The analysis of continuous data from n-of-1 trials using paired cycles: a simple tutorial

Stephen Senn stephen@senns.uk

Summary

N-of-1 trials are defined and the popular paired cycle design is introduced, together with an explanation as to how suitable sequences may be constructed.

Various approaches to analysing such trials are explained and illustrated using a simulated data set. It is explained how choosing an appropriate analysis depends on the question one wishes to answer. It is also shown that for a given question various equivalent approaches to analysis can be found, a fact which may be exploited to expand the possible software routines that may be used.

Sets of N-of-1 trials are analogous to sets of parallel group trials. This means that software for carrying out meta-analysis can be used to combine results from N-of-1 trials. In doing so, it is necessary to make one important change, however. Because degrees of freedom for estimating variances for individual subjects will be scarce, it is advisable to estimate local standard errors using pooled variances. How this may be done is explained and fixed and random effect approaches to combining results are illustrated.

1. Introduction

This paper provides a simple tutorial on analysing continuous data from n-of-1 trials using paired cycles. This design (to be described below) leads to various simple possible analyses and is an efficient way to compare two treatments on a within-patient basis where the nature of the disease and other practical considerations make this possible. The general framework that will be applied is that in which data are treated as if sampled from some hyper-population with Normally distributed values. This is the usual basis for 'parametric analysis', is a common device for modelling data and is known to yield (usually) similar results to an alternative framework in which the treated units are regarded as being fixed but the population is that of all possible random allocations[1, 2]. However, this correspondence works best when the sample size is large and this is often not the case when series of n-of-1 trials are being discussed. This reservation should be noted and in particular if the parametric analysis yields highly significant results, it may be the case that a randomisation test would be incapable, of yielding similar results. This is not necessarily a reason for abandoning the parametric approach. In data-poor contexts, such as often apply for the study of rare diseases, accepting the necessary assumptions may be the lesser of two evils. Nevertheless, the limitation should be born in mind.

The objectives of the tutorial are to provide simple justifications and instructions for various possible analyses of such trials and also explain for which purposes they are suited. Use of algebra is kept to a minimum and graphical and tabular representation of data and analyses are stressed.

For readers who require more technical detail, a general model for data from N-of-trials is presented and discussed in an appendix. It is explained how the way in which the overall treatment effect is regarded, either as a mean effect for the subjects studied or as a mean effect of the hypothetical

population of subjects of whom they might be considered to be a random sample, will effect the way that analysis proceeds.

2. The design

A common design for n-of-1 trials comparing two treatments is to organise allocation in such a way that within any given pair of periods, each treatment is used once[3-5]. Such pairs of periods have been referred to as *cycles*[6]. A possible scheme for a design in three cycles is given in Table 1 . Patients would then be allocated at random to one of the eight possible sequences.

	Periods					
	1	2	3	4	5	6
Sequence	Cycle 1		Cycle2		Cycle 3	
1	A	B	A	B	A	B
2	B	A	A	B	A	B
3	A	B	B	A	A	B
4	B	A	B	A	A	B
5	A	B	A	B	B	A
6	B	A	A	B	B	A
7	A	B	B	A	B	A
8	B	A	B	A	B	A

Table 1 Set of sequences for a design using six periods arranged in three cycles. Pairs with A followed by B are shaded yellow. Pairs with B followed by A are shaded blue.

In general, if there are *k* possible cycles in which patients can be treated, there will be 2^k possible sequences. A *canonical set* of possible sequences can be constructed as follows using the basic pair AB and BA. When moving between successive sequences in a list of sequences, for cycle 1 switch AB and BA after every sequence. For cycle 2, double the number of sequences before switching. For each successive cycle, double the sequences before switching.

This design, is relatively simple to organise, efficient and lends itself to various simple analyses. As regards organisation, a simple way to implement randomisation to sequences is just to randomise patients independently for each cycle. As regards the second, the close temporal control that is offered by randomising in pairs makes it efficient. It could be argued that if carry-over is likely and one wishes to guard against it, various other designs might be preferable, but the solutions these offer depend on implausible modelling assumptions and the best advice as regards carry-over is to ensure adequate washout between treatments, if necessary limiting measurement of the effect of each treatment towards the end of periods in which they are given[7]. As regards analysis, it is the purpose of this note to explain how this may be achieved. For advice on reporting n-of-1 trials see the CENT statement [8].

3. Illustrative data for analysis

N-of-1 trials lend themselves to addressing a number of different questions that might arise naturally in connection with studying the effects of treatments[9]. The questions are

Q1. Was there an effect of treatment in the trials?

Q2. What was the average effect of treatment in the trials that were run?

Q3. Was the treatment effect identical for all patients in the trials?

Q4. What was the effect for individual patients in the trials?

Q5. What will be the effect of treatment when used more generally (in future)?

The suggested analyses will be organised in terms of these questions. We shall use the simple simulated data that were presented in Araujo et al[6] to illustrate these analyses.

It is supposed that a trial in asthma has been carried out comparing two treatments, A and B, each given as a single dose. Twelve, patients have been randomised in pairs of cycles as described above. The first ten have completed all three cycles of treatment. However, patients 11 has only completed two cycles of treatment and patient 12 has only completed 1. This has been done to illustrate a complication in analysis that may arise in practice. We thus have data from $(10\times 3)+2+1=33$ cycles and therefore from $2\times 33=66$ episodes. In all the analyses that follows we shall assume that the fact that some values are missing is uninformative and that reasonable inferences may be based on the values that remain.

	Treatment					
Patient	A	B	A	B	A	B
1	1	2	3	4	6	5
	2394	2686	2515	2675	2583	2802
2	2	1	3	4	6	5
	2746	2726	2592	2867	2743	2742
3	1	2	3	4	6	5
	2668	2560	2542	2584	2491	2737
4	1	2	3	4	6	5
	2397	2696	2411	2895	2499	2760
5	2	1	3	4	5	6
	3179	3221	2952	3096	2600	3192
6	1	2	4	3	5	6
	2643	2496	2759	2847	2651	2860
7	1	2	3	4	5	6
	2678	2843	2492	2763	2801	2890
8	2	1	3	4	5	6
	2887	2862	2875	3083	2689	2967
9	2	1	3	4	6	5
	2490	2841	2648	3044	2688	2914
10	2	1	3	4	6	5
	2268	2576	2413	2493	2344	2699
11	2	1	4	3	6	5
	2617	2923	2629	2832		
12	1	2	4	3	5	6
	2627	2759				

Table 2 A simulated trial in asthma. 12 Patients have been randomised in three cycles to treatment A followed by B or B followed by B. The table gives the periods in which the patients received A or B and the FEV_1 in mL below. For example, Patient 1 received treatment A in periods 1, 3 and 6 and treatment B in periods 2,4 and 5.

The results are measurements of forced expiratory volume in one second, FEV_1, in mL taken 12 hours after treatment. The data are presented in Table 2 sorted by treatment within cycle (that is to say A followed by B). The period in which A or B was administered is given within Table 2 and this reflects the randomisation used. The data are also available to download from https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0167167#sec009 . Note, however, that those data include values for cycles 3 from patients 11 and 12 and cycle 2 from patient 12, which are assumed missing here.

A useful plot of the data is given in Figure 1, which is a *trellis plot*. Each window represents the results for a given patient. The result for each cycle is represented by a blue circle plotting the value under B (Y axis) against that under A (X axis). The diagonal line represents equality between the two treatments. The average values over all cycles are represented by red asterisks. It is noticeable that the blue circles are generally above and to the left of the line of equality suggesting that B has a bigger effect than A.

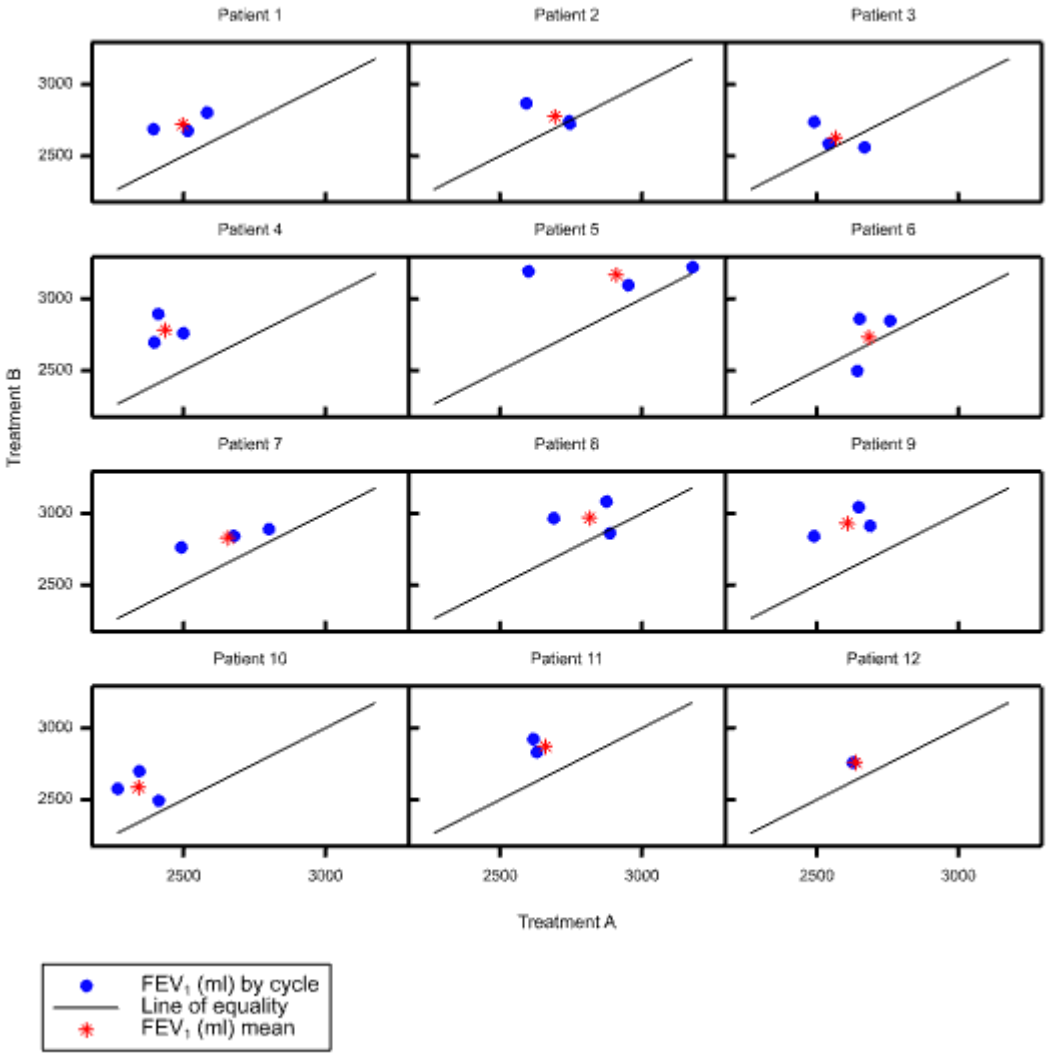

Figure 1 Trellis plot of the results from the simulated example.

4. Demonstrating that there can be a difference between treatments

This limited question leads to a very simple analysis. The relevant null hypothesis is that there is no difference between treatments *for any of the patients*. If that is the case, under the null hypothesis, it does not matter which patient is studied, the result may be expected to be the same. This renders the *differences* between A and B as being independent over patients *by hypothesis*. That being so, we can carry out a matched pair analysis on the 33 cycles.

Cycle	1	2	3
Patient			
1	292.0	160.0	219.0
2	-20.0	275.0	-1.0
3	-108.0	42.0	246.0
4	299.0	484.0	261.0
5	42.0	144.0	592.0
6	-147.0	88.0	209.0
7	165.0	271.0	89.0
8	-25.0	208.0	278.0
9	351.0	396.0	226.0
10	308.0	80.0	355.0
11	306.0	203.0	*
12	132.0	*	*

Table 3 Differences (Treatment B – treatment A) per cycle arranged by patient.

The data have been reduced to differences by cycle and patient and are presented in Table 3. These differences can be analysed by a one-sample t-test for which the statistics in Table 4 are produced.

Statistic	**Value**	**Explanation**
n	33	Number of cycles
Mean	194.55 mL	Mean of the 33 cycles
Variance	26188 mL^2	Sample variance of the 33 cycles
SD	161.8 mL	Standard deviation = $\sqrt{}$Variance
SE	28.17 mL	Standard error = SD/$\sqrt{}$n
DF	32	Degrees of freedom = *n*-1
t	6.91	t-statistic =194.55mL/28.17mL
P-value	<0.001	Probability under H_0 a t-statistic with 32 DF will be $\geq$ 6.91 or $\leq$-6.91

Table 4 Summary statistics to perform a one-sample t-test based on differences per cycle.

5. Putting bounds on the mean effect for the patients studied

The critical value at the 5% level for a t-statistic with 32 degrees of freedom is 2.037. If this is multiplied by the standard error the product is $2.037 \times 28.17mL = 57.5mL$. If this is subtracted and added to the mean of 194.5mL, then we obtain a 95% confidence interval for the mean effect of (137.2mL, 251.9mL).

This particular calculation can be criticised. Whereas it is reasonable to assume by hypothesis that the treatment effect is constant for all patients when we are testing that this effect is zero for them

all, as soon as we allow that the effect is *not* zero, it becomes plausible that it might vary from patient to patient. If we regard the patients as being fixed, that is to say that we are only making a statement about these patients, then we could claim that this source of variation would not contribute to the treatment estimate changing were we to repeat the experiment. However it will contribute to the overall estimate of variation that we have used.

This source of variation can be eliminated by constructing variance estimates patient by patient. The calculations are given in Table 5;

Patient	DF	Variance	Sum of Squares
1	2	4372.3	8744.7
2	2	27260.3	54520.7
3	2	31572.0	63144.0
4	2	14233.0	28466.0
5	2	85601.3	171202.7
6	2	32767.0	65534.0
7	2	8356.0	16712.0
8	2	25166.3	50332.7
9	2	7758.3	15516.7
10	2	21636,3	43272.7
11	1	5304.5	5304.5
12	0	0.0	0.0
Total	21		522750.5

Table 5 Intermediate calculation for to estimate the common within-patient variance. Note that the units of variances and sums of squares are mL^2 of FEV_1.

Here the column labelled **DF** gives the degrees of freedom patient by patient and is equal to the number of cycles minus 1. The column labelled **Variance** gives the *local* estimate of the variance of the differences (B-A) patient by patient. For patient 12 the value is zero since the patient was only studied in one cycle and hence there is only one difference. The column headed Sum of Squares is obtained by multiplying the variance by the degrees of freedom. The overall sum of squares is 522,750.5 mL^2 and if this is divided by the total DF, 21, we obtain 24,893 mL^2, which is thus our estimate of the variance on the assumption that variability does not vary from patient to patient.

The consequent calculations are summarised in Table 6

Statistic	Value	Explanation
n	33	Number of cycles
Mean	194.55 mL	Mean of the 33 cycles
Variance	24893 mL^2	Sample variance of the 33 cycles
SD	157.8 mL	Standard deviation = $\sqrt{}$Variance
SE	27.47 mL	Standard error = SD/$\sqrt{}$n
DF	21	Degrees of freedom = n-1-11
t	7.08	t-statistic =194.55mL/27.47mL
P-value	<0.001	Probability under H_0 a t-statistic with 21 DF will be $\geq$ 7.08 or $\leq$--7.08

Table 6 Summary statistics to perform a one-sample t-test based on differences per cycle with the patient by treatment interaction removed from the variance estimate.

The other criticism will tend in the other direction. If we do not regard the patients as fixed then we have not reflected the variation from patient to patient enough, since our estimate is based on using

cycles as the unit of inference rather than patients. We now consider an analysis that uses patients as the unit of inference.

6. Putting more general bounds on the treatment effect

One way of proceeding is to reduce the differences to a mean per patient and then perform an analysis using these 12 means differences as our raw input. The data are presented in Table 7. We shall ignore the column labelled standard error for the moment. (We shall use this later.) Instead, we just base our analysis on the 12 *per patient estimates*.

Per patient estimate	Standard error
223.7	91.09
84.7	91.09
60.0	91.09
348.0	91.09
259.3	91.09
50.0	91.09
175.0	91.09
153.7	91.09
324.3	91.09
247.7	91.09
254.5	111.56
132.0	157.77

Table 7 Summary statistics per patient that may be used for various analyses

If we carry out a one-sample t analysis on these values, we can summarise the results as in Table 8

Statistic	**Value**	**Explanation**
n	12	Number of patients
Mean	192.74 mL	Mean of the 12 patient means
Variance	9895 mL^2	Sample variance of the 33 cycles
SD	99.48 mL	Standard deviation = √Variance
SE	28.72 mL	Standard error = SD/√n
DF	11	Degrees of freedom = *n*-1
t	6.71	t-statistic =192.7mL/28.72mL
P-value	<0.001	Probability under H_0 a t-statistic with 32 DF will be ≥ 6.71 or ≤-6.71

Table 8 Summary statistics to perform a one-sample t-test based on differences per patient.

The end result is very similar to that reached before. It is not surprising that the mean is scarcely different. The fact that the standard error is similar, however, reflects the fact that *for this particular example* the variation in effect from patient to patient over and above that to be expected by the random variation from cycle to cycle is small. Nevertheless, the analysis is conceptually different to that previously provided as it is relevant to a different question: *what can one say about the mean effect in general, not just for patients studied*. We shall revisit this question in section 7 below.

There are now 11 degrees of freedom and the critical value for the t-statistic is now slightly larger at 2.201. We thus have $2.201 \times 28.72\ mL = 63.2mL$ as the value that has to be subtracted from and added to the mean to get lower and upper 95% confidence limits. The resulting 95% confidence interval is (129.5, 255.9).

7. Meta-analytic approaches

A set of n-of-1 trials such as we have been considering is analogous to a collection of results from independent clinical trials, such as might be summarised in a meta-analysis. There is an extensive theory of how such results should be analysed and software routines exist within all the major statistical packages that may be used to perform a meta-analysis, This means that tools are available that may be simply adapted to perform the analysis of a set-of n-of-I trials.

There is one important change in data-preparation that is, however, necessary. Standard meta-analytic approaches assume that the standard errors used to calculate the weights are themselves calculated without error. This is, of course, not true. Estimated standard errors are random variables, not known parameters. However, if the associated degrees of freedom are reasonably large, this assumption does not matter. For n-of-1 trials, however, the degrees of freedom are typically small. In our example, there are no more than two per patient. Naively estimating the variances independently is unwise[10]. It is better to use a pooled variance to do so.

Thus, we impose an assumption that the within-patient *variation* between estimates per cycle is constant across patients. We then proceed to estimate the variance.

For this purpose we can use the approach illustrated in Table 5 and Table 6 above. For each patient the degrees of freedom are calculated as the number of cycles in which they were treated minus one. The values are shown in column two of Table 5. The sample variance of the estimated treatment effect for each patient is calculated and given in column three. The product of the values in columns two and three gives the sums of squares (corrected by the mean), which is shown in column four. (If the available statistical software package has a standard function available for the *corrected sum of squares*, it may be easier simply to calculate column four directly.) The sum of the values in column four is 522750.5 mL^2. Dividing the total sum of squares by the total degrees of freedom, 21, yields an estimated variance of 24892.9 mL^2 and the square root of this is 157.77 mL

Note that since patient 12 was only treated in one cycle, it is impossible anyway to estimate a variance for them. However, using the data from other patients we assume that the estimated standard deviation for them is the same as for all patients and is thus 157.8 mL. Since the estimate for this patient is only based on one cycle, the standard error for them is the same as the standard deviation since, trivially, $157.77\ mL/\sqrt{1} = 157.77\ mL$. In general if a patient was treated in k cycles we have $SE = s/\sqrt{k}$ where *s* is the estimated pooled standard deviation ($157.77\ mL$ for this example) . For patient 11 we have *k = 2* and for all other patients *k = 3*. Substituting these values of *k* yields the standard errors given in Table 7.

We can now apply standard meta-analytic approaches to the data in Table 7. There is a wide choice of packages to do this. Here we illustrate the analysis using the **meta** package of Guido Schwarzer's[11]. The results of using the **metagen( )** and **forest( )** functions are displayed in Figure 2

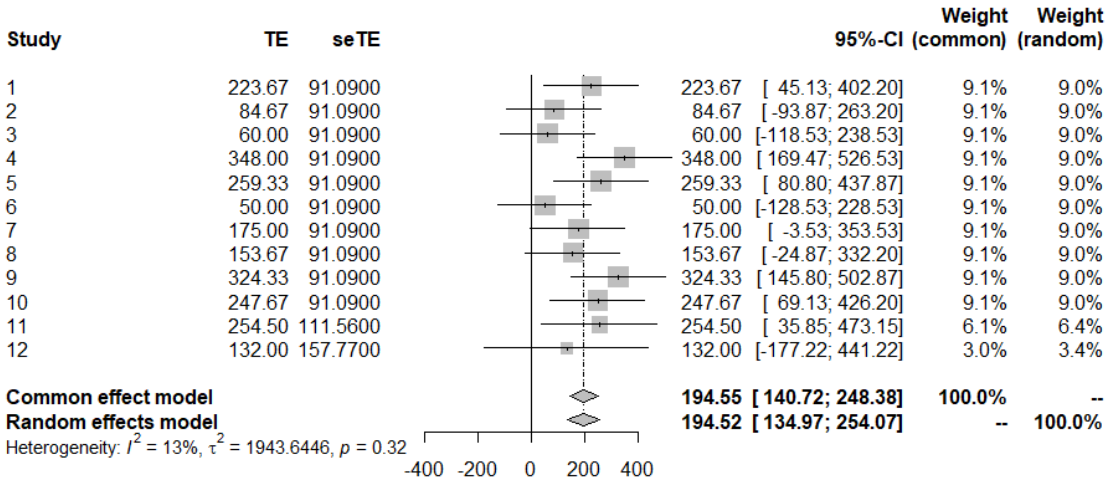

*Figure 2 Results of analysis using the **meta** package.*

This provides both a fixed and a random effects analysis. For this example, the results are very similar. Furthermore, the point estimate of 194.55 is identical to that reached for the matched pairs analysis of the 33 cycles. This is no coincidence. Since the standard errors patient by patient have been calculated using the same variance, the difference between them merely reflects the numbers of cycles for which information was obtained. The **metagen( )** function is a *generic inverse variance meta-analysis* function. It weights results proportionately to the inverse of the square of the standard error, that is to say proportionately to the number of cycles.

The standard error is different however. This is based on 21 degrees of freedom rather than 32. The extent to which results vary from patient to patient have been removed from the estimate of the variance. The difference is 11 degrees of freedom and these are the degrees of freedom that correspond to the treatment-by-patient interaction. As has been pointed out elsewhere, this point is frequently misunderstood[12]. More generally, both a fixed and random effect interaction fit a treatment-by-trial interaction. (In this case, the analogy of *trial* is *patient*.) It is what they *do* with it that makes the difference.

The random effects meta-analysis estimate has a slightly wider confidence interval. This is because it provides an estimate of the treatment effect that would apply were it the case that the patients that have been studied were no longer fixed but could be regarded as a random sample from a wider but 'similar' population. Thus the terms that the interaction measures are no longer regarded as being fixed but values that might be different from one occasion to another. Thus, this uncertainty is incorporated in the confidence intervals. In favour of the random effects analysis is the fact that it addresses a more important question. Against it is the fact that strong assumptions (the similarity of patients studied with those in the target population) have to be made.

8. Estimates of effects for individual patients

It may surprise some that superior estimates of the effects from individual patients can be obtained by also using the results from others. However, a little reflection shows that using results from others is exactly what happens when data from parallel group trials provide predictions of the effect of treatments. Therefore, a series of n-of-1 trials will provide two sorts of information for a given individual, namely *personal* and *global*, the former only using a given patient's data and the latter all

the data. Each of these is an unbiased estimate of the effect for a patient and they may be combined to produce a so-called *shrunk* estimate as follows

$$shrunk = w \times personal + (1 - w)\, global\ ,$$

where *w* is a weight between 0 and 1. The greater the value of *w,* the greater attention we pay to the information from the given patient. The estimate is referred to as *shrunk* because the result will lie between personal and global and so may be regarded as having shrunk towards the latter compared to the former. An alternative term is *best linear unbiased predictor*(BLUP).

Just at we combine information from a meta-analysis by weighting the trial proportionately to the inverse of the variances of their estimates, so we weight these two sorts of information inversely according to their variances.

A plot of the shrunk estimates is provided in Figure 3, which exhibits strong shrinkage, The reason that this is so is because there is little evidence of differences in the effect of treatment from one patient to another what observed differences there are being largely due to within-patient variation, that is to say that observed effects vary randomly from cycle to cycle. For patients 1 to 10 the degree of shrinkage is the same, so that their points line on a straight line. Patients 11 and 12 are labelled because they have different (stronger) shrinkage since their results are based on two cycles and one cycle respectively rather than on three.

We do not need to go into the theory of this more deeply, it is covered, for example, in the paper by Araujo et al[6], already cited and also in Senn (2019)[13]. Fortunately, this sort of question is addressed in various meta-analytic packages. For example, the **metafor** package[14] within R has a **blup( )** function that will do this. It is, of course, necessary to have prepared the data in the way described at the beginning of this section.

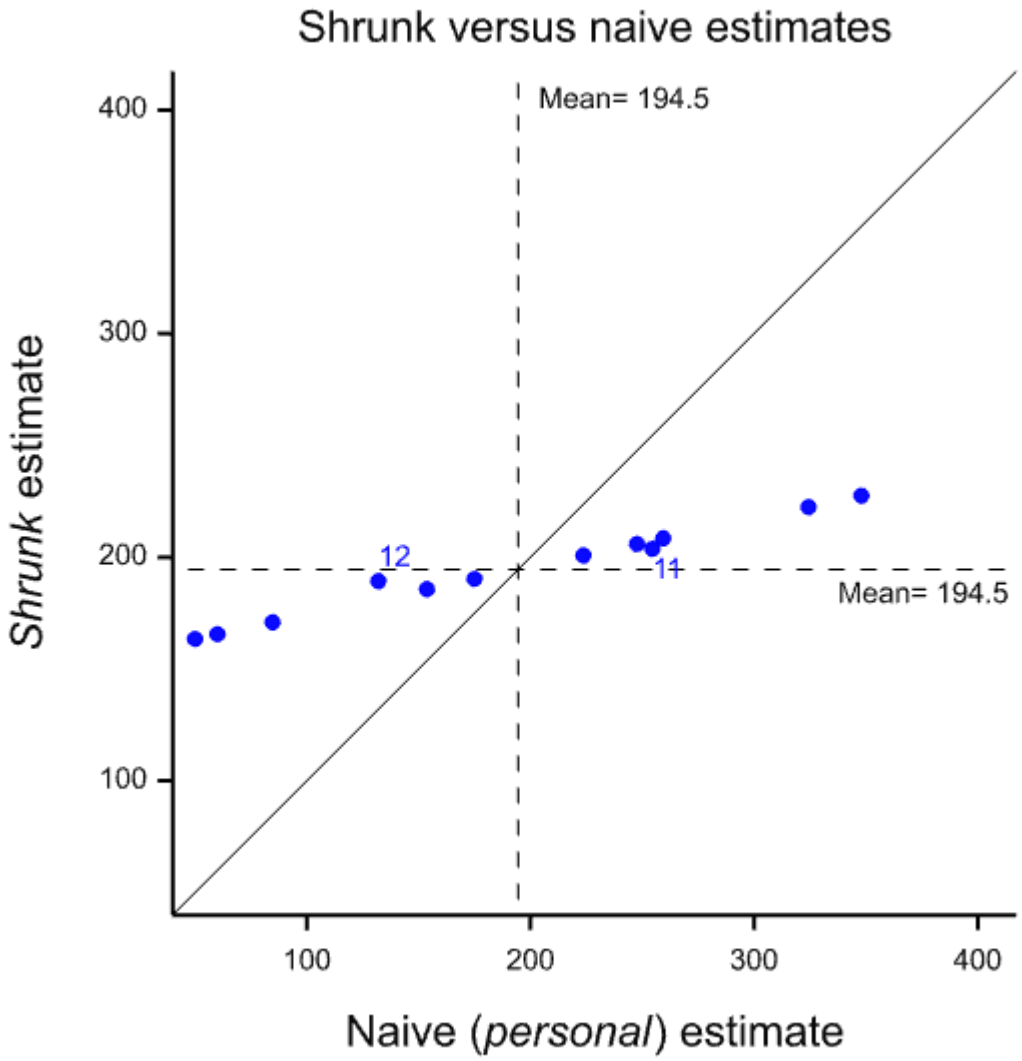

Figure 3 Shrunk estimates for FEV_1 in mL based on a weighted combination of global and personal estimates versus the naïve estimate based on personal information only. The diagonal line is the line of equality.

9. Linear mixed effect and non-linear mixed effect models

The analyses shown so far can all be regarded as special cases of so called *linear mixed effects models*. Such models provide a powerful, flexible framework for analysis but do require greater statistical skill in their handling and are not covered in this simple tutorial. For more information about them, see the paper by Araujo et al[6] and also Zucker et al(2010)[5] .

For other outcomes, for example binary outcomes, non-linear mixed effect models may be used. Their application to crossover trials is covered in Senn, (2002)[7], Senn, (2021[15]) and Jones and Kenward (2015)[16].

10. Analysis when there is only one patient

The techniques discussed so far are applicable when results can be obtained from a number of patients, which means that these results can be combined not only for the purpose of examining

average effects of treatment but also for the purpose of producing superior shrunk estimates for individual patients. It is sometimes the case, however, that the rarity of the disease or other practical difficulties mean that very few patients, and in the limit only one, can be recruited.

Given the possibility of treating the patient for many cycles, a reasonable analysis could still be carried out. However, if only a few cycles are available, a severe difficulty presents itself. Suppose that, as was the case in our simulated example, only three cycles can be used. In that case, not only will the mean effect be estimated poorly, the variance of the effect will be estimated extremely poorly, since only two degrees of freedom will be available. This is what might be called a matter of *second order efficiency*: the effect is on the estimate of variability not on the variability of the estimate. This has a catastrophic effect on calculating confidence intervals or significance. For the simulated example, by estimating the variance from all the patients, we had a variance estimate with 21 degrees of freedom. The 97.5% quantile on the t-distribution with 21 degrees of freedom is 2.080. On the other hand, with only two degrees of freedom it is 4.303, more than twice as large. Hence, other things being equal, confidence intervals for treatment effects would be more than doubled were we to use the local (to each patient) values for estimating the variance.

One possibility is to try and use an external estimate for the variance of the effect, even if it is accepted that the estimate of the effect itself must be limited to the patient. This is very much in the spirit of post-hoc ANOVA tests, where variances are often pooled across treatments even if only two of them are being compared. This habit originated in agriculture where degrees of freedom are scarce and, not always logically, is often used in multi-armed parallel group trials, pooling the variance from all treatments, even when only two are being compared, despite the fact that degrees of freedom are abundant[9, 17].

Even if a treatment is being trialled for the first time, it may be the case that the disease has been studied previously. One solution would be to use a suitable variance estimate from such studies to calculate the standard error for the n-of-1 trial. Care needs to be taken to match like with like. It has to be a within-patient variance and a trap must be avoided. The variance of the difference between two observations on a given subject is *twice* the within-subject variances as usually defined by statisticians. It might be appropriate to cap the number of degrees of freedom for such a historical variance at some relatively low number, say 10, even where many subjects have been studied, and pool accordingly with the data from the n-of-1 trial.

Such an approach is illustrated in Figure 4. It is assumed that only patient number 5 of those previously considered is being measured. However, information on variability of results is available from other historical patients. (Here the data from the remaining 11 patients has been used.) These data are combined with those from patient 5 to form a weighted variance, where the weights are the two degrees of freedom available for patient 5 and the assumed 'prior degrees of freedom' varying from 0 to 10 for the remaining patients. (Note that this is a deliberate choice and is not the same as the actual degrees of freedom used in estimating this prior variance.) The resulting 'posterior degrees of freedom' will be the sum of the two and thus vary from 2 to 12. The critical value of the t-distribution is calculated accordingly, as is the standard error and hence the confidence limits are obtained.

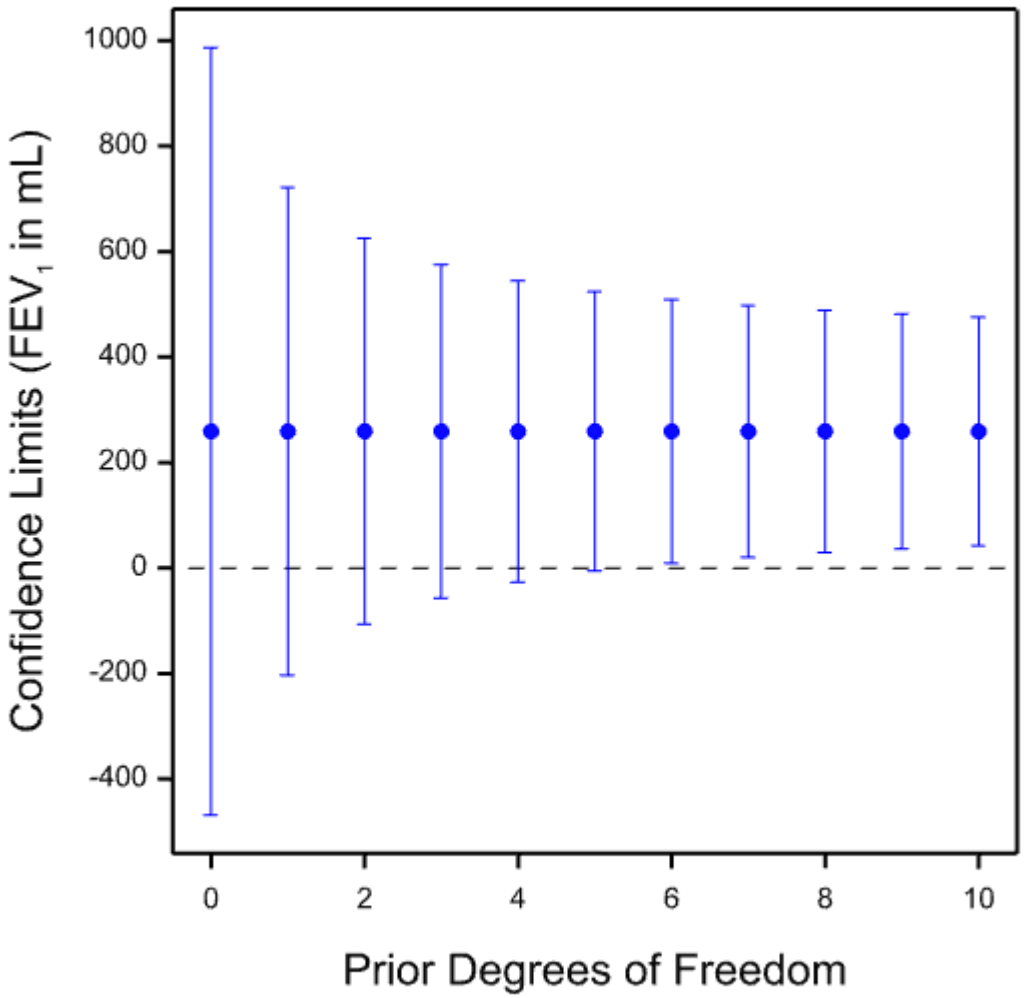

Figure 4 Illustration of technique of pooling a prior variance with the variance from a given patient (in this case patient number 5). The 95% confidence limits are shown. Information from the other patients is assumed to be available and various possible weights in terms of 'prior degrees of freedom' are considered. The point estimate is unaffected but depending on prior degrees of freedom assumed the estimated variance and the critical value of the t-distribution with change.

If the prior degrees of freedom are 0, then the result is equivalent to just using the data from patient 5. Prior information is not used to calculate the point estimate, which thus remains unchanged. The variance will change and this might increase or decrease depending on whether the variance for the patient under consideration is smaller or larger than that from the historic data. Here patient 5 had a larger than average value. Whether the variance and hence the standard error increases or reduces, the critical value of the t-distribution for calculating the 95% limits will shrink towards the asymptotic value of 1.96 that applies to the Normal distribution. For two (posterior) degrees of freedom the value is 4.30 and for 12 it is 2.18.

Of course this is all very speculative but desperate remedies may be needed when data are scarce.

11. Conclusions

N-of-1 trials encourage us to look at treatment effects at the lowest level, that of patients themselves. Of course, this is the level at which decisions are made and so, ideally, it is the level at which we should like to estimate effects of treatment. Nevertheless, random variability will still affect our estimates and combining local and global information will often lead to worthwhile improvements in precision. The scarcity of data may make some compromise as regards standards inevitable but what should not be compromised are the standards employed in explaining what has been done. Assumptions should be stated and the aim should be to make it as clear as possible what choices have been made and how they have been implemented.

It is hoped that this tutorial has succeeded in explaining how this may be done.

Acknowledgements

This research was started when working on the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant Agreement no. 602552 for the IDEAL project and completed when working on the National Institute for Health and Care Research (NIHR) DIAMOND project. Support from both is gratefully acknowledged as is collaboration with Artur Araujo, Robin Chatters, Andrew Cooke, Liv Hawksworth, Steven Julious and Sonia Leite. I am grateful to Guido Schwarzer for helpful comments on the **meta** package.

References

1. Rosenberger WF, Uschner D, Wang Y. Randomization: The forgotten component of the randomized clinical trial. *Stat Med*. Jan 15 2019;38(1):1-12. doi:10.1002/sim.7901
2. Ludbrook J, Dudley H. Why permutation tests are superior to t and F tests in biomedical research. *American Statistician*. 1998;52(2):127-132.
3. Jaeschke R, Adachi J, Guyatt G, Keller J, Wong B. Clinical usefulness of amitriptyline in fibromyalgia: the results of 23 N-of-1 randomized controlled trials. *J Rheumatol*. Mar 1991;18(3):447-51.
4. Zucker DR, Schmid CH, McIntosh MW, D'Agostino RB, Selker HP, Lau J. Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *Journal of clinical epidemiology*. Apr 1997;50(4):401-10. doi:S0895-4356(96)00429-5 [pii]
5. Zucker DR, Ruthazer R, Schmid CH. Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. *Journal of clinical epidemiology*. Dec 2010;63(12):1312-23. doi:S0895-4356(10)00194-0 [pii]

10.1016/j.jclinepi.2010.04.020
6. Araujo A, Julious S, Senn S. Understanding Variation in Sets of N-of-1 Trials. Research. *PloS one*. 2016;11(12):e0167167. doi:10.1371/journal.pone.0167167
7. Senn SJ. *Cross-over Trials in Clinical Research*. Second ed. Wiley; 2002.
8. Vohra S, Shamseer L, Sampson M, et al. CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. *BMJ*. 2015;350:h1738. doi:10.1136/bmj.h1738
9. Senn SJ. Added Values: Controversies concerning randomization and additivity in clinical trials. Research paper. *Statistics in Medicine*. Dec 6 2004;23(24):3729-3753.
10. Senn SJ. Letter to the Editor: in defence of the linear model. Letter. *Controlled clinical trials*. 2000;21:589 - 592.
11. Schwarzer G. meta: An R package for meta-analysis. *R news*. 2007;7(3):40-45.

12. Senn SJ. The many modes of meta. Research Paper. *Drug Information Journal*. 2000;34:535-549.
13. Senn SJ. Sample size considerations for n-of-1 trials. Research. *Statistical methods in medical research*. 2019;28(2):372-383. doi:10.1177/0962280217726801
14. Viechtbauer W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*. 2010;36(3):1-48.
15. Senn SJ. *Statistical Issues in Drug Development*. 3rd ed. vol 69. John Wiley & Sons; 2021:616.
16. Jones B, Kenward MG. *Design and Analysis of Cross-over Trials*. 3rd ed. Monographs on statistics and applied probability. CRC Press; 2015:412.
17. Julious SA. Why do we use pooled variance analysis of variance? Viewpoint. *Pharm Stat*. 2005;4(1):3-5.

Appendix

A general model for outcomes from n-of-1 trials arranged in cycles can be expressed as follows.

$$Y_{irs} = \lambda_i + \beta_{ir} + \varepsilon_{irs} + Z_{irs}\tau_i, \tag{1}$$

where Y_{irs} is the measured outcome for occasion $s, s = 1,2$ of cycle $r, r = 1,2 \ldots k_i$ for patient $i, i = 1,2 \cdots n$. Here $\lambda_i \sim N\left(\Lambda, \phi^2\right)$ is a random effect for patient i, $\beta_{ir} \sim N\left(0, \gamma^2\right)$ is a random effect for cycle r within patient i, ε_{irs} is a random error term for occasion s of cycle r for patient i and $\tau_i \sim N\left(\mathrm{T}, \psi^2\right)$ is a random treatment effect for patient i, with $Z_{irs} = -½, ½$, depending on whether the patient was assigned A or B on that occasion in that cycle. All stochastic terms are assumed independent of each other.

If we reduce everything to within-cycle differences first, then the random patient and cycle terms are eliminated and only σ^2, γ^2 are relevant to calculating our estimates. We can have

$$\hat{\tau} = \frac{\sum_{i=1}^{n}\sum_{r=1}^{k_i}\frac{Y_{ir2} - Y_{ir1}}{Z_{ir2} - Z_{ir1}}}{\sum_{i=1}^{n} k_i} \tag{2}$$

as an estimate of T. Since $Z_{ir2} - Z_{ir1} = 1, -1$ depending on whether A is given on the first occasion in a cycle or the second, this is simply the sum of all the within-cycle differences for treatment B minus treatment A divided by the total number of cycles. If we have the same number of cycles, k, per patient this simplifies to

$$\hat{\tau} = \frac{\sum_{i=1}^{n}\sum_{r=1}^{k}\frac{Y_{ir2} - Y_{ir1}}{Z_{ir2} - Z_{ir1}}}{nk}. \tag{3}$$

What the appropriate variance of this estimator is depends on what we consider it is an estimate of, that is to say, what we consider T to be. For example, if we take it to be an estimate of the mean

treatment effect for these patients, then this is fixed for the sample. We shall refer to this as the *local purpose*. We then have that the variance is

$$Var\left(\hat{\tau}\right)=\frac{2\sigma^2}{\sum_{i=1}^{n}k_i}. \tag{4}$$

In the balanced case where $k_i=k,\forall i$ then we have

$$Var\left(\hat{\tau}\right)=\frac{2\sigma^2}{nk}. \tag{5}$$

On the other hand, if we take T to be the mean treatment effect in a population of patients from whom the patients studied may be taken to be a random sample, then we have

$$Var\left(\hat{\tau}\right)=\frac{\psi^2}{n}+\frac{2\sigma^2}{\sum_{i=1}^{n}k_i}, \tag{6}$$

with, in the balanced case,

$$Var\left(\hat{\tau}\right)=\frac{\psi^2}{n}+\frac{2\sigma^2}{nk}. \tag{7}$$

We refer to this as the *global purpose*. Note that for the global purpose a) this estimator is only optimal in the unbalanced case or if $\psi^2=0$ and b) whether or not this is optimal, the variance for the global is only the same as for the local purpose if $\psi^2=0$.

An alternative approach to estimation starts with the individual patient estimates,

$$\hat{\tau}_i=\frac{\sum_{r=1}^{k_i}\frac{Y_{ir2}-Y_{ir1}}{Z_{ir2}-Z_{ir1}}}{k_i}. \tag{8}$$

For the global purpose these have variances

$$Var\left(\hat{\tau}_i\right)=\psi^2+\frac{\sigma^2}{k_i}. \tag{9}$$

These estimates may then be combined in a weighted sum to produce an estimate

$$\hat{\mathrm{T}}_{global}=\sum_{i=1}^{n}w_i\hat{\tau}_i, \tag{10}$$

where

$$w_i=\frac{\frac{1}{\psi^2+\sigma^2/k_i}}{\sum_{i=1}^{n}\left(\frac{1}{\psi^2+\sigma^2/k_i}\right)}, \tag{11}$$

that is to say, with weights inversely proportional to the variance and summing to one. Note that(9), (10) and(11) define an estimate that has the same general form as a random effects meta-analysis estimator, the only practical difference being that σ^2 should be estimated globally, rather than individually patient by patient. The variance of (10) is given by

$$Var\left(\hat{T}_{global}\right)=\frac{1}{\sum_{i=1}^{n}\frac{1}{Var\left(\hat{\tau}_i\right)}}. \tag{12}$$

Note also, that if $k_i=k,\forall i,i=1\cdots n$ we have from(10) that $\hat{\mathrm{T}}_{global}=\sum_{i=1}^{n}\hat{\tau}_i\Big/n$ and from (12) that $Var\left(\hat{\mathrm{T}}_{global}\right)=Var\left(\hat{\tau}\right)\big/n$.

In section 8 the formula for shrunk estimates was given as

$$shrunk=w\times personal+\left(1-w\right)global. \tag{13}$$

If we assume that a suitably large number of patients have been studied, then the global estimate as a prediction for the long term average may be assumed to have a variance of ψ^2 whereas the local estimate for patient i may be assumed to have a variance of $2\sigma^2\big/k_i$. These two estimates should be weighted proportionately to the inverse of their variances, so we have

$$w=\frac{\psi^2}{\psi^2+2\sigma^2\big/k_i},1-w=\frac{2\sigma^2\big/k_i}{\psi^2+2\sigma^2\big/k_i}. \tag{14}$$

Since w is the weight for the personal element and ψ^2 is the variation in the true treatment effect from patient to patient, we can see that, other things being equal, as this variation becomes more important more weight is given to the global estimate. Similarly, since $1-w$ is the weight for the global estimate, we can see that as the within patient variation σ^2 gets larger, then more weight will be given to the global estimate, although this can be reduced by increasing the number of cycles k_i in which the patient is observed.

Finally, we have as a formula for the variance of the shrunk estimate,,

$$Var(shrunk)=\frac{2\psi^2\sigma^2}{k_i\psi^2+2\sigma^2}. \tag{15}$$

Note that if we have no local information on patient i so that $k_i=0$ we have that (15) is equal to ψ^2, which, since we must rely on global information only, is to be expected. On the other hand, as $\psi^2\to\infty$ we have that (15) $\to 2\sigma^2\big/k_i$ which is the personal variance, which again is only to be expected, since the results from other patients contribute no information. In general, however, (15) is lower than either the global or the personal variance.