2 Endpoint: Number of patients treated

Table 6 shows the descriptive statistics for the endpoint of total number of patients for D=9, as estimated by the simulations. Holding beta constant, high P(D) has the lowest mean patient count, while low P(D) has the highest patient count. Because lower P(D) values are correlated with number of dose levels tested as seen by the histograms in Figure 11, more simulations will reach dose level 9 which means more patients need to be recruited to test the previous dose levels. Higher P(D) values have a lower maximum dose level tested which leads to fewer patients tested. Holding P(D) constant, high beta values have the lowest mean patient count values, while low beta values have the highest mean patient count values. Explained in the descriptive section for highest dose level tested, lower beta values are more likely to have a higher max dose level tested which leads to more patients being required, illustrated by the mean values in Table 6. Minimum values are similar throughout the nine categories, ranging from 9 to 12. However, maximum values are lower for high P(D) and higher for low P(D) categories, ranging from 30 to 39 patients. When comparing dose level simulations (Appendix 16), D=9 has the highest mean values of patient count, while D=3 has the lowest mean values of patient count. Since the interval between each dose level is smaller for nine dose levels, more patients must be enrolled at each dose level to find the MTD compared to three dose levels. Mean values for D=5,7, and 9 follow the pattern described previously where high P(D) values have lower mean values compared to low P(D) values; however, D=3 simulations do not follow this pattern. The patient count remains the same or increases as the P(D) category increases. The pattern for D=5,7, and 9 is intuitive because there is enough precision between the dose levels to accurately find the MTD. However, the pattern for D=3 is likely due to the low precision caused by the low number of dose levels and having many missing values. Figure 14 illustrates the patterns described above as a boxplot for D=9. Boxplots for D=3,5,7 can be found in Appendix 28-30.

Figure 11 - Boxplot of total number of patients by beta x P(D) categories for D=9

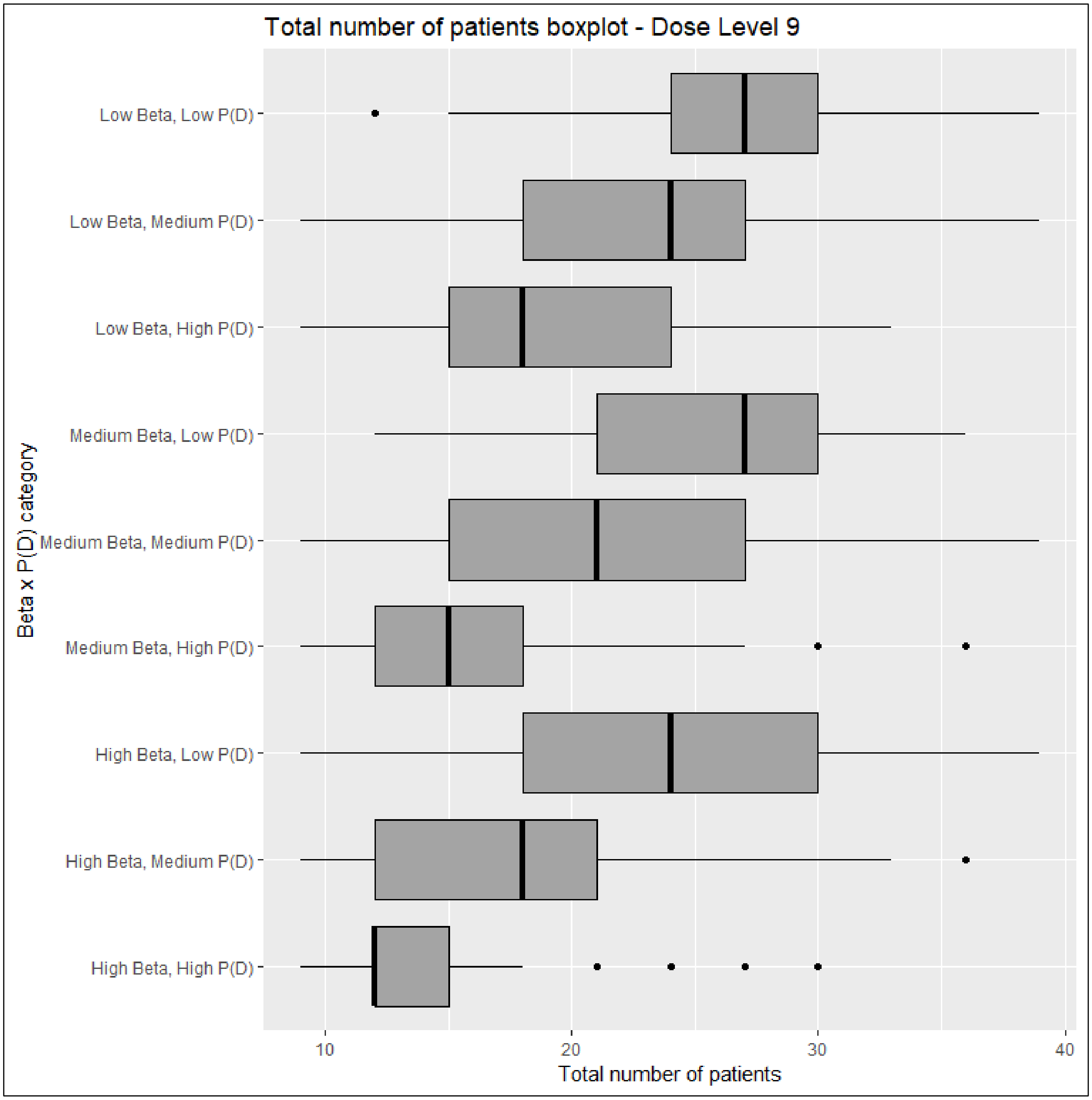


Table 6 – Total number of patients for ***D=9***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
mean	14	24	18	19	26	23	16	25	21
sd	4.6	6.8	6.2	5.6	4.9	6.6	5.2	5.9	6.9
median	12	24	18	18	27	24	15	27	21
q1	12	18	12	15	24	18	12	21	15
q3	15	30	21	24	30	27	18	30	27
min	9	9	9	9	12	9	9	12	9
max	30	39	36	33	39	39	36	36	39

4.2.4 Endpoint: Percent of times that a trial fails to identify a MTD (1st dose level has 2+ DLTs)

Finally, Table 7 shows the descriptive statistics for the endpoint of missing MTD for D=9, as estimated by the simulations. The MTD not being found is always a concern, and the fewer dose levels a design has, the more likely it is that the MTD will not be found. For D=9, there are very few missing values for the MTD, ranging from 0 to 2 for the missing count which is less than 1% of their respective simulations, since the lower dose levels tend to be associated with lower probabilities of DLT. However, as the maximum dose level count decreases (Appendix 19-21), the number of missing values increases. D=3 simulations had 39 missing values for the high beta, high P(D) category, which is 26% of the 150 simulations. It is easier to see for D=3 how P(N) and beta can affect missing values. Increasing P(D) leads to a higher probability that the MTD will not be found and increasing the beta parameter also leads to a higher probability that the MTD will not be found.

Overall, these descriptive statistics for each endpoint, the DLT probability curves and MTD histograms give a general idea of how the 3+3 operates under the 9 different scenarios and 4 different maximum dose levels. However, to fully understand how P(D) and beta are related to the MTD, linear regression will be used to determine any linear relationships, interaction terms, R^2 values, and the best model that can be used to analyze and compare each category.

Table 7 - Missing MTD for ***D=9***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
count	2	0	1	0	0	0	1	0	0
proportion	0.013	0	0.0067	0	0	0	0.0067	0	0

4.3 Linear Regression Analysis Results

By the design of this study, P(D) and beta are very likely to impact the probability of DLT at the MTD. To determine the magnitude that these two parameters have on the outcome, linear regression was run. For the initial linear regression model, P(D) and beta were read in as continuous variables, and categorical versions (catBeta and catP(D)) using the cutoff points in Table 1 were created of both variables to check for association. The probability of DLT at the MTD was used as the outcome for this linear model. MTD values at dose level 0 were considered missing values, and D=9 was used as the set dose level for this analysis. Various diagnostic tests were used to verify that the linear regression assumptions of normality, linearity, homoscedasticity, and independence were met. Independence was met due to the nature of the data – each simulation was only evaluated once, and no paired data were used. The assumption of normality was not met when examining the histogram distribution and the Q-Q plot of the regression residuals because there were many outliers and it was left skewed (Figure 15), although it had acceptable skewness and kurtosis values of 0.37 and 0.04, respectively.

To fix this normality issue, three different transformations were considered: arcsine (Figure 16), log (Figure 17), and square root (Figure 18). The log transformation did not help with normality and exacerbated the issue, skewing the data to the right and creating a non-linear Q-Q plot. The arcsine transformation helped with normality by fixing the skewness issue, as well as creating a more linear QQ plot with fewer outliers. This transformation had a skewness and kurtosis value of -0.08 and -0.32, respectively. The square root transformation also helped with normality by fixing the skewness issue, as well as creating a more linear QQ plot with fewer outliers. This transformation had a skewness and kurtosis value of -0.20 and -0.38. Although the square root transformation had worse skewness and kurtosis values compared to the arcsine, the results are very similar, but the square root transformation allows for easier interpretation of the linear regression results. Furthermore, the arcsine transformation makes it difficult to treat the outcome as a measurement variable and typically recommends logistic regression, while the square root transformation is often used for count variables and recommends linear or Poisson regression. Therefore, I chose the square root transformation for the outcome of MTD.

Figure 12 – Untransformed normality test

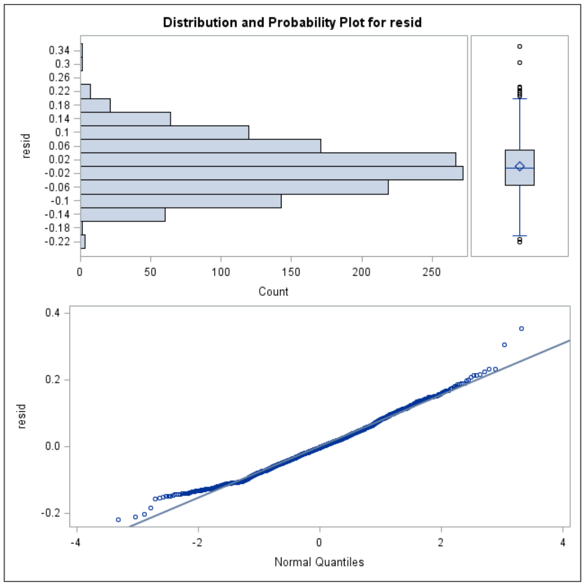


Figure 13 - Arcsine transformation

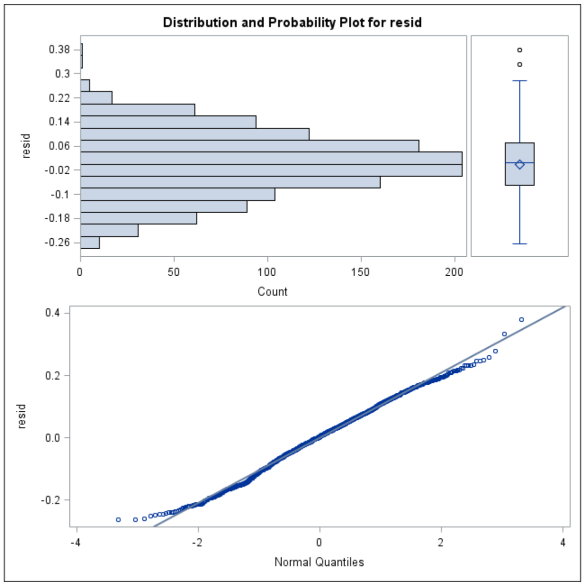


Figure 14 - Log transformation

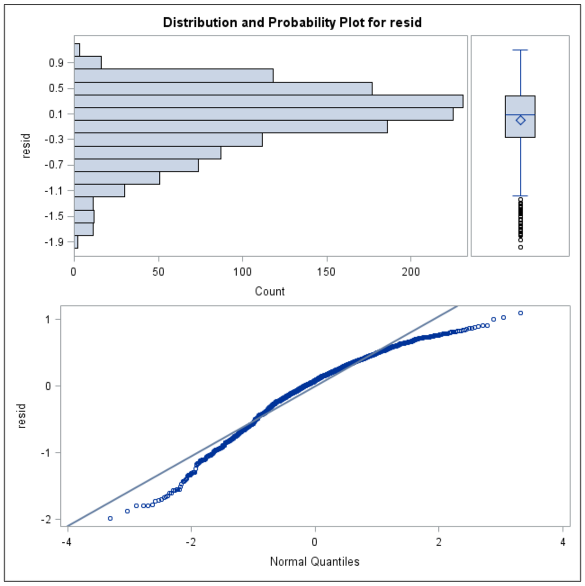
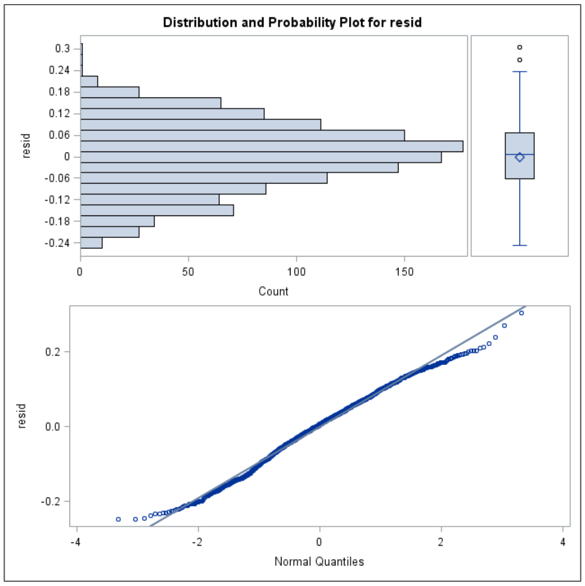


Figure 15 – Square root transformation



A scatterplot of the predicted vs. residuals from the linear regression before the outcome transformation was examined to evaluate homoscedasticity and linearity assumptions (Figure 19). The scatterplot was color-categorized by the number of doses tested in that simulation for clearer identification of each point. Regarding homoscedasticity, there is a large issue throughout the entire graph where the variance of each x value is unequal. The variance on the left side of the graph is significantly smaller than the right side. Similarly, the assumption of linearity is also violated although not as severely. On the left side of the graph, the width of negative values is larger than the width of positive values. To fix this, the square root transformation was used (Figure 20). There remains a slight issue on the left side of the graph (predicted value range of 0.34 to 0.36) where values are concentrated for the dose level 9 points around 0 on the y axis. However, as this area is a small proportion of the overall graph, it is less of an issue. The overall plot appears to be reasonably homoscedastic and is improved over the untransformed graph. The assumption of linearity was also helped with the square root transformation, and all values except from the predicted value range of 0.35 to 0.37 look linear. Therefore, all four linear regression assumptions are met after the square root transformation.

Figure 16 - Predicted vs. Residual scatterplot categorized by doses tested before transformation

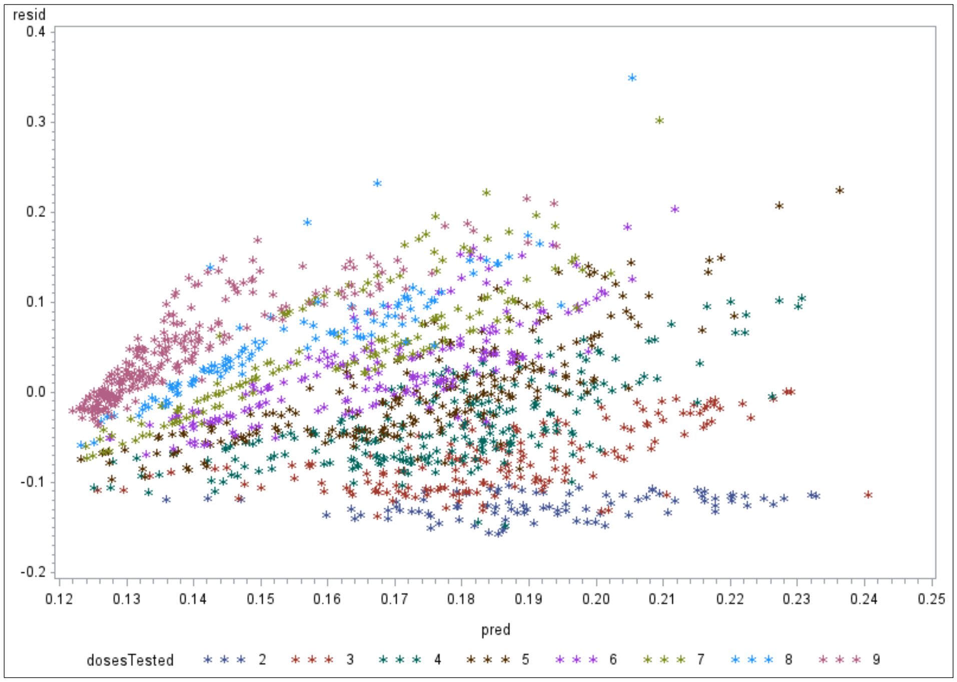
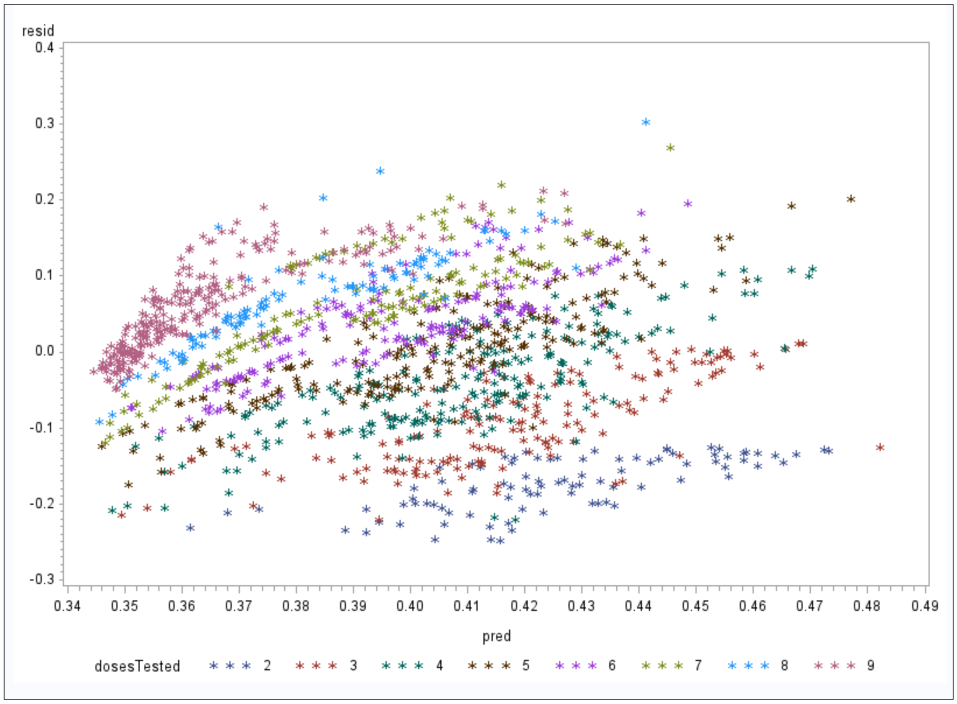


Figure 17 - Predicted vs. Residual scatterplot categorized by doses tested after square root transformation



To understand how P(D) and beta affected the predicted and residual values even further, figure 18 was split up by catBeta and catP(D) (Figure 21). The larger the beta value, the higher the variance is for the predicted values, ranging from 0.35 to 0.53. The residual variance is very similar between the three catBeta values. Similarly, the larger the P(D) value, the higher the variance is for the predicted values. However, this pattern is much more apparent when categorizing by P(D) as the lowest P(D) category is very concentrated in the first third of the graph, while the highest P(D) category is more varied but has higher predicted values.

Figure 18 - Predicted vs. Residual scatterplot categorized by catBeta and catP(D) after square root transformation

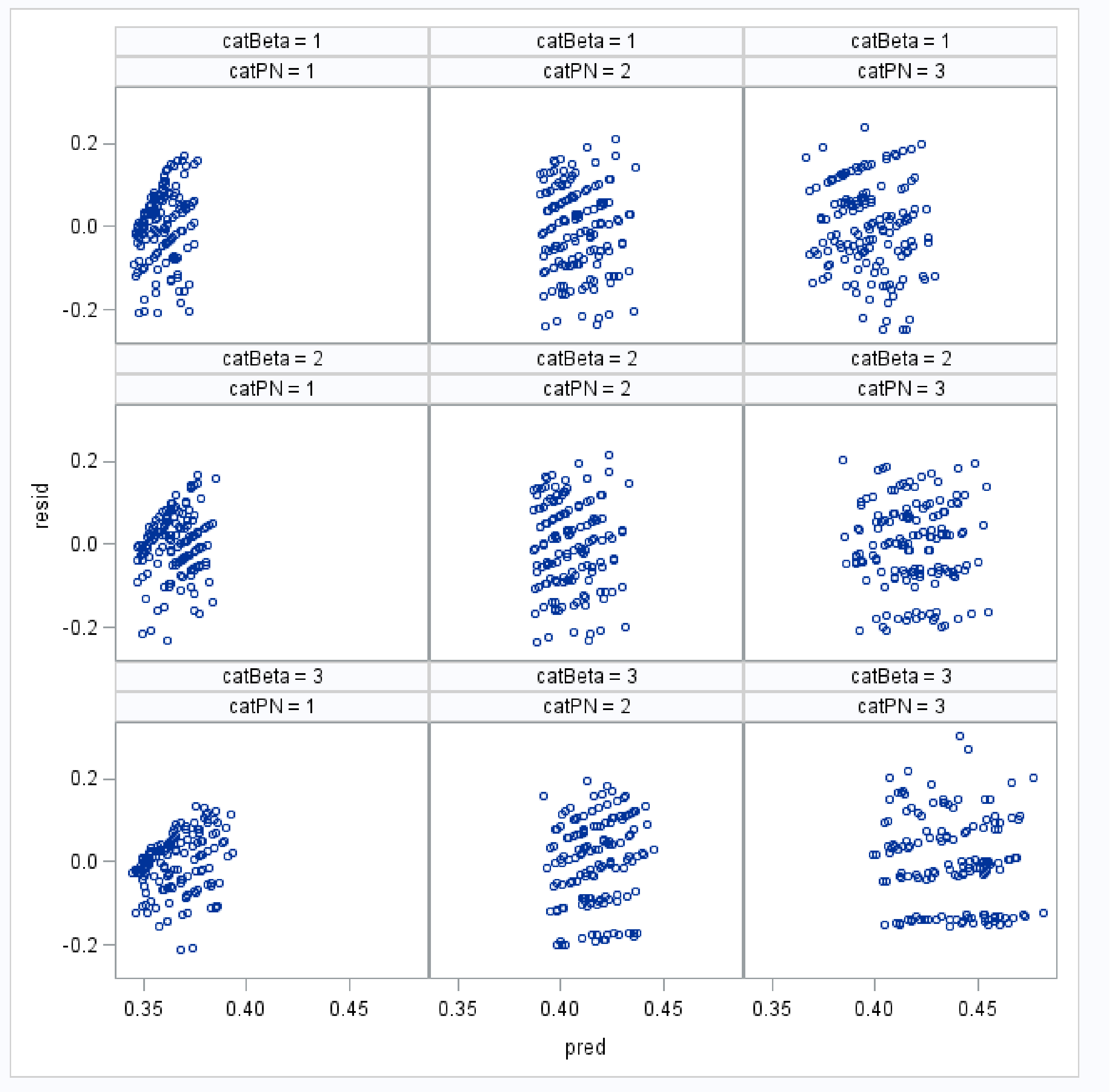


Table 8 - ANOVA Table – Square root-transformed linear regression

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	R^2
Model	6	1.15	0.19	21.07	<.0001	0.09
Error	1339	12.23	0.009			
Corrected Total	1345	13.38				

Table 9 - Type III Sums of Squares – Square root-transformed linear regression

Source	DF	Type III SS	Mean Square	F Value	Pr > F
P(D)	1	0.07	0.07	7.87	0.0051
Beta	1	0.11	0.11	12.54	0.0004
CatBeta	2	0.006	0.003	0.33	0.7221
CatP(D)	2	0.105	0.05	5.77	0.0032

Table 10 - Parameter estimates – Square root-transformed linear regression

Parameter	Estimate	95% CI	Standard Error	t-Value	Pr > |t|
Intercept	0.30	(0.25, 0.35)	0.025	11.82	<.0001
P(D)	0.11	(0.034, 0.194)	0.041	2.80	0.0051
Beta	0.73	(0.327, 1.14)	0.021	3.54	0.0004
CatBeta – low	0.008	(-0.012, 0.028)	0.010	0.81	0.4198
CatBeta – medium	0.004	(-0.011, 0.020)	0.008	0.57	0.5698
CatBeta – high	Reference	-	-	-	-
CatP(D) - low	0.024	(-0.011, 0.060)	0.018	1.37	0.1715
CatP(D) – medium	0.032	(0.009, 0.054)	0.012	2.73	0.0064
CatP(D) - high	Reference	-	-	-	-

The results of this initial linear regression can be found in Tables 8-10. Table 8 shows the ANOVA table results with a F-value of 21.07 and a p-value of <.0001, indicating a strong relationship between the variables in the model and the outcome. However, the model only explains 9.0% of the variability in the outcome which is relatively small. Table 9 shows the type III sums of squares of the transformed model, showing how much each variable contributes to the model when all other variables are already included. Type III sums of squares were used rather than type I because I wanted to examine the overall effect each variable had on the model, rather than one at a time. Beta has the highest F-value of 12.54, while P(D) is slightly smaller with an F-value of 7.87. Both variables are extremely significant with p-values <0.05 and contribute a lot to the model. CatBeta has a small F-value of 0.33 and is not statistically significant (p = 0.72), contributing little to the model. CatP(D), although not as significant as beta and P(D), has a F-value of 5.77 (p=0.003) and still contributes significantly to the model. Table 10 shows the parameter estimate table of the regression. As it uses the Type III sums of squares p-values, the results for the continuous variables are the same. P(D) is statistically significantly associated with the probability of DLT at the MTD (p=0.005). The following interpretations are transformed back to the original scale from the square root scale. For every unit increase in P(D), the probability of DLT at the MTD increases by 0.01 (95% CI 0.001, 0.004). Similarly, beta is statistically significantly associated with the probability of DLT at the MTD (p=<.0001). For every unit increase in beta, the probability of DLT at the MTD increases by 0.53 (95% CI 0.11, 1.30). The different levels of categorical beta are not significantly different from one another, with low beta and medium beta having p-values of 0.42 and 0.57, respectively, when compared to the base level of high beta. The low level of P(D) is not significantly different from the base level of high P(D) (p=0.17), while the medium level of P(D) is statistically significantly different from the base level (p=0.006).

After verifying all linear regression assumptions and running the initial linear model, four different prediction methods were used to select the best model: forward selection, backwards elimination, stepwise and all subsets. Variables included in the full model included P(D), beta, catBeta, catP(D), P(D)*catBeta, catP(D)*beta, and catP(D)*catBeta.

Backwards elimination selected a model (Tables 11-12) with P(D), beta, catBeta, catP(D), beta*catP(D), and catBeta*catP(D). This model had a R^2 of 0.082, but it had a high variable count making it less parsimonious. It had a high F-value of 19.95 and an overall p-value of <.0001, and all variables in the model had significant p-values except for catP(D) where p=0.18. Forward selection and stepwise selected a model (Tables 13-14) with only P(D) and beta in the model with a R^2 of 0.066. This model had a larger F-value of 47.42 compared to the backwards selection model's F-value of 19.95. As the adjusted R^2 of the two models were very similar and the P(D) and beta model was much more parsimonious, this model was chosen as the final prediction model. All parameters were statistically significant (all p<0.0001), with P(D) having a beta estimate of 0.08 and beta having a beta estimate of 0.66. However, all models had a relatively low R^2 value of under 0.1, indicating that the model explained very little variability in the outcome of the probability of DLT at the MTD.

Further model diagnostics such as jackknife and studentized residuals, Cook's D values, and D-Betas were used to evaluate the model, but only 13 observations out of 1350 were found with residual values over 2, but none were over 2.5 or statistically significant. There were no Cook's D values over 0.5, and no D-Beta values stood out. Collinearity of P(D) and beta were also checked and a VIF value of 1.36 indicated no issues.

Overall, there was a statistically significant relationship between P(D), beta, and the probability of DLT at the MTD, while other categorical terms and interaction terms were less significant.

Table 11 – Backwards elimination ANOVA table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	Adjusted R^2
Model	6	1.10	0.18	19.04	<.0001	0.074
Error	1343	12.91	0.01			
Corrected Total	1349	14.01				

Table 12 – Backwards elimination parameter estimates

Parameter	Estimate	Standard Error	t-Value	Pr > |t|
Intercept	0.37	0.019	19.41	<.0001
P(D)	0.09	0.042	2.13	0.337
Beta	3.10	0.56	5.56	<.0001
CatBeta	-0.04	0.012	-3.60	0.0003
CatP(D)	-0.02	0.014	-1.29	0.20
Beta*CatP(D)	-1.00	0.22	-4.61	<.0001
CatBeta*CatP(D)	0.02	0.007	2.86	0.0043

Table 13 - Forward selection and stepwise selection ANOVA table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	Adjusted R^2
Model	2	0.87	0.44	44.81	<.0001	0.061
Error	1347	13.14	0.01			
Corrected Total	1349	14.02				

Table 14 - Forward selection and stepwise selection parameter estimates

Parameter	Estimate	Standard Error	t-Value	Pr > |t|
Intercept	0.35	0.006	55.21	<.0001
P(D)	0.067	0.017	3.89	<.0001
Beta	0.58	0.132	4.42	<.0001

5 Discussion

The goal of this study is to evaluate how well the 3+3 design performs on five different outcomes in a variety of scenarios. Rather than comparing the 3+3 to other designs, the goal of this study is to find scenarios where the 3+3 performs well, finding a low average probability of DLT so researchers can justify the usage of the 3+3 design if their drug fits a similar profile.

The main outcome of interest is the selection of the MTD for recommendation in phase II trials, typically the most important result of a phase I clinical trial. Secondary outcomes of interest are related to the safety of these trials, including the number of DLTs in a trial, the highest dose tested in a trial, and how many patients on average are enrolled in a trial. Additionally, as the 3+3 may fail to determine the MTD due to a high starting dose, the outcome of missing MTD values is also considered. To evaluate the effectiveness of the 3+3 design using these outcomes of interest, 5400 simulations were run with varying DLT probability curves to emulate different drug scenarios and number of dose levels.

The simulations outputted an average probability of DLT at the MTD for D=3,5,7, and 9 ranging between 0.14 and 0.23, depending on the parameters P(D) and beta. This range is relatively low for a probability of DLT but is what researchers have experienced and seem to desire clinically for phase I trials. Histograms detailing the distribution of MTD values for the simulations explain the effect of the parameters even further. For low P(D) categories, or simulations where the probability of DLT at the highest dose level is below 30%, the highest dose level probability may be too low as most of the simulations end at the highest dose level. Going beyond the highest dose level, D, is potentially considered when no dose was found due to lower than expected toxicity; however, this simulation study does not escalate the dosage further and selects the highest dose as the MTD. Although it is better for patients to not experience a DLT, the fact that the simulations end at the last dose level gives less information about a drug's side effects and tolerability overall. It is possible that the correct MTD is the last dose level in this situation, but it is more likely that a better MTD could be selected if the dosage were increased further. Therefore, allowing the dosage to be escalated past the highest set dose level opens other potential simulation outcomes and may be considered in the future. Similarly, the lowest dose level possible in this simulation study was dose level 0, allowing for one level of de-escalation if two or more patients experienced DLTs at dose level 1. When dose level 0 is reached, this is considered a trial where no dose was found due to unacceptable toxicity. Going beyond dose level 0 is sometimes recommended for promising drugs and can be done so by modifying the protocol but will not be considered in this simulation trial. This study does consider dose level 0 as a failure to determine the MTD but does not consider dose level 9 as a failure and accepts it as the MTD.

Outcomes related to safety also were relatively low. The average number of DLTs in a trial ranged between 2 and 4, while the average number of patients in a trial was around 24. The highest dose level tested depended heavily on the P(D) category but was expected – the lower the probability of DLT, and the more likely the simulation will end on a higher dose level.

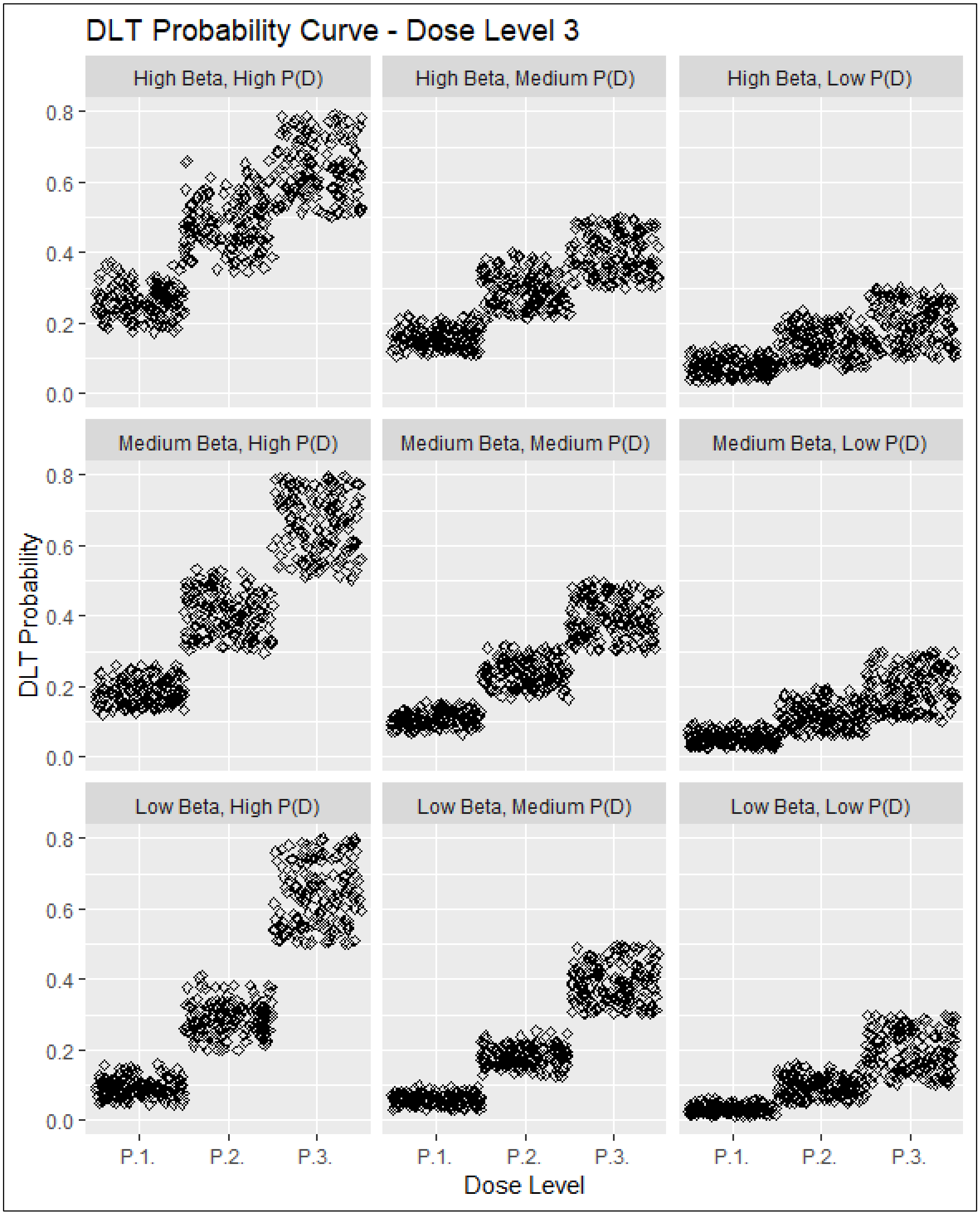
Although researchers want to accurately pinpoint the MTD, the 3+3 never specifies a desired probability of DLT as an outcome. More statistically powerful designs such as the CRM and other model-based designs allow researchers to specify a desired probability to close in on allowing for more flexibility, yet researchers still prefer the 3+3's intrinsically low probabilities of DLT. "In a review of more than 1200 phase I studies from 1991 to 2006, more than 98% of trials utilized the 3+3 dose escalation scheme" (Hansen et al, 2014). The simplicity and effectiveness in choosing a relatively low probability of DLT, potentially ranging from 10% to 25%, are why the 3+3 design remains popular today. However, this desired probability range also heavily depends on the definition of a DLT for that trial. For example, milder toxicities such as neutropenia may be considered a DLT in one phase I trial, while more severe toxicities that immediately threaten a patient's life may be considered a DLT in another trial. Milder DLT definitions may allow for slightly higher probabilities of DLT ranging up to 35%, while more severe DLTs may encourage lower DLT probabilities. Another issue with the definition of a DLT is how it affects the patient's quality of life during and after the trial. Prolonged moderate DLTs may not be taken as seriously during the trial but may seriously affect the patient's quality of life in the long-term. Additionally, late DLTs that happen after a trial may not be considered when recommending a dose.

Some caveats for the simulations include the sample size, randomization seed, and outcome transformation. When calculating the sample size, an estimated sample size of 147 was recommended using the variability table in Figure 5. However, this sample size was rounded up to 150 to add additional power to the study at no cost, while having a large enough sample size to observe a difference between each scenario. This increase of sample size had no cost due to it being a simulation, but it is acknowledged that there is an ethical and financial cost of changing this sample size recommendation in real trials. The randomization seed to generate the data was also not saved, which limits the reproducibility of the results. However, the raw data and the program used to generate the data were saved and can reproduce relatively similar results. Finally, the outcome was unable to be evaluated as untransformed or log-transformed, which makes interpretation of the linear regression results more meaningful. Choosing to use the square root transformation, although it fulfills the linear regression assumptions, makes it harder to interpret the regression results, although they provide a better understanding of how P(D), beta, and their interaction terms change the final probability of DLT at the MTD.

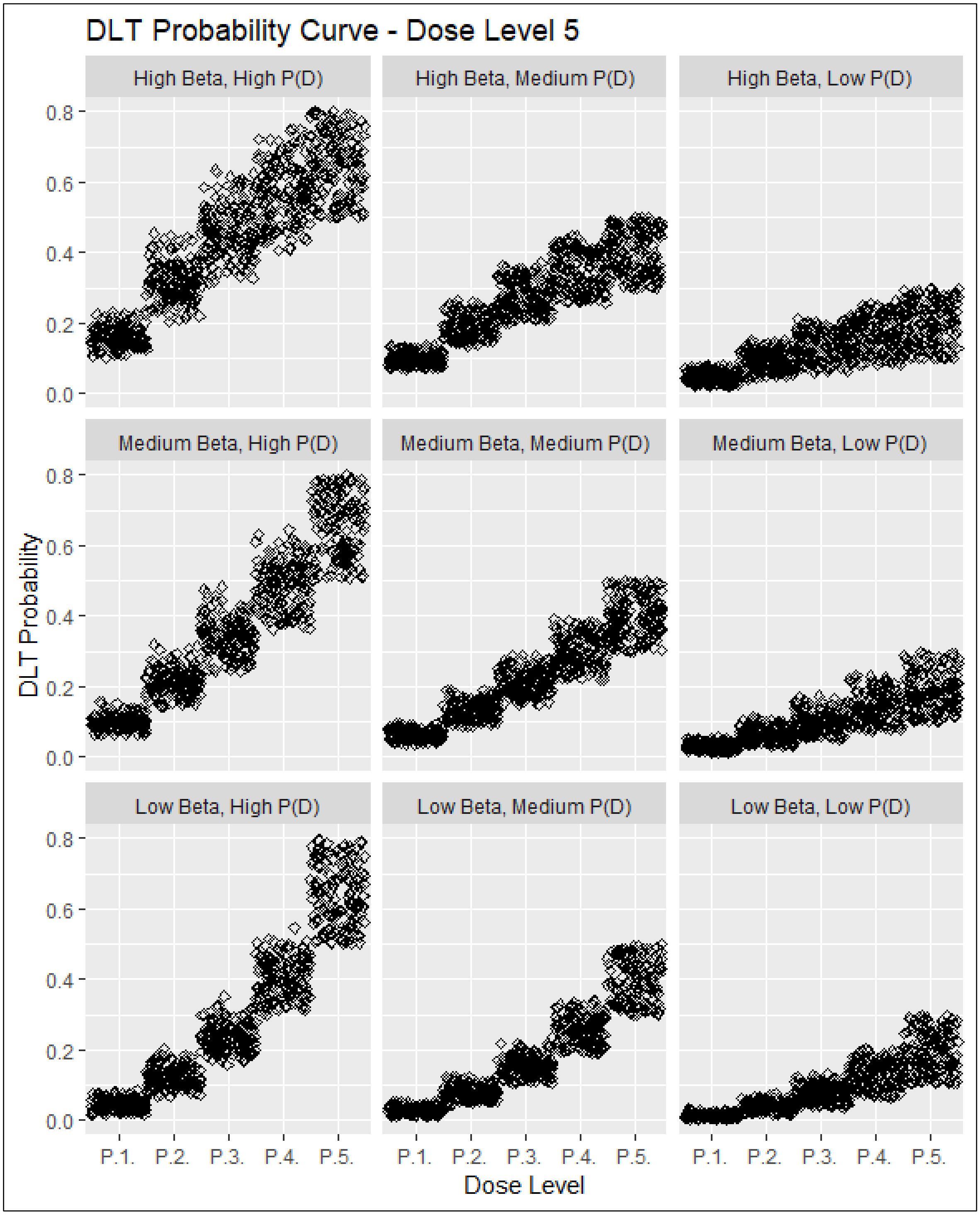
The 3+3 rules-based design is a tried and true method still being used by a majority of phase I clinical trials. It provides a low probability of DLT in most scenarios, capping out at 25% for this study’s simulations, along with a low average count of DLTs and patient recruitment. More statistically powerful methods such as CRM are slowly being adopted due to their flexibility and ability to specify a desired DLT; however, it is a more complicated design requiring a prior model and adaptation of that model with new data. Researchers who have been using the much simpler 3+3 and obtaining desirable results may be discouraged from adopting these more complicated designs, which may cater towards more specific scenarios. Therefore, the 3+3 will remain the most effective general model for phase I clinical trials until a more statistically powerful method which is equally as simplistic is discovered.

6 Appendix

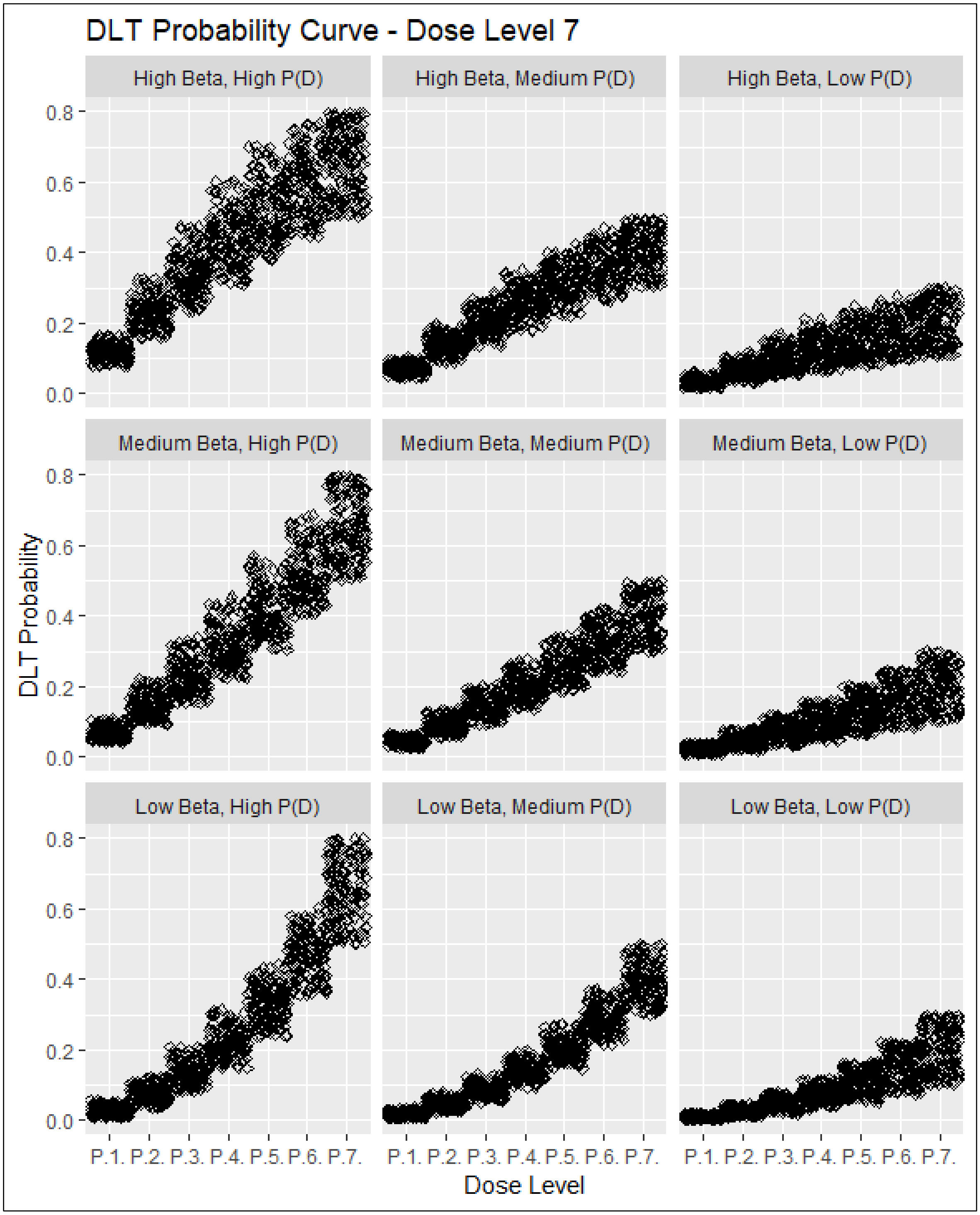
Appendix 1 - DLT probability curve for each category for D=3 with jitter



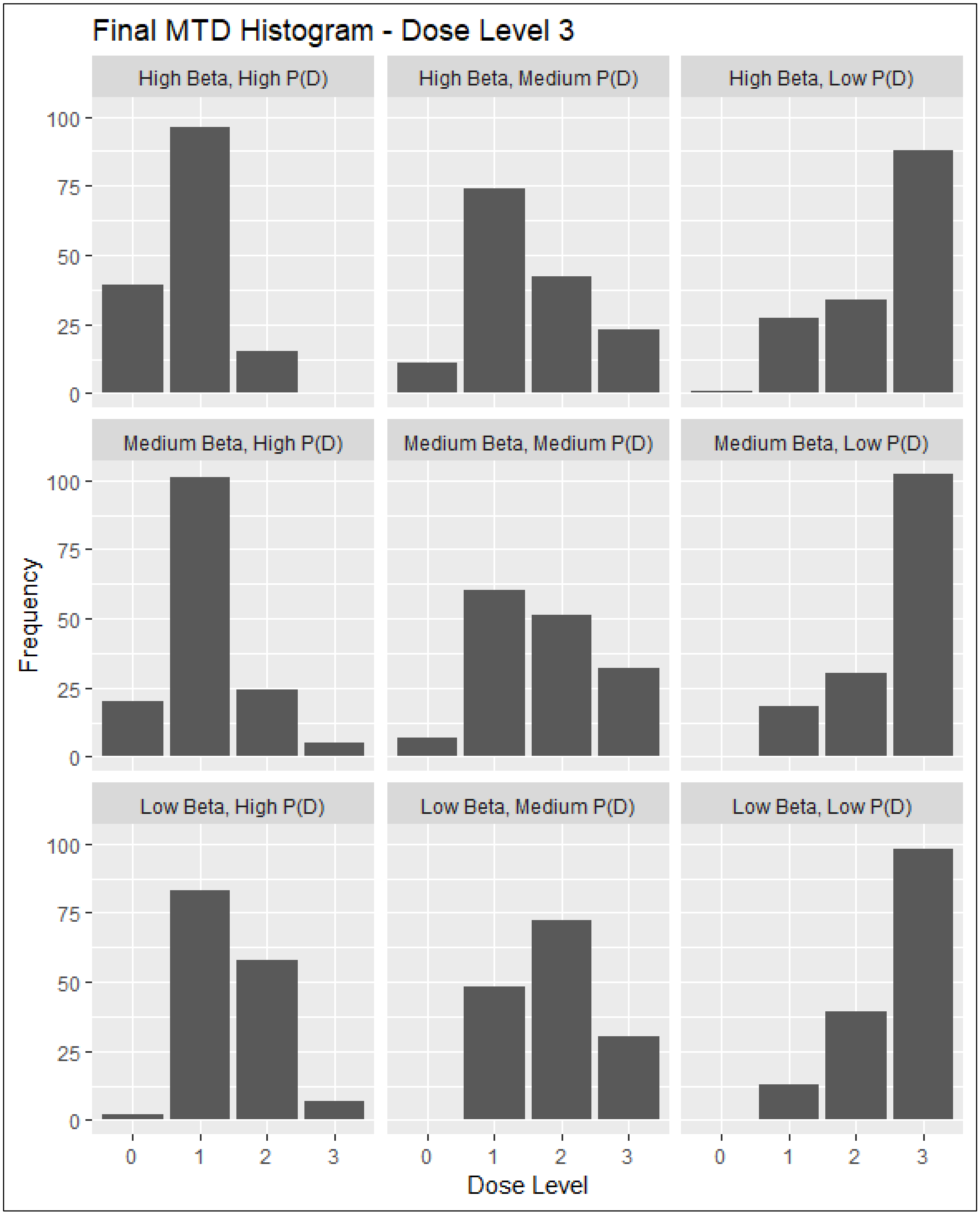
Appendix 2 - DLT probability curve for each category for D=5 with jitter



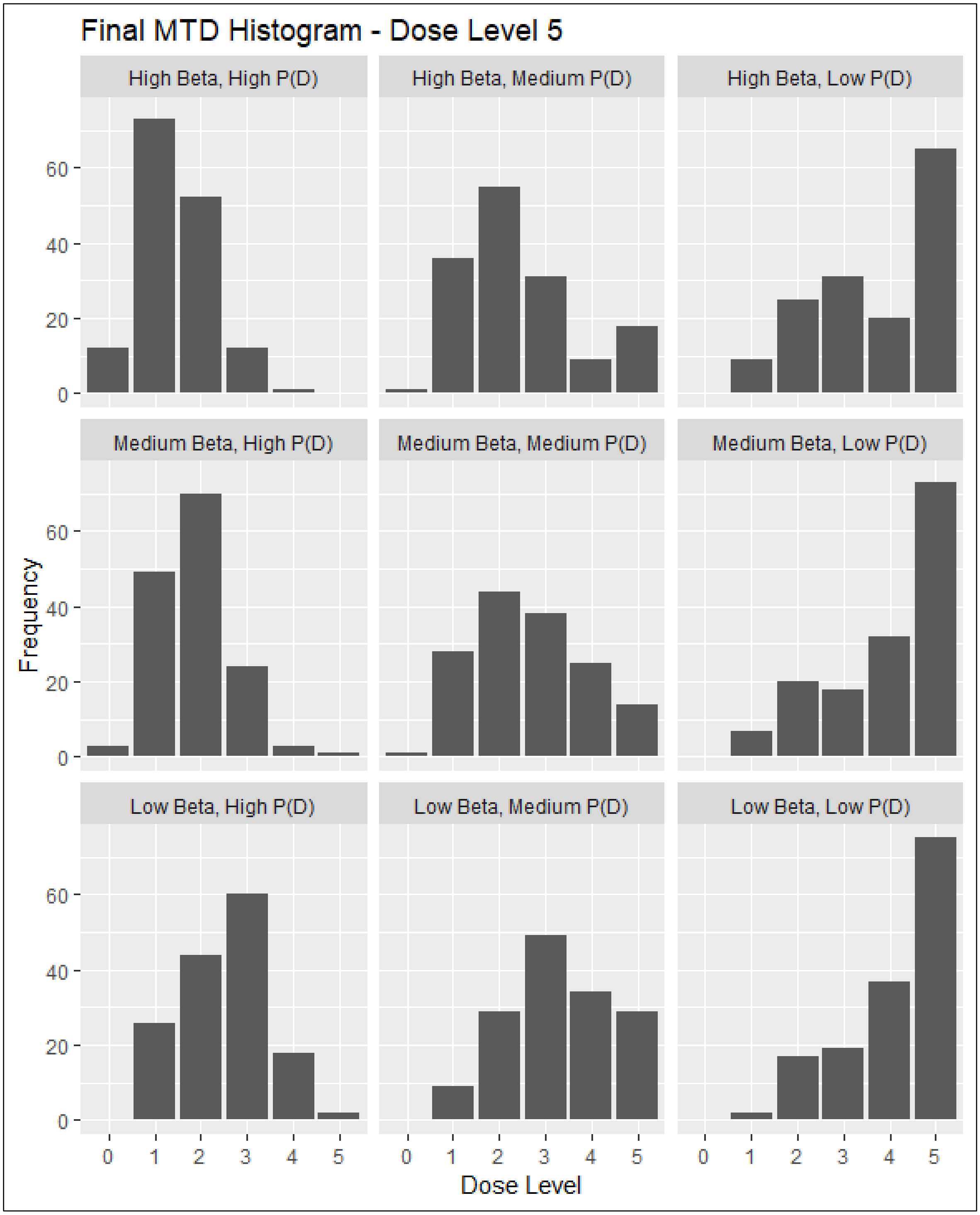
Appendix 3 - DLT probability curve for each category for D=7 with jitter



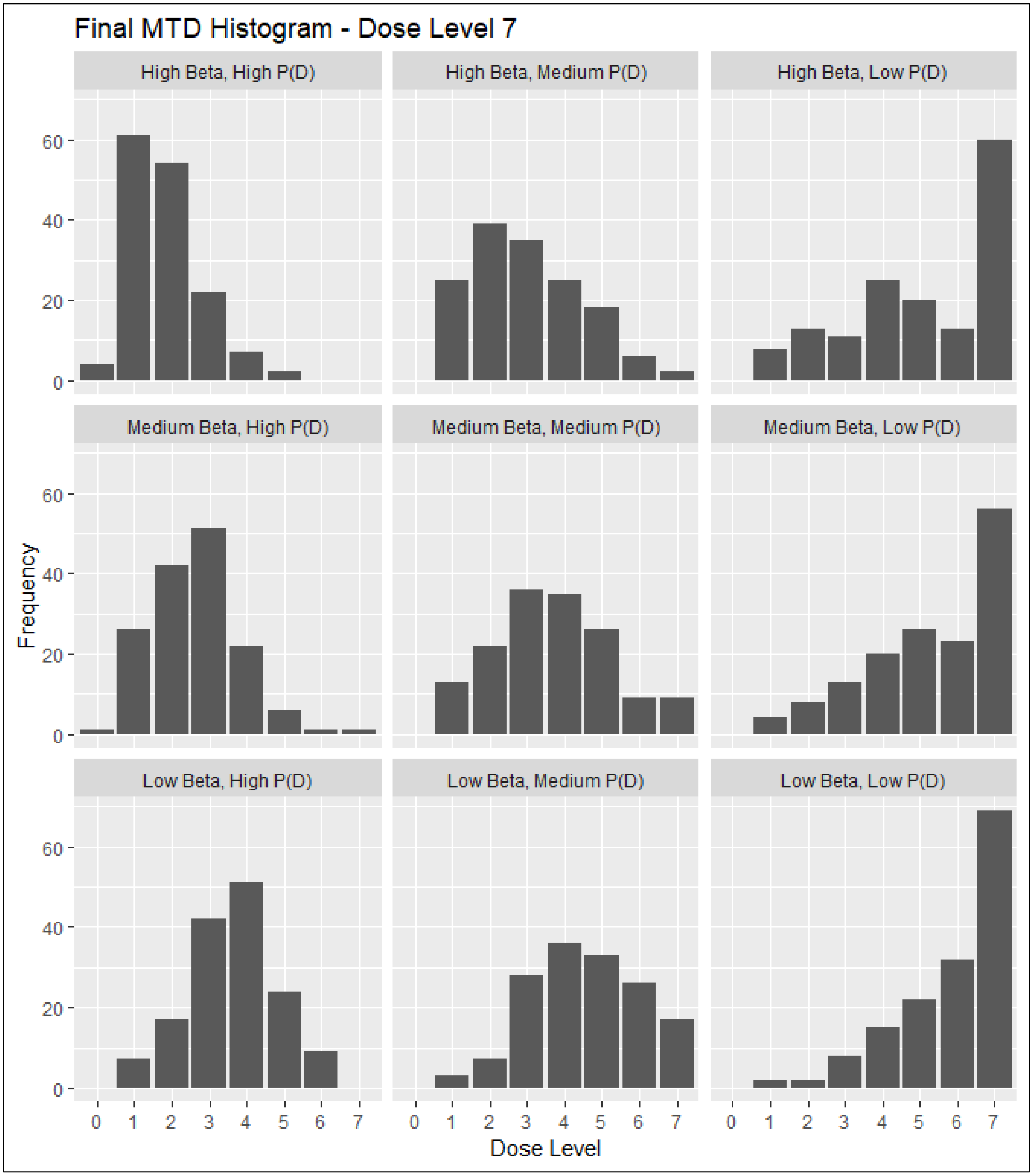
Appendix 4 - MTD Histogram for all categories for D=3



Appendix 5 - MTD Histogram for all categories for D=5



Appendix 6 - MTD Histogram for all categories for D=7



Appendix 7 - Probability of DLT at MTD for ***D=3***

	High Beta	High Beta	High Beta	Low Beta	Low Beta	Low Beta	Med Beta	Med Beta	Med Beta
	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)
mean	0.21	0.16	0.22	0.19	0.15	0.18	0.21	0.16	0.21
sd	0.14	0.062	0.11	0.12	0.07	0.11	0.13	0.064	0.11
median	0.24	0.14	0.2	0.14	0.15	0.17	0.19	0.15	0.19
q1	0	0.11	0.15	0.094	0.11	0.071	0.15	0.12	0.11
q3	0.3	0.2	0.3	0.28	0.2	0.22	0.24	0.19	0.28
min	0	0	0	0	0.024	0.031	0	0.036	0
max	0.57	0.3	0.47	0.56	0.3	0.49	0.68	0.3	0.48

Appendix 8 - Probability of DLT at MTD for ***D=5***

	High Beta	High Beta	High Beta	Low Beta	Low Beta	Low Beta	Med Beta	Med Beta	Med Beta
	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)
mean	0.23	0.15	0.22	0.19	0.14	0.2	0.2	0.14	0.19
sd	0.12	0.057	0.095	0.11	0.063	0.11	0.096	0.061	0.095
median	0.2	0.14	0.21	0.18	0.13	0.17	0.2	0.14	0.17
q1	0.15	0.12	0.14	0.11	0.098	0.12	0.11	0.1	0.12
q3	0.31	0.19	0.29	0.26	0.18	0.28	0.25	0.18	0.25
min	0	0.036	0	0.028	0.012	0.019	0	0.021	0
max	0.58	0.3	0.48	0.55	0.3	0.45	0.52	0.3	0.41

Appendix 9 - Probability of DLT at MTD for ***D=7***

	High Beta	High Beta	High Beta	Low Beta	Low Beta	Low Beta	Med Beta	Med Beta	Med Beta
	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)
mean	0.2	0.15	0.19	0.2	0.15	0.19	0.19	0.14	0.18
sd	0.1	0.058	0.088	0.1	0.065	0.095	0.093	0.055	0.091
median	0.19	0.15	0.18	0.19	0.14	0.18	0.19	0.14	0.18
q1	0.13	0.11	0.13	0.12	0.11	0.12	0.13	0.11	0.12
q3	0.26	0.18	0.25	0.26	0.19	0.25	0.25	0.17	0.24
min	0	0.019	0.055	0.015	0.016	0.013	0	0.016	0.035
max	0.61	0.28	0.4	0.54	0.3	0.49	0.54	0.29	0.46

Appendix 10 - Highest dose level tested for ***D=3***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
mean	2.1	2.8	2.5	2.5	2.9	2.7	2.3	2.9	2.6
sd	0.35	0.38	0.5	0.5	0.26	0.45	0.45	0.32	0.5
median	2	3	2	2	3	3	2	3	3
q1	2	3	2	2	3	2	2	3	2
q3	2	3	3	3	3	3	3	3	3
min	2	2	2	2	2	2	2	2	2
max	3	3	3	3	3	3	3	3	3

Appendix 11 - Highest dose level tested for ***D=5***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
mean	2.6	4.3	3.4	3.6	4.6	4.2	3	4.5	3.6
sd	0.78	0.94	1.1	0.96	0.74	0.92	0.84	0.9	1.1
median	2	5	3	4	5	4	3	5	4
q1	2	4	3	3	5	4	2	4	3
q3	3	5	4	4	5	5	3	5	5
min	2	2	2	2	2	2	2	2	2
max	5	5	5	5	5	5	5	5	5

Appendix 12 - Highest dose level tested for ***D=7***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
mean	2.9	5.7	4.1	4.7	6.4	5.5	3.7	6	4.7
sd	0.96	1.6	1.4	1.2	1.1	1.3	1.1	1.4	1.5
median	3	6	4	5	7	6	4	7	5
q1	2	5	3	4	6	5	3	5	4
q3	3	7	5	6	7	7	4	7	6
min	2	2	2	2	2	2	2	2	2
Max	6	7	7	7	7	7	7	7	7

Appendix 13 - Number of DLTs for ***D=3***

	High Beta	High Beta	High Beta	Low Beta	Low Beta	Low Beta	Med Beta	Med Beta	Med Beta
	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)
mean	3.9	1.4	2.8	2.9	1.2	2.4	3.4	1.1	2.7
sd	1.2	1.3	1.3	1.1	1.2	1.2	1.3	1.1	1.4
median	4	1	3	3	1	3	3	1	3
q1	3	0	2	2	0	2	3	0	2
q3	5	2	4	3	2	3	4	2	4
min	2	0	0	0	0	0	0	0	0
max	7	7	7	8	5	6	8	4	6

Appendix 14 - Number of DLTs for ***D=5***

	High Beta	High Beta	High Beta	Low Beta	Low Beta	Low Beta	Med Beta	Med Beta	Med Beta
	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)
mean	3.6	1.9	2.8	3.2	1.8	2.7	3.2	1.8	2.8
sd	1.1	1.3	1.2	1.2	1.3	1.3	1.1	1.2	1.2
median	3	2	3	3	2	3	3	2	3
q1	3	1	2	2	1	2	2	1	2
q3	4	3	3	4	3	3	4	3	3
min	2	0	0	0	0	0	2	0	0
max	7	5	7	7	5	7	8	5	7

Appendix 15 - Number of DLTs for ***D=7***

	High Beta	High Beta	High Beta	Low Beta	Low Beta	Low Beta	Med Beta	Med Beta	Med Beta
	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)	High P(D)	Low P(D)	Med P(D)
mean	3.2	2.4	3.3	3.4	2.1	3	3.2	2.3	3.2
sd	1.1	1.4	1.1	1.1	1.4	1.2	0.96	1.5	1.4
median	3	2	3	3	2	3	3	2	3
q1	2	2	2	3	1	2	3	1	2
q3	4	3	4	4	3	4	4	3	4
min	2	0	1	2	0	0	2	0	0
max	7	7	7	8	6	6	6	7	9

Appendix 16 - Total number of patients for ***D=3***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
mean	11	8.9	11	10	8.8	10	11	8.5	10
sd	2.8	2.5	2.7	2.2	2.5	2.4	2.7	2.4	2.3
median	12	9	12	9	9	9	12	9	12
q1	9	6	9	9	6	9	9	6	9
q3	12	9	12	12	12	12	12	9	12
min	9	6	6	6	6	6	6	6	6
max	21	21	18	21	18	18	21	15	18

Appendix 17 - Total number of patients for ***D=5***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
mean	12	14	13	14	15	15	12	15	14
sd	2.7	2.6	3	3.1	3	3.4	3.3	3.2	3.6
median	12	15	12	12	15	15	12	15	12
q1	9	12	12	12	12	12	9	12	12
q3	12	15	15	15	18	18	12	18	15
min	9	9	9	9	9	9	9	9	9
max	21	21	24	24	21	24	21	24	24

Appendix 18 - Total number of patients for ***D=7***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
mean	12	20	16	17	21	20	14	20	17
sd	2.8	4.9	5.2	4.5	3.9	4.6	3.8	4.7	5.5
median	12	21	15	18	21	21	12	21	18
q1	9	18	12	15	18	18	12	18	12
q3	12	24	18	21	24	21	17	24	21
min	9	9	9	9	9	9	9	9	9
max	21	30	33	30	30	30	27	30	33

Appendix 19 - Missing MTD for ***D=3***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
count	39	1	11	2	0	0	20	0	7
proportion	0.26	0.0067	0.073	0.013	0	0	0.13	0	0.047

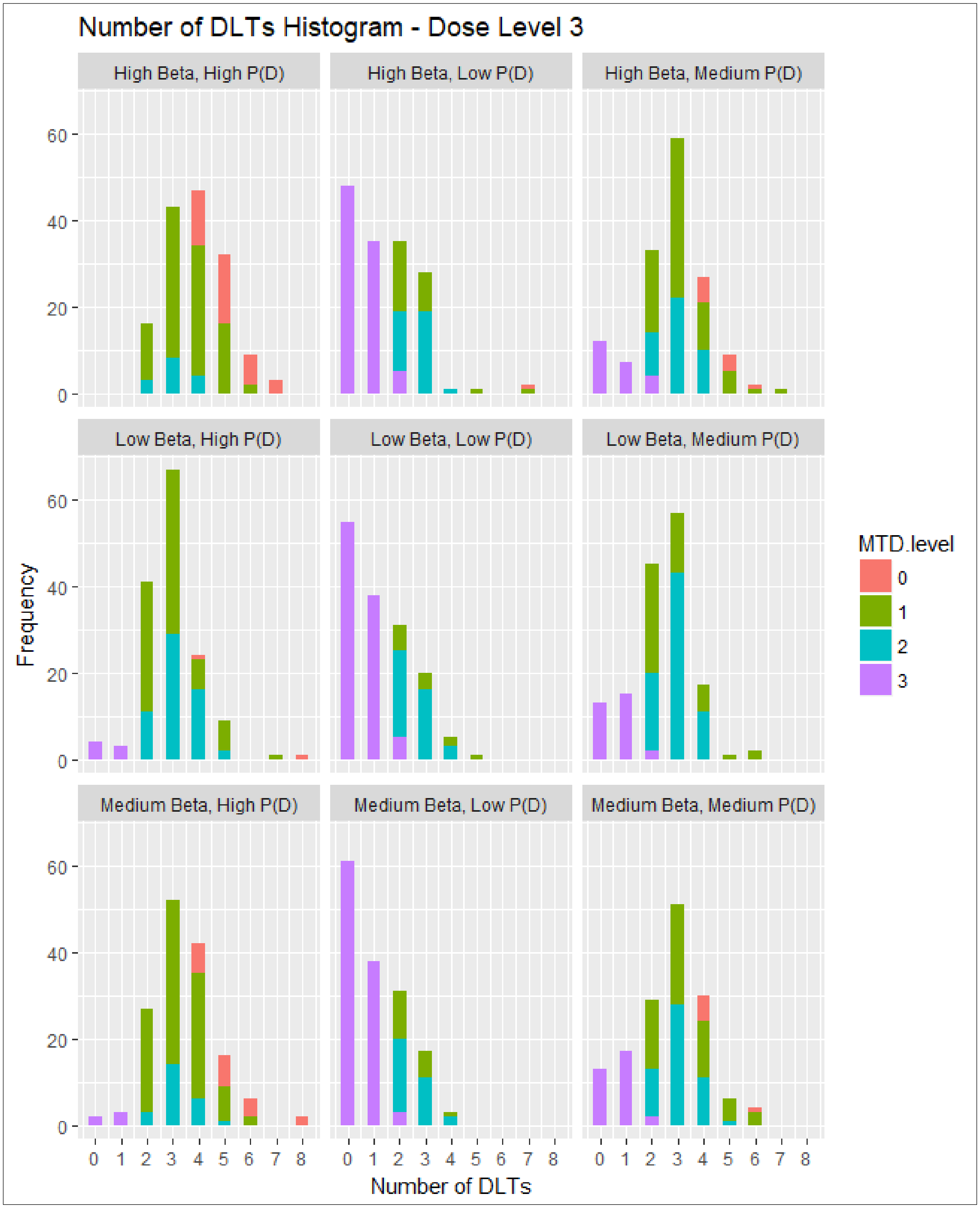
Appendix 20 - Missing MTD for ***D=5***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
count	12	0	1	0	0	0	3	0	1
proportion	0.08	0	0.0067	0	0	0	0.02	0	0.0067

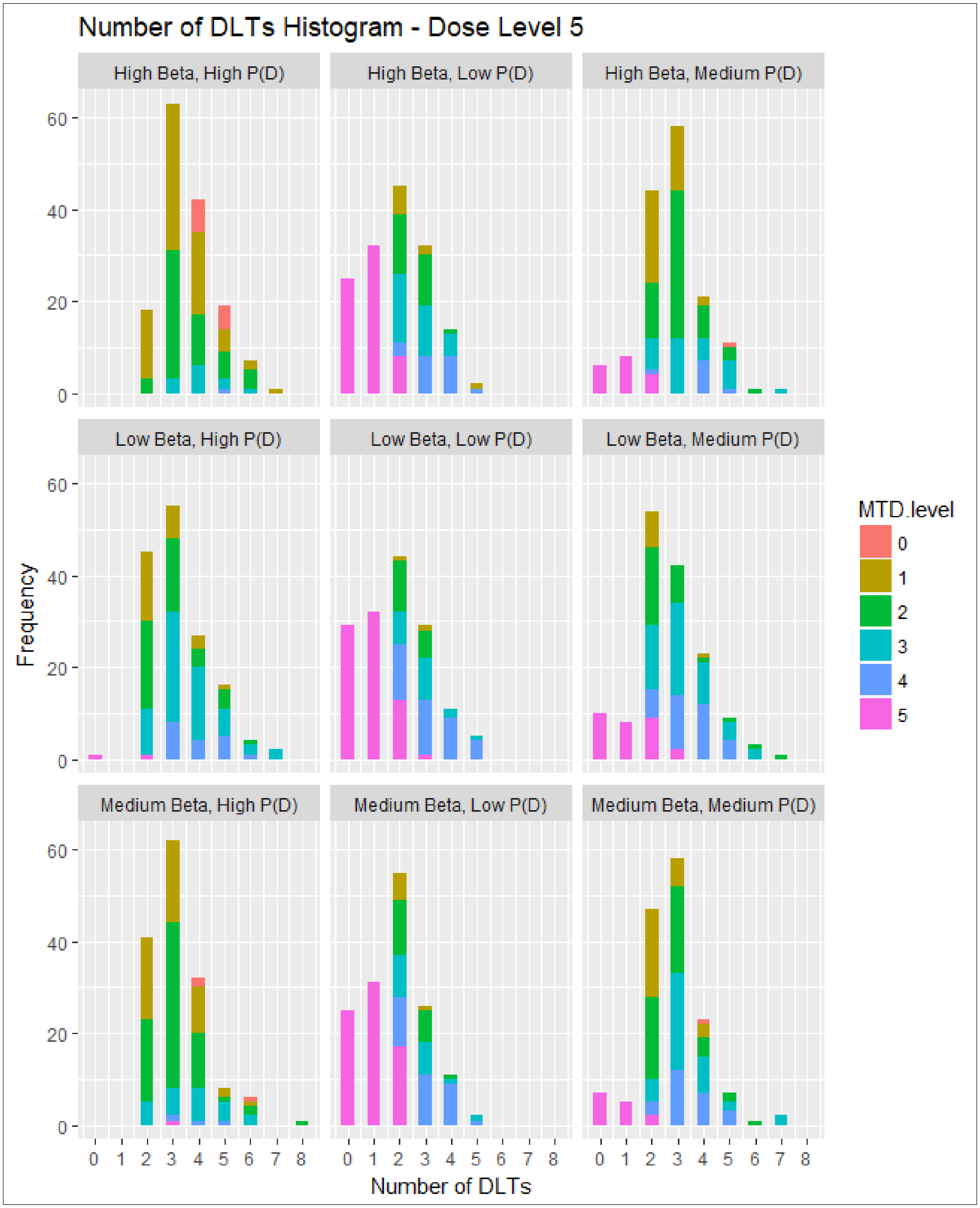
Appendix 21 - Missing MTD for ***D=7***

	High Beta High P(D)	High Beta Low P(D)	High Beta Med P(D)	Low Beta High P(D)	Low Beta Low P(D)	Low Beta Med P(D)	Med Beta High P(D)	Med Beta Low P(D)	Med Beta Med P(D)
count	4	0	0	0	0	0	1	0	0
proportion	0.027	0	0	0	0	0	0.0067	0	0

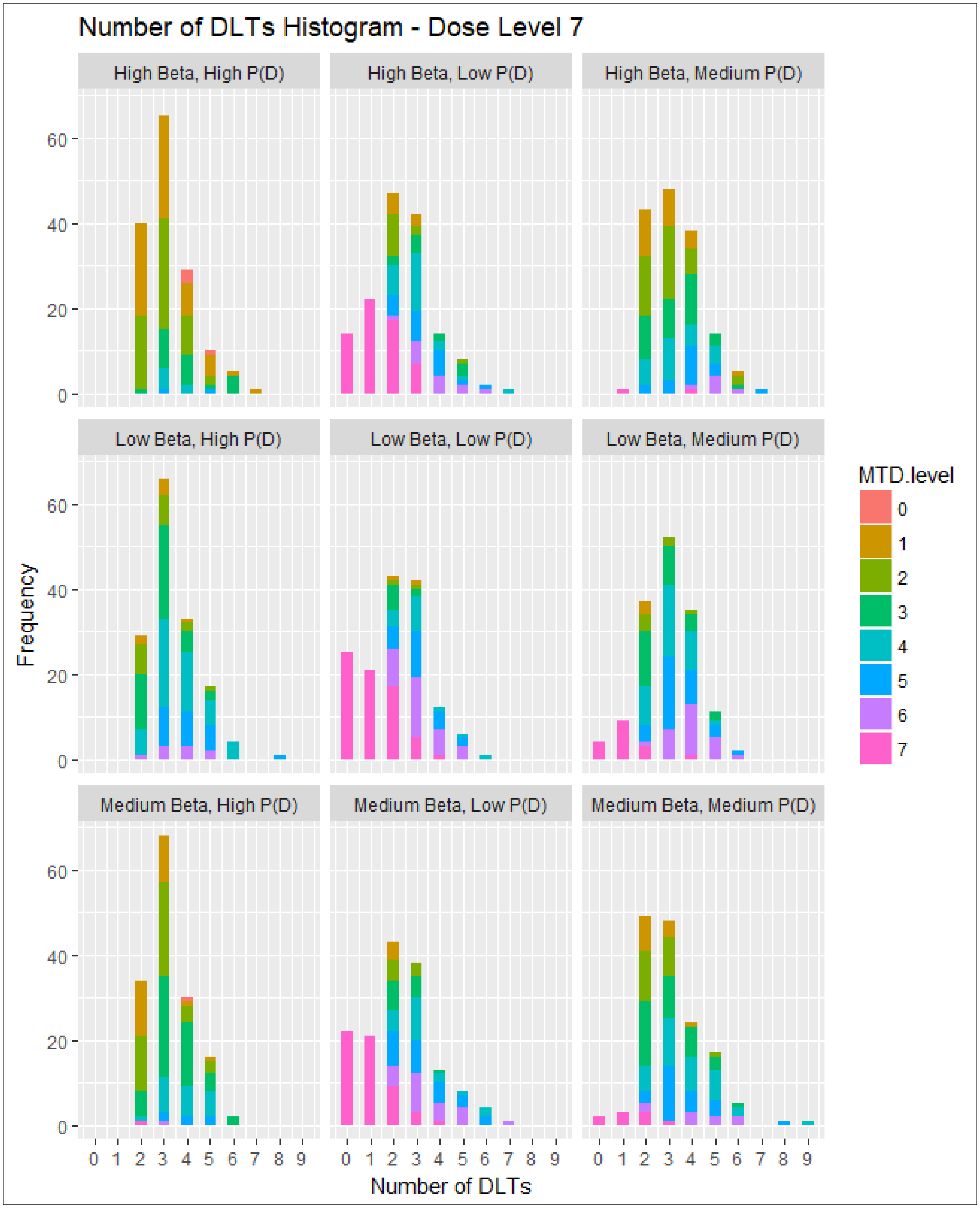
Appendix 22 - Histogram of Number of DLTs categorized by MTD level for D=3



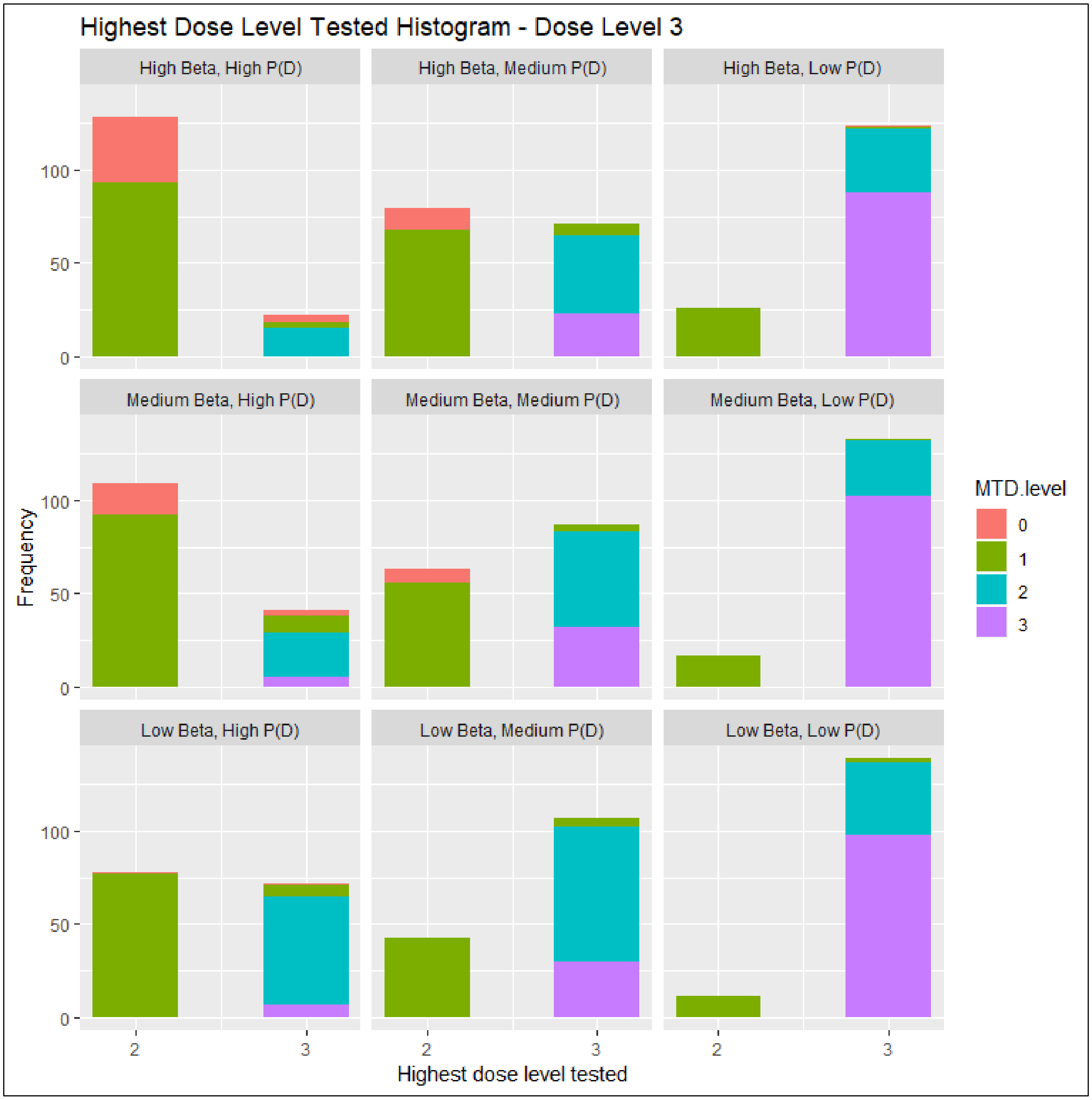
Appendix 23 - Histogram of Number of DLTs categorized by MTD level for D=5



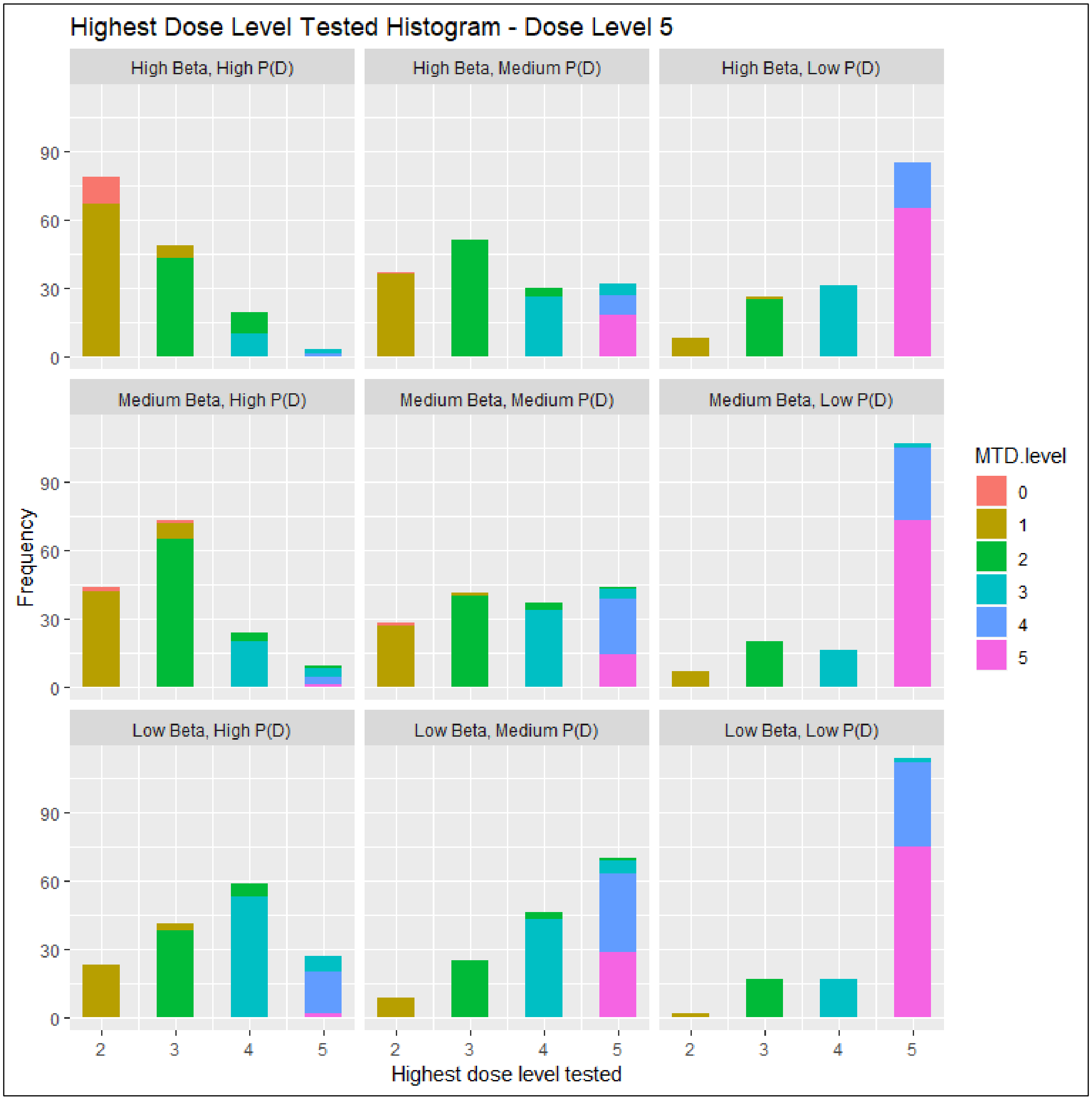
Appendix 24 - Histogram of Number of DLTs categorized by MTD level for D=7



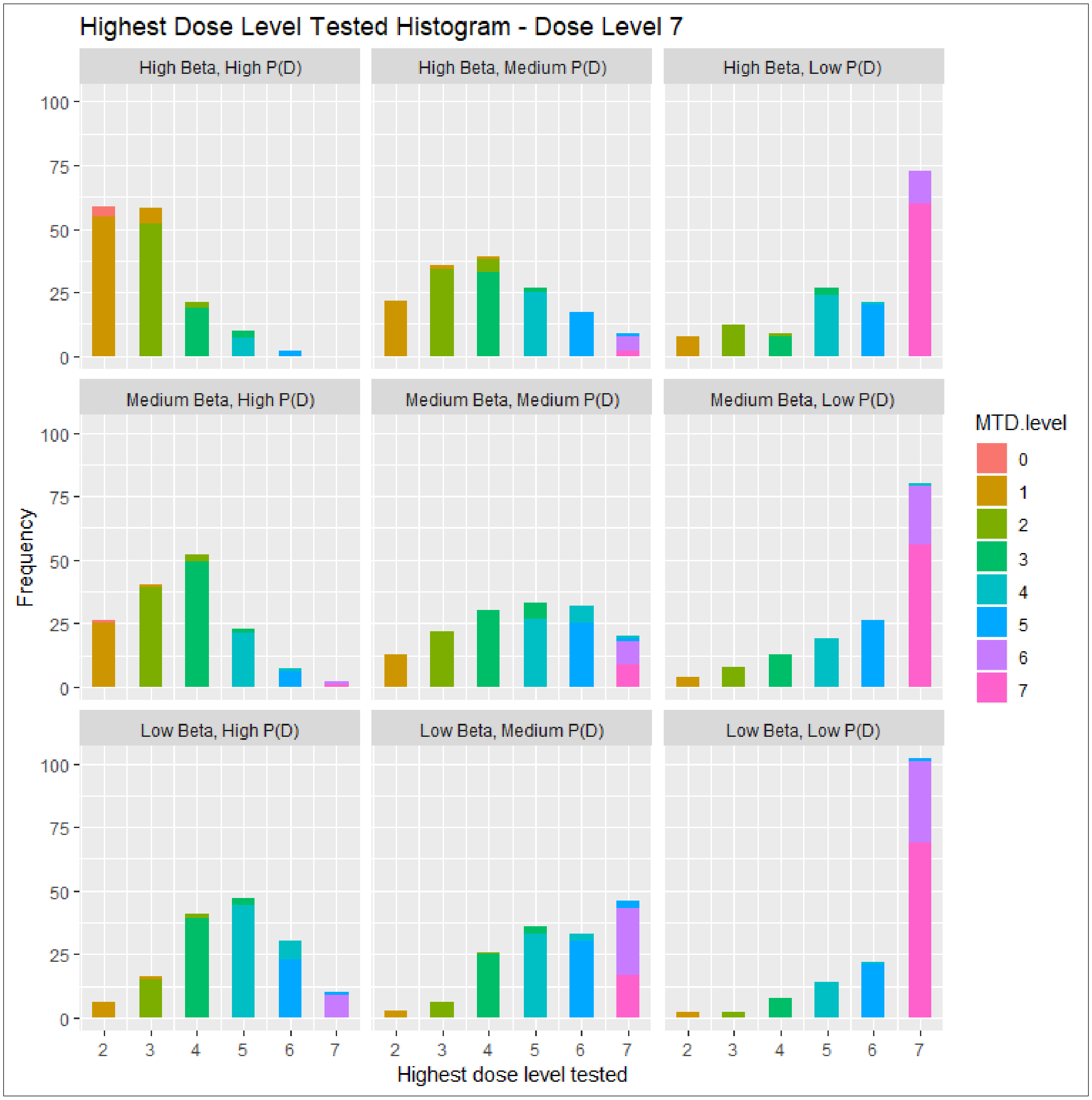
Appendix 25 - Histogram of highest dose level tested categorized by MTD level for D=3



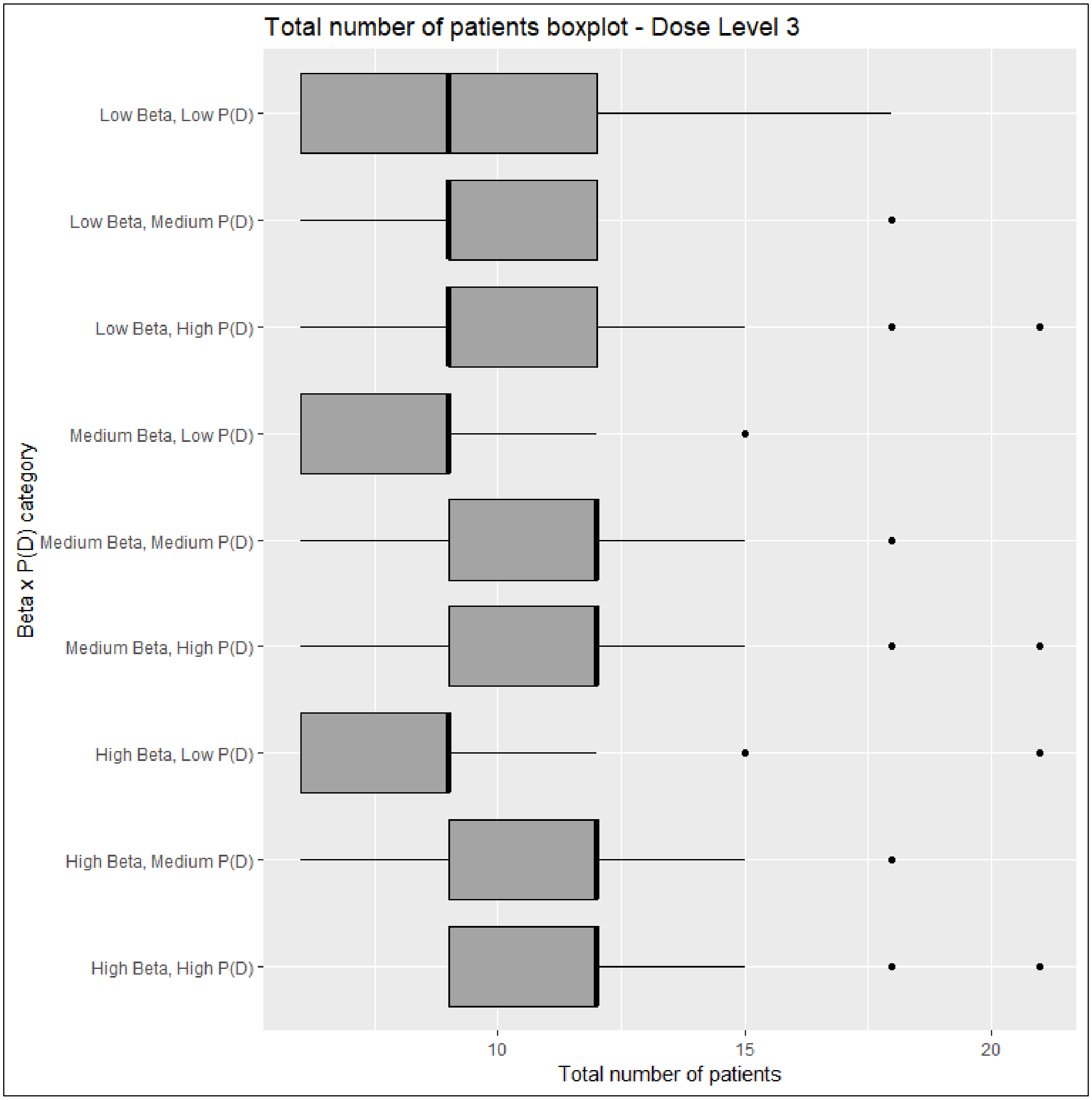
Appendix 26 - Histogram of highest dose level tested categorized by MTD level for D=5



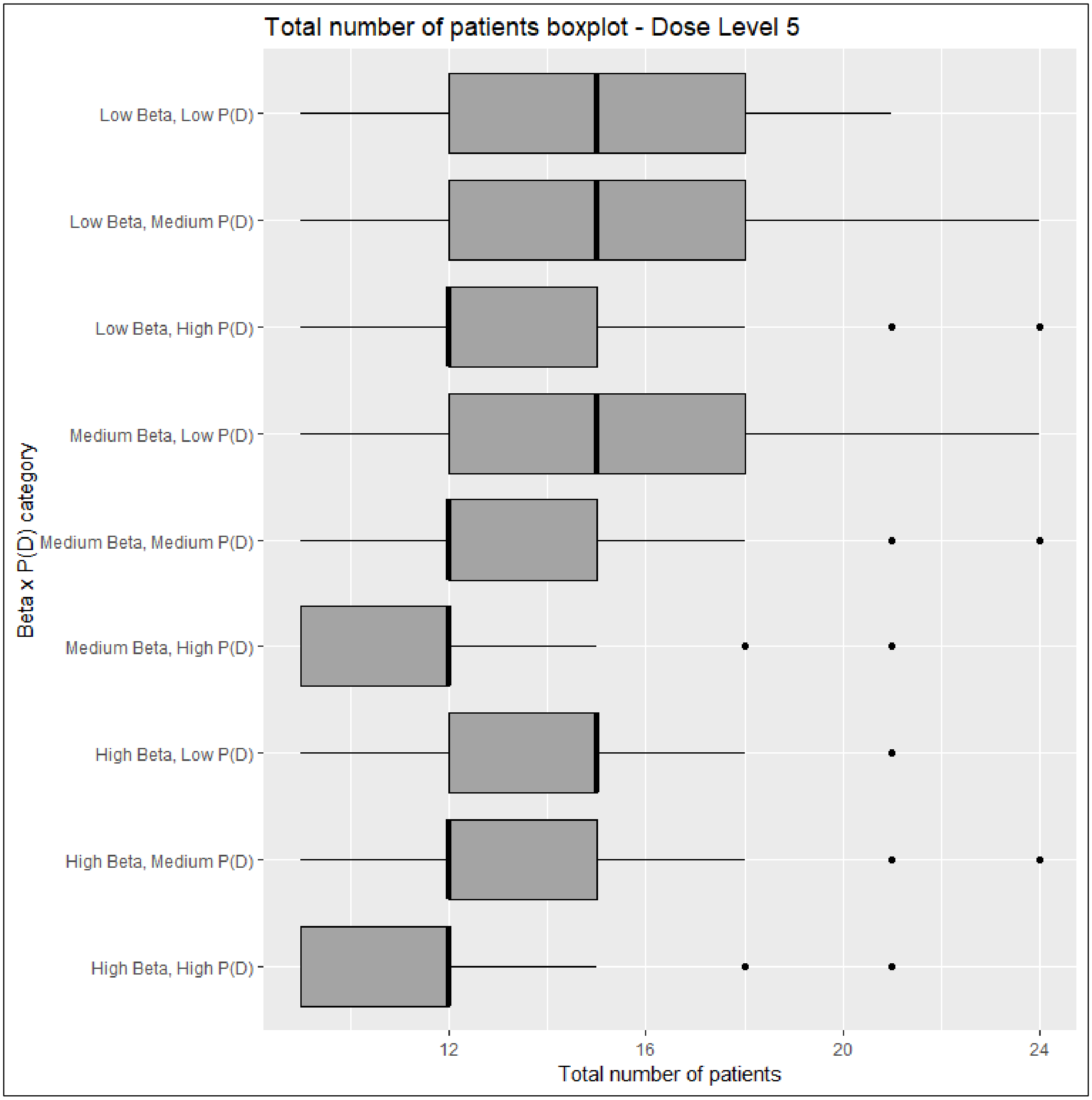
Appendix 27 - Histogram of highest dose level tested categorized by MTD level for D=7



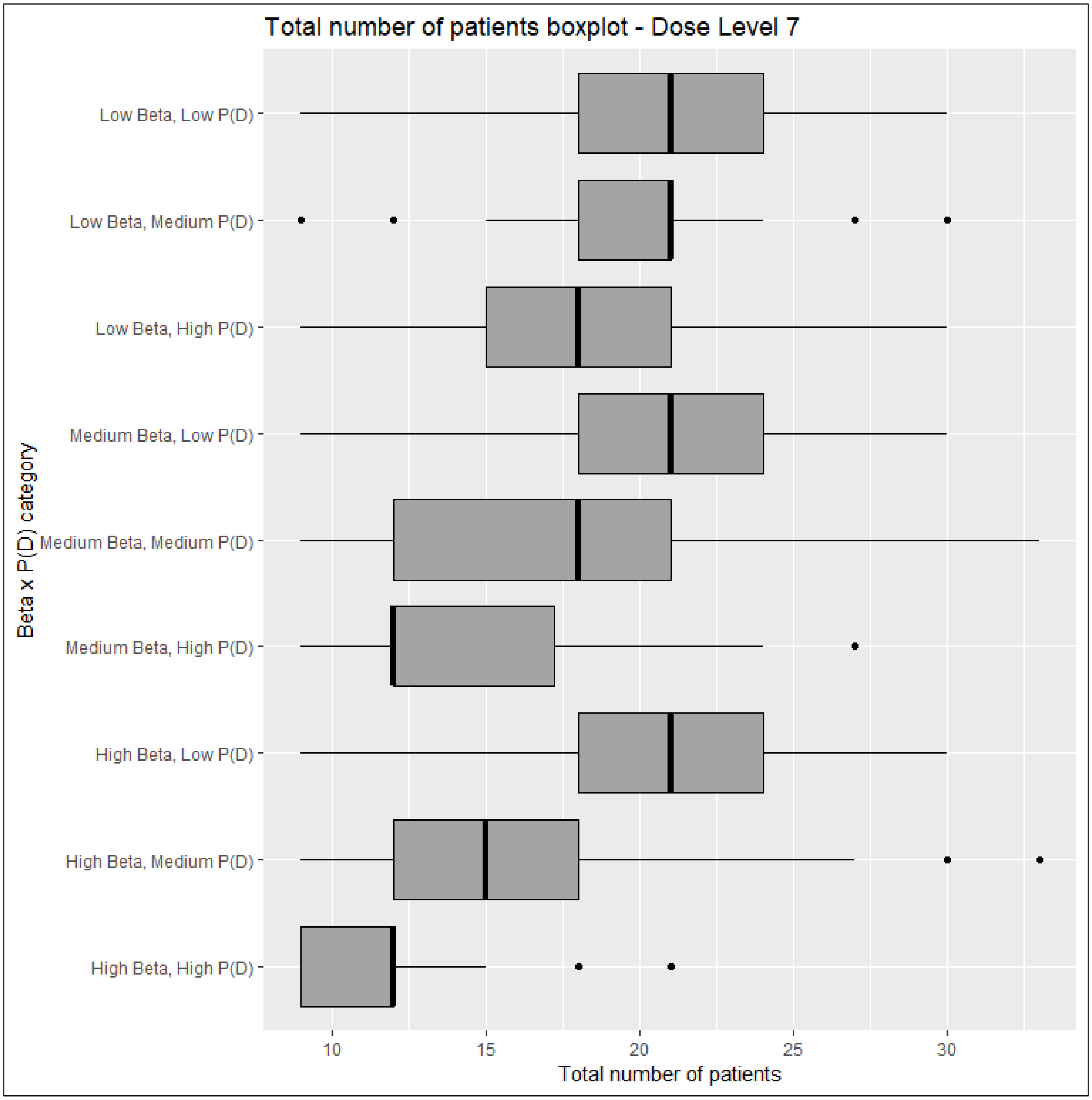
Appendix 28 - Boxplot of total number of patients by beta x P(D) categories for D=3



Appendix 29 - Boxplot of total number of patients by beta x P(D) categories for D=5



Appendix 30 - Boxplot of total number of patients by beta x P(D) categories for D=7



7 References

Hansen, A.R., Graham, D.M., Pond, G.R., Siu, L.L.: “Phase 1 Trial Design: Is 3+3 the Best?” Cancer Control, 2014, 21, (3), pp. 200–208.

Jaki, T., Clive, S., Weir, C.J.: “Principles of dose finding studies in cancer: a comparison of trial designs” Cancer Chemotherapy and Pharmacology, 2013, 71, (5), pp. 1107–1114.

Ji, Y., Wang, S.-J.: “Modified Toxicity Probability Interval Design: A Safer and More Reliable Method Than the 3+3 Design for Practical Phase I Trials” Journal of Clinical Oncology, 2013, 31, (14), pp. 1785–1791.

Morgensztern, D., LoRusso P., Boerner SA, Herbst RS, Eder JP: The Molecular Basis of Cancer (Fourth Edition), Phase I Trials Today, 2014, pp. 661-676.

Nie, L., Rubin, E.H., Mehrotra, N., et al.: “Rendering the 3+3 Design to Rest: More Efficient Approaches to Oncology Dose-Finding Trials in the Era of Targeted Therapy” Clinical Cancer Research, 2016, 22, (11), pp. 2623–2629.

Pocock, S.J.: “Clinical Trials: A Practical Approach.” Biometrics, 1984, 40, (4), p. 1211.

Quigley, J., Pepe, M., Fisher, L.: “Continual Reassessment Method: A Practical Design for Phase 1 Clinical Trials in Cancer” Biometrics, 1990, 46, (1), p. 33.

Storer, B.E.: “Design and Analysis of Phase I Clinical Trials” Biometrics, 1989, 45, (3), p. 925.

Sposto, R., Groshen, S.: “A wide-spectrum paired comparison of the properties of the Rolling 6 and 3+3 Phase I study designs” Contemporary Clinical Trials, 2011, 32, (5), pp. 694–703.