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

Turner, R.J.

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.