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

Turner, R.J.

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

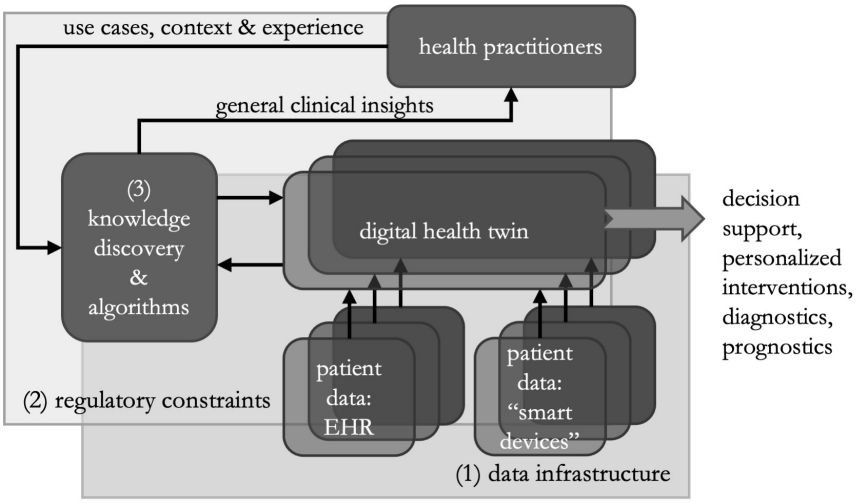

Figure 1.1: Schematic representation of a *digital health twin* framework. Adapted from The EPI Consortium [2019].

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].

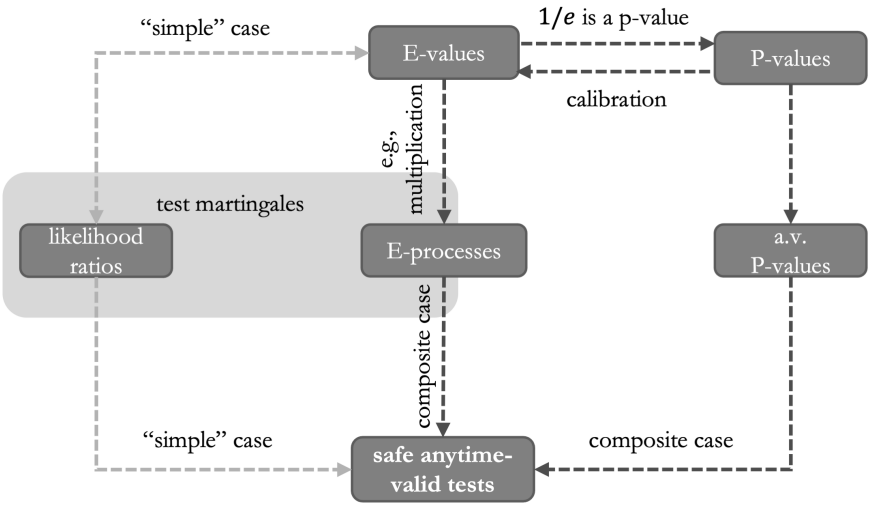

Figure 1.2: The connections between important concepts in safe anytime-valid testing. In the "simple" case where we consider a point null and alternative hypothesis, e–values and likelihood ratios are closely connected and even coincide when optimizing with respect to expected growth rate. When we consider 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.