The non-existent average individual

Blaauw, Frank

Chapter 10

Discussion: Personalization in Psychopathology Research

We set out to explore various methods to collect and analyze data on different levels. We collected data at the level of the individual and of the group by designing and implementing online platforms to measure both psychological and physiological variables. Further we investigated and developed methods to perform (semi-)automated analysis and, more importantly, used them for the creation of personalized feedback.

The terms 'personalized medicine' and 'precision medicine' have gained significant interest over the past years, and are used whether appropriate or not. We explored different methods to enable a *truly* personalized medicine, in which 'truly' signifies the use of personalized methods of analysis. To foster the use of such individual methods, the collection of large amounts of data on the level of the individual is vital. We created several applications that automate self-assessment protocols, data gathering, complex statistical procedures, and feedback.

10.1 HowNutsAreTheDutch and Leefplezier

Part I provided insight into and a description of our two e-mental health platforms: HowNutsAreTheDutch (HND) and Leefplezier. One of the strengths of the HND and Leefplezier projects is that we involve the general public in mental health research (viz., crowdsourcing). The HND website and Leefplezier App provide participants with the opportunity to gain insight into their mental health, whether or not by comparing their scores to scores of other participants. Moreover, the combination of (i) measuring mental symptoms and strengths and (ii) our longitudinal time-intensive design may allow for a more detailed and micro-level view of the dynamics of mental health and ill-health than most studies are able to provide (Keyes, 2007; Lamiell, 1998; Lee Duckworth et al., 2005; Molenaar & Campbell, 2009; Piantadosi et al., 1988). This broad range of assessed mental strengths as well as the use of automated feedback set these studies apart from previous studies such as Netherlands Mental Health Survey and Incidence Study (NEMESIS) and Lifelines.

Our use of Web applications and mobile applications enabled all inhabitants of the Netherlands (and essentially globally) to access the diary studies, but only small groups of people were actively informed about the studies. In HND, an active approach seems to have been crucial in the recruitment, given that the number of subscriptions increased noticeably after presentations and other advertising activities by the research team (see Figure 5.4 on page 67 for a graph showing the acquisition of new participants over time). A limitation of these studies is that only a fraction of the total Dutch population, which is approximately seventeen million (Centraal Bureau voor de Statistiek, 2017), participated. This indicates that the potential of diary studies for national health promotion through self-assessment, in the format that we applied, is limited. Nonetheless, the crowdsourcing methodology resulted in the collection of valuable data sets that allow for group-level and idiographic analyses that can shed light on etiological processes and may contribute to the development of empirical-based health promotion solutions. Moreover, these data sets can provide insight into novel and personal factor-context interactions, and as such, could help to see mental health symptoms as more than symptoms alone, and view mental symptoms as a person's very strengths.

Nearly half of the population will meet current Diagnostic and Statistical Manual of Mental Disorders (DSM) criteria for a mental disorder at some time in their life, but this does not mean that they will all need treatment (Kessler et al., 2005). Health is best defined by the person (rather than by the doctor), according to his or her functional needs, which can be the meaning of 'personalized medicine' (Perkins, 2001; The Lancet, 2009). Clinicians, then, are partners in delivering those needs. Diary studies can play a role in this process of empowerment, as they enable for personalized models that can shed light on etiology and personal dynamics, as well as personalized solutions, and merit the perspective of health as people's ability to adapt to their environments and to self-manage (Huber et al., 2011; Lee Duckworth et al., 2005; Solomon, 2012). Goals and criteria for 'treatment success' often differ substantially between clinicians and their patients, which may explain part of the high drop-out rates ($25\,\%$ to $60\,\%$) for most psychiatric interventions (Perkins, 2001; Tehrani, Krussel, Borg, & Munk-Jorgensen, 1996).

Mental symptoms may be seen as more than 'defects to be corrected', as individuals' differences may be their very strengths. For example, individuals with autism may be great scientists, mathematicians, or software testers (Mottron, 2011; Solomon, 2012), while anxious individuals may be rather creative, sensitive, and agreeable, thus perfect employees for social job tasks (George & Jones, 2008; Stossel, 2014). Antagonistic, mistrustful, uncooperative and rude people, in contrast, may be excellent drill sergeants or bill collectors (George & Jones, 2008). Furthermore, genes that increase vulnerability for schizophrenia and bipolar disorder also underly cre-

ativity (Power et al., 2015). Everyone has both strengths and weaknesses, but which is which is a function of the context in which we live and grow (Darwin, 1859). A reconsideration of diversity and focus on individual strengths and resources that may compensate for — or buffer against — the expression of mental symptoms, may help people to preserve acceptable levels of mental well-being despite the presence of psychopathology (Harkness & Luther, 2001; Sheldon, Kashdan, & Steger, 2011; Solomon, 2012). This would fit in with a concept of mental health as a hybrid of absence of mental illness and the presence of well-being and mental resources (Keyes, 2007; Lee Duckworth et al., 2005).

10.2 Automated Impulse Response Analysis

In Chapter 6, we provided the description and implementation of Automated Impulse Response Analysis (AIRA). AIRA provides tailored advice about mental health and well-being based on data collected from diary studies or sensor data. AIRA allows participants to interactively inspect relations between their psychological factors, and how they interact over time by means of impulse response function (IRF). We performed several experiments to demonstrate the performance of AIRA. In our experiments, presented in Section 6.5.2, we showed that the results from the automated analysis of AIRA are comparable to earlier manual work. In Section 6.6 we presented an application of AIRA in research, demonstrating its potential in psychopathology research projects.

In mental health research, few case studies (generally consisting of one to five participants) have previously applied IRF analysis to diary data (e.g., Hoenders et al., 2011; Rosmalen et al., 2012). AIRA includes several methods to perform an IRF analysis similar to the one used previously, but instead of using manual analysis, AIRA applies an automated technique. To the best of our knowledge, AIRA is the first approach for automatically generating IRF-based advice.

The analysis of IRF and vector autoregression (VAR) models leaves room for ample discussion. Firstly, a controversial point is whether or not to include contemporaneous effects in IRF analysis (Brandt & Williams, 2007; Sims, 1980). Because the directionality of contemporaneous effects has no clear foundation in reality and therefore lacks conceptual justification, the use of contemporaneous effects is controversial. Although some of these directions can be determined using theoretical domain knowledge, this might not hold for the individual. As such, AIRA currently does not base its advice on contemporaneous effects.

Secondly, the use of IRF and VAR models in the context of psychopathology research can be a point of discussion. Both VAR models and IRF analysis are not orig-

inally developed to be used with ecological momentary assessment (EMA) data. In AIRA we assume that a VAR model fit to EMA data is representative for a period that exceeds the EMA study itself; that it is representative for future events, and essentially that it estimates the true data generating distribution using the parametric VAR technique. Treating the VAR model as a linear and time invariant system enables us to perform such analysis. We based this assumption on several previous studies also applying VAR analysis to psychological research data (e.g., Hoenders et al., 2011; Rosmalen et al., 2012; van Gils et al., 2014). We should note that, as for many statistical techniques, for AIRA to work a key requirement is that the model comprises relevant variables. When the VAR model merely consists of variables not having meaningful connections, not having meaningful variance, or variables that cannot be influenced AIRA will not function properly.

Thirdly, when a VAR model has more than one lag, it might happen that the sign of a VAR coefficient (i.e., the direction of the effect) is not equal for all lags. For example, a variable may have a negative effect on another variable in one lag but a positive effect in a different lag. In this scenario, it is difficult to determine whether the effect of the variable should be considered positive or negative when basing this decision solely on the VAR model. In AIRA, we circumvent this issue by using IRF, and moreover, by summing the cumulative IRF values to obtain a net effect value. This net effect value considers the entire horizon for determining an effect to be either a net gain or a net loss. That is, if the positive effects outweigh the negative effects, the relation is considered positive (and vice versa for negative effects). This approach is novel and specific to AIRA, and can be considered essential when inspecting VAR models (Brandt & Williams, 2007).

Fourthly, a point of discussion is the creation of the VAR models. Although the creation of VAR models is deliberately left out of this dissertation (see Emerencia, 2014, for a dissertation that partly discusses this topic), there are a few important points one needs to take into account. The estimation of a VAR model entails various caveats, such as the selection of the correct lag length and the equidistant measuring of the EMA data set. When fitting a VAR model to data, these are properties to take into account. For AIRA, we relied on previous work allowing the creation of VAR models to be performed automatically (Emerencia et al., 2016).

The performance of AIRA with respect to the calculation time is sufficient for practical use. Basic experiments show that the calculation of an advice using a VAR model consisting of six variables and a horizon of twenty steps happens on a modern laptop in less than a second. Furthermore, our time complexity analysis implies that this performance scales well within acceptable boundaries for practical use. AIRA is implemented in both the R-language and in JavaScript, and can therefore run in any browser on any modern operating system (mobile or desktop). It should

be noted that the running time for computing advice using bootstrapped IRF analysis increases linearly with the (constant) number of bootstrap iterations.

Objectively establishing the correctness of any algorithm that provides suggestions or advice is a complex issue. AIRA forms no exception. While the implemented formulas may be mathematically correct, we have not yet evaluated the practical utility of the advice generated by AIRA in terms of clinical accuracy, understandability, and helpfulness for the individual. Hence, whether the advice actually contributes to the well-being of an individual remains to be shown in future research. Moreover, in the evaluations we performed on AIRA, we focused on data sets collected using the EMA methodology. To further investigate the accuracy and effectiveness of AIRA, the use of simulated data sets would be an interesting option.

In our real-world application of AIRA in Section 6.6, we aimed to answer questions about the relation between anhedonia and major depressive disorder (MDD). The study we performed had several notable strengths. First, our EMA design ensured that emotional dynamics we studied were measured in the participants' daily lives and their natural environments, and thus ecologically valid. Second, we distinguished between different dimensions of positive and negative affect, thereby shedding light on the relevant differences in emotional dynamics that would have been overlooked in studies excluding this dimensionality.

The findings of this application of AIRA should also be considered in light of several limitations. First, the presence of anhedonia was indicated by endorsement of a single questionnaire item on loss of interest, but items related to the other hallmark of anhedonia, loss of pleasure, were not available in the data set used and therefore not included in the analysis. Second, our time frame of six hours was relatively long, which may explain why the associations under study were only present in a small subset of the sample. Third, given that the HND sample consisted mostly of highly educated women, our results may not generalize to other populations. Finally, stress experience was measured indirectly by assessing level of distress, rather than the direct impact of stressors. Thus, while the different role of positive affect (PA) in the anhedonic versus non-anhedonic group stands out more clearly and reliably, it remains difficult to unravel the difference in associations between negative affect (NA) and stress experience between the two groups. Our results suggest different emotional dynamics may underlie depressive symptomatology. Anhedonic depression may be characterized by individuals that exhibit lowered favorable impact of PA high arousal on affect and behavior, and heightened reactivity to NA. On the other hand, non-anhedonic depression may be characterized by individuals that experience heightened stress reactivity. The large heterogeneity in the extent to which these pathways were present in individuals advocates a personalized approach to gain insight in how depressive symptomatology is maintained in daily

life. Future studies may relate different pathways of emotional dynamics to future course of depression.

In a broader perspective, AIRA could be a useful asset in the current field of mental health research and clinical practice. AIRA could be used on large scale platforms as a decision support tool, and as a means to give automatic personalized advice on mental health and well-being, for instance in national scale platforms such as HND and Leefplezier.

10.3 Machine Learning for a More Precise Medicine

Besides the time series approach, we explored various machine learning techniques to provide a more personalized and precise medicine. We approached personalizing medicine from two perspectives: (i) the interindividual perspective, in which we used several cross-sectional variables to create relatively high dimensional and data adaptive estimators, and (ii) the intraindividual perspective, in which we used EMA data to determine both population and personal statistical parameters.

10.3.1 Interindividual Perspective

The machine learning approach described in Chapter 7 was designed to create several classifiers for early prediction and classification of above clinical threshold depression. Our work builds on the large body of knowledge already available related to machine learning and the application thereof. Previous work already showed the feasibility of using machine learning in the psychopathology research (e.g., Chekroud et al., 2016; Perlis, 2013), and the present work adds the prediction of above clinical threshold levels of depression. We used a set of nine machine learning algorithms and showed their usefulness for doing predictions on the data provided. Determining the predictive qualities of a large set of machine learning methods with highly adaptive hyperparameter optimization in the field of psychology has, to the best of our knowledge, never been done before.

The developed classification approach can assist clinicians in their decision making process. We showed that our machine learning based classifiers achieved moderate to high performance levels (the top-three classifiers had an average performance ranging from 0.713 to 0.742). However, there are several important considerations to take into account before applying these classifiers to a broader, clinical context. Firstly, the use of machine learning is not standard practice in a clinical setting. As such, a cultural change might be needed before machine learning can become part of clinicians' toolboxes. Secondly, the fact that these classifiers mostly

work as 'black-box' systems, of which the inner workings are only partly understood, could hinder their acceptance by experts (Sittig, Krall, Dykstra, Russell, & Chin, 2006). Thirdly, although the performance of the classifiers presented might be relatively high from a technical perspective, comparing them with the accuracy of a manual approach is still to be explored in future work. Such a comparison could help give an estimate of the clinical performance of the used machine learning approach, and could help determine whether the achieved performance is high enough to be usable for medical decision making.

We determined the performance of the classifiers based on the five performance measures and the average thereof, as we chose not to rely on a single outcome measure. It is clear that due to the skewed distribution of the outcome variable, the use of the accuracy measure and F1-measure are often not informative, as there is little to no variation in scores and they might give a distorted view on the performance of the classifier (e.g., the constant dummy classifier had both a high F1-score and accuracy score). The area under the curve (AUC), Kappa score, and geometric mean measures showed more variation and were useful in selecting the best performing classifier. The overview provided by the confusion matrices provides a clear and accurate overview of the performance of a classifier. Note that despite the fact that the Random Forest algorithm performed best on this particular data set, this does not necessary generalize to other data sets. Every machine learning algorithm has its own benefits and pitfalls, causing some algorithms to perform well on a certain data set while others do not (as is illustrated by the variability in the test scores of our algorithms). These performance differences can be attributed to the diverse internal methods applied by each machine learning algorithm, and is oftentimes a consideration between various variables; for example, some methods assume a linear relation in the data and are fast to train, while other algorithms can deal with non-linear relations, but require more training time, or have more hyperparameters to optimize. A method in which a large number of flexible machine learning classifiers are trained to find the single optimal algorithm (or combination of algorithms; ensemble learning) is generally the best, data-adaptive way to go forward (e.g., Dietterich, 2000; Gashler et al., 2008; Lemke et al., 2015; van der Laan et al., 2007).

By applying a data-adaptive approach to statistical modeling (viz., machine learning), we overcome strong parametric assumptions on the statistical model, as we use the data to decide which model could best be used for making predictions. If we were to rely strictly on a parametric model (e.g., a logistic regression), we make the (relatively strong) assumption that the relationship between the input and output can be expressed by a small, finite number of parameters in a linear form. As we do not have knowledge whether or not this is the case, we should not impose such strong restrictions on our statistical model (Petersen & van der Laan, 2014). By

applying a large number of different machine learning classifiers (both of parametric and non-parametric nature), cross-validation (CV), and out-of-sample validation, we can draw relatively strong conclusions about how well our estimators would perform on new, unseen data.

We applied the same data-adaptive approach to select the best tuning parameters for our machine learning classifiers. Although extensive hyperparameter selection can greatly improve classifier performance, it also increases the training time of our final machine learning algorithm. The fact that each algorithm needs to be retrained for each new combination of hyperparameters and needs to be cross-validated is time consuming. For example, running a single instance of our application with 100 iterations of random search on a standard $3.5\,\mathrm{GHz}$ Intel core i7 processor took approximately 1.5 hours to run. We applied a random search method with several predefined parameter distributions to explore a different hyperparameter combinations. Because this approach is purely random, prior knowledge of accurately predicting configurations are not taken into consideration when new hyperparameters are drawn. Although previous studies have shown that random search performs well (e.g., Bergstra & Bengio, 2012), there exist different approaches that apply more sophisticated ways for finding well performing configurations. Such approaches could help make the hyperparameter selection procedure more efficient and more effective. An example of a more sophisticated approach could be found in the area of genetic algorithms. In genetic algorithms, a notion of natural selection is applied that causes the optimization procedure to explore new configurations along the path of previously well-performing configurations (e.g., Forrest, 1993). Genetic algorithms have been shown to outperform traditional search methods in some cases and seem to be a reasonable alternative to the current, random approach (P.-W. Chen, Wang, & Hahn-Ming Lee, 2004).

The applied machine learning approach is only as good as the machine learning algorithms it includes. As such, the logical next step for the present work is to increase the number of selected machine learning algorithms. We used an initial set of nine machine learning algorithms, all implemented using the scikit-learn python package. However, there is a plethora of machine learning algorithms available that have not yet been used and could easily be included in our application (e.g., multivariate adaptive regression splines [MARS; Friedman, 1991], Deletion / Substitution / Addition algorithm [D/S/A; Sinisi & van der Laan, 2004], or Bayesian methods). Furthermore, in the recent years, there has been an increasing interest in the field of *deep-learning* algorithms (e.g., as can be implemented using the TensorFlow or Theano libraries; Abadi et al., 2016; The Theano Development Team, 2016) and optimized boosting algorithms (e.g., XGBoost; T. Chen & Guestrin, 2016). Incorporating more implementations could further improve our current implementation both in

speed and accuracy. Besides the use of different machine learning algorithms, one could also apply different feature selection methods. The used feature selection procedure was based on previous psychopathology research (e.g., Chekroud et al., 2016), but is not the only option for performing feature selection. For example, one could fit univariate logistic regressions for each of the predictors and pick the ones that predict best / are significant, or rely on regularized machine learning methods (Chandrashekar & Sahin, 2014). These different methods could affect the initial features selected and could therefore influence the performance of the machine learning procedure.

10.3.2 Intraindividual Perspective

In Chapter 8, we described the theoretical foundation, initial implementation, and preliminary results of the Online SuperLearner (OSL). We showed that the OSL is a novel technique that can be used to perform causal inference based on time series data. In combination with the online one-step estimator it aims to become an essential tool for analyzing sequentially dependent data. The results of the OSL and online one-step estimator (OOS) combination, however, leave room for discussion. As shown in Table 8.2, the prediction of the OSL seems to converge reasonably well to the truth, and generally yields acceptable approximations. However, the one-step estimator's performance is not yet optimal. In all cases it actually worsens the initial estimation. We speculate that both the fluctuations in OSL performance and the OOS performance are caused by an erroneous implementation in the corresponding R implementations, or that a larger number of approximation iterations is needed for the OSL and OOS to converge well. More research is necessary to determine the root of these differences.

Any ensemble machine learning approach has a computational intensity that significantly increases with the number of candidate learners and the OSL forms no exception. Training a large number of candidate estimators on large data sets is time consuming. The time-complexity for K algorithms each having d hyperparameters is at least $\mathcal{O}(K^d)$ when trying all combinations of hyperparameters (the so called *grid-search*, see Chapter 7 for more information). The online learning approach circumvents retraining all learners for every new observation or set of observations and therefore requires less computing resources per iteration, eventually reducing the run time of the algorithm as data accumulates (Bottou & Le Cun, 2005). Although we have not yet performed analysis on the run times, the fact that with online learning only a small update needs to be performed instead of retraining a full algorithm makes this a reasonable assumption.

The OSL depends on the availability of online machine learning algorithms for

it to perform in an online fashion. The current implementation has a system to fall-back to batch learning whenever a learner does not support online training. With this fall-back, the OSL scans if any of the algorithms is not online. If this is the case, it will keep track of all previously seen data. The algorithms flagged as being online will still be trained in an online fashion. This fall-back should only be used when working with relatively small data sets and fast algorithms, as it defeats the purpose of the online learning procedure. For true online performance, one should only include online algorithms.

The online learning procedure is a vital component of the OSL, in the sense that it is needed to calculate a reliable estimate of the CV risk. For calculating and updating the CV risk, we rely (for each iteration of online training) on a single subsequent block as validation. One could also choose to increase this number of blocks, and use for example the m subsequent blocks for calculating the CV risk. Doing so might result in a more reliable estimate of the CV risk.

Our current implementation of the OSL is available as a free and open-source R-package. The open-source availability of this package could aid the adoption of the OSL methodology in the community, and might incline other researchers to adopt and improve upon this implementation. We strove to make our implementation of OSL extensible, which is reflected in the ease with which one can add new learners, summary measures, and data sources.

The performance of the OSL could be improved by introducing some changes in the underlying density estimation procedure. We apply a two-step procedure in which we first find which bin a continuous outcome is in (the discretization step), and when this bin is selected, we sample a value from a uniform distribution within that bin. This sampling step can be improved by fitting a bounded parametric density within each bin and sampling from that distribution, for example a β-distribution or triangle distribution.

OSL is based on several important assumptions. One of them is the assumption of stationarity. That is, the assumption that the data generating distributions on the level of the individual are ergodic (i.e., that the data generating distributions are shared over time). In order to alleviate this assumption, we have to treat each block of data independently, and assume a data generating distribution is shared at the group level. We argue that this is another, even stronger assumption to make. As such, we currently chose to relax both assumptions and use both data related to the group as data related to the individual in isolation for training the best estimator.

In the case study showing the applicability of OSL and OOS on the HND data set, we used data sets that had been imputed beforehand and as such, did not contain any missing values. The method we used for this is a well validated approach, but leaves room for improvement. As we are inherently treating the current estimation

problem as a missing data problem (viz., using the notion of counterfactual worlds), methods exist to use the same techniques for actual missing data (van der Laan & Rose, 2011). This method of imputation is currently not supported within our implementation of the OSL, and could be a direction for future work.

By applying OSL to the HND data set, we showed its usefulness in time series analysis using psychological data. We can use it to devise a more personalized way of analyzing data, whilst still taking advantage of data retrieved on the group level. The analysis we performed serves as an initial example, and more complex questions and analysis should be designed to further explore the possibilities of the OSL. The intention is that the open-source availability of the package serves as a catalyst for other researchers to use this package to explore their own questions.

10.4 Ecological Momentary Assessments and Wearables

Physiqual is a novel approach for processing sensor data for its use in EMA studies (Chapter 9). With Physiqual, we enable the use of sensor data from commercially available wearable devices in EMA mental health research by interfacing with service providers to export data in applicable formats. Physiqual is a way to manage data and it fills a niche with the rising interest in Quantified Self (QS) fueled by the increasing popularity of wearable devices. Our case study showed how Physiqual can be useful in adding physiological data to EMA data, potentially enabling new insights in psychophysiological research at the individual level. Currently Physiqual supports two service providers, but the platform can be easily extended to interact with other service providers in the future.

As with every new development, Physiqual has its limitations. While sensors sample data at a high frequency, EMA data is collected over longer intervals. To compensate for this discrepancy, heart rate data is downsampled, adhering to the low frequency of the EMA data. By downsampling, the most frequently occurring heart rate is presented, which we consider to be most in line with an EMA study. However, due to this downsampling we lose information about short but possibly intense shifts in heart rate. These intense changes could conceal short physiological (stressful or pleasant) events which might have a considerable influence on mental phenomena (Myrtek, 2004).

In addition, the downsampled data, as extracted from a sensor, is a summary of the measurements within a predefined period of time. This summary can be based on a varying number of measurements. Hence, the reliability of the exported measurements can differ. Currently, the format in which the data is exported does not accommodate a representation for the notion of reliability.

Practical limitations of Physiqual include the type of access allowed and the data exported by the service providers. For example, Fitbit permits intraday access to measurements only on a per-project basis, and the number of requests allowed has an hourly limit. Physiqual can only process data of a wearable sensor when this data is accessible. Some wearable platforms currently have limited options for data extraction by third party applications such as Physiqual. One of the most popular smartwatches at the time of writing is the Apple Watch (Statista, 2016). Although the Apple Watch provides several sensors useful for EMA research, Physiqual currently is not able to support it. At present, the Apple Watch does not provide an application programming interface (API) accessible via the Internet, nor does the *Apple HealthKit* platform. These platforms currently only provide a mobile iOS — the mobile operating system by Apple Inc. — API to retrieve data from the watch. No method for exposing this data directly to Physiqual is therefore available. To support the Apple Watch in an external platform like Physiqual, a third party mobile application must be developed which is capable of uploading the Apple Watch data to either one of the existing supported service providers or to a new platform.