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

Turner, R.J.

Safe Anytime-Valid Inference: from Theory to Implementation in Psychiatry Research

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 14 november 2023
klokke 11:15 uur

door

Rosanne Jane Turner

geboren te Castricum, Nederland

in 1992

Promotores:	Prof.dr. P.D. Grünwald	(CWI, Leiden University)
	Prof.dr. F.E. Scheepers	(UMC Utrecht)
Co-promotor:	Dr. A. Härmä	(Maastricht University)

Promotiecommissie:

Dr. R. de Heide	(VU Amsterdam)
Dr. A. Henzi	(University of Bern)
Prof.dr. O.L. Cremer	(UMC Utrecht)
Prof.dr. M.R. Spruit	
Prof.dr. H. Putter	
Prof.dr.ir. G.L.A. Derks	

The work for this thesis was carried out jointly at CWI, the research center for mathematics and computer science in The Netherlands, and at University Medical Center Utrecht. The work was supported by the Dutch Research Council (NWO), as part of the Enabling Personalized Interventions (EPI) project in the Commit2Data–Data2Person program under contract 628.011.028.

Origins of the material

This dissertation is based on the following papers. The author of this dissertation contributed substantially to each of these papers.

Chapter 2 is based on the paper with a minor revision pending and available as a technical report as:

> Rosanne J. Turner, Alexander Ly and Peter D. Grünwald. Generic E-Variables for Exact Sequential k-Sample Tests that allow for Optional Stopping. arXiv:2106.02693. June 2021.

Chapter 3 is based on the paper that is published in *Statistics and Probability Letters* as:

> Rosanne J. Turner and Peter D. Grünwald. Exact Anytime-valid Confidence Intervals for Contingency Tables and Beyond. Statistics & Probability Letters, 109835. 2023.

Chapter 4 is based on the paper that is published in *BMC Psychiatry* as:

> Rosanne J. Turner, Femke Coenen, Femke Roelofs, Karin Hagoort, Aki Härmä, Peter D. Grünwald, Fleur P. Velders and Floortje E. Scheepers. Information extraction from free text for aiding transdiagnostic psychiatry: constructing NLP pipelines tailored to clinicians' needs. BMC Psychiatry. June 2022. doi: 10.1186/s12888-022-04058-z.

Chapter 5 is based on the paper that is published in *Scientific Reports* as:

> Rosanne J. Turner, Karin Hagoort, Femke Coenen, Rosa J. Meijer and Floortje E. Scheepers. Bayesian network analysis of antidepressant treatment trajectories. Scientific Reports, 13(1), 8428. 2023.

Chapter 6 is based on the paper that is published in revised version in *Psychiatry Research* as:

> Yuri van der Does, Rosanne J. Turner, E.J.H. Bartels, Karin Hagoort, Aäron Metselaar, Floortje E. Scheepers, Peter D. Grünwald, Metten Somers and Edwin van Dellen. Outcome prediction of electroconvulsive therapy for depression. Psychiatry Research, 326, 115328. 2023.

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.

Contents

Chapter 1

Introduction

Classical research methods, such as p-value hypothesis testing, have come under intense scrutiny over the past decade [Wasserstein and Lazar, 2016, Benjamin et al., 2018]. It has proven very difficult for researchers to apply them correctly: the prevailing methods taught to applied researchers are actually too rigid for performing research in a modern environment, especially when working in a dynamic manner with lots of collaborations. Sadly, this leads to faulty use of the aforementioned methods and subsequent invalidity of experiment conclusions, which has even led to a call to abandon significance testing altogether [Amrhein et al., 2019].

Partly as a consequence of the above, recently, interest in sequential testing and particularly *safe, anytime-valid inference* (SAVI) with *e-values* has emerged [Wang and Ramdas, 2020, Waudby-Smith and Ramdas, 2020, Vovk and Wang, 2021, Shafer et al., 2021, Orabona and Jun, 2021, Henzi and Ziegel, 2022, Grünwald et al., 2022a]. This framework potentially offers the same functionality as the classical significance testing methods and also provides researchers with plenty of flexibility, for example through enabling *optional stopping*, *optional continuation*, *anytime-valid effect size estimation* and *federated learning*.

In this thesis, the theory of e-values is further developed for performing SAVI in scenarios applicable to healthcare (specifically, for several use-cases in psychiatry), where one wants to estimate treatment effects for small subgroups of patients. It is then explored how one could actually set up a real-time inference process in practice in an automated manner, combining text mining with network analysis techniques for data preparation and exploration and then confirming hypotheses with SAVI [Tukey, 1980]. The overall aim of this work is to contribute to answering the research question "how can one perform real-time research in healthcare using routinely collected clinical data?".

This introductory chapter starts with a sketch of the bigger scope of the research in this thesis: the Enabling Personalized Interventions project, a Dutch nation-wide project with the goal of working toward a *digital health twin* in section 1.1. This section also discusses the potential importance of *federated learning* for the construction of such a digital health twin. Next, an important potential solution for inference in the online, federated setting, the *SAVI* framework, is introduced in section 1.2. In section 1.3, the psychiatry use-case for the methods developed in this thesis is introduced, together with an overview of the current

state of the art for knowledge discovery in psychiatry. The content of this thesis is introduced in detail in sections 1.4, 1.5 and 1.6. Section 1.4 describes chapters 2 and 3, where the development of new instantiations of SAVI are discussed. In section 1.5 chapters 4, 5 and 6 are discussed, where knowledge discovery in psychiatry through network analysis plays a central role. Finally, section 1.6 contains an introduction of chapter 7, where the SAVI techniques are refined specifically for setting up confirmatory (i.e., with the goal of inference) experiments in psychiatry.

1.1 Toward a digital health twin: on the potential role of federated learning and SAVI

The work in this thesis is part of the Dutch nation-wide Enabling Personalized Interventions (EPI) project. The EPI consortium recognizes three current limitations for using the full potential of healthcare data: data and knowledge extracted from data remain in their original location and are not shared, (the correct type of) data is not analyzed to arrive at useful clinical insights, and insights that are generated are not available to clinicians and patients. The goal of this project is to *"liberate, analyze and action (healthcare) data in a trustworthy way"* [The EPI Consortium, 2019]. To this end, EPI strives to develop a framework that will facilitate the development and use of a *digital health twin (DHT)* framework [Bruynseels et al., 2018].

Digital (health) twins are *"in silico representations of an individual that dynamically reflect molecular status, physiological status and life style over time"* (Bruynseels et al. [2018], p. 1). In more detail, a complete DHT in practice would comprise of a patient's health records from all their care providers, amplified with for example wearable data, data from their mobile devices and smart devices. The DHT is *updated in real time* each time new data becomes available in one of the data sources. The added value of the DHT lies within the potential for *continuous learning and providing feedback*: data from many (possibly similar) individuals can be used to learn *patterns* in the data, in particular to learn about the effects of certain interventions.

One can imagine that realizing a DHT framework in practice would be a complicated task, both from a data-infrastructure and a legal perspective. A schematic representation of one possible realization of a DHT framework is depicted in figure 1.1. The first component needed is a data infrastructure that links the EHR and other devices with patient data to their corresponding DHT, and that links the DHTs to the learning algorithms that eventually will produce the clinical insights. The second component are the regulatory constraints placed on these links. Patients should be able to withdraw their consent to transfer (part of) their data to the DHT, or to transfer data from their DHT to the learning algorithms, or even just to generate general clinical insights from their data. The researchers providing the learning algorithms and the health practitioners providing the use cases should not be allowed to access all data in the DHT, but only the data they contractually have access to for their specific projects. The third component consists of the actual knowledge discovery process and the corresponding algorithms that learn from the DHT data: these can receive input from the DHTs and health

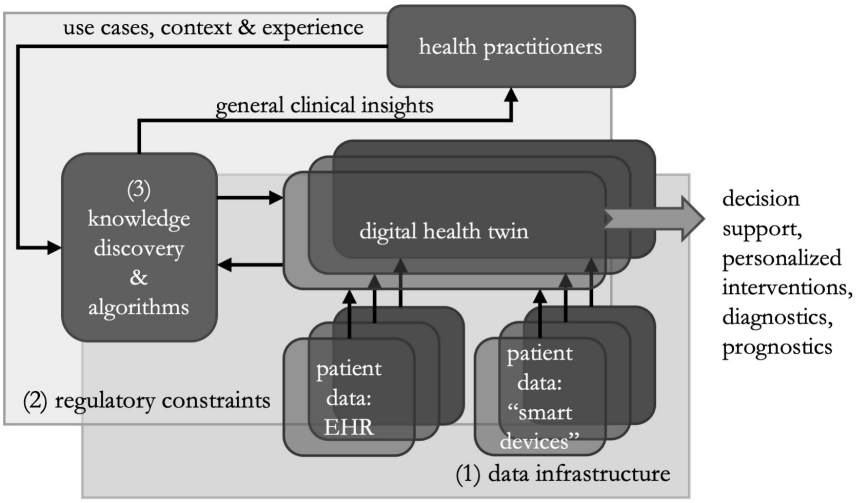


Figure 1.1: Schematic representation of a *digital health twin* framework. Adapted from The EPI Consortium [2019].

practitioners, who can enrich the DHTs' data with existing knowledge from literature. Note that this is a continuous process: each time DHT data is updated or new clinical context is provided, the algorithms are updated. The resulting trained algorithms are then sent back to the digital health twin, to finally enable generating personalized clinical insights to offer decision support and enable personalized interventions through shared decision making.

Key part of the DHT are these learning algorithms that can learn from and make predictions for patients in (near-)real-time. Particularly in the healthcare domain, training such algorithms raises some interesting challenges regarding privacy of patients. During the past decade, two seemingly paradoxical developments have taken place. On the one hand, there has been a rise in initiatives to make research more democratized, accessible and transparent, for example through the development of EU-wide regulations on data availability [Nederlandse Rijksoverheid, 2021]. On the other hand, (European) privacy laws have become much more strict, prohibiting sharing identifiable data without explicit consent for each specific instance [Otto, 2018]. These laws complicate learning in a patient-tailored manner, as learning tailored to smaller and smaller groups of patients (i.e., patients stratified according to more and more characteristics) requires learning from increasing numbers of examples. Collecting all this data in one place and learning from it centrally is often not possible, because of infeasibility in obtaining consent from patients to share data.

Possible solutions lie within not sharing the patient data, but only (parts of) algorithms trained on the data. This is called federated learning [Konečnỳ et al., 2016]. There are two major federated learning scenarios: in the first one, we have

"incomplete" digital health twin versions for single patients stored in separate locations, for example when part of patient data is stored in the general practitioner's system, and part at an academic hospital, and we want to learn from both sources to predict a course of treatment. This is called vertically partitioned data. In the second scenario, data are partitioned horizontally. We do have complete digital health twins, but they are stored at multiple locations, for example in a setting where multiple academic centers are collaborating to train an algorithm for personalized recommendations and need lots of examples.

The work in this thesis focuses on developing learning methods for SAVI for the real-time analysis of horizontally partitioned patient data. The other parts of the DHT and the EPI framework are described in work of S. Amiri on differential privacy (see for example Amiri et al. [2021] and Amiri et al. [2022]), the work of C. Allaart on learning from vertically partitioned data [Allaart et al., 2022], the work of M. Kebede on access control [Kebede, 2021] and the work of J. Kassem on developing an adaptive computing infrastructure that enables implementation all of the aforementioned methods [Kassem et al., 2021].

1.2 Safe, anytime-valid inference

In this section, current methods for confirmatory (inferential) research are described, and it is explained why they are not particularly suitable for implementation in frameworks for distributed, real-time learning such as the EPI framework and the DHT. Next, e-values and their extension to anytime-valid e-processes, the federated learning setting and confidence sequences are described. Throughout this section, we will use a running example of testing and estimating the mean value of the height of a population.

Definitions We will use notation analogously to Ramdas et al. [2022] throughout this introduction. We define Π, a set of distributions on our sample space Ω, and assume that some distribution $P \in \Pi$ generates our data, for example a stream of observations $Y_1, Y_2, Y_3, ...$, where we will abbreviate $Y^n = (Y_1, Y_2, Y_3, ..., Y_n)$. Typically, we want to test whether P aligns with some *null hypothesis* that we have, or if we can reject this null hypothesis for some alternative hypothesis. For example, our null hypothesis might be that the height of people in the Netherlands is distributed according to a normal distribution with a mean of 175 (cm) and an arbitrary standard deviation, and our alternative might be that the height is distributed according to any other normal distribution. Formally we define the set of distributions $\mathbf{P}$ reflecting our null hypothesis $\mathcal{H}_0$ and the set of distributions $\mathbf{Q}$ reflecting our alternative $\mathcal{H}_1$ as (often non-intersecting) subsets of Π. When the set of distributions corresponding to a hypothesis comprises of only one distribution, we refer to the hypothesis as *simple*; otherwise, we call it *composite*. Often, we will consider distributions P_θ (or Q_θ) parameterized by some $\theta \in \Theta$, with parameter space $\Theta_0 \subset \Theta$ corresponding to $\mathcal{H}_0$ and $\Theta_1 \subset \Theta$ to $\mathcal{H}_1$. Uppercase will be used to indicate probability distributions and lowercase for the corresponding probability mass functions or densities.

Current practices and developments in confirmatory research As briefly mentioned in the previous sections, there are a lot of difficulties with the confirmatory phase of research [Peterson, 2021]. One major contributor to these problems is the hypothesis testing methodology, and fundamental disagreements thereon. In the field of statistics, roughly four (partly overlapping) views on hypothesis testing can be recognized. They will be briefly introduced in this subsection, alongside the most important "ingredients" in hypothesis testing, and later the SAVI will be placed in perspective of these practices.

The Fisherian point of view places the emphasis on rejecting a null hypothesis [Fisher, 1925]. Within this Fisherian view, we would set up a study and then calculate a *P-value*:

Definition 1.1 (P-value)**.** A P-value for $\mathbf{P}$ is a random variable PVAL such that $P(\text{PVAL} \leq \alpha) \leq \alpha$ for all $P \in \mathbf{P}$ and $\alpha \in [0, 1]$.

We thus have $P(\text{PVAL} \leq \alpha) = \alpha'$ for some $\alpha' \leq \alpha$, with α' possibly depending on P; for standard p-values, usually $\alpha' = \alpha$, or α' is very close to α. For *conservative* p-values, α' is substantially smaller than α.

In words, this definition implies that the lower the P-value, the less compatibility the data have with the null hypothesis. For example, if a (well-designed and executed) study and analysis to test the null hypothesis that the mean value of the height distribution equals 175 cm revealed a p-value of 0.024, the probability of this occurring under the null hypothesis is *at most* 0.024. Now imagine another study, organized independently of the first, revealing a p-value of 0.0011 for testing the same null hypothesis: a Fisherian would say that in the second study, more evidence against the null hypothesis has been collected, as the probability of observing the second p-value under the null hypothesis would be a lot smaller (at most 0.0011). Another example: if we assume a normal distribution with fixed variance, $\text{PVAL} \leq \alpha$ means that the data have fallen in the $1/2\alpha$ left-tail or right-tail of the distribution. Note that there is no mention of the *alternative hypothesis* in this view of hypothesis testing.

Closely related is the Neyman-Pearsonian view on testing [Neyman and Pearson, 1933]. This is a binary view with a focus on the probability (and penalty) of making an erroneous decision: upper bounds for acceptable error probabilities of wrongly rejecting the null hypothesis (α, "type-I error") and failing to reject the null hypothesis while the alternative is true (β, "type-II" error) are specified *before* each experiment. Experiments are planned based on the α and β thresholds, and only the decision whether the null is rejected or not (rejecting iff $\text{PVAL} \leq \alpha$) is reported. Hence the name "frequentist statistics" that is often used to refer to these methods: they are entirely based on the hypothetical scenario where many experiments are carried out, and the highest acceptable frequency of erroneous decisions in such a collection of experiments. Continuing the height example, an experiment could be planned for testing the null hypothesis that the mean value of the height distribution equals 175 (cm). Planning this experiment with analysis with a classical t-test in mind reveals that when a type-I error probability of 0.05 and type-II error probability of 0.15 are deemed acceptable, the height

of 326 Dutch people would have to be collected to detect a deviation of at least 1 cm to the mean value of 175 cm in at least 85 percent of experiments. After collecting the heights of these 326 people we would perform one t-test, and only report whether the p-value was smaller than or equal to ("reject $\mathcal{H}_0$"), or bigger than 0.05 ("accept $\mathcal{H}_0$").

The actual observed p-value does not give extra information in this view of testing. Interestingly, in applied research, often a mixture of the two is used: the decision to reject the null hypothesis is for example often requested to be reported alongside the p-value in medical journals [Lang and Altman, 2014], complicating the (intended) interpretation of study results.

Note that, irrespective of whether we use Fisherian or NP p-value testing, calculating a p-value requires very precise definitions of the stopping rule and the corresponding experiment setup. In practice, p-values are often used wrongly: for example in an interview study, 56 percent of psychology researchers admitted to "deciding whether to collect more data after looking to see whether the results were significant" [John et al., 2012]. In this scenario, the distribution under the null hypothesis has shifted because the researcher peeks at the data and based on that observation decides to continue sampling. A p-value designed for the null hypothesis where data is collected and only analyzed once (i.e., the ones used in the most well-known frequentist hypothesis tests, such as the t-test or Fisher's exact test) is no longer valid in this scenario. Type-I error can blow up quickly under this kind of malpractice, yielding interpretation of experiment results impossible. See for example an experiment from chapter 2 in this thesis: after collection of 1000 samples and "peeking" at the p-value after each new sample, the type-I error probability increased to 0.30.

The third view of hypothesis testing discussed here is Bayesian, which leaves the frequentist principles and error probabilities behind and instead focuses on updating prior *beliefs* based on new *evidence*. Central roles in Bayesian statistics are played by *prior distributions* and *Bayes marginal distributions*.

Definition 1.2 (Bayes marginal distribution)**.** A prior distribution W_j with density w_j corresponding to hypothesis $\mathcal{H}_j$ is a probability distribution on Θ_j associated with $\mathcal{H}_j$. The Bayes marginal distribution for data Y, where Y could be a single data point or a vector Y^n as above, is defined as

$$p_{W_j}(Y) = \int_\theta p_\theta(Y) w_j(\theta) d\theta.$$

When we have formulated prior distributions (beliefs) for the null hypothesis (W_0) and the alternative hypothesis (W_1), we can define a *Bayes factor* to represent the evidence in favour of the alternative, against the null:

$$\mathrm{BF}_{10}(Y) = \frac{p_{W_1}(Y)}{p_{W_0}(Y)}. \tag{1.1}$$

In contrast with the p-values seen above, the value of the Bayes factor directly represents evidence for the hypotheses: the higher the value, the more evidence

present in the data for the alternative hypothesis. Standardized "levels of evidence" and cut-off values have been proposed for evaluating study results with Bayes factors [Jamil et al., 2017].

Evidently, the choice of prior distributions plays an essential role in the value and generalizability of the Bayes factor. The value of a Bayes factor calculated by one research group will offer little useful information to another research group that does not agree with the prior beliefs the first group incorporated in the Bayes factor. Unfortunately, how to choose prior distributions is still a major topic of discussion within the Bayesian field. On one end of the scale, there are subjective Bayesians, who argue that it only makes sense to express probability as one's pure beliefs in the likeliness of outcomes [De Finetti, 2017, Ramsey, 1931]. No or little weight should be placed on outcomes that in the belief of the researcher are unlikely to ever occur. Returning to our height example, most people would find it very unlikely to find an average length of 190 cm in a random sample of Dutch people, so almost no prior mass should be assigned to this parameter value. Someone who has played a lot of basketball might have a different view of the world and might disagree, and would put more mass on this outcome. On the other end of the scale, there are the objective Bayesians, who strife to define *informationless* prior distributions that attach equal weight to all distributions in the hypotheses [Berger et al., 1998, Jeffreys, 1998, Jaynes, 1957, Savage, 1954]. Looking at the length example again, in this scenario, one might put equal prior mass on the average length being 190 cm, 173 cm, 90 cm, and any other possible human length. Applying such a prior would make it easier to collaborate with research groups with disagreeing views on a subject. However, defining such priors is an intricate process and there exist critical appraisals of objective Bayesianism, arguing that the principles of informationless priors conflict with the factorization of conditional probabilities central in Bayes' theorem [Seidenfeld, 1979].

The last view of hypothesis testing places even more emphasis on evidence in the data collected: this view advocates abolishing the testing process altogether and replace this by estimation with an emphasis on confidence intervals [Berner and Amrhein, 2022].

Definition 1.3 (Confidence set)**.** A set CI is a confidence set for some parameter of interest $\phi : \Pi \to \Delta$ (for example, an odds ratio or mean difference) if:

$$P(\phi(P) \in \text{CI}) \geq 1 - \alpha \text{ for all } P \in \Pi.$$

That is, the probability that we exclude the parameter value corresponding to distribution P when the data are generated by that same P is bounded by α. Usually, $\Delta \subset \mathbb{R}$, and CI are confidence intervals, hence the abbreviation "CI". For example, returning to the length example, our entire set of distributions Π comprises of all normal distributions with mean μ and standard deviation σ: $\Pi = \{P_{\mu,\sigma} : (\mu, \sigma) \in \Theta\}$, $\Theta = \{(\mu, \sigma) : \mu \in \mathbb{R}, \sigma > 0\}$. We might want to create a confidence interval around the mean, and would have the mean length as our measure of interest: we then set $\phi(P_{\mu,\sigma}) = \mu$. When the heights in the population in reality follow some normal distribution $P_{\mu',\sigma'}$, a valid confidence interval at level $\alpha = 0.05$ would include the true mean length μ' in $100 \times (1 - \alpha) = 95$ percent of

experiments.

However, with this approach, we run into the same problems as before: we need an exact definition of our experimental setup to define valid confidence intervals, which means that we again need a very strict description of our study design including setting the final sample size in advance, as with the calculation of p-values described above. Standard confidence intervals cannot be applied for federated, anytime-valid learning, and hence cannot be implemented in settings such as the DHT.

***e*-values** We will now introduce the e-value, the key player in SAVI, and illustrate how it relates to the concepts introduced above. The idea of using e-values for testing hypotheses was originally introduced a long time ago, in the field of information theory by Leonid Levin: he named them *tests of randomness* [Levin, 1976]. However, the theory was not further developed and translations to the field of statistics in terms of interpretation, type-I error guarantee, power and optimality remained non-existent. Around 2019, interest in e-values from a statistical viewpoint suddenly rose, first through separate independent initiatives [Grünwald et al., 2022a, Vovk and Wang, 2021, Shafer et al., 2021, Wasserman et al., 2020], and later through joint work by the pioneers [Ramdas et al., 2022].

Definition 1.4 (e-value)**.** An e-value[1] for null hypothesis $\mathbf{P}$ is a nonnegative random variable E such that the expected value $\mathbb{E}_P[E]$ is at most 1 for all $P \in \mathbf{P}$.

Definition 1.4 says that under the null hypothesis, we do not expect to observe big e-values, as under the null, their expected value is at most 1. We may think of the realized e-value as a betting score: we buy a ticket for 1 euro, and retrieve e euro as the outcome of the bet. Definition 1.4 expresses that we do not expect to gain money under the null hypothesis. This betting score can thus directly be used as a measure of evidence against the null hypothesis [Shafer et al., 2021]: if our score is unexpectedly high, i.e., much higher than 1, we make a large profit in the betting game, and we might reject our null hypothesis. The reader might notice that this interpretation has a lot of parallels to the hypothesis testing with Bayes factors described earlier. In fact, in the case where we have a simple null hypothesis $\mathbf{P} = \{P_0\}$ with corresponding density or mass function p_0, the Bayes factor $p_{W_1}(Y)/p_0(Y)$, for *any* choice of W_1, is an e-value for $\{P_0\}$, as

$$\mathbb{E}_{P_0}\left[\frac{p_{W_1}(Y)}{p_0(Y)}\right] = \int_Y p_0(Y)\frac{p_{W_1}(Y)}{p_0(Y)}dY = \int_Y p_{W_1}(Y)dY = 1. \tag{1.2}$$

However, (1.2) will for most Bayes factors not hold for *composite null* hypotheses, as most Bayes factors for composite null hypotheses are not e-values. Nevertheless, interestingly, further on we will see that in a certain sense *optimal* e-values also take on the form of Bayes factors. Besides this evidential interpretation, there

[1] Throughout the other chapters in this thesis we will make a distinction between the random variables, e-variables, and their realized values, e-values, but to improve readability of this introductory chapter we will use the term e-values for both concepts here, analogous to the way in which we refer to p-values.

is also a connection to frequentist testing and p-values. By Markov's inequality, it is straightforward that we can also use e-values in a frequentist manner, in a hypothesis test with type-I error probability guarantee at level α:

$$P\left(E \geq \frac{1}{\alpha}\right) \leq \alpha \mathbb{E}_P[E] \leq \alpha.$$

Similarly, it can be derived that $1/E$ is a conservative p-value (see definition 1.1, the conservativeness resulting from the trading of some of the test's power for the improved flexibility of e-values, as we will see below. Interestingly, with e-values, we now are able to combine the frequentist and Fisherian views discussed earlier, as they allow for post-hoc determination of the type-I error probability threshold, allowing for better utilization of extreme observations in frequentist hypothesis testing scenarios [Grünwald, 2022].

From e-values to e-processes The introduction on e-values so far only considered single tests: now, we will extend the e-values to safe, anytime-valid e-processes, which will be the main concern in this thesis. Let us again consider the sample space Ω, now equipped with filtration $\mathbf{F}$.[2] We define our "starting capital" (as in the betting interpretation) $E_0 = 1$. The stream of e-values $(E_0, E_1, E_2, E_3, ..., E_t)$ calculated on data stream $Y^t = (\emptyset, Y_1, Y_2, Y_3, ..., Y_t)$ is then a *conditional e-process* if:

$$\mathbb{E}_P[E_t | \mathbf{F}_{t-1}] \leq 1. \tag{1.3}$$

The collection of e-values in equation (1.3) are called a *conditional e-process*. Each e-value E_t for a new batch of data Y_t can be calculated taking into account *any combination of information* available up to (not including) time t. Multiplication of the elements of a conditional e-process also yields an e-value, which is key:

$$E^{(t)} = \Pi_{j=1}^{t} E_j.$$

The collection $(E^{(1)}, E^{(2)}, E^{(3)}, ...)$ is an (unconditional) *e-process* (proposition 2 in Grünwald et al. [2022a])[3]. Combining this fact with Ville's inequality shows that we can use these e-processes to perform *safe anytime-valid tests* ([Ville, 1939], corollary 1 from Grünwald et al. [2022a]):

$$\text{For all } P \in \mathbf{P} : P\left(\text{ there exists } t \text{ s.t. } E^{(t)} \geq \frac{1}{\alpha}\right) \leq \alpha. \tag{1.4}$$

[2] This is a measure theoretic concept. $\mathbf{F}_t$ can be interpreted as all possible combinations of information we may have observed during our experiments up to and including time t. This may also include side information we do not necessarily directly incorporate in our hypothesis test, for example our research budget or information about the work of other research groups. In standard cases, $\mathbf{F}_t$ will often simply coincide with the data observed up to and including time t, Y^t.

[3] An e-process is a generalization of a test martingale: all e-processes that we encounter in this thesis are test martingales.

Hence, the probability that we will ever reject the null hypothesis, while data are in fact generated under the null, is bounded by α. These findings offer some very useful potential applications. No matter the stopping rule we apply in our study design (e.g., sampling until all of the research budget has been spent, sampling until a prespecified date or number of participants), the e-process can be applied in an *anytime-valid test* with type-I error guarantee at level α. The definition of the conditional e-process in equation (1.3) even allows us to look at each $E^{(t)}$ to decide whether to continue data collection for batch Y_{t+1}: we can now test each time a new data entry has become available. This is fundamentally different from methods such as *alpha spending*, where testing moments really have to be committed to in advance, and changing testing moments on the fly is a costly process [Demets and Lan, 1994].

Returning to the heights example, we could instantiate an e–value for testing the null hypothesis that the height in a population is distributed according to a normal distribution with mean 175 and an arbitrary standard deviation. We could then start calculating e-values and testing our null hypothesis as soon as we have measured the height of the first subject: after our first subject, we calculate E_1, peek if $E^{(1)} \geq 1/\alpha$, and decide if we want to continue data collection. If we move on to the second subject, we calculate E_2, peek if $E^{(2)} = E_1 \times E_2 \geq 1/\alpha$, and so forth. In chapter 2 it can be observed that in certain cases, studies can be finished a lot quicker due to this *optional stopping*.

"Good" e-values: the simple case The definitions above so far only mentioned the null hypothesis. However, of course we want e-values with good *power* ($1 - \beta$, with β the type-II error mentioned earlier) under the alternative. Taking into account the multiplicative definition of e-processes, one would at all cost want to avoid observing $E_j = 0$ under the alternative, as this would mean all further experiments would then be futile and the value of the e-process would stay zero from there on. In terms of betting, we have lost all our capital in this scenario. To avoid this, Grünwald et al. [2022a] proposed to design e-values that maximize *expected logarithmic return*, also called *growth rate*, a concept introduced by Kelly [1956].

Definition 1.5 (Growth rate optimal (GRO) (Grünwald et al. [2022a], theorem 1))**.** Let Y be a given random variable. Let Q be a distribution for Y with given mass or density function q. Grünwald et al. [2022a] show that there always exists a probability mass or density function p_0^* such that $E(Y) = q(Y)/p_0^*(Y)$ is (a) an e-value and (b) it achieves the following supremum:

$$\sup_{E \in \mathcal{E}(\mathbb{P})} \mathbb{E}_{Y \sim Q}[\log E],$$

where $\mathcal{E}_Y(\mathbf{P})$ is the set of all possible e-values for $\mathbf{P}$ that can be written as a function of the given random variable Y. We call this e-value the Q-GRO e-value.

By using the logarithm as optimality criterion, we avoid choosing e-values that can take on the value 0 (as would happen in the case where we would directly

optimize for power), as this would imply a growth rate toward minus infinity. We also have an idea of the evidence we expect to collect under the alternative if $Y_1, Y_2, ... \sim$ i.i.d. Q: $E^{(t)}$ will up to first order in the exponent converge to $\exp(t\mathbb{E}_Q[\log E_{(j)}])$ [Kelly, 1956]. More elaborate discussions on other advantages of optimizing growth rate can be found in Grünwald et al. [2022a] and Ramdas et al. [2022].

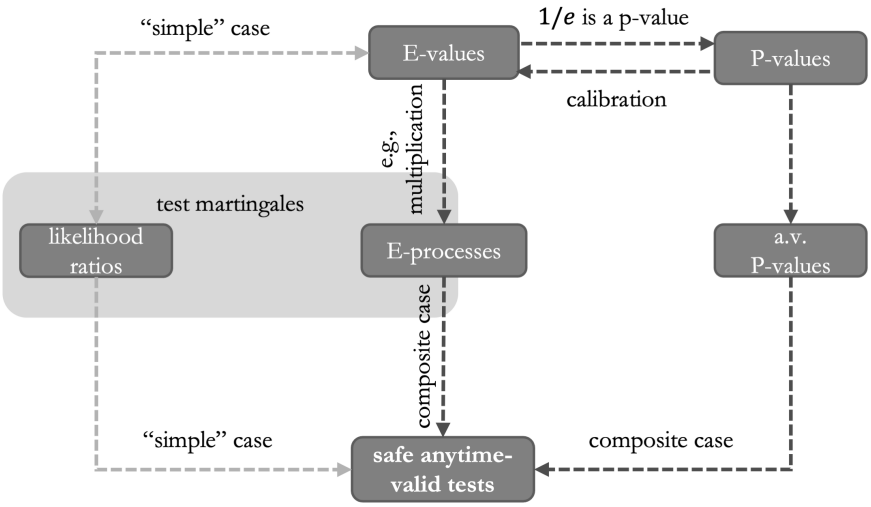


Figure 1.2: The connections between important concepts in safe anytime-valid testing. In the "simple" case where we consider a point null and alternative hypothesis, e–values and likelihood ratios are closely connected and even coincide when optimizing with respect to expected growth rate. When we consider the case where we have a composite null and/ or alternative however, simple likelihood ratios no longer provide valid sequential tests. All concepts and connections are explained in detail in the text of section 1.2.

As we already saw in equation (1.2), in case of a simple, singleton null hypothesis $\{P\}$ the likelihood ratio between any Bayes marginal distribution and P is an e-value, that can be used to build an e-process. It even turns out that in the case where we also have a simple alternative hypothesis $\mathbf{Q} = \{Q\}$, the likelihood ratio of Q and P (i.e., equation (1.2) with W_1 a point prior such that $P_{W_1} = Q$) coincides with the GRO e-value. This is also depicted schematically in figure 1.2: in the simple case, the likelihood ratio is an e-process, coincides with (a good choice of) an e-value and can be used for sequential testing. In these scenarios with a simple null, GRO e-values are closely related to and have been studied before but under different names, for example as Wald's sequential probability ratio test and in Royall's work on the universal bound for likelihood ratios [Royall, 1997].

"Good" e-values: the composite case In case of a composite null hypothesis, defining "good" e-values is not straightforward anymore. The Bayes factor $p_{W_1}(Y)/p_{W_0}(Y)$ is in general not an e-value in this case, as we do not necessarily have $\mathbb{E}_P[p_{W_1}(Y)/p_{W_0}(Y)] \leq 1$ for all $P \in \mathbf{P}$ as in equation (1.2). For an elaborate discussion on the potential use and difficulties of Bayesian statistics for anytime-valid inference, see for example De Heide and Grünwald [2021].

So far, two major distinguishable approaches toward defining e-values for composite null hypotheses have been proposed. The first one, introduced by Wasserman et al. [2020], is named *universal inference*: as its name implies, it is applicable to a wide variety of parametric and nonparametric settings. The idea is based on calculating the maximum likelihood estimator for the null distribution P_t based on all data *seen up to and including time t*. When plugging this into a likelihood ratio, one ends up with a process that is by construction dominated by other test martingales, which is then by definition an e-value at each time t and a building block of an e-process.

In this thesis, we will instead focus on the second approach, based on extending the GRO-criterion introduced above and a process called *reverse information projection* (RIPr). Restating theorem 1 by Grünwald et al. [2022a]:

Theorem 1.1 (RIPr)**.** For a given alternative distribution $\mathbf{Q} = \{Q\}$ and composite null $\mathbf{P}$ parameterized by some Θ_0, there exists a Q-GRO e-value $E(Y) = q(Y)/p_0^*(Y)$ as in definition 1.5 that uniquely can be found through *reverse information projection* of Q onto $\mathbb{P}$. That is, it satisfies:

$$\sup_{E \in \mathcal{E}(\mathbb{P})} \mathbb{E}_{Y \sim Q}[\log E] = \inf_{W_0} D(Q||P_{W_0}),$$

where the infimum is over all distributions on Θ_0, and $D(.||.)$ is the Kullback-Leibler divergence ("relative entropy").

In other words, for composite null, the Q-GRO e-value can be found through minimizing the Kullback-Leibler divergence between Q and P_{W_0} with respect to W_0. This concept can also be extended to a composite alternative: for example when a prior on Θ_1 is available, a W_1-GRO e-value can be defined. In absence of such a prior, to provide practical alternatives to the discussions on objective and subjective views on Bayesianism, an e-value could be optimized for *worst-case* GRO (for example see [Grünwald et al., 2022a] section 3, or [Turner, 2019] for an implementation), or the GRO e-value *relative* to the information we are missing about the true $Q \in \mathbb{Q}$, called *REGROW* (see [Grünwald et al., 2022a] section 4, and chapter 2 in this thesis).

Concurrently with the emerging work on e-values, there have been developments around anytime-valid p-values [Johari et al., 2022]. Interestingly, the two are in fact connected (as can be observed in figure 1.2): as stated before, $1/E$ is a conservative p-value, but p-values can also be converted into e-values by a process called *calibration* [Vovk and Wang, 2021]. This makes this e-value always substantially smaller than $1/\text{PVAL}$: this calibration comes at a cost. It is however unclear what the costs of this calibration would be for specific implementations,

and how these p/e-values would compare in terms of power to GRO e-values. Because anytime-valid p-values lack the nice combination properties of e-values in the federated setting, they are beyond the scope of this thesis.

Applications of e-processes: federated setting and confidence sequences Especially the property that the product of e-values and e-processes again yields e-values and e-processes, with the same "safety" guarantees (type-I error guarantees), makes them interesting potential candidates for implementation in a federated learning scenario with horizontally partitioned data. Traditionally, in healthcare research, study results from separate medical centers are combined through meta-analysis. However, most of the classical meta-analysis methods are actually not valid under the "shifting" null hypothesis scenario described earlier, where decisions to perform more studies are based on peeking at other study results. A striking example of this "gold rush" is given in Ter Schure and Grünwald [2019], where it is also illustrated clearly how meta-analysis with e-values can guarantee type-I error probability control. Using e-processes for meta-analysis would even enable meta-analysis "on-the-fly": each time a new data point has become available in one of the participating centers, the global e-process value can be updated, in theory leading to much faster and robust decisions compared to classical meta-analyses [Ter Schure, 2022]. Such processes would also be ideal to implement in DHT scenarios such as in figure 1.1: e-values based on new data entries could be computed locally (in local centers or even within patients' data storage systems), with only the need to share small floating point numbers with the central algorithmic node to update the estimates used for patient recommendations.

The e–value based hypothesis tests described so far can also be extended to anytime-valid confidence sequences (CS) [Howard et al., 2021, Pace and Salvan, 2020]. These extend the definition of confidence intervals above, and can be constructed by inverting e-value-based tests for testing a whole set of null hypotheses, each for a specific value of $\delta = \phi(P)$:

$$\mathrm{CS}_t = \{\delta \in \Delta : E_\delta^{(t)} < 1/\alpha\}.$$

For example, returning to our height example one last time, we now define a set of null hypotheses, for a grid of possible mean values of the distribution of the population height. We define the corresponding set of e-values, i.e. $E_{\mu'}$ is an e-value for the null hypothesis that the data are generated by a normal distribution with mean μ' (and arbitrary standard deviation). Each time a new data point has come in we update $E_{\mu'}^{(t)}$ for every value of μ': once $E_{\mu'}^{(t)} \geq 1/\alpha$ at any t we exclude that μ' from the confidence sequence.

These confidence sequences could again easily be applied to obtain safe, anytime-valid estimations in the federated setting described in the previous paragraph: instead of one e-value, now the individual e-values for a grid of values of $\phi(P)$ are shared with a central algorithmic node. These ideas will further be explored in chapters 3 and 7.

1.3 Knowledge discovery in psychiatry: current state of the art and the potential role of machine learning

Progress made over the past decades has not been equal for all fields of medicine [Krumholz, 2014]. Especially in psychiatry, clear clinical progress has come to a halt [Dean, 2017]. Over the past century, focus has shifted from a psychoanalytical view to a more *biological* view of psychiatry, especially with the introduction of psychopharmacology, imaging techniques and genomics. The concurrent introduction of the *Diagnostic and Statistical Manual of Mental Disorders (DSM)* for classification of mental disorders strengthened this biological view: each patient should match with at least one mental disorder from the DSM, which in theory has one specific biological cause that can be treated, predicted or even prevented in some way. However, plenty of evidence suggests that this biological approach toward psychiatry has not lead to an improvement to psychiatry's global burden of disease [Dean, 2017]. Over the past decade, this has led to an emerging number of calls for paradigm shifts and transitioning to completely new, less biologically oriented, diagnostic systems.

The complex nature of psychiatry research One possible explanation for this halt in progress could be that the complex, multi-faceted nature of psychiatric pathology does not match the traditional gold-standard research methods well. Within this evidential framework, most value is put on randomized controlled clinical trials, where treatment arms are compared between homogeneous groups of patients with well-defined, well-framed syndromes [Burns et al., 2011]. As a first consequence, definitions of study populations in these trials are strict and narrow, resulting in them being not representative for the heterogeneous presentations of psychiatric illness [Lee et al., 2007]. This leads to *selection bias*, with a mismatch between study populations and the clinical population, and a discrepancy of drugs' performance in clinical trials versus performance in daily clinical practice [Hernán et al., 2004]. Second, the relatively simple statistical models used to detect treatment effects in these trials might not capture the complex interplay between mental health disorders, patient characteristics and psychotropic drugs. As per standard, most trials are analyzed with the classical p-value based null-hypothesis testing described in section 1.2, only able to capture (linear) effects on mean changes on (semi-)continuous outcome measures, such as standardized questionnaires.

Fully utilizing the EHR One very rich source of information that until recently remained vastly underused are the electronic health records: the entire corpus of data generated during routine (and, optionally, trial) clinical care. Using EHR data for developing clinical insights offers lots of potential benefits when compared to databases specifically set up for clinical trials: less information remains *hidden*, the burden on clinical staff is significantly relieved through a reduction in administration and patients' consent is easier to manage [Coorevits et al., 2013].

Already thirty years ago, the potential value of using databases for *knowledge discovery* was recognized. Knowledge discovery is described by Frawley as the dis-

covery of *patterns* among data entries in a database: once the pattern is *interesting to a user* and (probabilistically) *certain*, it is new knowledge [Frawley et al., 1992]. Unfortunately, knowledge discovery processes are not yet part of routine reflection and improvement processes at (academic) clinical institutes, perhaps because of the lack of infrastructure and appropriate algorithms as described in section 1.1. Recently, Menger and others made first steps toward adapting Frawley's ideas to and implementing them structurally at several mental health institutes throughout the Netherlands in his PhD dissertation [Menger, 2019].

Besides developments on analyzing EHR data, over the past years, incorporating smartphones and other smart devices as data sources for running algorithms to improve mental health has emerged as a new promising topic of research (for example, among many others, De Looff et al. [2019] and Susaiyah et al. [2021]). Unfortunately, these devices are not part of routine clinical care or even most clinical trials in the Netherlands, because of many technical and legal hurdles. Perhaps some of the infrastructural innovations proposed in 1.1 can contribute to future implementations, but for the exploratory and confirmatory research in this theses these types of data were not available yet.

Algorithmic learning in psychiatry Clinical applications of machine learning in psychiatry have scarcely been implemented in actual clinical practice so far. A recent review of applications for predicting in-patient violence by Parmigiani et al. [2022] highlighted that the wide variety of (often black-box) algorithms used resulted in non-intersecting sets of predictors in 8 independent studies, complicating generalizablity and interpretation of results. They especially advocate the need for large, insightful studies into learning from data. Ermers *et al.* also recognize that the black-box nature of many machine learning models could hinder adaptation in practice. They distinguish several additional potential caveats for implementing machine learning in psychiatry [Ermers et al., 2020]. Machine learning models could interfere with self-reflection and critical thinking of clinicians during the decision making process. Further, a potential demise of context could create biased models, only utilizing information that can be used for machine learning in decisions. And lastly, the ground-truth problem might hinder training well-performing models [Liang et al., 2017].

To enable learning for small groups of patients, or even individual patients, studying large groups of patients is key. However, large-scale studies into (severe) mental disorders are limited. Treatment is often divided over large-scale, highly specialized centers, and data sharing is often completely off the table to ensure patients' privacy, especially of rich data sources such as clinical notes. The e-values and safe anytime-valid effect estimation methods described in section 1.2 could potentially offer a solution: federated learning enables learning locally from psychiatry patients' data, and only require sharing the locally trained algorithms between mental health institutes.

However, e-values for complex effect estimation scenarios such as logistic (penalized) regression have not been established yet. Therefore, another method especially suitable for transparent and federated learning in the exploratory phase of research will also be studied in this thesis: Bayesian network analysis [Brig-

anti et al., 2022]. Bayesian networks flexibly offer the possibility to incorporate prior knowledge on associations and effect sizes based on earlier research. Over the past decade, Bayesian network analysis has been an emerging technique in the field of mental health, because such networks are especially suitable for modeling the complex interplay between symptoms of mental health disorders [Borsboom, 2017]. In chapter 5, an extensive introduction is given into the composition of Bayesian networks.

1.4 Chapters 2 and 3: implementations of safe, anytime-valid inference

The SAVI paradigm is still relatively novel and had, before the work in this thesis was started (2019), mainly been developed theoretically. For example, theory as described in section 1.2 about methods to define e-values with good properties for discovering evidence for an alternative hypothesis has been well-developed, but the actual development of optimal e-values, corresponding software implementations and feasibility studies for specific hypothesis testing scenarios were still lacking. To work toward integrating e-values and SAVI as a core component of common research practice, it is essential that such software and illustrations of implementations exist. Such implementations are the subject of chapters 2 and 3, and the corresponding R software is available on CRAN [Ly et al., 2022].

Setting In this thesis, GRO e-values and corresponding confidence sequences are developed for a common hypothesis testing scenario: the comparison of multiple treatments. In this scenario, multiple groups of patients (or potentially other units of analysis: the e-values presented in this paper are also relevant as an alternative to traditional *A/B testing methods*, commonly used in econometric and marketing research [Kaufmann et al., 2014]), are treated with various strategies, classically placebo versus treatment, or gold standard versus new treatment. Formally, we consider k data streams of data blocks with stream index $g \in (1, \ldots, k)$, where $Y^{(t),g} = (Y_{(1),g}, Y_{(2),g}, \ldots, Y_{(t),g})$, with a different treatment for each stream g. The outcomes in each stream are distributed according to some distribution P_{θ_g}, with $\theta_g \in \Theta$. According to the null hypothesis, the distributions of the outcomes Y coincide over the streams:

$$\mathcal{H}_0 : \theta_1 = \theta_2 = ... = \theta_k = \theta \text{ for some } \theta \in \Theta. \tag{1.5}$$

With the e-values, we can gather evidence or test to investigate whether the outcome distributions are similar under the different treatments, and with the confidence sequences, we can estimate the magnitude of the difference in outcomes (for example a mean difference or relative risk ratio) between the treatments. Because we use an e-process for the tests and confidence intervals, we can gather this evidence each time a new block of data is complete, where we have prespecified only the number of observations we are going to collect for each treatment arm in this specific block. For example, in a balanced design, we could test each time one new observation has been made for each treatment.

Contributions In chapter 2, a general definition of an e-value for the above-described hypothesis testing scenario for two or more data streams is presented. The e-value definition presented there offers a lot of flexibility, as it presents a simple analytical definition that can be implemented for arbitrary data streams. Further, it closely resembles the relative GRO e-value (see section 1.2) in some testing scenarios with a compound alternative, for example in the scenario of $k \times 2$ contingency table testing. Concisely, to construct the e-value, one makes a point estimate of the alternative distribution $\hat{Q}$ based on data seen in the data stream and/or expert knowledge, and further constructs the $\hat{Q}$-relative GRO e-value through RIPr onto $\mathbf{P}$. The faster our estimate of Q converges to the truth during data collection, the closer we get to the real GRO e-value and the more *powerful* the test. We illustrate the *power* of the sequential test based on our e-value through simulations and a comparison to classical methods in a clinical study performed previously. In chapter 3, we extend the simple e-value to anytime-valid confidence sequences. We also implemented the e-values and confidence sequences in a software package for the statistical programming language R [Ly et al., 2022].

1.5 Chapters 4 – 6: data preparation and exploratory analysis in clinical psychiatry research

Research into Bayesian network analysis of the complex interplay between symptoms in mental health disorders has really taken flight over the past decade. However, most research again concerns well-defined, homogeneous groups of patients, and uses long, standardized questionnaires, yielding models that cannot be implemented straightforwardly in routine clinical psychiatry. To work toward Bayesian networks that can be implemented in a clinical decision support process in routine practice, in the work in this thesis, we built on the previous work on exploratory and predictive analysis of existing EHR data at UMC Utrecht and Parnassia Groep by Vincent Menger to discover new, possibly causal patterns in psychiatry data using Bayesian network analysis.

To discover such patterns across heterogeneous groups of patients, first we needed to define a *treatment outcome measure* with clinical relevance for the entire spectrum of mental health disorders. Gold standards that are registered in the EHR during routine clinical care for such an outcome measure were currently lacking. However, to enable learning from the EHR for large, heterogeneous groups of patients in a retrospective manner, the information covering these treatment outcome themes needed to be extracted from free text.

Contributions In chapter 4, we define psychiatry treatment outcome measures applicable to the entire spectrum of mental health disorders through a combination of systematic review, interviews with clinical staff and qualitative analysis. We then develop an information extraction pipeline that combines rule-based and deep-learning based text mining techniques that can recover phrases regarding these outcome measures from free clinical text, and convert these retrieved texts into *scores* for each patients on the outcome topics.

In chapter 5, we combine this information extraction pipeline with data from structured (tabular) sources in the EHR to develop a Bayesian network of patient characteristics, treatment characteristics and treatment outcomes. We do this for a relatively large and heterogeneous patient population of patients who received antidepressants during an admission at UMC Utrecht or Parnassia Groep, the second line mental health institute for the entire west of the Netherlands. Patterns of associations found for specific small patient groups are clinically assessed.

In chapter 6, we look at a clinical question of a slightly different nature and investigate how incorporating expert clinical knowledge and summary statistics from other centers can improve Bayesian network analysis. This chapter focuses on modeling outcomes of electroconvulsive treatment for depressive episodes at UMC Utrecht. For this select population, plenty of clinical trial data is already available, and we set out to investigate how incorporating this data affects predictions and prediction accuracy of a Bayesian model.

1.6 Chapter 7: stratified anytime-valid effect estimation and application to a psychiatry use-case

The retrospective, exploratory findings from chapters 4 and 5 revealed interesting new patterns in the data of UMCU and PG. For patients and clinicians at these specific institutes, these patterns in itself might be of enough added value to incorporate them in decision support models. However, before these results can be generalized, confirmatory research in a prospective (perhaps even randomized) manner is essential. In chapter 7, we develop safe anytime-valid tests and confidence sequences for these kinds of settings, where we want to estimate treatment effects in data streams stratified according to one or more characteristics. To achieve this, we extend the e-values and confidence sequences developed for testing Bernoulli streams in chapters 2 and 3. We then illustrate through simulations how a prospective, federated trial design to test some of the hypotheses formulated based on chapters 4 and 5 with these tests could be planned, and how many patients we expect to include in such a design.

Setting In this chapter, we specifically focus on count data in a stratified contingency table setting. The outcomes in the streams now not only depend on their treatment, but also on certain stratification characteristics. We now (purely for simplicity) focus on the case of two treatment groups, $g \in \{a, b\}$. We will now use g to indicate the treatment groups, and from now on k stands for the number of strata $k \in 1, \ldots, K$. Outcomes in treatment group g and stratum k are Bernoulli distributed according to $P_{\theta_{g,k}}$. Under the null hypothesis, we then have:

$$\mathcal{H}_0 : \theta_{a,k} = \theta_{b,k} \text{ for all } k. \tag{1.6}$$

This is also the underlying idea of the Cochran-Mantel-Haenszel test, the classical frequentist method for analyzing stratified (count) data [Mantel and Haenszel, 1959]. Giving a clinical example: a clinical researcher might suspect that the fact whether a patient was admitted to ward A or B, and whether they were younger or

older than 65 might interact with recovery probabilities and treatment allocation, thus being a confounder in the relation between the treatment patients receive and patient recovery. For analysis, the data thus need to be stratified according to the four possible combinations of properties (the ward and the age). Under our null hypothesis, the probability of the outcome does not depend on the treatment after stratification: none of the treatments is superior with respect to the outcome measure.

The tests and confidence sequences developed in chapter 7 are again valid under optional continuation and especially apt for learning in a federated setting. Each time a data block consisting of a prespecified number of observations in both treatment arms is complete within one stratum, results can be calculated based on only that block of data and previously stored summary statistics. To compute the global e-value and confidence sequences, only the e-values corresponding to the individual data blocks have to be shared with a central computing unit.

Contributions In chapter 7, we illustrate the development of an e-value for testing (1.6). As mentioned above, the value of this e-value is computed by calculating e-values for data blocks within the separate strata separately; we show that through implementing *cross-talk* techniques from the field of machine learning the power of the e-value can be improved. In more detail: as we can see in equation (1.3) we are allowed to look back at all information that we had before we started collecting data for our current block. We use the data of all previously seen strata and determine the best *mix* of information across the strata for each stratum to determine the hyperparameters of our e-value: for example, we can share the success rate or odds of success between certain strata.

We next show that, as a substantial novelty, we can also incorporate this cross-talk to construct confidence sequences for arbitrary effect sizes for each stratum. We also show that we can combine and invert our e-values to construct confidence sequences for the minimal, maximal and mean effect, even when success rates and treatment effects are heterogeneous over strata.

1.7 The composition of this dissertation

Chapters 2 throughout 7 have all been written as stand-alone publications in scientific journals or conference proceedings and can therefore be read as self-contained papers. An overview of the papers corresponding to the chapters can be found on pages i and ii. As the work in this thesis is of multidisciplinary nature, the chapters were written for different audiences, and different background knowledge is required to read them.

Chapter 2

Generic E-Variables for Exact Sequential k-Sample Tests that allow for Optional Stopping

Rosanne J. Turner[1,2], Alexander Ly[1,3], Peter D. Grünwald[1,4]

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

Abstract

We develop e-variables for testing whether two or more data streams come from the same source or not, and more generally, whether the difference between the sources is larger than some minimal effect size. These e-variables lead to exact, nonasymptotic tests that remain safe, i.e., keep their type-I error guarantees, under flexible sampling scenarios such as optional stopping and continuation. In special cases our e-variables also have an optimal 'growth' property under the alternative. While the construction is generic, we illustrate it through the special case of $k \times 2$ contingency tables, i.e. k Bernoulli streams, allowing for the incorporation of different restrictions on the composite alternative. Comparison to p-value analysis in simulations and a real-world 2×2 contingency table example show that e-variables, through their flexibility, often allow for early stopping of data collection — thereby retaining similar power as classical methods — while also retaining the option of extending or combining data afterwards.

2.1 Introduction

We develop hypothesis tests that remain statistically valid under flexible sampling scenarios, in which one is allowed to engage in optional continuation and optional stopping. We focus on the setting with data coming from several groups (often: treatment(s) versus control), with the goal of testing whether the underlying distributions are all the same. We design a family of tests for this scenario based on e-variables and test martingales that preserve type-I error guarantees under optional stopping. Hence, if the level α-test is performed and the null hypothesis holds true, the probability that the null will *ever* be rejected is bounded by α. Our tests can be implemented, and are exact, for composite null and alternative hypotheses, arbitrary distributions and in combination with arbitrary divergence measures. While our e-variable construction works for general parametric models, in the practical part of this paper we restrict ourselves to sequential categorical data, i.e. Bernoulli streams, for which we provide explicit implementation details and test scenarios.

Relevance Even in this age of big data and huge models, simple tests for comparing two populations are still used as heavily as ever in clinical trials, psychological studies and so on — areas heavily plagued by the *reproducibility crisis* [Pace and Salvan, 2020]. In a by-now notorious questionnaire [John et al., 2012], more than 55% of the interviewed psychologists admitted to the practice of 'adding data until the results look good'. While classical methods lose their type-I error guarantee if one does this (an example of this is provided in Appendix S2.D of the Supplementary Material), e-variable based tests allow for it, while, due to the option of stopping early, remaining competitive in terms of sample sizes needed to obtain a desired power. We illustrate the practical advantage of our test in Section 2.7 using the recent real-world example of the SWEPIS trial which was stopped early for harm [Wennerholm et al., 2019]. Their analysis being based on a p-value (by definition designed for fixed sampling plan), the question whether there was indeed sufficient evidence available to stop early is very hard to answer, since the sampling plan was not followed, and consequently the p-value based on which they stopped the study was by definition incorrectly calculated. This also makes it very difficult to combine the test results with results from earlier or future data while keeping anything like error control. We show that with our e-variable based methodology we would have obtained sufficient evidence to stop for harm after the same number of events had occurred, because we are allowed to perform an interim analysis each time one pair of treatment and control samples have been collected. Additionally, this e-variable, even though based on a stopped trial, can be effortlessly combined with e-variables from other trials while retaining error guarantees. Also, our results are of interest beyond mere testing: the e-variables we develop in this paper can be used to obtain *anytime-valid confidence intervals* [Howard et al., 2021] that also remain valid under optional stopping [Turner and Grünwald, 2023].

In Section 2.4 and 2.5 we refine our generic test to the 2×2 and $k \times 2$ model. An advantage of focusing on this simple setting is that it is arguably the simplest

and clearest example in which there is a nuisance parameter (the proportion under the null) that does not admit a group invariance. Nuisance parameters that satisfy such an invariance (such as the variance in the t-test, or the grand mean in the two-sample t-test) are quite straightforward to turn into e-variables and test martingales via the method of maximal invariants, as explained by Grünwald et al. [2022a] and already put into practice by e.g. Robbins [1970], Lai [1976]. The present paper shows that the proportion under the null can also be handled in a clean and simple manner. As explained below, the resulting instantiated 2×2 test appears to be quite different from existing sequential and Bayesian approaches. Thus, more than 85 years after *the lady tasting tea*, we are able to still say something quite new about the age-old problem of contingency table testing.

Related Work A sequential test for the 2×2 setting has been suggested as early as 1947 by Wald (1947). Wald's test statistic can be viewed as a product of e-variables and hence his test can be modified so as to remain valid under optional stopping. Yet, as explained in Section 2.8.2, in the 2×2 setting, Wald's e-variables lack the optimality property of the ones we introduce here, and they cannot be generalized to arbitrary models or effect size notions. Other earlier approaches (e.g. [Siegmund, 2013, Section V.2] and [Johari et al., 2022]) are based on asymptotic approximations, or consider a somewhat different problem in which the null is simple [Lindon and Malek, 2022] (and then standard likelihood ratio tests [Royall, 1997] can be used). In contrast, our e-variable based tests are exact and nonasymptotic, meaning they are valid in (even the smallest) finite samples, and hold for general composite null and alternative hypotheses. e-variables also offer a lot more flexibility than traditional α-spending and group sequential methods: although these methods allow for interim looks at the data, most often at pre-specified moments, a maximum sample size still needs to be set in advance, which does not truly allow for optional stopping and optional continuation (a more elaborate comparison of the two methods can be found in Ter Schure et al. [2020, Section 1]).

In fact our tests are more closely related to, yet still different from, Bayes factor tests: in the case of simple null hypotheses, e-variable based tests coincide with Bayes factors [Grünwald et al., 2022a]. However, in the 2×2 setting the null is not simple, and while the Bayes factor is a ratio of two Bayes marginal likelihoods, our e-variables are ratios of more general, 'prequential' [Dawid, 1984] likelihood ratios. In some special cases, the numerator is still a Bayes marginal likelihood, but the denominator, in the 2×2 setting, almost never is (Section 2.3.2) . Thus, while similar in 'look', our approach is in the end quite different from the default Bayes factors for tests of two proportions that were proposed by Kass and Vaidyanathan [1992] and by Jamil et al. [2017], the latter based on early work by Gunel and Dickey [1974]. To illustrate, in Appendix S2.C (Supplementary Material) we show that none of the variants of the Gunel-Dickey Bayes factor that are applicable in our set-up yield valid e-variables (are anytime-valid).

Another recent approach that bears some similarity to ours are the two-sample tests from Manole and Ramdas [2023], Shekhar and Ramdas [2021]. They focus on a nonparametric setting and their test martingales satisfy optimality properties

as the sample size gets large. Instead, we focus on the parametric case and, for this case, manage to derive e-variables that are equal to or closely approximate to "optimal" (see section 2.2.2) e-variables, thus optimizing for the small-sample case (in principle, our tests could be used in a nonparametric setting as well, but since they rely on using a prior on the alternative, the test martingales of Manole and Ramdas [2023], Shekhar and Ramdas [2021] might be easier to use in that case). Another general nonparametric two-sample approach with a sequential flavor, but without optional stopping error guarantees, is Lhéritier and Cazals [2018].

Contents In section 2.2 we formally introduce the notation used throughout this paper and restate the concepts of e-variables, optional stopping and the Growth Rate Optimality (GRO) criterion, GRO being the analogue of 'optimal power' in our optional continuation setting. In Section 2.3 we propose our generic e-variable for tests of two streams in general and investigate when it has the GRO property. In Section 2.4 and 2.5 we specifically show how these general e-variables can be applied in the setting of a test of two proportions, with and without restrictions on the alternative hypothesis. In Sections 2.6 and 2.7 we provide, through simulations and a real-world example, comparisons of various e-variables and Fisher's exact test with respect to GRO and power. In Section 2.8 we compare our generic approach to other e-variables one might define for this problem, including the ones based on Wald's test. We end with a conclusion. All proofs are in the Supplementary Material.

2.2 Setup, notation and preliminaries

In this section we describe our setup and notation in detail, and cover the necessary preliminaries from the theory of safe anytime-valid inference with e-variables. We refer to Ramdas et al. [2022], Grünwald et al. [2022a], Shafer et al. [2021], respectively, for an extensive introduction to this theory, to the use of e-variables in 'optional continuation' over several studies in particular, and to their enlightening betting interpretation.

2.2.1 Setup

Suppose we collect samples from two distinct groups, denoted a and b. In both groups, data are i.i.d. and come in sequentially — even though, as explained underneath (2.2) below, our approach can also be fruitfully used in the fixed design case. We thus have two data streams, $Y_{1,a}, Y_{2,a}, \ldots$ i.i.d. $\sim P_{\theta_a}$ and $Y_{1,b}, Y_{2,b}, \ldots$ i.i.d. $\sim P_{\theta_b}$ with $\theta_a, \theta_b \in \Theta$, $\{P_\theta : \theta \in \Theta\}$ representing some parameterized underlying family of distributions, all assumed to have a probability density or mass function denoted by p_θ on some outcome space $\mathcal{Y}$. We will use notation $P_{(\theta_a,\theta_b)}$ (density $p_{(\theta_a,\theta_b)}$) to represent the joint distribution of both streams. Since it considerably simplifies notation and treatment, we focus on two-sample tests throughout the paper, pointing out at the relevant places how to extend our results to the k-sample setting for $k > 2$. We further assume that all streams are mutually fully independent, so that (returning to $k = 2$), the (marginal) probability of the

first $t = t_a + t_b$ outcomes, given that t_a of these are in group a and t_b in group b, and writing $y^t = (y_1, \ldots, y_t)$, is given by the probability density (or mass function)

$$p_{\theta_a,\theta_b}(y_a^{t_a}, y_b^{t_b}) \coloneqq p_{\theta_a}(y_a^{t_a}) p_{\theta_b}(y_b^{t_b}) = \prod_{t=1}^{t_a} p_{\theta_a}(y_{t,a}) \prod_{t=1}^{t_b} p_{\theta_b}(y_{t,b}). \tag{2.1}$$

To indicate that random vector $(Y_a^{t_a}, Y_b^{t_b}) \coloneqq (Y_{1,a} \ldots, Y_{t_a,a}, Y_{1,b}, \ldots, Y_{t_b,b})$ has a distribution represented by (2.1) we write '$Y_a^{t_a}, Y_b^{t_b} \sim P_{\theta_a^*,\theta_b^*}$'. According to the *null hypothesis* $\mathcal{H}_0 = \{P_{\theta_a,\theta_b} : (\theta_a, \theta_b) \in \Theta_0\}$, $\Theta_0 = \{(\theta, \theta) : \theta \in \Theta\}$, both processes coincide. Thus, we have that $\theta_a^* = \theta_b^* = \theta_0$ for some $\theta_0 \in \Theta$ and then the density of data $y_a^{t_a}, y_b^{t_b}$ is given by $p_{\theta_0}(y_{1,a}, \ldots, y_{t_a,a}, y_{1,b}, \ldots, y_{t_b,b})$. The alternative $\mathcal{H}_1$ expresses that $d(\theta_a, \theta_b) > \delta$ for some divergence measure d and some effect size $\delta \geq 0$.

To enable sequential application of our e-variables, we define a block $Y_{(j)}$ as a set of data consisting of n_a outcomes in group a and n_b outcomes in group b, for some pre-specified n_a and n_b. The n_a and n_b used for the j-th block $Y_{(j)}$ are allowed to depend on past data, but they must be fixed before the first observation in block j occurs (this rule can be loosened to some extent, see Section 2.3.1 and Appendix S2.E). A classical paired one-sample test corresponds to the special case with $n_a = n_b = 1$ and data coming in in the order $a, b, a, b, \ldots$.

2.2.2 e-variables and test martingales

While to some extent going back as far as Darling and Robbins [1967], interest in e-variables has exploded only very recently [Howard et al., 2021, Ramdas et al., 2020, Vovk and Wang, 2021, Shafer et al., 2021, Grünwald et al., 2022a, Pace and Salvan, 2020, Manole and Ramdas, 2023, Henzi and Ziegel, 2022]. In its simplest form, an *e-variable* is a nonnegative random variable S such that under all distributions P in the null hypothesis,

$$\mathbf{E}_P[S] \leq 1. \tag{2.2}$$

We use the term e-value for the realized value of S, analogously to its classical counterpart, the p-value. Our test works by first designing e-variables for a *single block* of data, and then later extending these to sequences of blocks $Y_{(1)}, Y_{(2)}, \ldots$ by multiplication. At each point in time, the running product of block e-values observed so far is itself an e-variable, and the random process of the products is known as a *test martingale*:

Definition 2.1. Let $\{Y_{(j)}\}_{j \in \mathbf{N}}$, with all $Y_{(j)}$ taking values in some set $\mathcal{Y}$, represent a discrete-time random process. Let $\mathcal{H}_0$ be a collection of distributions for the process $\{Y_{(j)}\}_{j \in \mathbf{N}}$. For all $j \in \mathbf{N}$, let $S_{(j)}$ be a non-negative random variable that is adapted to $\sigma(Y^{(j)})$, with $Y^{(j)} = (Y_{(1)}, \ldots, Y_{(j)})$, i.e. there exists a function s such that $S_{(j)} = s(Y^{(j)})$.

1. We say that $S_{(j)}$ is an e-variable for $Y_{(j)}$ conditionally on $Y^{(j-1)}$ if for all

$P \in \mathcal{H}_0$,

$$\mathbf{E}_P\left[S_{(j)} \mid Y_{(1)}, \ldots, Y_{(j-1)}\right] \leq 1. \tag{2.3}$$

That is, for each $y^{(j-1)} \in \mathcal{Y}^{j-1}$, all $P_0 \in \mathcal{H}_0$, (2.2) holds with $S = s(y_{(1)}, \ldots, y_{(j-1)}, Y_{(j)})$ and P set to $P_0 \mid Y^{(j-1)} = y^{(j-1)}$.

2. If, for each j, $S_{(j)}$ is an e-variable conditional on $Y_{(1)}, \ldots, Y_{(j-1)}$, then we call the process $\{S_{(j)}\}_{j \in \mathbf{N}}$ a *sequential e-variable process* relative to the given $\mathcal{H}_0$ and $\{Y_{(j)}\}_{j \in \mathbf{N}}$ and we call $\{S^{(m)}\}_{m \in \mathbf{N}}$ with $S^{(m)} = \prod_{j=1}^{m} S_{(j)}$ the corresponding *test martingale*.

Henceforth, we omit the phrase 'relative to $\mathcal{H}_0$ and $\{Y_{(j)}\}_{j \in \mathbf{N}}$' whenever it is clear from the context. By the tower property of conditional expectation, one verifies that for any process of conditional e-variables $\{S_{(j)}\}_{j \in \mathbf{N}}$, we have for all m that the product $S^{(m)}$ is itself an 'unconditional' e-variable as in (2.2), i.e. $\mathbf{E}_P[S^{(m)}] \leq 1$ for all $P \in \mathcal{H}_0$. Definition 2.1 adapts and slightly modifies terminology from [Ramdas et al., 2022, Shafer et al., 2011].

Safety The interest in e-variables and test martingales derives from the fact that we have type-I error control irrespective of the stopping rule used: for any test martingale $\{S^{(j)}\}_{j \in \mathbf{N}}$, Ville's inequality [Shafer et al., 2021] tells us that, for all $0 < \alpha \leq 1$, $P \in \mathcal{H}_0$,

$$P(\text{there exists } j \text{ such that } S^{(j)} \geq 1/\alpha) \leq \alpha. \tag{2.4}$$

Thus, if we measure evidence against the null hypothesis after observing j data units by $S^{(j)}$, and we reject the null hypothesis if $S^{(j)} \geq 1/\alpha$, then our type-I error will be bounded by α, no matter what stopping rule we used for determining j. We thus have type-I error control even if we use the most aggressive stopping rule compatible with this scenario, where we stop at the first j at which $S^{(j)} \geq 1/\alpha$ (or we run out of data, or money to generate new data). We also have type-I error control if the actual stopping rule is unknown to us, or determined by external factors independent of the data $Y_{(j)}$. We will call any test based on $\{S^{(j)}\}_{j \in \mathbf{N}}$ and a (potentially unknown) stopping time τ that, after stopping, rejects iff $S^{(\tau)} \geq 1/\alpha$ a *level α-test that is safe under optional stopping*, or simply a *safe test*.

GRO-Optimality, Simple $\mathcal{H}_1$ Grünwald et al. [2022a] (in the first version of their paper put on arXiv in 2019) introduced a definition of e-variable optimality that has by now become standard. To explain it, first consider a simple $\mathcal{H}_1 = \{Q\}$ and consider

$$\mathbf{E}_Q[\log S_{(j)}] \quad ; \quad \mathbf{E}_Q[\log S^{(m)}] \tag{2.5}$$

where $S_{(j)}$ and $S^{(m)}$ are e–variables (i.e. non-negative random variables satisfying (2.2)) that, respectively, can be written as a function of $Y_{(j)}$ and $Y^{(m)} = (Y_{(1)}, \ldots, Y_{(m)})$. The e-variable which maximizes the quantity on the left among all e-variables that can be written as a function of $Y_{(j)}$, assuming it exists, is called the *Growth Rate Optimal* e-variable for $Y_{(j)}$ relative to Q, or simply 'Q-GRO for

$Y_{(j)}$', and denoted as $S_{\text{GRO}(Q),(j)}$. Similarly, the e-variable maximizing the quantity on the right, among all e-variables that can be written as function of $Y^{(m)}$, is called Q-GRO for $Y^{(m)}$. Grünwald et al. [2022a], Shafer et al. [2021], Ramdas et al. [2022] explain why the logarithm is the appropriate function to use here.

In 'nice' cases, the Q-GRO e-variable for m outcomes can be obtained by multiplying the individual Q-GRO e-variables:

Proposition 1. Let $\mathcal{H}_1 = \{Q\}$ be simple and $\mathcal{H}_0$ be potentially composite, and 'nondegenerate' in the sense that for some $P \in \mathcal{H}_0$, $D(Q\|P) < \infty$, $D(\cdot\|\cdot)$ denoting the KL divergence. We define the following condition, with q, p the density of Q and P, respectively:

$$\text{There exists a } P \in \mathcal{H}_0 \text{ such that } S_{(1)} = q(Y_{(1)})/p(Y_{(1)}) \text{ is an } e\text{-variable.} \quad (2.6)$$

When this condition holds, $S_{(1)} = S_{\text{GRO}(Q),(1)}$ is the Q-GRO e-variable for $Y_{(1)}$. An e-variable of this form automatically exists if $\mathcal{H}_0$ is simple. If we further assume that $Y_{(1)}, Y_{(2)}, \ldots$ are i.i.d. according to all distributions in $\mathcal{H}_0 \cup \mathcal{H}_1$, then $S^{(m)}_{\text{GRO}(Q)} = \prod_{j=1}^m S_{\text{GRO}(Q),(j)}$.

If Condition (2.6) holds and $Y_{(1)}, Y_{(2)}, \ldots$ are i.i.d. according to all distributions in $\mathcal{H}_0 \cup \mathcal{H}_1$, it thus makes sense to define the *Q-GRO test martingale* to be the test martingale $(S^{(j)}_{\text{GRO}(Q)})_{j\in\mathbf{N}}$. We will then have that $S_{\text{GRO}(Q),(j)} = s_Q(Y_{(j)})$ for a fixed function $s_Q : \mathcal{Y} \to \mathbf{R}_0^+$.

In Section 2.3 (Theorem 2.1) we develop functions s_Q (denoted $s(\cdot; n_a, n_b, \theta_a^*, \theta_b^*)$ there) for simple $\mathcal{H}_1 = \{Q\}$ so that $S_{Q,(1)} = s_Q(Y_{(1)})$ is an e–variable even though $\mathcal{H}_0$ is composite and not convex, so that Proposition 1 does not apply. Since we invariably assume the $Y_{(j)}$ are i.i.d., $S_{Q,(j)} := s_Q(Y_{(j)})$ is an e–variable as well and with $S_Q^{(m)} := \prod_{j=1}^m S_{Q,(j)}$, $(S_Q^{(m)})_{m\in\mathbf{N}}$ is a test martingale. The construction works for the general setting of two data streams discussed in the introduction, and for some special $\mathcal{H}_0$ (even though composite), the $S_{Q,(j)}$ will in fact be Q-GRO and $(S_Q^{(m)})_{m\in\mathbf{N}}$ will be the Q-GRO test martingale. These include the $\mathcal{H}_0$ that arise in the 2×2 setting, our main application. For other $\mathcal{H}_0$, the e-variables $S_{Q,(j)}$ will not necessarily have the Q-GRO-property; they are designed to have (2.5) large, but it may be even larger for other e-variables.

2.2.3 From simple to composite setting: choice of the e-variable and optimality

In case $\mathcal{H}_1$ is composite, no direct analogue of the GRO-criterion for designing e-variables exists, since it is not clear under what distribution $Q \in \mathcal{H}_1$ we should maximize (2.5). In this paper, we deal with this situation by *learning* Q from the data in a Bayesian fashion. It is now convenient to write $\mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\}$ in a parameterized manner (accordingly, henceforth we shall write θ_1-GRO e-variable instead of P_{θ_1}-GRO e–variable and $S_{\text{GRO}(\theta),(j)}$ instead of $S_{\text{GRO}(P_\theta),(j)}$). We will assume i.i.d. data, thus, if $\mathcal{H}_1$ were true, then data would be i.i.d. $\sim P_{\theta_1^*}$ for some $\theta_1^* \in \Theta_1$. Starting with a distribution W on Θ_1, i.e. a prior, at each point

in time j, we determine the Bayesian posterior $W \mid Y^{(j-1)}$ and use the Bayes predictive $P_{W|Y^{(j-1)}} := \int_{\Theta_1} P_\theta dW(\theta \mid Y^{(j-1)})$ as an estimate for the 'true' $P_{\theta_1^*}$. As is well-known, under conditions on W and $\mathcal{H}_1$ (which, if $\mathcal{H}_1$ is finite-dimensional parametric, are very mild), the posterior will concentrate around θ^* and hence $P_{W|Y^{(j-1)}}$ will resemble $P_{\theta_1^*}$ more and more, with very high probability, as more data becomes available.

At each point in time j, we use our current estimate $P_{W|Y^{(j-1)}}$ to design a conditional e-variable $S_{(j)}$. Note that even though our test depends on the choice of a prior distribution on the alternative, the choice of prior does not affect the type-I error safety guarantee, hence it is fine, even from a frequentist point of view, if such a prior is chosen based on vague prior knowledge. On an informal level, as long as $P_{W|Y^{(j-1)}}$ converges to the 'true' $P_{\theta_1^*}$, the $S_{(j)}$ will in fact also start to more and more resemble the e–variables $S_{\mathrm{GRO}(\theta_1^*),(j)}$ we designed for $\mathcal{H}_1 = \{P_{\theta_1^*}\}$ and which were designed to have a large expected growth under the 'true' $P_{\theta_1^*}$. If we had known the true $P_{\theta_1^*}$ all along, the best test martingale we could have used is $S^{(m)}_{\mathrm{GRO}(\theta_1^*)} = \prod_{j=1}^m S_{\mathrm{GRO}(\theta_1^*),(j)}$, which maximizes $\mathbf{E}_{Y^{(m)} \sim P_{\theta_1^*}}[\log S]$ over all e-variables S for $Y^{(m)}$. Assuming the convergence happens fast, we expect the following quantity to be small:

$$\mathbf{E}_{Y^{(m)} \sim P_{\theta_1^*}} \left[\log S^{(m)}_{\mathrm{GRO}(\theta_1^*)} - \log \prod_{j=1}^m S_{(j)} \right], \tag{2.7}$$

i.e., we may expect that the test martingale $\prod_{j=1}^m S_{(j)}$ grows not much slower than $S^{(m)}_{\mathrm{GRO}(\theta_1^*)}$.

2.3 Two-stream safe tests

2.3.1 A generic e-variable for 2-stream–blocks

We first consider the case in which the alternative hypothesis is simple: $\Theta_1 = \{\theta_1\}$ for some fixed $\theta_1 = (\theta_a^*, \theta_b^*) \in \Theta^2$. Consider a fixed sample size of size n, and assume that we will observe a block of n_a outcomes in group a and n_b outcomes in group b. In this case, we can define an e-variable as the likelihood ratio between $p_{\theta_a^*, \theta_b^*}$ and a carefully chosen distribution that is a product of mixtures of distributions from Θ_0: for $n_a, n_b \in \mathbf{N}$, $n := n_a + n_b$ and $y_a^{n_a} = (y_{1,a}, \ldots, y_{n_a,a}) \in \mathcal{Y}^{n_a}$ and $y_b^{n_b} = (y_{1,b}, \ldots, y_{n_b,b}) \in \mathcal{Y}^{n_b}$, we define:

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*) := \frac{p_{\theta_a^*}(y_a^{n_a})}{\prod_{i=1}^{n_a} \left(\frac{n_a}{n} p_{\theta_a^*}(y_{i,a}) + \frac{n_b}{n} p_{\theta_b^*}(y_{i,a}) \right)} \cdot \frac{p_{\theta_b^*}(y_b^{n_b})}{\prod_{i=1}^{n_b} \left(\frac{n_a}{n} p_{\theta_a^*}(y_{i,b}) + \frac{n_b}{n} p_{\theta_b^*}(y_{i,b}) \right)}. \tag{2.8}$$

Theorem 2.1. The random variable $S_{[n_a, n_b, \theta_a^*, \theta_b^*]} := s(Y_a^{n_a}, Y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*)$ is

an e-variable, i.e. we have:

$$\sup_{\theta \in \Theta} \mathbf{E}_{V^n \sim P_\theta} \left[s(V^n; n_a, n_b, \theta_a^*, \theta_b^*) \right] \leq 1.$$

Moreover, if $\{P_\theta : \theta \in \Theta\}$ is a convex set of distributions, then $S_{[n_a, n_b, \theta_a^*, \theta_b^*]}$ is the (θ_a^*, θ_b^*)-GRO e-variable: for any non-negative function s' on $\mathcal{Y}^{n_a+n_b}$ satisfying $\sup_{\theta \in \Theta} \mathbf{E}_{V^n \sim P_\theta} [s'(V^n)] \leq 1$, we have:

$$\mathbf{E}_{Y_a^{n_a}, Y_b^{n_b} \sim P_{\theta_a^*, \theta_b^*}} [\log s(Y_a^{n_a}, Y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*)] \geq \mathbf{E}_{Y_a^{n_a}, Y_b^{n_b} \sim P_{\theta_a^*, \theta_b^*}} [\log s'(Y_a^{n_a}, Y_b^{n_b})].$$

Crucially, in the second part of the theorem, we do not require convexity of $\mathcal{H}_0$, a set of distributions over $\mathcal{Y}^{n_a+n_b}$ (if $\mathcal{H}_0$ were convex, the GRO property would already follow automatically [Koolen and Grünwald, 2022]), but instead of $\{P_\theta : \theta \in \Theta\}$, a set of distributions on $\mathcal{Y}$. In the 2×2 case $\mathcal{H}_0$ is not convex, since the set of i.i.d. Bernoulli distributions over $n_a + n_b > 1$ outcomes is not convex. Nevertheless, $\{P_\theta : \theta \in \Theta\}$ is just the Bernoulli model on one outcome, which is convex, so in this setting, we get the GRO e-variable.

To illustrate, consider the basic case in which data comes in in fixed batches $Y_{(1)}, Y_{(2)}, \ldots$, with each batch $Y_{(j)} = ((Y_{(j-1)n_a+1,a}, Y_{(j-1)n_a+2,a}, \ldots, Y_{jn_a,a})$ $, (Y_{(j-1)n_b+1,b}, Y_{(j-1)n_b+2,b}, \ldots, Y_{jn_b,b}))$, having exactly n_a outcomes in group a and n_b outcomes in group b, and let $n = n_a + n_b$. This case would obtain, for example, in a sequential clinical trial in which patients come in one by one, each odd patient is given the treatment and each even patient is given the placebo. Then $n = 2$, $n_a = n_b = 1$. We may then measure the evidence against the null hypothesis by the product E variable

$$S^{(m)}_{[n_a, n_b, \theta_a^*, \theta_b^*]} := \prod_{j=1}^{m} S_{(j),[n_a, n_b, \theta_a^*, \theta_b^*]} \quad ; \quad S_{(j),[n_a, n_b, \theta_a^*, \theta_b^*]} := s(Y_{(j)}; n_a, n_b, \theta_a^*, \theta_b^*). \tag{2.9}$$

By Ville's inequality (2.4), the probability under any distribution in the null that there is an m with $S^{(m)}_{[n_a, n_b, \theta_a^*, \theta_b^*]}$ larger than $1/\alpha$, is bounded by α, hence, type-I error guarantees are preserved under optional stopping if we perform the test based on $\{S^{(m)}_{[n_a, n_b, \theta_a^*, \theta_b^*]}\}_{m \in \mathbf{N}}$ as defined underneath (2.4), as long as we stop between and not 'within' batches (if we stop within a batch, the E-variable $S^{(m)}_{[n_a, n_b, \theta_a^*, \theta_b^*]}$ is undefined).

If the data do not come in batches of equal size, we may proceed as follows. First, we need to fix some $n_a \geq 1$ and $n_b \geq 1$ of our own choice. The treatment below will give valid e-variables irrespective of our choice of n_a and n_b, but it will be seen that some choices are much more reasonable (will lead to much more evidence against the null, if the null is false) than others.

Thus, fix n_a and n_b, set $n = n_a + n_b$. At each time t, we will have observed, so far, some number t_a of outcomes in group a, and t_b in group b. Now let m_t be the largest m such that $mn_a \leq t_a$ and $mn_b \leq t_b$. Now, for $m = 1, 2, \ldots$, define

$Y_{(m)}$ as above. At any given time t, $Y_{(1)}, Y_{(2)}, \ldots, Y_{(m_t)}$ will have been observed, and there may be a number n'_j remaining observations in group $j \in \{a, b\}$ so that either $n'_a < n_a$ or $n'_b < n_b$ or both. Since the $\{Y_{(j)}\}_{j \in \mathbf{N}}$ determine a test martingale in the sense of Definition 2.1, optional stopping while preserving type-I error guarantees is then possible at any point in time t, as long as the e-variable is calculated as (2.9) above for $m = m_t$, thus ignoring the final $n'_a + n'_b$ outcomes.

How should n_a and n_b be chosen in practice? For example, consider a variation of the clinical trial setting above in which the treatment-control assignment is randomized: for each incoming patient, a fair coin is flipped to decide treatment (a) or placebo (b). Then at any given time the number of patients in group a and b will not be precisely equal, but if we choose $n_a = n_b = 1$ as above it is highly unlikely that the amount of data we have to ignore at any given time t is very large. Similarly, if G_t, the group membership of the t-th observation, is itself i.i.d. according to some distribution P^*, we might have some idea of the probability $p^*(a)$ assigned to group a; if $p^*(a) = 2/5$ (say), we would choose $n_a = 2, n_b = 3$.

We can add a significant amount of extra flexibility by allowing for variable group sizes, i.e., the chosen n_a and n_b may depend on the past. Appendix S2.E in the supplementary material describes how to do this. In this way, one can in principle *learn* $p^*(a)$ from the data, changing group sizes n_a and n_b flexibly as data come in. For simplicity, we have not followed this approach here, but all our results readily extend to this case.

Extension to k-sample streams It is entirely straightforward to extend (2.8) to the scenario where we do not compare 2, but k i.i.d. data streams. Indeed, in the supplementary material we state and prove the generalization of Theorem 2.1 to k data streams. We again consider some fixed $\vec{\theta} = (\theta_a, \theta_b, ..., \theta_k) \in \Theta^k$. The probability of the first $t = \sum_{g=1}^{k} t_g$ outcomes is now given by the density or mass function $p_{\vec{\theta}} := p_{\theta_a}(y_a^{t_a}) p_{\theta_a}(y_b^{t_b}) ... p_{\theta_k}(y_k^{t_k})$. We now need to fix the k group outcome numbers $\vec{n} := (n_a, n_b, ..., n_k)$ in advance, which allows us to define the extended e-variable as a function of the data $\vec{y}^n = (y_a^{n_a}, y_b^{n_b}, ..., y_k^{n_k})$, with $n = \sum_{g=1}^{k} n_g$ for testing the null where $\theta_a = \theta_b = ... = \theta_k$:

$$s(\vec{y}^n; \vec{n}, \vec{\theta}^*) := \prod_{g=1}^{k} \frac{p_{\theta_g^*}(y_g^{n_g})}{\prod_{i=1}^{n_g} \left(\sum_{g'=1}^{k} \frac{n_{g'}}{n} p_{\theta_{g'}^*}(y_{i,g}) \right)}. \tag{2.10}$$

This e-variable is again GRO if $\{P_\theta : \theta \in \Theta\}$ is convex. To keep notation as clear as possible, we now return to the simpler 2-sample case except for a short example of an application of this extension as a flexible and exact (non-asymptotic) alternative to the chi-square test in section 2.6.

2.3.2 The generic e-variable with Bayesian alternative

Now fix some prior W_1 with density w_1 on the alternative $\Theta_1 \subseteq \Theta^2$. We can trivially extend the definition of our generic e–variable relative to singleton (θ_a^*, θ_b^*) to an e–variable relative to arbitrary prior W_1 on (θ_a^*, θ_b^*): define $p_{W_1,a}(y) :=$

$\int p_{\theta_a}(y) dW_1(\theta_a)$, the integration being over the marginal prior distribution over θ_a, and similarly, $p_{W_{1,b}}(y) := \int p_{\theta_b}(y) dW_1(\theta_b)$. Then, as a corollary of Theorem 2.1, the following is also an e–variable:

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, W_1) := \frac{\prod_{i=1}^{n_a} p_{W_{1,a}}(y_{i,a})}{\prod_{i=1}^{n_a} \left(\frac{n_a}{n} p_{W_{1,a}}(y_{i,a}) + \frac{n_b}{n} p_{W_{1,b}}(y_{i,a})\right)} \cdot \frac{\prod_{i=1}^{n_b} p_{W_{1,b}}(y_{i,b})}{\prod_{i=1}^{n_b} \left(\frac{n_a}{n} p_{W_{1,a}}(y_{i,b}) + \frac{n_b}{n} p_{W_{1,b}}(y_{i,b})\right)}. \tag{2.11}$$

This follows from applying Theorem 2.1 with a 'meta'-set of distributions, which is possible since we made no assumptions at all on the set Θ in Theorem 2.1: we replace Θ by $\mathcal{W}(\Theta)$, the set of distributions on Θ; we replace the background set of distributions $\{p_\theta : \theta \in \Theta\}$ by the set of distributions $\{p_W : W \in \mathcal{W}(\Theta)\}$; we replace the simple $\mathcal{H}_1 = \{P_{\theta_a^*, \theta_b^*}\}$ by a 'simple' $\mathcal{H}_1' = \{P_{W_a, W_b}\}$ for some distributions W_a and W_b on Θ. Such W_1-based generic e–variables can be used to *learn* the parameters θ_a^*, θ_b^* as more data in both streams come in, and this is how we will use them in a sequential context with optional stopping. Thus, assume again that data comes in batches $Y_{(1)}, Y_{(2)}, \ldots$ with each $Y_{(j)}$ consisting of n_a outcomes in group a and n_b outcomes in group b (generalization to flexible group sizes changing in time and depending on the past as described at the end of Section 2.3.1 is straightforward). We start with some prior W_1 for the first batch $Y_{(1)}$ but we now use, for the j-th batch $Y_{(j)}$, the *Bayesian posterior* $W_1 \mid Y^{(j-1)}$ as prior to define the j-th e–variable with:

$$S^{(m)}_{[n_a, n_b, W_1]} := \prod_{j=1}^{m} S_{(j),[n_a,n_b,W_1]} \quad ; \quad S_{(j),[n_a,n_b,W_1]} := s(Y_{(j)}; n_a, n_b, W_1 | Y^{(j-1)}). \tag{2.12}$$

Again, $\{S_{(j),[n_a,n_b,W_1]}\}_{j \in \mathbf{N}}$ is a sequential e–variable process, so testing based on the corresponding test martingale is safe under optional stopping by (2.4). If data are sampled from some alternative hypothesis (θ_a^*, θ_b^*), then as data accumulates, the posterior W_1 will, with high probability, concentrate narrowly around (θ_a^*, θ_b^*) and so $S_{(j),[n_a,n_b,W_1]}$ will behave more and more similarly to the 'best' (θ_a^*, θ_b^*) e-variable. Still, with the exception of a special case we indicate below, in general we cannot expect it to be the W_1-GRO E-variable. But we are not particularly concerned by this: our experiments in Section 2.6 indicate that, at least in the 2×2 table setting, it behaves quite well in terms of power, which is often the main practical interest.

Simplification when $\{P_\theta : \theta \in \Theta\}$ is Convex and $\mathcal{Y}$ is finite Denoting $W_{1,g}|Y^{(m)}$ as the marginal posterior for θ_g, for $g \in \{a, b\}$, we can rewrite (2.12) as

$$S^{(m)}_{[n_a,n_b,W_1]} = \prod_{j=1}^{m} \frac{\prod_{i=1}^{n_a} p_{W_{1,a}|Y^{(j-1)}}(Y_{(j-1)n_a+i,a}) \prod_{i=1}^{n_b} p_{W_{1,b}|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}{\prod_{g\in\{a,b\}} \prod_{i=1}^{n_g} \left(\frac{n_a}{n} p_{W_{1,a}|Y^{(j-1)}}(Y_{(j-1)n_g+i,g}) + \frac{n_b}{n} p_{W_{1,b}|Y^{(j-1)}}(Y_{(j-1)n_g+i,g})\right)}$$

$$\overset{\text{if } \{P_\theta : \theta \in \Theta\} \text{ convex, } \mathcal{Y} \text{ finite}}{=} \prod_{j=1}^{m} \prod_{i=1}^{n_a} \frac{p_{W_{1,a}|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})}{p_{\breve{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})} \prod_{i=1}^{n_b} \frac{p_{W_{1,b}|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}{p_{\breve{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}. \tag{2.13}$$

Here we define $\breve{\theta}_0|Y^{(j-1)} \in \Theta$ s.t. $p_{\breve{\theta}_0|Y^{(j-1)}} = (n_a/n)p_{W_{1,a}|Y^{(j-1)}} + (n_b/n)p_{W_{1,b}|Y^{(j-1)}}$, the existence of $\breve{\theta}_0|Y^{(j-1)}$ being guaranteed if $\{P_\theta : \theta \in \Theta\}$ is convex and the sample space is finite (for then, by Carathéodory's Theorem [Eckhoff, 1993], for any distribution W on Θ there is a distribution W' on Θ with finite support such that $p_W = p_{W'}$, and by convexity, there is θ° such that $p_{W'} = p_{\theta^\circ}$). This rewrite will enable several additional results for such Θ.

Connection to Bayes Factors Consider W_1 such that θ_a and θ_b are independent under W_1 with marginal distributions W_a and W_b, and now further take $n_a = n_b = 1$. By basic telescoping, and using that if θ_a and θ_b are independent under the prior, they must also be independent under the posterior, we can then further rewrite (2.12) as

$$\frac{\int p_{\theta_a}(Y_a^m) dW_a(\theta_a) \int p_{\theta_b}(Y_b^m) dW_b(\theta_b)}{\prod_{j=1}^{m} \prod_{g\in\{a,b\}} \left(\frac{1}{2} p_{W_{1,a}|Y^{(j-1)}}(Y_{j,g}) + \frac{1}{2} p_{W_{1,b}|Y^{(j-1)}}(Y_{j,g})\right)} \overset{\text{if } \{P_\theta : \theta \in \Theta\} \text{ convex}}{=} \tag{2.14}$$

$$\frac{\int p_{\theta_a}(Y_a^m) dW_a(\theta_a) \int p_{\theta_b}(Y_b^m) dW_b(\theta_b)}{\prod_{j=1}^{m} \prod_{g\in\{a,b\}} p_{\breve{\theta}_0|Y^{(j-1)}}(Y_{j,g})}. \tag{2.15}$$

The equality holds if $\{P_\theta : \theta \in \Theta_0\}$ is convex and $\mathcal{Y}$ is finite so that (2.13) holds. As seen from (2.14), even without finiteness or convexity, the numerator of the generic product e-variable is now equal to the Bayesian marginal likelihood of the data based on prior W_1. Thus, in this special case (i.e. $n_a = n_b = 1$, prior independence; the derivation breaks down if these do not hold), if the denominator could also be written as a Bayes marginal likelihood, then our e-variable would really be a Bayes factor. Yet, even if $\{P_\theta : \theta \in \Theta\}$ is convex, it cannot be written in this way, though it is very 'close': each of the m factors in the denominator in (2.15) is the product density function of two identical distributions for one outcome, and Proposition 2 below shows that, in the special case of the 2×2 model with W_a and W_b independent beta priors, this distribution may itself be the Bayes predictive

distribution obtained by equipping Θ_0 with another beta prior. Still, for a real Bayes factor corresponding to $\mathcal{H}_0$, for each j, the two outcomes $Y_{j,a}, Y_{j,b}$ in the j-th block would not be independent given $Y^{(j-1)}$, whereas in (2.15) they are, so we may conclude that in general, our e-variables are not equivalent to any Bayes factor.

2.4 Safe tests for two proportions

We assume the setting above and, for now, assume that both streams are Bernoulli. This will substantially simplify the formulae. Thus, $\Theta = [0,1]$ and (2.1) now specializes to

$$\begin{aligned} p_{\theta_a,\theta_b}(y_a^{t_a}, y_b^{t_b}) &:= p_{\theta_a}(y_{1,a}, \ldots, y_{t_a,a}) p_{\theta_b}(y_{1,b}, \ldots, y_{t_b,b}) \\ &= \theta_a^{t_{a1}}(1-\theta_a)^{t_a - t_{a1}} \theta_b^{t_{b1}}(1-\theta_b)^{t_b - t_{b1}}. \end{aligned} \tag{2.16}$$

t_{a1} represents the number of outcomes 1 in stream a among the first t_a ones, and t_{b1} the number of outcomes 1 in stream b among the first t_b ones. According to the null hypothesis, we have that $\theta_a^* = \theta_b^* = \theta_0$ for some $\theta_0 \in \Theta = [0,1]$. (2.16) now simplifies to:

$$p_{\theta_0}(y_a^{t_a}, y_b^{t_b}) := \theta_0^{t_1}(1-\theta_0)^{t_0}.$$

t_1 represents the number of ones in the sequence $y^{t_a+t_b} = y_1, \ldots, y_{t_a+t_b}$, and similarly for t_0.

We now run through the results of the previous section for this instantiation of our test. Again, we start with the case of a simple $\mathcal{H}_1 = \{P_{\theta_a^*,\theta_b^*}\}$. (2.8) can now be written as:

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*) := \frac{p_{\theta_a^*}(y_a^{n_a})}{p_{\theta_0}(y_a^{n_a})} \cdot \frac{p_{\theta_b^*}(y_b^{n_b})}{p_{\theta_0}(y_b^{n_b})} \quad ; \quad \theta_0 = \frac{n_a}{n}\theta_a^* + \frac{n_b}{n}\theta_b^*. \tag{2.17}$$

Theorem 2.1 tells us that this is an e-variable. Since $\{P_\theta : \theta \in \Theta\}$, the Bernoulli model, is convex, the theorem also tells us that in this case the generic e-variable with simple alternative is always (θ_a^*, θ_b^*)-GRO.

We now turn to the generic e–variable relative to arbitrary prior W_1. For the Bernoulli model the Bayes posterior predictive distribution is itself a Bernoulli distribution, with its parameter equal to the posterior mean. Therefore, while the generic e–variable relative to prior W_1 is still given by (2.11), this now simplifies to:

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, W_1) = s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, \theta_a^*, \theta_b^*) \quad ; \quad \theta_g^* = \mathbf{E}_{\theta_g \sim W_1}[\theta_g], \ g \in \{a, b\}. \tag{2.18}$$

Combining this with (2.13) we infer that

$$S^{(m)}_{[n_a,n_b,W_1]} = \prod_{j=1}^{m} \prod_{i=1}^{n_a} \frac{p_{\check{\theta}_a|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})}{p_{\check{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_a+i,a})} \prod_{i=1}^{n_b} \frac{p_{\check{\theta}_b|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})}{p_{\check{\theta}_0|Y^{(j-1)}}(Y_{(j-1)n_b+i,b})} \tag{2.19}$$

where $\breve{\theta}_a|Y^{(j-1)} = \mathbf{E}_{\theta_a \sim W|Y^{(j-1)}}[\theta_a]$ and $\breve{\theta}_b|Y^{(j-1)} = \mathbf{E}_{\theta_b \sim W|Y^{(j-1)}}[\theta_b]$ and $\breve{\theta}_0|Y^{(j-1)} = (n_a/n)\breve{\theta}_a \mid Y^{(j-1)} + (n_b/n)\breve{\theta}_b \mid Y^{(j-1)}$.

Simplified Calculations with Independent Beta Priors Now take the special case in which θ_a and θ_b are independent under the prior W_1 with marginals W_a and W_b. In this case, θ_a and θ_b are also independent under the posterior, and we can simplify $\breve{\theta}_a|Y^{(j-1)} = \mathbf{E}_{\theta_a \sim W_a|Y_a^{(j-1)n_a}}[\theta_a]$, the expectation of θ_a under the posterior W_a given all data so far in group a, and similarly for group b. Using beta priors, this expectation is easy to calculate and we get:

Proposition 2. Let θ_a, θ_b be independent under W_1, with marginals W_a and W_b respectively. Suppose that these are beta priors with parameters (α_a, β_a) and (α_b, β_b) respectively. Then, upon defining $U_a = \sum_{i=1}^{(j-1)n_a} Y_{i,a}$, $U_b = \sum_{i=1}^{(j-1)n_b} Y_{i,b}$, $U = \sum_{i=1}^{(j-1)n}(Y_{i,a} + Y_{i,b})$ we have that $\breve{\theta}_a, \breve{\theta}_b, \breve{\theta}_0$ as above satisfy: $\breve{\theta}_a|Y^{(j-1)} = (U_a + \alpha_a)/((j-1)n_a + \alpha_a + \beta_a)$, $\breve{\theta}_b|Y^{(j-1)} = (U_b + \alpha_b)/((j-1)n_b + \alpha_b + \beta_b)$ respectively, and $\breve{\theta}_0|Y^{(j-1)}$ is as further above. In the special case that we fix the prior parameters in the groups proportional to the group size fraction $\kappa := n_b/n_a$, i.e we fix $\alpha_b = \kappa\alpha_a$, $\beta_b = \kappa\beta_a$, the expression for $\breve{\theta}_0$ simplifies to $\breve{\theta}_0|Y^{(j-1)} = (U + (1+\kappa)\alpha_a)/((j-1)n + (1+\kappa)\alpha_a + (1+\kappa)\beta_a)$.

2.5 (Un)Restricted composite $\mathcal{H}_1$ in the 2×2 setting

In this section we describe the main instantiations of the 2×2 stream testing scenario that are relevant in practice. These differ in the choice of $\mathcal{H}_1$: the choice can be fully unrestricted (we simply want to find whether there is any discrepancy from $\mathcal{H}_0$ at all); restricted in terms of effect size; or restricted because we have prior knowledge about either θ_a^* or θ_b^*. We consider each in turn, the second and third scenario in a separate subsection. Section 2.6 provides extensive numerical simulations for all three scenarios.

In the first scenario, a researcher wants to perform a *two-sided test*; they simply aim to find any discrepancy from $\mathcal{H}_0$ if it exists, with no restrictions are placed on $\mathcal{H}_1$. In this case, if we choose W_1 as independent beta priors on θ_a and θ_b, we can simply proceed as described in Proposition 2 above, taking a beta prior for simplicity. We will develop a reasonable 'default' choice for the hyper parameters by experiment in Section 2.6.

2.5.1 Dealing with Effect Sizes

In the second scenario we really want to test $\mathcal{H}_0$ against a restricted $\mathcal{H}_1$ consisting of those hypotheses that have a certain minimal *effect size* δ. This would then be a one-sided test. For example, a researcher might know that a new treatment must cure at least a certain number of patients more compared to a control treatment to provide *a clinically relevant treatment effect* δ. In this case, $\mathcal{H}_1$ could be restricted

to either of the sets $\Theta(\delta)$ or $\Theta^+(\delta)$, where

$$\Theta(\delta) = \left\{\theta \in [0,1]^2 : d(\theta) = \delta\right\} \quad ; \quad \Theta^+(\delta) = \begin{cases} \left\{\theta \in [0,1]^2 : d(\theta) \geq \delta\right\} & \text{if } \delta > 0 \\ \left\{\theta \in [0,1]^2 : d(\theta) \leq \delta\right\} & \text{if } \delta < 0, \end{cases} \tag{2.20}$$

where we set $d((\theta_a, \theta_b)) = \theta_b - \theta_a$. A second notion of effect size that often will be applicable in this sort of research is the *log odds ratio* between θ_b and θ_a, with restricted parameter space again given by (2.20) but d set to

$$d((\theta_a, \theta_b)) = \log\left(\frac{\theta_b}{1-\theta_b} \cdot \frac{1-\theta_a}{\theta_a}\right). \tag{2.21}$$

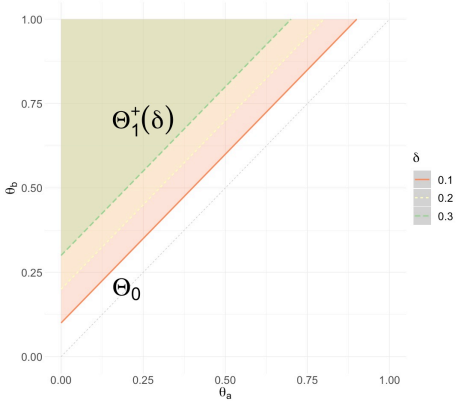
These are the two effect size notions that will feature in our experiments. An illustration of both divergence measures and the resulting restricted parameter spaces is given in Figure 2.1. A third popular notion of effect size, the relative risk, behaves, for small θ_a and $\delta > 0$, very similarly to the odds ratio, and will therefore not be separately considered in our experiments.

If we pick $\mathcal{H}_1$ restrict to $\Theta(\delta')$, then we could simply use the beta prior mentioned before with support conditioned on this set. What about the more realistic case of a $\mathcal{H}_1$ with $\delta \in \Theta^+(\delta')$? A first, intuitive (and certainly defensible) approach would be to use a prior W_1' that is spread out over $\Theta^+(\delta')$, e.g. (if $\delta' > 0$) the beta prior as above conditioned on $\delta \geq \delta'$. However, in terms of the GRO criterion, there are good reasons to still use a prior W_1^* that puts all prior mass on $\Theta(\delta')$, the boundary of the real parameter space $\Theta(\delta^+)$. Namely, for the resulting e-variable process $S^{(1)}_{[n_a,n_b,W_1^*]}, S^{(2)}_{[n_a,n_b,W_1^*]}, \ldots$, it holds for every m that

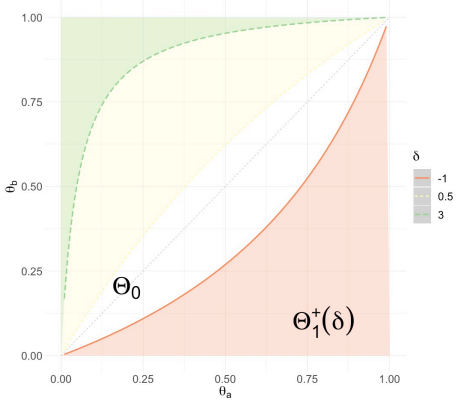
$$\text{for all } (\theta_a, \theta_b) \text{ with } d((\theta_a, \theta_b)) > \delta',\ \mathbf{E}_{Y^{(m)} \sim P_{(\theta_a,\theta_b)}}[\log S^{(m)}_{[n_a,n_b,W_1^*]}] \geq \min_{\theta \in \Theta(\delta')} \mathbf{E}_{Y^{(m)} \sim P_\theta}[\log S^{(m)}_{[n_a,n_b,W_1^*]}]. \tag{2.22}$$

Thus, we might want to use the prior W_1^* also if δ can be more extreme than δ', since if δ is actually more extreme, the expected (log-) evidence against $\mathcal{H}_0$ using W_1^* (even though designed for δ') will actually get larger anyway.

The advantage of the first approach is that it will lead to a much higher growth rate ($\mathbf{E}_{P_{(\theta_a,\theta_b)}}[\log S^{(m)}_{[n_a,n_b,W_1']}]$ much larger than $\mathbf{E}_{P_{(\theta_a,\theta_b)}}[\log S^{(m)}_{[n_a,n_b,W_1^*]}]$) if we are 'lucky' and $|d(\theta_a, \theta_b)| \gg |\delta'|$. The price to pay is that it will lead to somewhat smaller growth if $d((\theta_a, \theta_b))$ is (still arger than but) close to δ' (experiments omitted). It is easy to see why: the prior W_1' must spread out its mass over a much larger subset of $[0,1]^2$ than W_1^*. Therefore, the E-variables based on W_1' will perform somewhat worse than those based on W_1^* if the data are sampled from a point (θ_a^*, θ_b^*) in the support of W_1^*, simply because W_1^* gives much larger prior support in a neighborhood of (θ_a^*, θ_b^*). For this reason, and also because it is computationally a lot simpler, we decided to focus our experiments on the second approach rather than the first.



(a) $d((\theta_a, \theta_b)) = \theta_b - \theta_a$



(b) $d((\theta_b, \theta_a)) = \log\left[\frac{\theta_b}{(1-\theta_b)}\frac{(1-\theta_a)}{\theta_a}\right]$

Figure 2.1: Examples of restricted alternative hypothesis parameter spaces for several values of two divergence measures; the difference between group means and the log odds ratio. Θ_0 denotes the null hypothesis parameter space; $\Theta_1^+(\delta)$ the restricted alternative hypothesis parameter space.

Calculating the prior and posterior for restricted $\mathcal{H}_1$ For both notions of effect size, θ_a and θ_b can no longer be independent for any prior on $\Theta(\delta)$. Hence, the prior and posterior do not longer admit the composition in terms of beta densities as in Proposition 2. For example, when putting a prior on $\Theta(\delta)$ with the additive effect size notion, we know the new domain of θ_a would be $[0, 1 - \delta]$. θ_b is completely determined by θ_a and δ in this case. We will still use a beta prior on $\Theta(\delta)$ and calculate posteriors by a numerical approach, explained in Appendix S2.B in the Supplementary Material.

2.5.2 Working with Restrictions on event rate

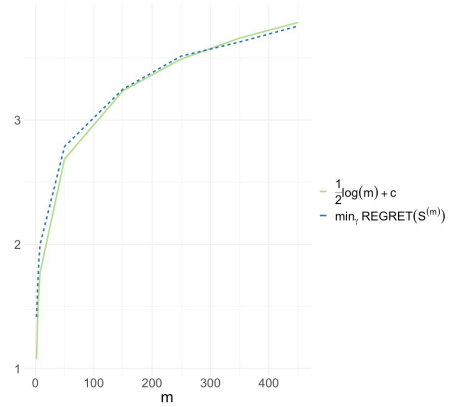
In practice, researchers often already have estimates of the occurrence rate of events in the control group in their experiments; for example, estimates of the proportion of patients that recover from a disease under standard care are known, and researchers investigate whether the proportion of recovered patients is higher in a group receiving an experimental treatment. This restriction on θ_a can be incorporated in the e-variable. This incorporation becomes especially easy if $\mathcal{H}_1$ is already restricted to a set $\Theta^+(\delta')$ with minimal relevant effect size δ'. For then $\Theta(\delta')$ contains just one point (θ_a^*, θ_b^*) (in the case of the linear effect size, this is $(\theta_a, \theta_a + \delta)$), and the e–variable constructed according to the guidelines of the previous subsection, which puts all its mass on δ' even though we allow $\delta \geq \delta'$, would be the generic e–variable corresponding to putting prior mass 1 on (θ_a^*, θ_b^*).

2.6 Illustration via simulated data

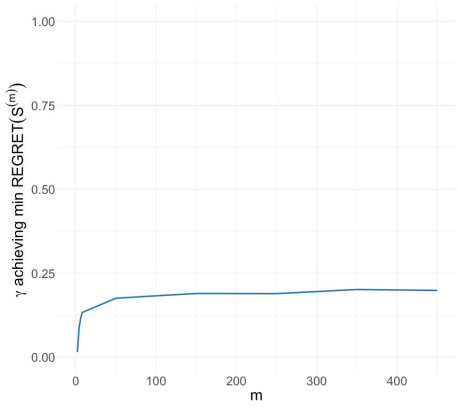
In this section, we illustrate properties of our e-variables for 2×2 application through simulated data, generated with our software package [Ly et al., 2022]. First, we determine a reasonable choice of beta prior hyper-parameter to use in (2.19) in terms of the GRO-criterion. Thereafter, we show by more simulations that our proposal for the beta prior hyper-parameter based on GRO also performs well in terms of power. Finally, we compare the power of our e-variable with this default prior choice and different restrictions on $\mathcal{H}_1$ to Fisher's exact test.

REGROW For simplicity, in all our experiments we will invariably set the beta prior hyper-parameters to $\alpha_a = \alpha_b = \beta_a = \beta_b = \gamma$ for some $\gamma > 0$ (recall that any such choice leads to a valid e-variable). We will aim for the γ that minimizes (2.7) in the worst-case over all $\theta_1^* \in [0, 1]^2$, thereby following the REGROW *(relative growth-rate optimality in worst-case)* criterion of Grünwald et al. [2022a], who give a minimax regret motivation for this choice. In essence, the prior minimizing, among all distributions over $[0, 1]^2$, the maximum of (2.7) over all θ_1^* can be viewed as the prior that allows us to learn θ_1^* as fast as possible (based on a minimal sample) in the worst-case. Here we are contented to adopt a sub-optimal but computationally convenient prior by restricting the minimum to be over a 1-dimensional family of beta priors with hyper parameter γ. We find the minimizing γ through experiments: results are depicted in Figure 2.2. It depends on the number of data blocks m, which is unknown in advance, but for large m, in

the setting with $n_a = n_b = 1$, it converges to $\gamma \approx 0.18$, and this is the value we will take as our default choice — our experiments below indicate that it remains a good choice, also when our main concern is power, and also under restrictions on $\mathcal{H}_1$.



(a) $\min_\gamma$ REGRET$S^{(m)}$



(b) $\arg\min_\gamma$ REGRET$S^{(m)}$

Figure 2.2: Minimized regret w.r.t. Beta prior hyperparameter γ for the two-sample stream e-variable for two proporions (2.18). Relative growth rate (see (2.7)) was estimated through 10000 simulations and REGRET was calculated as the maximum over θ_1^*.

Power Whereas growth rate is the natural performance measure in experiments that may always be continued at some point in the future, traditionally oriented researchers may be more interested in power. The question is then whether the optimal asymptotic choice $\gamma \approx 0.18$ in terms of the relative GRO property for unrestricted $\mathcal{H}_1$ is also the optimal choice in terms of power (which is usually considered in combination with some minimal effect size, i.e. a restricted $\mathcal{H}_1$).

The following experiment shows that by and large it is. For simplicity we only illustrate the case $n_a = n_b = 1$ and a desired power of 0.8. For various effect sizes δ, and various values of γ, we first determined the smallest sample size (number of blocks) m such that, under optional stopping up until and including m, the power is ≥ 0.8 in the worst case over all (θ_a, θ_b) with $\delta = \theta_b - \theta_a$. Here by 'optional stopping up until and including m', we mean 'we stop and reject the null iff $S^{(m')}_{[n_a,n_b,W_{[\gamma]}} > \alpha^{-1}$ for some $m' \in \{1, 2, \ldots, m\}$, and we stop and accept the null if this is not the case (so m is the maximal sample size we consider)'. We call this m the *worst-case* sample size needed for 80% power at effect size δ with prior parameter γ. The reason for calling it worst-case is that in practice, by engaging in optional stopping with a fixed maximal sample size, the *expected sample size* of this procedure is smaller: if, for $m' < m$, we already have $S^{(m')}_{[n_a,n_b,W_{[\gamma]}} > \alpha^{-1}$ then we stop and reject early; if not, we go on until we have seen m blocks and then stop (and reject iff $S^{(m)}_{[n_a,n_b,W_{[\gamma]}} > \alpha^{-1}$). We thus performed two simulation experiments: first, to estimate the worst-case sample size (at $\alpha = 0.05$), and second, to estimate the expected sample size. Again, the estimates were obtained by re-simulating a sequence of data blocks K times for a large number of K, making sure the bias and variance of the estimates were sufficiently small.

In Figure 2.3 results of these experiments are depicted. We make two observations: first, almost no difference in sample sizes to plan for between $\gamma = 0.18$ and $\gamma = 0.05$ was observed for distributions with small expected sample sizes (represented by the triangles and the dots, which overlap for most data points), and other values of γ obtained smaller power, indicating that the relative growth-optimal $\gamma = 0.18$ could in practice be used as a default setting for our e-variable — and as a consequence, we recommend it as such. Second, in the rightmost panel we see that for distributions with *very* small relative differences between θ_a and θ_b, e.g. $P_{0.5,0.58}$, values of γ higher than 0.18 yielded a higher power, whereas for such δ, the relative GROW criterion was optimized for $\gamma = 0.18$ for the corresponding (very large) stopping times in our simulation experiments. This is not surprising given what is known for simple $\mathcal{H}_0 = \{P_{\theta_0}\}$: when testing a point null θ_0 with a 1-dimensional exponential family alternative, safe tests based on Bayes factors with standard Bayesian (e.g. Gaussian or conjugate) priors do not obtain optimal power in an asymptotic sense: they reject if $|\hat{\theta} - \theta_0|^2 \gtrsim (\log n)/n$ (with $\hat{\theta}$ denoting the MLE; see the example on Z-tests by Grünwald et al. [2022a]) whereas based on nonstandard 'switching' [Van der Pas and Grünwald, 2018] or 'stitching' methods [Howard et al., 2021], corresponding to special priors with densities going to infinity as effect size goes to 0, one can get rejection if $|\hat{\theta} - \theta_0|^2 \gtrsim (\log \log n)/n$. However, there is a significant price to pay in terms of the constants hidden in the asymptotics, and in practice, 'standard' priors may very well perform better at all but very large sample sizes [Maillard, 2019]. Given that the higher γ, the more the beta prior behaves like a switch prior, we conjecture that what we see in Figure 2.3 on the right at very small δ is a version of the switching/stitching phenomenon with a composite null; since it only kicks in at very large sample sizes, we prefer $\gamma = 0.18$ as the default choice after all.

Finally, we compared the performance of our e-variables with the "default"

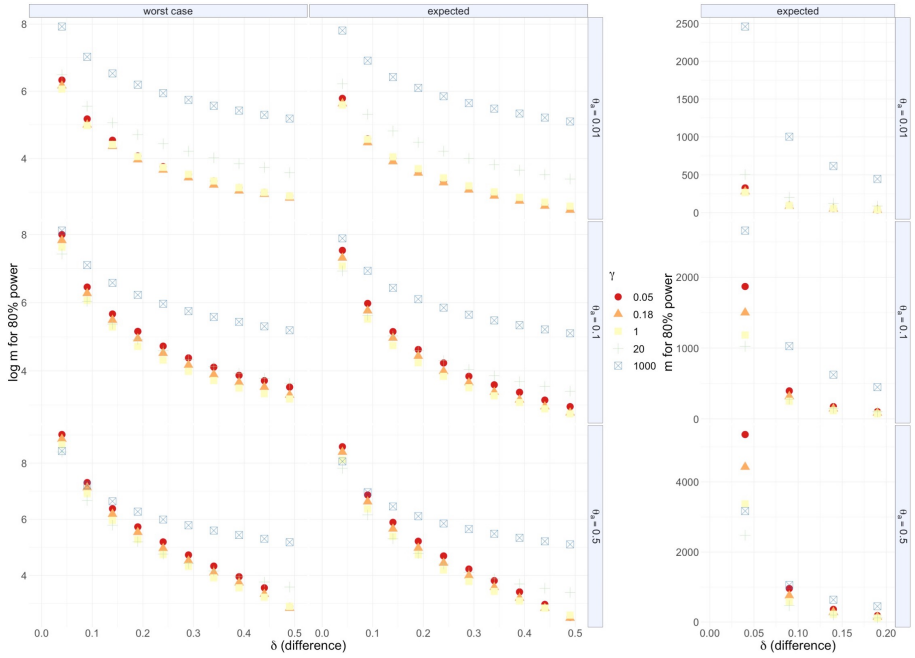


Figure 2.3: In 2000 simulations the natural logarithm, left, or identity, right, of the number of data blocks m ("sample sizes") needed for achieving 80% power while testing at $\alpha = 0.05$ for distributions with varying group means and varying differences between group means were estimated for different beta prior parameter values.

beta priors with $\gamma = 0.18$ with their classical counterpart, Fisher's exact test. We show that with Fisher's exact test, type-I error probability guarantee is lost, whereas with the e-variables it remains bounded — since these results are exactly as would be expected from the theory they have been placed in the supplementary material (Figure S2.2 in the Supplementary Material). In the main text below, we compare worst-case and expected stopping times of the e-variables with- and without restrictions on $\mathcal{H}_1$ for sample sizes one would need to plan for when analyzing experiment results with Fisher's exact test; see Figure 2.4. We noticed that the expected sample sizes achieved under optional stopping with the e-variable with unrestricted $\mathcal{H}_1$ were very similar to the sample sizes needed to plan for with Fisher's exact test. When using a correctly specified restriction on $\mathcal{H}_1$ (the leftmost data points in the second and third subfigures), this expected number of samples is even considerably lower than the sample size to plan for with Fisher's exact test. However, under misspecification, when the difference or log odds ratio used in the design of the e-variable turns out to be a lot smaller than the real difference present in the data generating machinery, one should expect to collect more samples (the data points towards the right in the second subfigure). This effect would disappear if we were to put a prior on the full $\Theta^+(\delta)$ rather than the boundary $\Theta(\delta)$, at the price of slightly worse behaviour in the well-specified case when data is sampled from $\Theta(\delta)$. Note that in Figure 2.4 we used the default beta prior parameters $\gamma = 0.18$ found optimal for the unrestricted case for the restricted cases as well; some first experiments revealed that changing the prior parameter values did not lead to significant changes in power for the restricted e-variables (results not shown). We do however offer the possibility in our software package [Ly et al., 2022] to run similar experiments for users to determine the optimal prior parameter γ for a given expected sample size and $\Theta^{(+)}(\delta')$.

Beyond Two-Stream Data: Safe Tests for k Proportions We also compared the performance of the extended version of our e-variable for k Bernoulli data streams to the corresponding classical, nonsequential counterpart, the chi-squared test [McHugh, 2013]. In this setting, we have a $k \times 2$ contingency table test, where we test whether k Bernoulli data streams come from the same source. The extension of (2.19) to k data streams analogously to (2.10) is straightforward. In simulation experiments, it was observed that our e-variable with uniform priors significantly outperforms the chi-square test for small sample sizes and large effect sizes (see Figure 2.5). For absolute differences of at least $\delta_{\max} = 0.45$, the expected sample size becomes significantly smaller than the fixed sample size needed for the chi-squared test. This is probably partially explained by the fact that the statistic used for the chi-squared test only asymptotically follows a chi-squared distribution, in contrast to our e-variable test, which is exact, valid under finite sample sizes. This means that for expected cell counts smaller than 5 the chi-square test should not be used, reflected in an increased number of samples needed for similar power [McHugh, 2013].

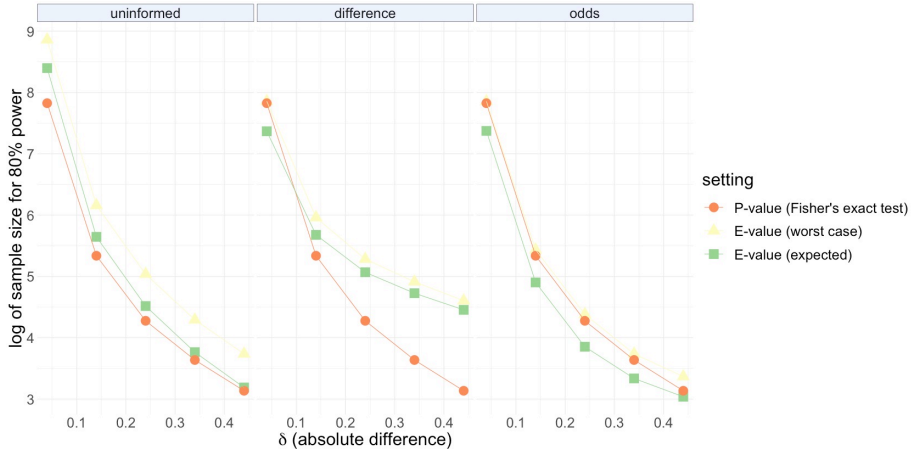


Figure 2.4: Estimates from 1000 simulations of worst-case and expected sample sizes for achieving 80% power estimated for three types of e-variables with different restrictions on $\mathcal{H}_1$, and the sample size to plan for with Fisher's exact test. Hypothesized effect sizes were 0.04 for the e-variables with prior information on the absolute difference and were converted equivalently for the log odds ratio prior information case, and we set $\gamma = 0.18$ for the beta priors.

2.7 Illustration via real world data

We will now demonstrate the approach through a real-world example: the SWEPIS study on labor induction [Wennerholm et al., 2019]. Wagenmakers and Ly [2020] have used this example before to illustrate how using single p-values to make decisions can hide valuable information in research data.

In the SWEPIS study, two groups of pregnant women were followed. In the first group labor was induced at 41 weeks, and in the second labor was induced after 42 weeks. The study was stopped early, as 6 cases of stillbirth were observed in the 42-weeks group (at $n_b = 1379$), as compared to 0 in the 41-weeks group (at $n_a = 1381$). These data yield a significant Fisher's exact test, $\text{P} \approx 0.015$, for testing that the number of stillbirths in the 42-weeks group is higher, when (wrongly) assuming that n_a and n_b were fixed in advance to the above values.

If we had used e-variables for continuously analyzing this data, would we then have found evidence for superiority of the 41 weeks approach, and would we have stopped the study earlier? As the e-variables we propose are not exchangeable, i.e., their values change under permutations of the data sequences, a direct comparison to the results of the SWEPIS study is not possible as the exact data stream is not available. To simulate a "real-time" scenario equivalent to the SWEPIS study, we assume we collect a total of 1380 data blocks, with $n_a = n_b = 1$, with a total of 2760 observations. We already know that in group a, 0 events are observed. In group b, 6 events are observed, of which we know that the last event was observed in data block 1380, directly before the study was stopped. Hence, we can simulate

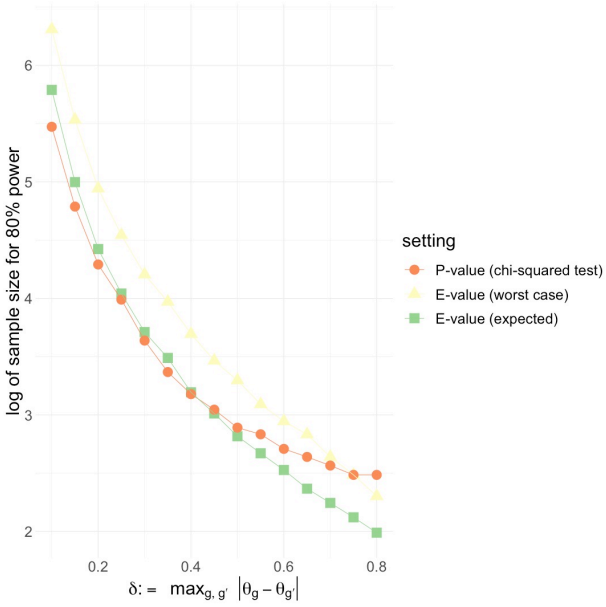


Figure 2.5: Estimates from 1000 simulations of worst-case and expected sample sizes for achieving 80% power estimated for testing with the k-stream e-variable, and the sample size to plan for with the chi-square test. Data were simulated with balanced data blocks, $\vec{n} = (1, 1, 1, 1)$ and $\vec{\theta}$ was set as an equally spaced grid from $\theta_a = 0.1$ to $\theta_k = \theta_a + \delta_{\max}$. We set $\gamma = 1$ for the beta priors.

the "real-time" data by permuting the indices of the observations in group b in the 1379 first data blocks.

Four different approaches for analyzing the data with e-variables were explored: without any restriction on $\mathcal{H}_1$, with a restriction based on the additive divergence measure (the minimal difference between the groups), with a restriction based on the log odds ratio, and with a restriction on the event rate in the control group *and* on the minimal difference. The minimal difference, log odds ratio and event rate used were chosen based on a large recent meta-analysis on stillbirths [Muglu et al., 2019]; we used $\delta = 0.00318$ as a restriction on the difference between the groups, $\log(2)$ for the log odds ratio and 0.0001 as the event rate. For all e-variables, the default beta prior hyperparameters with $\gamma = 0.18$ as earlier were used.

In Figure 2.6 the spread of the evidence collected with the four types of e-variables in 1000 simulations analogous to the SWEPIS setting is depicted. Because the observed effect size was higher than expected, e-values obtained with the (too low) restriction on the effect size were lower than the e-values obtained with the e-variable without restrictions. Adding the restriction on the event rate increased the e-values, and in all 1000 simulations, the SWEPIS study would have been stopped before the occurrence of the sixth stillbirth. Figure 2.6 also depicts results of a second simulation experiment, where we sampled 1000 data streams

from $P_{0,6/1380}$ and recorded the stopping times while analyzing the streams with the four e-variables with different restrictions on $\mathcal{H}_1$. With the e-variables without restriction, or with a restriction on the event rate and difference between the groups, we would have often stopped data collection earlier than in the SWEPIS setting.

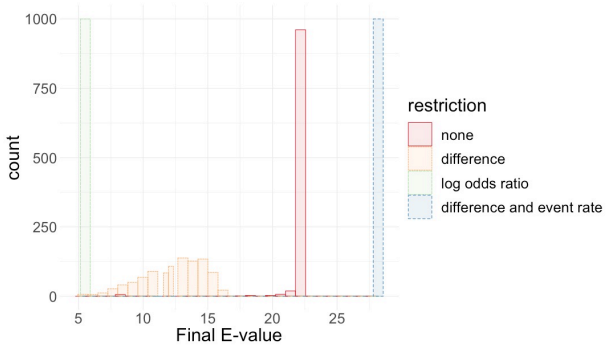
Wagemakers and Ly with their method also found evidence for the existence of a difference between the two groups, but not nearly of the same degree: they reported Bayes factors that varied, depending on the choice of the prior, between 1 and 5.4 (note that whenever we reject, our product of e-values, which like a Bayes factor can be thought of as a prequential likelihood ratio, must be ≥ 20). A possible explanation for this difference could be that the Bayes factors used for collecting evidence in their study are not designed for analyzing stream data. As we also saw in our experiments, choosing the wrong prior or restriction on $\mathcal{H}_1$ can make a large difference for the evidence collected.

We can thus conclude that, would the monitoring of the study have been performed with e-variables instead of p-values, first of all we would have collected *correct* evidence for a higher proportion of stillbirths in the 42-weeks group, and second, the degree of evidence is quite similar to that collected with the (incorrectly determined) p-value: both are significant at the 0.05 level. The study design with e-variables could effortlessly follow the classical flow of clinical trial design: before the start of the trial, a power analysis could be carried out to determine the minimum sample sizes that one needs to arrange resources for under the desired sampling scheme (balanced or unbalanced, see [Ly et al., 2022, Vignettes]). In collaboration with experts, a restriction could be put on the event rate or difference between the groups to potentially improve the power. During the study, because the SWEPIS design is balanced, an e-value is calculated each time a new patient has come in in the control and treatment groups, and the researchers and data safety monitoring boards are allowed to look at the results and decide to stop the study at any time, not affecting Type-I error probability guarantees. After the study or in case the study is stopped early because of reasons beyond rejecting the null hypothesis, because e-values were used, one can always continue a study later or combine e-values across multiple studies in an anytime-valid meta-analysis [Ter Schure, 2022].

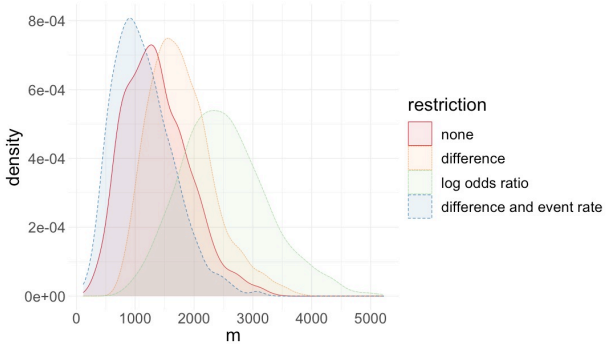
2.8 Other e-Variables for two data streams

2.8.1 The GRO e-variable for some Exponential and Location Families

The simplification (2.17) shows that in the Bernoulli case with simple $\Theta_1 = \{(\theta_a^*, \theta_b^*)\}$, we can take in our denominator p_{θ_0} with $\theta_0 = \frac{n_a}{n}\theta_a^* + \frac{n_b}{n}\theta_b^*$ — which can also be interpreted as the distribution in the null corresponding to a mixture of the means, rather than the mixture of two distributions in the null. The Bernoulli model is a special case of 1-parameter exponential families which can all be parameterized in terms of their means so that $\Theta \subset \mathbf{R}$ and $\mathbf{E}_{P_\theta}[Y] = \theta$; this is also possible for some location families that are not of exponential form.



(a) Simulated e-values in SWEPIS setting, stopping at $m = 1380$ or when $E \geq 20$



(b) Simulated stopping times in setting with continuing until $E \geq 20$

Figure 2.6: Spread of e-values and stopping times observed with safe analysis of 1000 simulations of data streams analogous to the SWEPIS scenario, with four different types of restrictions on $\mathcal{H}_1$.

This suggests that, for all such models, instead of (2.8) we might also consider the likelihood ratio (2.17). For the Bernoulli model, both definitions will coincide, but for general 1-parameter exponential families they do not since their corresponding set of densities is not convex. The question is now whether (2.17) defines an e-variable for general exponential families. It turns out that the answer is *no* in general, but *yes* in some special cases. For a negative example, consider the case with $\Theta = \mathbf{R}^+$ representing the family of exponential distributions in their mean-value parameterization, i.e. $p_\theta(y) = \lambda \exp(-\lambda y)$ with $\lambda = 1/\theta$ and take $n_a = n_b = 1$. A simple calculation shows that for any $\theta_a^* \neq \theta_b^* \in \Theta$, we have $\lim_{\theta \to \infty} \mathbf{E}_{Y_a, Y_b \text{ i.i.d.} \sim P_\theta}[p_{\theta_a^*}(Y_a) p_{\theta_b^*}(Y_b) / p_{(\theta_a^* + \theta_b^*)/2}(Y_a, Y_b)] = \infty$. The negative binomial families provide, by a similar calculation, another negative example. For a positive example, consider the case with $\Theta = \mathbf{R}$ representing the Gaussian location family with fixed variance 1 and again take $n_a = n_b = 1$. A simple calculation shows that (2.17) is equal to the likelihood ratio for testing whether the difference $Z = Y_a - Y_b$ is a Gaussian with variance $\sqrt{2}$ with either mean 0 or mean $\theta_b - \theta_a$. This is in fact the standard paired-sample Z-test that would normally be advised in this situation. In fact it is the GRO e-variable for this situation:

Proposition 3. Let $\{P_\theta : \theta \in \Theta\}$ represent a family of probability distributions with densities p_θ, with Θ a convex set in $\mathbf{R}^k$ for some $k \geq 1$. For any $\theta_a^*, \theta_b^* \in \Theta$ we have: if (2.17) is an e-variable for $\Theta_1 = \{(\theta_a^*, \theta_b^*)\}$ then it is the GRO e-variable for $\Theta_1 = \{(\theta_a^*, \theta_b^*)\}$.

The proof is immediate from Proposition 1. The proposition implies that in the special cases in which (2.17) does provide an e-variable, it is to be preferred (achieves better growth) above our original construction (2.8). (2.8) has the advantage that it provides an e-variable relative to arbitrary models. We plan to study the cases in which (2.17) can be used instead in future work.

2.8.2 The Conditional e-variable for Tests of Two Proportions

Wald [1947] proposed a 2-sample sequential probability ratio test (SPRT) for the 2×2 setting. Since SPRTs can be written in terms of products of e-variables (although products of e-variables often do not give SPRTs; see the discussion by Grünwald et al. [2022a]), let us see what e-variables Wald's test corresponds to. The setting is restricted to size-2 blocks with $n_a = n_b = 1$. We measure effect size with d the log-odds ratio (2.21) and consider an alternative with a $d(\theta_a, \theta_b)$ that is at least some given δ. Using that, for all $(\theta_a, \theta_b) \in (0,1)^2$, $z \in \{0, 1, 2\}$, the conditional probability mass function $p_{\theta_a, \theta_b}(Y_a, Y_b \mid \sum Y_a + Y_b = z)$ only depends on the log-odds ratio, we can write it, as $q_\delta(y_a, y_b | z)$ where q_δ is a probability mass function whose definition depends on (θ_a, θ_b) only via $\delta = d((\theta_a, \theta_b))$. We then take as our e-variable $S_{\text{COND},\delta} := q_\delta(Y_a, Y_b \mid Y_a + Y_b) / q_0(Y_a, Y_b \mid Y_a + Y_b)$. Since the conditional distribution $q_0(Y_a, Y_b \mid Z)$ is the same for all distributions in the null, this conditional likelihood gives an e-variable and can be used instead of our generic e-variable. Since for this Bernoulli case, our e-variable is in fact GRO, we would expect this new conditional e-variable to perform worse in terms

of GRO (and for the reasons given in Section 2.2 also in terms of the amount of data needed before one can reject at a desired power), and experiments (not reported here) confirm that it indeed performs slightly worse for δ close to 0, and substantially worse for larger δ. This is already suggested by the fact that, unlike the GRO e-variable, $S_{\text{COND},\delta}$ takes on value 1 whenever $y_a = y_b$, effectively ignoring data blocks in which both outcomes are the same. Another disadvantage is that it can only be used in combination with effect size given by the odds ratio or any monotonic transformation thereof; whereas the GRO e-variable can also be combined with the difference $\theta_b - \theta_a$ or any other desirable notion of effect size.

2.9 Conclusion

We have established e-variables and test martingales for the general i.i.d.-data streams problem. We have demonstrated, using theory, simulations and a real-world example that, for tests of two proportions, by choosing an appropriate prior on Θ_1, the method can be made competitive with classical methods that do not allow for optional stopping. Whereas in this paper, we have focused on testing, our e-variables can also be extended to get *anytime-valid confidence sequences* [Howard et al., 2021, Lai, 1976], i.e. confidence sequences for effect sizes that are valid even under optional stopping. This requires us to first extend the testing to scenarios with $\delta \geq \delta_1$ vs. $\delta \leq \delta_0$ for $\delta_0 \neq 0$, that is, null hypotheses with $\theta_a \neq \theta_b$. We have reported on this extension in Turner and Grünwald [2023]. Our work also suggests a question for future work that is practically relevant, easy to state but hard to answer: to what extent do our findings generalize to logistic regression?

Chapter 3

Exact Anytime-valid Confidence Intervals for Contingency Tables and Beyond

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

E-variables are tools for retaining type-I error guarantee with optional stopping. We extend E-variables for sequential two-sample tests to general null hypotheses and anytime-valid confidence sequences. We provide implementations for estimating risk difference, relative risk and odds-ratios in contingency tables.

3.1 Introduction

We consider a setting where we collect samples from two distinct groups, denoted a and b. In both groups, data come in sequentially and are i.i.d. We thus have two data streams, $Y_{1,a}, Y_{2,a}, \ldots$ i.i.d. $\sim P_{\theta_a}$ and $Y_{1,b}, Y_{2,b}, \ldots$ i.i.d. $\sim P_{\theta_b}$ where we assume that $\theta_a, \theta_b \in \Theta$, $\{P_\theta : \theta \in \Theta\}$ representing some parameterized underlying family of distributions, all assumed to have a probability density or mass function denoted by p_θ on some outcome space $\mathcal{Y}$.

e-variables [Grünwald et al., 2022a, Vovk and Wang, 2021] are a tool for constructing tests that keep their Type-I error control under optional stopping and continuation. Previously, Turner et al. [2021] developed e-variables for testing equality of both data streams, i.e. with null hypothesis $\vec{\Theta}_0 := \{(\theta_a, \theta_b) \in \Theta^2 : \theta_a = \theta_b\}$. Here we first generalize these e-variables to more general null hypotheses in which we may have $\theta_a \neq \theta_b$. We then use these generalized e-variables to construct *anytime-valid* confidence sequences; these provide confidence sets that remain valid under optional stopping [Darling and Robbins, 1967, Howard et al., 2021].

As in [Turner et al., 2021], we first design e-variables for a *single block* of data $(Y_a^{n_a}, Y_b^{n_b})$, where a block is a set of data consisting of n_a outcomes $Y_a^{n_a} = (Y_{a,1}, \ldots, Y_{a,n_a})$ in group a and n_b outcomes $Y_b^{n_b} = (Y_{b,1}, \ldots, Y_{b,n_b})$ in group b, for some pre-specified n_a and n_b. An *e-variable* is then, by definition, any nonnegative random variable $S = s'(Y_a^{n_a}, Y_b^{n_b})$ such that

$$\sup_{(\theta_a,\theta_b)\in\vec{\Theta}_0} \mathbf{E}_{Y_a^{n_a}\sim P_{\theta_a}, Y_b^{n_b}\sim P_{\theta_b}} [s'(Y_a^{n_a}, Y_b^{n_b})] \leq 1. \tag{3.1}$$

Turner et al. [2021] first defined such an e-variable for $\vec{\Theta}_0 = \{(\in \Theta^2 : \theta_a = \theta_b\}$ so that it would tend to have high power against a given simple alternative $\vec{\Theta}_1 = \{(\theta_a^*, \theta_b^*)\}$. Their e-variable is of the following simple form (with $n = n_a + n_b$):

$$s'(Y_a^{n_a}, Y_b^{n_b}) = \frac{p_{\theta_a^*}(Y_a^{n_a})}{\prod_{i=1}^{n_a}(\frac{n_a}{n} p_{\theta_a^*}(Y_{a,i}) + \frac{n_b}{n} p_{\theta_b^*}(Y_{a,i}))} \cdot \frac{p_{\theta_b^*}(Y_b^{n_b})}{\prod_{i=1}^{n_b}(\frac{n_a}{n} p_{\theta_a^*}(Y_{b,i}) + \frac{n_b}{n} p_{\theta_b^*}(Y_{b,i}))}. \tag{3.2}$$

These e-variables can be extended to sequences of blocks $Y_{(1)}, Y_{(2)}, \ldots$ by multiplication, and can be extended to composite alternatives by sequentially learning (θ_a^*, θ_b^*) from the data, for example via a Bayesian prior on $\vec{\Theta}_1$. The n_a and n_b used for the j-th block $Y_{(j)}$ are allowed to depend on past data, but they must be fixed before the first observation in block j occurs. For simplicity, in this note we only consider the case with n_a and n_b that remain fixed throughout; extension to the general case is straightforward.

By a general property of e-variables, at each point in time, the running product of block e-variables observed so far is itself an e-variable, and the random process of the products is known as a *test martingale* [Grünwald et al., 2022a, Shafer et al., 2021]. An e-variable-based test at level α is a test which, in combination with any stopping rule τ, reports 'reject' if and only if the product of e-values correspond-

ing to all blocks that were observed at the stopping time and have already been completed, is larger than $1/\alpha$. Such a test has a type-I error probability bounded by α irrespective of the stopping time τ that was used; see the aforementioned references for much more detailed introductions and, for example [Henzi and Ziegel, 2022], for a practical application.

In case $\{P_\theta : \theta \in \Theta\}$ is convex, the e-variable (3.2) has the so-called GRO- (*growth-rate-optimality*) property: it maximizes, over all e-variables (i.e. over all nonnegative random variables $S = s'(Y_a^{n_a}, Y_b^{n_b})$ satisfying (3.1)) the logarithmic growth rate

$$\mathbf{E}_{Y_a^{n_a} \sim P_{\theta_a^*}, Y_b^{n_b} \sim P_{\theta_b^*}} [\log S], \tag{3.3}$$

which implies that, under (θ_a^*, θ_b^*), the expected number of data points before the null can be rejected is minimized [Grünwald et al., 2022a].

Below, in Theorem 3.1 in section 3.2, which generalizes Theorem 1 in Turner et al. [2021], we extend (3.2) to the case of general null hypotheses, $\vec{\Theta}_0 \subset \Theta^2$, allowing for the case that the elements of $\vec{\Theta}_0$ have two different components, and provide a condition under which it has the GRO property. From then onwards we focus on what we call 'the 2×2 contingency table setting' in which both streams are Bernoulli, θ_j denoting the probability of 1 in group j. For this case, Theorem 3.2 gives a simplified expression for the e-variable and shows that the GRO property holds if $\vec{\Theta}_0 \subset [0,1]^2$ is convex. Then we will extend this e-variable to deal with composite $\vec{\Theta}_1$ and use this to define anytime-valid confidence sequences. We illustrate these through simulations. All proofs are in Appendix S3.A.

3.2 General Null Hypotheses

In this section, we first construct an e-variable for general null hypotheses that generalizes (3.2). We then instantiate the new result to the 2×2 case. The following development and results require $\{P_\theta : \theta \in \Theta\}$ to be 'nondegenerate' in the sense that there exists $\theta \in \Theta$ such that for all $\theta' \in \Theta$, $D(P_\theta \| P_{\theta'}) < \infty$. This mild condition holds, for example, for exponential families; we tacitly assume nondegeneracy from now on.

Our goal is thus to define an e-variable for a block of $n = n_a + n_b$ data points with n_g points in group g, $g \in \{a, b\}$. For notational convenience we define, for $\theta_a, \theta_b \in \Theta$, P_{θ_a,θ_b} as the joint distribution of $Y_a^{n_a} \sim P_{\theta_a}$ and $Y_b^{n_b} \sim P_{\theta_b}$, so that $p_{\theta_a,\theta_b}(y_a^{n_a}, y_b^{n_b}) = \prod_{i=1}^{n_a} p_{\theta_a}(y_{a,i}) \prod_{i=1}^{n_b} p_{\theta_b}(y_{b,i})$ so that we can write the null hypothesis as $\mathcal{H}_0 := \{P_{\theta_a,\theta_b} : (\theta_a, \theta_b) \in \vec{\Theta}_0\}$. Our strategy will be to first develop an e–variable for a *modified* setting in which there is only a single outcome, falling with probability n_a/n in group a and n_b/n in group b. To this end, for $\vec{\theta} = (\theta_a, \theta_b)$, we define $p'_{\vec{\theta}}(Y|a) := p_{\theta_a}(y), p'_{\vec{\theta}}(Y|b) := p_{\theta_b}(y)$, all distributions with a $'$ refering to the modified setting with just one outcome. We let $\mathcal{W}^\circ(\vec{\Theta}_0)$ be the set of all distributions on $\vec{\Theta}_0$ with finite support. For $W \in \mathcal{W}^\circ(\vec{\Theta}_0)$, we define $p'_W(Y|g) = \int p'_{\vec{\theta}}(Y|g) dW(\vec{\theta})$. We set $p'_W(y^k|g) := \prod_{i=1}^k p'_W(y_i|g)$. We further define, for given alternative $\vec{\Theta}_1 = \{(\theta_a^*, \theta_b^*)\}$, $p^\circ(\cdot|g)$, $g \in \{a, b\}$ to be, if it exists,

the conditional probability density satisfying

$$\mathbf{E}_{G\sim Q'}\mathbf{E}_{Y\sim P_{\theta^*_G}}[-\log p^\circ(Y \mid G)] = \inf_{W\in\mathcal{W}^\circ(\vec{\Theta}_0)} \mathbf{E}_{G\sim Q'}\mathbf{E}_{Y\sim P_{\theta^*_G}}[-\log p'_W(Y \mid G)] \tag{3.4}$$

with $Q'(G)$ the distribution for $G \in \{a, b\}$ with $Q'(G = a) = n_a/n$. Clearly we can rephrase (3.4) equivalently as:

$$D(Q'(G,Y)\|P^\circ(G,Y)) = \inf_{W\in\mathcal{W}^\circ(\vec{\Theta}_0)} D(Q'(G,Y)\|P'_W(G,Y)), \tag{3.5}$$

where D is the KL divergence. Here we extended the conditional distributions $P'_W(Y|G)$ and $P^\circ(Y|G)$ (corresponding to densities $p'_W(Y|G)$ and $p^\circ(Y|G)$) to a joint distribution by setting $P'_W(G,Y) := Q'(G)P'_W(Y|G)$ (and similarly for P°) and we extended $Q'(G,Y) := Q'(G)P_{\theta^*_G}(Y)$. We have now constructed a modified null hypothesis $\mathcal{H}'_0 = \{P'_{\vec{\theta}}(G,Y) : \vec{\theta} \in \vec{\Theta}_0\}$ of joint distributions for a single 'group' outcome $G \in \{a, b\}$ and 'data' outcome $Y \in \mathcal{Y}$. We let $\bar{\mathcal{H}}'_0 = \{P_W(G,Y) : W \in \mathcal{W}^\circ(\vec{\Theta}_0)\}$ be the convex hull of $\mathcal{H}'_0$.

The p° satisfying (3.5) is commonly called the *reverse information projection* of Q' onto $\bar{\mathcal{H}}'_0$. Li [1999] shows that p° always exists under our nondegeneracy condition, though in some cases it may represent a sub-distribution (integrating to strictly less than one); see [Grünwald et al., 2022a, Theorem 1] (re-stated for convenience in the supplementary material) who, building on Li's work, established a general relation between reverse information projection and e–variables. Part 1 of that theorem establishes that if the minimum in (3.4) (or (3.5)) is achieved by some $W^\circ \in \mathcal{W}^\circ$ then $p^\circ(\cdot|\cdot) = p'_{W^\circ}(\cdot|\cdot)$ and, with $\vec{\theta}^* = (\theta^*_a, \theta^*_b)$, for all $\vec{\theta} \in \vec{\Theta}_0$,

$$\mathbf{E}_{G\sim Q'}\mathbf{E}_{Y\sim P'_{\vec{\theta}}|G}\left[\frac{p'_{\vec{\theta}^*}(Y|G)}{p^\circ(Y|G)}\right] = \mathbf{E}_{G\sim Q'}\mathbf{E}_{Y\sim P'_{\vec{\theta}}|G}\left[\frac{p'_{\vec{\theta}^*}(G,Y)}{p^\circ(G,Y)}\right] \le 1. \tag{3.6}$$

This expresses that $p'_{\vec{\theta}^*}(Y|G)/p^\circ(Y|G)$ is an e-variable for our modified problem, in which within a single block we observe a single outcome in group g, with g chosen with probability n_g/n. If we were to interpret the e–variable of the modified problem as in (3.6) as a likelihood ratio for a single outcome, its corresponding likelihood ratio for a single block of data in our original problem with n_g outcomes in group g would be:

$$\begin{aligned} s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, (\theta^*_a, \theta^*_b); \vec{\Theta}_0) &:= \frac{p'_{(\theta^*_a,\theta^*_b)}(y_a^{n_a}|a)p'_{(\theta^*_a,\theta^*_b)}(y_b^{n_b}|b)}{p^\circ(y_a^{n_a}|a)p^\circ(y_b^{n_b}|b)} \\ &= \frac{p_{\theta^*_a}(y_a^{n_a})p_{\theta^*_b}(y_b^{n_b})}{p^\circ(y_a^{n_a}|a)p^\circ(y_b^{n_b}|b)}. \end{aligned} \tag{3.7}$$

The following theorem expresses that this 'extension' of the e-variable in the modified problem gives us an e-variable in our original problem:

Theorem 3.1. $S_{[n_a,n_b,\theta^*_a,\theta^*_b;\vec{\Theta}_0]} := s(Y_a^{n_a}, Y_b^{n_b}; n_a, n_b, (\theta^*_a, \theta^*_b); \vec{\Theta}_0)$ as in (3.7) is an

E-variable, i.e. with $s'(\cdot) = s(\cdot; n_a, n_b, (\theta_a^*, \theta_b^*); \vec{\Theta}_0)$, we have (3.1). Moreover, if $\mathcal{H}_0' = \{P_{\vec{\theta}}' : \vec{\theta} \in \vec{\Theta}_0\}$ (the null hypothesis for the *modified* problem) is a convex set of distributions and $\mathcal{Y}$ is finite (so that $\mathcal{H}_0' = \bar{\mathcal{H}}_0'$) and furthermore $\mathcal{H}_0'$ is compact in the weak topology, then (a) $p^\circ(\cdot|\cdot) = p'_{\vec{\theta}}(\cdot|\cdot)$ for some $\vec{\theta} \in \vec{\Theta}_0$ and (b) $S_{[n_a, n_b, \theta_a^*, \theta_b^*; \vec{\Theta}_0]}$ is the (θ_a^*, θ_b^*)-GRO e-variable for the *original* problem, maximizing (3.3) among all e-variables.

In the case that $\mathcal{H}_0'$ is not convex and compact, we do not have a simple expression for p° in general, and we may have to find it numerically by minimizing (3.4). In the 2×2 table (Bernoulli Θ) case though, there are interesting $\mathcal{H}_0$ for which the corresponding $\mathcal{H}_0'$ is convex, and we shall now see that this leads to major simplifications.

3.2.1 General Convex $\vec{\Theta}_0$ for the 2×2 contingency table

In this subsection and the next, $\{P_{\theta_a, \theta_b}\}$ refers to the 2×2 model again, with $\mathcal{Y} = \{0, 1\}$ and θ denoting the probability of 1. We now let $\vec{\Theta}_0$ be any closed convex subset of $[0, 1]^2$ that contains a point in the interior of $[0, 1]^2$. Again, note that the corresponding $\mathcal{H}_0 = \{P_{\vec{\theta}} : \vec{\theta} \in \vec{\Theta}_0\}$ need not be convex; still, $\mathcal{H}_0'$, the null hypothesis for the modified problem as defined above, must be convex if $\vec{\Theta}_0$ is convex, and this will allow us to design e-variables for such $\vec{\Theta}_0$. Let $\mathcal{H}_1 = \{P_{\theta_a^*, \theta_b^*}\}$ with (θ_a^*, θ_b^*) in the interior of $[0, 1]^2$, and let

$$\mathrm{KL}(\theta_a, \theta_b) := D(P_{\theta_a^*, \theta_b^*}(Y_a^{n_a}, Y_b^{n_b}) \| P_{\theta_a, \theta_b}(Y_a^{n_a}, Y_b^{n_b})) = \sum_{y_a^{n_a} \in \{0,1\}^{n_a}, y_b^{n_b} \in \{0,1\}^{n_b}} p_{\theta_a^*}(y_a^{n_a}) p_{\theta_b^*}(y_b^{n_b}) \log \frac{p_{\theta_a^*}(y_a^{n_a}) p_{\theta_b^*}(y_b^{n_b})}{p_{\theta_a}(y_a^{n_a}) p_{\theta_b}(y_b^{n_b})} \quad (3.8)$$

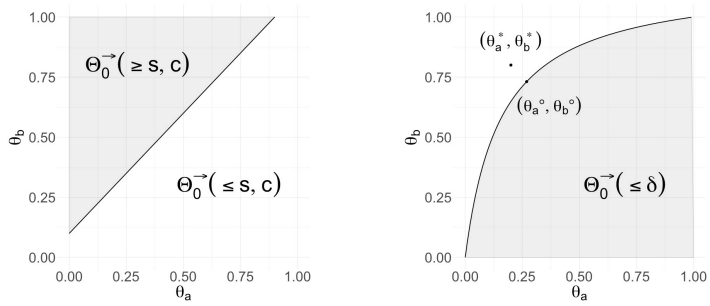
stand for the KL divergence between $P_{\theta_a^*, \theta_b^*}$ and P_{θ_a, θ_b} restricted to a single block (note that in the previous subsection, KL divergence was defined for a single outcome Y). The following result builds on Theorem 3.1:

Theorem 3.2. $\min_{(\theta_a, \theta_b) \in \vec{\Theta}_0} \mathrm{KL}(\theta_a, \theta_b)$ is uniquely achieved by some $(\theta_a^\circ, \theta_b^\circ)$. If $(\theta_a^*, \theta_b^*) \in \vec{\Theta}_0$, then $(\theta_a^\circ, \theta_b^\circ) = (\theta_a^*, \theta_b^*)$. Otherwise, $(\theta_a^\circ, \theta_b^\circ)$ lies on the boundary of $\vec{\Theta}_0$, but not on the boundary of $[0, 1]^2$. The e–variable (3.7) is given by the distribution W that puts all its mass on $(\theta_a^\circ, \theta_b^\circ)$, i.e.

$$s(y_a^{n_a}, y_b^{n_b}; n_a, n_b, (\theta_a^*, \theta_b^*); \vec{\Theta}_0) = \frac{p_{\theta_a^*}(y_a^{n_a}) p_{\theta_b^*}(y_b^{n_b})}{p_{\theta_a^\circ}(y_a^{n_a}) p_{\theta_b^\circ}(y_b^{n_b})} \quad (3.9)$$

is an e-variable. Moreover, this is the (θ_a^*, θ_b^*)-GRO e-variable relative to $\vec{\Theta}_0$.

We can extend this e-variable to the case of a composite $\mathcal{H}_1 = \{P_{\theta_a, \theta_b} : (\theta_a, \theta_b) \in \vec{\Theta}_1\}$ by *learning* the true $(\theta_a^*, \theta_b^*) \in \vec{\Theta}_1$ from the data [Turner et al., 2021]. We thus replace, for each $j = 1, 2, \ldots$, for the block $Y_{(j)}$ consisting of n_a points $Y_{(j),a,1}, \ldots, Y_{(j),a,n_a}$ in group a and n_b points $Y_{(j),b,1}, \ldots, Y_{(j),b,n_b}$ in group



(a) linear boundary (b) odds boundary

Figure 3.1: Examples of null hypothesis parameter spaces for two types of boundaries.

b, the 'true' θ_g^* for $g \in \{a, b\}$ by an estimate $\breve{\theta}_g \mid Y^{(j-1)}$ based on the previous $j-1$ data blocks. The e-variable corresponding to m blocks of data then becomes

$$S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0]} = \prod_{j=1}^{m} \prod_{i=1}^{n_a} \frac{p_{\breve{\theta}_a|Y^{(j-1)}}(Y_{(j),a,i})}{p_{\breve{\theta}_a^\circ|Y^{(j-1)}}(Y_{(j),a,i})} \prod_{i=1}^{n_b} \frac{p_{\breve{\theta}_b|Y^{(j-1)}}(Y_{(j),b,i})}{p_{\breve{\theta}_b^\circ|Y^{(j-1)}}(Y_{(j),b,i})} \tag{3.10}$$

where, for $g \in \{a, b\}$, $\breve{\theta}_g|Y^{(j-1)}$ can be an arbitrary estimator (function from $Y^{(j-1)}$ to θ_g) and $(\breve{\theta}_a^\circ \mid Y^{(j-1)}, \breve{\theta}_b^\circ \mid Y^{(j-1)})$ is defined to achieve $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0} D(P_{\breve{\theta}_a|Y^{(j-1)},\breve{\theta}_b|Y^{(j-1)}}(Y_a^{n_a}, Y_b^{n_b}) \| P_{\theta_a,\theta_b}(Y_a^{n_a}, Y_b^{n_b}))$.
No matter what estimator we choose, (3.10) gives us an e-variable. In Section 3.3, as in [Turner et al., 2021], we implement this estimator by fixing a prior W and using the Bayes posterior mean, $\breve{\theta}_g|Y^{(j-1)} := \mathbf{E}_{\theta_g \sim W|Y^{(j-1)}}[\theta_g]$. Let us now illustrate Theorem 3.2 for two choices of $\vec{\Theta}_0$.

$\vec{\Theta}_0$ with linear boundary First, we let $\vec{\Theta}_0(s, c)$, for $s \in \mathbf{R}, c \in \mathbf{R}$, stand for any straight line through $[0,1]^2$: $\vec{\Theta}_0(s, c) := \{(\,) \in [0,1]^2 : \theta_b = s + c\theta_a\}$. This can be extended to $\vec{\Theta}_0(\leq s, c) := \bigcup_{s' \leq s} \vec{\Theta}_0(s', c)$ and similarly to $\vec{\Theta}_0(\geq s, c) := \bigcup_{s' \geq s} \vec{\Theta}_0(s', c)$. For example, we could take $\vec{\Theta}_0 = \vec{\Theta}_0(s, c)$ to be the solid line in Figure 3.1(a) (which would correspond to $s = 0.1, c = 1$), or the whole area underneath the line ($\vec{\Theta}_0(\leq s, c)$) including the line itself, or the whole area above it including the line itself ($\vec{\Theta}_0(\geq s, c)$). Now consider a $\vec{\Theta}_0(s, c)$ that has nonempty intersection with the interior of $[0,1]^2$ and that is separated from the point alternative (θ_a^*, θ_b^*), i.e. $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0} \mathrm{KL}(\theta_a, \theta_b) > 0$. Utilizing the independence of the observations, we can rewrite (3.8) as follows:

$$\mathrm{KL}(\theta_a, \theta_b) := n_a \mathbf{E}_{Y \sim p_{\theta_a^*}} \left[\log \frac{p_{\theta_a^*}(Y)}{p_{\theta_a}(Y)} \right] + n_b \mathbf{E}_{Y \sim p_{\theta_b^*}} \left[\log \frac{p_{\theta_b^*}(Y)}{p_{\theta_b}(Y)} \right].$$

As we defined θ_b to be completely determined as $\theta_b = s + c\theta_a$, substituting and combining with simple differentiation w.r.t. θ_a gives that the minimum is achieved by the unique $(\theta_a^\circ, \theta_b^\circ) \in \vec{\Theta}_0$ satisfying:

$$n_a \left(-\frac{\theta_a^*}{\theta_a^\circ} + \frac{1-\theta_a^*}{1-\theta_a^\circ} \right) + n_b \cdot c \cdot \left(-\frac{\theta_b^*}{\theta_b^\circ} + \frac{1-\theta_b^*}{1-\theta_b^\circ} \right) = 0. \tag{3.11}$$

This can now be plugged into the e-variable (3.9) if the alternative is the simple alternative, or otherwise into its sequential form (3.10). In the basic case in which $\vec{\Theta}_0 = \{(^) \in [0,1]^2 : \theta_a = \theta_b\}$, the solution to (3.11) reduces to the familiar $\theta_a^\circ = \theta_b^\circ = (n_a\theta_a^* + n_b\theta_b^*)/n$ from Turner et al. [2021].

If (θ_a^*, θ_b^*) lies above the line $\vec{\Theta}_0(s,c)$, then by Theorem 3.2, $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0(\leq s,c)} \mathrm{KL}(\theta_a, \theta_b)$ must lie on $\vec{\Theta}_0(s,c)$. Theorem 3.2 gives that it must be achieved by the $(\theta_a^\circ, \theta_b^\circ)$ satisfying (3.11). Similarly, if (θ_a^*, θ_b^*) lies below the line $\vec{\Theta}_0(s,c)$, then $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0(\geq s,c)} \mathrm{KL}(\theta_a, \theta_b)$ is again achieved by the $(\theta_a^\circ, \theta_b^\circ)$ satisfying (3.11).

$\vec{\Theta}_0$ with log odds ratio boundary Similarly, we can consider $\vec{\Theta}_0(\delta)$, $\vec{\Theta}_0(\leq\delta)$, $\vec{\Theta}_0(\geq\delta)$ that correspond to a given log odds effect size δ. That is, we now take

$$\begin{aligned}
\vec{\Theta}_0(\delta) &:= \left\{ (\theta_a, \theta_b) \in [0,1]^2 : \log \frac{\theta_b(1-\theta_a)}{(1-\theta_b)\theta_a} = \delta \right\} \\
\vec{\Theta}_0(\leq\delta) &:= \left\{ (\theta_a, \theta_b) \in [0,1]^2 : \log \frac{\theta_b(1-\theta_a)}{(1-\theta_b)\theta_a} \leq \delta \right\} \\
\vec{\Theta}_0(\geq\delta) &:= \left\{ (\theta_a, \theta_b) \in [0,1]^2 : \log \frac{\theta_b(1-\theta_a)}{(1-\theta_b)\theta_a} \geq \delta \right\}.
\end{aligned}$$

For example, we could now take $\vec{\Theta}_0 = \vec{\Theta}_0(\leq\delta)$ to be the area under the curve (including the curve boundary itself) in Figure 3.1(b), which would correspond to $\delta = 2$. Now let δ and point alternative (θ_a^*, θ_b^*) be such that $\delta > 0$ and $\vec{\Theta}_0(\leq\delta)$ is separated from (θ_a^*, θ_b^*), i.e. $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0(\leq\delta)} \mathrm{KL}(\theta_a, \theta_b) > 0$. Let $(\theta_a^\circ, \theta_b^\circ) := \arg\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0(\delta)} \mathrm{KL}(\theta_a, \theta_b)$. As Figure 3.1(b) suggests, $\vec{\Theta}_0(\leq\delta)$ is convex. Theorem 3.2 now tells us that $\min_{(\theta_a,\theta_b)\in\vec{\Theta}_0(\leq\delta)} \mathrm{KL}(\theta_a, \theta_b)$ is achieved by $(\theta_a^\circ, \theta_b^\circ)$. Plugging these into (3.9) thus gives us an e-variable. $(\theta_a^\circ, \theta_b^\circ)$ can easily be determined numerically. Similarly, if $\delta < 0$, $\vec{\Theta}_0(\geq\delta)$ is convex and closed and if (θ_a^*, θ_b^*) is separated from $\vec{\Theta}_0(\geq\delta)$, the $(\theta_a^\circ, \theta_b^\circ)$ minimizing KL on $\vec{\Theta}_0(\delta)$ gives an e-variable relative to $\vec{\Theta}_0(\geq\delta)$.

3.3 Anytime-valid confidence sequences for the 2×2 case

We will now use the e-variables defined above to construct anytime-valid confidence sequences. Let $\delta = \delta(\theta_a, \theta_b)$ be a notion of effect size such as the log odds ratio (see above) or absolute risk $\theta_b - \theta_a$ or relative risk θ_b/θ_a. A $(1-\alpha)$-*anytime-valid*

(AV) confidence sequence [Darling and Robbins, 1967, Howard et al., 2021] is a sequence of random (i.e. determined by data) subsets $\mathrm{CS}_{\alpha,(1)}, \mathrm{CS}_{\alpha,(2)}, \ldots$ of Γ, with $\mathrm{CS}_{\alpha,(m)}$ being a function of the first m data blocks $Y^{(m)}$, such that for all $(\theta_a, \theta_b) \in [0,1]^2$,

$$P_{\theta_a,\theta_b}\left(\exists m \in \mathbf{N} : \delta(\theta_a,\theta_b) \notin \mathrm{CS}_{\alpha,(m)}\right) \leq \alpha.$$

We first consider the case in which for all values $\gamma \in \Gamma$ that δ can take, $\vec{\Theta}_0(\gamma) := \{(\theta_a,\theta_b) \in [0,1]^2 : \delta(\theta_a,\theta_b) = \gamma\}$ is a convex set, as it will be for absolute and relative risk. Fix a prior W_1 on $[0,1]^2$. Based on (3.10) we can make an *exact* (nonasymptotic) AV confidence sequence

$$\mathrm{CS}_{\alpha,(m)} = \left\{\delta : S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\delta)]} \leq \frac{1}{\alpha}\right\} \tag{3.12}$$

where $S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\delta)]}$ is defined as in (3.10) and is a valid e-variable by Theorem 3.2. To see that $(\mathrm{CS}_{\alpha,(m)})_{m\in\mathbf{N}}$ really is an AV confidence sequence, note that, by definition of the $\mathrm{CS}_{\alpha,(m)}$, we have
$P_{\theta_a,\theta_b}\left(\exists m \in \mathbf{N} : \delta(\theta_a,\theta_b) \notin \mathrm{CS}_{\alpha,(m)}\right)$ is given by

$$P_{\theta_a,\theta_b}\left(\exists m \in \mathbf{N} : S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\delta)]} \geq \frac{1}{\alpha}\right) \leq \alpha,$$

by Ville's inequality [Grünwald et al., 2022a, Turner et al., 2021]. Here the $\mathrm{CS}_{\alpha,(m)}$ are not necessarily intervals, but, potentially losing some information, we can make a AV confidence sequence consisting of intervals by defining $\mathrm{CI}_{\alpha,(m)}$ to be the smallest interval containing $\mathrm{CS}_{\alpha,(m)}$. We can also turn any confidence sequences $(\mathrm{CS}_{\alpha,(m)})_{m\in\mathbf{N}}$ into an alternative AV confidence sequence with sets $\mathrm{CS}'_{\alpha,(m)}$ that are always a subset of $\mathrm{CS}_{\alpha,(m)}$ by taking the *running intersection*

$$\mathrm{CS}'_{\alpha,(m)} := \bigcap_{j=1..m} \mathrm{CS}_{\alpha,(j)}.$$

In this form, the confidence sequences $\mathrm{CS}'_{\alpha,(m)}$ can be interpreted as *the set of δ's that have not yet been rejected* in a setting in which, for each null hypothesis $\vec{\Theta}_0(\delta)$ we stop and reject as soon as the corresponding e-variable exceeds $1/\alpha$. The running intersection can also be applied to the intervals $(\mathrm{CI}_{\alpha,(m)})_{m\in\mathbf{N}}$. To simplify calculations, it is useful to take W_1 a prior under which θ_a and θ_b have independent beta distributions with parameters $\alpha_a, \beta_a, \alpha_b, \beta_b$. We can, if we want, infuse some prior knowledge or hopes by setting these parameters to certain values — our confidence sequences will be valid irrespective of our choice [Howard et al., 2021]. In case no such knowledge can be formulated (as in the simulations below), we advocate the prior, which, among all priors of the simple form asymptotically achieves the REGROW criterion (a criterion related to minimax log-loss regret, see [Grünwald et al., 2022a]), i.e for the case $n_a = n_b = 1$ we set W_1 to an independent beta prior on θ_a and θ_b with $\gamma = 0.18$ as was empirically found to be the 'best'

value [Turner et al., 2021].

Log Odds Ratio Effect Size The situation is slightly trickier if we take the log odds ratio as effect size, for $\vec{\Theta}_0(\delta)$ is then not convex. Without convexity, Theorem 3.2 cannot be used and hence the validity of AV confidence sequences as constructed above breaks down. We can get nonasymptotic anytime-valid confidence sequences after all as follows. First, we consider a one-sided AV confidence sequence for the submodel of positive effect sizes $\{(\theta_a, \theta_b) : \delta(\theta_a, \theta_b) \geq 0\}$, defining

$$\mathrm{CS}^+_{\alpha,(m)} = \{\delta \geq 0 : S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\leq\delta)]} \leq \alpha^{-1}, \}$$

where we note that $\vec{\Theta}_0(\leq \delta)$ is convex (since $\delta \geq 0$) and also contains (θ_a, θ_b) with $\delta(\theta_a, \theta_b) < 0$. This confidence sequence can give a lower bound on δ. Analogously, we consider a one-sided AV confidence sequence for the submodel $\{(\theta_a, \theta_b) : \delta(\theta_a, \theta_b) \leq 0\}$, defining

$$\mathrm{CS}^-_{\alpha,(m)} = \{\delta \leq 0 : S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\geq\delta)]} \leq \alpha^{-1}\},$$

and derive an upper bound on δ. By Theorem 3.2, both sequences $(\mathrm{CS}^+_{\alpha,(m)})_{m=1,2,\ldots}$ and $(\mathrm{CS}^-_{\alpha,(m)})_{m=1,2,\ldots}$ are AV confidence sequences for the submodels with $\delta \geq 0$ and $\delta \leq 0$ respectively. Defining $\mathrm{CS}_{\alpha,(m)} = \mathrm{CS}^+_{\alpha,(m)} \cup \mathrm{CS}^-_{\alpha,(m)}$, we find, for (θ_a, θ_b) with $\delta(\theta_a, \theta_b) > 0$,

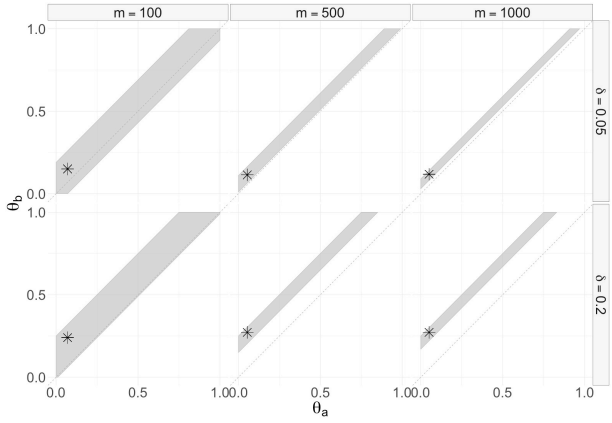
$$P_{\theta_a,\theta_b}\left(\exists m \in \mathbf{N} : \delta(\theta_a, \theta_b) \notin \mathrm{CS}_{\alpha,(m)}\right) = P_{\theta_a,\theta_b}\left(\exists m \in \mathbf{N} : \delta(\theta_a, \theta_b) \notin \mathrm{CS}^+_{\alpha,(m)}\right) \leq \alpha,$$

and analogously for (θ_a, θ_b) with $\delta(\theta_a, \theta_b) < 0$. We have thus arrived at a confidence sequence that works for all δ, positive or negative.

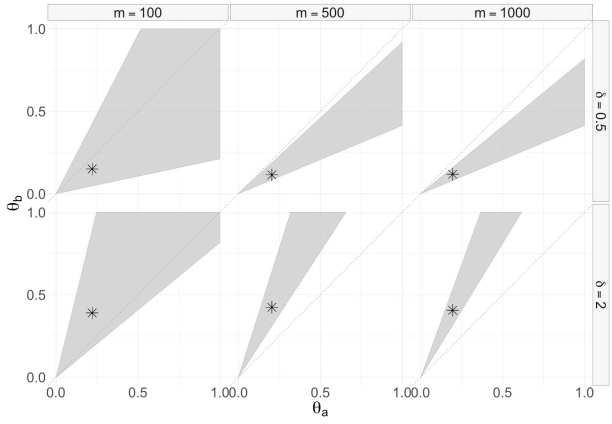
3.3.1 Simulations

In this section some numerical examples of confidence sequences for the two types of effect sizes are given. All simulations were run with code available in our software package [Ly et al., 2022].

Risk difference Risk difference is defined as the difference between success probabilities in the two streams: $\delta = \theta_b - \theta_a$. Figure 3.2 shows running intersections of confidence sequences with δ as the risk difference for simulations for various distributions and stream lengths. These sequences are constructed by testing null hypotheses based on $\vec{\Theta}_0(s, c)$, with $c = 1$ and $s = \delta$. $\mathrm{CI}_{\alpha,(m)}$ for the risk difference on $\vec{\Theta}_0$ is an interval, corresponding to the 'beam' of $(\theta_a, \theta_b) \in [0, 1]^2$ bounded by the lines $\theta_b = \theta_a + \delta_{\mathrm{L}}$ and $\theta_b = \theta_a + \delta_{\mathrm{R}}$ with $\delta_{\mathrm{L}} > \delta_{\mathrm{R}}$ being values such that $S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\delta_{\mathrm{L}})]} = S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0(\delta_{\mathrm{R}})]} = 1/\alpha$. In Appendix S3.B we illustrate the



(a) Risk difference



(b) Relative risk

Figure 3.2: Depiction of parameter space with running intersection of confidence sequence for data generated under various effect sizes, at different time points m in a data stream. The asterisks indicate the maximum likelihood estimator at that time point. The significance threshold was set to 0.05. The design was balanced, with data block sizes $n_a = 1$ and $n_b = 1$.

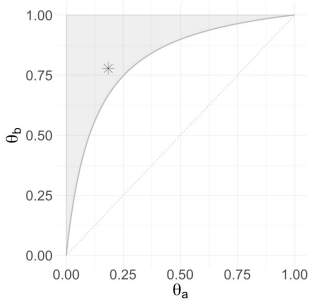
calculations leading to Figure 3.2. Figure S3.1 in the Appendix illustrates that the running intersection indeed improves the confidence sequence, albeit slightly.

Relative risk Relative risk is defined as the ratio between the success probabilities in group b and a: $\delta = \theta_b/\theta_a$. Hence, confidence sequences for this effect size measure can again be constructed using the linear boundary form $\vec{\Theta}_0(s, c)$ again, but now with $s = 0$ and $c = \delta$. Figure 3.2 shows running intersections of confidence sequences with δ as the relative risk.

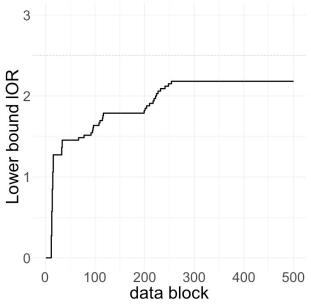
Log odds ratio boundary If the maximum likelihood estimate based on $Y^{(m)}$ lies in the upper left corner as in Figure 3.3(a), the confidence sets $\mathrm{CS}_{(m)}$ we get at time m have a one-sided shape such as the shaded region, or the shaded region in Figure 3.3(c), if the estimate lies in the lower right corner. Again, we can improve these confidence sequences by taking the running intersection; running intersections over time are illustrated in Figures 3.3(b) and 3.3(d).

3.4 Conclusion

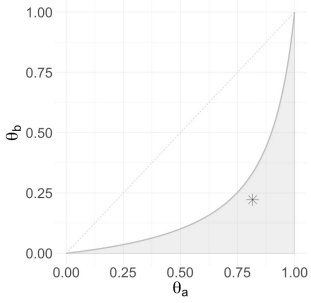
We have shown how e-variables for data streams can be extended to general null hypotheses and non-asymptotic always-valid confidence sequences. We specifically implemented the confidence sequences for the 2×2 contingency tables setting; the resulting confidence sequences are efficiently computed and show quick convergence in simulations. For estimating risk differences or relative risk ratios between proportions in two groups, to our knowledge, such exact confidence sequences did not yet exist. For the log odds ratio we could also have used the sequential probability ratio (SPR) in Wald's SPR test [Wald, 1945] test, which can be re-interpreted as a (product of) e-variables [Grünwald et al., 2022a]. However, the SPR does not satisfy the GRO property making it sub-optimal (see also [Adams, 2020]); moreover, as should be clear from the development, our method for constructing confidence sequences can be implemented for any effect size notion with convex rejection sets $\vec{\Theta}_0(\leq \delta)$ and $\vec{\Theta}_0(\geq \delta)$, not just the log odds ratio. A main goal for future work is to use Theorem 3.2 to provide such sequences for sequential two-sample settings that go beyond the 2×2 table.



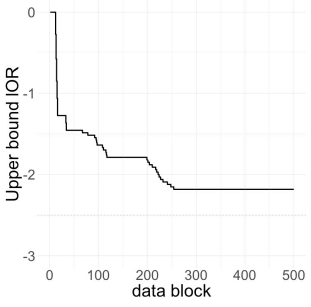
(a) CS^+ at $n = 500$, true lOR 2.5



(b) Running lower bound CS^+, true lOR 2.5



(c) CS^- at $n = 500$, true lOR -2.5



(d) Running upper bound CS^-, true lOR -2.5

Figure 3.3: One-sided confidence sequences for odds ratios. 500 data blocks were generated under P_{θ_a,θ_b} with $\theta_a = 0.2$ and log of the odds ratio (lOR) 2.5 for figures a and b, and $\theta_a = 0.8$ and lOR -2.5 for figures c and d. The asterisks indicate the maximum likelihood estimator at $n = 500$. The significance threshold was set to 0.05. The design was balanced, with data block sizes $n_a = 1$ and $n_b = 1$. Note that CS^- is empty for (a) and (b) and CS^+ for (c) and (d) in these confidence sequences.

Chapter 4

Information Extraction from Free Text for Aiding Transdiagnostic Psychiatry: constructing NLP Pipelines Tailored to Clinicians' Needs

Dr. Rosanne J. Turner[1,2], Femke Coenen[1], Femke Roelofs[1], Karin Hagoort[1], Dr. Aki Härmä[3], Prof. Peter D. Grünwald[2,4], Dr. Fleur P. Velders[1], Prof. Dr. Floortje E. Scheepers[1]

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

Abstract

Background Developing predictive models for precision psychiatry is challenging because of unavailability of the necessary data: extracting useful information from existing electronic health record (EHR) data is not straightforward, and available clinical trial datasets are often not representative for heterogeneous patient groups. The aim of this study was constructing a natural language processing (NLP) pipeline that extracts variables for building predictive models from EHRs. We specifically tailor the pipeline for extracting information on outcomes of psychiatry treatment trajectories, applicable throughout the entire spectrum of mental health disorders ("transdiagnostic").

Methods A qualitative study into beliefs of clinical staff on measuring treatment outcomes was conducted to construct a candidate list of variables to extract from the EHR. To investigate if the proposed variables are suitable for measuring treatment effects, resulting themes were compared to transdiagnostic outcome measures currently used in psychiatry research and compared to the HDRS (as a gold standard) through systematic review, resulting in an ideal set of variables. To extract these from EHR data, a semi-rule based NLP pipeline was constructed and tailored to the candidate variables using Prodigy. Classification accuracy and

F1-scores were calculated and pipeline output was compared to HDRS scores using clinical notes from patients admitted in 2019 and 2020.

Results Analysis of 34 questionnaires answered by clinical staff resulted in four themes defining treatment outcomes: symptom reduction, general well-being, social functioning and personalization. Systematic review revealed 242 different transdiagnostic outcome measures, with the 36-item Short-Form Survey for quality of life (SF36) being used most consistently, showing substantial overlap with the themes from the qualitative study. Comparing SF36 to HDRS scores in 26 studies revealed moderate to good correlations (0.62 - 0.79) and good positive predictive values (0.75 - 0.88). The NLP pipeline developed with notes from 22170 patients reached an accuracy of 95 to 99 percent (F1 scores: 0.38 – 0.86) on detecting these themes, evaluated on data from 361 patients.

Conclusions The NLP pipeline developed in this study extracts outcome measures from the EHR that cater specifically to the needs of clinical staff and align with outcome measures used to detect treatment effects in clinical trials.

4.1 Background

In psychiatry, it is still difficult to choose the best treatment for individual patients based on their specific characteristics. For example, in major depressive disorder, only one third of patients achieves remission after first-line treatment [Rybak et al., 2021]. This is why there is a plethora of attempts at developing machine learning models that support shared decision making and precision psychiatry (for example see Ermers et al. [2020] for a recent overview of machine learning models in major depressive disorder, and Sanfelici et al. [2020] for psychosis). However, as patient needs are personal and treatment outcomes are never binary in psychiatry [Wigman et al., 2013], choosing a representative outcome measure on which the machine learning models should report is key, but not straightforward.

In clinical trials, diagnosis-specific symptom rating scales are frequently used to detect treatment effects. However, these measures restrict developing decision support models to just one group of patients with the same "diagnostic label", whereas in practice, there almost never is a one-to-one correspondence between diagnostic labels and patients [Meiseberg and Moritz, 2020]. In addition, availability of patients' scores on rating scales in the electronic health records (EHR) is limited in practice, as they are mostly registered structurally in the clinical trial setting. Lastly and perhaps most importantly, symptom rating scales may not cover all information patients and clinicians are actually interested in with regard to recovery, for example insights into daily and social functioning.

Hence, alternative outcome measures for machine learning models to support patients and clinicians in (shared) decision making seem warranted. One alternative could be using scores that represent the patient's functioning, as they can be used to follow up treatment effectiveness in patients with different psychiatric disorders. This way, predictive models in which patients from a wide spectrum of mental disorders are included could utilize these outcome measures. Functional outcome measures may also better reflect added value for patients and the community [Glied et al., 2015], making machine learning models' predictions more insightful in comparison to predicting improvements on symptom rating scales.

This kind of information is not registered in a structured manner in the EHR, and extracting such outcome variables from clinical free text is a time-consuming process. On the other hand, it is unwarranted to introduce new questionnaires to clinical staff to collect data prospectively in a structured format for each predictive model that is built, as this would disproportionally increase administrative burden. Therefore, the aim of this study was to build a natural language processing (NLP) pipeline that can easily be tailored towards extracting specific information from clinical notes, and to show a specific application for extracting transdiagnostic outcome measures for mental health disorders.

To investigate which information would be valuable to report on in psychiatric clinical practice, psychiatry clinical staff of an academic hospital answered questionnaires to assess which outcome measures they would find appropriate to determine the effectiveness of treatment throughout the entire spectrum of mental health disorders. So far, most predictive models in psychiatry have been built around diagnosis-specific outcome measures [Fusar-Poli et al., 2018], hence

it is currently unknown whether treatment effects could be reflected adequately through more transdiagnostic and functional outcome measures, and whether it would be sensible to construct predictive models for these outcome measures at all. Therefore, to assess which transdiagnostic outcome measures resulting from the questionnaires were candidates, an overview of transdiagnostic measures used for detecting treatment effects in the research setting was created through systematic review. Second, the aptness of the found transdiagnostic measures for measuring treatment effects was assessed through comparing transdiagnostic domain scores in depression clinical trials with the gold standard in depression, the Hamilton Depression Rating Scale (HDRS), also through systematic review [Hamilton, 1960, Williams, 2001].

The results of the questionnaires and systematic reviews were combined into a list of candidate transdiagnostic outcome measures. Finally, it was assessed whether these could be accurately extracted from the EHR data with our proposed NLP pipeline. To compare the extracted outcomes to a gold standard measure in a subgroup of patients with symptoms of depression, analogously to the comparison of the outcome measures and HDRS through the systematic review, the association between the outcome measures constructed with the NLP pipeline and HDRS scores of patients at the academic hospital was assessed.

4.2 Methods

Determining which information on treatment outcomes is valuable in clinical practice To investigate which transdiagnostic outcome measures contain useful information for clinical practice, online questionnaires were developed and distributed among clinical staff at the Psychiatry department of UMCU (through Castor EDC, Ciwit B.V.). Questionnaires contained a combination of seven closed and seven open questions on defining recovery and treatment goals relevant for clinical decision making. For the analysis of the open questions, the framework for thematic analysis by Braun and Clarke was used [Braun and Clarke, 2006]. Detailed methods can be found in the additional information file, section 2.

Identifying transdiagnostic outcome measures used in research To further assess which outcome measures would be potential candidate measures for measuring treatment effects throughout the entire spectrum of mental health disorders, we aimed to find all transdiagnostic outcome measures that have been used in clinical trials from 2015 up to July 2020 through systematic review. The six-year cutoff was chosen to be able to focus on currently relevant outcome measures applicable to the Diagnostic and Statistical Manual of Mental Disorders, fifth Edition [American Psychiatric Association, 2013]. Studies concerning adult patients primarily diagnosed with a psychiatric disorder where at least one transdiagnostic outcome measure was used were included (details in additional information, section 3).

Assessing transdiagnostic outcomes for measuring treatment effects In the second review, the aptness of a transdiagnostic outcome measure to measure treatment effects was investigated through comparing changes in the 36-item Short-Form Survey for quality of life (SF36) with the gold standard in depression, the HDRS. All clinical trials up to July 2020 concerning patients with depression where both the HDRS and the SF36 were utilized as primary or secondary outcome measures were included. Mean SF36 subcomponent score changes were compared to the mean HDRS score changes through weighted correlation, and a confusion matrix was created to investigate the ability of the SF36 to reveal a significant treatment effect (details in additional information, section 4).

Assessing routinely collected information in the EHR as information sources To find sources to extract information on candidate themes after systematic review and qualitative analysis, the full spectrum of EHR data available at the psychiatry department of UMCU until 2020 was assessed, which included data from 22170 patients: de-identified doctors' and nurses' notes [Menger et al., 2018b], referral and dismissal letters, standardized forms containing treatment and prevention plans, standardized questionnaires performed (semi-)structurally, juridical status, destination after dismissal, lab measurements and prescribed medication. These sources were qualitatively assessed with regard to frequency of availability, relevance and quality.

Constructing an NLP pipeline To extract outcome measures from the unstructured data sources, the doctors' and nurses' notes, an NLP pipeline for analyzing Dutch clinical notes was developed, using as many available clinical text as possible, including notes from 5664 inpatient trajectories and from 18689 patients that were treated ambulatory. The main aim of the pipeline was to find for each patient all sentences that contain clinically relevant information about the candidate themes resulting from the qualitative study and reviews, and to attach a sentiment score for each theme to the sentences to be able to see if observations were positive or negative. As there is often a lot of repetition in daily written clinical notes (e.g., "Situation has not changed, patient still lacks initiative and still has a depressed mood"), we aimed to let the pipeline only filter and score sentences that contained an indicator of change in the patient's situation. This would probably give clinicians information that is more relevant to the course of treatment, compared to including sentences without change indicators in the scores.

A schematic overview of the proposed NLP pipeline with a hypothetical example of the analysis of a piece of clinical text can be found in figure 4.1. The five steps of analysis are briefly described in the next two paragraphs. Main units of analysis in the pipeline are sentences: in the first step, clinical notes are preprocessed by splitting them into sentences with a spaCy tokenizer [Honnibal et al., 2020]. In the second step, the sentences pass the theme filter, passing only when at least one phrase corresponding to one of the candidate themes is detected. In the third step, sentences pass through the change filter when they contain a phrase indicating a moment of change. This could either be a word directly describing

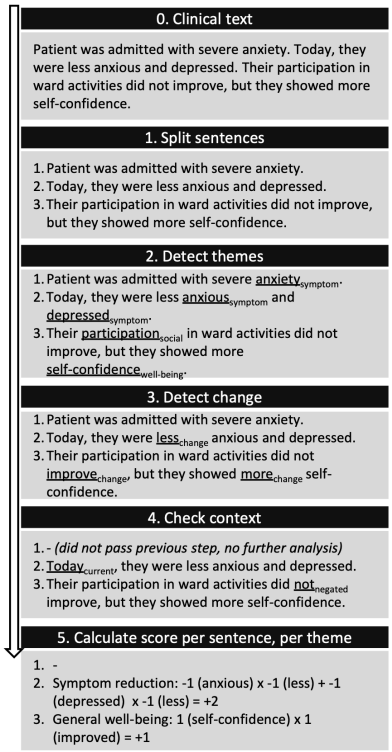


Figure 4.1: Schematic depiction of the NLP pipeline for extracting moments of change for each patient from clinical notes with a hypothetical example of a clinical text passing through all steps. Note that because in step 3 no change word was detected in sentence 1, further analysis of that sentence is cancelled. Note also that in step 4, a negated context is detected for the word "improve" in sentence 3, hence this change word and the corresponding theme word are not passed further through the analysis.

change (e.g. "improvement"), or a comparative form of an adjective (e.g. "angrier"). For the theme and change filters, lists of phrases for rule-based filtering (in Dutch) were needed. These were composed with the annotation tool Prodigy by authors RJT and FC (Prodigy, ExplosionAI, Berlin, Germany). Prodigy takes as input a spaCy model and a list of seed terms, and based on these seed terms and the word embeddings in the model efficiently suggests new phrases to add to the list. One of the major advantages of this method for composing a phrase list is that frequent spelling errors are included. Examples of parts of the composed lists with translations to English can be found in additional table 1, and complete composed phrase lists (in Dutch) can be found in the online repository for this project, available on GitHub [Turner, 2021].

In the fourth step, a context filter was applied to check if the theme phrase and change phrase were mentioned in a correct sentence context. Five checks were performed: whether the phrases were current, not hypothetical, concerned the patient, not negated, and whether the change concerned the theme (e.g., we need to detect "Today, anxiety symptoms increased", but not "We increased the medication doses but the patient's anxiety did not respond"). This filter uses part-of-speech and dependency tagging based on a previously developed spaCy model, regular expressions and literal phrases; details and a tutorial of the software can be found in Menger [2020] and Menger et al. [2018a]. In the fifth and final step, the sentences received scores for all themes that passed the filter for that specific sentence. This was done by RJT and FC through assigning sentiment scores to the theme and change phrases. For this project, we chose to assign negative phrases (e.g. "anxiety", "anger") the value -1, and positive phrases (e.g., "joy", "hygiene") the value 1. Change words indicating an increase were assigned the value 1, and those indicating a decrease the value -1. Final sentiment scores per sentence were calculated by multiplying each theme phrase score with its corresponding change phrase score. This way, an increase in something with a negative connotation, such as "more anxiety", would result in a score of -1, and an increase in something with a positive connotation, such as "participation improved", would result in a score of +1 (see also figure 4.1). When a sentence contained multiple theme phrases with a corresponding change phrase, e.g., "The patient was more anxious and sad", the scores were added, this example sentence resulting in a score of -2.

To assess if this pipeline could accurately extract sentences containing a moment of change with respect to the themes (regardless of sentiment), four validation datasets, one for each theme, were composed efficiently with the use of spaCy and Prodigy. As validation data, clinical notes from adult patients with one or more inpatient treatment trajectories at UMCU in 2020 were used, which were unseen during the phrase list development process described above. Using the theme phrases as a warm start, Prodigy selected sentences from the total data pool to label based on classification difficulty. Sentences were then labelled manually by RJT and FC, labelling a sentence as "accept" when it was judged that it should pass through all filters, and "reject" when it was judged to not contain a moment of change concerning the theme, in the correct contexts. The pipeline was also applied to this validation set, also labelling a sentence as "accept" when it passed through all filters for that specific theme, and "reject" when it did not pass. Given

and predicted labels were then compared, and classification accuracy, precision, recall and F1-scores were calculated.

Comparison of our candidate transdiagnostic outcomes to a gold standard Finally, to compare our structured and unstructured transdiagnostic outcome measures to a more symptom-specific gold standard, the NLP theme scores and scores on domains extracted from structured sources (e.g., juridical status and medication prescriptions) were compared to HDRS scores for patients admitted in 2019 and 2020 through linear regression with stepwise AIC-based model selection in R. Summary scores for each patient were obtained by calculating the mean sentiment over all sentences that passed the filters for that patient for each theme. E.g., if for the theme "symptoms" three sentences passed the filter for a patient, with scores -2, 1 and 2, the mean symptom sentiment score for this patient would be 1/3. Only complete cases, with clinical notes and information from all selected structured sources available, were analyzed. The scores from structured sources were incorporated as categorical data, either having worsened (e.g. more benzodiazepine prescriptions at the end of an admission compared to the start), having stayed the same or having improved. In potential, such a linear model trained to reflect gold standard HDRS scores could be used to in the end compose a combined weighted score from the NLP scores and information from structured data.

For all analyses, R (version 4.0.3) and Python (version 3.7.4) were used.

4.3 Results

Clinicians' views on outcomes Between June 23, 2020, and July 27, 2020, 38 healthcare professionals gave consent to participate in a survey on defining goals of treatment and recovery. 34 completed at least one item of the questionnaire. The group comprised 12 nurses, 3 nurse practitioners, 9 residents in psychiatry, and 10 psychiatrists. Through qualitative analysis, four distinct themes were identified that comprise the concepts "goals of treatment" and "recovery of a patient": personalization, symptom reduction, general well-being and social functioning. Detailed descriptions of the themes are depicted in table 4.1.

Table 4.1: Qualitative analysis of clinical staff's responses to a questionnaire on defining goals of treatment and recovery

Theme	**Description**	**Examples**
Personalization	Recovery is a highly personal process that is shaped by the patient's goals, story and views. Therefore, the treatment goals are dependent on the needs and goals of the patient. A situation is pursued in which professional care is no longer needed and the patient returns to his usual environment and position before illness.	"The patient's request, what he/she requires to function to his/her own needs..."; "In this respect it is always necessary to look at the patient's position before his illness, what he/she aims to accomplish, and which other factors are hindering, respectively facilitating the patient."
Symptom reduction	Treatment goals include reduction of symptoms, encompassing both psychiatric and somatic complaints. This reduction ranges from complete remission to mere stabilization in the acute phase of the illness. The recovery process is hard work and sometimes involves an initial aggravation (e.g., side effects). The aim is that the symptoms are diminished in a way that the patient is not restricted by them anymore (e.g., in daily functioning), or that the patient can function on his previous level again.	"Supporting patients in their recovery by treatment of psychiatric illness or symptoms.; "Reduction or recovery of symptoms."; "...as symptom-free as possible..."; "Recovery to the level of premorbid functioning and reduction of symptoms to premorbid"
General well-being	Another treatment goal is to raise general well-being and quality of life. The treatment stimulates that the patient gains insight into his illness and learns to cope with it and the vulnerability that remains when the symptoms are reduced. A new balance is established between the patient's capacities and the burden of the illness. This gives room for positive experiences, joy and a regained purpose in life.	"Improvement of quality of life."; "Feeling like living and being able to experience life satisfaction again."; "Regaining a purpose and a balance between the patient's capacities and the burden of the illness."

Table 4.1, continued

Theme	Description	Examples
Social functioning	Finally, treatment aims to improve the patient' social and societal functioning. The healthcare professionals try to enhance autonomy and self-sufficiency, so that the patient becomes able to participate in society again. This entails e.g., living independently, engaging in activities that are important to the patient, having a job and meaningful relationships with others.	"Treatment of complaints, that give severe hinder in daily life, of the patient so that the patient is able to gradually resume his/her life and participate in society again."; "Recovery of healthy functioning on life domains like work, relationships, living and spare time."

Transdiagnostic outcome measures in research The search for clinical trials where transdiagnostic outcomes were used yielded 1962 studies, of which 362 were included (details of exclusion criteria and an overview of included studies can be found in additional information, section 3 and additional table 2). In these studies, 242 different transdiagnostic outcome measures were applied. The most prevalent outcome measures were the Clinical Global Impression (CGI), Short Form Health Survey (SF), Global Assessment of Functioning (GAF), EuroQol 5d (EQ-5D) and World Health Organization Quality of Life (WHOQOL) questionnaires [Busner and Targum, 2007, Brazier et al., 1992, Jones et al., 1995, Rabin and de Charro, 2001, World Health Organization, 1995]. An overview of the ten most-used outcome measures is provided in the additional information, additional table 3. The CGI and the GAF concern very short surveys, but the SF, EQ-5D and WHOQOL all three concern longer, detailed questionnaires with overlapping themes concerning physical, mental and emotional well-being, and social and societal functioning.

In figure 4.2 the frequency of usage of these outcome measures per diagnosis is illustrated. The SF is used in a substantial portion of studies for all diagnoses, whereas for the other questionnaires the usage varies depending on the specific diagnosis. The 36-item, most widely used version of the SF (SF36) consists of eight subcomponents; physical health, physical role perception, bodily pain, general health perception, mental health, emotional role perception, vitality and social functioning, which together roughly cover the spectrum of topics covered by the other most-used questionnaires. As the SF36 also is the most widely-used method to quantify health-related quality of life [Cordier et al., 2018], these SF36 subcomponents were used for further investigation of the extent to which a transdiagnostic outcome measure is as sensitive to changes over the course of treatment compared with diagnosis-specific questionnaires.

Transdiagnostic outcomes for measuring treatment effects Systematic review yielded 26 studies where both SF36 and HDRS were measured during treatment trajectories of patients with depression; detailed results can be found in the

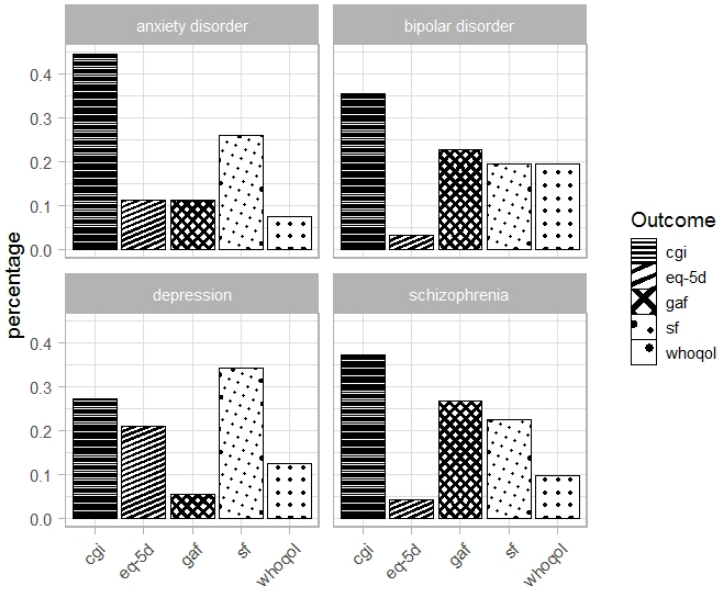


Figure 4.2: Top 5 most-used transdiagnostic outcome measures during the past five years. The prevalence of usage of the top 5 most-used transdiagnostic outcome measures for the most prevalent diagnoses for which general outcome measures were used during the past 5 years in clinical trials are shown.

additional information, additional table 4. The strength of the Pearson correlation coefficients between SF36 subscores and changes in HDRS scores varied from moderate to strong with bodily pain to be the lowest, and physical health perception to be the highest (R = -0.601, and R = -0.786, respectively), i.e., better scores on the subcomponents of the SF36 indicate an improvement of symptoms of depression (additional table 5). The positive predictive value of most of the SF36 subscores was high, indicating that the SF36 is apt for detecting treatment effects (additional table 6).

Interestingly, the themes resulting from the interviews with psychiatry staff show substantial overlap with the subcomponents of the ten most prevalent questionnaires found in our systematic reviews, which also mainly focused on (physical and) mental health symptoms, social and societal functioning and more general emotional well-being. Specifically comparing them to the SF36, the theme "symptom reduction" corresponds to the subcomponents physical health, bodily pain and mental health, the theme "social functioning" directly to its social functioning counterpart, "general well-being" to vitality and general health perception and "personalization" to physical role perception, general health perception and emotional role perception. With the SF36 subcomponents deemed as good alternatives for measuring treatment effects compared to a syndrome-specific gold standard in

the systematic review above, it was hypothesized that these four themes would be good candidate treatment effect measures as well, while also catering to the specific needs of clinicians in psychiatry practice.

Extraction from the EHR To report on outcomes on the four found themes for individual patient treatment trajectories, all EHR sources available at the psychiatry department of UMCU until 2020 (22170 patients) were assessed with regard to availability of information on the themes, and for relevance and quality of this information. An overview of assessed sources and their aptness is given in the additional information, additional table 7. For the information extraction, the definition of the "personalization" theme was narrowed down to "patient experience", and for this theme, the EHR sources were searched for information on the thoughts and remarks of patients about their treatment trajectory. Sources selected as feasible for calculating theme scores were clinical notes (for each theme), juridical status (for symptoms and social functioning), medication prescriptions during admission (for symptoms) and destination after dismissal (for symptoms and social functioning).

NLP pipeline assessment To validate the NLP pipeline for extracting information from unstructured EHR sources, validation sets were composed for each theme with all clinical notes of admitted, adult patients at the UMCU in 2020. This set comprised 439 trajectories of 361 patients with a mean duration of 57 days; 39 percent was admitted to emergency care, 31 percent to a ward specialized in the diagnosis of first episode psychosis and 26 percent to a ward specialized in affective and psychotic disorders. In table 4.2, the average number of sentences containing a phrase for each of the four themes per inpatient treatment trajectory, the number of sentences selected by the pipeline as mentioning a change in the theme in the correct context and some example sentences can be found. On validation sets with 663, 292, 328 and 269 sentences for symptoms, social, well-being and patients' experience, respectively, 0/1 accuracies between 95 and 99 percent were achieved on each of the themes (also see table 4.2). Remarkable is the high precision, but low recall for the symptom reduction and general well-being themes; reviewing the false negative sentences revealed that a large part could be contributed to missed verb conjugations in the change phrases, and specifically conjugation breaks, which occur a lot in Dutch. Also notable is the low precision but high recall for the patient experience theme.

Table 4.2: Results of the NLP pipeline applied to all clinical notes of 2020.

Theme	**Mean number of sentences concerning theme per trajectory (sd)**	**Mean number of sentences with relevant change in theme per trajectory (sd)**	**Examples of sentences translated from Dutch marked as correct**	**Classification accuracy of pipeline**	**Precision**	**Recall**	**F1-score**
Symptom reduction	103 (99.8)	8.0(10.2)	"Nervousness increased over the course of the day","The patient appears drowsier than before"	0.988	0.857	0.461	0.6
Social functioning	131 (121)	4.0 (5.1)	"Friendly, more interaction than yesterday"	0.997	0.750	1.00	0.857
General well-being	119 (126)	6.4 (8.9)	"This afternoon, the patient felt less well", "Had less energy"	0.951	0.833	0.250	0.385
Patient experience	164 (159)	9.4 (11.0)	"Says that it is going well, has the idea that it is going better and better"	0.993	0.333	1.00	0.5

As an example of clinical applicability, to assess the aptness of these scores, in addition to the scores from the structured sources (medication prescriptions, juridical status and destination after dismissal) to reflect treatment effects during an inpatient treatment trajectory, they were compared to changes in HDRS scores in patients with symptoms of depression. These were available for 120 patients in 2019 and 2020; 80 of these patients were admitted to the ward for affective and psychotic disorders, and 40 to other wards. The mean HDRS score at the end of inpatient treatment trajectories was 14, with a minimum of 2 and a maximum of 33. On average, 2802 sentences of clinical notes were available for each patient, and 88 change sentences passed the filter. Linear regression with stepwise model selection revealed that the most parsimonious model (based on AIC) for predicting HDRS scores at the end of inpatient treatment trajectories included mean sentiment scores for the symptom and social functioning themes, juridical status, benzodiazepine prescriptions and other psychiatric medication prescriptions as covariates (table 4.3). Negative model coefficients were found for the sentiment of psychiatric core symptoms and a decrease (i.e. improvement) in benzodiazepine prescriptions, implying that improvements on these themes are associated with improvement of depression symptoms.

Table 4.3: The most parsimonious linear regression model after stepwise model selection with AIC for predicting HDRS scores at the end of treatment

Predictor	**Coefficient**	**Standard error**	**P-value**
(Intercept)	8.549	3.151	0.00777
Mean sentiment psychiatry symptoms	-3.711	1.336	0.00645
Mean sentiment social functioning	2.354	1.318	0.07692
No change in juridical status	5.496	0.920	0.35966
Juridical status improved	5.412	2.698	0.04739
No change in benzodiazepine prescriptions	-3.769	1.774	0.03589
Decrease in benzodiazepine prescriptions	-3.467	2.288	0.13264
No change in other psychiatry medication prescriptions	2.346	1.628	0.15243
Decrease in other psychiatry medication prescriptions	3.403	1.881	0.07315

4.4 Discussion

With the research described in this paper, we aimed to identify useful and real-time extractable outcome measures for machine learning models in psychiatry. Through systematic review, transdiagnostic outcome measures concerning core symptoms, social functioning and general well-being were identified. Comparison of scores on these themes with Hamilton scores through systematic review showed that

these themes appropriately reflect outcomes of treatment trajectories. Themes defined by clinicians at the academic hospital that together cover the spectrum of defining successful treatment trajectories were symptom reduction, general well-being, social functioning and personalization, which show substantial overlap with the themes found through systematic review. Through combining structured and unstructured EHR data that was already available, an NLP pipeline was developed through which scores on the subthemes could be extracted from the EHR, with good F1-scores for detecting information on symptoms and social functioning. . The symptom reduction and social functioning themes were associated with HDRS scores for patients admitted in 2019 and 2020.

In this study we composed the phrase lists for text mining each theme ourselves, tailored to the current specific problem and clinician writing styles, which required a substantial time investment. In future research the performance of our phrase lists could be compared with existing medical ontologies like Systemized Nomenclature of Medicine Clinical Terminology (SNOMED-CT) [Stearns et al., 2001]. However, existing medical ontologies do not address issues like spelling mistakes and form variability, which might decrease their sensitivity. Similarly, the rule-based nature of our pipeline did not allow for enough flexibility to cover all verb conjugations in Dutch, possibly explaining the low recall on the symptom and well-being themes. The pipeline also had a low precision for the patient experience theme, probably also explained by the rule-based nature; because of the broad, unspecific nature of this theme, many generic phrases were included in the filtering lists. A tool which could potentially handle more flexibility is the open-source Medical Concept Annotation Toolkit (MedCAT) [Kraljevic et al., 2021]. This is a novel self-supervised machine learning algorithm that uses concept vocabulary (including SNOMED-CT) for extracting concepts and also supports contextualization through unsupervised learning, matching ambiguous concepts to the best fitting overarching concepts.

To enable comparing the transdiagnostic measures we selected based on the qualitative study into clinical staff's beliefs and literature review to an objective measure, a linear model with stepwise model selection was fitted with the transdiagnostic measures as predictors, and HDRS scores as outcomes. Ideally, one would compare the candidate outcome measures to an existing transdiagnostic outcome questionnaire such as SF36 to be able to extend this comparison beyond depressive symptoms, but these are not often part of routine clinical care and were not available for our retrospectively collected cohort. The second systematic review performed in this study however revealed that changes in HDRS scores are correlated with changes in SF36 scores. The HDRS quantifies depression symptoms; with this analysis, we have shown that several of the candidate outcome measures are associated with this gold standard. For the NLP themes, the themes reflecting social functioning and psychiatric symptoms were associated, perhaps reflecting the symptom-oriented nature of the HDRS. These associations might indicate that the theme scores developed in this study could potentially be used to measure treatment effects transdiagnostically, but to prove this, comparison with an objective transdiagnostic standard such as the SF36 or syndrome specific gold standards reflecting other mental illnesses would be necessary.

Not for all patients for whom clinical notes were available, sentences with changes were present for every theme in the clinical notes. This highlights the possibility of the existence of bias in these retrospective clinical notes: possibly, only more remarkable changes during treatment trajectories are denoted. When trying to gain qualitative insights into treatment trajectories for individual patients these "noisy" observations being omitted might actually be helpful, but when trying to create quantitative overviews or to find associations results could be misleading. This is an unavoidable challenge when trying to use existing data to develop predictive models, and warrants the need for prospective studies into the coherence between transdiagnostic outcomes measured through standardized questionnaires, and the content of clinical notes.

The research in this paper emphasizes the need for standardized outcome measures for comparing and combining machine learning models in mental health. The Core Outcome Measures in Effectiveness Trials (COMET) initiative has initiated this sort of work with the goal of streamlining clinical trial initiatives [Prinsen et al., 2014]; it would be interesting to further formalize standards for prediction models as well, as this would certainly aid working towards FAIR use of data and initiatives for sharing and collectively training machine learning models in healthcare [Wilkinson et al., 2016, Deist et al., 2020].

4.5 Conclusions

This paper highlights information extraction from clinical notes as a good alternative for standardized questionnaires when one aims to gain insight into treatment outcomes at their facility. We have shown that it is not only feasible to extract information on outcome measures of interest from clinical text, but we also validated that these transdiagnostic themes might accurately reflect treatment outcomes in a subgroup of patients with symptoms of depression, as compared with the Hamilton questionnaire. This approach has a closer connection to clinical practice and individual patients, as it is directly based on real data and clinical practice as opposed to measuring instruments for clinical research. From here forward, pipelines like this could be used to generate better insights into treatment outcomes for all patients in a cohort for which clinical notes are available, as opposed to only patients for which standardized questionnaires are available, a possible source of selection bias. Clinicians could for example be offered real-time insights into treatment outcomes for diverse patient groups at their department through a dashboard with summary statistics of all the outcome measures. An interesting addition would be the construction of a combined weighted outcome score, with weights for example based on a linear regression model, such as the one trained in this paper, with the HDRS scores at outcomes.

Chapter 5

Bayesian Network Analysis of Antidepressant Treatment Trajectories

Rosanne J. Turner[1,2], Karin Hagoort[1], Rosa J. Meijer[3], Femke Coenen[1], Prof. Dr. Floortje E. Scheepers[1]

1: University Medical Center Utrecht, Brain Center, Netherlands
2: CWI, Machine Learning group, Netherlands
3: Data Science Department, Parnassia Groep, Den Haag, Netherlands

Abstract

It is currently difficult to successfully choose the correct type of antidepressant for individual patients. To discover patterns in patient characteristics, treatment choices and outcomes, we performed retrospective Bayesian network analysis combined with natural language processing (NLP).

This study was conducted at two mental healthcare facilities in the Netherlands. Adult patients admitted and treated with antidepressants between 2014 and 2020 were included. Outcome measures were antidepressant continuation, prescription duration and four treatment outcome topics: core complaints, social functioning, general well-being and patient experience, extracted through NLP of clinical notes. Combined with patient and treatment characteristics, Bayesian networks were constructed at both facilities and compared.

Antidepressant choices were continued in 66% and 89% of antidepressant trajectories. Score-based network analysis revealed 28 dependencies between treatment choices, patient characteristics and outcomes. Treatment outcomes and prescription duration were tightly intertwined and interacted with antipsychotics and benzodiazepine co-medication. Tricyclic antidepressant prescription and depressive disorder were important predictors for antidepressant continuation.

We show a feasible way of pattern discovery in psychiatry data, through combining network analysis with NLP. Further research should explore the found patterns in patient characteristics, treatment choices and outcomes prospectively, and the possibility of translating these into a tool for clinical decision support.

5.1 Introduction

Patients seeking treatment for severe depression symptoms often have a long trajectory ahead of them; only approximately one third continues their medication of first choice and about 30 percent has still not achieved remission after four treatment steps [Rybak et al., 2021, Rush et al., 2006, Gaynes et al., 2009]. Meanwhile, the contribution of mental health disorders to the global burden of disease is substantial [Whiteford et al., 2015]. Despite the limitations, pharmacological treatment of severe depression is still the most common treatment choice. Since it is still difficult to predict the response to a specific antidepressant type in an individual, the prescription process is one of trial and error. For a patient this can result in unnecessary and possibly harmful side effects and delayed recovery. Especially challenging in the prescription of antidepressants is that both the choice of the antidepressant and the response are influenced by multiple variables relating to the prescriber, the patient, illness characteristics and the drug itself [Bayes and Parker, 2019]. Insights into the interactions between these factors and their effects on treatment outcomes could allow greater precision in the choice of an antidepressant for a given patient, but are currently lacking [Pradier et al., 2020].

To empower patients and clinician during treatment choices, the multi-faceted, non-binary aspects of psychiatric care are hard, but essential to account for [Kirtley et al., 2022]. During the last decade many machine learning models with the aim of personalizing treatment recommendations for patients with symptoms of depression have been developed [Ermers et al., 2020]. However, little has changed in actual clinical psychiatry practice yet, perhaps because of the "black box" nature of most clinical machine learning models [Kundu, 2021].

Network analysis is a promising candidate from the joint field of statistical learning and machine learning that could potentially offer the desired multi-faceted insights into psychopathology in an explainable and transparent manner [Borsboom and Cramer, 2013]. It comprises of methods of data analysis where dependencies and/or causal pathways between all variables in a dataset are learned and visualized [Scutari and Strimmer, 2011]. Because mental health syndromes often present as a collection of tightly intertwined signs and symptoms, which sustain and influence each other and can be intervened on through multiple pathways, network analysis appears especially apt for capturing these concepts.

Previous studies on network analysis in mental health mainly focused on modelling symptom networks and yielded promising results. A study with a penalized Gaussian graphical model, including 1029 participants and 16 depression and anxiety symptoms, resulted in stable networks [Beard et al., 2016]. In a greedy search Bayesian network approach with a relatively small sample size of 353 subjects where relations between 10 stress-related variables were investigated, moderate classification accuracy of the network was achieved [Lee et al., 2019]. Network analyses studies with similar sample sizes and numbers of variables on obsessive-compulsive disorder and depression, and suicidal ideation (408 and 336) also revealed key gateway symptoms influencing symptom clusters [McNally et al., 2017, De Beurs et al., 2021]. A pilot with personalized feedback to patients through symptom network analysis showed promising results with respect to increasing a

patient's understanding of their psychopathology [Kroeze et al., 2017].

The above-mentioned studies illustrate the aptness of network analysis for showing and interpreting associations between symptoms. In the future, a tool for explainable personalized insights into antidepressant recommendations based on these kinds of networks could potentially be of significant value in clinical decision making. To work toward this goal, for this study, we intended to explore if treatment characteristics (antidepressant choices and co-medication), patient characteristics and treatment outcomes in addition to symptom scores can be incorporated in network analysis, using retrospective data from two mental healthcare facilities in the Netherlands. To extract information on mental health symptoms and treatment outcomes from the retrospective data, the network analysis was combined with a natural language processing (NLP) model [Turner et al., 2022].

Since our end goal is to develop a tool for explainable personalized insights into antidepressant recommendations we were primarily interested in causal paths and discovering (conditional) dependence relations among patient characteristics, treatment choices and outcomes. Hence, we have chosen to perform a Bayesian network (BN) analysis instead of a partial correlation network analysis or Markov random field analysis [Briganti et al., 2022]. The final BN, the found dependencies and predictions for hypothetical patients were compared to expert opinion to assess the potential of the model for future implementation in a tool for clinical decision support.

5.2 Methods

Main units of analysis Main units of analysis were first-time inpatient antidepressant treatment trajectories at participating mental health care facilities; consecutive prescriptions for one type of antidepressant were viewed as a single treatment trajectory. New prescriptions for the same type of antidepressant that started within 30 days after the old prescription were viewed as belonging to the same trajectory. Antidepressant treatment trajectories between 2014 and 2020 at two mental healthcare facilities involved were included. The first mental healthcare facility, Parnassia Group (PG), provides basic and specialized services for prevention, treatment (inpatient and outpatient), support and care after treatment. The second facility, UMC Utrecht (UMCU), is an academic specialized facility for tertiary care. As PG and UMCU deliver care in different regions in the Netherlands, the probability of overlap in patient populations is negligible.

To ensure a homogeneous patient population, only trajectories with (partial) inpatient treatment were included. We ultimately aim to assist a broader group of patients than only those with a clear-cut classification fitting the Diagnostic and Statistical Manual of Mental Disorders (DSM) categories. Therefore, all antidepressant trajectories were included regardless of DSM classification [American Psychiatric Association, 1994]. However, to keep the populations from both facilities comparable we did not include patients with addiction as a primary diagnosis, since PG includes a few clinics specialized in addiction treatment and UMCU does not, and addiction as a primary diagnosis has a dominant impact on all interventions [Naglich et al., 2018]. Patients with addiction as a secondary diagnosis were

included, to still enable investigating the possible interactions between depressive symptoms, choice of antidepressant and treatment with disulfiram.

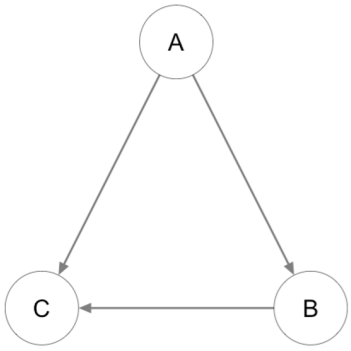
Predictor variables Predictor variables available at the start of (or becoming available during) the treatment trajectories comprised of gender, age, antidepressant type, co-medication, psychiatric (co-)morbidities according to the DSM classification system and global assessment of functioning (GAF) scores as registered in the DSM classification system. Antidepressant types were grouped into selective serotonin reuptake inhibitors (SSRI), non-selective serotonin reuptake inhibitors (nSSRI), tricyclic antidepressants (TriCA), tetracyclic antidepressants such as mirtazapine and mianserine (TetraCA), monoamine oxidase inhibitors (MAOI) and a remainder category (other), including for example bupropion (for a full overview, see supplementary table S5.1). Co-medication subgroups included in analysis were lithium, antipsychotics, tranquilizers (benzodiazepines) and disulfiram. Disulfiram was included because of its strong interactions with TriCAs [Ciraulo et al., 1985]. As information on other forms of treatment running concurrently, such as psychotherapy, was not available in a homogeneous format within and across treatment facilities, we did not incorporate these other treatments as predictor variables. (Co-)morbidities included were depression, anxiety disorder, personality disorder and problems in the social environment. Information on all variables except GAF scores was complete; missing data on GAF scores were imputed using the MICE software in R [Van Buuren and Groothuis-Oudshoorn, 2011].

Outcome variables Acceptability and efficacy are the main categories of outcome variables in antidepressant research. In this study acceptability is operationalized in prescription duration (>= 5 weeks indicating an "effective" duration, i.e., long enough for a treatment effect to be observed), and continuation of the antidepressant type (the final type prescribed at the mental healthcare facility during consecutive treatment for that patient). Efficacy was measured in terms of change scores on four mental health recovery themes: psychiatric core complaints, general well-being, social functioning and patient's experience. These last four scores were extracted from doctors' and nurses' notes with an NLP model, as described in previous work [Turner et al., 2022]. We explicitly chose not to use Hamilton scores as outcomes in this study, as those only focus on symptom reduction and are not systematically registered during routine clinical care in the Netherlands. Concisely, all available clinical notes during the patients' antidepressant treatment trajectories were screened for sentences concerning moments of change on one of the four themes, mentioned in a correct context (including, for example, "Today, the patient's mood significantly improved" but not "last year, after their grandmother died, the patient's mood declined"). The detected words were then combined with a sentiment score (1 or -1 for each detected word), a positive score indicating a positive change and vice versa, and a mean score for the entire treatment trajectory was calculated for each theme.

Medication doses At least 24 different antidepressants were prescribed at PG and UMC Utrecht between 2014 and 2020. To ensure faulty entries in the electronic

patient files were not included in the dataset, prescriptions where less than half of the minimal therapeutic dose according to the Dutch national standards of care was prescribed were excluded [Dutch National Healthcare Institute, 2020]. These minimal doses as listed May 2020 are summarized in supplementary table S5.2. Further, prescriptions exceeding five times the maximal therapeutic dose were excluded as well, as these can only be faulty entries in the electronic patient records.

Bayesian network analysis All analyses were performed with R (version 4.0.0) using the "bnlearn" package [Scutari, 2010], and network visualizations were constructed with the "qgraph" package [Epskamp et al., 2012]. A BN is a representation of the presence and strength of dependencies between all variables in a dataset (these could be predictive and/ or causal, see further below). A BN represents qualitative and quantitative information. It includes the structure of the dependencies (sometimes also called relations or associations), often depicted in a schematic figure called a graph, and the corresponding quantitative model built with these dependencies. For example, in the toy BN depicted in figure 5.1, the dependencies between the three variables are indicated with arrows in the graph on the left, and the corresponding model quantification in the form of a conditional probability table is depicted on the right. Learning a BN from data and/ or expert knowledge also follows these two stages: first one performs structure learning, identifying the relevant dependencies and their direction; and secondly parameter learning, estimating the parameters that quantify the dependencies [Briganti et al., 2022, Scutari, 2010].



	Conditional probability	
Variable A	0.7	
Variable B	*A absent*	*A occurred*
	0.4	0.7
Variable C	*A absent*	*A occurred*
B absent	0.4	0.4
B occurred	0.3	0.9

Figure 5.1: Toy example of a depiction of the structure of a Bayesian network with three binary variables and the corresponding predictive model: in this case, a conditional probability table.

Learning the structure of the BN teaches us which other variables in the dataset influence the probability that a variable takes on a certain value. Going back to the toy example, variable B has one incoming arrow, from variable A, indicating that we can write our prediction of the probability that B will occur in terms of A, "B

depends on A". Variable C has two incoming arrows, indicating that "C depends on both A and B". Conveniently, this dependency also gives us information in the other direction [Scutari and Strimmer, 2011]: if we have observed that B has taken on the value 1, we have a better estimate of which value A has compared to the scenario where we do not have any information at all, which we will also use for making predictions in the final part of this paper.

In general, there are two approaches for structure learning: constraint-based and score-based. Constraint-based learning aims to optimize the learning process to discover conditional dependencies between variables based on statistical inference and hypothesis testing. Score-based structure learning is aimed at optimizing the predictions the model makes for the data (formally: the likelihood of the joint probability distribution of all variables in the dataset). In this paper, both approaches were applied ("pcStable" was used as a constraint-based method, and "tabu" as a score-based method). To reduce the possibility of including spurious dependencies, model averaging was performed with bootstrap resampling, with 100 iterations. To ensure stability of the found associations, only edges that appeared in more than 85% of bootstrap samples were included in the averaged network [Briganti et al., 2022]. To investigate stability of associations across the two mental healthcare facilities, bootstrapped averaged networks were obtained at both facilities with both methods and compared with respect to found dependencies between variables.

To quantify the model, the current version of the Bayesian network software we used only offers the possibility of incorporating discrete variables as predictors of discrete variables; discrete variables cannot depend on continuous variables, potentially limiting the structures that can be found. To account for this, all variables were converted to binary variables. Age and GAF scores were compared to their respective median value at PG to ensure comparability across locations, with age being converted to older or younger than 48 (the median age in the PG data), and GAF being higher or lower than 50. Antidepressant prescriptions durations were converted to >=5 weeks or < 5 weeks (minimum time for an expected clinical effect), and mean sentiment scores being positive (>= 0) or negative (< 0).

Model parameters were fitted for the average model according to their maximum likelihood estimators and the resulting conditional probability tables were recorded. A toy example of such a table is included in figure 5.1. For example, because B depends on A, it can be observed that the probability of B occurring increases from 0.4 to 0.7 if we know A has occurred. C depends on both A and B, and it can also be observed that the model captures an interaction between A and B: in the absence of B, the effect of A on C disappears. Such model predictions in the presence or absence of information on specific hypothetical patient and treatment characteristics were also generated for the final network using the logic sampling functionality in the bnlearn package and compared to expert (FS) knowledge.

Ethics statement This study (number 22–705/DB) was assessed and approved to not fall under the scope of the Medical Research Involving Human Subjects

Act (WMO) by MREC NedMec: a recognized medical research ethics committee to which the Antoni van Leeuwenhoek, the Princess Máxima Center for pediatric oncology and the UMC Utrecht are affiliated. Complying with the guidelines issued by the MREC NedMec for research not falling under the Medical Research Involving Human Subjects Act (WMO), informed consent was waived by a quality officer from the research and ethics board of the Brain Center of UMC Utrecht on behalf of the MREC NedMec. It was deemed a disproportional effort to obtain informed consent of each individual patient because of the retrospective nature of the study and number of patients, of which many could not be contacted anymore because they continued their treatment elsewhere. However, the centers where this study was carried out uses an opt-out policy for patients who do not want their data to be used for research. Only data from patients who did not object to the use of their routinely collected electronic health record data were analyzed. According to Dutch national guidelines, the board of each university medical center is responsible for research quality control [Netherlands Federation of University Medical Centers, 2020]. For this study, the protocol was approved by a quality officer from the research and ethics board of the Brain Center of UMC Utrecht, appointed by the board of UMC Utrecht. This study conforms to the declaration of Helsinki for ethical principles involving human participants. To assure patients' privacy data were de-identified, for which the DEDUCE software was used [Menger et al., 2018b].

5.3 Results

4808 and 735 trajectories of patients with first-time inpatient antidepressant prescriptions were included in PG and UMCU respectively. In table 5.1 summary statistics of included trajectories are depicted. At PG, there is generally a long outpatient follow-up of patients, as the facility offers a wide range of levels of care: the mean period during which follow-up treatment was given at PG after the start of a first inpatient antidepressant trajectory was 1175 days (median follow-up duration: 866 days). At UMCU however, care is very specialized and patients are referred to other care facilities after dismissal: one month after dismissal, for 261 trajectories where patients were released into ambulatory care there still was an (ambulatory) care path at UMCU. For 174 trajectories, patients were referred to inpatient care at another facility. This resulted in a mean (inpatient) follow up duration of only 52 days after the start of antidepressant prescriptions.

Table 5.1: Overview of patient and treatment characteristics of included treatment trajectories

	PG (n = 4808)	**UMCU (n = 735)**
Variable	Mean (sd) or proportion	Mean (sd) or proportion
Follow-up from start prescription (days)	1175 (804)	52.7 (59.8)

Table 5.1, continued

	PG	UMCU
Age	48.42 (17.97)	43.861 (17.02)
Sex: female	0.577	0.615
Prescription group MAOI	0.013	0.061
Prescription group other	0.047	0.020
Prescription group SSRI	0.467	0.430
Prescription group nSSRI	0.161	0.200
Prescription group TCA	0.177	0.261
Prescription group TetraCA	0.172	0.060
Benzodiazepine prescription	0.643	0.848
Lithium prescription	0.081	0.165
Antipsychotics prescription	0.423	0.574
Disulfiram prescription	0.015	0.003
DSM: Depression	0.401	0.571
DSM: Personality disorder	0.268	0.242
DSM: Anxiety disorder	0.077	0.125
DSM: Social problems	0.025	0.211
GAF score at start treatment	48.03 (9.463)	33.33 (13.56)
Medication trajectory duration (days)	162.7 (234.1)	109.4 (236.0)
Continuation of antidepressant	0.663	0.894
Mean change sentiment core complaints	-0.152 (0.752)	-0.206 (0.703)
Mean change sentiment social functioning	0.321 (0.698)	0.464 (0.696)
Mean change sentiment well-being	0.293 (0.618)	0.162 (0.696)
Mean change sentiment experience	-0.094 (0.681)	-0.160 (0.590)

Duration and continuation The average prescription duration of the first-time antidepressant trajectories at PG was 163 days and 109 days at UMCU. At PG, 33.7% of patients switched to a different antidepressant type during follow-up, whereas at UMCU, only 10.6% of patients switched. This could partially be explained by the shorter follow-up period at UMCU, or the fact that more patients at UMCU had a history of ineffective antidepressant use. Note that the average prescription duration at UMCU exceeds the mean follow-up time, as many patients were dismissed with a prescription to continue their antidepressant use as at home or in another clinic. Outcome measures are summarized in detail for each type of initially prescribed antidepressant in supplementary table S5.3. At both facilities, patients were most likely to continue their prescription when they started with SSRI or TriCA. Patients were most likely to switch when they started with an "other" type of antidepressant (often bupropion), a MAO inhibitor (UMCU) or a

tetracyclic antidepressant (PG). Prescription durations were also the shortest for tetracyclic antidepressants, and relatively long for MAO inhibitors and nSSRIs.

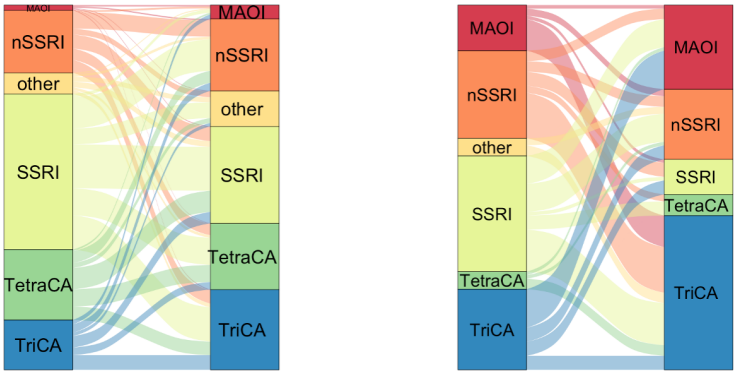


Figure 5.2: Flow diagram of antidepressant type switches for patients who did not continue their initial prescription for PG (left) and UMCU (right). Note the existence of flows from an antidepressant type to that same type during follow-up: these occur when a patient started with a single type of antidepressant and later switched to a combination of types or switched and thereafter returned to the original type. Pauses in prescriptions of the same antidepressant type were not regarded as switches.

In figure 5.2 (constructed using the ggalluvial package [Brunson and Read, 2020]), antidepressant type prescription switches for patients who did not continue their initial type(s) of antidepressant are depicted. At PG, SSRI is the biggest group that patients switch to, and at UMCU, patients more often switch to a tricyclic antidepressant. At UMCU, MAO inhibiters form a significant fraction of follow-up medication, including patients that tried an nSSRI, SSRI or tricyclic antidepressant first.

Mental health recovery outcome measures Examples of (translated) found sentences for each of the analyzed themes, core complaints, social functioning, well-being and patient's experience, are "Anxiety complaints less than Tuesday last week and manageable", "Patient says they have the feeling they are improving every day", "Patient likes working more, because it improves daily structure" and "Patient experiences more peaceful feelings". At PG, on average 4.06, 1.65, 4.04 and 5.69 sentences indicating a moment of change with respect to complaints, social functioning, well-being and experience were detected during antidepressant prescription periods. At UMCU, 11.2, 4.63, 10.0 and 16.7 sentences were on average detected for the respective themes. A possible explanation for this difference could be the nature of the patient reports at both centers: at the UMCU, follow-up was entirely inpatient, reflected in a higher mean number of days with available clinical notes (34) and total length of all clinical notes combined (mean 82.691 characters per patient). At PG the follow-up was mostly outpatient, where clini-

cal notes were on average available on 28 days with a mean total length of 26.674 characters per patient.

Bayesian networks With the constraint-based algorithm, the bootstrapped averaged network contained 24 arcs connecting the variables in the network for the PG data (figure 5.3). The UMCU bootstrapped averaged network only contained 9 arcs, of which 2 were also present in the PG network. Interestingly, many dependencies were found between the text-mined outcome measures and prescription duration nodes, indicating that trajectory outcomes are tightly intertwined. In the PG data, the use of benzodiazepines and antipsychotics during antidepressant prescriptions were directly linked to these outcomes. Whether a patient switched antidepressant types was directly dependent on TetraCA prescriptions and a DSM diagnosis of a (type of) depressive disorder.

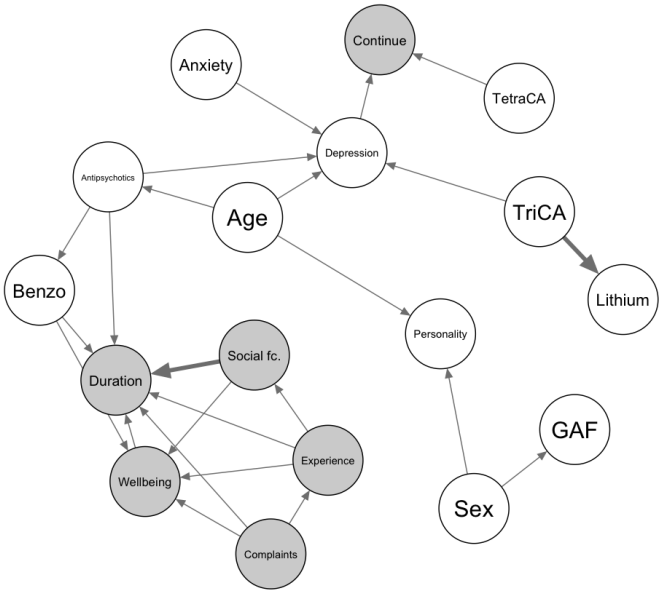


Figure 5.3: Bayesian networks found through constraint-based estimation (with the pcStable algorithm) on the PG data. Outcome variables are highlighted in grey. Dependencies that were also present in the UMCU network are highlighted with bold arrows. Abbreviations: Personality; personality disorder; Social fc: social functioning; Social pr.: social problems according to DSM.

The analysis was repeated with a score-based algorithm (tabu). In the PG data, 28 dependencies were found in the bootstrapped averaged network (see figure 5.4), of which 19 overlapped with the constraint-based network (irrespective of the direction of the dependency). In the UMCU data, only 9 dependencies were

found, of which 5 overlapped with the PG network. The direct dependencies between benzodiazepines, antipsychotic prescriptions and the outcome measures remained present. A direct effect of tricyclic antidepressants on the probability of switching to another type of antidepressant was found in this network, instead of an indirect effect through the DSM diagnosis of a specific depressive disorder, and no dependency on the prescription of tetracyclic antidepressants was found.

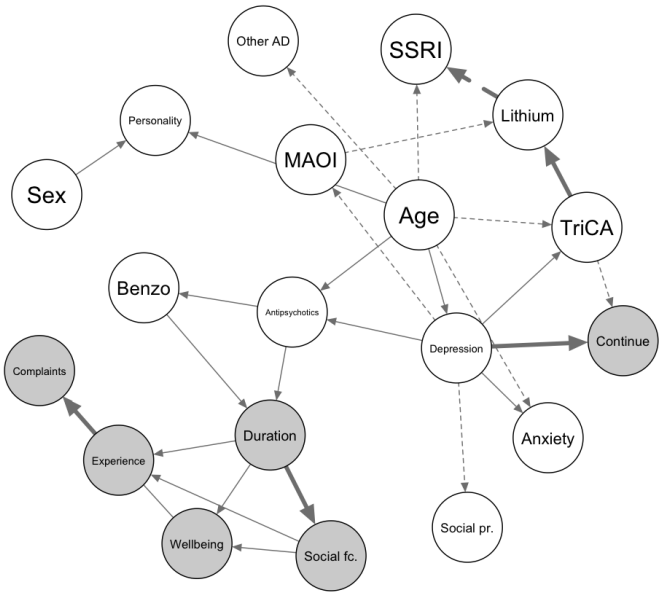


Figure 5.4: Bayesian networks found score-based estimation (with the tabu algorithm) on the PG data. Outcome variables are highlighted in grey. Dependencies that were also found for the UMCU data are highlighted in bold. Connections that were not found with the constraint-based method on the PG data are indicated with dashed lines. Abbreviations: Personality; personality disorder; Social fc: social functioning; Social pr.: social problems according to DSM.

Found dependencies and comparison to expert opinion Below a few examples are given of hypothetical patients and quantification of the dependencies found in the network. These patterns could give some interesting pointers for future confirmatory research. Note that even though sometimes variables are not direct neighbors in the network, they can depend on each other indirectly, in the absence of information on other nodes. For example, if we are unsure whether we are going to diagnose a patient as having depressive disorder as a main diagnosis, the patient's age and DSM classification for social problems can give us extra information about whether the patient will switch antidepressants according to the

score-based network. When we decide on the patient's depression diagnosis, these paths become "blocked" and these variables do not give us extra information on a possible future switch anymore [Peters et al., 2017].

Building on this example, we see that the network predicts that for patients with social problems that are also prescribed antipsychotics, the probability of choosing the correct antidepressant type (not having the switch) increases with 6% when they are prescribed a TriCA instead of another antidepressant type. This could potentially be explained by the fact that the prescription of anti-psychotics suggests severe, possibly psychotic, depression and TriCAs are more effective in more severe depression states, possibly due to anticholinergic effects that more strongly reduce stress or anxiety. For patients without social problems and without antipsychotics prescriptions, this "profit" after choosing a TriCA is even bigger and increases to 11%. A possible rationale could be the beneficial effect of TriCAs being explained by its anticholinergic, sedating properties, which would have a smaller effect on patients already taking antipsychotics [Schneider et al., 2019].

Focusing on the other outcome measures, we see that these are completely determined by each other and the decision to prescribe benzodiazepines and antipsychotics. If our hypothetical patient is treated with both benzodiazepines and antipsychotics, the probability of sufficient prescription duration to experience a clinical effect would be 76%. The probability of well-being having a positive sentiment score in clinical notes would be 52%. However, if our hypothetical patient is not going to take benzodiazepines and antipsychotics, the probability of continuing the initial antidepressant prescription beyond 5 weeks drops to 52%, and the probability of positive well-being scores drops to 46%. There appears to exist some interaction between benzodiazepine and antipsychotic use that strengthens or dominates the effect of the antidepressant prescribed which makes switching less necessary.

The outcome measures prescription duration, core complaints, social functioning, well-being and patient experience were tightly connected in the found networks. For example, a net positive well-being score improved the probability of obtaining a positive score on the core complaints domain with 15.2%. Incorporating the different domains in the network also allows for possible personalized recommendations in future decision support tools. For example, in the score-based network, adding benzodiazepine and antipsychotics to a treatment regimen only improved the probability of a positive net mean score on the complaints domain with 1.4%, but the social domain score improved with 6.9%. A patient that is especially interested in improving social functioning might find information on these outcome domains presented separately in a decision support tool especially useful, in contrast to a single pooled outcome measure.

5.4 Discussion

This work in this manuscript concerned using Bayesian network analysis combined with NLP for pattern discovery in patient characteristics, treatment choices and outcomes during antidepressant trajectories. Several interesting trends were observed in the routinely collected clinical data studied. In the secondary and tertiary

care settings studied, antidepressant choices had a higher continuation rate (66% and 89%) than expected from literature. Bayesian networks based on the data from PG, the secondary care mental healthcare facility, revealed 28 (predictive) dependencies between treatment choices, patient characteristics and outcomes. At UMCU, the tertiary care mental healthcare facility, most of these dependencies could not be replicated.

We have shown that using NLP to preprocess routinely collected clinical data can allow pattern discovery through Bayesian network analysis in a relatively big corpus of patient data. The combination with NLP enables large-scale studies without burdening clinical staff with extra administrative tasks for research purposes, such as separately registering patient prescriptions and specific treatment outcome measures. These tools could be combined in future research for investigating similar exploratory research questions. The possibility of using Bayesian network analysis for confirmatory research is discussed further below.

In patients who switched their antidepressant type during follow-up, switches were quite evenly distributed over other antidepressant types, although switching occurred more often after describing tetracyclic antidepressants or "other" antidepressants (for example buproprion) at PG. Interestingly, these types of antidepressants, frequently prescribed as third-line therapy options, appear slightly less effective in actual clinical practice. This leads to the hypothesis that specific subtypes of depression, perhaps not studied in clinical trials, need different antidepressant working mechanisms. This makes it even more relevant to search for patterns that can predict the right prescription in an early phase of treatment. Unfortunately, follow-up at the tertiary care facility UMCU was limited, possibly explaining the absence of most dependencies found at PG. As the network we are estimating here appears to be sparse and we do not expect variables in the network to have more than five predictive factors, the 735 patient trajectories used for learning at the tertiary care facility should have sufficed [Briganti et al., 2022], thus not explaining the missing dependencies. Another possible explanation could be the specialized nature of the care given at UMCU, with the different types of patients really having another network graph underlying the antidepressant trajectory data, where perhaps different variables should be included.

Patterns discovered in this study should be purely interpreted in an exploratory manner, as Bayesian networks require several assumptions to be met to enable causal interpretations [Briganti et al., 2022]. Two important assumptions are that there should be no selection bias in the data, and there should be no (hidden) confounding variables. These are two assumptions that are very difficult to verify when working with retrospective data. Exploratory analysis did not show a confounding effect of treatment location on model outcomes (data not shown), but to truly fulfill these assumptions a randomized controlled trial should be performed where patients are randomly allocated to a (combination of) antidepressant(s). Such a trial would probably be unattainable in clinical practice because of ethical constraints (assigning a MAO inhibitor with severe potential side effects to someone with mild symptoms of depression, for example).

Nevertheless, a prospective study design with a more transparent process of data collection and structured questionnaires could already offer a lot of improve-

ment on data quality. We studied first-time inpatient antidepressant treatment trajectories, but limited information on the trajectory leading up to this inpatient treatment period was available in the retrospective data, routinely collected during inpatient care. It could possibly add a lot of value to incorporate antidepressant types tried before inpatient treatment into the model. As it is known that individual psychiatrists make treatment choices for therapy resistant depression based on personal experience, studies on the interplay between the specific psychiatrist prescribing the antidepressant, treatment choices made and treatment outcomes are warranted as well [Zimmerman et al., 2004]. Moreover, a prospective study would enable registering which antidepressants were actually administered, in addition to knowing which were prescribed, and a comparison with standard measuring instruments of depressive disorder to formalize findings.

Such a prospective study could also enable the opportunity to include more detailed information on different symptoms patients are experiencing, and possible side effects of antidepressants. Since side effects are an important cause of discontinuation of antidepressants, the incorporation of different types of side effects could be of great value in the decision making about the type of antidepressant and medication adherence. Cipriani et al. (2018) found significant differences between types of antidepressants and continuation rates and also highlighted the importance of strategies to distinguish differences in response to antidepressants on an individual level [Cipriani et al., 2018]. The NLP model used in this study was specifically developed to model broadly defined treatments for diverse groups of patients.

In the further retrospective studies or in the absence of the possibility to collect more detailed information on symptoms or side effects, more specific NLP models such as MedCAT could be used to extract this type of data from routinely collected clinical notes [Kraljevic et al., 2021]. A recent pilot study on using MedCAT for extracting information on cognitive side effect during depression treatment with electroconvulsion therapy showed promising results [Schepper et al., 2022]. These approaches with reuse of clinical data could be of great value for personalized medicine because it will enable learning for a wider spectrum of patient types. For personalized modelling this is needed because the current strict selection criteria for patients to be included in clinical trials limits the extrapolation of study outcomes to individuals in daily practice [Bayes and Parker, 2019].

In the future, networks like the one described in this study could be translated to decision support tools in clinical practice. Individual patients could for example choose the outcomes they are most interested in, and the characteristics that influence the predicted outcomes for individual patients the most could be highlighted [Sevilla, 2021]. A systematic review of Samalin (2018) showed positive effects of shared decision making interventions on medication adherence and depression outcome. A personalized tool to facilitate this process would be of great value [Samalin et al., 2018]. Essential for such a decision support tool is the incorporation of prospectively collected data, and the incorporation of uncertainty estimates. Recent advances in the field of statistics have revealed new possibilities to give these kind of estimates, for example confidence sequences for discrete (conditional) independence relations that are robust under sequential testing [Turner

and Grünwald, 2023]. Incorporating these into a Bayesian-network based clinical decision tool that is prospectively updated would enable offering patients and clinicians robust and up-to-date estimates.

Chapter 6

Outcome Prediction of Electroconvulsive Therapy for Depression using a Bayesian Network Model based on Clinical Information

Yuri van der Does[1], Rosanne J. Turner[1,2], Miel J.H. Bartels[1], Karin Hagoort[1], Aäron Metselaar[1], Floortje E. Scheepers[1], Peter D. Grünwald[2,3], Edwin van Dellen[1,4]

1: University Medical Center Utrecht, Brain Center, Netherlands
2: CWI, Machine Learning group, Netherlands
3: Leiden University, Department of Mathematics, Netherlands 4: Department of Neurology, UZ Brussel and Vrije Universiteit Brussel, Brussels, Belgium

Abstract

We developed and tested a Bayesian network(BN) model to predict ECT remission for depression, with non-response as a secondary outcome.

We performed a systematic literature search on clinically available predictors. We combined these predictors with variables from a dataset of clinical ECT trajectories (performed in the University Medical Center Utrecht) to create priors and train the BN. Temporal validation was performed in an independent sample.

The systematic literature search yielded three meta-analyses, which provided prior knowledge on outcome predictors. The clinical dataset consisted of 248 treatment trajectories in the training set and 49 trajectories in the test set at the same medical center. The AUC for the primary outcome remission was 0.783(95%CI 0.647-0.921), accuracy 0.78, sensitivity 0.67, specificity 0.81, after temporal validation in the independent sample. Prior literature information marginally increased AUC and reduced CI width.

A BN model comprised of prior knowledge and clinical data can predict remission of depression after ECT with reasonable performance. This approach can be used to make outcome predictions in psychiatry, and offers a methodological framework to weigh additional information, such as patient characteristics, symptoms and biomarkers. In time, it may be used to improve shared decision-making in clinical practice.

6.1 Introduction

Depression is a leading cause of disability according to the World Health Organization, affecting one in six people during their lifetime [Kessler et al., 2005, World Health Organization, 2022]. Electroconvulsive therapy (ECT) is the most effective available treatment for severe depression [Lisanby, 2007]. In practice, ECT is usually reserved for patients who show insufficient response to antidepressant medications and psychotherapy, in part because of stigma and anticipated cognitive side effects [Leiknes et al., 2012]. Although highly effective on a group level, a substantial number of patients show no or insufficient response to ECT. There are several factors associated with response to ECT, including age and presence of psychotic symptoms [Van Diermen et al., 2018]. However, in current psychiatric practice, neither systematic assessment of these independent predictors, nor assessment of cumulative predictive value of multiple predictors are routinely used in the decision to initiate ECT for individual patients. As a result, treatment outcome on the individual level remains largely unpredictable.

Clinical decision support systems (CDSSs) are computerized tools which provide clinicians individualized information based on various sources of information, for instance demographic characteristics and information from electronic health records (EHRs). CDSSs make use of prediction models or algorithms for systematic assessment of information. CDSSs are used in several clinical specialties, such as in cardiovascular medicine [ESC Cardiovasc Risk Collaboration et al., 2021, Hageman et al., 2022]. In psychiatry, the availability of CDSSs is modest at best, as was illustrated by Koposov and colleagues [Koposov et al., 2017]. Bright and colleagues give an overview of clinically implemented CDSSs across all medical specialties in a systematic review and meta-analysis of randomized controlled trials. They found that clinicians more likely to appoint the appropriate treatment when informed by CDSSs compared to clinicians who did not use these systems, based on 46 studies across diverse venues and systems (OR 1.57, 95%CI 1.35 – 1.82) [Bright et al., 2012]. A recent Cochrane review by Stacey and colleagues, which assessed the effect of decision aids, reported that patients who used CDSSs were better informed on treatment options, felt more knowledgeable, and were likely to have more accurate risk perceptions [Stacey et al., 2017]. A CDSS which can predict the effect of ECT for individual patients could be useful to inform patients and facilitate shared decision making before treatment is initiated. In order to realize this, a prediction model for ECT outcome is required.

In this study, we developed a personalized effect prediction model for the prediction of remission after ECT, and, secondary, ECT non-response, using a Bayesian network (BN) model. BNs are a combination of intuitive graphical representations of causal or predictive dependencies between variables, and the corresponding underlying quantitative model (for an insightful tutorial aimed at psychopathology researchers see Briganti et al. [2022]). The presence and underlying quantitative model of these dependencies can be derived from data, can be obtained from expert opinion, or both [Arora et al., 2019]. The aim was to predict effect of an ECT trajectory using data which was clinically available before ECT was initiated. We used a systematic review to identify predictors of ECT

outcome to inform a BN with prior knowledge from literature. Subsequently, we used expert knowledge to further design the BN, and tested its performance in clinical data. Finally, we validated the performance of this prediction model in a validation dataset.

6.2 Methods

We used a stepwise approach to create the BN model for ECT outcome prediction: 1) acquiring prior knowledge by performing a literature search for clinically obtainable predictors; 2) Creation and training of a BN using prior knowledge from literature and a clinical dataset; and 3) validation of the trained model in a validation cohort.

6.2.1 Acquiring prior knowledge

Systematic review To acquire high quality prior knowledge from literature, we performed a systematic review on predictors of ECT outcome, in which we only searched for high quality meta-analyses. We performed a literature search in the online libraries MEDLINE and EMBASE according to the PRISMA guidelines for systematic reviews [Page et al., 2021]. The protocol of this review was not registered in advance and is not available for review as such. The question this systematic review addressed was formulated as: "for adult patients undergoing ECT, what are clinical predictors for outcome (response or remission) of ECT". Search terms used were: (ect OR electroconv*) AND predict* AND (remission OR respons* OR outcome*); all published articles available before 16-11-2022 were reviewed. Articles were excluded in screening of title and abstract if they were: non-human, non-English language, when treatment was not ECT and if study design was not a meta-analysis. Systematic reviews without meta-analysis were excluded. When the full-text studies were not available, the authors were contacted. Eligibility criteria for inclusion were studies on predictors of ECT outcomes of which data were readily available at baseline in most patients. These were defined as demographic predictors, clinical assessment predictors, comorbidity predictors, pharmacological or technical ECT aspects predictors. The definition did not include MRI findings predictors, because MRI scans are not performed as standard practice at the start of ECT.

The screening and quality assessment of articles were performed by two independent reviewers (YD and AM), without automation tools. Discrepancies in results were resolved by consensus, or by a third reviewer in case of disagreement (ED). We performed a ROBIS quality assessment to assess the risk of bias in the identified meta-analyses [Whiting et al., 2016]. Only studies with an overall low risk of bias were included in model development.

Data were collected from each individual predictor for both remission and response. Standardized mean differences with 95% confidence intervals (95%CI) were collected for continuous predictors, odds ratios (ORs) with 95%CI were collected for dichotomous predictors. Outcome was defined as "remission" or "response", without further specification, in order to include all relevant studies. When two or

more meta-analyses provided data for a single predictor and outcome, the authors would, after consensus, only extract data from the most relevant meta-analysis available, based on date of publication and quality assessment. Results of the data extraction were used as priors for the BN model.

6.2.2 Bayesian network model development

Study population For the BN model, we used individual patient data from patients who were treated with ECT in the University Medical Center Utrecht (UMCU) in the Netherlands between 1 January 2008 and 27 September 2019. We included all patients receiving ECT for a depressive episode, including patients with bipolar and schizoaffective disorders, who had a discharge letter with conclusion of ECT trajectory available When patients had multiple ECT trajectories, a subsequent trajectory was only included after a time interval of at least 90 days. As many clinical patient characteristics structurally collected in routine clinical care were acquired retrospectively and were used as predictors of outcome. These were age, sex, somatic comorbidity, age of onset of symptoms, duration of depressive episode, ECT naivety, co-morbid personality disorder, severity of depression, psychotic features, catatonic features, and diagnostic context of ECT indication (e.g. depressive disorder, bipolar disorder, or schizoaffective disorder). Clinical patient characteristics were extracted from Electronic Health Records (EHRs). Patient data were anonymized using DEDUCE and therefore the institutional medical ethics review board waived informed consent [Menger et al., 2018b]. Validation of the model was performed in a cohort which consisted of patients who received ECT in the UMCU between 17 July 2018 and 22 October 2021.

Outcomes The primary outcome was remission after ECT. Remission outcome was assessed by deriving the conclusion from the psychiatrist's discharge letter , stating "remission", indicating an absence of depressive symptoms. This dichotomous outcome has been used in meta-analytic research and has clinical usefulness, because it is informative and understandable for both clinicians and patients [Pagnin et al., 2004]. Non-response was assessed as secondary outcome, and was defined as absence of any amelioration of symptoms of depression. This was also assessed by deriving the conclusion from the discharge letter.

Statistical analysis To explore the data used for training the model, group means or proportions for the predictor variables were compared between remission and non-remission using the appropriate hypothesis tests, where we used Bonferroni correction to correct for multiple testing. In case of missing data we used multiple imputation, using IterativeImputer (Python) for the training set and MICE for the validation set (R). To gain insight into the associations and/or causal relations between predictors on multiple levels, a BN was fitted on the UMCU data with the "bnlearn" package in R [Scutari, 2010]. Prior to learning the structure of the network, black- and whitelists were created based on the data derived from the meta-analyses combined with expert knowledge from authors YD, MS and ED. Associations on these lists were either by default included (for the whitelist)) in, or

excluded (for the blacklist) from the network. Adding this sort of prior knowledge can vastly improve the stability of overparameterized networks [Briganti et al., 2022]. On the blacklist, response or remission was excluded as a predictor of other variables in the network, and age and gender were excluded as being dependent on other variables in the network. Somatic comorbidity and cognitive disorder were also excluded as possible direct predictors of non-response or remission. On the whitelist, all predictors except catatonic symptoms, forced care and first ECT trajectory were included as direct predictors of response or remission. Personality disorder was included as a direct predictor of somatic comorbidity, age of onset, relapse, episode duration and psychotic and catatonic symptoms. The structure of the BN was determined through applying the score-based (i.e., aimed at optimizing the predictive performance of the network) "Hill-Climbing" algorithm on the data 100 times through bootstrapping: to improve stability, only dependencies occurring in at least 85% of the bootstrapped networks were included in the final structure [Briganti et al., 2022].

Based on the dependencies found in the BN a hierarchical Bayesian logistical regression model was fitted specifically for predicting response to ECT with the "arm" package in R, as the bnlearn package did not offer fitting such models with prior information. As predictors, all variables found to be associated with response to ECT from the meta-analyses and all variables included as predictors of the outcome variable in the BN were included. Priors on coefficients were chosen to be normal, with mean and standard deviation either estimated through the meta-analysis, or set to 0 and 1 in the absence of prior information; all prior settings can be found in table 6.2.

To generate insights into model performance, 5-fold cross-validation was performed and mean accuracy, ROC-AUC (receiver operator characteristics area under the curve) and corresponding 95% confidence intervals were calculated. ROC-AUC can be interpreted as the ability of the predictor to distinguish between true positive and negative cases (or probability that it will do so correctly). The model was subsequently validated in the independent dataset, where ROC-AUC curves, sensitivity, specificity and calibration curves were calculated to assess the external validity of the model. The model creation and validation were in accordance with the TRIPOD statement [Collins et al., 2015].

6.3 Results

6.3.1 Acquiring prior knowledge

Systematic review We found a total of 1638 articles after removing duplicates. Of these, 1614 were excluded after screening of title and abstract. A total of 24 articles were sought for retrieval, of which six were eventually not available. Full text screening for eligibility was performed in 18 articles, of which five were included [Van Diermen et al., 2018, Oxlad and Baldwin, 1996, Haq et al., 2015, Havaki-Kontaxaki et al., 2006, Kho et al., 2003]. One additional article was found while screening manually for relevant references in the included articles [Heijnen et al., 2010]. Flowchart and quality assessment summary are reported in the supplemen-

tary material. We included three meta-analyses with an overall low risk of bias [Van Diermen et al., 2018, Haq et al., 2015, Heijnen et al., 2010]. All meta-analyses showed overlap of included studies. For the predictors psychotic symptoms, age, melancholic symptoms and depression severity, two studies reported data on response outcome [Van Diermen et al., 2018, Haq et al., 2015]. For these predictors, we only extracted data from the meta-analysis of Van Diermen and colleagues, because this meta-analysis was more recent and was assessed as having an overall lower risk of bias in the ROBIS quality assessment [Van Diermen et al., 2018]. For the predictor medication failure, data for remission outcome from Heijnen and others and data for response outcome was extracted from the meta-analysis from Haq and others [Heijnen et al., 2010, Haq et al., 2015]. Extracted data of predictors of ECT outcomes were reported in supplementary material.

6.3.2 Bayesian network model development

Training and validation datasets We included a total of 248 treatment trajectories of patients receiving ECT at the UMCU between 2009 and 2019 in the training dataset, of which 90 (36%) were classified as remission and 63 (25%) as non-response. The validation set consisted of 49 independent treatment trajectories, of which 12 (24%) were classified as remission and 6 (12%) as non-response. Summary statistics (mean values or proportions for patients with and without remission) of both datasets can be found in table 6.1. In the training data, nine treatment trajectories had an unknown episode duration. In the independent validation cohort, 28 cases had missing data, for the relapse, episode duration, catatonic symptoms and age of onset predictors. All missing variables were imputed. Age was the only predictor with statistically significant differences after Bonferroni correction between non-responders and patients with response/remission, with higher age being associated with higher remission rate, in both the training set and in the test set ($p = 0.0005$ and $p = 0.0016$ respectively).

Table 6.1: Summary statistics. Mean values (continuous variables) or proportions (categorical variables) of predictors of patients with the corresponding property (dichotomous variables) and p-values based on a t-test for continuous data or Fisher's exact test for dichotomous data of patients with or without remission in the train and test data. Missing variables were excluded column-wise. For dichotomous variables, counts of patients with the corresponding property are included between brackets. For the test set, non-imputed data are shown.

	Training set (n = 248) Mean		p-value	**Validation set (n = 49)** Mean		p-value
	No remission (n = 158)	Remission (n = 90)		No remission (n = 37)	Remission (n = 12)	
Relapse (y/n)	0.80 (126)	0.84 (76)	0.40	0.882 (30)	1 (11)	0.558
Episode duration (months)	24.5	18.0	0.039	27.9	14.2	0.109
Age (years)	48.5	55.9	0.000527	50.4	69.2	0.0016
First ECT (y/n)	0.772 (122)	0.733 (66)	0.538	0.667 (22)	0.700 (7)	1
Psychotic symptoms(y/n)	0.241 (38)	0.333 (30)	0.139	0.344 (11)	0.364 (4)	1
Catatonic symptoms (y/n)	0.070 (11)	0.100 (9)	0.469	0.094 (3)	0 (0)	0.558
Severe depressive episode (y/n)	0.791 (125)	0.888 (79)	0.119	0.405 (15)	0.333 (4)	0.743
Forced care (y/n)	0.045 (7)	0.033 (3)	0.751	0.135 (5)	0 (0)	0.315
ECT trajectory number (count)	1.20	1.18	0.792	1.32	1.167	0.424
Somatic comorbidity (y/n)	0.430 (68)	0.400 (36)	0.689	0.378 (14)	0.333 (4)	1
Female (y/n)	0.639 (101)	0.733 (66)	0.159	0.730 (27)	0.667 (8)	0.721
Age of onset (years)	33.8	36.6	0.166	38.81	49.00	0.211
Medication failure (y/n)	0.082 (13)	0.133 (12)	0.272	0.162 (6)	0.083 (1)	0.665
Personality disorder (y/n)	0.310 (49)	0.178 (16)	0.0246	0.297 (11)	0.333 (4)	1
Bipolar disorder (y/n)	0.095 (15)	0.089 (8)	1	0.108 (4)	0.0833 (1)	1
Cognitive impairment (y/n)	0 (0)	0.011 (1)	0.363	0.081 (3)	0.0833 (1)	1

Table 6.1, continued

	Training set Mean		p-value	**Validation set** Mean		p-value
	No remission	Remission		No remission	Remission	
Major depressive disorder (y/n)	0.867 (137)	0.856 (77)	0.849	0.865 (32)	0.833 (10)	1
Schizoaffective disorder (y/n)	0.0380 (6)	0.0444 (4)	1	0 (0)	0.0833 (1)	0.245

A total of five patients with were included in both the training set and the validation set, because they had multiple ECT trajectories. For completeness, we also ran the analyses without these trajectories in the validation set (n=44) (summary table is available in supplementary data).

Bayesian network model and hierarchical model for predicting remission
The Bayesian network found with the Hill-Climbing algorithms revealed no new direct dependencies between predictor variables and outcome variable remission that were not already present on the whitelist provided by the experts and meta-analysis (supplementary figure 1.)

The model containing solely priors from literature had an AUC of 0.63 (95% CI 0.56 – 0.70) and an accuracy of 0.63 for predicting remission of UMCU patients in the training set. After updating the model coefficients using the data of UMCU patients, the AUC was 0.629 (values 0.505 – 0.763 observed in 5-fold cross-validation) and the classification accuracy estimated through 5-fold cross-validation was 0.637. The trained hierarchical Bayesian logistic regression model and an overview of priors can be found in table 6.2. For completeness, a model containing only patient derived data with no prior information, showed a mean AUC of 0.59 (values 0.53 – 0.82 observed in 5-fold cross-validation) and a mean accuracy for remission of 0.66 (values 0.60 – 0.81 observed in 5-fold cross-validation).

Table 6.2: The final logistic regression model for predicting remission, and the priors used for fitting the model. NA indicates "not available": prior estimates of mean and sd were available for four out of 18 predictors. 13 predictors were selected to be included in the final model through the Bayesian network analysis.

Predictor	**Mean estimate**	**Sd estimate**	**Coefficient**	**Coefficient SE**
Medication failure	-0,65393	0,143841	-0,55991	0,13831
Severe depressive Episode	-0,097	0,05	-0,08764	0,049603
Age	0,258	0,063	0,052953	0,013273
Psychotic symptoms	0,383901	0,116449	0,388671	0,109663
Personality disorder	NA	NA	-0,50342	0,34052
Bipolar disorder	NA	NA	-0,46309	0,657538
Relapse	NA	NA	-0,37544	0,41086
ECT trajectory number	NA	NA	-0,2943	0,270555
Major depressive disorder	NA	NA	-0,18883	0,618723
Schizoaffective disorder	NA	NA	-0,10911	0,700235
Age of onset	NA	NA	-0,02874	0,013801
Episode duration	NA	NA	-0,01756	0,007355
Female	NA	NA	0,489148	0,295852
First ECT	NA	NA	NA	NA
Catatonic symptoms	NA	NA	NA	NA
Forced care	NA	NA	NA	NA
Somatic comorbidity	NA	NA	NA	NA

Table 6.2, continued

Predictor	**Mean estimate**	**Sd estimate**	**Coefficient**	**Coefficient SE**
Cognitive impairment	NA	NA	NA	NA

Validation of the updated model on the 49 patients in the validation set resulted in an AUC of 0.783 (95%CI 0.647-0.921), with an accuracy of 0.78, with for predicting remission. Remission occurred in 12 patients: there were 30 true negatives (61.2% of the validation set), 8 true positives (16.3% of the validation set), 4 false negatives (8.2% of the validation set) and 7 false positives (14.3% of the validation set). The corresponding sensitivity of the model assessed on the validation set was 0.67 and the specificity 0.81. A model without prior information resulted in an AUC of 0.773 (95%CI 0.623-0.922) and an accuracy of 0.75. For completeness, validation of the model excluding five patients who also had a ECT trajectory in the training set resulted in an AUC of 0.686 (95%CI 0.513-0.859). Summary data are reported in the supplementary material.

An overview of data of misclassified cases of remission is given below in Table 6.3. False negative cases (patients predicted as not achieving remission after ECT while in reality they did), were generally younger, without psychotic symptoms. False positive cases were generally older, with psychotic symptoms, and did not have personality disorders, which were strong predictors in the final model for remission (see Table 6.2). These false positives could possibly explain the decreasing trend in the calibration plot in the bins with the highest predicted probabilities of remission, where the model overestimates the success probabilities (see Figure 6.1).

Table 6.3: Group means (for continuous data) or proportions with the corresponding property (for dichotomous data) for misclassified cases in the validation set, split based on false negative or false positive misclassification.

	False negative (n = 4)	**False positive** (n = 7)
Relapse	1	1
Episode duration	24	10
Age	53.3	76.7
First ECT	0.667	0.714
Psychotic symptoms	0	0.571
Catatonic features	0	0.143
Severe depressive episode	0.500	0.143
Forced care	0	0.286
ECT trajectory number	1.00	1.14
Somatic comorbidity	0.500	0.429
Female	0.500	1.0
Age of onset	33.7	65.6
Medication failure	0	0
Personality disorder	0.750	0.143

Table 6.3, continued

	False negative	**False positive**
Bipolar disorder	0.25	0.00
Cognitive impairment	0	0.143
Major depressive disorder	0.750	1.00
Schizoaffective disorder	0	0

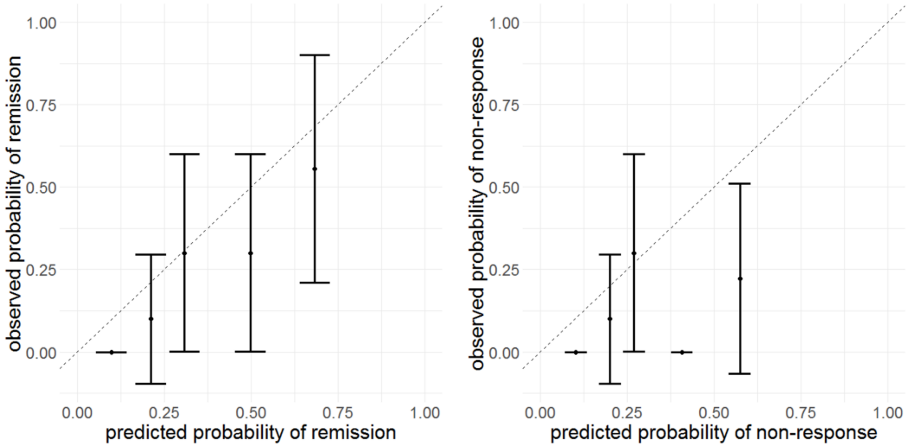


Figure 6.1: Calibration plots of the model for prediction remission (left) and non-response (right) on the validation set. Patients in the validation set were divided into 5 or 4 equal bins, depending on the probability of remission or non-response as predicted by the model. For those bins, the observed probability of remission or non-response and corresponding upper- and lower confidence bounds were estimated based on the patient data, resulting in the figures depicted above.

Bayesian network model and hierarchical model for predicting secondary outcome non-response In the training set, 63 (25%) of trajectories was classified as non-response, in the . The AUC for the model for predicting the secondary outcome, ECT non-response, was 0.644 (values 0.603-0.675 observed in 5-fold cross-validation), with a classification accuracy estimated through 5-fold cross-validation of 0.746. In the validation set, non-response occurred in 6 out of 49 patients. Validation of the updated model resulted in an AUC of 0.624 (95%CI 0.377-0.871) for predicting (non-)response, with an accuracy of 0.78, a sensitivity of 0.33 and a specificity of 0.84. The trained hierarchical Bayesian logistic regression model and an overview of priors for non-response can be found in supplementary files.

6.4 Discussion

In this study, we created and temporally validated BN model to predict outcome after ECT for depression, using prior knowledge from literature combined with single center clinical patient data. We found a mean AUC of 0.629 (values 0.505 – 0.763 observed in 5-fold cross-validation) for the training set and an AUC for the validation set of 0.783 (95%CI 0.647-0.921) for predicting remission to ECT. These findings suggest that probability of remission of a depressive episode using ECT can be reasonably well estimated with readily available clinical predictors for individual patients. For non-response, we found a mean AUC of 0.644 and an AUC for the validation set of 0.624 (95%CI 0.377-0.871).

High-quality meta-analyses are considered as the highest level of evidence in evidence-based medicine. However, one of the downsides of meta-analyses is that the aggregated data have no direct clinical value to individual patients [Berlin and Golub, 2014]. In this study we used the knowledge from the best meta-analyses available in a BN model to create a clinical decision support system which calculates personalized outcome predictions for ECT. Although these methods have been studied before, this study is, to our knowledge, the first to investigate the outcome of ECT using a BN. Previous studies of BNs in psychiatry focused on dementia and cognitive impairment [Jin et al., 2016, Gross et al., 2018, Moreira and Namen, 2018]. BNs are mostly used in the fields of cardiology and oncology, but have not yet been adopted as a standard technique in medical decision making. One explanation is that previous publications on BNs mostly emphasized technical aspects instead of clinical usefulness [McLachlan et al., 2020, Kyrimi et al., 2021]. We found that the addition of prior information to our model increased the AUC for remission marginally and marginally reduced CI width, from an AUC of 0.773 (95%CI 0.623 – 0.922) in the no priors model, to an AUC of 0.783 (95%CI 0.647 – 0.921) in the final model. Based on these findings, including prior information hypothetically decreases the sampling variability in a model, by increasing the number of samples of which data is derived. An additional value of priors is that they can be used as an extra validation of findings in a study cohort. If significant discrepancies are observed, further investigation on bias is warranted.

Our findings showed that the presence of psychotic symptoms was a strong predictor for remission, as well as the absence of a personality disorder and the absence of medication failure. These findings were expected because previous studies which identified these variables were used as prior knowledge in our study [Van Diermen et al., 2018, Heijnen et al., 2010]. Several studies found reduced effectiveness of ECT in patients with personality disorders [Yip et al., 2021, Prudic et al., 2004]. Interestingly, higher age was no statistically significant predictor for remission in our study, which is contrary to previous research [Van Diermen et al., 2018]. In our analysis of misclassification, younger patients did not have psychotic symptoms, and many elderly patients did. However, our findings were based on a single sample, and selection bias may have had an effect here. The secondary outcome of non-response did not yield significant results.

In the misclassification analyses and calibration plots for both remission and non-response, we found a decreasing trend in the plots in the higher predicted prob-

abilities, resulting in an overestimation of success observed probabilities (figure 1). Specificity was relatively high, but several cases were falsely positive, resulting in low sensitivity. We infer that the dataset may be confounded. However, because of the small sample size of the validation cohort, we cannot assess to what extent. Exploratory analyses of potential confounders, preferably in a larger validation cohort, may yield additional clinical predictors. Next to clinical and demographical parameters, several previous studies reported on biomarkers as predictors of ECT outcome, including MRI, EEG and genetic findings [Luykx et al., 2022, Levy et al., 2019, Simon et al., 2021]. Hypothetically, the accuracy of our model could be increased by including these predictors. However, the problem with these data is that these are not routinely obtained in clinical practice, and therefore often unavailable for the treatment decision about ECT. Therefore, although we were unable to include biomarkers in the model due to unavailability in our data, the clinical model presented here may be easier to implement in clinical practice than a model based on biomarker data.

A hypothetical "real-life" CDSS for ECT outcome prediction would be available for all patients who were eligible for ECT, including patients who were treated previously. Therefore, such a model would include patient data of all previously performed ECT trajectories. As an illustration, we performed the additional analysis of remission for the validation set with the exclusion of 5 patient trajectories of patients who already were included in the training set with a previous ECT trajectory. This resulted in an AUC of 0.686 (95%CI 0.513-0.859) without these trajectories, compared to 0.783 (95%CI 0.647-0.921). The increase in AUC may be attributable to the fact that ECT treatment was repeated in these specific patients (resulting in a new validation set trajectory) because they responded successfully to ECT before (in the training set). Prospective replication of these findings is necessary to investigate the effect of selection bias in these findings.

Although outcome prediction of ECT may benefit shared decision-making, prospective studies are necessary before this model can be implemented as CDSS in standard practice. For example, the subjective experience and needs of individual patients are essential for treatment decisions but were not included here. Moreover, in our sample, the decision to initiate ECT was already made. This resulted in a selected population of patients who were willing to undergo ECT. To assess clinical usefulness, it is necessary to also analyze the patients who decide not to start ECT, and why this decision is made. Misclassification bias may arise after implementation if treatment decisions are made differently because they are informed by a CDSS, and this adaptive change in decision making is not accounted for. One solution for this potential bias is a stepped-wedge cluster randomized controlled trial, in which the CDSS intervention (and its impact on outcomes) is gradually introduced and evaluated at sites [Hemming et al., 2018]. Another factor is the unknown generalizability of findings from our single center study at a university hospital to other treatment settings. We speculate that this could have resulted in an increased severity of depression in our sample, and maybe in other unknown selection biases. An (inter)national, multicenter trial could increase generalizability of our current findings.

Our model did not include adverse effects of ECT. This was due to the fact

that adverse effects were not recorded systematically, which may have led to a reporting bias. Adverse effects of ECT consist of amnesia, headache and nausea and occur in most patients during treatment [Andrade et al., 2016]. Adverse events may be mild, but can also be a reason to halt ECT prematurely, for example in the case of severe amnesia or delirium. Halting treatment may consequently influence the outcome. We hypothesize that there may also be dependencies between these predictors, and that these could be incorporated to the BN model. Additionally, the inclusion of data generated during each session of ECT, such as seizure duration could be used to predict outcomes more accurately during the treatment. However, this would require a model with repeated measurements, with updated probabilities after each session. This approach could guide psychiatrists and patients in their decision to continue, stop or alter frequency of ECT. We aim to expand our model to include these factors and to further test for generalizability in future work.

We used a systematic review of meta-analyses for the collection of prior knowledge. A downside of this method is missing data of recent studies which are not yet included in a meta-analysis. Another problem is that a single study is included in more than one meta-analysis, and that meta-analyses on the same subject reported different outcomes. We considered risk of bias smallest if we analyzed the searches of multiple research groups and selected the one meta-analysis with the highest quality, with the potential risk of sacrificing some recency of data. We used clinical discharge letters with the final outcome of ECT to define the outcomes remission and response. Quantitative assessment of depression, for instance using the Hamilton Rating Scale for depression (HRSD) or Montgomery–Asberg Depression Rating Scale (MADRS), is often used in clinical trials [Van Diermen et al., 2018, Hamilton, 1967, Montgomery and Åsberg, 1979]. Outcomes remission and (partial) response are defined using a reduction of the score by a certain percentage, or below an arbitrary threshold. The potential upside of this approach is that, in theory, treatment can be evaluated objectively. However, there is an ongoing debate about the use of the reliability and validity of depression instruments [Fried et al., 2022]. One of the hypothetical downsides of depression instruments is that the score is comprised of several symptom clusters. An equal reduction in scores of two patients after ECT may not resemble the same effect. Additionally, in clinical practice, standardized application of quantitative assessments requires additional time and training of staff. Therefore, we chose to use the most clinically relevant outcome assessment, the conclusion of the discharge letter. This outcome included both clinician assessment and subjective patient experience. In 23 cases, we had missing data on clinical variables. We used multiple imputation to make optimal use of data. Although multiple imputation is superior to complete case analysis regarding potential bias, it may influence model performance [Steyerberg, 2009].

Conclusion In this study, we found that a BN model comprised of prior knowledge and clinical data can predict remission of depression after ECT with reasonable performance. This approach can be used to make outcome predictions in psychiatry, and offers a methodological framework to weigh additional informa-

tion, such as patient characteristics, symptoms and biomarkers. In time, it may be used improve shared decision-making in clinical practice.

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.

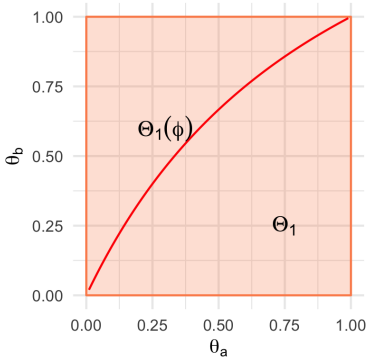


Figure 7.1: Parameter space $\breve{\theta}_a|Y^{(j-1)}$ and $\breve{\theta}_b|Y^{(j-1)}$ are estimated in, in 2×2 table without strata; either through placing a beta prior on the entire unit square (in light orange) and calculating the posterior mean with all data up to and including time $j-1$ or through 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.

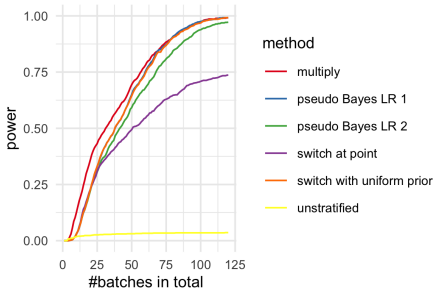


Figure 7.2: Power for rejecting the null at level $\alpha = 0.05$ that the odds ratio in all strata equals 1 estimated with 1000 repeated experiments for various e-variable combination methods. 40 batches were collected in each of three strata (so maximum sample size was $m = 120$) and sampling was stopped 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 $P_{\theta^*_{a,k},\theta^*_{b,k}}$.

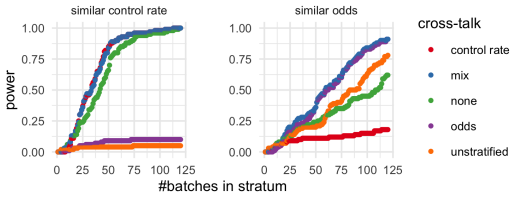


Figure 7.3: Power for rejecting the null hypothesis at level $\alpha = 0.05$ that the odds ratio in all strata equals 1 estimated with 100 repeated experiments for various types of cross-talk. 40 batches were collected in each stratum and sampling was stopped as soon as $E^{(m)} \geq \frac{1}{\alpha}$. 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 $4, 4.01, 2.95$.

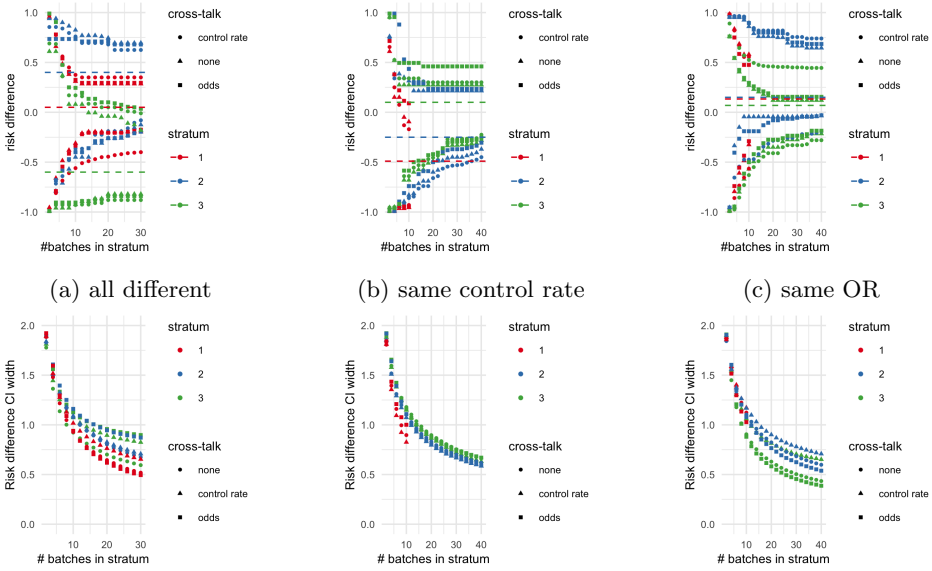
Then, with $\mathbf{E}$ denoting expectation under this distribution, the e-process $E^{(m)}$ defined by multiplication as in (7.8) and the MIX e-process $E^{(m)}$ as above with constituent e-processes defined multiplicatively as in (7.8) both achieve:

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

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.

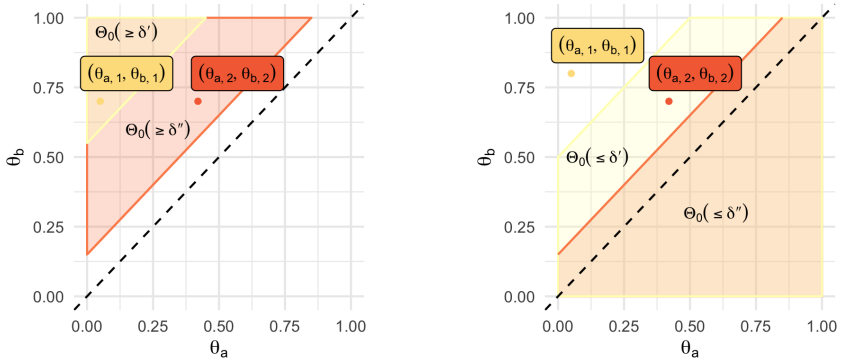
Let us denote the e-process consisting of the e-variables for testing $\theta_k \in \Theta_0(\geq \delta)$ in each stratum combined, using any of the methods described above in Section 2, as $E_\delta^{*(m)}$. The one-sided confidence interval for the minimum effect can be defined as:

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

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).

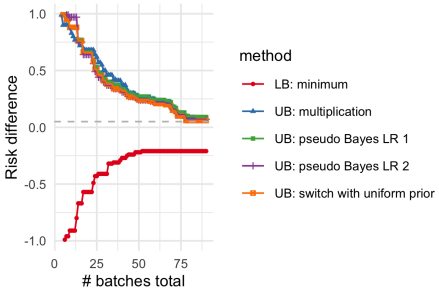


Figure 7.6: Example of confidence sequences for the lower- (LB) and upper (UB) bounds of the minimum effect. 30 observations were made in each stratum, and the real differences were 0.5, 0.4 and 0.05. With the switch method, a uniform prior ranging from $m_{\text{switch}} = 5$ until 30 was applied. With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

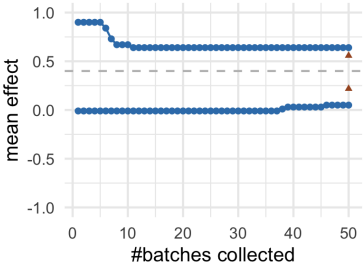


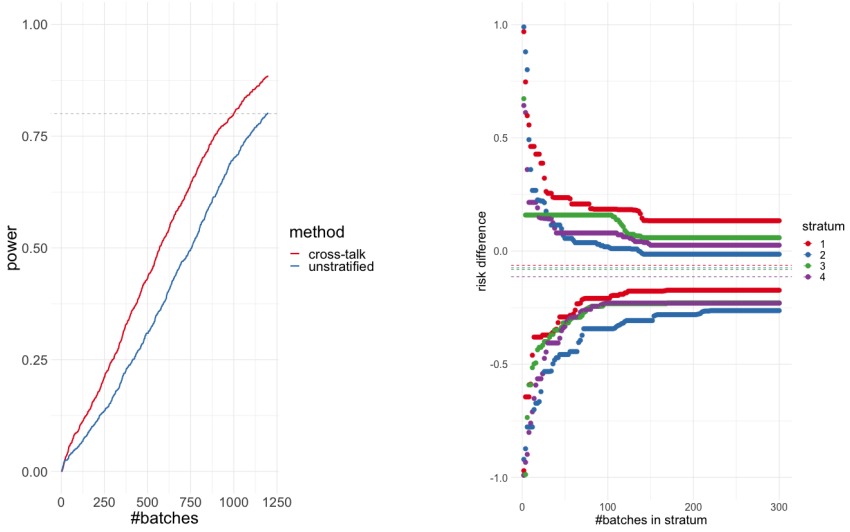
Figure 7.7: Simulated example of 95% confidence sequences for the mean effect across subpopulations. 25 observations were made in each stratum, and the real risk difference of 0.4 was homogeneous across subpopulations. The confidence sequence for the mean effect is plotted alongside the Miettinen-Nuninen 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.

Chapter 8

Discussion

In this chapter, the work described in the other chapters of this thesis is reviewed concisely, and placed in the context of other recent and related developments and the overarching research question "how can one perform real-time research in healthcare using routinely collected clinical data?". Open problems and directions for future work are discussed.

8.1 Implementations of safe, anytime-valid inference

To work toward enabling inferential statistics for real-time research, in chapters 2, 3 and 7 we studied and developed a generic analytical type of e-variables and the corresponding confidence sequences for comparing two or more data streams. We specifically implemented these for studying *categorical* data, for example for the well-known 2×2 contingency table test setting and the *Cochran-Mantel-Haenszel* test setting. For these settings, our "simple" e-variable proved to come very close to the GRO measure, depending on the hyperparameter settings chosen, which determine the speed at which the e-variables "learns" the true data generating distribution (in case of a simple alternative, our e-variable coincides with the relative GRO measure). Directions for future work should concern studying the performance of this generic simple e-variable outside the categorical setting, for example for count data or continuous data, where we know that it does not provide the GRO measure. As in these settings, calculating a GRO e-variable analytically is often impossible and approximating it can be computationally heavy, our simple definition might provide an interesting feasible alternative in some scenarios (a first step was made by Hao et al. [2023]). An extensive overview of testing scenarios, and comparison of available (GRO or universal-inference based [Wasserman et al., 2020]) approaches with respect to power would be of significant added value for applied researchers wanting to apply safe, anytime-valid inference.

Other developments around GRO e-variables for categorical outcomes In the work in this thesis, the problem of sequential testing on categorical stream data was approached in a block-wise manner, conditioning on the number of data entries per group collected in a block. A different approach was developed concurrently by Adams [2020] and a variation is considered by Hao et al. [2023], among

others. They instead condition on ("fix in advance") the number of successful outcomes observed, which yields an especially elegant analytical expression for this conditional GRO e-variable. This approach might be less applicable to common substantive research designs, where funds are allocated for the inclusion of a set number of participants or study units in advance. On the other hand, for companies executing A/B testing on a large scale, it might be especially interesting, for example re-evaluating the performance of two web page designs after a certain number of sales has been made. A detailed comparison concerning power and expected sample sizes of the methods in this thesis and conditional e-values and development of an accompanying tutorial would be of great added value to substantive researchers that need to choose between the two approaches when setting up an inferential study with real-time monitoring.

The work in chapter 7, where e-variables for stratified data and confidence sequences for subgroups of patients were developed, already hinted at the need for *safe, anytime-valid logistic regression*; a setting very common in clinical research. A very interesting first step toward this, using an idea similar to the "simple" e-variable presented in this thesis, was recently presented by Grünwald et al. [2022b]. In this work, an e-variable for testing conditional independence of any outcome variable Y of some (treatment) characteristic variable X given other variables Z is presented and illustrated in a logistic regression setting. The e-variable relies on an accurate "Model-X assumption": the full model or an accurately enough estimate of the distribution of X given Z should be available for the e-variable to remain valid. Nevertheless, since in healthcare practice X and Z often would be treatment and patient characteristics, this is something that can be estimated with retrospectively collected data, outside the costly clinical trial settings. Extension to full logistic regression, including confidence sequences of the model parameters, is still an open problem.

Computational limitations Despite the "simple", analytical form of the e-variables studied in this thesis, we did run into some computational limitations during the work in chapters 3 and 7 which could be improved upon in future work. For example for the confidence sequences described in this thesis, upper- and lower bounds were determined by calculating e-values for a precise grid of divergence parameter values. For the universal-inference based minimization for the confidence sequences over several strata in chapter 7, iterative minimization over multiple parameters for each stratum still limits the number of strata we can implement our ideas for. In future work, statistics and computer science experts should join forces to explore how these calculations and optimizations could be carried out in a more efficient way, enabling more precise results and more flexibility in study designs, required for analyzing sets of clinical data with many predictor or stratification variables.

8.2 Knowledge discovery in psychiatry

Chapters 4, 5 and 6 describe exploratory analysis of the EHR data of the clinical psychiatry departments at UMC Utrecht (UMCU) and Parnassia Groep (PG),

with the goal of exploring how a wide array of routinely collected clinical data can be used for knowledge discovery, eventually in an automated, real-time setting. In chapter 4, the focus was on defining clinically relevant psychiatric outcome measures for information extraction and knowledge discovery processes in close collaboration with clinicians, and the development of a corresponding text mining model based on word embeddings [Menger et al., 2018a]. The selected topics concerned psychiatric core complaints, social functioning, general well-being and patient experience. In chapter 5, a Bayesian network analysis was performed at UMCU and PG in a very heterogeneous group of patients who all were treated with antidepressants during their admission. This analysis combined the information extraction pipeline developed in chapter 4 with patient and treatment characteristics available from structured (tabular) data sources in the EHR.

The exploratory analysis at PG highlighted several interesting possible associations and showed that treatment outcome topics were closely connected. These findings point towards the existence of a *tipping point* in the mental health state of psychiatry patients: if one would succeed in positively influencing one of the aspects of a patients mental health, such as suddenly having many positive social interactions, this might further positively influence the other aspects of one's mental state and the probability of recovery. Nevertheless, besides the strong connections between treatment outcomes, many of the associations between patient characteristics and treatment outcomes found at PG could not be replicated at UMCU. Possibly the study at UMCU was underpowered to find the relatively small effects of patient characteristics on treatment outcomes. Another explanation could be that the nature of mental illness of patients at UMCU is substantially more severe than at PG, and that in these severely ill patients other processes play an important role in determining treatment outcomes than basic patient and treatment characteristics, for example strict supervision in upholding activities of daily living, or a certain interactions with particular types of caregivers.

In chapter 6, a different approach toward network analysis was taken. Here, a more homogeneous group of patients was studied, namely patients receiving electroconvulsive therapy for a depressive episode. For this select group of patients, plenty of prior studies were performed and these were, together with expert knowledge, incorporated in the modelling process through systematic review. Adding this prior information to the Bayesian network and underlying logistic regression model for predicting remission improved prediction accuracy and resulted in good performance for predicting remission in a temporal sample for validation.

Future, prospective studies or even clinical trials to confirm the findings from these exploratory studies are warranted for at least two major reasons. First, using texts written during routine clinical care might be a source of *reporting bias*: association may appear especially positive or negative for certain groups of patients due to under- or overreporting. Second, to be able to report actual causal associations instead of predictive associations, the possibilities of selection bias and the presence of hidden variables should be excluded, which is nearly impossible in a retrospective setting [Briganti et al., 2022].

Innovations in clinical psychiatry Nevertheless, despite the wide array of predictors included in the models in chapter 6, predictive accuracy could still potentially be improved upon, indicating that important predictive information was still missing in the datasets extracted from the routinely collected data in the electronic health records. An important future development could be linking data from wearables and other smart devices to the DHT [De Looff et al., 2019]. This would also enable asking patients for feedback about their mental state and thoughts about the treatment process in a fast and accessible manner, information that was now incorporated in for example the models described in chapters 4 and 5 entirely as written down by a third person, the clinical staff.

Another possibility could be the improvement of information extraction for knowledge discovery in routinely written clinical text: recently, a lot of exciting new possibilities have emerged, such as improvements of the open-source, easily shareable MedCAT (medical concept annotation tool) model [Van Es et al., 2023]. MedCAT is based on (often standardized) medical concept databases, such as SNOMED and UMLS [Spackman et al., 1997, Bodenreider, 2004], and offers the possibility to refine these concept databases based on a local text corpus in an accessible web-based interface. This makes it especially suitable for collaborating on an information extraction project with clinicians, where clinicians through the web-based interface can actively take part in the model training and evaluation process.

It is evident that the final product (the fitted Bayesian networks) of the research described above does not complete the entire process of clinical knowledge discovery: the causal graphical models and conditional probability tables comprising the Bayesian networks are far too complex to directly use in clinical decision support tools. Future research should concern converting these Bayesian networks combined with patient characteristics *and interests* into a tool for patient-tailored advice. One interesting solution for this could be to focus on the amount of information that is passed through the various patient characteristics in the network, selecting the most important paths and converting this information into natural language [Sevilla, 2021]. Combining these kinds of natural language generation models with uncertainty estimates would be a logical next step in working toward Bayesian networks as decision support tools.

8.3 Federated learning in Psychiatry and healthcare in general

To investigate the suitability of the methods described in chapters 2 and 3 for more complex (and realistic) medical research questions, in chapter 7 we studied implementing anytime-valid confidence intervals for a psychiatry use-case where we stratify patients into small groups, based on the hypothesis developed in chapter 5. The confidence sequences we developed offer exciting new possibilities, such as sequentially estimating a mean, minimal or maximal treatment effect (for any effect size notion, such as relative risk, risk difference, odds ratio, and so forth) across subpopulations, and sequentially estimating many confidence intervals in separate subpopulations. Our new algorithms in itself showed very promising results:

through combining safe, anytime-valid inference with machine learning techniques such as cross-talk and pseudo-Bayesian averaging we achieved clinically *realistic* sample sizes with corresponding precise enough, anytime-valid effect estimations.

In future work, these algorithms could vastly alleviate the complexity of multi-center clinical trials and research projects, as the anytime-valid property not only ensures that study results are valid *within* one trial, but also when *combining safe confidence sequences between study centers*. Implementation in such a *federated* setting is straightforward: e-values that summarise the evidence for the hypotheses tested based on all local patient data combined (stored as floating point numbers) to construct confidence sequences can computed locally. Only these numbers have to be shared with a central location to compute the study-level confidence sequences, in principle omitting any identifiable patient data leaving the local study centers. A first "living" meta-analysis using this setup has already been performed to investigate the effect of preventive vaccination of healthcare workers to protect them against COVID-19 infections [Ter Schure et al., 2022].

Future of the digital health twin The steps taken so far within the works in this thesis, for enabling research with healthcare data in real-time, and in the EPI consortium have hopefully brought us a little bit closer to a world where personalized recommendations are standard, while still ensuring privacy of patients. One major limitation in achieving this, that was also encountered during the substantive work in chapters 4, 5 and 6 of this thesis, is the "data-readiness" of mental health centers. Although at UMC Utrecht and Parnassia Groep an established pipeline from EHR to research data was already present, data were not stored in a homogeneous format, which caused the preprocessing to become an elaborate process. Developments such as implementation of the FHIR (Fast Healthcare Interoperability Resources) framework, a standard for information exchange between healthcare providers, could improve the threshold for data-readiness at healthcare institutes significantly [Leroux et al., 2017]. Hopefully, the soon-to-arrive first proofs of concept of the EPI framework and similar initiatives will entice other clinical institutes to work towards data-readiness as well, such that we can really start working toward personalized recommendations in clinical practice.

Nederlandse Samenvatting

Dit proefschrift, *veilige altijd valide inferentie: van theorie naar implementaties voor wetenschappelijk onderzoek in de psychiatrie*, gaat over het doorontwikkelen van een nieuw paradigma in de statistiek, veilige altijd valide inferentie (SAVI), en het toewerken naar een toepassing hiervan in wetenschappelijk onderzoek in de psychiatrie. De gangbare, meest gebruikte methoden voor statistische inferentie (het doen van uitspraken over de gehele populatie op basis van een kleine steekproef) zijn niet geschikt voor flexibele onderzoeksopzetten. Voorbeelden van dit soort opzetten zijn continue analyse van onderzoeksdata en projecten waarbij algoritmes gedeeld worden tussen verschillende instanties. In de eerste twee hoofdstukken van dit proefschrift wordt onderzocht hoe SAVI doorontwikkeld kan worden voor enkele veel voorkomende onderzoeksvragen in dit soort flexibele onderzoeksopzetten. In de drie daaropvolgende hoofdstukken wordt onderzocht hoe data verzameld tijdens psychiatrische zorg geschikt kan worden gemaakt voor continue analyse in meerdere ziekenhuizen, en wordt exploratief gezocht naar patronen in deze retrospectieve data met Bayesiaanse netwerkanalyse. Tot slot worden in het afsluitende hoofdstuk de nieuwe statistische methoden verder ontwikkeld voor het beantwoorden van onderzoeksvragen met een vorm zoals de vragen die volgden uit de Bayesiaanse netwerkanalyse, en wordt geillustreerd hoe een flexibele onderzoeksopzet om een van deze vragen te beantwoorden er uit zou kunnen zien.

Leren in meerdere instellingen Data verzameld tijdens routinematige zorg en/ of medisch onderzoek blijft nu meestal exclusief binnen de zorginstellingen waar de zorg is verleend. Dit is goed voor de privacy van patiënten, maar maakt het lastig om te *leren* en om *patronen* te ontdekken in de data, om later de zorg te verbeteren. Vooral als het doel is te leren voor kleine groepen patiënten, om uiteindelijk *gepersonaliseerde aanbevelingen* te kunnen doen, is data van grote aantallen patiënten nodig voor algoritmes om patronen te ontdekken en bevestigen. Om leren van data bij verschillende zorginstellingen mogelijk te maken terwijl ook de privacy van patiënten gewaarborgd blijft is het *Enabling Personalized Interventions* consortium (Nederlandse vertaling van de naam: faciliteren van gepersonaliseerde behandelingen) opgericht, waar het werk verricht voor dit proefschrift onderdeel van uitmaakt. In dit consortium wordt gewerkt aan een kader en softwarepakket waarbinnen ten eerste patiëntdata bij zorginstellingen op een veilige manier bereikt kan worden, ten tweede regelgevende beperkingen opgelegd kunnen worden op basis van toestemming die zorginstellingen en indi-

viduele patiënten hebben gegeven en ten derde algoritmes worden ontwikkeld die gefedereerd kunnen leren. Dit houdt in dat als meerdere zorginstellingen samen een algoritme willen trainen de data niet verzameld hoeft te worden op een centraal punt. In plaats daarvan wordt het algoritme naar de veilige omgevingen in de aparte instellingen gestuurd, waarbij de (niet-herleidbare) getrainde algoritmes worden teruggestuurd en gecombineerd op een centraal punt. In dit proefschrift worden methoden ontwikkeld en onderzocht die bij uitstek geschikt zijn voor dit soort toepassingen, waarbij de data verdeeld zijn over meerdere instellingen.

Datastromen Het bovengenoemde SAVI, een centraal concept in dit proefschrift, is een van deze methoden. Resultaten van SAVI-analyse van verschillende zorginstellingen kunnen met elkaar gecombineerd worden door simpele vermenigvuldiging. Daarnaast is SAVI-analyse ook geschikt voor *continue analyse van datastromen*; het biedt de mogelijkheid data opnieuw te analyseren na de inclusie van iedere nieuwe patiënt. Na iedere nieuwe analyse kan een *veilige* beslissing genomen worden over de uitkomst van het onderzoek. Dit houdt in dat de methode een zogenoemde type-I fout garantie geeft op ieder moment: de kans dat we onterecht de nulhypothese van het onderzoek verwerpen (bijvoorbeeld de kans dat we onterecht besluiten dat een nieuw middel superieur is vergeleken met een placebo) is begrensd door een door de onderzoeker ingestelde maximaal acceptabele kans op een fout, op welk moment we de beslissing ook nemen. Dit is iets dat met de traditioneel toetsen waarbij bijna altijd gebruik wordt gemaakt van p-waardes niet kan: hierbij moet van tevoren vastgesteld worden met hoeveel patiënten de analyse verricht zal worden. Traditionele p-waarde toetsen zijn dus niet geschikt voor continue analyse van onderzoeksdata. In hoofdstuk 2 van dit proefschrift worden algemene SAVI toetsen voor het vergelijken van twee of meer datastromen ontwikkeld, en wordt een specifieke implementatie met bijbehorende software voor binaire datastromen ontwikkeld. Deze wordt ook vergeleken met de klassieke p-waarde tegenhanger, Fishers exacte toets.

Effecten schatten in datastromen Naast het uitvoeren van hypothesetoetsen zoals "is het nieuwe middel beter dan een placebo" is het voor het nemen van beslissingen essentieel een goede effectschatting te hebben, om bijvoorbeeld te kunnen zeggen hoeveel winst een patiënt zou kunnen verwachten bij gebruik van het nieuwe middel. Traditionele manieren om effecten te schatten, zoals klassieke betrouwbaarheidsintervallen, hebben last van dezelfde beperkingen als p-waardes: deze zijn alleen valide als ze gebruikt worden met een van tevoren vastgestelde hoeveelheid data. In hoofdstuk 3 worden de in hoofdstuk 2 ontwikkelde SAVI toetsen uitgebreid en gebruikt als bouwsteen voor het maken van *altijd valide betrouwbaarheidsintervallen*. Deze altijd valide betrouwbaarheidsintervallen zouden bijvoorbeeld kunnen worden gebruikt in onderzoeksdashboards, waarbij datastromen in werkelijke tijd geanalyseerd worden en de effectschatting direct door onderzoekers en clinici gevolgd kan worden. Hierbij kunnen de onderzoekers op ieder moment beslissen dat de effectschatting precies genoeg is om de studie te kunnen stoppen en beleid aan te passen.

Leren van vrije tekst Voordat flexibele onderzoeken met als doel inferentie zoals hierboven beschreven kunnen worden opgezet, is het onmisbaar dat onderzoeksdata in het juiste format ontsloten wordt en geëxploreerd wordt voor hypothesevorming. Binnen de psychiatrie wordt veel informatie opgeslagen in de vorm van vrije tekst, wat zowel exploratief onderzoek als statistische inferentie bemoeilijkt. Tegelijkertijd is het onmogelijk van clinici te verwachten dat zij alle gegevens in tabulaire vorm in het elektronisch patiëntdossier registreren. Om deze redenen spelen *tekstmining* en informatie extractie een steeds belangrijkere rol in de exploratieve fase van medisch onderzoek. In hoofdstuk 4 wordt onderzocht of met deze technieken informatie over behandeluitkomsten uit routinematig geschreven klinische teksten geëxtraheerd kan worden, om later patronen te kunnen herkennen in de combinatie van deze uitkomsten, patiëntkarakteristieken en behandelkeuzes.

Netwerkanalyse Een methode voor het ontdekken van patronen bij uitstek passend bij de complexiteit van psychiatrische pathologie is Bayesiaanse netwerkanalyse. Met deze analysemethode kunnen causale of voorspellende verbanden tussen patiëntkarakteristieken, behandelkeuzes en een of meerdere uitkomsten ontdekt worden. De uitkomst van de analyse kan inzichtelijk gevisualiseerd worden in netwerkvorm. In hoofdstuk 5 van dit proefschrift wordt Bayesiaanse netwerkanalyse gecombineerd met de tekstmining methode ontwikkeld in hoofdstuk 4 om patronen te ontdekken en hypotheses te ontwikkelen over de uitkomsten na behandeling met verschillende typen antidepressiva op basis van routinematig verzamelde klinische data in twee instituten voor mentale gezondheidszorg. Bayesiaanse netwerkern bieden ook de mogelijkheid voor het integreren van *voorkennis*: in hoofdstuk 6 wordt voorkennis uit medische literatuur gecombineerd met expertkennis van psychiaters om een Bayesiaans netwerk voor het modelleren van uitkomsten na behandeling met electroconvulsietherapie te verbeteren.

Leren voor specifieke groepen patiënten Het onderzoek in hoofdstuk 5 en 6 bracht meerdere kansrijke hypotheses over behandeleffecten bij specifieke, kleine groepen patienten voort, die verdere validatie middels prospectief onderzoek behoeven om een robuuste effectschatting te kunnen maken. In hoofdstuk 7 worden de methodes ontwikkeld in hoofdstuk 2 en 3 uitgebreid voor onderzoeksvragen van deze vorm, waarbij behandeleffecten in verschillende datastromen prospectief woren geanalyseerd en ook nog rekening gehouden wordt met het type patiënt: de data wordt *gestratificeerd* naar het type patiënt. Om de mogelijkheid om behandeleffecten te detecteren met deze onderzoeksmethode te verbeteren, wordt het effect van informatieuitwisseling tussen de patiëntgroepen op een zelflerende manier onderzocht. Uiteindelijk wordt geïllustreerd hoe een onderzoeksopzet om een van de resulterende hypothesen uit hoofdstuk 5 er uit zou kunnen zien.

Makkelijker valide medisch onderzoek In dit proefschrift zijn nieuwe methoden ontwikkeld om medisch onderzoek op een meer flexibele en makkelijkere manier op te zetten: met deze nieuwe methoden is het mogelijk om onderzoeksresultaten in werkelijke tijd, al tijdens te studie te analyseren en eventueel eerder te beslissen *of* langer door te gaan als meer informatie verzameld moet worden. Ook zijn deze

methoden direct uit te breiden naar continue analyse over meerdere onderzoeksinstellingen. Continuering van het werk van het Enabling Personalized Interventions consortium en soortgelijke initiatieven in de komende jaren zal hopelijk de complexe data-infrastuctuur die nodig is voor dit soort flexibele studies met meerdere deelnemende ziekenhuizen mogelijk maken, zodat daadwerkelijk toegewerkt kan gaan worden naar gepersonaliseerde behandelingen door valide effectschattingen in specifieke groepen patiënten.

Acknowledgments

Without my supervisors Peter Grünwald, Floortje Scheepers and Aki Härmä it would have been impossible to complete the diverse body of work presented in this thesis. Peter introduced me to many new topics in mathematics and theoretical statistics. I really enjoyed our collaboration, combining further developing the safe statistics theory with actual software implementations and simulations concurrently. Floortje was incredibly fast at adapting to all complicated analysis techniques we proposed, and picking out exactly the right things to move psychiatry research along. Aki introduced me to the work of many interesting researchers at Philips, widening my scope and peeking my interest for the potential of the role of NLP in healthcare.

I want to give special mention to Karin Hagoort, teamlead innovation at the Psychiatry department of UMC Utrecht and head of the PsyData team. Apart from helping me with so many organizational challenges during the project and introducing me to many other interesting people in the UMC Utrecht and other mental healthcare institutes, we had many insightful discussions on the implementation of AI in healthcare and Karin's contributions to the three psychiatry-focused chapters in this thesis have been very valuable.

These three chapters would also have been impossible to write without the collaboration with and support from my PsyData colleagues at UMC Utrecht, especially Femke, Saskia, Kees, Vincent, Zimbo, and Willem. I also particularly want to mention the data scientists I collaborated with at Parnassia Groep: Roel, Rosa and Eline. Even though some of you have moved on to different jobs and positions by now, I would be happy to work together or exchange ideas on other data science projects in the future with each of you.

I also want to mention the other Psychiatrists of UMC Utrecht I collaborated with for the work in this thesis: Fleur Velders, Edwin van Dellen, Metten Somers and Yuri van der Does, and of course all of your colleagues who attended my presentations and with whom I worked on smaller data science questions over the years. It has been great to collaborate with clinicians with such an affinity for statistics and computer science, who could really think along critically with the work in this thesis.

My colleagues at CWI - over the years it have been too many to list them all here - greatly contributed toward my development as a statistician. I loved being able to get an insight in the work of theoretical mathematics and machine learning researchers: it was a real eye opener to follow all your work, this stimulated me

to think beyond the (small collection of) machine learning methods most-used in clinical research. I especially want to highlight the collaboration I had with Alexander Ly, collaborating on the safestats software package, who taught me lots of nifty coding tricks.

Lastly I want to mention my colleagues from the EPI consortium, Tim, Corinne, Saba, Jamila, Milen and in particular the PIs and project lead Paola Grosso, Cees de Laat and Sander Klous. I loved that my work was part of a bigger project and to learn about each of your respective fields of research.

Curriculum Vitae

From 2010 until 2013 Rosanne studied Medicine at the Leiden University Medical Center (LUMC). She was admitted to the Honours College of Leiden University, where she was awarded a grant to start a research project at the pathology department of LUMC under supervision of Dr. Hans Baelde, Prof. Kitty Bloemenkamp and Prof. J.A. Bruijn. In 2014 she got the opportunity to continue this research full time directly after obtaining her Bachelor degree, which resulted in the PhD thesis *Endothelial Pathology in Preeclampsia* at the Faculty of Medicine of Leiden University.

During her time as a researcher at LUMC, Rosanne discovered that scientific research and particularly methodology and mathematics interested her the most. Therefore she decided to continue her Master studies in this direction: from 2017 until 2019 she studied Statistical Science for the Life and Behavioral Sciences at the faculty of Science at Leiden University, for which she graduated *cum laude*. During her studies she also worked part-time as a software engineer at El Nino development. She wrote her Master thesis *Safe tests for 2 x 2 contingency tables and the Cochran-Mantel-Haenszel test* under supervision of Prof. Peter Grünwald at CWI, which was awarded the Jan Hemelrijk Award by the Dutch society for statistics and operations research (VVSOR).

After graduating she continued the research on safe statistics started during her master thesis as part of a second PhD trajectory, this time in Mathematics. She worked in the *Enabling Personalized Interventions* consortium under supervision of Prof. Peter Grünwald, Prof. Floortje Scheepers (UMC Utrecht) and Dr. Aki Härmä (Philips research) on implementations of safe statistics and other methods suitable for real-time, federated learning. Rosanne spent half her time focusing on developing statistical methodology in the machine learning group at CWI, and the other half implementing new methods and working as a data scientist at the data science team PsyData at the Psychiatry department of UMC Utrecht. Since finishing her second PhD project, Rosanne has continued working at the Psychiatry department of UMC Utrecht as a clinical data scientist.

Bibliography

R. J. Adams. Safe hypothesis tests for the 2 × 2 contingency table. Master's thesis, Delft University of Technology, 2020.

C. G. Allaart, B. Keyser, H. Bal, and A. Van Halteren. Vertical split learning-an exploration of predictive performance in medical and other use cases. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2022.

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (4th ed.)*. Washington, DC: American Psychiatric Association, 1994.

American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (5th ed.)*, volume 21. American Psychiatric Publishing, 2013.

S. Amiri, A. Belloum, S. Klous, and L. Gommans. Compressive differentiallyprivate federated learning through universal vector quantization. In *AAAI Workshop on Privacy-Preserving Artificial Intelligence*, 2021.

S. Amiri, A. Belloum, E. Nalisnick, S. Klous, and L. Gommans. On the impact of non-iid data on the performance and fairness of differentially private federated learning. In *2022 52nd Annual IEEE/IFIP International Conference on Dependable Systems and Networks Workshops (DSN-W)*, pages 52–58. IEEE Computer Society, 2022.

V. Amrhein, S. Greenland, and B. McShane. Scientists rise up against statistical significance, 2019.

C. Andrade, S. S. Arumugham, and J. Thirthalli. Adverse effects of electroconvulsive therapy. *Psychiatric Clinics*, 39(3):513–530, 2016.

P. Arora, D. Boyne, J. J. Slater, A. Gupta, D. R. Brenner, and M. J. Druzdzel. Bayesian networks for risk prediction using real-world data: a tool for precision medicine. *Value in Health*, 22(4):439–445, 2019.

A. Bayes and G. Parker. How to choose an antidepressant medication. *Acta Psychiatrica Scandinavica*, 139(3):280–291, 2019.

C. Beard, A. J. Millner, M. J. Forgeard, E. I. Fried, K. J. Hsu, M. Treadway, C. V. Leonard, S. Kertz, and T. Björgvinsson. Network analysis of depression and

anxiety symptom relationships in a psychiatric sample. *Psychological medicine*, 46(16):3359–3369, 2016.

D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, C. Camerer, et al. Redefine statistical significance. *Nature human behaviour*, 2(1):6–10, 2018.

J. O. Berger, L. R. Pericchi, and J. A. Varshavsky. Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 307–321, 1998.

J. A. Berlin and R. M. Golub. Meta-analysis as evidence: building a better pyramid. *Jama*, 312(6):603–606, 2014.

D. Berner and V. Amrhein. Why and how we should join the shift from significance testing to estimation. *Journal of Evolutionary Biology*, 35(6):777–787, 2022.

D. de Beurs, C. Bockting, A. Kerkhof, F. Scheepers, R. O'Connor, B. Penninx, and I. van de Leemput. A network perspective on suicidal behavior: Understanding suicidality as a complex system. *Suicide Life Threat. Behav.*, 51(1):115–126, 2021. doi: 10.1111/sltb.12676.

O. Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32:D267–D270, 2004.

D. Borsboom. A network theory of mental disorders. *World psychiatry*, 16(1): 5–13, 2017.

D. Borsboom and A. O. Cramer. Network analysis: an integrative approach to the structure of psychopathology. *Annual review of clinical psychology*, 9:91–121, 2013.

V. Braun and V. Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

J. E. Brazier, R. Harper, N. M. Jones, A. O'Cathain, K. J. Thomas, T. Usherwood, and L. Westlake. Validating the sf-36 health survey questionnaire: new outcome measure for primary care. *BMJ*, 305(6846):160–4, 1992. doi: 10.1136/bmj.305.6846.160.

G. Briganti, M. Scutari, and R. J. McNally. A tutorial on bayesian networks for psychopathology researchers. *Psychological methods*, 2022.

T. J. Bright, A. Wong, R. Dhurjati, E. Bristow, L. Bastian, R. R. Coeytaux, G. Samsa, V. Hasselblad, J. W. Williams, M. D. Musty, et al. Effect of clinical decision-support systems: a systematic review. *Annals of internal medicine*, 157 (1):29–43, 2012.

J. Brunson and Q. Read. ggalluvial: Alluvial plots in 'ggplot2'. r package., 2020. URL `http://corybrunson.github.io/ggalluvial/`.

K. Bruynseels, F. Santoni de Sio, and J. Van den Hoven. Digital twins in health care: ethical implications of an emerging engineering paradigm. *Frontiers in genetics*, 9, 2018. doi: 10.3389/fgene.2018.00031.

P. B. Burns, R. J. Rohrich, and K. C. Chung. The levels of evidence and their role in evidence-based medicine. *Plastic and reconstructive surgery*, 128(1):305, 2011.

J. Busner and S. D. Targum. The clinical global impressions scale: applying a research tool in clinical practice. *Psychiatry (Edgmont)*, 4(7):28–37, 2007.

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011. ISSN 1548-7660.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, Cambridge, UK, 2006.

A. Cipriani, T. A. Furukawa, G. Salanti, A. Chaimani, L. Z. Atkinson, Y. Ogawa, S. Leucht, H. G. Ruhe, E. H. Turner, J. P. T. Higgins, M. Egger, N. Takeshima, Y. Hayasaka, H. Imai, K. Shinohara, A. Tajika, J. P. A. Ioannidis, and J. R. Geddes. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet*, 391(10128):1357–1366, 2018. doi: 10.1016/s0140-6736(17)32802-7.

D. A. Ciraulo, J. Barnhill, and H. Boxenbaum. Pharmacokinetic interaction of disulfiram and antidepressants. *Am. J. Psychiatry*, 142(11):1373–4, 1985. doi: 10.1176/ajp.142.11.1373.

G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Annals of internal medicine*, 162(1):55–63, 2015.

P. Coorevits, M. Sundgren, G. O. Klein, A. Bahr, B. Claerhout, C. Daniel, M. Dugas, D. Dupont, A. Schmidt, P. Singleton, et al. Electronic health records: new opportunities for clinical research. *Journal of internal medicine*, 274(6): 547–560, 2013.

R. Cordier, T. Brown, L. Clemson, and J. Byles. Evaluating the longitudinal item and category stability of the sf-36 full and summary scales using rasch analysis. *Biomed Res Int*, 2018:1013453, 2018. doi: 10.1155/2018/1013453.

D. Darling and H. Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1):66–68, 1967.

A. P. Dawid. Present position and potential developments: Some personal views statistical theory the prequential approach. *Journal of the Royal Statistical Society: Series A (General)*, 147(2):278–290, 1984.

C. E. Dean. Social inequality, scientific inequality, and the future of mental illness. *Philosophy, Ethics, and Humanities in Medicine*, 12(1):1–12, 2017.

T. M. Deist, F. Dankers, P. Ojha, M. Scott Marshall, T. Janssen, C. Faivre-Finn, C. Masciocchi, V. Valentini, J. Wang, J. Chen, Z. Zhang, E. Spezi, M. Button, J. Jan Nuyttens, R. Vernhout, J. van Soest, A. Jochems, R. Monshouwer, J. Bussink, G. Price, P. Lambin, and A. Dekker. Distributed learning on 20 000+ lung cancer patients - the personal health train. *Radiother Oncol*, 144: 189–200, 2020. doi: 10.1016/j.radonc.2019.11.019.

D. L. Demets and K. G. Lan. Interim analysis: the alpha spending function approach. *Statistics in medicine*, 13(13-14):1341–1352, 1994.

L. van Diermen, S. van den Ameele, A. M. Kamperman, B. C. Sabbe, T. Vermeulen, D. Schrijvers, and T. K. Birkenhäger. Prediction of electroconvulsive therapy response and remission in major depression: meta-analysis. *The British journal of psychiatry*, 212(2):71–80, 2018.

B. Duan, A. Ramdas, and L. Wasserman. Interactive rank testing by betting. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *PMLR*, pages 201–235, 2022.

Dutch National Healthcare Institute. Farmacotherapeutisch kompas, 2020. URL `https://www.farmacotherapeutischkompas.nl/`.

J. Eckhoff. *Helly, Radon, and Carathéodory type theorems, Handbook of Convex Geometry Part A*, pages 389–448. Elsevier, 1993.

S. Epskamp, A. O. Cramer, L. J. Waldorp, V. D. Schmittmann, and D. Borsboom. qgraph: Network visualizations of relationships in psychometric data. *Journal of statistical software*, 48:1–18, 2012.

N. J. Ermers, K. Hagoort, and F. E. Scheepers. The predictive validity of machine learning models in the classification and treatment of major depressive disorder: State of the art and future directions. *Frontiers in Psychiatry*, 11:472, 2020.

T. van Erven, P. Grünwald, and S. de Rooij. Catching up faster in bayesian model selection and model averaging. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

ESC Cardiovasc Risk Collaboration, SCORE2 working group, et al. Score2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in europe. *European Heart Journal*, 42(25):2439–2454, 2021.

B. van Es, L. C. Reteig, S. C. Tan, M. Schraagen, M. M. Hemker, S. R. Arends, M. A. Rios, and S. Haitjema. Negation detection in dutch clinical texts: an evaluation of rule-based and machine learning methods. *BMC bioinformatics*, 24(1):10, 2023.

B. de Finetti. *Theory of probability: A critical introductory treatment*, volume 6. John Wiley & Sons, 2017.

R. A. Fisher. *Statistical methods for research workers*. Oliver and Boyd, 1925.

W. J. Frawley, G. Piatetsky-Shapiro, and C. J. Matheus. Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57–57, 1992.

E. I. Fried, J. K. Flake, and D. J. Robinaugh. Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*, 1(6):358–368, 2022.

P. Fusar-Poli, Z. Hijazi, D. Stahl, and E. W. Steyerberg. The science of prognosis in psychiatry: A review. *JAMA Psychiatry*, 75(12):1289–1297, 2018. doi: 10.1001/jamapsychiatry.2018.2530.

B. N. Gaynes, D. Warden, M. H. Trivedi, S. R. Wisniewski, M. Fava, and A. J. Rush. What did star*d teach us? results from a large-scale, practical, clinical trial for patients with depression. *Psychiatr Serv*, 60(11):1439–45, 2009. doi: 10.1176/ps.2009.60.11.1439.

S. A. Glied, B. D. Stein, T. G. McGuire, R. R. Beale, F. F. Duffy, S. Shugarman, and H. H. Goldman. Measuring performance in psychiatry: A call to action. *Psychiatr Serv*, 66(8):872–8, 2015. doi: 10.1176/appi.ps.201400393.

T. J. Gross, R. B. Araújo, F. A. C. Vale, M. Bessani, and C. D. Maciel. Dependence between cognitive impairment and metabolic syndrome applied to a brazilian elderly dataset. *Artificial intelligence in medicine*, 90:53–60, 2018.

P. Grünwald. *The Minimum Description Length Principle*. MIT Press, Cambridge, MA, 2007.

P. Grünwald. Beyond Neyman-Pearson. *arXiv preprint arXiv:2205.00901*, 2022.

P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *accepted, pending minor revision, for publication in Journal of the Royal Statistical Society: Series B*, 2022a.

P. Grünwald, A. Henzi, and T. Lardy. Anytime valid tests of conditional independence under model-x. *arXiv preprint arXiv:2209.12637*, 2022b.

E. Gunel and J. Dickey. Bayes factors for independence in contingency tables. *Biometrika*, 61(3):545–557, 1974.

S. H. Hageman, A. J. McKay, P. Ueda, L. H. Gunn, T. Jernberg, E. Hagström, D. L. Bhatt, P. G. Steg, K. Läll, R. Mägi, et al. Estimation of recurrent atherosclerotic cardiovascular event risk in patients with established cardiovascular disease: the updated smart2 algorithm. *European Heart Journal*, 43(18):1715–1727, 2022.

M. Hamilton. A rating scale for depression. *J Neurol Neurosurg Psychiatry*, 23:56–62, 1960.

M. Hamilton. Development of a rating scale for primary depressive illness. *British journal of social and clinical psychology*, 6(4):278–296, 1967.

Y. Hao, P. Grünwald, T. Lardy, L. Long, and R. Adams. E-values for k-sample tests with exponential families. *arXiv preprint arXiv:2303.00471*, 2023.

A. U. Haq, A. F. Sitzmann, M. L. Goldman, D. F. Maixner, and B. J. Mickey. Response of depression to electroconvulsive therapy: a meta-analysis of clinical predictors. *The Journal of clinical psychiatry*, 76(10):18164, 2015.

B. J. Havaki-Kontaxaki, P. P. Ferentinos, V. P. Kontaxakis, K. G. Paplos, and C. R. Soldatos. Concurrent administration of clozapine and electroconvulsive therapy in clozapine-resistant schizophrenia. *Clinical Neuropharmacology*, 29 (1):52–56, 2006.

R. de Heide and P. D. Grünwald. Why optional stopping can be a problem for bayesians. *Psychonomic Bulletin & Review*, 28:795–812, 2021.

W. T. Heijnen, T. K. Birkenhäger, A. I. Wierdsma, and W. W. van den Broek. Antidepressant pharmacotherapy failure and response to subsequent electroconvulsive therapy: a meta-analysis. *Journal of clinical psychopharmacology*, 30(5): 616–619, 2010.

K. Hemming, M. Taljaard, J. E. McKenzie, R. Hooper, A. Copas, J. A. Thompson, M. Dixon-Woods, A. Aldcroft, A. Doussau, M. Grayling, et al. Reporting of stepped wedge cluster randomised trials: extension of the consort 2010 statement with explanation and elaboration. *BMJ*, 363, 2018.

A. Henzi and J. F. Ziegel. Valid sequential inference on probability forecast performance. *Biometrika*, 109(3):647–663, 2022.

M. Herbster and M. K. Warmuth. Tracking the best expert. *Machine learning*, 32 (2):151–178, 1998.

M. A. Hernán, S. Hernández-Diaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology*, pages 615–625, 2004.

M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd. spacy: Industrial-strength natural language processing in python, 2020. URL `https://doi.org/10.5281/zenodo.1212303`.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 2021.

T. Jamil, A. Ly, R. D. Morey, J. Love, M. Marsman, and E.-J. Wagenmakers. Default "Gunel and Dickey" Bayes factors for contingency tables. *Behavior Research Methods*, 49:638–652, 2017.

E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106 (4):620, 1957.

H. Jeffreys. *The theory of probability*. Oxford University Press, 1998.

Y. Jin, Y. Su, X.-H. Zhou, S. Huang, and A. D. N. Initiative. Heterogeneous multimodal biomarkers analysis for alzheimer's disease via bayesian network. *EURASIP Journal on Bioinformatics and Systems Biology*, 2016:1–8, 2016.

R. Johari, P. Koomen, L. Pekelis, and D. Walsh. Always valid inference: Continuous monitoring of a/b tests. *Operations Research*, 70(3):1806–1821, 2022.

L. K. John, G. Loewenstein, and D. Prelec. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 23(5):524–532, 2012.

S. H. Jones, G. Thornicroft, M. Coffey, and G. Dunn. A brief mental health outcome scale-reliability and validity of the global assessment of functioning (gaf). *Br J Psychiatry*, 166(5):654–9, 1995. doi: 10.1192/bjp.166.5.654.

R. E. Kass and S. K. Vaidyanathan. Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(1):129–144, 1992.

J. A. Kassem, O. Valkering, A. Belloum, and P. Grosso. Epi framework: Approach for traffic redirection through containerised network functions. In *2021 IEEE 17th International Conference on eScience (eScience)*, pages 80–89. IEEE, 2021.

E. Kaufmann, O. Cappé, and A. Garivier. On the complexity of a/b testing. In *Conference on Learning Theory*, pages 461–481. PMLR, 2014.

M. G. Kebede. Automating normative control for healthcare research. In *International Workshop on AI Approaches to the Complexity of Legal Systems, International Workshop on AI Approaches to the Complexity of Legal Systems, International Workshop on Explainable and Responsible AI and Law*, pages 62–72. Springer, 2021.

J. L. Kelly. A new interpretation of information rate. *The bell system technical journal*, 1956.

R. C. Kessler, W. T. Chiu, O. Demler, and E. E. Walters. Prevalence, severity, and comorbidity of 12-month dsm-iv disorders in the national comorbidity survey replication. *Archives of general psychiatry*, 62(6):617–627, 2005.

K. H. Kho, M. F. van Vreeswijk, S. Simpson, and A. H. Zwinderman. A meta-analysis of electroconvulsive therapy efficacy in depression. *The journal of ECT*, 19(3):139–147, 2003.

O. J. Kirtley, K. van Mens, M. Hoogendoorn, N. Kapur, and D. de Beurs. Translating promise into practice: a review of machine learning in suicide research and prevention. *Lancet Psychiatry*, 9(3):243–252, 2022. doi: 10.1016/s2215-0366(21)00254-6.

B. Klingenberg. A new and improved confidence interval for the mantel–haenszel risk difference. *Statistics in Medicine*, 33(17):2968–2983, 2014.

J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

W. M. Koolen and T. van Erven. Freezing and sleeping: Tracking experts that learn by evolving past posteriors. *arXiv preprint arXiv:1008.4654*, 2010.

W. M. Koolen and P. Grünwald. Log-optimal anytime-valid e-values. *International Journal of Approximate Reasoning*, 141:69–82, 2022.

W. M. Koolen and S. de Rooij. Universal codes from switching strategies. *IEEE Transactions on Information Theory*, 59(11):7168–7185, 2013.

R. Koposov, S. Fossum, T. Frodl, Ø. Nytrø, B. Leventhal, A. Sourander, S. Quaglini, M. Molteni, M. de la Iglesia Vayá, H.-U. Prokosch, et al. Clinical decision support systems in child and adolescent psychiatry: a systematic review. *European Child & Adolescent Psychiatry*, 26:1309–1317, 2017.

Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M. P. Richardson, R. Stewart, A. D. Shah, W. K. Wong, Z. Ibrahim, J. T. Teo, and R. J. B. Dobson. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artif Intell Med*, 117:102083, 2021.

R. Kroeze, D. C. van der Veen, M. N. Servaas, J. A. Bastiaansen, R. C. O. Voshaar, D. Borsboom, H. G. Ruhe, R. A. Schoevers, and H. Riese. Personalized feedback on symptom dynamics of psychopathology: A proof-of-principle study. *Journal for Person-Oriented Research*, 3(1):1, 2017.

H. M. Krumholz. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs*, 33(7):1163–1170, 2014.

S. Kundu. AI in medicine must be explainable. *Nature Medicine*, 27(8):1328–1328, 2021.

E. Kyrimi, K. Dube, N. Fenton, A. Fahmi, M. R. Neves, W. Marsh, and S. McLachlan. Bayesian networks in healthcare: What is preventing their adoption? *Artificial Intelligence in Medicine*, 116:102079, 2021.

T. L. Lai. On confidence sequences. *The Annals of Statistics*, 4(2):265–280, 1976.

T. A. Lang and D. G. Altman. Statistical analyses and methods in the published literature: The SAMPL guidelines. *Guidelines for reporting health research: A user's manual*, pages 264–274, 2014.

J. Lee, R. Henning, and M. Cherniack. Correction workers' burnout and outcomes: A bayesian network approach. *International journal of environmental research and public health*, 16(2):282, 2019.

W. Lee, J. Bindman, T. Ford, N. Glozier, P. Moran, R. Stewart, and M. Hotopf. Bias in psychiatric case–control studies: literature survey. *The British Journal of Psychiatry*, 190(3):204–209, 2007.

K. A. Leiknes, L. J.-v. Schweder, and B. Høie. Contemporary use and practice of electroconvulsive therapy worldwide. *Brain and behavior*, 2(3):283–344, 2012.

H. Leroux, A. Metke-Jimenez, and M. J. Lawley. Towards achieving semantic interoperability of clinical study data with FHIR. *Journal of biomedical semantics*, 8(1):1–14, 2017.

L. A. Levin. Uniform tests of randomness. *Soviet Mathematics Doklady*, 17(2): 337–340, 1976.

A. Levy, S. Taib, C. Arbus, P. Péran, A. Sauvaget, L. Schmitt, and A. Yrondi. Neuroimaging biomarkers at baseline predict electroconvulsive therapy overall clinical response in depression: a systematic review. *The journal of ECT*, 35(2): 77–83, 2019.

A. Lhéritier and F. Cazals. A sequential non-parametric multivariate two-sample test. *IEEE Transactions on Information Theory*, 64(5):3361–3370, 2018.

J. Li. *Estimation of Mixture Models*. PhD thesis, Yale University, New Haven, CT, 1999.

J. Li and A. Barron. Mixture density estimation. In S. Solla, T. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12, pages 279–285, Cambridge, MA, 2000. MIT Press.

J. J. Liang, C.-H. Tsou, and M. V. Devarakonda. Ground truth creation for complex clinical nlp tasks–an iterative vetting approach and lessons learned. *AMIA Summits on Translational Science Proceedings*, 2017:203, 2017.

M. Lindon and A. Malek. Anytime-valid inference for multinomial count data. *Advances in Neural Information Processing Systems*, 35:2817–2831, 2022.

S. H. Lisanby. Electroconvulsive therapy for depression. *New England Journal of Medicine*, 357(19):1939–1945, 2007.

P. de Looff, M. L. Noordzij, M. Moerbeek, H. Nijman, R. Didden, and P. Embregts. Changes in heart rate and skin conductance in the 30 min preceding aggressive behavior. *Psychophysiology*, 56(10):e13420, 2019.

J. J. Luykx, D. Loef, B. Lin, L. van Diermen, J. O. Nuninga, E. van Exel, M. L. Oudega, D. Rhebergen, S. N. Schouws, P. van Eijndhoven, et al. Interrogating associations between polygenic liabilities and electroconvulsive therapy effectiveness. *Biological Psychiatry*, 91(6):531–539, 2022.

A. Ly, R. J. Turner, and J. ter Schure. R-package `safestats`, 2022. CRAN.

O.-A. Maillard. Mathematics of statistical sequential decision making, 2019. Thèse de Habilitation.

T. Manole and A. Ramdas. Martingale methods for sequential estimation of convex functionals and divergences. *IEEE Transactions on Information Theory*, 2023.

N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, 22(4): 719–748, 1959.

M. L. McHugh. The chi-square test of independence. *Biochemia medica*, 23(2): 143–149, 2013.

S. McLachlan, K. Dube, G. A. Hitman, N. E. Fenton, and E. Kyrimi. Bayesian networks in healthcare: Distribution by medical condition. *Artificial intelligence in medicine*, 107:101912, 2020.

R. McNally, P. Mair, B. Mugno, and B. Riemann. Co-morbid obsessive–compulsive disorder and depression: A bayesian network approach. *Psychological medicine*, 47(7):1204–1214, 2017.

J. Meiseberg and S. Moritz. Biases in diagnostic terminology: Clinicians choose different symptom labels depending on whether the same case is framed as depression or schizophrenia. *Schizophr Res*, 222:444–449, 2020. doi: 10.1016/j.schres.2020.03.050.

V. Menger. Psynlp, 2020. URL `https://github.com/vmenger/psynlp`.

V. Menger, F. Scheepers, and M. Spruit. Comparing deep learning and classical machine learning approaches for predicting inpatient violence incidents from clinical text. *Applied Sciences*, 8(6):981, 2018a.

V. Menger, F. Scheepers, L. M. van Wijk, and M. Spruit. DEDUCE: A pattern matching method for automatic de-identification of Dutch medical text. *Telematics and Informatics*, 35(4):727–736, 2018b.

V. J. Menger. *Knowledge Discovery in Clinical Psychiatry*. PhD thesis, Utrecht University, 2019.

S. A. Montgomery and M. Åsberg. A new depression scale designed to be sensitive to change. *The British journal of psychiatry*, 134(4):382–389, 1979.

L. B. Moreira and A. A. Namen. A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. *Computer methods and programs in biomedicine*, 165:139–149, 2018.

J. Muglu, H. Rather, D. Arroyo-Manzano, S. Bhattacharya, I. Balchin, A. Khalil, B. Thilaganathan, K. S. Khan, J. Zamora, and S. Thangaratinam. Risks of stillbirth and neonatal death with advancing gestation at term: A systematic review and meta-analysis of cohort studies of 15 million pregnancies. *PLoS medicine*, 16(7):e1002838, 2019.

A. C. Naglich, A. Lin, S. Wakhlu, and B. H. Adinoff. Systematic review of combined pharmacotherapy for the treatment of alcohol use disorder in patients without comorbid conditions. *CNS Drugs*, 32(1):13–31, 2018. doi: 10.1007/s40263-017-0484-2.

Nederlandse Rijksoverheid. Wet hergebruik van overheidsinformatie, 2021. URL `https://wetten.overheid.nl/BWBR0036795/`.

Netherlands Federation of University Medical Centers. Guideline quality assurance of research involving human subjects, 2020.

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

F. Orabona and K.-S. Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *arXiv preprint arXiv:2110.14099*, 2021.

M. Otto. Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation–GDPR). In *International and European Labour Law*, pages 958–981, 2018.

S. Oxlad and M. Baldwin. Multiple case sampling of ect administration to 217 minors: Review and meta-analysis. *Journal of Mental Health*, 5(5):451–464, 1996.

L. Pace and A. Salvan. Likelihood, replicability and robbins' confidence sequences. *International Statistical Review*, 88(3):599–615, 2020.

M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021.

D. Pagnin, V. de Queiroz, S. Pini, and G. B. Cassano. Efficacy of ECT in depression: a meta-analytic review. *The journal of ECT*, 20(1):13–20, 2004.

T. Pandeva, T. Bakker, C. A. Naesseth, and P. Forré. E-valuating classifier two-sample tests. *arXiv preprint arxiv:2210.13027*, 2022.

G. Parmigiani, B. Barchielli, S. Casale, T. Mancini, and S. Ferracuti. The impact of machine learning in predicting risk of violence: A systematic review. *Frontiers in psychiatry*, 13, 2022.

S. van der Pas and P. Grünwald. Almost the best of three worlds: Risk, consistency and optional stopping for the switch criterion in nested model selection. *Statistica Sinica*, 28(1):229–255, 2018.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

D. Peterson. The replication crisis won't be solved with broad brushstrokes. *Nature*, 594(7862), 2021.

M. F. Pradier, T. H. McCoy Jr, M. Hughes, R. H. Perlis, and F. Doshi-Velez. Predicting treatment dropout after antidepressant initiation. *Translational psychiatry*, 10(1):1–8, 2020.

C. A. Prinsen, S. Vohra, M. R. Rose, S. King-Jones, S. Ishaque, Z. Bhaloo, D. Adams, and C. B. Terwee. Core Outcome Measures in Effectiveness Trials (COMET) initiative: protocol for an international Delphi study to achieve consensus on how to select outcome measurement instruments for outcomes included in a 'core outcome set'. *Trials*, 15:247, 2014. doi: 10.1186/1745-6215-15-247.

J. Prudic, M. Olfson, S. C. Marcus, R. B. Fuller, and H. A. Sackeim. Effectiveness of electroconvulsive therapy in community settings. *Biological psychiatry*, 55(3): 301–312, 2004.

S. Qiu, W. Poon, M. Tang, and J. Tao. Construction of confidence intervals for the risk differences in stratified design with correlated bilateral data. *Journal of Biopharmaceutical Statistics*, 29(3):446–467, 2019.

R. Rabin and F. de Charro. EQ-5D: a measure of health status from the EuroQol Group. *Ann Med*, 33(5):337–43, 2001. doi: 10.3109/07853890109002087.

A. Ramdas, J. Ruf, M. Larsson, and W. Koolen. Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv preprint arXiv:2009.03167*, 2020.

A. Ramdas, P. Grünwald, V. Vovk, and G. Shafer. Game-theoretic statistics and safe anytime-valid inference. *arXiv preprint arXiv:2210.01948*, 2022.

F. P. Ramsey. Truth and probability. In *The Foundations of Mathematics and other Logical Essays*. Routledge & Kegan Paul Ltd, 1931.

H. Robbins. Statistical methods related to the law of the iterated logarithm. *The Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

R. Royall. *Statistical Evidence: A Likelihood Paradigm*. Chapman and Hall, 1997.

A. J. Rush, M. H. Trivedi, S. R. Wisniewski, A. A. Nierenberg, J. W. Stewart, D. Warden, G. Niederehe, M. E. Thase, P. W. Lavori, and B. D. Lebowitz. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: a STAR* D report. *American Journal of Psychiatry*, 163 (11):1905–1917, 2006.

Y. E. Rybak, K. S. P. Lai, R. Ramasubbu, F. Vila-Rodriguez, D. M. Blumberger, P. Chan, N. Delva, P. Giacobbe, C. Gosselin, S. H. Kennedy, H. Iskandar,

S. McInerney, P. Ravitz, V. Sharma, A. Zaretsky, and A. M. Burhan. Treatment-resistant major depressive disorder: Canadian expert consensus on definition and assessment. *Depress Anxiety*, 38(4):456–467, 2021. doi: 10.1002/da.23135.

L. Samalin, J.-B. Genty, L. Boyer, J. Lopez-Castroman, M. Abbar, and P.-M. Llorca. Shared decision-making: a systematic review focusing on mood disorders. *Current psychiatry reports*, 20(4):1–11, 2018. ISSN 1535-1645.

R. Sanfelici, D. B. Dwyer, L. A. Antonucci, and N. Koutsouleris. Individualized diagnostic and prognostic models for patients with psychosis risk syndromes: A meta-analytic view on the state of the art. *Biol Psychiatry*, 88(4):349–360, 2020. doi: 10.1016/j.biopsych.2020.02.009.

L. J. Savage. *The foundations of statistics*. John Wiley & Sons, Inc., 1954.

C. W. M. Schepper, R. J. Turner, L. Schaijk, and K. Hagoort. Bijwerkingen van ECT detecteren in klinische teksten, conference presentation at NVVP Voorjaarscongres 2022, 2022.

J. Schneider, M. Patterson, and X. F. Jimenez. Beyond depression: Other uses for tricyclic antidepressants. *Cleveland Clinic Journal of Medicine*, 86(12):807–814, 2019.

J. ter Schure. *ALL-IN Meta-Analysis*. PhD thesis, Leiden University, 2022.

J. ter Schure, M. F. Pérez-Ortiz, A. Ly, and P. Grunwald. The safe logrank test: Error control under continuous monitoring with unlimited horizon. *arXiv preprint arXiv:2011.06931*, 2020.

J. ter Schure, A. Ly, L. Belin, C. S. Benn, M. J. Bonten, J. D. Cirillo, J. A. Damen, I. Fronteira, K. D. Hendriks, A. P. Junqueira-Kipnis, et al. Bacillus calmette-guerin vaccine to reduce covid-19 infections and hospitalisations in healthcare workers: a living systematic review and prospective all-in meta-analysis of individual participant data from randomised controlled trials. *medRxiv*, pages 2022–12, 2022.

J. A. ter Schure and P. Grünwald. Accumulation bias in meta-analysis: the need to consider time in error control. *F1000Research*, 8, 2019.

M. Scutari. Learning bayesian networks with the bnlearn r package. *Journal of Statistical Software*, 35:1–22, 2010.

M. Scutari and K. Strimmer. Introduction to graphical modelling. *Handbook of Statistical Systems Biology*, 2011.

T. Seidenfeld. Why I am not an objective Bayesian; some reflections prompted by Rosenkrantz. *Theory and Decision*, 11(4):413–440, 1979.

J. Sevilla. Finding, scoring and explaining arguments in bayesian networks. *arXiv preprint arXiv:2112.00799*, 2021.

S. Shaer, G. Maman, and Y. Romano. Model-free sequential testing for conditional independence via testing by betting. *arXiv preprint arXiv:2210.00354*, 2022.

G. Shafer, A. Shen, N. Vereshchagin, and V. Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, pages 84–101, 2011.

G. Shafer et al. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(2):407–431, 2021.

S. Shekhar and A. Ramdas. Nonparametric two-sample testing by betting. *arXiv preprint arXiv:2112.09162*, 2021.

D. Siegmund. *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media, 2013.

L. Simon, M. Blay, F. Galvao, and J. Brunelin. Using EEG to predict clinical response to electroconvulsive therapy in patients with major depression: a comprehensive review. *Frontiers in Psychiatry*, 12:643710, 2021.

K. A. Spackman, K. E. Campbell, and R. A. Côté. SNOMED RT: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association, 1997.

D. Stacey, F. Légaré, K. Lewis, M. J. Barry, C. L. Bennett, K. B. Eden, M. Holmes-Rovner, H. Llewellyn-Thomas, A. Lyddiatt, R. Thomson, et al. Decision aids for people facing health treatment or screening decisions. *Cochrane database of systematic reviews*, 2017.

M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang. SNOMED clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association, 2001.

E. Steyerberg. *Clinical prediction models*. Springer New York, 2009.

A. Susaiyah, A. Härmä, E. Reiter, and M. Petković. Neural scoring of logical inferences from data using feedback. *International Journal of Interactive Multimedia and Artificial Intelligence*, 6(5):90–99, 2021.

The EPI Consortium. Epi: Enabling personalized interventions., 2019. URL `https://enablingpersonalizedinterventions.nl`. Last accessed 18 January 2023.

J. W. Tukey. We need both exploratory and confirmatory. *The American Statistician*, 34(1):23–25, 1980.

R. J. Turner. Safe tests for 2 x 2 contingency tables and the Cochran-Mantel-Haenszel test. Master's thesis, Leiden University, 2019.

R. J. Turner. PsyNLP, 2021. URL `https://github.com/rosanneturner/psynlp_outcome_measures`.

R. J. Turner. Netwerkanalyse van antidepressiva behandeltrajecten. In *NVVP Voorjaarscongres*, 2022.

R. J. Turner. safeSequentialTestingAISTATS2023, 2023. Code corresponding to AISTATS Paper, accessible at `https://github.com/rosanneturner/safeSequentialTestingAISTATS2023`.

R. J. Turner and P. D. Grünwald. Exact anytime-valid confidence intervals for contingency tables and beyond. *Statistics & Probability Letters*, page 109835, 2023.

R. J. Turner, A. Ly, and P. D. Grünwald. Generic e-variables for exact sequential k-sample tests that allow for optional stopping. *arXiv preprint arxiv:2106.02693*, 2021.

R. J. Turner, F. Coenen, F. Roelofs, K. Hagoort, A. Härmä, P. D. Grünwald, F. P. Velders, and F. E. Scheepers. Information extraction from free text for aiding transdiagnostic psychiatry: constructing nlp pipelines tailored to clinicians' needs. *BMC Psychiatry*, 22(1):407, 2022. doi: 10.1186/s12888-022-04058-z.

J. Ville. Étude critique de la notion de collectif, 1939.

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021.

E.-J. Wagenmakers and A. Ly. Bayesian Scepsis About SWEPIS: Quantifying the Evidence That Early Induction of Labour Prevents Perinatal Deaths. *PsyArXiv*, 2020. doi: 10.31234/osf.io/5ydpb.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186, 1945.

A. Wald. *Sequential Analysis*. Wiley, 1947.

R. Wang and A. Ramdas. False discovery rate control with e-values. *arXiv preprint arXiv:2009.02824*, 2020.

L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020.

R. L. Wasserstein and N. A. Lazar. The ASA statement on p-values: context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*, 2020.

U. Wennerholm, S. Saltvedt, A. Wessberg, M. Alkmark, C. Bergh, S. B. Wendel, H. Fadl, M. Jonsson, L. Ladfors, V. Sengpiel, et al. Induction of labour at 41 weeks versus expectant management and induction of labour at 42 weeks (SWEdish Post-term Induction Study, swepis): multicentre, open label, randomised, superiority trial. *British Medical Journal*, 367, 2019.

H. A. Whiteford, A. J. Ferrari, L. Degenhardt, V. Feigin, and T. Vos. The global burden of mental, neurological and substance use disorders: an analysis from the global burden of disease study 2010. *PLoS One*, 10(2):e0116820, 2015. doi: 10.1371/journal.pone.0116820.

P. Whiting, J. Savović, J. P. Higgins, D. M. Caldwell, B. C. Reeves, B. Shea, P. Davies, J. Kleijnen, R. Churchill, et al. ROBIS: a new tool to assess risk of bias in systematic reviews was developed. *Journal of clinical epidemiology*, 69: 225–234, 2016.

J. T. Wigman, J. van Os, E. Thiery, C. Derom, D. Collip, N. Jacobs, and M. Wichers. Psychiatric diagnosis revisited: towards a system of staging and profiling combining nomothetic and idiographic parameters of momentary mental states. *PLoS One*, 8(3):e59559, 2013. doi: 10.1371/journal.pone.0059559.

M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J. W. Boiten, L. B. da Silva Santos, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3:160018, 2016. doi: 10.1038/sdata.2016.18.

J. B. Williams. Standardizing the hamilton depression rating scale: past, present, and future. *Eur Arch Psychiatry Clin Neurosci*, 251 Suppl 2:II6–12, 2001. doi: 10.1007/BF03035120.

World Health Organization. The World Health Organization Quality of Life assessment (WHOQOL): position paper from the World Health Organization. *Soc Sci Med*, 41(10):1403–9, 1995. doi: 10.1016/0277-9536(95)00112-k.

World Health Organization. Fact sheet depression 2022, 2022. URL `https://www.who.int/news-room/fact-sheets/detail/depression`.

A. G. Yip, K. J. Ressler, F. Rodriguez-Villa, S. H. Siddiqi, and S. J. Seiner. Treatment outcomes of electroconvulsive therapy for depressed patients with and without borderline personality disorder: a retrospective cohort study. *The Journal of Clinical Psychiatry*, 82(2):28439, 2021.

M. Zimmerman, M. Posternak, M. Friedman, N. Attiullah, S. Baymiller, R. Boland, S. Berlowitz, S. Rahman, K. Uy, and S. Singer. Which factors influence psychiatrists' selection of antidepressants? *Am. J. Psychiatry*, 161(7): 1285–9, 2004. doi: 10.1176/appi.ajp.161.7.1285.

Appendix with Supplementary Material

Supplementary material for chapter 2

Appendix S2.A contains detailed proofs. Appendix S2.B contains a detailed description of the numerical approach to calculating e-variables for restricted $\mathcal{H}_1$. Appendix S2.C contains a detailed description of Gunel-Dickey Bayes factors. Appendix S2.D contains optional stopping experiments. Appendix S2.E explains how to adapt the block group sizes n_a and n_b based on past data.

S2.A Proofs

The proofs below repeatedly use Theorem 1 of Grünwald et al. [2022a] and a direct corollary (called Corollary 2 by Grünwald et al. [2022a]), which we restate here for convenience, combined as a single statement. We use the notation adopted later in the paper: for $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ and, for W a distribution on Θ_0, we write $P_W = \int P_\theta dW(\theta)$.

Theorem (Theorem 1 of Grünwald et al. [2022a]) Let Y be a random variable taking values in a set $\mathcal{Y}$. Suppose Q is a probability distribution for Y with density q that is strictly positive on all of $\mathcal{Y}$ and let $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ be a set of distributions for Y where each P_θ has density p_θ. Let $\mathcal{W}_0$ be the set all distributions on Θ_0. Assume $\inf_{W_0 \in \mathcal{W}_0(\Theta_0)} D(Q\|P_{W_0}) < \infty$. Then (a) there exists a (potentially sub-) distribution P_0^* with density p_0^* such that

$$S^* := \frac{q(Y)}{p_0^*(Y)}$$

is an e-variable (p_0^* is called the *Reverse Information Projection (RIPr) of q onto* $\{p_W : W \in \mathcal{W}_0\}$ [Li, 1999, Li and Barron, 2000, Grünwald et al., 2022a]). Moreover, (b), S^* satisfies

$$\sup_{S \in \mathcal{E}(\Theta_0)} \mathbf{E}_{Y\sim Q}[\log S] = \mathbf{E}_{Y\sim Q}[\log S^*] = \inf_{W_0 \in \mathcal{W}_0(\Theta_0)} D(Q\|P_{W_0}) = D(Q\|P_0^*). \tag{A.1}$$

and is thus the Q-GRO e-variable for Y. If the minimum is achieved by some W_0^*, i.e. $D(Q\|P_0^*) = D(Q\|P_{W_0^*})$, then $P_0^* = P_{W_0^*}$. Moreover, (c), if there exists an e-variable S of the form $q(Y)/p_{W_0}(Y)$ for some $W_0 \in \mathcal{W}_0$ then W_0 must achieve the infimum in (A.1) and S must be essentially equal to S^* in the sense that for all $P \in \mathcal{H}_0 \cup \{Q\}$, $P(S^* = q(Y)/p_{W_0}(Y)) = 1$. Similarly (d), if there exists a $W_0^* \in \mathcal{W}_0$ that achieves the infimum in (A.1) then $S = q(Y)/p_{W_0^*}(Y)$ is an e-variable and S is again essentially equal to S^*.

S2.A.1 Proof of Propositions

Proof of Proposition 1 Below we state and prove a slight generalization of Proposition 1.

Proposition 4 (generalization)**.** Let $\mathcal{H}_1 = \{Q\}$ be a singleton and let $\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\}$ be such that for some distribution W on Θ_0, $D(Q\|P_W) < \infty$. For general $\theta \in \Theta_0$ and distributions W on Θ_0, define $S_{\theta,(j)} := q(Y_{(j)})/p_\theta(Y_{(j)})$ and $S_{W,(j)} = q(Y_{(j)})/p_W(Y_{(j)})$. We have:

1. Suppose there exists a distribution W on Θ_0 such that $S_{W,(1)}$ is an e-variable. Then $S_{W,(1)}$ is the Q-GRO e-variable for $Y_{(1)}$. In particular, if W puts mass 1 on a particular $\theta^\circ \in \Theta_0$, then $S_{W,(1)} = S_{\theta^\circ,(1)}$ is the Q-GRO e-variable.

2. If $\Theta_0 = \{\theta_0\}$ is simple then, with the prior W_0 putting mass 1 on θ_0, $S_{W_0,(1)} = S_{\theta_0,(1)}$ is an e-variable and hence, by the above, also the Q-GRO e-variable.

3. If, for some $\theta^\circ \in \Theta_0$, $S_{\theta^\circ,(1)}$ is an e-variable and we further assume that $Y_{(1)}, Y_{(2)}, \ldots$ are i.i.d. according to all distributions in $\mathcal{H}_0 \cup \mathcal{H}_1$, then $S^{(m)}_{\mathrm{GRO}(Q)} = \prod_{j=1}^m S_{\theta^\circ,(j)}$; that is, the Q-GRO optimal (unconditional) e-variable for $Y^{(m)}$ is the product of the individual Q-GRO optimal e-variables.

Proof. Part 1 The theorem above, part (b), implies, with $Y = Y_{(1)}$, that some Q-GRO e-variable S^* for $Y_{(1)}$ exists. Part (c) then implies that we can take S^* to be equal to $S_{W,(1)}$. This implies the statement.

Part 2 is immediate.

Part 3 We assume that $S_{\theta^\circ,(1)}$ is an e–variable. Then the i.i.d. assumption implies that $S^{(m)}_{\theta^\circ} := \prod_{j=1}^m S_{\theta^\circ,(j)} = \prod q(Y_{(j)})/p_{\theta^\circ}(Y_{(j)})$ is also an e-variable. But [Grünwald et al., 2022a, Theorem 1], part (c) as stated above implies (by taking a distribution W putting mass 1 on θ) that for $\mathcal{H}_0$ for which data are i.i.d., for each $m \geq 1$, that if a $\theta \in \Theta_0$ exists such that $S^{(m)}_\theta$ is an e-variable, then $S^{(m)}_\theta$ must be the Q-GRO e-variable for $Y^{(m)}$. This proves the statement. □

Proof of Proposition 2 The formulae for $\breve{\theta}_a|Y^{(j-1)}$ and $\breve{\theta}_b|Y^{(j-1)}$ are standard expressions for the Bayes predictive distribution based on the given beta priors; we omit further details. As to the expression for $\breve{\theta}_0|Y^{(j-1)}$ in terms of $\kappa = n_b/n_a$: Straightforward rewriting gives, for general $\alpha_a, \alpha_b, \beta_a, \beta_b$:

$$\breve{\theta}_0|Y^{(j-1)} = \frac{1}{1+\kappa}\breve{\theta}_a|Y^{(j-1)} + \frac{\kappa}{1+\kappa}\breve{\theta}_b|Y^{(j-1)}. \tag{A.2}$$

If we plug in the expressions for $\breve{\theta}_a|Y^{(j-1)}, \breve{\theta}_b|Y^{(j-1)}$ and we instantiate to $\alpha_b = \kappa\alpha_a$, and $\beta_b = \kappa\beta_a$, this becomes

$$\begin{aligned}\breve{\theta}_0|Y^{(j-1)} &= \frac{1}{1+\kappa}\frac{U_a+\alpha_a}{n_a(j-1)+\alpha_a+\beta_a} + \frac{\kappa}{1+\kappa}\frac{U_b+\alpha_b}{\kappa(n_a(j-1)+\alpha_a+\beta_a)} \\ &= \frac{1}{1+\kappa}\frac{U_a+U_b+(1+\kappa)\alpha_a}{n_a(j-1)+\alpha_a+\beta_a} = \frac{U+(1+\kappa)\alpha_a}{n(j-1)+(1+\kappa)\alpha_a+(1+\kappa)\beta_a},\end{aligned}$$

which is what we had to prove.

S2.A.2 Proof of Theorem 1

We first restate Theorem 1 in its extended version that holds for $k \geq 2$ data streams. Let $\vec{n} = (n_1, \ldots, n_k), n = \sum_{g=1}^k n_g, \vec{\theta} = (\theta_a, \ldots, \theta_k) \in \Theta^k$ and $\vec{y}^n$ be as defined in the main text (3.3). We use '$\vec{Y}^n \sim P_{\theta^*}$' as an abbreviation for '$Y_1^{n_1} \sim P_{\theta_1^*}; \ldots; Y_k^{n_k} \sim P_{\theta_k^*}$'.

Theorem .1 (extended)**.** Let

$$s(\vec{y}^n; \vec{n}, \vec{\theta}^*) := \prod_{g=1}^k \frac{p_{\theta_g^*}(y_g^{n_g})}{\prod_{i=1}^{n_g}\left(\sum_{g'=1}^k \frac{n_{g'}}{n} p_{\theta_{g'}^*}(y_{i,g})\right)}.$$

The random variable $S_{[\vec{n},\vec{\theta}^*]} := s(\vec{Y}^n; \vec{n}, \vec{\theta}^*)$ is an e-variable, i.e. we have:

$$\sup_{\theta\in\Theta} \mathbf{E}_{V^n\sim P_\theta}\left[s(V^n; \vec{n}, \vec{\theta}^*)\right] \leq 1.$$

Moreover, if $\{P_\theta : \theta \in \Theta\}$ is a convex set of distributions, then $S_{[\vec{n},\vec{\theta}^*]}$ is the $(\vec{\theta}^*)$-GRO e-variable: for any non-negative function s' on $\mathcal{Y}^n$ satisfying $\sup_{\theta\in\Theta} \mathbf{E}_{V^n\sim P_\theta}[s'(V^n)] \leq 1$, we have:

$$\mathbf{E}_{\vec{Y}^n\sim P_{\theta^*}}[\log s(\vec{Y}^n; \vec{n}, \vec{\theta}^*)] \geq \mathbf{E}_{\vec{Y}^n\sim P_{\theta^*}}[\log s'(\vec{Y}^n)].$$

Proof of Theorem .1 The following fact plays a central role in the proof:

Fact For $g \in (1, ..., k)$, let $n_g \in \mathbf{N}, n := \sum_{g=1}^k n_g$ and let $u_g \in \mathbf{R}^+$. Suppose that $\sum_{g=1}^k n_g u_g \leq n$. Then $\prod_{g=1}^k u_g^{n_g} \leq 1$.

This result follows from the following standard generalization of Young's inequality to k numbers: for any k numbers $u_1, \ldots, u_k \in \mathbf{R}_0^+$ and any k nonnegative numbers $p_1, \ldots, p_k$ with $\sum_{g=1}^k p_g = 1$, we have $\prod_{g=1}^k u_g^{p_g} \leq \sum_{g=1}^k p_g u_g$. Applying this with $p_g = n_g/n$ to u_g and n_g as above, we get $\prod_{g=1}^k u_g^{n_g/n} \leq \sum_{g=1}^k (n_g u_g)/n \leq 1$, and the result follows by exponentiating to the power n.
Part 1 For $y \in \mathcal{Y}$, set set $p^\circ(y) := \sum_{g=1}^k (n_g/n) p_{\theta_g^*}(y)$ and $p^\circ(y^m) = \prod_{i=1}^m p^\circ(y_i)$.

For all $\theta \in \Theta$ we have:

$$\mathbf{E}_{V^n \sim P_\theta}\left[s(V^n; \vec{n}, \vec{\theta^*})\right] = \prod_{g=1}^{k} \mathbf{E}_{Y_g^{n_g} \sim P_\theta}\left[\frac{p_{\theta_g^*}(Y_g^{n_g})}{p^\circ(Y_g^{n_g})}\right] = \prod_{g=1}^{k}\left(\mathbf{E}_{Y\sim P_\theta}\left[\frac{p_{\theta_g^*}(Y)}{p^\circ(Y)}\right]\right)^{n_g}. \tag{A.3}$$

We also have

$$\sum_{g=1}^{k} \frac{n_g}{n} \mathbf{E}_{Y\sim P_\theta}\left[\frac{p_{\theta_g^*}(Y)}{p^\circ(Y)}\right] = \mathbf{E}_{Y\sim P_\theta}\left[\sum_{g=1}^{k} \frac{n_g}{n} \cdot \frac{p_{\theta_g^*}(Y)}{\sum_{g'=1}^{k} \frac{n_{g'}}{n} p_{\theta_{g'}^*}(Y)}\right] = 1. \tag{A.4}$$

The result now follows by combining (A.3) with (A.4) using the Fact further above.

Part 2 By convexity of $\{P_\theta : \theta \in \Theta\}$, there exists $\theta^\circ \in \Theta$ such that $p_{\theta^\circ} = \sum_{g=1}^{k}(n_g/n)p_{\theta_g^*}$ and then the numerator in (A.4) can we rewritten as $p_{\theta^\circ}(\vec{y})$. The GRO-property is now an immediate consequence of Proposition 4, Part 1.

S2.B Numerical approach to calculating e-variables for restricted $\mathcal{H}_1$

In this subsection we describe how we propose to approximate the beta prior and posterior on the restricted $\mathcal{H}_1$ with parameter space $\Theta(\delta)$, as defined in (5.1). Note that we limit ourselves to $\delta > 0$ in this detailed description; for $\delta < 0$ one can apply an entirely equivalent approach, with an extra term in the reparameterization. We define

$$\zeta = \begin{cases} \delta \text{ if } d((\theta_a, \theta_b)) = \theta_b - \theta_a, \\ 0 \text{ if } d((\theta_a, \theta_b)) = \text{log-odds-ratio}(\theta_a, \theta_b), \end{cases}$$

such that we have $\theta_a \in (0, 1-\zeta)$ and in both cases, θ_b is completely determined by θ_a: $\theta_b = d^{-1}(\delta; \theta_a)$. Hence, our density estimation problem now becomes one-dimensional, which enables us to put a discretized prior on the restricted parameter space.

First, we discretize the parameter space Θ_a to a grid (a vector) with precision $K, K \in (0, 1-\zeta)$ and $1/K \in \mathbb{N}^+$: $\bar{\boldsymbol{\theta}}_{\boldsymbol{a}} = (K, 2K, 3K, \ldots, 1-\zeta)$. Then, we reparameterize $\theta_a = (1-\zeta)\rho$, with $\rho \in (0,1)$. Then, we have $\bar{\boldsymbol{\rho}} = (K/(1-\zeta), 2K/(1-\zeta), \ldots, 1)$. For the discretized grid $\bar{\boldsymbol{\rho}}$, we compute the prior $W = \text{Beta}(\alpha, \beta)$ densities and normalize them, which also gives us the discretized densities for each $\theta_a^i \in \bar{\boldsymbol{\theta}}_{\boldsymbol{a}}$ (with $i \in (1, 2, \ldots, 1/K)$):

$$\pi_{\alpha,\beta,\zeta}(\theta_a^i) = \frac{\text{Beta}(\frac{\theta_a^i}{1-\zeta}; \alpha, \beta)}{\sum_{k=1}^{\frac{1}{K}} \text{Beta}(\frac{\theta_a^k}{1-\zeta}; \alpha, \beta)}.$$

For all elements of $\bar{\boldsymbol{\theta}}_{\boldsymbol{a}}$, the corresponding θ_b is retrieved and the likelihood of incoming data points $p_{\theta_a,\theta_b}(Y^{(j-1)})$ is calculated. We can then estimate the posterior

density of $\theta_a^i \in \bar{\boldsymbol{\theta}}_{\boldsymbol{a}}$:

$$p(\theta_a^i|Y^{(j-1)}) = \frac{\pi_{\alpha,\beta,\zeta}(\theta_a^i)p_{\theta_a^i,\theta_b^i}(Y^{(j-1)})}{\sum_{k=1}^{\frac{1}{K}} \pi_{\alpha,\beta,\zeta}(\theta_a^k)p_{\theta_a^k,\theta_b^k}(Y^{(j-1)})}.$$

We can then estimate $\breve{\theta}_a|Y^{(j-1)} = \mathbf{E}_{\theta_a \sim W|Y^{(j-1)}}[\theta_a]$ as $\sum_{i=1}^{\frac{1}{K}} p(\theta_a^i|Y^{(j-1)})\theta_a^i$, and $\breve{\theta}_b|Y^{(j-1)} = d^{-1}(\delta; \theta_a|Y^{(j-1)})$.

S2.C The Gunel-Dickey Bayes Factors do not give rise to e–variables

Sampling	Fixed	Bayes factor (10) for 2x2 table
Poisson	none	$\frac{8(n+1)(n_1+1)}{(n+4)(n+2)}\left[\frac{n_{a1}!n_{b1}!n_{a0}!n_{b0}!n!}{(n_1+1)!n_0!n_a!n_b!}\right]$
Joint multinomial	n	$\frac{6(n+1)(n_1+1)}{(n+3)(n+2)}\left[\frac{n_{a1}!n_{b1}!n_{a0}!n_{b0}!n!}{(n_1+1)!n_0!n_a!n_b!}\right]$
Independent multinomial	n_a, n_b	$\binom{n}{n_1}/\left(\binom{n_a}{n_{a1}}\binom{n_b}{n_{b1}}\right)\frac{(n+1)}{(n_a+1)(n_b+1)}$
Hypergeometric	n_a, n_b, n_1	$\frac{n_{a1}!n_{b1}!n_{a0}!n_{b0}!n!}{\prod_{i\in\{a,b,0,1\}}(n_i+\mathbb{I}n_i=min(n_a,n_b,n_0,n_1))!}$

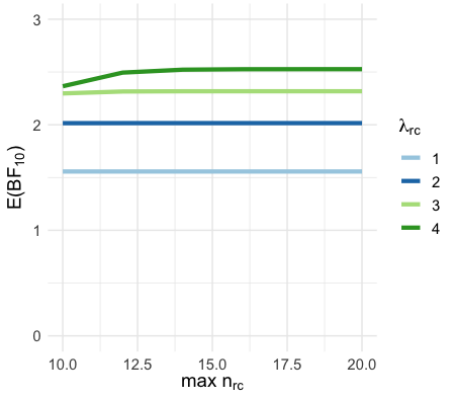
Table S2.1: Overview of (objective) Bayes factors for contingency table testing provided by Gunel and Dickey [1974] and Jamil et al. [2017].

We will not consider the hypergeometric and joint multinomial scenarios for this paper, where the number of successes n_1 is fixed, as they do not match the block-wise data design in this paper. The Bayes factor for the Poisson sampling scheme is not an e-variable, as the expectation under the null hypothesis with Poisson distributions on individual cell counts exceeds 1 for rates $\lambda \geq 1$:

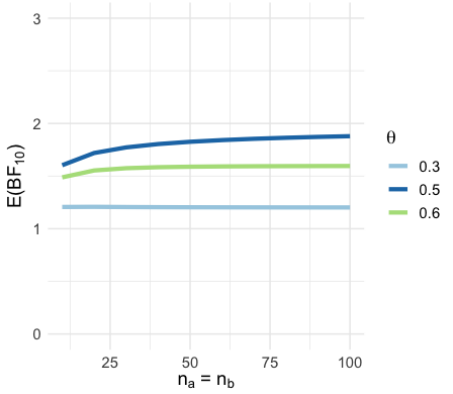
$$\mathbb{E}_{n_{rc}\sim\text{Poisson}(\lambda_{rc})}\left[BF_{10}(N_{a1}, N_{b1}, N_{a0}, N_{b0})\right] =$$
$$\sum_{n_{a1}=0}^{\infty} \dots \sum_{n_{b0}=0}^{\infty} \pi_{\lambda_{a1}}(n_{a1}) \dots \pi_{\lambda_{b0}}(n_{b0}) BF_{10}(n_{a1}, n_{b1}, n_{a0}, n_{b0}) =$$
$$\frac{8}{\exp(\lambda_{a1} + \ldots + \lambda_{b0})} \sum_{n_{a1}=0}^{\infty} \dots \sum_{n_{b0}=0}^{\infty} \lambda_{a1}^{n_{a1}} \dots \lambda_{b0}^{n_{b0}} \frac{(n+1)(n_1+1)}{(n+4)(n+2)} \frac{n!}{(n_1+1)!n_0!n_a!n_b!},$$

as illustrated numerically in Figure S2.1 for increasing limits for the sums $\sum_{n_{rc}=1}^{\max n_{rc}}$.

For the independent multinomial sampling scheme, let, without loss of gener-



(a) The Gunel-Dickey Bayes factor for the Poisson sampling scheme is not an e-variable: $\sum_{n_{a1}=0}^{\max n_{rc}} \dots \sum_{n_{b0}=0}^{\max n_{rc}} \pi_{\lambda_{a1}}(n_{a1}) \dots \pi_{\lambda_{b0}}(n_{b0}) BF_{10}(n_{a1}, n_{b1}, n_{a0}, n_{b0})$ for various $\max n_{rc}$ and λ_{rc}.



(b) The Gunel-Dickey Bayes factor for the independent multinominal sampling scheme is not an e-variable: $\mathbb{E}_{N_{a1}, N_{b1} \sim \text{Binomial}(\theta)} [BF_{10}(N_{a1}, N_{b1} | n_a, n_b)]$ for various choices of θ and n_g.

Figure S2.1: GD

ality, $n_a < n_b$. We get, with $n_0 = n - n_1$,

$$\mathbb{E}_{N_{a1},N_{b1}\sim\text{Binomial}(\theta)}\left[BF_{10}(N_{a1}, N_{b1}|n_a, n_b)\right] =$$
$$\sum_{n_{a1}=0}^{n_a}\sum_{n_{b1}=0}^{n_b}\binom{n_a}{n_{a1}}\binom{n_b}{n_{b1}}\theta^{n_1}(1-\theta)^{n_0}\frac{\binom{n}{n_1}}{\binom{n_a}{n_{a1}}\binom{n_b}{n_{b1}}}\frac{(n+1)}{(n_a+1)(n_b+1)} =$$
$$\frac{(n+1)}{(n_a+1)(n_b+1)}\sum_{n_{a1}=0}^{n_a}\sum_{n_{b1}=0}^{n_b}\binom{n}{n_1}\theta^{n_1}(1-\theta)^{n_0}$$

Numerical simulations show that, for a range of choices for n, n_a and θ this exceeds 1; see Figure S2.1.

S2.D Type-I error guarantee under optional stopping

Type-I Error In Figure S2.2 type-I error rates of several e-variables and Fisher's exact test estimated through a simulation experiment are depicted. 2000 samples of length 1000 were drawn according to a Bernoulli(0.1) distribution to represent 1000 data streams in two groups. After each complete block $m \in \{1, \ldots, 1000\}$ an e-value or p-value was calculated and the proportion of rejected experiments up until m with each test type was recorded. As the stream lengths increase, the type-I error rate under (incorrectly applied) optional stopping with Fisher's exact test increases quickly. The type-I error rate of the e-variables remains bounded.

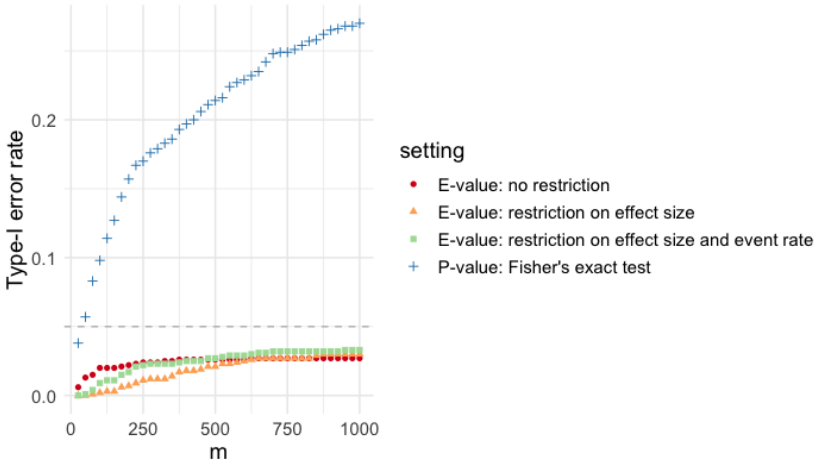


Figure S2.2: Type-I error rates for various e-variables and Fisher's exact test under optional stopping estimated with 1000 simulations of two Bernoulli(0.1) data streams of length 1000, with $n_a = n_b = 1$. Significance level $\alpha = 0.05$ was used (grey dashed line). For the safe tests, beta prior parameter values used were $\gamma = \alpha_a = \beta_a = \alpha_b = \beta_b = 1/2$ ($\gamma = 0.18$ gave comparable results). For the e-variables with restrictions on $\mathcal{H}_1$, we used $\delta = 0.05$ and $\theta_a = 0.1$.

S2.E Adjusting n_a and n_b based on past data

To see how to choose n_a and n_b for subsequent blocks based on past data, we first need to formalize the fact that data in different streams may arrive asynchronously. Thus, let $t = 1, 2, \ldots$ represent global ('calendar') time, and introduce corresponding random variables V_t and G_t: at each t, we obtain an outcome V_t in $\mathcal{Y}$ in group $G_t \in \{a, b\}$. We make no assumptions about the relative ordering of outcomes from the two groups. At time t, we have that t_a, the number of a's that are observed so far, and t_b, the number of b's observed so far, satisfy $t_a + t_b = t$, but subject to this constraint we allow them coming in any order. We now introduce a function $f : \bigcup_{t \geq 0} \mathcal{Y}^t \times \{0, 1\}^t \to \{\text{STOP-BLOCK}, \text{CONTINUE}\}$ that, at each point in time t,

decides whether the current block should end ($f(V^t, G^t) = \text{STOP-BLOCK}$) or not ($f(V^t, G^t) = \text{CONTINUE}$). As long as the value of this function does not depend on the actual outcomes V_t observed after the last block that was completed, all requirements for having a test martingale and thus for safe optional stopping are met. For example, suppose that on data $V_1, G_1, V_2, G_2, \ldots, V_t, G_t$ observed so-far, f has output STOP-BLOCK at m occasions, the last time at $t' = t-k$ for some $k > 0$. Then $f(t)$ is allowed to depend on $Y^{(m)}$ and G^t, but for any fixed $Y^{(m)} = y^{(m)}, G^t = g^t$, for all $y^k, y'^k \in \mathcal{Y}^k$, we must have $f((y^{(m)}, y^k), g^t) = f((y^{(m)}, y'^k), g^t)$.

Supplementary material for chapter 3

Appendix section S3.A contains proofs and section S3.B contains extended simulation results.

S3.A Proofs

Both proofs below use Theorem 1 of Grünwald et al. [2022a] and a direct corollary (called Corollary 2 by Grünwald et al. [2022a]), which we re-state here, for convenience, combined as a single statement. Recall that we use notation $P_W := \int P_{\vec{\theta}} dW(\vec{\theta})$.

Theorem (Theorem 1 of Grünwald et al. [2022a]) Let Y be a random variable taking values in a set $\mathcal{Y}$. Suppose Q is a probability distribution for Y with density q that is strictly positive on all of $\mathcal{Y}$ and let $\mathcal{H}_0 = \{P_{\vec{\theta}} : \vec{\theta} \in \vec{\Theta}_0\}$ be a set of distributions for Y where each $P_{\vec{\theta}}$ has density $p_{\vec{\theta}}$. Let $\mathcal{W}_0$ be the set of all distributions on $\vec{\Theta}_0$. Assume $\inf_{W_0 \in \mathcal{W}_0(\vec{\Theta}_0)} D(Q\|P_{W_0}) < \infty$. Then (a) there exists a (potentially sub-) distribution P_0^* with density p_0^* such that

$$S^* := \frac{q(Y)}{p_0^*(Y)}$$

is an e-variable (p_0^* is called the *Reverse Information Projection (RIPr) of q onto* $\{p_W : W \in \mathcal{W}_0\}$). Moreover, (b), S^* satisfies

$$\sup_{S \in \mathcal{E}(\vec{\Theta}_0)} \mathbf{E}_{Y\sim Q}[\log S] = \mathbf{E}_{Y\sim Q}[\log S^*] = \inf_{W_0 \in \mathcal{W}_0(\vec{\Theta}_0)} D(Q\|P_{W_0}) = D(Q\|P_0^*). \tag{A.5}$$

(where $\mathcal{E}(\vec{\Theta}_0)$ is the set of all e-variables relative to null hypothesis $\mathcal{H}_0$) and S^* is thus the Q-GRO e-variable for Y. If the minimum is achieved by some W_0^*, i.e. $D(Q\|P_0^*) = D(Q\|P_{W_0^*})$, then $P_0^* = P_{W_0^*}$. Moreover, (c), if there exists an e-variable S of the form $q(Y)/p_{W_0}(Y)$ for some $W_0 \in \mathcal{W}_0$ then W_0 must achieve the infimum in (A.5) and S must be essentially equal to S^* in the sense that for all $P \in \mathcal{H}_0 \cup \{Q\}$, $P(S^* = q(Y)/p_{W_0}(Y)) = 1$. Similarly (d), if there exists a $W_0^* \in \mathcal{W}_0$ that achieves the infimum in (A.5) then $S = q(Y)/p_{W_0^*}(Y)$ is an e-variable and S is again essentially equal to S^*.

Proof of Theorem 3.1 *Part 1* The real idea behind the proof is the formulation of the modified testing problem in which only a single outcome per block is observed. This we already did in the main text. Linking the two is simply the last, very simple step, with analogies to the proof of Part 1 of Theorem 1 in Turner et al. [2021].

Let $n_a, n_b \in \mathbf{N}, n := n_a + n_b$ and let $u, v \in \mathbf{R}^+$. Suppose that $n_a u + n_b v \leq n$. Then $u^{n_a} v^{n_b} \leq 1$, which follows immediately from applying Young's inequality to

$u^{n_a/n}, v^{n_b/n}$ but can also be derived directly by writing v as function of u and differentiating $\log(u^{n_a} v^{n_b})$ to u.

Further, by independence, for $(\theta_a, \theta_b) \in \vec{\Theta}_0$,

$$\begin{aligned}
&\mathbf{E}_{Y_a^{n_a} \sim P_{\theta_a}, Y_b^{n_b} \sim P_{\theta_b}} [s'(Y_a^{n_a}, Y_b^{n_b})] = \\
&\mathbf{E}_{Y_a^{n_a} \sim P_{\theta_a}} \left[\frac{p_{\theta_a^*}(Y_a^{n_a})}{p^\circ(Y_a^{n_a}|a)} \right] \cdot \mathbf{E}_{Y_b^{n_b} \sim P_{\theta_b}} \left[\frac{p_{\theta_b^*}(Y_b^{n_b})}{p^\circ(Y_b^{n_b}|b)} \right] = \\
&\left(\mathbf{E}_{Y \sim P_{\theta_a}} \left[\frac{p_{\theta_a^*}(Y)}{p^\circ(Y|a)} \right] \right)^{n_a} \cdot \left(\mathbf{E}_{Y \sim P_{\theta_b}} \left[\frac{p_{\theta_b^*}(Y)}{p^\circ(Y|b)} \right] \right)^{n_b} = \\
&\left(\mathbf{E}_{Y \sim P'_{\theta}|a} \left[\frac{p'_{\theta^*}(Y|a)}{p^\circ(Y|a)} \right] \right)^{n_a} \cdot \left(\mathbf{E}_{Y \sim P'_{\theta|b}} \left[\frac{p'_{\theta^*}(Y|b)}{p^\circ(Y|b)} \right] \right)^{n_b}. \qquad \text{(A.6)}
\end{aligned}$$

Combining the two facts stated above, (3.6) implies that the latter quantity is bounded by 1.

Part 2 By lower-semicontinuity of the KL divergence in its second argument (Posner's theorem, used as in Grünwald et al. [2022a]) the infimum in (3.4) is achieved by some prior distribution W° so that by Theorem 1 of Grünwald et al. [2022a] (part (b) in the formulation above), $p^\circ(\cdot \mid \cdot) = p'_{W^\circ}(\cdot \mid \cdot)$ and hence also $P^\circ(G, Y) = P'_{W^\circ}(G, Y)$. By convexity of $\mathcal{H}'_0$ and finiteness of the support of $P'_{\vec{\theta}}(G, Y)$, there must be some $\vec{\theta}$ such that $P'_{W^\circ}(G, Y) = P_{\vec{\theta}}(G, Y)$ and hence also $p'_{W^\circ}(\cdot \mid \cdot) = p'_{\vec{\theta}}(\cdot \mid \cdot)$, which shows (a). This means that we have now created an e-variable for the original problem which can be written as $p_{\theta_a^*, \theta_b^*} / p_{W_0}$ with p_{W_0} a prior distribution on $\vec{\theta}_0$ (namely, the one that puts mass 1 on $\vec{\theta}$). (b) is then an immediate consequence of Theorem 1 of Grünwald et al. [2022a] (part (c) in the formulation above). (note that we *cannot* draw this conclusion if $\mathcal{H}'_0$ is not convex; for then the distribution p'_{W° may not correspond to the distribution p_{W° in the original problem — this correspondence is only guaranteed if p'_{W° coincides with some $p'_{\vec{\theta}}$.

Proof of Theorem 3.2 Recall that we assume that $\vec{\Theta}_0$ is convex and compact. We set $\textsc{kl}'(\theta_a, \theta_b) := D(P'_{\theta_a^*, \theta_b^*} \| P'_{\theta_a, \theta_b})$ where D is the KL divergence as in (3.5), i.e. for the modified setting in which P'_{θ_a, θ_b} is a distribution on a single outcome, as discussed before Theorem 3.1. For the 2×2 model this KL divergence can be written explicitly as

$$\begin{aligned}
D(P'_{\theta_a^*, \theta_b^*} \| P'_{\theta_a, \theta_b}) &= \mathbf{E}_{G \sim Q'} \mathbf{E}_{Y \sim P'_{\vec{\theta}^*}|G} \left[\log \frac{p'_{\vec{\theta}^*}(Y|G)}{p'_{\vec{\theta}}(Y|G)} \right] \qquad \text{(A.7)} \\
&= \frac{n_a}{n} \mathbf{E}_{Y \sim P'_{\theta_a^*}} \left[\log \frac{p_{\theta_a^*}(Y)}{p_{\theta_a}(Y)} \right] + \frac{n_b}{n} \mathbf{E}_{Y \sim P'_{\theta_b^*}} \left[\log \frac{p_{\theta_b^*}(Y)}{p_{\theta_b}(Y)} \right] \\
&= \frac{n_a}{n} \sum_{y_a \in \{0,1\}} p_{\theta_a^*}(y_a) \log \frac{p_{\theta_a^*}(y_a)}{p_{\theta_a}(y_a)} + \frac{n_b}{n} \sum_{y_b \in \{0,1\}} p_{\theta_b^*}(y_b) \log \frac{p_{\theta_b^*}(y_b)}{p_{\theta_b}(y_b)}.
\end{aligned}$$

From (3.8) we now see that $n\text{KL}'(\theta_a, \theta_b) = \text{KL}(\theta_a, \theta_b)$. We will prove the theorem with KL replaced by KL$'$ and $\mathcal{H}_0$ by $\mathcal{H}_0'$; since the two KL's agree up to a constant factor of n, all results transfer to the KL mentioned in the theorem statement.

Since $\vec{\Theta}_0$ is compact in the Euclidean topology and all distributions in $\mathcal{H}_0'$ can be represented as 2-dimensional vectors, i.e. they have common and finite support, we must have that $\mathcal{H}_0$ is compact in the weak topology so we can use the lower-semicontinuity of KL divergence in its second argument (Posner's theorem) as in [Grünwald et al., 2022a] to give us that the minimum KL divergence $\min \text{KL}'(\theta_a, \theta_b)$ is achieved by some $(\theta_a^\circ, \theta_b^\circ)$. Since KL divergence is strictly convex in its second argument and $\mathcal{H}_0'$ is convex (this is the place where we need to use KL$'$ rather than KL: $\mathcal{H}_0$ may *not* be convex!), the minimum must be achieved uniquely. Since KL divergence $\text{KL}'(\theta_a, \theta_b)$ is nonnegative and 0 only if $(\theta_a, \theta_b) = (\theta_a^*, \theta_b^*)$, it follows that $(\theta_a^\circ, \theta_b^\circ) = (\theta_a^*, \theta_b^*)$ if $\min \text{KL}(\theta_a, \theta_b) = 0$. Otherwise, since we assume (θ_a^*, θ_b^*) to be in the interior of $[0,1]^2$, $\text{KL}(\theta_a, \theta_b) = \infty$ iff (θ_a, θ_b) lies on the boundary of $[0,1]^2$. Thus, $(\theta_a^\circ, \theta_b^\circ)$ must lie in the interior of $[0,1]^2$ as well. $(\theta_a^\circ, \theta_b^\circ)$ cannot lie in the interior of $\vec{\Theta}_0$ though: for any point (θ_a, θ_b) in the interior of $\vec{\Theta}_0$ we can draw a line segment between this point and (θ_a^*, θ_b^*). Differentiation along that line gives that $\text{KL}'(\theta_a, \theta_b)$ monotonically decreases as we move towards (θ_a^*, θ_b^*), so the minimum within the closed set $\vec{\Theta}_0$ must lie on its boundary.

It remains to show that (3.9) is the (θ_a^*, θ_b^*)-GRO e-variable relative to $\mathcal{H}_0$. To see this, note that, by convexity of $\mathcal{H}_0'$, from Theorem 3.1, we must have that the GRO e-variable for this original problem is of the form

$$\frac{p_{\theta_a^*}(y_a^{n_a}) p_{\theta_b^*}(y_b^{n_b})}{p_{\theta_a^+}(y_a^{n_a}) p_{\theta_b^+}(y_b^{n_b})}$$

for some (θ_a^+, θ_b^+). The result then follows again by Theorem 1 of Grünwald et al. [2022a] (part (c) in the formulation above): this shows that the distribution W_0 that puts mass 1 on (θ_a^+, θ_b^+) minimizes, among all distributions W on $\vec{\Theta}_0$, $D(P_{\theta_a^*, \theta_b^*} \| P_W)$. Since the set of such distributions includes all distributions that put mass 1 on *some* $(\theta_a, \theta_b) \in \vec{\Theta}_0$, we must have that $(\theta_a^+, \theta_b^+) = (\theta_a^\circ, \theta_b^\circ)$.

S3.B Extended simulation results

Numerical example We here give a small numerical example to illustrate the construction of our confidence sequences. For this example, we will look in detail at the data used to generate the second row of Figure 3.2a, the second panel, where we have observed 500 data blocks, with 27 "successes" ($y = 1$) in group a, and 136 "successes" in group b. To estimate δ_{L} and δ_{R}, $S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0]}$ as in (7.14) was calculated for that specific data stream, for a grid of possible δ, each defining one $\vec{\Theta}_0$; here, a grid with size 100 and a precision of 0.02 on $[-1, 1]$ was applied. The prior W_1 for the posterior mean was chosen as a Beta prior with $\alpha = \beta = 0.18$ according to Turner et al. [2021]. The area corresponding to values of δ for which $S^{(m)}_{[n_a,n_b,W_1;\vec{\Theta}_0]} < \frac{1}{0.05}$ after block $m = 500$ represents the confidence interval. For example, for the lower bound, δ_{L}, the smallest value of δ that did not lead to rejection was 0.15, with a corresponding e-value of 2.23. The e-value corresponding to $\delta = 0.13$ was 24.17, hence this risk difference was excluded from the confidence interval.

Running intersection In Figure S3.1, confidence sequence width is compared with and without applying the running intersection.

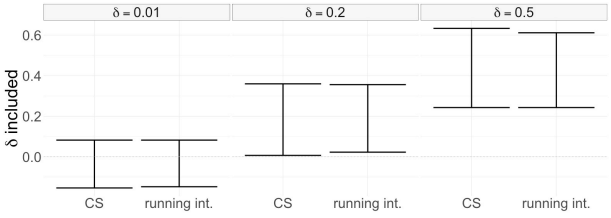


Figure S3.1: Confidence sequence with and without running intersection, for data generated under $P_{\theta_a,\theta_a+\delta}$ with $\theta_a = 0.05$, for a data stream of length 100. The significance threshold was set to 0.05. The design was balanced, with data block sizes $n_a = 1$ and $n_b = 1$.

Supplementary material for chapter 4

The following contains Section 1, *examples of theme and change phrases used for filtering sentences in the NLP pipeline*, of the supplementary material for Chapter 4 in this thesis. The other sections of the supplementary material can be found online in the publication corresponding to this chapter in *BMC Psychiatry* [Turner et al., 2022].

Table S4.1: Examples from the lists used for rule-based filtering of the four themes and change phrases

Category	**Dutch**	**Translation to English**	**Sentiment score**
Symptom reduction	Angstiger	More anxious	-1
	Angstigheid	Anxiety	-1
	Agressie	Aggression	-1
	Agresie	Aggression (misspelled)	-1
	Somber	Sad	-1
	Somer	Sad (misspelled)	-1
	Rotgevoel	Bad feeling	-1
	Doelloosheid	Aimlessness	-1
Social functioning	Zelfstandig	Independent	1
	Zelfstandige	Independent (conjugation)	1
	Zelfstandigheid	Independence	1
	Resocialiseren	Resocialize	1
	Participeert	Participates	1
	Vriendinnen	Girlfriends	1
	Vriendschappen	Friendships	1
	Verantwoordelijkheid	Responsibility	1
General well-being	Welbevinden	Well-being	1
	Welzijn	Well-being (synonym)	1
	Hoop	Hope	1
	Zingeving	Meaning	1
	Zinvol	Meaningful	1
	Zelfwaardering	Self-esteem	1
	Eigenwaarde	Self-esteem (synonym)	1
	Zelfvertrouwen	Self-confidence	1
	Zelfvetouwen	Self-confidence (misspelled)	1
Patient experience	Voelde	Felt	1
	Nez	In their own words (abbreviated)	1
	Voelt	Feels	1

Table with examples, continued

Category	**Dutch**	**Translation to English**	**Sentiment score**
	Uitte	Expressed	1
	Verwoorde	Articulated	1
	Constateert	Noted	1
	Merkt	Notes	1
	Mekrt	Notes (misspelled)	1
Change indicator	Afnam	Decreased	-1
	Afname	Decrease	-1
	Afgenomen	Decreased (conjugation)	-1
	Toenemende	Increasing	1
	Toenemde	Increasing (misspelled)	1
	Verbeter	Improve	1
	Verminder	Reduce	-1
	Vermindern	Reduce (misspelled)	-1

Supplementary material for chapter 5

Table S5.1: Overview of antidepressant prescription groups and specific antidepressants present in the data

Group	Antidepressant
MAOI	Tranylcypromine
	Moclobemide
	Phenelzine
nSSRI	Trazodone
	Duloxetine
	Venlafaxine
Other	Bupropion
	Vortioxetine
	Agomelatine
	Hyperici herba
SSRI	Sertraline
	Citalopram
	Escitalopram
	Fluoxetine
	Paroxetine
	Fluvoxamine
TetraCA	Mirtazapine
	Mianserine
TriCA	Nortriptyline
	Amitriptyline
	Clomipramine
	Imipramine
	Doxepine
	Maprotiline
	Dosulepine

Table S5.2: Overview of therapeutic dose range for selection of antidepressant treatment trajectories

antidepressant	Minimal dose	Maximal dose
tranylcypromine	10	60
phenelzine	8	120
moclobemide	100	600
clomipramine	10	250
nortriptyline	20	250
amitriptyline	10	150
imipramine	10	300

Table with dose ranges, continued

antidepressant	**Minimal dose**	**Maximal dose**
dosulepin	50	225
doxepin	25	300
trimipramine	NA	NA
venlafaxine	75	375
mirtazapine	15	45
trazodone	100	400
bupropion	150	300
duloxetine	60	120
agomelatine	25	50
vortioxetine	5	20
hyperici herba	NA	NA
sertraline	50	200
citalopram	10	40
fluoxetine	20	60
escitalopram	5	20
paroxetine	20	50
fluvoxamine	50	300

Table S5.3: Detailed summary of outcome measures per antidepressant prescription group.

AD Type	Facility	N	Continuation	Med. dur. until switch	Prescription duration	Core complaints	Social	Well-being	Experience
SSRI	PG	2244	0.680	77	162	-0.166	0.337	0.301	-0.084
	UMCU	316	0.924	16	92	-0.344	0.386	0.094	-0.217
nSSRI	PG	774	0.625	97	188	-0.174	0.324	0.302	-0.117
	UMCU	147	0.878	21	143	-0.119	0.567	0.229	-0.1128
TriCA	PG	853	0.742	86	175	-0.117	0.322	0.257	-0.077
	UMCU	192	0.901	42	122	-0.098	0.493	0.201	-0.079
TetraCA	PG	827	0.573	45	126	-0.115	0.280	0.308	-0.115
	UMCU	44	0.886	51	49	-0.182	0.689	0.140	-0.222
MAOI	PG	62	0.613	122	212	-0.102	0.167	0.250	0.101
	UMCU	45	0.733	14.5	137	-0.057	0.390	0.187	-0.121
Other	PG	224	0.558	85	170	-0.180	0.399	0.263	-0.147
	UMCU	15	0.733	14.5	54	-0.340	0.432	0.105	-0.317

Note that 171 out of 4808 trajectories at PG and 24 at UMCU concerned trajectories where two types of antidepressants were started on the same day. At PG, 106 concerned combinations of a tetracyclic antidepressant with another type; at UMCU this concerned 12 of the 24 cases. The remainder mainly consisted of combined prescriptions of tricyclic antidepressants, SSRIs and nSSRIs, possibly discontinuation schemes started at the beginning of the admission of the patient. For the outcome measure summaries in this table, if a patient started two types of antidepressants at the same day, this data is incorporated in the two separate corresponding rows in the table. This separation into two entries is offered here purely with the purpose of keeping this table concise. In the Bayesian network analyses in this manuscript, these types of trajectories are viewed as one trajectory with a combination of antidepressant types: the Bayesian network can handle learning such interactions between variables in the model.

Supplementary material for chapter 6

The supplementary material for Chapter 6 can be found online in the publication corresponding to this chapter in Psychiatry Research as: *Yuri van der Does, Rosanne J. Turner, Miel J.H. Bartels, Karin Hagoort, Aaron Metselaar, Floortje E. Scheepers, Peter D. Grünwald, Metten Somers and Edwin van Dellen. Outcome prediction of electroconvulsive therapy for depression. Psychiatry Res. 2023 Aug;326:115328. doi: 10.1016/j.psychres.2023.115328*

Supplementary material for chapter 7

Appendix section S7.A contains detailed proofs and section S7.B additional experiments and figures.

S7.A Proofs

Proof. (of theorem 7.2.1). First consider the basic case with $E^{(m)}$ as in (7.8). As we show below, we have, with $\mathbf{E} \equiv \mathbf{E}_{P_{\theta^*}}$,

$$\mathbf{E}\left[\log E^{(m)}\right] = \mathbf{E}\left[\sum_{j=1}^{m} \log S_j\right] = \mathbf{E}\left[\sum_{j=1..m} \sum_{x\in\{a,b\}} \sum_{i=1..n_x} \log \frac{p_{\breve{\theta}_{x,k_j}|Y^{(j-1)}}(Y_{j,x,i})}{p_{\breve{\theta}_{0,k_j}|Y^{(j-1)}}(Y_{j,x,i})}\right] \geq$$

$$\mathbf{E}\left[\sum_{j=1..m} \sum_{x\in\{a,b\}} \sum_{i=1..n_x} \log \frac{p_{\breve{\theta}_{x,k_j}|Y^{(j-1)}}(Y_{j,x,i})}{p_{\tilde{\theta}_{0,k_j}}(Y_{j,x,i})}\right] \geq$$

$$\mathbf{E}\left[\sum_{\substack{j=1..m\\x\in\{a,b\}\\i=1..n_x}} \log \frac{p_{\theta^*_{x,k_j}}(Y_{j,x,i})}{p_{\tilde{\theta}_{0,k_j}}(Y_{j,x,i})} - \sum_{\substack{k=1..K\\x\in\{a,b\}}} \log\left(n_x m_k\right)\right] + O(1) =$$

$$\sum_{k=1..K} m_k \cdot D(P_{\theta^*_{a,k},\theta^*_{b,k}} \| P_{\tilde{\theta}_{0,k},\tilde{\theta}_{0,k}})) + O(\log m) \tag{A.8}$$

where we use notation $D(P_{\theta^*_a,\theta^*_b} \| P_{\theta_0,\theta_0})$ as in (7.4); and $\tilde{\theta}_{0,k}$ is defined as $\arg\min_{\theta\in[0,1]} D(P_{\theta^*_{a,k},\theta^*_{b,k}} \| P_{\theta,\theta})$ which by the same calculation as the one leading up to (7.4, is given by $\tilde{\theta}_{0,k} = (n_a/n)\theta^*_{a,k} + (n_b/n)\theta^*_{b,k}$, and m_k denotes the number of times that an instance of block k was observed in the first m blocks, and we remind the reader that $+O(\log m)$ may also indicate a negative difference of order $\log m$. (A.8) immediately implies the result, using (7.6).

The first two equalities in (A.8) are immediate. The first inequality follows because $P_{\tilde{\theta}_{0,k_j},\tilde{\theta}_{0,k_j}}$ minimizes KL divergence to $P_{\theta^*_{a,k_j},\theta^*_{b,k_j}}$ among all $\theta \in [0,1]$, within each block j. The final equality follows by independence and basic calculus. It remains to show the second inequality. This one follows because we use a prior $W(\theta_{a,k}, \theta_{b,k}$ under which θ_a and θ_b are independently beta distributed with strictly positive densities on $(0,1)$. We can then use a standard Laplace approximation of the Bayesian marginal likelihood to obtain, for each fixed $k \in \{1,\ldots,K\}$, where

the expectation $\mathbf{E}$ is over $Y'_{(1)}, \ldots, Y'_{(m')} \sim P_{\theta^*_{a,k}, \theta^*_{b,k}}$:

$$\mathbf{E}\left[-\log \prod_{j=1}^{m'} \prod_{x \in \{a,b\}} \prod_{i=1}^{n_x} p_{\breve{\theta}_{x,k} | Y^{(j-1)}}(Y_{j,x,i}) \right] =$$

$$\mathbf{E}\left[-\log \left(\int \prod_{j=1}^{m'} \prod_{x \in \{a,b\}} \prod_{i=1}^{n_x} p_{\theta_{x,k}}(Y_{j,x,i}) \right) dW(\theta_{a,k}, \theta_{b,k}) \right]$$

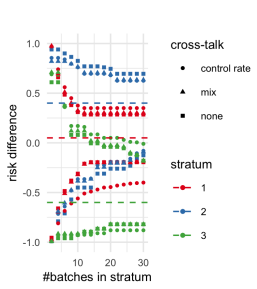
$$\leq \mathbf{E}\left[\sum_{j=1}^{m'} -\log p_{\theta^*_{a,k}, \theta^*_{b,k}}(Y_{(j)}) \right] + \log(n_a + n_b) m' + O(1).$$

Here the equality is standard telescoping of the Bayesian marginal likelihood, and the inequality is the Laplace approximation, i.e. the same calculation as the one leading up to the $(d/2) \log n$ BIC approximation of Bayesian marginal likelihood for a d-parameter exponential family; here $d = 2$ since we have two free parameters, $\theta^*_{a,k}$ and $\theta^*_{b,k}$; see [Grünwald, 2007, Chapter 8] for proof and detailed explanation).

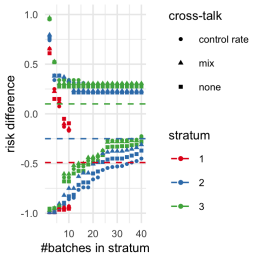
This shows the result for the basic case that $E^{(m)}$ is arrived at by multiplication, (7.8). The case for $E^{(m)}_{\text{MIX}}$ follows similarly by noting that, by construction, $E^{(m)}_{\text{MIX}} \geq E^{(m)}_{\text{NONE}}/3$, where $E^{(m)}_{\text{NONE}}$ denotes the standard e-process with multiplication and without cross-talk, for which we have already (just) shown the result.

□

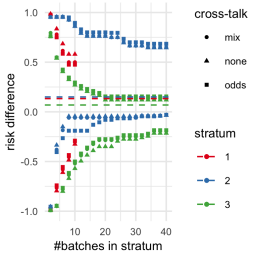
S7.B Additional experiments



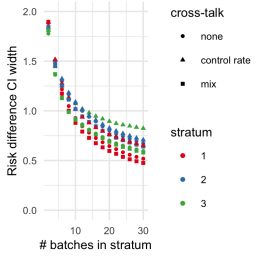
(a) all different



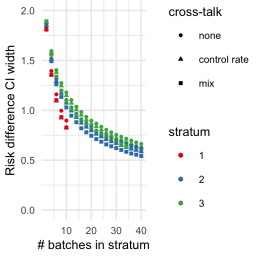
(b) same control group rate



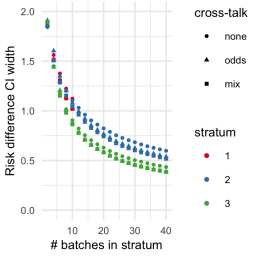
(c) same OR



(d) all different

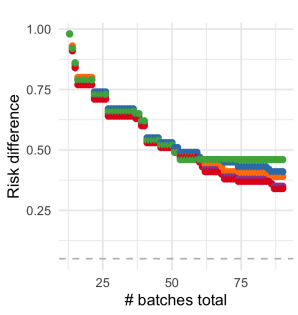


(e) same control group rate

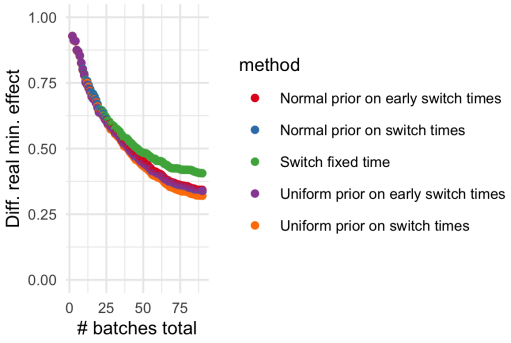


(f) same OR

Figure S7.1: 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, including mixing 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).



(a) Upper bound sequence example



(b) Average difference with true minimum

Figure S7.2: Example of a confidence sequence and average difference from upper bound to true minimal effect size value through 100 simulations, for different switch priors on j^*. 30 observations were made in each stratum, and the real differences were 0.5, 0.4 and 0.05. For the priors on early switch times, all prior mass was distributed between batch numbers 5 up to 10.α was set to 0.05.

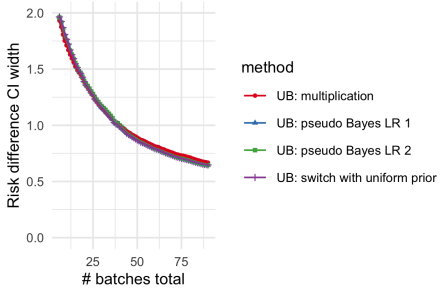
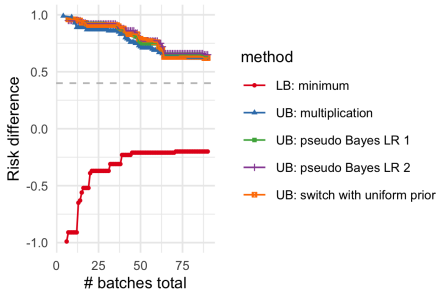
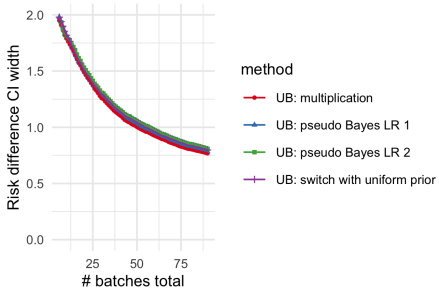


Figure S7.3: Average interval width (upper bound for the respective methods minus lower bound estimated with the minimum method) of confidence sequences for the lower- (LB) and upper (UB) bounds of the minimum effect and estimated through 100 simulations. 30 observations were made in each stratum, and the real differences were 0.5, 0.4 and 0.05. With the switch method, a uniform prior ranging from $j^* = 5$ 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.



(a) Confidence sequence example



(b) Average width

Figure S7.4: Example of confidence sequences for the lower- (LB) and upper (UB) bounds of the minimum effect, and average interval width (upper bound for the respective methods minus lower bound estimated with the minimum method). 30 observations were made in each stratum, and the real differences were 0.4, 0.4 and 0.5. 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.

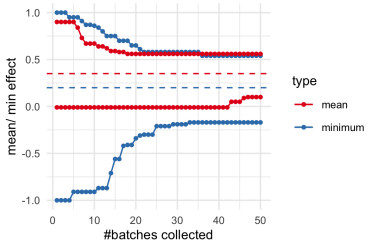


Figure S7.5: Simulated example of a confidence sequence for the mean effect across subpopulations. 25 observations were made in each stratum, and the real risk differences were 0.2 and 0.5. The confidence sequence for the mean difference is plotted alongside the confidence sequence for the minimum of the differences, estimated with pseudo-Bayesian averaging and a uniform switch prior. α was set to 0.05.