Safe anytime-valid inference: from theory to implementation in psychiatry research

Turner, R.J.

Chapter 7 is based on the paper that is accepted for oral presentation AISTATS 2023. Chapter 7 in this thesis contains the extra section 7.4, linking it to the psychiatry use-case studied in chapter 5. The paper without extra section 7.4 is available in the AISTATS conference proceedings as:

Rosanne J. Turner and Peter D. Grünwald. Safe Sequential Testing and Effect Estimation in Stratified Count Data. PMLR 206. February 2023. URL: https://proceedings.mlr.press/v206/turner23a.html

Chapter 7

Safe Sequential Testing and Effect Estimation in Stratified Count Data

Rosanne J. Turner[1,2], Peter D. Grünwald[1,3]

1: CWI, Machine Learning group, Netherlands
2: University Medical Center Utrecht, Brain Center, Netherlands
3: Leiden University, Department of Mathematics, Netherlands

Abstract

Sequential decision making significantly speeds up research and is more cost-effective compared to fixed-n methods. We present a method for sequential decision making for stratified count data that retains Type-I error guarantee or false discovery rate under optional stopping, using *e-variables*. We invert the method to construct stratified anytime-valid confidence sequences, where cross-talk between subpopulations in the data can be allowed during data collection to improve power. Finally, we combine information collected in separate subpopulations through pseudo-Bayesian averaging and switching to create effective estimates for the minimal, mean and maximal treatment effects in the subpopulations.

7.1 Introduction

Fixed-n hypothesis tests and confidence intervals limit research opportunities and quick decision making, as they rely on static research designs where data are only evaluated at one time point. We aim to develop hypothesis tests for conditional independence and anytime-valid confidence sequences for stratified treatment effects in subpopulations that retain a guarantee on the probability of falsely rejecting the null hypothesis and coverage of the true effect under continuous monitoring of data. To this end we use *e-values*, tools for constructing tests that keep the *type-I error rate* (or *false positive rate*) controlled under sequential testing with optional stopping. Over the last four years, e-values have become the standard tools (essentially, the appropriate alternative for p-values) for dealing with such settings. Below we summarize the essentials; for much more background on the budding field of e-processes (also known as 'testing by betting' and 'safe testing') see the recent overview [Ramdas et al., 2022] and specifically for details on e-values refer to Grünwald et al. [2022a], Vovk and Wang [2021]. In this paper, we develop e-processes for stratified 2×2 tables, enabling, in Section 7.2, anytime-valid (i.e. valid under optional stopping) conditional independence (CI) tests for Bernoulli streams for two groups a and b (e.g. a is control, b is treatment), where the test is conditional on a third variable, the stratum. Based on these CI tests, we then, in Section 7.3, develop anytime-valid confidence sequences (henceforth just called 'confidence sequences') for a notion of effect size representing divergence from CI. The importance of our tests is ubiquitous in e.g. medical statistics — we can think of the CI test in Section 7.2 as an an anytime-valid sequential version of the Cochran-Mantel-Haenzel test, a work-horse in the field of epidemiology. Our e-processes are generalizations of those designed for 2×2 tables (same setting as ours, but with just a single stratum) by Turner et al. [2021], Turner and Grünwald [2023]. To achieve the generalization, we employ tools from the theoretical machine learning literature, most notably the literature on *prediction with expert advice* [Cesa-Bianchi and Lugosi, 2006], which extends Bayesian learning techniques with ideas such as 'sleeping', 'switching' and the like. Moreover, inspired by these ideas, we develop the novel notion of *cross-talk* between strata, which allows us to make confidence intervals *narrower* if outcomes in various strata are interrelated, while nevertheless remaining *valid* even if they are not. While for many statistical models, anytime-valid tests need more data to reach a desired conclusion than fixed n methods and anytime-valid confidence intervals are somewhat wider than standard ones [Ramdas et al., 2022, Grünwald et al., 2022a], we find in this paper that we can partially counteract this difference by employing the cross-talk strategy (which is not available for fixed-n methods), as is illustrated by comparing our confidence sequences to fixed-n confidence intervals for Mantel-Haenszel risk differences in Section 7.3.

E-Processes Consider a random process $Y_1, Y_2, \ldots$ and let $\mathcal{H}_0$, the *null hypothesis*, be a set of distributions for this process. An e-variable for $Y_j, Y_{j+1}, \ldots, Y_m$ conditional on $Y^{(j-1)} = (Y_1, \ldots, Y_{j-1})$ for testing $\mathcal{H}_0$ is any non-negative random variable S that can be written as function of $Y^{(m)} = (Y_1, \ldots, Y_m)$

such that

$$\forall P \in \mathcal{H}_0 : \mathbb{E}_P[S \mid S^{(j-1)}] \leq 1; \tag{7.1}$$

for $j = 1$ we set $\mathbb{E}_P[S \mid S^{(0)}] := \mathbb{E}_P[S]$ and call S an *unconditional* e-variable; an e-value is the value an e-variable takes on a realized sample. It is easily shown that for any sequence $S_1, S_2, \ldots$ where S_j is an e-variable for $Y_{(j)}$ conditional on $Y^{(j-1)}$, the product $E^{(m)} := \prod_{j=1}^{m} S_j$ is an unconditional e-variable for $Y^{(m)}$. $E^{(1)}, E^{(2)}, \ldots$ is called a *test martingale* or *e-process* (see Ramdas et al. [2022] on how e-processes strictly generalize test martingales). Via Ville's inequality, it is shown that e-processes have the remarkable property that, for any $0 < \alpha < 1$, *the probability that there exists an m such that $E^{(m)} \geq 1/\alpha$ is bounded by α.* As a consequence, if we look at the data at some time m and reject if $E^{(m)} \geq 1/\alpha$, the probability under the null of falsely rejecting the null is at most α no matter how we chose m; it may be determined by external circumstances (do we have money to experiment further?) or by aggressive stopping rules such as 'keep sampling until you can reject the null', or even by peeking into the future. Tests with this property are called *safe under optional stopping* and Ramdas et al. [2020] show that, in essence, all reasonable such tests should be based on e-processes. Just like p-values can be converted into confidence intervals, e-process can be converted into anytime-valid confidence tests, also known as *confidence sequences* — we will explore these in Section 7.3.

Setting We consider the *stratified contingency table* setting/model. Under the *global null* hypothesis (we consider more complicated nulls later), outcomes $Y \in \{0, 1\}$ are independent of groups $X \in \{a, b\}$ (e.g. representing interventions) given their stratum $k \in [K] := \{1, ..., K\}$. We formalize this by measuring time in terms of *blocks*: we assume that at each time $j = 1, 2, \ldots$, we are given a stratum indicator $k_j \in [K]$ and we observe a block of $n = n_a + n_b$ outcomes, with n_a outcomes in group a and n_b in group b, all in the same stratum k_j. We write $Y^{(m)} = (Y_1, \ldots, Y_m)$ with Y_j the data vector corresponding to the j-th block arriving. Hence $Y_j = (Y_{j,a,1}, \ldots, Y_{j,a,n_a}, Y_{j,b,1}, \ldots, Y_{j,b,n_b})$ is a vector in $\{0, 1\}^n$ denoting $n = n_a + n_b$ outcomes in k_j. Under both null and alternative, all blocks are assumed independent, with each outcome in group x in stratum k independently $\sim$ Bernoulli$(\theta_{x,k})$. Formally, the null hypothesis then expresses that

$$\mathcal{H}_0 : \theta_{a,k} = \theta_{b,k} \text{ for all } k. \tag{7.2}$$

We will assume $n_a = n_b = 1$ for all strata in simulation examples in this paper, but these can be chosen freely in practice and can even be adapted in between data blocks — as long as they are set at or before the beginning of a data block, they are allowed to depend on the past. Of course, in practice, we often deal with $2K$ i.i.d. streams of data, one for each group-stratum combination, with data not necessarily coming in at the same rate for different strata/groups. While superficially different, we can still recast this setting in terms of blocks: for example, participant may sequentially enter a study and are each independently randomized with probability $1/2$ to receive 'treatment' (group b) or 'control/placebo' (a). We then wait until

the first time t_1 that we have seen n_a outcomes in group a and n_b outcomes in group b in the same stratum; we call this stratum k_1, denote these n outcomes Y_1, and proceed observing outcomes in the various streams until the first time t_2 that there is another stratum k_2 (potentially $k_2 = k_1$) so that we have seen n_a outcomes in group a, n_b in group b in stratum k_2; we denote these n outcomes Y_2, and so on. If we want to stop at any time t, we take as data all blocks that have been completed so far, and ignore all started-yet-unfinished blocks.

Related Work The first paper to use e-processes for conditional independence testing is [Lindon and Malek, 2022], but their tests are very different from ours and involve a *simple* null hypothesis, allowing them to use Bayes factors for their e-processes. Further, Turner et al. [2021], Turner and Grünwald [2023] develop independence tests and confidence sequence for 2×2 tables; our paper is a direct extension of theirs, extending their techniques to the stratum-conditional case. Very recently four other related papers have appeared: [Pandeva et al., 2022, Grünwald et al., 2022b, Shaer et al., 2022, Duan et al., 2022]: these papers all differ from ours in that they assume data are jointly i.i.d. (i.e. one observes a single i.i.d. stream $(X_1, Y_1, K_1), (X_2, Y_2, K_2), \ldots$). The latter three also make the so-called *Model-X* assumption (the distribution of $X_i \mid K_i$ is assumed known). Our paper is complementary: we do not need the i.i.d. or Model-X assumption and as explained above, our setting does not just capture data in blocks (such as paired data) but also data in the form of $2K$ i.i.d. streams, one for each group in each stratum, with no stochastic assumptions about what group or what stratum arrives at what time. The price we have to pay is that we can only deal with a small number of strata and with finite sets of outcomes and number of groups (in this paper we focus on 2 but extension to the finite case is straightforward); aforemetioned references can deal with arbitrary covariate and outcome random variables K_i and Y_i. Nevertheless, small-strata-count-studies are highly common in the medical statistics world, and we show here how to construct efficient sequential tests for them.

The code used for experiments in this paper will initially be placed on the repository linked to this publication [Turner, 2023], and will later be integrated in the safestats R package [Ly et al., 2022].

7.2 E-variables for testing the global null

We first consider the case where there is only one stratum, $k_j = k^*$ for each each j. The problem is then reduced to testing whether two Bernoulli data streams come from the same source. Turner et al. [2021] showed that in this case, for arbitrary estimators $\breve{\theta}_a|Y^{(j-1)}, \breve{\theta}_b|Y^{(j-1)}$, the following is an e-variable for Y_j conditional on $Y^{(j-1)}$, i.e. (7.1) holds with $S := S_j$ given by

$$S_j = \prod_{i=1}^{n_a} \frac{p_{\breve{\theta}_a|Y^{(j-1)}}(Y_{j,a,i})}{p_{\breve{\theta}_0|Y^{(j-1)}}(Y_{j,a,i})} \prod_{i=1}^{n_b} \frac{p_{\breve{\theta}_b|Y^{(j-1)}}(Y_{j,b,i})}{p_{\breve{\theta}_0|Y^{(j-1)}}(Y_{j,b,i})}, \tag{7.3}$$

where $p_\theta(Y) = \theta^Y(1-\theta)^{1-Y}$ denotes the Bernoulli(θ) probability of $Y \in \{0,1\}$), as long as we pick $\breve{\theta}_0 \in \Theta_0 = [0,1]$ as follows:

$$\breve{\theta}_0 = \breve{\theta}_0 | Y^{(j-1)} := \arg\min_{\theta \in [0,1]} D(P_{\breve{\theta}_a, \breve{\theta}_b} \| P_{\theta,\theta}) \stackrel{(a)}{=} \frac{n_a}{n}\breve{\theta}_a | Y^{(j-1)} + \frac{n_b}{n}\breve{\theta}_b | Y^{(j-1)}. \tag{7.4}$$

Here and in the sequel, P_{θ_a,θ_b} represents the distribution on $n_a + n_b$ independent binary outcomes with the first n_a outcomes $\sim$ Bernoulli(θ_a) and the subsequent n_b outcomes $\sim$ Bernoulli(θ_b), i.e. the distribution of outcomes in a single block according to (θ_a, θ_b), and $D(P_{\theta_a,\theta_b} \| P_{\theta'_a,\theta'_b})$ abbreviates the KL divergence between two such distributions. Equality (a) follows by simple calculus.

Importantly, in (7.3), $(\breve{\theta}_a, \breve{\theta}_b) \in \Theta_1 = [0,1]^2$ *can be chosen as a function of past data anyway we like, not affecting the Type-I error guarantee.* Nevertheless, if we were given the true probabilities θ_a^* and θ_b^* of the two groups in block j, then we could set $\breve{\theta}_a = \theta_a^*$ and $\breve{\theta}_b = \theta_b^*$ and this choice is special: the e-variable (7.3) then has, among all e-variables, the largest expected logarithm under the true alternative $P_{\theta_a^*,\theta_b^*}$. We then say it is *growth-rate optimal* (GRO) for collecting evidence against the null hypothesis [Grünwald et al., 2022a]. Formally, we define

$$\text{GRO}(\theta_a^*, \theta_b^*) := \sup_S \mathbf{E}_{Y_j \sim P_{\theta_a^*,\theta_b^*}}[\log S] \tag{7.5}$$

where the supremum is over all random variables S that are e-variables for Y_j under $\mathcal{H}_0$. It directly follows from [Grünwald et al., 2022a, Theorem 1] that, if we plug in $\breve{\theta}_a = \theta_a^*$ and $\breve{\theta}_b = \theta_b^*$ into (7.3), then the resulting S_j is GRO and its growth rate is equal to the KL divergence, i.e.

$$\mathbf{E}_{Y_j \sim P_{\theta_a^*,\theta_b^*}}[\log S_j] = \text{GRO}(\theta_a^*, \theta_b^*) = D(P_{\theta_a^*,\theta_b^*} \| P_{\tilde{\theta},\tilde{\theta}}), \tag{7.6}$$

where $\tilde{\theta} = (n_a/n)\theta_a^* + (n_b/n)\theta_b^*$. Growth-rate optimality is the analogue of statistical *power* in the sequential setting: if we plug in these 'true' $\breve{\theta}_a = \theta_a^*, \breve{\theta}_b = \theta_b^*$, we expect the product $E^{(m)}$ to increase as fast as possible in m, enabling us to reach $1/\alpha$ and reject the null hypothesis as fast as possible, compared with all other possible e-processes. In practice though, θ_a^* and θ_b^* are unknown, but to get near-grow-optimal e-variables, we can *estimate* $\breve{\theta}_a$ and $\breve{\theta}_b$ based on all data seen before data block j — then $\breve{\theta}_a$ and $\breve{\theta}_b$ converge to θ_a^*, θ_b^* and our e-variables S_j get better and better in the GRO sense. We follow Turner et al. [2021] who successfully chose to place a beta prior on the parameter space and took the Bayesian posterior mean as an estimate.

In treatment/ control test settings, there often exists prior knowledge of a minimal clinically relevant or expected odds ratio $\text{OR}(\theta_a, \theta_b) := (\theta_b/(1-\theta_b))((1-\theta_a)/\theta_a)$, i.e. it is known that $\text{OR}(\theta_a, \theta_b) = \phi$ for some given ϕ. In that case, one can restrict estimating $\breve{\theta}_a$ and $\breve{\theta}_b$ to $\Theta_1(\phi) = \{(\theta_a, \theta_b); \text{OR}(\theta_a, \theta_b) = \phi\}$, possibly improving power and growth-rate of the test [Turner et al., 2021]. Both search spaces are illustrated in Figure 7.1.

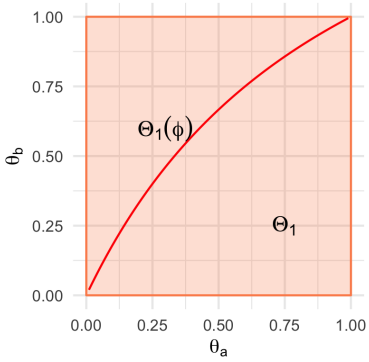


Figure 7.1: Parameter space $\breve{\theta}_a|Y^{(j-1)}$ and $\breve{\theta}_b|Y^{(j-1)}$ are estimated in, in 2×2 table without strata; either through placing a beta prior on the entire unit square (in light orange) and calculating the posterior mean with all data up to and including time $j-1$ or through restricting the posterior estimation to a particular odds ratio value ϕ and placing a beta prior on all pairs (θ_a, θ_b) corresponding to this odds ratio value (for example the red curve, for $\phi = 2$).

Combining e-variables from individual strata We can use the e-variable in (7.3) to calculate e-process values $E^{(m),k}$ for each stratum k separately. To be precise, we set S_j^k to the equivalent of (7.3) if $k = k_j$,

$$S_j^k = \prod_{i=1}^{n_a} \frac{p_{\breve{\theta}_{a,k}|Y^{(j-1)}}(Y_{j,a,i})}{p_{\breve{\theta}_{0,k}|Y^{(j-1)}}(Y_{j,a,i})} \prod_{i=1}^{n_b} \frac{p_{\breve{\theta}_{b,k}|Y^{(j-1)}}(Y_{j,b,i})}{p_{\breve{\theta}_{0,k}|Y^{(j-1)}}(Y_{j,b,i})}, \tag{7.7}$$

and $S_j^k = 1$ otherwise, i.e. if $k_j \neq k$, and $E^{(m),k} := \prod_{j=1}^m S_j^k$ — note that at each 'time j', the product e-variable only changes for the k such that j-th block was a block of outcomes in stratum k.

We now need to combine the e-processes-per-stratum into a single e-process for (7.2) to measure evidence against $\mathcal{H}_0$ and allowing tests with type-I error probability guarantee on (7.2), the global null hypothesis that the odds ratio of the success probabilities equals 1 in each stratum. There are several ways to do this. The first and most straightforward option is to *multiply* the individual e-values across the strata:

$$E^{(m)} = \prod_{j=1}^{m} S_j^{k_j} = \prod_{j=1}^{m} \prod_{k=1}^{K} S_j^k. \tag{7.8}$$

To see that $E^{(1)}, E^{(2)}, \ldots$ is an e-process, simply note that each $S_j^{k_j}$ is a conditional e-variable (i.e. it satisfies (7.1) with $S = S_j^{k_j}$) since, given that S_j in (7.3) is a conditional e-variable, $S_j^{k_j}$ must be an e-variable as well. When $\theta_{a,k} \approx \theta_{b,k}$ in a few of the strata, this might be a data-inefficient approach, as one would need to collect a lot of extra evidence in the strata where the success probabilities

are substantially different to counteract the expected small e-values in the other strata. A second option that possibly better handles these cases is to create a *convex combination*, i.e. a mixture, of e-values at each time point j (any convex combination of e-variables is also an e-variable [Vovk and Wang, 2021]). A simple first option is to pick some prior distribution on the strata $\pi(k)$, and to use that distribution for calculating the mixture after each batch comes in:

$$S_j := \sum_{k=1}^{K} \pi(k) S_j^k \ ; \ E^{(m)} = \prod_{j=1}^{m} S_j \text{ so that also}$$
$$E^{(m)} = \prod_{k=1}^{K} E^{(m),k} \text{ with } E^{(m),k} = \prod_{j=1}^{m} S_j^k. \tag{7.9}$$

Extending the simple averaging above, we could replace the prior $\pi(k)$ in (7.9) with a distribution $\pi(k|y^{(j-1)})$ that depends on previous data $y^{(j-1)}$, since, since we assume the data itself in each block are independent, dependency of π on past data will not affect guarantee (7.1). Such an approach is called the *method of mixtures* in the anytime-valid testing literature [Ramdas et al., 2022]. Thus, any distribution on $[K]$ that depends on the past is allowed here, but an intuitive choice is a *pseudo-Bayesian posterior*

$$\pi(k|y^{(j-1)}) := \frac{\pi(k)(E^{(j-1),k})^\eta}{\sum_{k'} \pi(k')(E^{(j-1),k'})^\eta}, \tag{7.10}$$

where by definition, $E^{(0)} = 1$ and we pick η beforehand as a *learning rate*: if we set it to a higher value, we will focus on strata with higher e-values more quickly; with $\eta = 1$, (7.10) becomes similar to a Bayesian posterior. Just as the beta-posterior used to determine $\breve{\theta}_{x,k}$ in (7.7) allows us to learn $\theta^*_{x,k}$, this new posterior allows us to learn which strata can help us most to reject the null. However, even for $\eta = 1$ the analogy to Bayes only goes so far — for example, at each j, only the e-variable S^{k_j} for stratum k_j changes; the other S^k 'sleep' [Koolen and Van Erven, 2010] and thus $E^{(j-1),k}$ behaves differently from a likelihood. This more general past-determined updating originates in the area of machine learning called *prediction with expert advice* where many other such 'posterior'-updates have been considered [Herbster and Warmuth, 1998, Van Erven et al., 2007, Koolen and De Rooij, 2013]. These include the more extreme approach called *switching*. With this approach, we calculate (7.9) with $\pi(k)$ replaced by any distribution we like (the choice is again allowed to depend on $Y^{(j-1)}$) up to and including a particular batch j^*. Thereafter, for $j \geq j^*$, we set

$$\pi^*(k|y^{(j)}) = \begin{cases} 1 & \text{if } k = k^* \text{ with } k^* = \arg\max_k E^{(j^*),k} \\ 0 & \text{otherwise} \end{cases} \tag{7.11}$$

creating a new E-process $E^{(1)}_{[j^*]}, E^{(2)}_{[j^*]}, \ldots$ such that, for $m \leq j^*$, $E^{(m)}_{[j^*]} = E^{(m)}$ and,

for $m > j^*$,

$$E_{[j^*]}^{(m)} = E^{(j^*)} \cdot \prod_{j=j^*+1}^{m} E^{(j),k^*} \tag{7.12}$$

j^* could arbitrarily be picked prior to the study, or we could also place a prior on the moment of switching and take a weighted average over (7.12) for various values of j^* for each δ, thereby obtaining yet another e-process with j^* 'integrated out' (see Figure S7.2 in the supplementary material for a more elaborate comparison of switch priors in a simulation experiment for confidence sequences).

In Figure 7.2, the three different methods for combining e-variables for testing $\mathcal{H}_0$ are compared with respect to *power*: the expected probability of rejecting $\mathcal{H}_0$ under some fixed data generating distribution. For Figure 7.2, data were sampled from a distribution where risk differences and control group rates all differed between strata. It can be observed that all methods that took the stratification into account outperformed the unstratified approach, where just one sequential e-variable was calculated for all strata combined. The three different methods will be re-compared for confidence sequences in Figure 7.6.

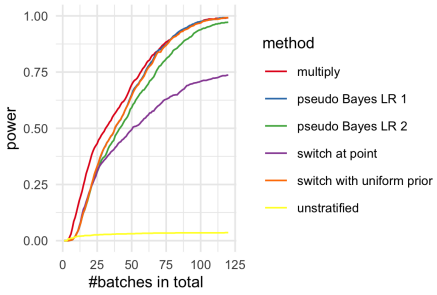


Figure 7.2: Power for rejecting the null at level $\alpha = 0.05$ that the odds ratio in all strata equals 1 estimated with 1000 repeated experiments for various e-variable combination methods. 40 batches were collected in each of three strata (so maximum sample size was $m = 120$) and sampling was stopped as soon as $E^{(m)} \geq \frac{1}{\alpha}$. Real control group success rates were $0.1, 0.2, 0.8$ and real risk differences were $0.05, 0.4, -0.6$. Pseudo-Bayesian approaches were implemented with learning rates (LR) 1 and 2. Switch approaches were implemented for switching at point $j^* = 10$, or with a uniform prior on switch times $j = 5$ until $m - 5$.

Cross-talk between strata To further improve power of the hypothesis test, we will allow for *cross-talk* between strata while estimating $\breve{\theta}_{a,k}$ and $\breve{\theta}_{b,k}$ based on data seen so far. In the current simple setting of testing the global null, 'cross-talk' simply amounts to design S_j that grow faster (allowing for faster rejecting of the null) if the alternative satisfies certain *constraints*. For example, if one expects treatment effects (say, measured as odds ratios) to be stable (identical) throughout different strata, but control group recovery rates to vary, one would like cross-talk about the odds ratios between strata. Practically, this means that to arrive at the

estimates $\breve{\theta}_{x,k} \mid Y^{(j-1)}$, we first limit the parameter space to $\Theta_1(\hat{\phi}^{(j-1)})$, i.e. all vectors $\theta_{x,k}$ with odds ratio $\hat{\phi}^{(j-1)}$, set to be the maximum likelihood odds ratio based on all previous data in all strata, i.e. calculated by ignoring strata. We then calculate $\breve{\theta}_{x,k} \mid Y^{(j-1)}$ as posterior means using beta priors conditioned on the parameters being in $\Theta_1(\hat{\phi}^{(j-1)})$. Similarly, when one expects control group recovery rates to be stable, but the treatment effects to vary because of a possible interaction with stratum characteristics, allowing cross-talk about control group recovery rates might improve power. In practice, we achieve this by using as beta prior parameters for the control group rate $\breve{\theta}_{a,k}|Y^{(j-1)}$ the total counts of failures and successes aggregated over all strata (summed with some initial prior parameters to ensure stable estimates at time point $j = 1$; we set initial prior values 0.18 for both the fail and success rate based on a suggestion by [Turner et al., 2021]). In the odds-ratio cross-talk scenario, we effectively constrain the parameters of the alternative $\breve{\theta}_{x,k} \mid Y^{(j-1)}$ at each j to share the same odds-ratio; in the control-group cross-talk scenario, we constrain these parameters to share the same θ_a, i.e. $\breve{\theta}_{a,k}|Y^{(j-1)} = \breve{\theta}_{a,k'}|Y^{(j-1)}$ for each k, k'. Would one be unsure whether cross-talk would improve power at all, and if so, whether one should cross-talk on the odds ratios or the cross ratios, one could put prior mass 1/3 on each of the corresponding three e-values, say $E_\rho^{(m)}$ for $\rho \in \{\text{NONE}, \text{ODDS}, \text{CONTROL RATE}\}$, where NONE stands for the standard e-variable without cross-talk. One could then, for each block j, use a mixture e-variable, where the three e-values are mixed as in (7.10) with $\eta = 1$, k replaced by ρ and '$E^{(j-1),k}$' replaced by '$E_\rho^{(j-1)}$' giving a new 'MIX' e-process. All four cross-talk scenarios are explored in simulations in Figure 7.3, where data were generated from strata with similar control group success rates, but different risk differences, and different control group success rates, but similar odds ratios showing that allowing for cross-talk on control rate or odds ratio improves power in the respective scenarios. The cross-talk mixture performs comparably to the optimal cross-talk options in both cases. Cross-talk can be expected to improve power even if, in truth, under the alternative, the odds-ratio resp. control-group rate is just similar, but not exactly the same under all groups; and the confidence sequences of the next section remain valid (but will get wider) even if the odds-ratios resp. control-group rates happen to be completely different. Thus, the method described here cannot really be viewed as a constraint on the model, and we chose to call it *cross-talk* instead: data in one stratum informs, 'talks to' estimates for other strata.

A GRO-Sanity Check While the simulations above and below show encouraging empirical results regarding the power of our methods, it is still useful to have some theoretical assurance that, no matter the 'true' alternative generating the data, all methods we consider produce e-values that grow fast (i.e. achieve good power) under this alternative. We now provide a simple theorem to this end. As usual in the e-value and safe-testing literature, and for reasons explained by Grünwald et al. [2022a], we concentrate on GRO (7.5) rather than power.

Theorem 7.2.1. Suppose that we observe $m = m_1 + \ldots + m_K$ blocks, with m_k blocks lying in stratum k, each such block sampled independently from $P_{\theta^*_{a,k},\theta^*_{b,k}}$.

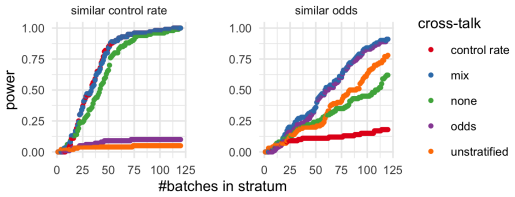


Figure 7.3: Power for rejecting the null hypothesis at level $\alpha = 0.05$ that the odds ratio in all strata equals 1 estimated with 100 repeated experiments for various types of cross-talk. 40 batches were collected in each stratum and sampling was stopped as soon as $E^{(m)} \geq \frac{1}{\alpha}$. On the left, real control group success rates were $0.49, 0.5$ and 0.51 in each stratum; risk differences were $-0.09, -0.49, 0.39$. On the right, real odds ratios were $4, 4.01, 2.95$.

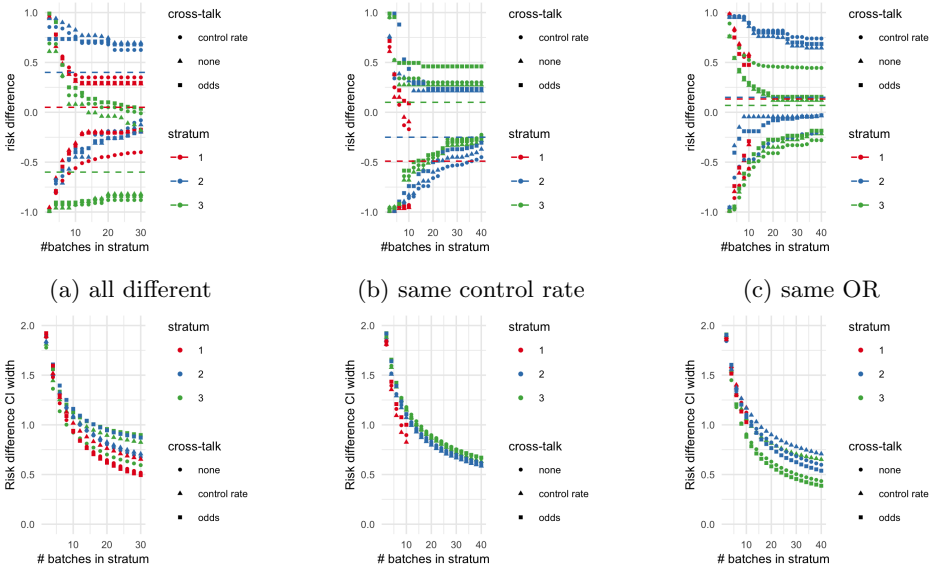
Then, with $\mathbf{E}$ denoting expectation under this distribution, the e-process $E^{(m)}$ defined by multiplication as in (7.8) and the MIX e-process $E^{(m)}$ as above with constituent e-processes defined multiplicatively as in (7.8) both achieve:

$$\sum_{k=1}^{K} m_k \text{GRO}(\theta^*_{a,k}, \theta^*_{b,k}) = \mathbf{E}\left[\log E^{(m)}\right] + O(\log m). \tag{7.13}$$

To interpret the result, note that, if an oracle were to supply us with $\theta^*_a = (\theta^*_{a,1}, \ldots, \theta^*_{a,k}), \theta^*_b = (\theta^*_{b,1}, \ldots, \theta^*_{b,k})$ i.e. if we were told 'if the alternative were true, then its parameters would be $P_{\theta^*_a, \theta^*_b}$', then we could use the GRO (growth optimal e-variable) which, conditional on observing a block in stratum k, would obtain the optimal, largest possible expected growth $\text{GRO}(\theta^*_{a,k}, \theta^*_{b,k})$. Since we assume data to be independent, the best growth we could obtain with such an oracle is given by the left-hand side of (7.13). The theorem expresses that the price for learning (via Bayes predictive distributions $\breve{\theta}_{x,k}$ based on beta-priors) rather than knowing θ^*_a, θ^*_b is modest, namely logarithmic in m whereas the growth itself is linear in m; this is the standard situation for parametric settings, described in detail by Grünwald et al. [2022a]. We may expect the constant hidden in the $O(\log m)$ to become substantially smaller if the preconditions for effective cross-talk hold as described above, e.g. odds ratios or group recovery rates are identical or similar across strata; but determining this constant precisely across cases, as well as extending the analysis to pseudo-Bayesian and switch e-processes, is complicated and will be left for future research. The proof of this theorem can be found in the appendix.

7.3 Extension to confidence sequences

Turner and Grünwald [2023] showed that (7.3) in the 2×2-table (single stratum) can be generalized, to test null hypotheses $\mathcal{H}_0 := \{P_{(\theta_a, \theta_b)}; (\theta_a, \theta_b) \in \Theta_0\}$ beyond



(a) all different (b) same control rate (c) same OR

(d) all different (e) same control rate (f) same OR

Figure 7.4: Examples of 95% stratified confidence intervals ((a), (b) and (c)) and mean confidence interval widths estimated over 100 runs ((d), (e) and (f)) with different types of cross-talk. In (a), (b) and (c) the true risk difference of the data generating distribution in each stratum is indicated by a dashed line. For (a) and (d), the data were generated by distributions with different control group success rates (0.1, 0.2 and 0.8) and risk differences (0.05, 0.4 and −0.6) in each stratum. For (b) and (e), strata sizes were unbalanced: as can be seen for stratum 1, the red points, data collection stopped after 10 batches. Control group success rates were all 0.5 and risk differences were different (−0.49, −0.25 and 0.1). For (c) and (f), strata sizes were unbalanced as well, and now odds ratios were the same in each stratum (2), but control group rates differed again (0.2, 0.25 and 0.85).

'$\theta_a = \theta_b$':

$$S_{j,[\Theta_0]} = \prod_{i=1}^{n_a} \frac{p_{\breve{\theta}_a|Y^{(j-1)}}(Y_{j,a,i})}{p_{\breve{\theta}^\circ_a|Y^{(j-1)}}(Y_{j,a,i})} \prod_{i=1}^{n_b} \frac{p_{\breve{\theta}_b|Y^{(j-1)}}(Y_{j,b,i})}{p_{\breve{\theta}^\circ_b|Y^{(j-1)}}(Y_{j,b,i})} \tag{7.14}$$

is an e-variable, as long as $\Theta_0 \subset [0,1]^2$ is convex and closed. Here $(\breve{\theta}^\circ_a \mid Y^{(j-1)}, \breve{\theta}^\circ_b \mid Y^{(j-1)})$ is defined to minimize KL divergence, i.e. is the pair $(\theta_a, \theta_b) \in \Theta_0$ that minimizes, over Θ_0,
$D(P_{\breve{\theta}_a|Y^{(j-1)},\breve{\theta}_b|Y^{(j-1)}}(Y_a^{n_a}, Y_b^{n_b}) \| P_{\theta_a,\theta_b}(Y_a^{n_a}, Y_b^{n_b}))$. (7.3) is a special case since with $\Theta_0 = \{(\theta,\theta) : \theta \in [0,1]\}$, this KL divergence is minimized by $(\breve{\theta}^\circ_0, \breve{\theta}^\circ_0)$ with $\breve{\theta}^\circ$ as defined underneath (7.3). Again, $\breve{\theta}_a$ and $\breve{\theta}_b$ are estimated based on past data $Y^{(j-1)}$ as in (7.3). Based on (7.14) one can construct an *exact* (nonasymptotic) confidence sequence (CS)

$$\mathrm{CS}_{\alpha,(m)} = \left\{ \delta : E^{(m)}_{[\Theta_0(\delta)]} \leq \frac{1}{\alpha} \right\}, \tag{7.15}$$

with $\Theta_0(\delta) \subset [0,1]^2$ a null hypothesis determined by a divergence measure. By construction, such a confidence sequence is *always-valid* [Ramdas et al., 2022] in the sense that for any δ, any $\theta \in \Theta_0(\delta)$, the P_θ-probability that there will *ever* be an m such that $\delta \notin \mathrm{CS}_{\alpha,(m)}$ is at most α. This means that we can take the *running intersection* of the confidence sequence while retaining coverage, which will be used throughout the simulation experiments in this paper. In this paper, we are going to construct confidence sequences for risk differences as examples, where we are going to test hypotheses of the form $\Theta_0(\delta) := \left\{(\theta_a, \theta_b) \in [0,1]^2 : \theta_b - \theta_a = \delta\right\}$ — below we extend this to the case that differentiates in terms of the strata. Still, everything could also easily be adapted to construct confidence intervals for other divergence measures, such as odds and risk ratios [Turner and Grünwald, 2023].

7.3.1 One CS per stratum

If we expect the effect size values to differ between the strata, one could decide to report a separate confidence sequence for each stratum using (7.15) above. To reach a better estimate sooner, we could however still allow cross-talk on control group success rates or odds ratios between subpopulations, as described in section 2 above. In this setup, we would end up with a *collection* of k confidence sequences:

$$\mathrm{CS}^k_{\alpha,(m)} = \left\{ \delta : E^{(m),k}_{[\Theta_0(\delta)]} \leq \frac{1}{\alpha} \right\}, \tag{7.16}$$

with $\breve{\theta}_a$ and $\breve{\theta}_b$ in $E^{(m),k}$ estimated based on data seen up to time m and $E^{(m),k}$ defined as in (7.9) with S^k_j replaced by $S^k_{j,[\Theta_0]}$ as in (7.14), calculated for stratum k. Illustrations of confidence intervals over time with the three options for cross-talk are depicted in Figure 7.4. As can be observed there, not allowing cross-talk gives the best results when the true data generating distributions in the strata have different control group success rates and odds ratios (see the circle-shaped

points in Figure 7.4d, especially in the third stratum, where the effect size has a different sign). However, when control group rates or odds ratios are similar across strata, allowing cross-talk improves results. See for example Figure 7.4e, where interval width decreases much faster in the smaller stratum 1 while allowing cross-talk about the control group rate. Similar experiments for comparing confidence sequences with and without the mixture of cross-talk methods can be found in the supplementary material, Figure S7.1.

7.3.2 CS for the minimum or maximum

In some scenarios, for example when we do not have the means to collect a large data sample, or when data is very unbalanced in one or more strata, it could be more informative to create one CS for the minimum or maximum effect size value over all strata. To achieve this, we introduce two new forms of null hypotheses and corresponding e-variables that will subsequently be inverted to create two one-sided confidence sequences, for lower and upper bounds on the minimum or maximum.

One-sided CS: upper bound We will first illustrate how to estimate an upper bound on some minimal effect size value over strata[1]. To this end, we consider a null hypothesis of the form $\mathcal{H}_{0,\delta} : \forall k : \theta_k \in \Theta_0(\geq \delta)$ (i.e. for risk difference effect size, $\Theta_0(\geq \delta) = \{(\theta_a, \theta_b) \in [0,1]^2 : \theta_b - \theta_a \geq \delta\}$) and aim to design e-variables to test it. E.g. in the example depicted in Figure 7.5a, we aim to design an e-variable that will reject $\mathcal{H}_{0,\delta''}$ at any batch j with probability less than α (i.e., that offers type-I error guarantee), when the data in the strata are in reality generated by $(\theta_{a,1}, \theta_{b,1})$ and $(\theta_{a,2}, \theta_{b,2})$. We do eventually want to reject $\mathcal{H}_{0,\delta'}$ as $\delta((\theta_{a,2}, \theta_{b,2})) < \delta'$. As we collect more and more data, we can reject null hypotheses corresponding to values of δ' for which $\delta' - \delta((\theta_{a,2}, \theta_{b,2}))$ gets closer and closer to 0.

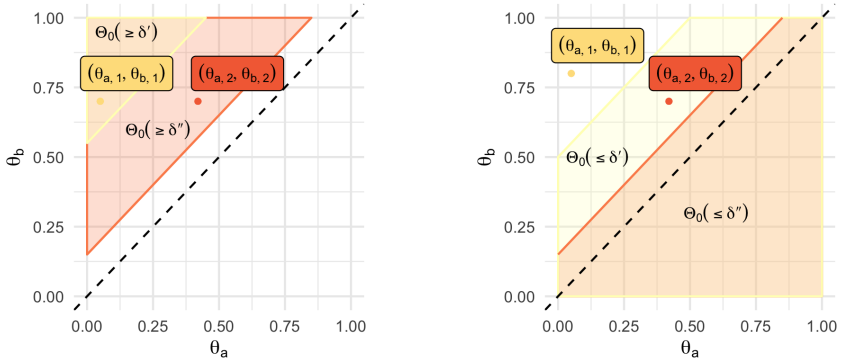
Let us denote the e-process consisting of the e-variables for testing $\theta_k \in \Theta_0(\geq \delta)$ in each stratum combined, using any of the methods described above in Section 2, as $E_\delta^{*(m)}$. The one-sided confidence interval for the minimum effect can be defined as:

$$\mathrm{CS}^{+}_{\alpha,(m)} := \left[-1, \min\left\{\delta : E_\delta^{*(m)} \geq \frac{1}{\alpha}\right\}\right]. \tag{7.17}$$

All possible approaches for combining e-variables from separate strata, as described in Section 2 above, to find an upper bound for the minimal effect size value are compared in the confidence intervals in the paragraph below.

One-sided CS: lower bound We now also aim to estimate a lower bound for the minimal effect size value (or, analogously, an upper bound for the maximal effect size value). To achieve this, we now consider a null hypothesis of the form $\mathcal{H}_{0,\delta} : \exists k : \theta_k \in \Theta_0(\leq \delta)$. Looking at Figure 7.5b as an example, where data are

[1] Analogously, with this method a lower bound on some maximal effect size value can be estimated by reversing all signs.



(a) Examples of parameter spaces for $\mathcal{H}_{0,\delta} : \forall k : \theta_k \in \Theta_0(\geq \delta)$.

(b) Examples of parameter spaces for $\mathcal{H}_{0,\delta} : \exists k : \theta_k \in \Theta_0(\leq \delta)$.

Figure 7.5: Parameter space examples for hypotheses tested to construct upper and lower bounds on minima and maxima of effect size values

generated by $(\theta_{a,1}, \theta_{b,1})$ and $(\theta_{a,2}, \theta_{b,2})$, we aim to design an e-variable that will reject $\mathcal{H}_{0,\delta'}$ at any batch j with probability less than α (i.e., we again want type-I error guarantee if $\mathcal{H}_{0,\delta'}$ is true), as $\delta((\theta_{a,2}, \theta_{b,2})) < \delta'$. We do want to reject as quickly as possible $\mathcal{H}_{0,\delta''}$, as $\forall k, \delta(\theta^{(k)}) > \delta''$. As we collect more data, we can reject null hypotheses with values of δ'' for which $\delta((\theta_{a,2}, \theta_{b,2})) - \delta''$ gets closer and closer to 0.

To build our one-sided confidence interval $\text{CS}^-_{\alpha,(m)}$, we again want to construct a compound e-variable $E^{*(m)}_{\delta}$ testing the null hypothesis corresponding to each value of δ, but now take $\max\{\delta : E^{*(m)}_{\delta} \geq 1/\alpha\}$ as our lower bound. To test $\mathcal{H}_{0,\delta}$ we will use the *minimum* of $E^{(j),k}_{\Theta_0(\leq\delta)}$ over all k, which provides an e-variable for $\mathcal{H}_{0,\delta}$. To see this, let us assume $\mathcal{H}_{0,\delta}$ is true an that for some k^*, $\theta_{k^*} \in \Theta_0(\leq \delta)$; the other data generating distributions might or might not come from $\Theta_0(\leq \delta)$. Then: $\mathbb{E}(\min_k S^k) \leq \min_k \mathbb{E}(S^k) \leq \mathbb{E}(S^{k^*}) \leq 1$.

Combining into confidence interval We now combine the lower bound and upper bound estimation methods established above to build confidence intervals for the minimal effect size value. This can be achieved through taking the intersection of the one-sided confidence sequences introduced above: $\text{CS}_{\alpha,(m)} := \text{CS}^-_{\alpha,(m)} \cap \text{CS}^+_{\alpha,(m)}$. Results from an experiment where in one of the strata the treatment effect was substantially smaller than in the others are depicted in Figure 7.6 (with average interval widths in the supplementary material, Figure S7.3). In early phases of data collection, multiplication gives the quickest convergence, but as more data is collected, the "sequential learning" methods converge quicker. When risk differences where about the same across all strata, multiplication converged the quickest (see Figure S7.4 in the Supplementary material).

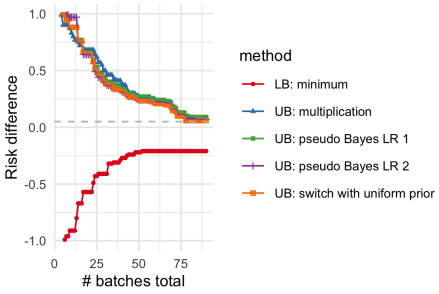


Figure 7.6: Example of confidence sequences for the lower- (LB) and upper (UB) bounds of the minimum effect. 30 observations were made in each stratum, and the real differences were 0.5, 0.4 and 0.05. With the switch method, a uniform prior ranging from $m_{\text{switch}} = 5$ until 30 was applied. With the pseudo-Bayesian approach, the learning rate η was set to 1 and 2. α was set to 0.05.

7.3.3 CS for the mean effect

In addition to estimating the minimum or maximum effect in one of the strata, one might be interested in estimating the mean effect an intervention will have on an entire population, given the existence of subpopulations. For example, one might want to estimate the effect a vaccination will have on the probability of people being contaminated with a disease, taking into account that a certain proportion of the population concerns elderly or immunocompromised citizens.

Assuming we have a trustworthy estimate of the proportion of subjects belonging to each stratum k in the population of interest, π_k, we aim to estimate the mean risk difference (mean expected effect of the intervention) $\delta^* := \sum_k \pi_k \delta_k$. We can build a confidence sequence for δ^* by constructing an e-variable for the set of all possible success probability distributions satisfying this δ^*, $\mathcal{H}_{0,\delta^*} : \{P_{\vec{\theta}}; d(\vec{\theta}) = \sum_k \pi_k d((\theta_{a,k}, \theta_{b,k})) = \delta^*\}$. It is not directly clear what an optimal e-variable could look like; one option that offers both the type-I error guarantee with potentially good power is to combine the growth-rate optimal e-variable (7.3) for a specific δ_k in each stratum with the *universal inference* [Wasserman et al., 2020] method for designing e-processes. Based on this strategy, we look at the set of all vectors $\vec{\delta} := (\delta_1, ..., \delta_K)$ that satisfy $\sum_k \pi_k \delta_k = \delta^*$. For one member of the set, we can calculate the e-variable *based on all batches of data seen up to and including time* m according to (7.3):

$$E_{[\vec{\delta}]}^{(m)} = \prod_k E_{[\Theta_0(\delta_k)]}^{(m),k},$$

where $E_{[\Theta_0(\delta_k)]}^{(m),k}$ can be calculated using estimates for $\breve{\theta}_{a,k}$ and $\breve{\theta}_{b,k}$ as before, only including data seen up to *and not including* batch m. The e-variable for $\mathcal{H}_{0,\delta^*}$ can then be calculated as [Wasserman et al., 2020]: $E_{\delta^*}^{*(m)} = \min_{\vec{\delta}} E_{[\vec{\delta}]}^{(m)}$, and the corresponding confidence sequence can be constructed as before, analogously to (7.17).

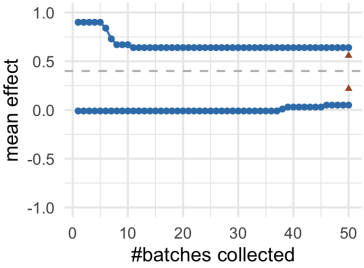


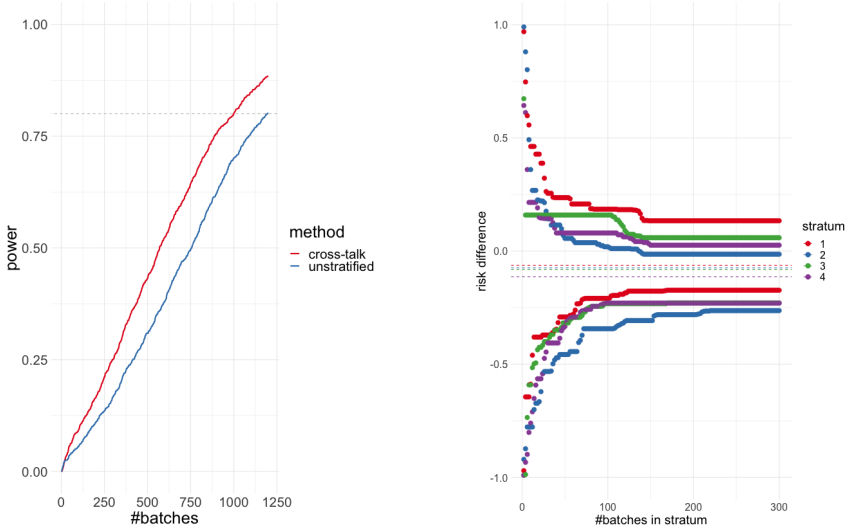
Figure 7.7: Simulated example of 95% confidence sequences for the mean effect across subpopulations. 25 observations were made in each stratum, and the real risk difference of 0.4 was homogeneous across subpopulations. The confidence sequence for the mean effect is plotted alongside the Miettinen-Nuninen confidence interval, a fixed-n confidence interval method, at batch number 50 (the purple triangles). In the supplementary materials, figure S7.5, the mean effect CS is further illustrated for heterogeneous risk differences in strata.

Comparison to fixed-n CI for Mantel-Haenszel risk difference Much of the research into estimating stratified risk differences with coverage guarantee has considered Mantel-Haenszel risk differences, where risk differences or odds ratios are *homogeneous* across strata but control group rates can vary (see for example [Qiu et al., 2019]), with fixed-n designs. This is a strong assumption, and we do not make it ourselves; but we *can* use cross-talk on the risk difference to tailor our confidence sequences so that they adapt (get narrow) if the risk difference is indeed homogeneous. One recent fixed-n approach for this setting was described and implemented by Klingenberg [2014]. In Figure 7.7, our confidence sequence for the mean effect is compared to the Miettinen-Nuninen (MN) confidence interval from Klingenberg [2014] at fixed time 50 in a setting where risk differences were homogeneous. The MN-interval is slightly narrower, but because we are allowed to continuously monitor the confidence interval while retaining coverage with the confidence sequence, we can exclude 0 from the CS considerably earlier than with the fixed-n method — which is remarkable because unlike the MN fixed-n confidence interval, our anytime-valid confidence sequences are also valid if in fact risk differences are not homogeneous.

7.4 Application in psychiatry use-case

We will now illustrate the process of planning and analyzing a study with the stratified, safe anytime-valid tests described in this paper. As a use-case we will look at a recent exploratory study at two major mental healthcare facilities in the Netherlands. Data from 4808 and 735 patients in their first clinical antidepressant treatment trajectory was analyzed in an exploratory Bayesian network analysis [Turner, 2022] (this thesis, chapter 5). This retrospectively collected data set revealed several potential interesting associations between patient characteristics,

treatment choices and treatment outcomes. However, because of this retrospective setup, these patterns cannot yet be interpreted as causal relations (an overview of potential confounders is given in this thesis, chapter 5). Therefore, before these results can be used in clinical applications, further inferential analysis is needed to confirm the formed hypotheses and generate robust uncertainty estimates. Each hypothesis and uncertainty estimates could then be investigated in a randomized controlled trial or prospective study with safe anytime-valid inference. An example of a power analysis and simulation of an anytime-valid confidence interval for one of these hypotheses is given below.



(a) Power estimation

(b) Confidence interval example

Figure 7.8: Power for rejecting the null hypothesis at level $\alpha = 0.05$ that the odds ratio in all strata equals 1 estimated with 1000 repeated experiments, and an example of a resulting confidence interval in such a sampling scheme. 300 batches were sampled in each stratum according to the alternative hypothesis described in the main text and sampling in the power analysis was stopped as soon as $E^{(m)} \geq \frac{1}{\alpha}$. The cross-talk used was of the mixture type described in section 7.2.

One association to explore was that choosing a particular type of antidepressant, a tricyclic antidepressant, increases treatment success, but at a different rate for different groups of patients. For patients without social problems and without antipsychotics prescriptions the probability increased from 63 to 74 percent (stratum 4 in figure 7.8). This effect was smaller for patients with a different combination of characteristics: the success probability increased from 68 percent to 74.3 percent for patients with social problems *and* antipsychotics prescriptions (stratum 1), for patients with only social problems the increase was from 67 to 74.4 percent (stratum 2) and for patients with only antipsychotics from 66 to 74.1 percent (stratum 3). A power analysis to plan an experiment with safe, stratified

analysis and a balanced design was performed, of which the result is depicted in figure 7.8. In the power analysis, we took as the alternative hypothesis one based on the numbers above. The balanced design implies that each time a patient has been treated with and without a tricyclic antidepressant within one of the strata, an interim analysis is performed, and a decision to stop the study or continue can be taken. In figure 7.8 it can be observed that with this design and analysis, we need 194 less batches of data (388 less patients) when we use the stratified analysis with cross-talk, compared to an unstratified safe anytime-valid analysis. In the example of a confidence interval constructed with cross-talk on the right, it can be observed that we can already exclude 0 from the confidence interval in stratum 2 after observing only 140 batches of patients. Would this confidence interval have been displayed live on a dashboard during a study, clinicians could have decided on their recommendation to prefer tricyclic antidepressants for these patients way before the total 1000 batches of patients were seen: they could have decided already much earlier, after the first 280 patients of stratum 2 had been seen.

7.5 Conclusion and future work

We have introduced a new method for global null hypothesis testing and constructing exact anytime-valid confidence sequences in stratified count data. Our method is complementary to previously proposed methods for similar settings as we need no stochastic assumptions about the arrival times of the subgroups or strata, and no Model-X assumptions. We have shown that our tests and estimates are efficient in terms of power, and that precise effect size estimations can be reached with less strong model assumptions compared to pre-existing fixed-n methods, while retaining coverage guarantees and allowing sequential decision making. We have also shown that we can improve the traditional model of global null testing in the CMH-setting through incorporating ideas from machine-learning: allowing for cross-talk between strata, and incorporating pseudo-Bayesian learning and switching between strata for learning compound effect measures.

Our work extends that of Turner et al. [2021] and Turner and Grünwald [2023] to incorporate strata for count data. Their methods, however, are generally implementable for any convex null hypothesis, and future work should explore if they also can feasibly be extended to stratified sequential effect estimation for continuous outcome variables.