Handling data loss

In preclinical experiments it is not uncommon that not every sample or lab animal results in usable data. In some circumstances, this can be compensated by collecting more samples later. This is not always possible, e.g. because it is not straightforward to request replacement animals. The ability to compensate lost data significantly impacts the statistical design so both cases are discussed separately.

The aspect of the design that is most impacted by data loss is the power analysis, particularly if the lost samples cannot be compensated. In this case, even in the fixed sample design the expected data loss needs to be taken into account. In theory, the exact way to determine the power of the design would be to determine the power of each possible outcome and weigh them by their probability. This is a lot of work and a bit unnecessary, so in practice researchers simply estimate the drop-out rate and add enough mice to compensate the likely losses.

The same applies to group sequential designs. We could evaluate the power of all possible outcomes, however it is much more practical to add enough mice to compensate the likely loss of power. This can be done by looking at one likely 'bad' outcome, by estimating the likely losses and by assuming the losses occur in a way that causes an unbalanced design.

A factor that will determine the possible likely bad outcomes, is whether or not there is a limit on the sample size within a relevant time period or per batch. This will generally reduce our freedom to perform interim analyses whenever we like, as opposed to if data can be collected sequentially and at a reasonable speed.

The application will be illustrated on modified versions of the toy example. Unless stated otherwise, conditions, assumptions and input in GSDesigner are the same as in the original toy example. For convenience and to enable direct comparisons, we use the same error spending functions as in the main article. A complete tutorial for the toy example in GSDesigner can be found in 'S_2 Tutorial' on github.com/ICDS-vubUZ/gsd-dash-app.

Data loss can be compensated

No limit on sample size per batch/time period

This situation is by far the simplest since it does not impact the design. It is possible to simply continue collecting data until the target for the next interim analysis has been reached. No special actions need to be taken relative to the case where we do not expect data loss.

Limited sample size per batch/time period

In this case, the maximum possible sample size is still fixed, but the exact sample size at each interim analysis is uncertain. This can be handled by adapting the information ratio t according to the obtained sample size.[1] If we had planned to obtain a third of the maximum possible sample size, yet we only obtain a

[1] In clinical trials, the information ratio is sometimes not adapted to the obtained sample size because then the critical values do not need to be updated. This is not the case at small sample sizes. Even if the error spent is the same, the degrees of freedom of the t-distribution are changed and thus the critical values as well.

quarter, t is now 0.25 rather than the originally planned $t = 0.333$. This ensures that batches containing more data are given more weight than batches with a smaller sample size.

Since there is a limit on the sample size one can obtain per analysis, the most natural implementation is that the lost data is compensated in the last batch until that one is at capacity and a new batch needs to be created. Other possibilities exist however, so when and where the lost data gets compensated is a rule that should be decided before the start of the experiment.

Under these circumstances it is also possible that the number of interim analyses needs to be adapted. For example, we might want to add an analysis in case we need to create extra batches or if we need to split the extra data collection over several days. However, adding interim analyses usually decreases the power, so a rule on a maximum number of analyses is not out of place. It is also wise to include a rule to skip an analysis in case too few data points were collected. After all, if there is little information added, there is little added benefit to perform an interim analysis and the alpha and/or beta are better spent elsewhere.

Illustration on the toy example

In this version of the toy example, we assume we can only process maximum 6 mice per day, but if we lose mice along the way we can replace them later. We try to process the maximum number of mice each day either until we reach a conclusion through interim analysis or until the maximum possible sample size has been reached. From our experience with previous experiments, we estimate that on average one in ten mice does not produce usable data.

We have chosen to use both alpha and beta spending and we define the following rules for the design:

1. The information ratio is defined as the ratio between the obtained sample size and the maximum possible sample size.
2. In general, an analysis will be performed at the end of each day.
3. Mice added to compensate losses are added to the last analysis.
4. Skip an analysis if only 3 or less data points were collected since the previous analysis or if one group has a sample size of less than 3 in total.
5. Split the last analysis if data collection requires more than one day, unless this would cause the last analysis to be performed with only 3 data points added.

In this case the first rule might be a bit superfluous as this is the natural way to define the information ratio under these circumstances. However, it is different in other data loss examples, so we have chosen to include it for clarity. The reasoning behind the third rule is fairly straight-forward as well, since we cannot add more mice at any earlier days since those are already at maximum capacity.

The timing of the interim analyses should be defined, which is why the second rule exists. For the fourth rule, we have determined that 3 data points do not add enough information to warrant their own analysis. This corresponds to requiring more than half of the target added data points. This is a reasonable, but arbitrary choice and other fractions could have been chosen. The fifth rule was written to give us the

opportunity to perform an analysis at the end of each day, i.e. after each collected batch of data, as long as this is consistent with the previous rule's reasoning.

The reason for the last rule is that adding a fourth analysis would significantly reduce the power of the statistical design. This is illustrated in table 5, where the properties for a possible outcome without the last rule are shown. Note that this outcome is very unlikely if we only expect to lose one in ten mice, but could easily occur if higher levels of data loss can be expected.

Table 1 Possible scenarios if 4 analyses were allowed

Total sample size at analysis				**Information ratio *t* at analysis**				**Error spending function**	**Power**
1st	2nd	3rd	Final	1st	2nd	3rd	Final		
3 + 3	6 + 4	9 + 5	9 + 9	0.333	0.556	0.777	1	Pocock	0.676
								Compromise	0.704

Without rule 5, in a design with a maximum total sample size of 18 mice, it would in theory be possible to lose 2 data points in both the 2nd and 3rd analysis. These need to be compensated in a 4th analysis. Note that this would significantly lower the power of the design.

As with the original toy example, in the first phase of the power analysis we will assume no larger sample size is needed than for the fixed sample design. With N = 16 and an average loss of 1 in 10 mice, we can expect to lose 1 or 2 mice during the experiment. Given that we have already limited the maximum number of interim analyses, there are two remaining factors that can negatively affect the power of the design: unbalanced testing and spending more alpha earlier in the design.

A possible 'bad' outcome would be if the 2 mice are lost in the same group in the 2nd analysis, since that would result in an unbalanced test. If no mice are lost or the losses only occur in the last batch, then the last interim analysis has fewer or the same number of mice as the other analyses. Relative to other scenarios, this means more error is spent earlier and this situation can be expected to be less powerful.

Sample size per group per analysis

Analysis 1	Analysis 2	Analysis 3
3	6	8
3	4	8

Figure 1:Initial changed GSDesigner input for the toy example with lost data compensation and limited sample size per batch/ time period

We will therefore check this scenario, assuming that other likely outcomes have more power or a similar power level. The only input that has changed relative to the original toy example is the sample size per analysis.

The scenario where we have no data loss is exactly the same as for the original toy example. Here two of the error spending functions were underpowered at 8 mice per group. Raising the sample size to 9 mice per group for those two designs solved that issue. When we look at the power of our 'bad' outcome with a maximum total sample size of 8 mice per group, we can see in Table 2 that only one of our error spending functions is underpowered and once again adding one extra mouse per group solves the issue. To have sufficient power we therefore need only 8 mice per group if we choose the spending function mimicking the O'Brien-Fleming bounds, whereas we need 9 per group when using one of the other two spending functions.

Table 2 Power analysis of the toy example with data loss assuming compensation and with limited sample size per analysis

Total sample size at analysis			Information ratio t per analysis			Error spending function	Power
1st	2nd	Final	1st	2nd	Final		
3 + 3	6 + 4	8 + 8	0.375	0.625	1	OBF	0,8243
						Pocock	0,7788
						Compromise	0,8016
3 + 3	6 + 4	9 + 9	0.3333	0.5556	1	Pocock	0,8366
						Compromise	0,8537

The alpha and beta spending functions are the same as in table 3.

Data loss cannot be compensated

As opposed to the previous case, the total sample size at the final analysis is no longer certain. This means the definition of the information ratio is now no longer as clear as before. One option could be to use the fraction of processed samples or lab animals, regardless of whether the measurements were a success. However, this could cause similar weights to be attached to very different sample sizes. This usually does not benefit the power nor the expected sample size and hence is not recommended.

Another option is to divide the obtained measurements (N_i) by the maximum possible total sample size at that time. So if M_i is the number of maximum measurement attempts still available at interim analysis i, then the information ratio is $t_i = N_i/(N_i + M_i)$. This still does not guarantee a balanced weighing of error spent relative to sample size, especially if the drop-out rate is high, but it skews it in such a way that optimizes power. This comes at the cost of having higher expected sample sizes.

The third and most sensible option is to divide the obtained mice by the expected total possible sample size, i.e. if r is the drop-out rate then $t_i = N_i/(N_i + M_i)$. This gives us the best approximation of balancing the information ratio to the sample size. This has the benefit that when the obtained sample size is larger than expected, the alpha and/or beta spent are increased and we have a better chance of stopping early. When the obtained sample size is lower than expected, this causes the error spending to be shifted to a later analysis, thereby counteracting potential power loss. The weakness of this method is the estimate of the drop-out rate. If this is based on experience of similar models and methods, then this definition of the information ratio is the better choice. On the other hand, if the experiments involve procedures the researchers have not performed before and the value of r is a wild guess, then the definition of t_i as the ratio of obtained measurements relative to maximum possible sample size is likely a more cautious choice.

Limited sample size per analysis

This is fairly analogous to the subsection with compensation of lost data with a sample size limit per analysis. We still need rules about when to skip or move interim analyses, but we no longer need a rule as to when to add an interim analysis. Additionally, we likely need a larger sample size to compensate possible losses.

Illustration on the toy example

In this version of the toy example we assume we can only process maximum 6 mice per day and we cannot replace any of the lost mice. We try to process the maximum number of mice each day either until we reach a conclusion through interim analysis or until all available mice have been processed. From our experience with previous experiments, we estimate that on average one in ten mice does not produce usable data.

We have chosen to use both alpha and beta spending and we define the following rules to design:

1. The information ratio is defined as the ratio between the obtained sample size and the expected total sample size if we continue to the last possible analysis.
2. In general, an analysis will be performed at the end of each day.
3. Skip an interim analysis if only 3 or less data points were collected since the previous analysis or if one group has a total sample size of less than 3.

Since we do not know exactly how many mice we will lose during the experiment, we now need a more flexible definition of the information ratio. Since we have some experience with similar experiments, we have opted for this version. The other rules also featured in the previous example and the reasoning is the same.

Just as before, the simplest way to find the appropriate sample size, is to start at the sample size required for the fixed sample design and increase it as needed. Without data loss, the total required sample size was 16 mice. Taking into account the drop-out rate of $r = 1/10$, the fixed sample design requires a sample size of

$$N = 16/(1 - r) \approx 18.$$

Sample size per group per analysis

Analysis 1	Analysis 2	Analysis 3
3	6	9
3	4	7

Total costs at analysis

Default option: the costs equal the sample sizes

Analysis 1	Analysis 2	Analysis 3
6	12	18

Figure 2: Changed input on the 'Interim analyses' tab for the toy example without lost data compensation and with a sample size limit per batch/ time period

To estimate the power for the designs with interim analyses, we need to determine a likely 'bad' outcome, i.e. one with low power. With a sample size of 18 and an estimated loss of 10%, we will likely lose 1 or 2 mice. We will assume the data is lost early on and all in the same group, such that a maximum number of analyses will be unbalanced. If the data were lost in the first batch, we would simply skip the analysis on the first day, therefore we assume we lose two mice on the second day.

Information ratio

Use the default option: the sample size divided by the sample size at the final analysis

Analysis 1	Analysis 2	Analysis 3
0.3571	0.6494	1

Figure 3: Changed input on the 'Error spending' tab for the 2nd adapted toy example. For the calculation of the information ratios see Table 3

The input that changes relative to the original toy example is in this case the sample sizes and the information ratio. Since we no longer use the default definition of the information ratio, we need to uncheck the box and manually enter the information ratio per analysis. If we wish to get an accurate estimate of the expected number of mice we will

use, we should also uncheck the default option for the costs and add the mice that did not result in useable data to the 'Costs' input table. Because even though we did not obtain data from them, the mice were sacrificed all the same and therefore should still be counted.

For the various spending functions, the results can be seen in Table 3. As before, the error spending function mimicking the O'Brien-Fleming bounds is once again sufficiently powered, whereas the other two are not. This similarity to the other examples is not very surprising, since most of the properties are the same. It should therefore also not come as a surprise that adding one extra mouse per group results in a sufficient sample size, provided we specify that we do not add an extra analysis.

Table 3 Power analysis of the toy example with data loss assuming no compensation and with limited sample size per analysis

Total sample size at analysis			**Definition of the information ratio**	**Information ratio *t* at analysis**			**Error spending function**	**Power**
1st	2nd	Final		1st	2nd	Final		
3 + 3	6 + 4	9 + 7	Obtained mice/ expected total mice	6/(6+10.8) = 0.3571	10/(10+5.4) = 0.6494	16/(16+0) = 1	OBF	0.8151
							Pocock	0.7663
							Compromise	0.7895
			Obtained mice/ maximum possible mice	6/18 = 0.3333	10/16 = 0.625	16/16 = 1	OBF	0.8181
							Pocock	0.7718
							Compromise	0.7940
3 + 3	6 + 4	10 + 8	Obtained mice/ expected total mice	6/(6+12.6) = 0.3226	10/(10+7.2) = 0.5814	18/18 = 1	Pocock	0.8230
							Compromise	0.8411
			Obtained mice/ maximum possible mice	6/20 = 0.3	10/18 = 0.5556	18/18 = 1	Pocock	0.8274
							Compromise	0.8449

The alpha and beta spending functions are the same as in table 3. The table contains two different definitions of the information ratio. If for analysis i we use N_i for the obtained mice, U_i for the unprocessed mice and r for the estimated data loss ratio, then the first definition is $N_i/(N_i - (1 - r)U_i)$ and the second is $N_i/(N_i + U_i)$

The same is true in case the information ratio is defined as the ratio of the obtained sample size and the maximum possible size. The power for the toy example with this particular definition can also be found in table 7. This version of the information ratio has the advantage of achieving higher power, particularly if the data loss is severe. However, it has a lower probability of stopping early and saving mice and other resources, especially if the data loss was lower than expected.

No limit on sample size per analysis

If there is no sample size limit per analysis, there are more options to choose when to perform the analyses. One option is to perform analyses when a fixed number of measurement attempts have been made, e.g. every time when 6 mice have been processed. This is in essence equivalent to working with batches and the required rules and considerations are the same as in the previous subsection

Another option is to perform an analysis after a certain number of successful measurements were made, i.e. after a certain sample size has been achieved. In this case the decreased sample size would be concentrated at the final analysis, which has the benefit that all tests are performed with a balanced sample size except perhaps the final analyses. The downside is that it makes the functioning of the information ratio counterproductive. When working with batches, the information ratio works to counter the downside of data loss or to exploit the lack thereof. For this approach it does the exact opposite. It aggravates the lack of power at lower sample sizes and decreases the chance of stopping early at higher sample sizes. This drawback can be countered by choosing static values per analysis or by choosing a definition that subverts the general idea behind the information ratio. However, this makes it counter intuitive and difficult to interpret.

Illustration on the toy example

In this version of the toy example, we assume there is no real problem to switch between collecting data and analysing the data, but we cannot replace any of the lost mice. From our experience with previous experiments, we estimate that on average one in ten mice does not produce usable data.

We have chosen to use both alpha and beta spending and we define the following rules for the design:

1. Maximum 3 analyses will be performed: the first after each group has a sample size of 3, the second when each group has a sample size of 6 and the third after all remaining mice have been processed.
2. The second analysis is skipped if we expect the last analysis will contribute less than 3 data points.
3. For the information ratio, we will use the expected values prior to the beginning of the experiment. It is defined as the expected sample size per analysis divided by the expected sample size at the third analysis.

The first rule is necessary to define the moments to perform the analyses. The second rule exists since, like in the previous two examples, we consider three data points do not merit a separate analysis if it can be avoided. In this example, the information ratio has been defined to be static, since an information ratio that evolves proportional to the distribution of the sample size, will negatively affect the properties of the design. As in the previous example, the default option for the costs and information ratio no longer apply and need to be written in the input tables manually.

With regard to power, the number of possible outcomes is limited under this set-up. Since the information ratio cannot change and since all the data loss will be concentrated in the last scenario, the possible outcomes are determined by the total data loss. The required sample size is 9 mice per group for the fixed sample design including compensation for expected losses. For the first step of the power analysis we will therefore once again start with a maximum total sample size of 18 and a likely loss of two mice. As can be seen in table 8, the same two error spending functions as before require a slightly larger maximum sample size namely 10 mice per group. With a total sample size of 20, we are still most likely to lose a total of two mice.

Table 4 Power analysis of the toy example with data loss assuming no compensation and without limited sample size per analysis

Total sample size at analysis			Information ratio *t* at analysis			Error spending function	Power
1st	2nd	Final	1st	2nd	Final		
3 + 3	6 + 6	9 + 7	6/(0.9*18) = 0.375	12/(0.9*18) = 0.75	0.9*18/(0.9*18) = 1	OBF	0.8176
						Pocock	0.7592
						Compromise	0.7842
3 + 3	6 + 6	10 + 8	6/(0.9*20) = 0.3333	12/(0.9*20) = 0.6667	0.9*20/(0.9*20) = 1	Pocock	0.8143
						Compromise	0.8344

The alpha and beta spending functions are the same as in table 3

Unanticipated situations

The correct approach for the statistical design is to anticipate plausible losses and create rules before the start of the experiment to handle the various possible outcomes. Ad hoc changes to the design, even with the best intentions, may inadvertently introduce bias and should therefore not be performed.

Nevertheless, in practice it is impossible to foresee everything and there will be occasions where circumstances require changes to be made after the experiment has begun. For this reason, we include this section on making changes while attempting to minimize the introduction of bias. Even so we wish to stress that it is in the researcher's best interest to try to avoid this situation.

First, the person changing the statistical design should not have seen the data, nor been involved with collecting the data. This person should not have access to any information that was not available prior to the beginning of the experiment other than the external factor that caused the need to change the design and, if any interim analyses have already been performed, the sample size and the error spent at those analyses. This information is necessary since we cannot change the past, and these features should therefore remain the same for the new design.

There might be valid reasons to need to change the statistical design other than unforeseen data loss, however they need to be unrelated to the statistical hypothesis. Changes to the hypothesized means and standard deviations are not allowed. In order not to introduce bias, the assumptions with regard to effect sizes need to remain the same as during the original power analysis.