Applying Bayesian Hierarchical Models to N-of-1 Trials

Tuomo Kareoja

04/06/19

Contents

Chapter 1

Introduction

N-of-1 trials in clinical practice are multiple crossover trials conducted on a single patient, where treatment periods with different treatments are formed into multiple blocks each of which contains at least one period of each treatment under consideration[1]. By comparing the measurements taken during different treatments, a comparisons between treatment options can be made and the most suitable treatment option chosen for the particular patient in question.

In the following pages I give a concise explanation of the experimental design of N-of-1 trials and how we can statistically model these kinds of experiments given the kinds of challenges their design poses. I then follow up with a how we can apply Bayesian methods in estimating the effectiveness of different treatments highlighting how these methods can gives us needed flexibility when running N-of-1 trials by not relying on hypothesis testing depending on a prespecified design. I then consider the case when there are multiple similar N-of-1 trials and show how it is possible to pool the information from these with hierarchical Bayesian methods, without loosing sight of the goal of N-of-1 trials: to find the best treatment most suitable for each patients particularities. Finally I end by with a complete example of analyzing multiple N-of-1 trials with hierarchical Bayesian methods using RStan-package with simulated data.

Chapter 2

What Are N-of-1 Trials?

In the appraisal of any medical treatment the "gold standard" is a randomized controlled trial (RCT), where subjects are randomized to two or more groups that are given different treatments or no treatment at all. The measurement from theses groups are then compared and a result derived about which treatment is most effective on average. This design takes into account unknown factors that might make some patients more suspectable to one treatment by forming the groups randomly and thus, on average, distributing these patients evenly between groups. By comparing the groups against each other and not just to the same patients at the beginning and end of the study, it also takes into account time related effects like natural progression of disease. Despite their unquestionable value in finding general effects, this method can run into problems when we actually try to apply its results to individual patients in practice.

Knowing the best treatment in average might not help much in finding right treatment for a particular patient if the variability in treatment effectiveness between patients is high. Although we can use covariates like age, gender, or certain gene variant to explain the variance between patients in RCT:s, there usually is still much unexplained variance. Some of this variance is of course caused by random factors like measurement error, but a significant part might be caused real and significant differences between the patients. In other words there might be significant factors that explain the variability of the efficacy of different treatments between patients that might either be too specific to be considered in a RCT study or unknown and thus impossible to analyze in this kind of experimental design.

Another problem with RCT:s is the peculiarity of their participants. It is common practice to accept patients to RCT:s only if they don't suffer from any medical issue besides the one that is being studied. This lack of comorbidity makes it easier to get clear results by removing confounding factors, but at the same time this lowers the external validity of the results, because in the real world patients often suffer from multiple medical issues

simultaneously. Because of this it might be possible that even when the scientific literature provides clear results about the relative effectiveness of different treatments, these results don't actually generalize that well to the kinds of patients the clinical practitioner sees in her office.

N-of-1 trials can be used to patch the holes in the knowledge that RCT:s cannot fill by changing the focus of the study from group averages to individual patient. In N of 1 design multiple treatments are tried sequentially in periods that are formed into blocks each of which contains a treatment period of each treatment option under consideration at least once. Within each treatment period the effects of the treatments are measured in comparable ways. Depending on the ease of measuring the outcome of interest, the measurements could be done just once per treatment period or even multiple times per day or even. Each block includes a period of each treatment under consideration in random or balanced order to take into account time related effects. A simple example would be a ABBA design that includes two blocks with which the two treatments A and B are assigned in a balanced order. Depending on the treatments considered, there might be also be a so called "washout" period between treatments, where the patient does not receive any treatment and her state is allowed to return to baseline. This method is used to prevent treatment interactions, that could make it difficult to analyze the results or could be dangerous to the patient. If the treatments under study allow, N-of-1 trial can also use the double-blind method, where the patient and caregiver both don't know which treatment is used at which point, and placebo treatment, where one of the treatments only resembles treatment, but does not consider any active ingredient (pharmacological or otherwise). Below is schema of a more complex N-of-1 trial with three treatments, random assignment of treatments and a washout period.

picture

The stated aim of N-of-1 trials is quite different from RCT:s where the latter tries to generalize results to population and find which treatment is best in general, the N-of-1 trial tries to only generalize to the patient in question. This means that where as comorbidity and other factors that might cause systematic variation between the patients in treatment outcomes are a problem in RCT:s, these are not a problem in N-of-1 trials because there is no need to generalize farther that this one patient. There is also no actual to know what is the factor causing a certain treatment to work better or worse, as long as it is not because of measurement errors or time related effect.

Use of N-of-1 trials is appropriate in situations where there are multiple treatment options but there is no prior knowledge of which of these would be best, when there is known to be considerable variability between patients in treatment efficacy, or when there is reason to doubt that the results from scientific literature generalize to the patient in question. This would seem to make n-of-1 trials applicable to quite a few situations, but there are also multiple factor restricting their use.

Firstly, N-of-1 trials can only be used on illnesses that are chronic, progress slowly and are at least somewhat stable. Also the treatments options available must have a noticeable treatment responses within a short timeframe. This is because running the trial needs time to complete and fast changing illness or slowly effecting treatment make it either impossible to distinguish true effects from the natural progression of the disease or make the trials impracticably lengthy. N-of-1 trials are also unsuitable for testing of preventative treatments because the effects of these treatments are often also impossible to measure without comparisons to other patients without the same treatment.

Secondly, running a N-of-1 trial is costly because of added expenses of training the medical staff in the method, running the trial with all its measurements and analyzing the data. This is means that it can be hard to find cases where using this method is cost effective and studies making these kind calculations have come to pessimistic conclusions (1,2).

These limitations have kept the use of N-of-1 trials rare in clinical use, even though they could potentially both increase the life quality of the patients and lower the health-care costs by finding most suitable medications to patients that might end up potentially using them for years. This state of affairs might be changing fast though. Aging and environmental factors are changing worlds disease burden so that bigger and bigger proportion of it is constituted by chronic diseases (3), of which common ones like non-acute cardiovascular diseases and diabetes are excellent candidates for N-of-1 trials. Also the cost of administering N-of-1 trials are dropping with advent of cheap and reliable health sensors like smartwatches and connected blood pressure monitors. For example in diabetic patients it is now possible to get real time readings of blood insuline levels with minimal effort from the patient (4). These factors mean that the popularity of this method could potentially rise significantly in the future.

Chapter 3

Statistical Modeling of N-of-1 Trials

Even though the data created by the N-of-1 trials resembles traditional time series data with autocorrelation between observations and repeated measurements from the same study unit, there are additional complexities that are caused by the structure of repeating treatment periods. Trying to take into account all the peculiarities of the study design could end up with model too complicated to the small amount of data generated by a single study, so one must consider carefully what factors actually need to incorporated in to the model.

Simplest model that we could employ is to just count the number of blocks where a treatment is considered "better" than others. The precise definition of "better" doesn't matter here. This way we arrive to a simple binomial model where the number of "successes" X is the number of blocks where a treatment is considered the "best" one follows binomial distribution and the probability of each treatment option of having k successes is given by:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{(n-k)} \tag{3.1}$$

Where k = number of blocks, where the treatments is considered the "best", n = total number of blocks and p = the probability of being considered the "best".

This type of model is rudimentary at best, because it fails to consider the magnitude of the differences between treatments and does not take into account the actual number of measurements within each treatment period. To take these factors into account more complex models are in order.

3.1 Basic Models

Before going further I make the assumption that the measurements are continuous as this is probably the most common case. Lets first look at a model where we assume that there are no time-trends and no autocorrelation between measurements. Let y_{mbpt} represent the outcome measured while on treatment m within treatment block b within treatment period p at time t:

$$y_{mbpt} = \mu_m + \gamma_b + \delta_{p(b)} + \epsilon_{t(p(b))} \tag{3.2}$$

where $\gamma_b \sim N(0, \sigma_\gamma^2)$, $\delta_{p(b)} \sim N(0, \sigma_\delta^2)$, and $\epsilon_{t(p(b))} \sim N(0, \sigma_\epsilon^2)$

This model assumes all treatment effects mu_m to be constant. Between the normally distributed terms γ_b represents random block effects, $\delta_{p(b)}$ random period effects and $\epsilon_{t(p(b))}$ random within period errors. We could choose one of the blocks as a reference and set $\gamma_1 = 0$ and assume that within each block the between period effects follow a same pattern, e.g. difference between treatment period one and two is the same within each block.

The random effects between blocks and treatment periods could represent for example the random variations in the motivation of the patient and possible changes in treating personnel within each block and treatment period. The random within period errors represents the measurement error of a single measurements within treatment periods. The relative sizes of the these terms are important for effective design of the trial because they determine if it is more be beneficial for the statistical power of the study to add more measurements, treatment periods or blocks.

If the measurements within blocks and within periods do not correlate, the model 3.2 can be simplified by dropping γ_b and $\delta_{p(b)}$:

$$y_{mbpt} = \beta_m + \epsilon_{t(p(b))} \tag{3.3}$$

Where $\epsilon_{t(p(b))} \sim N(0, \sigma_\epsilon^2)$

This simple model is a natural fit in scenarios where there is just one measurement within each treatment period.

3.2 Incorporating time-trends into the model

As the symptoms of the patient might not be completely stable (e.g. because symptoms get worse with the progression of the disease) fitting some kind of time-trend to the model might be advisable. We can modify the model 3.3 from previous chapter to include a linear

time-trend by adding an intercept and slope of the time-trend. In this case we can express the model just in terms of the measurement y_t taken at time t.

$$y_t = \beta_0 + \beta_1 t + \mu_t + \epsilon_t \tag{3.4}$$

Where $\epsilon_t \sim N(0, \sigma^2)$

Here β_0 is the intercept, β_1 the slope of the time trend, μ_t the effect of the treatment given during time t and ϵ_t is the residual error at time t. More complex time-trends can be introduced by modifying the slope, for example by adding the term $\beta_2 t^2$ to introduce a quadratic trend.

Another effect dependent on time to take into consideration are period effects. There might be some part of the trial that is within a period that we presume to have its own effect. An example of this kind of effect is if we study asthma medications and part of the trial falls within the pollen season. A simple way to model this is to use a dummy variable that takes the value 1 within the period and 0 outside it. Extending the model 3.4 with a period of constant effect β_2 we end up with:

$$y_t = \beta_0 + \beta_1 t + \beta_2 Z_t + \mu_t + \epsilon_t \tag{3.5}$$

Where $\epsilon_t \sim N(0, \sigma^2)$ and dummy variable $Z_t = 1$ when $t \in (t_{period\,start} \ldots t_{period\,end})$ and 0 otherwise.

Lastly, to take into account that treatment effects themselves can vary with time, for example because treatment works better during periods of greater disease severity, we can add a time-by-treatment interaction effect into the model. For example in the case where we expect that the illness gets steadily (e.g. linearly) worse with time, but the treatments compensate this by being similarly more effective, we can extend the model 3.4 that includes a simple linear time-trend by adding an interaction term:

$$y_t = \beta_0 + \beta_1 t + \mu_t + \mu_t \beta_1 t + \epsilon_t \tag{3.6}$$

Where $\epsilon_t \sim N(0, \sigma^2)$

3.3 Autocorrelation from Carryover Effects and the Slow Onset of Treatment Effects

A common occurrence in time series data is the autocorrelation between measurements, so that there is similarity between observations defined by a function of time lag between

them. A common first response to this problem is adding is to a time-trend to the model like we did in the previous chapter. This detrending often removes a substantial proportion of the autocorrelation, that could be caused for example by the natural progression of disease or seasonal variations in its symptoms[2]. Unfortunately in N-of-1 trials the problems of carryover effects and slow onset of treatment effect can lead to very complex autocorrelation patterns that are hard to remove with simple time-trend.

Carryover effects refer to the lingering effects of the treatment even after it has beed stopped. This can make the treatment effects in next treatment period with different treatment seem larger (or smaller in the unfortunate and hopefully rare case where the previous treatment was actually harmful) than they actually are. Carryover effects also encompass the effects of interactions between sequential the treatments, which could even be dangerous depending on the nature of the treatments. On the other hand treatment effects that manifest slowly can often give the opposite effect of carryover effects by making the treatments look less effective than they really are during the first measurements of each treatment period.[2]

To deal with these more devious sources of autocorrelation in our model we can take two routes. First possibility is to use a autoregressive model where we express the measurement error at time t as function of one or more previous measurement errors:

$$y_t = \mu_m + \epsilon_t \tag{3.7}$$

Where $\epsilon_t = \rho\epsilon_{(t-1)} + \mu_t$ in which ρ is the correlation between consecutive errors and $\mu_t \sim N(0, \sigma^2)$ is the error term.

Instead of making the error dependent on just the previous error the model can be adjusted to include more complex lag by defining $\epsilon_t = \rho_1\epsilon_{t-1} + \rho_2\epsilon_{t-2} + \ldots + \rho_x\epsilon_{t-x} + \mu_t$, where ρ_x is the correlation between the errors separated by x time units.

Second approach is a dynamic model where we express the autocorrelation in the measurements themselves so that the measurement at time t is a function of the measurement at $t-1$:

$$y_t = \rho y_{t-1} + \mu_t + \epsilon_t \tag{3.8}$$

Where ρ is the correlation between consecutive measurements and $\epsilon_t \sim N(0, \sigma^2)$ is the error term.

Although it can make more intuitive sense to make the make whole observation dependent on the previous observation, it is important to recognize that in this case the treatment effects μ_t must be interpreted differently as they are now conditioned on the previous measurements.

Although we can try to solve problems created by carryover effect and slow manifestation of treatment effect with modelling, a better way could be to take measures to mitigate the effects in the study design itself. By having long enough washout period between treatments we can minimize the carryover effect. If there are no harmful interactions between the treatment the next treatment can be started within the washout period so that we minimize the problem of slow treatment effects. If there are interactions to be taken into consideration, the first few measurement at the beginning of each treatment period could also just be dropped. By doing this we are of course throwing away data, but we must be remember that if the measurements at the beginning of the treatment period are mangled, this will also mangle our parameter estimates. Trying to take these effects into account in our model will probably not eliminate these effects completely and will increase the complexity of our model.

3.4 Non-continuous Measurements

Up to this point I have assumed that the measurement used are continuous, but we could of course have measurements that are binary, categorical or counts. With these kind of measurement the models need reformatting so that they don't presume normal distributions. The models still don't have differ much in principal from the models presented above and the principals underlined before can be applied.

To modify previously presented models to work in these cases we need to to formulate them as generalized linear models. We do this by keeping the right-hand sides of the equations intact but expressing the left-hand side in terms of link function of the probability distribution for the outcomes.

With a binary outcome, the measurements follow the Bernoulli distribution $y_t \sim Bernoulli(p)$ where the mean of the is the probability of the event measured. The link function is the logit function $logit(p) = log_e(\frac{p}{1-p})$. To formulate

For binary measurements we use logistic regression, categorical measurements categorical logistic regression and measurement in numbers of event Poisson regression.

$$y_{mbpt} = \beta_m + \epsilon_{t(p(b))} \tag{3.9}$$

Where $\epsilon_{t(p(b))} \sim N(0, \sigma_\epsilon^2)$

and equivalently, after exponentiating both sides

$$y_{mbpt} = \beta_m + \epsilon_{t(p(b))} \tag{3.10}$$

Where $\epsilon_{t(p(b))} \sim N(0, \sigma_\epsilon^2)$

Chapter 4

Applying Bayesian Principles

Now that we have some models defined we need to move into the next part of the analysis and actually give estimates to the parameters in the models

The Bayes principle

$$p(\theta|D) \propto p(D|\theta)p(\theta) \tag{4.1}$$

Including previous knowledge by using informative priors

Uncertainty and communicating this to a lay audience

Because of computational difficulties usage of Bayesian estimation has been historically confined to simple likelihood-function with appriopriate conjucate priors where the integral is possible to solve analytically.

Applying Bayes principle to a simple model

Chapter 5

Combining Information From Several N-of-1 Trials With Hierarchical Models

Conducting a individual N-of-1 trials is expensive but if we notice that there are some treatment or certain population that benefits greatly from this kind of design, we could lower the cost significantly be conducting multiple similar trials.

If this is possible on top of lowered costs we also have the option to pool the information from the separate trials to achieve greater statistical power.

This gives as a new challenge: how to pool the information so that we still take into account the individuality of each patient and don't end up recommending the same treatment to everyone regardless of the variability between the patients?

By applying Bayesian statistics we could achieve this be using hierarchical models where each parameter is imagined to come from "higher" distributions that is controlled by its own parameters. For example we could assume that the effectiveness of a certain treatment to be normally distributed within a population and the effectiveness of this drug for a single patient

If have two parameters instead of one we can just slot them into Bayes' rule and see how it applies to the joint parameter space:

$$p(\theta, \omega|D) \propto p(D|\theta, \omega)p(\theta, \omega) \tag{5.1}$$

To make this actually useful we need to to factor the right-hand side futher into a a chain of dependencies:

$$p(\theta, \omega|D) \propto p(D|\theta, \omega)p(\theta, \omega) = p(D|\theta)p(\theta|\omega)p(\omega) \tag{5.2}$$

We can now see that the data D depend only on the value of θ, so that if the value of θ is set then the data are independent of all other parameter values. Similarly, the value of θ depends conditionally only the value of ω and ω is an independent variable. Any model that can be factored into this kind of dependency chain is a hierarchical model.

These kind of dependencies among parameters are useful in several respects. First, the dependencies are meaningful for the given application. Second, because of the dependencies across parameters, all the data can jointly inform all the parameter estimates. This reduction of variance in the estimators, relative to the data, is the general property referred to by the term "shrinkage".

In general, shrinkage in hierarchical models causes lower-level parameters to shift toward the modes of the higher-level distribution. If the higher-level distribution has multiple modes, then the low-level parameter values cluster more tightly around those multiple modes, which might actually pull some low-level parameter estimates apart instead of together. The most amazing thing is that if we don't explicitly set the parameter values of the higher-level distributions, the amount of shrinkage is actually informed by the data so that similar observed data points from lower-level distributions lead to "tighter" estimates for the higher-level distributions and in this in turn leads to greater shrinkage.

Lets formulate and example where we have a simple balanced N-of-1 trial with two treatments and two blocks each with one period of both treatments. With these restrictions the only two possible trial design are ABBA and BAAB. If we assume that there is only one continuous measurement within each period and that there are no time-trends and no autocorrelation a good model to use on a single trial would be:

$$y_{mbpt} = \beta_m + \epsilon_{t(p(b))} \tag{5.3}$$

Where $\epsilon_{t(p(b))} \sim N(0, \sigma_\epsilon^2)$ and β_m

Now lets imagine that instead of just one N-of-1 trial we have conducted 30 similar trials. To tie these trials together at the parameter level. Lets adjust our model by adding an additional layer of dependencies to the treatment effects β_m:

$$y_{mbpt} = \beta_m + \epsilon_{t(p(b))} \tag{5.4}$$

Where $\epsilon_{t(p(b))} \sim N(0, \sigma_\epsilon^2)$ and β_m

A picture make this much clearer

Chapter 6

Example of a Hierarchical Bayesian Analysis Using Simulated Data

Imagined experimental design
- How the data was simulated
- What makes the simulated data hierarchical
- picture of the hierarchical parameter structure used to create the data
- What prior information do we have?

6.1 Analyzing a Single Trial

Lets begin by analyzing just single trial by itself.

6.1.1 Defining the Model

Defining the priors and incoporating our previous knowledge.
- STAN-laskenta ja mita siina pitaa ottaa huomioon

6.1.2 Results

Ajatuksena ottaa edustava otos posteriorijakaumasta ja tehda tasta otoksesta paatelmia koko posteriorijakaumasta

6.2 Analyzing Multiple Trials With Hierarchical Models

Hierarkisessa mallissa, jossa parametrit korreloivat voimakkaasti keskenaan,

yksinkertaiset samplaystekniikat voivat johtaa hyvin hitaaseen laskentaan. Demonstraatio huonolla liikkuvuudella graafissa.

6.2.1 Defining the Model

More efficient algorithms are available

HMC kayttaa askeleiden ehdottamiseen menetelmaa, jossa ehdotusten todennakoisyys jakauma muuttuu kulloisenkin sijainnen mukaan, niin etta on suurempi todennakoisyys ehdottaa arvoja siina suunnassa, jossa posteriori jakauman arvot kasvavat tarkastelemalla posteriorijakauman negatiivista logaritmin gradientteja.

HMC:ssä hiearkisen mallin keskenaan korreloivat parametrit eivat muodostu niin suureksi laskennaliseksi ongelmaksi.

sampling-menetelmasta

6.2.2 Results

Visualisation of the shrinkage

Bibliography

[1] Richard L. Kravitz, Naihua Duan, Sunita Vohra, Jiang Li: Introduction to to N-of-1 Trials: Indications and Barriers in Design and Implementation of N-of-1 Trials: A User's Guide, AHRQ, 2014.

[2] Christop H. Schmid, Naihua Duan: Statistical Design and Analytic Considerations in Design and Implementation of N-of-1 Trials: A User's Guide, AHRQ, 2014.