

University of Groningen

Computing a Second Opinion

Emerencia, Ando

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2014

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Emerencia, A. (2014). Computing a Second Opinion: Automated Reasoning and Statistical Inference applied to Medical Data. [S.l.]: s.n.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

**Computing a Second Opinion:
Automated Reasoning and Statistical
Inference applied to Medical Data**

Ando Emerencia

Supported by the Netherlands Organization for Health Research and Development (ZonMW)
under contract number 300.020.011





**rijksuniversiteit
 groningen**

**Computing a Second Opinion:
Automated Reasoning and Statistical
Inference applied to Medical Data**

Proefschrift

ter verkrijging van de graad van doctor aan de
Rijksuniversiteit Groningen
op gezag van de
rector magnificus Prof. dr. E. Sterken
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
vrijdag 27 juni 2014
om 11.00 uur

door

Armando Celino Emerencia

geboren op 6 november 1983
te Groningen, Nederland

Promotores: Prof. dr. M. Aiello
Prof. dr. N. Petkov
Prof. dr. P. de Jonge

Beoordelingscommissie: Prof. dr. E. Wit
Prof. dr. C. Combi
Prof. dr. C.N. Schizas

ISBN: 978-90-367-7090-3 (book)
ISBN: 978-90-367-7089-7 (e-book)

Contents

| | |
|---|-----------|
| Acknowledgments | ix |
| 1 Introduction | 1 |
| 1.1 Context | 2 |
| 1.1.1 Schizophrenia | 2 |
| 1.1.2 Current schizophrenia treatment | 3 |
| 1.1.3 Problems | 3 |
| 1.2 Scope of this thesis | 5 |
| 1.2.1 Wegweis | 7 |
| 1.2.2 Autovar | 8 |
| 1.3 Thesis organization | 9 |
| 2 Artificial Intelligence in Medicine: a brief overview | 11 |
| 2.1 Symbolic approach | 11 |
| 2.2 Connectionist approach | 12 |
| 2.3 Measuring performance | 13 |
| 2.4 In the 2000s | 14 |
| 2.4.1 Case-based reasoning | 15 |
| 2.4.2 Temporal abstraction, representation, and reasoning | 15 |
| 2.4.3 Data mining and data analysis | 16 |
| 2.5 State of the art | 17 |
| 2.5.1 Standards and interoperability | 17 |
| 2.5.2 Ontology-based applications | 18 |
| 2.5.3 Ambient intelligence | 19 |
| 2.5.4 Patient-centered applications | 19 |
| 2.6 Schizophrenia and other psychotic illnesses | 20 |

| | | |
|----------|---|-----------|
| 2.6.1 | Wegweis | 21 |
| 3 | E-health self-management for psychotic disorders | 23 |
| 3.1 | Introduction | 23 |
| 3.2 | Methods | 24 |
| 3.2.1 | Search strategy | 24 |
| 3.2.2 | Definitions | 25 |
| 3.2.3 | Study selection criteria | 25 |
| 3.2.4 | Data extraction | 26 |
| 3.2.5 | Quality assessment | 26 |
| 3.2.6 | Statistical analysis | 26 |
| 3.3 | Results | 28 |
| 3.3.1 | E-mental health self-management interventions and outcome | 28 |
| 3.3.2 | Cost-effectiveness and user involvement | 34 |
| 3.4 | Discussion | 34 |
| 3.4.1 | Types of e-mental health self-management interventions . . . | 35 |
| 3.4.2 | Evidence base for clinical outcome and cost-effectiveness . . . | 36 |
| 3.4.3 | Orientation of self-management interventions | 37 |
| 3.4.4 | Limitations | 38 |
| 4 | A system for generating personalized advice | 39 |
| 4.1 | Wegweis system design | 40 |
| 4.2 | Wegweis user interface | 41 |
| 4.3 | Problem ontology | 42 |
| 4.4 | Selecting and ranking advice | 47 |
| 4.4.1 | An algorithmic overview | 47 |
| 4.4.2 | Calculating the activation strengths | 48 |
| 4.4.3 | Calculating the advice unit priorities | 50 |
| 4.4.4 | An example run | 52 |
| 4.5 | Implementation | 54 |
| 4.6 | Discussion | 57 |
| 5 | Evaluation of Wegweis | 59 |
| 5.1 | Usability Evaluation | 59 |
| 5.1.1 | Methods | 62 |
| 5.1.2 | Results | 68 |
| 5.1.3 | Discussion | 69 |
| 5.2 | Evaluation involving patients and clinicians | 73 |
| 5.2.1 | Evaluation measurements | 74 |
| 5.2.2 | Clinicians and problem severities | 75 |

Contents

| | | |
|----------|--|------------|
| 5.2.3 | Patients and advice relevance | 79 |
| 5.2.4 | Discussion | 82 |
| 6 | Automating vector autoregression | 85 |
| 6.1 | Vector autoregression | 87 |
| 6.2 | Autovar overview | 88 |
| 6.3 | Model configurations | 91 |
| 6.3.1 | Trend variable inclusion | 91 |
| 6.3.2 | Dummy variables for weekdays | 92 |
| 6.3.3 | The lag order | 93 |
| 6.3.4 | Log-transforming the data | 93 |
| 6.4 | Model validity | 94 |
| 6.4.1 | Stability test | 95 |
| 6.4.2 | Residual diagnostic tests | 95 |
| 6.5 | Handling invalid models | 95 |
| 6.5.1 | When the model is not stable | 95 |
| 6.5.2 | When the model fails residual diagnostic tests | 95 |
| 6.6 | Constraining valid models | 97 |
| 6.7 | Algorithm for model selection | 98 |
| 6.8 | Discussion | 102 |
| 7 | Evaluation of Autovar | 103 |
| 7.1 | Implementation | 103 |
| 7.1.1 | Imported, modified, or implemented functions | 103 |
| 7.1.2 | Input data and parameters | 104 |
| 7.1.3 | Exogenous variables | 105 |
| 7.1.4 | Web application output | 108 |
| 7.2 | Evaluation | 113 |
| 7.2.1 | Comparison with manual analysis | 113 |
| 7.2.2 | Performance | 117 |
| 7.3 | Related Work | 120 |
| 7.3.1 | PcGive | 121 |
| 7.3.2 | Comparison | 122 |
| 7.4 | Discussion | 123 |
| 8 | Conclusion | 125 |
| 8.1 | Summary | 125 |
| 8.2 | Future work and open issues | 126 |
| 8.3 | Outlook | 127 |

Contents

| | |
|---------------------|------------|
| Bibliography | 129 |
| Samenvatting | 151 |

Acknowledgments

Interdisciplinary research involves, by definition, two or more academic or scientific disciplines. In the past four years, I have worked on the boundary between psychology and computer science, between artificial intelligence and the knowledge of experts, and between statistics and automation. To work on a boundary, you need support from both sides. As such, I have had the fortunate experience to work together with people from various disciplines.

First and foremost, I would like to thank my promoters for their guidance. Marco Aiello has been essential in enabling me to fulfill the different aspects of my PhD. He knows the balance between coaching and supporting, between setting goals and delivering results, and between work and other activities in life. Where I take myself very seriously, Marco has at times insisted that I take time off and celebrate success. Thanks also to Nicolai Petkov, without whom I would not have applied for this position. Thank you for the internship during my M.Sc. studies, for recommending me for this position, and for supporting me during my PhD. In the second half of my PhD and, in particular, the time after my PhD, Peter de Jonge has provided ways for me to continue my research, which are much appreciated.

A special thanks goes out to Lian van der Krieke, with whom I worked in close cooperation for the majority of my PhD. Even though our backgrounds are perhaps diametrically opposed, I believe that we complemented each other's expertise and functioned well as a team. Regarding the projects that we did together, I do not think either of us would have been able to do them alone. Thank you for the pleasant collaboration over the years.

I would also like to mention the co-authors with whom I have published papers. Elske Bos has laid the theoretical foundation on which AutoVar is built. I would also like to thank Judith Rosmalen, Nynke Boonstra, and Harriëtte Riese for our joint publications. There are more people without whom this project would not have been possible. I express my thanks to Lex Wunderink, Cees Slooff, Frank van Es, Richard Bruggeman, and Durk Wiersma, for creating room for us to test and deploy our systems with a real user base.

I look back on memorable experiences with my colleagues throughout the years. I have always looked up to Andrea as an example of how to do your PhD. Even though Andrea is not native to the Netherlands, he has shown me how to get things done around here. Many thanks also to Frank, for listening to my rants and for providing great feedback and keen insights. I thank my former and current colleagues for the pleasant time we have shared. Pavel, Eirini, Elie, and Mahir were here early on. More recently, I have worked alongside Ehsan, Viktoriya, Heerko, Ilche, Tuan, Faris, Saleem, Fatimah, Doina, and Alexander. I would also like to thank George, Giannis, Kerstin, Petra, Aree, and everyone currently in the Intelligent Systems group.

For their technical support and guidance, I would like to thank the professionals from RoQua for years of pleasant cooperation. I thank Sjoerd Sytema and Erwin Veermans for hosting Wegweis and for their long-term technical outlook. Many thanks go out to Marten, Samuel, and Herman for supporting Wegweis and for creating the functionality that made it possible. Of course, I want to mention Jorn for his technical contributions to Wegweis. I also thank the administrative staff for looking out for me all this time. Thank you Desiree, Esmee, and Ineke.

Finally, I would like to thank my family for their support, for teaching me the importance of science and education, and for giving me the opportunity and the encouragement to strive for the highest attainable goals. Thank you.

Ando Emerencia
Groningen
June 10, 2014

Chapter 1

Introduction

Healthcare is a data-intensive process. At any time, in any hospital, patients are monitored, illnesses are diagnosed, medication is prescribed, assessments are performed, and questionnaires are filled out. Currently, most of this data is stored electronically, and we refer to such data as *electronic medical data*. Electronic medical data is bound by implicit and explicit properties. For example, some data lives temporarily while other data is stored persistently, and some data has geographical restrictions while other data should be readable only by a specific set of people.

In modern hospitals, one of the most comprehensive forms of electronic medical data is the *electronic medical record* (EMR, also *electronic health record*), which encompasses any data associated with a patient that should be stored persistently (Jensen et al. 2012). For example, EMRs store identifying information, medical histories, demographics, medication, and, if applicable, treatment plans, assessment results, and electronic patient diary data.

Most forms of electronic medical data are never analyzed outside of their original purpose (Miller and Sim 2004, Häyrinen et al. 2008). Reasons for this isolation include issues concerning security, privacy, and doctor-patient confidentiality (Safran et al. 2007). Moreover, certain forms of electronic medical data are compatible only with local infrastructure (Kohane et al. 1996). Different care organizations use different hospital information systems, sometimes with local adjustments, which may use incompatible data formats. The interoperability of medical data has become an issue in recent years (Walker et al. 2005, Brailer 2005).

We consider electronic medical data not as a collection of isolated patient histories but as sets of interconnected nodes in a network governed by an underlying ontology. By applying techniques from automated reasoning and statistical inference, we can gain knowledge about diagnosis, prognosis, decision support, cause-and-effect relations between symptoms, side effects of medication, and advice for patients.

In this thesis, we seek answers to the questions of which aspects of care that involve transferring knowledge can be automated and how this automation can be performed. The focus is on gaining knowledge from electronic medical records using ontological reasoning and statistical inference. We believe that automated anal-

ysis of medical data will play an important role in the future of healthcare for two reasons. First, automation, in this context, solves many of the issues related to privacy and confidentiality that would occur in manual analyses of medical data. Thus, on the assumption that the results of the data analysis are either fully anonymized or otherwise available only to the patients and their respective clinicians, effective use of medical data outside of its original purpose becomes feasible. Second, automation scales at the mere costs of computation. Once we develop automated ways to deduct knowledge from data, rapid dissemination and widespread application of these concepts incurs relatively low operational costs. However, the required organizational effort would be substantial.

1.1 Context

For most patients, interaction with a hospital or care facility is typically brief, with a paucity of data being generated. The treatment protocols for people suffering from chronic, long-term illnesses however, tend to generate more data, and these patients are also more likely to benefit from improved data analysis techniques. Our research was performed to improve care for patients suffering from psychotic illnesses such as schizophrenia. Schizophrenia patients partake in yearly, extensive assessments and may record patient diary data or have other interaction with care facilities that is stored in their electronic medical records. Thus, there is a rich quantity of data that can be analyzed to increase our knowledge about the disease and its symptoms. Many aspects of schizophrenia, such as the cause of psychosis or the effects and interaction between different types of medication and therapy, are still relatively unknown. Moreover, some schizophrenia patients may experience practical limitations when it comes to finding relevant information or viewing their treatment plan themselves.

1.1.1 Schizophrenia

Schizophrenia is a mental disorder that affects approximately 1% of the population. The illness is characterized by psychoses, which are episodes involving a loss of contact with reality. The symptoms of the illness are caused by impaired processing of information in the brain in combination with gene-environment interactions (Van Os and Sham 2003).

Schizophrenia is characterized by cognitive dysfunctions and abnormalities in the perception of reality. People diagnosed with schizophrenia often experience hallucinations, delusions, and disorganized speech and thinking, accompanied by significant social and occupational problems (American Psychiatric Association 2000). Due to the complexity of this disorder and the diversity of care needed, proper and

frequent evaluation of treatment is particularly vital. That is why *routine outcome monitoring*, i.e., yearly assessments, offers much potential for better care (Opler et al. 2002).

1.1.2 Current schizophrenia treatment

Current schizophrenia treatment in the Northern Netherlands is centered around patient assessments through Routine Outcome Monitoring (ROM). In recent years, ROM has become increasingly important as part of a growing belief in the need for standardization in order to evaluate and improve patient care. A ROM assessment for a patient is conducted every 6 months or every year. These assessments involve physical fitness tests as well as a number of questionnaires that assess psychiatric and psychosocial problems, satisfaction, and care needs. The ROM protocol makes use of a number of questionnaires, e.g., the Health of the Nation Outcome Scales (HoNOS) (Wing et al. 1998) and the Manchester Short Assessment of Quality of Life (MANSA) (Priebe et al. 1999).

A simplified abstraction of the current schizophrenia management life cycle is shown in Figure 1.1. The results of a ROM assessment form the basis for a long-term treatment plan that is determined in a meeting between patient and clinician. These meetings take place roughly six weeks after an assessment. During the meeting, a treatment plan is formulated that is followed until the next assessment.

During the rest of the year, i.e., when in ambulatory or in-patient care, patients may collect *electronic patient diary data*, which is data entered by patients in a (web) application. Not all forms of electronic patient diary data are suitable for analysis. Here, we restrict ourselves to electronic psychometric data, i.e., pre-formatted questionnaire data. The patient fills out the questionnaire using the application, and the calculated summary scores of the questionnaire are used as data points. Participating patients are asked to fill out the questionnaire either daily or multiple times per day, at set intervals. Electronic patient diary data can accurately reflect the state of various aspects of a patient. Analysis of this data can reveal how the symptoms of an individual evolve over time, how they can be predicted, and which factors contribute to effective treatment.

1.1.3 Problems

There is increasing concern that patients are not sufficiently engaged in their treatment meetings, because they are not always adequately prepared to have a discussion. Patients have no direct access to the assessment results prior to the meeting and hear these results only through their clinician. This scenario creates an inequality wherein the patient is highly dependent on the expertise of the clinician and

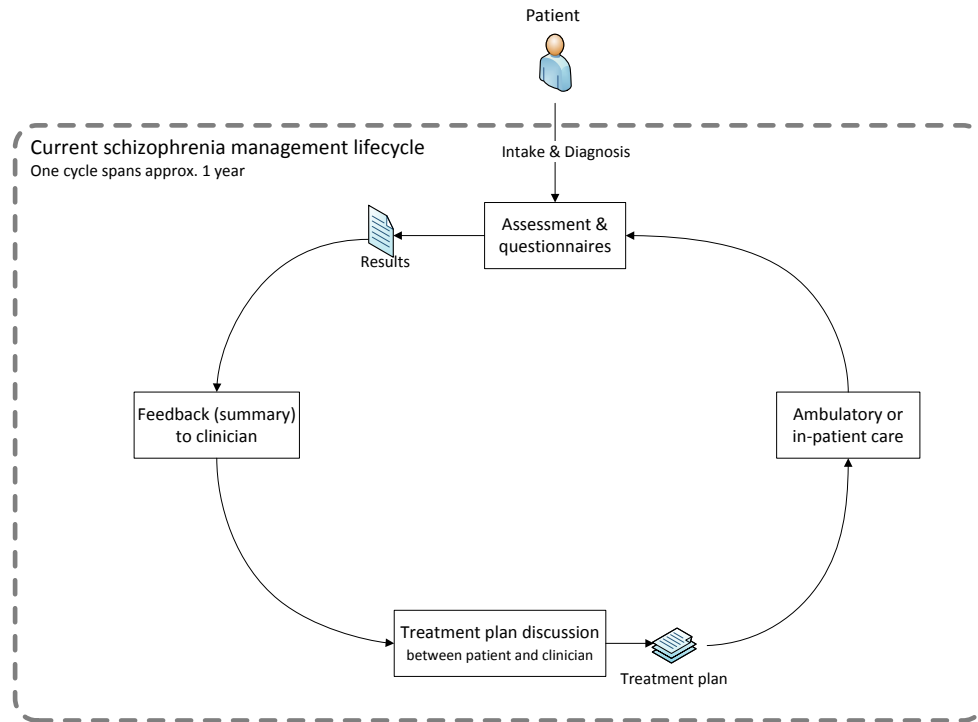


Figure 1.1: An abstraction of how schizophrenia in the Netherlands is currently managed. The events follow a yearly cycle where the treatment plan is adjusted according to assessment results.

cannot participate fully in medical decision making. In recent years, the ethics of such medical paternalism have been called into question (Deegan and Drake 2006).

To better prepare patients for meetings with their clinician, tools have recently been developed to support *shared decision making* (Godolphin 2009, Woltmann et al. 2011), which is considered an ethical imperative (Drake and Deegan 2009). Shared decision making is an approach in which patient and clinician are equal participants in deciding the treatment plan. Moreover, the approach emphasizes that patients should have access to the same information regarding their (mental) health as the clinician (Charles et al. 1997). Shared decision making is widely in use and has proved clinically successful for chronic illnesses (Duncan et al. 2008, Fullwood et al. 2013).

So far, however, sharing healthcare information with the patient in a direct and unsupervised manner, as part of shared decision making, has not been applied for schizophrenia patients. Moreover, to the best of our knowledge, there has been

no research on the automated translation of assessment results into relevant information for schizophrenia patients. There are a number of reasons for this. First, clinicians have traditionally subscribed to the belief that they need to protect their patients against potentially disturbing outcomes. Second, schizophrenia patients may experience a disturbed cognitive state and as a result clinicians may have been reluctant to gather data from them. Third, tools that facilitate shared decision making for schizophrenia patients require careful development because schizophrenia patients have special needs regarding the presentation of information, for example, via a simply structured and calm website using text for a low reading age (Schrank et al. 2010), that is, using text without difficult words.

Another issue is the cost of making effective use of electronic patient diary data. Currently, this type of data is collected only as part of small scale research projects involving few patients because the analysis requires time and effort from statisticians. Existing ways to automate this analysis still require statistical expertise to operate and thus do not scale well. We could gain knowledge and insight into long-term chronic illnesses and the interaction of their symptoms by applying a fully automated approach for analyzing electronic patient diary data.

1.2 Scope of this thesis

We researched, conceptualized, designed, implemented, and evaluated two systems. *Wegweis* is a web application that uses assessment data stored in the electronic medical records of schizophrenia patients to provide them with personalized advice. The advice is automatically generated and presented to patients without requiring human supervision and in accordance with guidelines and rules coded in a hierarchical ontology that is verified by experts. Our second project, *Autovar*, uses electronic patient diary data to identify cause-and-effect relationships between symptoms, medication use, and other activities, for individual patients. In *Autovar*, we automate all steps of vector autoregression that previously required statistical expertise.

Schizophrenia treatment is a complex affair and may involve different types of medication, psychotherapy, and forms of social support (Van Os and Kapur 2009). As a result of the side effects of medication and differing priorities of individual patients, it is currently impossible to predict which combination of medication and therapy is most desired for an individual patient. Hence, it is important for the patient to know what options and alternatives are available, and also to be able to evaluate their efficacy for themselves personally. We address the former issue with *Wegweis*, and the latter with *Autovar*. Thus, we find that both our systems use automated knowledge extraction applied to electronic medical data and affect the

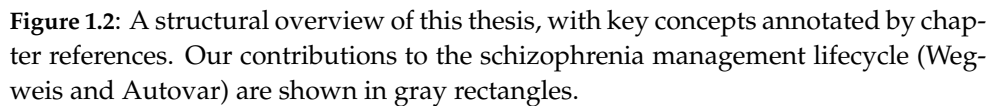


Figure 1.2 shows our contributions to current schizophrenia management. The yearly assessment data that was previously sent only to clinicians is now directly available for patients through Wegweis. To the best of our knowledge, Wegweis is the first system that provides schizophrenia patients with direct access to their assessment results and is also the first system for these patients to apply ontological reasoning in selecting personalized advice. The goal of Wegweis is to enable patients to better prepare themselves for discussing their treatment plan with their clinician, which is one of the principles of a patient-centered approach (Barry and Edgman-Levitan 2012). Figure 1.2 also shows that Autovar uses data collected by the patients. When a patient complains about problems to their clinician, it is often difficult to assess the frequency and severity of occurrence. By filling out daily questionnaires, patients can objectively monitor and report their condition in a way that allows for time series analysis. Analyzing this data normally requires statistical expertise.

Autovar embeds this expertise and enables automated analysis feasible for large scale exploitation. The results of this analysis can be used to determine the efficacy of different aspects of treatment for individual patients.

1.2.1 Wegweis

We propose an ontology-based approach for selecting and ranking information for schizophrenia patients based on their routine assessment results. Our approach ranks information by severity of associated schizophrenia-related problems and uses an ontology to decouple problems from advice, which adds robustness to the system because advice can be inferred for problems that have no exact match.

We have developed Wegweis, a web-based advice platform, to make the assessment data accessible and understandable for patients. We show that a fully automated explanation and interpretation of assessment results for schizophrenia patients, which prioritizes the information in the same way that a clinician would, is possible and is considered helpful and relevant by patients. The goal is not to replace the clinician but rather to function as a second perspective and to enable patient empowerment through knowledge.

We created a problem ontology, validated by a group of experts, to combine and interpret the results of multiple schizophrenia-specific questionnaires. We designed and implemented a novel ontology-based algorithm for ranking and selecting advice based on questionnaire answers. We designed, implemented, and evaluated Wegweis, a proof of concept for our algorithm, and, to the best of our knowledge, the first fully automated interpretation of assessment results for patients suffering from schizophrenia. We evaluated the system vis-à-vis the opinions of clinicians and patients in two experiments. For the task of identifying important problems based on MANSA questionnaires (the MANSA is a satisfaction questionnaire commonly used in schizophrenia assessments), our system corresponds to the opinion of clinicians 94% of the time for the first three problems and 72% of the time overall. Patients find two out of the first three advice topics selected by the system to be relevant and roughly half of the advice topics overall.

The main contribution of Wegweis is the construction of a robust framework that uses the electronic medical record for ranking and filtering information that is personalized for each patient. We show that a fully automated explanation and interpretation of ROM assessment results for schizophrenia patients that prioritizes the information in the same way that a clinician would is possible and is considered helpful and relevant by patients. This work forms an important step towards implementing shared decision making as part of the standardized approach in schizophrenia treatment.

1.2.2 Autovar

With the advances in portable consumer electronics, i.e., phones and tablets with internet access, the medical field has started using electronic patient diaries as an important means of collecting medical data. Recent studies have found these diaries suitable for time series analysis of patient symptoms using vector autoregression. Vector autoregression describes a specific set of statistical models used for modeling time series data of multiple variables. These models allow for forecasting, impulse-response analysis, and inferring the strength and direction of causality between variables.

Finding the best vector autoregression model for any data set, medical or otherwise, is a process that, to this day, is frequently performed manually in an iterative approach that requires time and expertise from statisticians. Very few software solutions for automating this process exist, and they still require statistical expertise to operate.

We propose a software solution called Autovar to automate the process of finding vector autoregression models for time series data, implementing an approach that closely resembles the way in which experts work manually. In our approach, we include improvements over the manual approach by leveraging the computing power that is made available through automation, e.g., by considering multiple alternatives instead of choosing just one.

In this thesis, we present our approach for automating vector autoregression, we describe the design and implementation of Autovar, we compare its performance against experts working manually, and we compare its features to those of the most used commercial solution available today. Our goal is to determine whether the approach of experts can be automated to an extent where vector autoregression no longer requires human supervision.

The main contribution of Autovar is to show that vector autoregression on a large scale can be feasible. We show that an exhaustive approach for model selection can be relatively safe to use. This work forms an important step toward making adaptive, personalized treatment available and affordable for all branches of health-care.

1.3 Thesis organization

Chapter 2 provides a brief overview of developments throughout the history of Artificial Intelligence (AI) that are relevant to our current work.

Chapter 3 narrows the scope and reviews the efficacy of e-health self-management applications for psychotic disorders to introduce the context of Wegweis (Van der Krieke et al. 2014).

Chapter 4 introduces our web application Wegweis (Emerencia et al. 2011). The chapter discusses the system design, user interface, and the custom problem ontology that forms the background knowledge used in the approach. We explain our algorithms for selecting and ranking advice in pseudocode and provide further implementational details.

Chapter 5 evaluates different aspects of Wegweis. We conduct a usability study of the system consisting of a heuristic evaluation, a qualitative evaluation and a quantitative evaluation (Van der Krieke et al. 2012). We also evaluate the functionality of the system in a study where we quantified how closely our method corresponds to the opinions of clinicians and to the opinions of patients (Emerencia et al. 2013).

Chapter 6 introduces our approach for automating vector autoregression on electronic patient diary data (Emerencia et al. 2014). We explain how models are constructed and tested, and how invalid models are handled. We explain our algorithms for automated model selection using pseudocode.

Chapter 7 evaluates Autovar. The chapter details the implementation aspects of Autovar, including the user interface of the web application front-end. We compare the performance of the system to statisticians working with STATA, and we compare the functionality to alternative software for automated model selection.

Chapter 8 concludes the thesis. We present a brief summary of the research and a collection of ideas for future work and investigation.

Chapter 2

Artificial Intelligence in Medicine: a brief overview

To provide context for the current work, this chapter presents a brief chronological selection of relevant developments in the history of artificial intelligence in medicine. The next chapter discusses applications for e-health self-management for psychotic disorders.

Traditionally, the application of computers and artificial intelligence in medicine was limited to the area of therapy recommendation and diagnosis (in particular, decision support systems). Computer-aided diagnosis can be seen as a classification problem where there is a fixed set of classes, called diagnoses, and knowledge embedded in a computer system in order to correctly label an unseen sample, called a case, with its diagnosis.

Therapy recommendation and diagnosis supported by computers has been achieved in a number of distinct approaches. In their implementations, all these approaches necessarily incorporate elements of *knowledge acquisition*, *knowledge representation*, *reasoning*, and (with varying degrees of comprehensibility) *explanation* (Lavrac et al. 2000).

Within the field of Artificial Intelligence in Medicine (AIM), each approach could be traced back to one of two schools of thought. These schools are the *symbolic approach* and the *connectionist approach*.

2.1 Symbolic approach

Expert systems are regarded as the first successful application of AIM (Musen 1999). In the 1970s, expert systems such as Mycin (Shortliffe 1976) tried to model the way in which a practitioner reasons about a problem. These systems work by asking questions to narrow down the search.

While Mycin is regarded as the first of its kind, other expert systems soon followed. Lavrac et al. (2000) give a partial list of expert systems that have successfully been applied in clinical practice: HODGKINS (1976), PIP (1976), CASNET (1978), HEADMED (1978), VM (1980), ONCOCIN (1981), EXPERT (1981), ABEL(1982),

INTERNIST-1 (1982), GALEN (1983), MDX (1983), CADUCEUS (1984), PUFF (1987) and CENTAUR (1997).

While expert systems were fairly successful in the years following their inception with new generations still being developed and used today, they only operate successfully in specific application areas, for a number of reasons. First, constructing an expert system requires a substantial time investment from the developers as well as the practitioners. To model the reasoning of a practitioner with an adequate level of detail, several rounds of interviews and testing are needed to identify all the edge cases. Classical expert systems do not incorporate aspects of machine learning and thus do not improve with use. Second, the efficacy of expert systems has primarily been demonstrated in environments where a myriad of special-purpose rules is in effect. If the depth of decision-making is relatively trivial, then implementing an expert system might not be worth the effort. This limits the scenarios where expert systems might be used in practice. Finally, expert systems may clash with existing clinical practice. In scenarios where there is no need for an approach that involves numerous steps of reasoning and testing, an expert system might end up solving problems that do not need solving while imposing changes on the way in which clinicians work.

We consider expert systems to be part of the *symbolic approach*. The symbolic approach expresses knowledge in a symbolic way, e.g., in *rules*. Rules have the undisputed advantages of simplicity, uniformity, transparency, and ease of inference, that over the years have made them one of the most widely adopted approaches for representing real world knowledge (Lavrac et al. 2000).

In the late 1980s/early 1990s, it became apparent that the most difficult step for the symbolic approach was *knowledge acquisition*. Thus, we see the introduction of several machine learning techniques to automate this process. Most notably, *rule induction* (CN2, C4.5rules, OneRule, Rule Learner, FOIL) and *decision trees* (ID3, ASSISTANT (1983), AQ, CN2, C4.5). In the late 1990s, symbolic approaches were used in data mining information from medical data sets, with an emphasis on relational learning through inductive logic programming (Lavrac et al. 2000).

2.2 Connectionist approach

After an initial surge in popularity following their inception in the late 1950s, *artificial neural networks* had little support left after Papert and Minsky illustrated the limitations of perceptrons, such as not being able to model the XOR function (Minsky and Seymour 1969). It was not until the mid 1980s that we see a return of the use of neural networks in artificial intelligence (and indeed in medicine), due to the back-propagation algorithm (Rumelhart et al. 1986, Werbos 1994), which did allow

for neural networks to learn more complex relations such as the XOR function.

Several algorithms for training neural networks became popular in the late 1980s and 90s, including naive Bayesian networks, Bayesian belief networks, feedforward-backpropagation neural networks and support vector machines. We refer to Kononenko (2001) for a comparison of these classifiers with respect to performance, transparency, explanation, reduction, and missing data handling capabilities. One interesting conclusion from that paper is that the more sophisticated Bayesian belief networks do not necessarily outperform the naive Bayesian classifier. Bayesian networks have been used for diagnostic reasoning, prognostic reasoning and treatment selection in biomedicine and healthcare (Lucas et al. 2004).

To optimize the performance of these algorithms and learning processes, several optimizations have been developed. Examples include ensemble learning, boosting, and expectation-maximization. The utility of the Bayesian network formalism was extended through *influence diagrams*, which take knowledge about decisions and preferences into account (Lucas et al. 2004).

Neural networks constitute the *connectionist approach*. Like the symbolic systems, connectionist systems have been used for diagnosis in medicine. While the symbolic approach intends to model knowledge on the level of human reasoning, connectionist systems, which are networks of interconnected simple units, were believed to operate at a subsymbolic level, providing more accurate accounts of cognition (Smolensky 1987). Unfortunately, accuracy was not always the most important goal in practice. Practitioners favored those systems that were able to show how an answer was derived. For many neural networks, this proved to be difficult since the internal knowledge representation of trained weights does not necessarily translate to real-world concepts.

There have been attempts to let the symbolic and connectionist approaches cooperate rather than compete with one another. Auramo and Juhola (1996) introduce a probabilistic expert system. Cooper (1993) notes that a probabilistic system can be naturally extended to a decision-theoretic system that recommends, for example, diagnostic tests to perform and therapies to administer. He deems it crucial that the field learns more about how to integrate belief networks and decision networks with other knowledge representations and inference methods.

2.3 Measuring performance

In discussions about the symbolic versus the connectionist approach, the question of which one is better has often been asked. The answer depends on the purpose of the system. The problem of comparing different approaches for a specific system is an instance of comparing classifiers over medical data sets, which is done based

on performance.

The performance of a system is not represented as a single quantity, rather there exist numerous qualitative and quantitative properties. The purpose of the system is used in establishing priorities for these criteria. With respect to the *quantitative properties*, the performance of different diagnostic methods is usually described by classification accuracy, sensitivity, specificity, ROC curve, and post-test probability (Kononenko 2001). Other than good performance, for a system to be useful in solving medical diagnostic tasks, the following *qualitative properties* are desired: the ability to appropriately deal with missing data and with noisy data (e.g., errors in the data), the transparency of diagnostic knowledge, the ability to explain decisions, and the ability of the algorithm to reduce the number of tests necessary to obtain a reliable diagnosis (Kononenko 2001).

2.4 In the 2000s

With the increasing availability of large storage devices and the internet in the 2000s, more systems started using direct digital information sources instead of requiring manual input. The internet is seen as a way to make the information sources available that are required by decision support systems (Horn 2000).

Coiera (2003) lists new applications of artificial intelligence in the medical domain: data mining techniques applied to patient data to generate alerts and reminders for practitioners; the field of medical imaging (e.g., CT and MRI scans) using image recognition and interpretation techniques from the fields of computer vision; laboratory analysis; therapy critiquing and planning; and electronic health records. In the following, we restrict our scope to the application domains most relevant to the work presented in this thesis.

In the 2000s, a new generation of decision support systems emerges: those that include the dimension of time in the reasoning process. For example, in *case-based reasoning* systems, a history of cases is taken into account. The problem of inferring the state of a system at various points in time as it changes in response to events is called *temporal reasoning* (Adlassnig et al. 2006, Fisher et al. 2005). Temporal reasoning over clinical data can be performed using data models and languages that support temporal queries. For medical applications to perform temporal reasoning, new data models and languages that supported temporal queries were needed. These were developed over the course of several decades and with varying degrees of complexity (Combi et al. 1997), sometimes as extensions to existing data models and query languages (e.g., SQL (Adlassnig et al. 2006)).

Digital information allows for automated knowledge extraction in a process called *data mining*. In the medical application domain, knowledge can be mined

for example from electronic health records or from patient monitoring equipment. Examples of applications used in practice include using HMMs to detect trends in vital signs in ICU monitoring (Stacey and McGregor 2007), and using case-based reasoning and data mining for monitoring and predicting blood sugar levels (Yuan et al. 2008).

2.4.1 Case-based reasoning

The need for case-based reasoning in medicine has been attributed to the fact that many diseases are not understood well enough for formal models or universally applicable guidelines to be available (Bichindaritz and Marling 2006). Case-based reasoning (CBR) is not a new concept, as it had already proven its use in the 1980s (Kolodner and Kolodner 1987). However, these early CBR systems did not model time explicitly (Augusto 2005).

CBR works by retrieving a set of similar cases for a new case and using those cases and their outcome to give advice to a domain expert. Any given advice is checked and repaired by the domain expert and stored in the database as well. CBR can show relevant features (e.g., causality), provide explanations and can make use of additional symbolic domain knowledge (Lavrac et al. 2000). Bichindaritz et al. (2011) give an overview of some of the early CBR systems in health sciences and note that the prototypical models used in CBR are better adapted to represent biomedical knowledge than other types of models.

When applying CBR to medical data analysis, one has to address several non-trivial questions, including the appropriateness of similarity measures used, the actuality of old cases, and how to handle different solutions (treatment actions) by different physicians (Lavrac et al. 2000). One weakness of case-based reasoning is not being able to associate probabilities and statistics with the results (Bichindaritz and Marling 2006). Furthermore, researchers in this field emphasize the need for standardization of case representations for the purpose of interoperability.

Researchers increasingly recognize the importance of embedding contextual knowledge in decision support systems (Pantazi et al. 2004). Montani (2011) gives a survey of the use of contextual knowledge in recent CBR implementations and concludes that contextual knowledge can make CBR systems more efficient, easier to maintain, and easier to adapt.

2.4.2 Temporal abstraction, representation, and reasoning

Temporal information is crucial in electronic health records and biomedical information systems (Zhou and Hripcsak 2007) for a number of reasons. Temporal information is required in order to derive causal relationships in medical data. Sys-

tems need to be able to interpret contextual statements such as “the last 3 days,” and “the 6th of November,” or specific intervals during which a patient was taking medication, in order to reason over such information. Knowledge structures used for this process of *temporal abstraction* should conform to general requirements for knowledge representation. These requirements include expressiveness, consistency, ease of verification, to be formally well-defined, and to be easily understood by domain experts (Horn 2001, Stacey and McGregor 2007).

Stacey and McGregor (2007) compare various systems for monitoring and managing temporal data in medicine (i.e., RÉSUMÉ, TrenDX, Asgaard, KNAVE) and remark that fusion with data mining processes is necessary to learn new knowledge from stored clinical data. Augusto (2005) gives a comprehensive overview of time-aware decision support systems, and identifies common concepts and terminology used in this field. Zhou and Hripcsak (2007) give an extensive overview of the topics, applications, and theories that exist within the field of temporal reasoning with medical data. They identify processing textual data as one of the more challenging tasks.

In recent years, the use of temporal information to derive causal relationships in medical data is exploited by the application of vector autoregression (VAR) on electronic patient diary data. Vector autoregression has its origins in the field of Econometrics (Sargent 1979) and is typically used in forecasting and analyzing financial models (Anderson 1979, Burbidge and Harrison 1984, Litterman 1986, Primiceri 2005). VAR on electronic patient diary data has been used to find cause and effect relationships between symptoms (Wild et al. 2010, Rosmalen et al. 2012, Hoenders et al. 2012). The results of VAR analysis can provide decision support or therapy recommendation.

2.4.3 Data mining and data analysis

Data mining is the process of finding patterns, trends, and regularities by sifting through large amounts of data (Fayyad et al. 1996, Pena-Reyes and Sipper 2000). Data mining is a collective term used to describe a category of techniques such as text mining, information mining, knowledge discovery in databases (KDD), data extraction, data cleansing, data reduction, model interpretation, model application, and many others (Bull et al. 2008). Data mining has been used to extract medical knowledge for diagnosis, screening, prognosis, monitoring, therapy support and overall patient management (Lavrac 1999).

There is a distinction between supervised and unsupervised data mining (Pena-Reyes and Sipper 2000, Perner 2006). The supervised approach can be seen as a classification problem in the sense that the description attributes of a set of labeled sam-

ples of a target concept are used in learning how to recognize members of that class. The unsupervised approach closely resembles an unsupervised clustering problem where the goal is to discover underlying regularities and patterns.

Lavrac (1999) mentions that KDD typically consists of the following steps: understanding the domain, forming the data set and cleaning the data, extracting of regularities in the form of patterns and rules, postprocessing discovered knowledge, and exploiting results. The popular concept of *intelligent data analysis* is described as an AI approach to KDD, taking domain knowledge into account.

The contention between the connectionist and symbolic approaches is apparent in the field of data mining as well. Zupan et al. (2006) remark that methods of data analysis and knowledge revision that explicitly rely on background knowledge have given way to sub-symbolic computational methods designed to maximize classification accuracy (e.g., neural networks, support vector machines, and HMMs). However, they note that this focus is changing, referring to the use of biomedical ontologies.

2.5 State of the art

Many of the technologies discussed in the previous section are still being used, developed, and implemented today. Aside from extensions to existing work, we can also identify several new trends.

2.5.1 Standards and interoperability

To improve care, independent health services need to be able to cooperate. This gives rise to new challenges. For example, there are many different (often locally customized) implementations of electronic health records (EHRs). A comparison of EHR approaches is given in Blobel and Pharow (2009). Standards have been developed as a requirement for the interoperability of healthcare applications. Examples include standards for messaging formats (HL7 v2.x, HL7 v3.x, ISO13606) (Vogt and Wittwer 2007), for patient summaries (HL7 CDA, CCR, CCD) (Ferranti et al. 2006), and for terminology (GALEN, UMLS, LOINC, SNOMED-CT, DICOM – for images) (Leong et al. 2007). Bender and Sartipi (2013) reflect on the development of HL7 v3, its criticism, and its evolution in the form of HL7 FHIR. The use of these standards is seen as a requirement for success in healthcare IT environments (Leong et al. 2007).

We also see a change from electronic health/medical records (for physicians), to personal health records (for patients). Personal health records (PHRs) are managed by the patients themselves instead of by practitioners and are stored at sites such as Microsoft HealthVault (Gorman and Braber 2008) instead of at a specific

hospital. Between 2008 and 2012, Google ran Google Health, but this service was discontinued due to lack of widespread adoption (Brown and Wehl 2011). Plastiras et al. (2014) compare the functionality of Several PHRs, including Microsoft HealthVault, Telemedical, NoMoreClipboard, Health Spek, and Health Companion and found that a major barrier for PHR adoption is the interoperability (or lack thereof) between PHRs and EHRs. Their solution is the development of an ontology-based information model for PHR to EHR interoperability.

The establishment of shared care must be supported by distributed, interoperable information systems (Globel 2006, Krummenacher et al. 2009). Globel (2006) concludes that for an open, user-centric, user-friendly, flexible, scalable, and portable EHR, a component-oriented model-driven architecture should be used.

For these interoperable systems, security and confidentiality of data are critical considerations for practitioner adoption (Hare et al. 2006). Adding security services into healthcare systems architectures and other suggestions for establishing trustworthiness are discussed in Globel et al. (2006), Globel (2007).

Internationally, governments are moving towards more interoperable architectures for eHealth ecosystems. The National E-Health Transition Authority (NEHTA) in Australia have used the Reference Model for Open Distributed Processing (RM-ODP) to build architectures and interoperability guidelines for eHealth systems at an enterprise level (Bond et al. 2013), and found that ODP standards are increasingly used in the HL7 standardization efforts. In Europe, the eSOS (Thorp 2010) and Renewing Health¹ projects are pilots for interoperable patient summaries, medication workflows, and telemonitoring on a large scale (Sauermann et al. 2013). European Ministerial eHealth conferences (e.g., eHealth2012 Copenhagen, eHealth2013 Dublin) learn from these projects to facilitate interoperability at the national, regional, and local levels.

2.5.2 Ontology-based applications

Medical ontologies give background knowledge, such as interpretations and relations, to data expressed in standardized formats. Dietterich et al. (2008) stress the need for using background knowledge. Dealing with background knowledge requires some way of effectively making use of logical knowledge in the form of relational schema and/or ontologies to constrain or bias the structure of the probabilistic model (Dietterich et al. 2008).

There are numerous examples of ontology-based applications in healthcare. For example, ontologies are used in the middleware of pervasive health systems for monitoring patients and managing alerts (Paganelli 2007) and for generating clini-

¹www.renewinghealth.eu

cal reminders for clinicians (Buranarach et al. 2009). Another example is TrialX, a web application that uses its own ontology to interpret and evaluate data stored in personal health records in order to match patients to clinical trials (Patel et al. 2010). More closely related to our project Wegweis is SEMPER, an interactive web-based platform that assists patients to self-manage work-related disorders and alcoholism. SEMPER uses ontologies for query expansion in text mining in documents (Maier et al. 2010). Kuriyama and colleagues (2007) developed an application for mobile devices for collecting and sending lifestyle data that are used to display health advice in a web application. They use an ontology to suggest exercises based on the goals of the patient.

2.5.3 Ambient intelligence

The next step in the evolution of AI, and the successor of ontology-based systems according to some researchers, is Ambient Intelligence (Ramos et al. 2008). Ambient Systems incorporate the operation of several related fields (e.g., ubiquitous computing and pervasive computing) combined with a higher level of artificial intelligence (Ramos et al. 2008).

Ambient intelligent (AmI) systems combine the following traits (Cook et al. 2009): sensitive, responsive, adaptive, transparent, ubiquitous, and intelligent. In contrast to previous techniques in AI, AmI applications are centered around the human user and focus more on local environments such as rooms, vehicles, or homes (Ramos 2007, Ramos et al. 2008, Augusto and McCullagh 2007). Augusto and McCullagh (2007) describe the scope of AmI, including several scenarios of application. They also stress the importance of safety critical AmI systems to be accepted by users and to be thoroughly tested to reduce the potential for error.

Ambient intelligence has endured criticism, especially related to security and privacy (Brey 2005, Crutzen 2007, Friedewald et al. 2007). Brey (2005) states that AmI has the potential to limit freedom and autonomy and warns for potential privacy risks. AmI technology goes beyond most of currently existing privacy-protecting borders (Friedewald et al. 2007).

2.5.4 Patient-centered applications

Until the 2000s, most applications of AIM were strictly practitioner-centered. The traditional applications of diagnosis and decision support were designed to support the practitioner. Currently, there is a paradigm shift toward more patient-centered (web) applications. The trend is that patients are granted more control over their treatment through personalized websites (Soto and Spertus 2007, Arsand and Demiris 2008, Andry et al. 2008, Gené Badia et al. 2009). Examples include hos-

pital websites where patients can schedule appointments and pharmacy websites where patients can order medication online (Sánchez et al. 2007). Buzzwords such as “E-Health” and “Health 2.0” have been coined to term this sentiment (Igras 2007, METU-SRDC 2007, Bos et al. 2008, Gorman and Braber 2008).

While patient-supporting web applications are already in use for mental illnesses such as anxiety, depression, and addiction (Proudfoot 2004), for schizophrenia and other severe mental illnesses, less has been achieved thus far (Kersting et al. 2009, Riper 2007, Välimäki et al. 2008).

There is a subtle difference in terminology between patient-centered and *patient-centric* care. In patient-centered care, the opinion and needs of the patient are more taken into account. In patient-centric care, the patient is the main source of the health care interactions and personalized data (Scher 2012).

2.6 Schizophrenia and other psychotic illnesses

In Finland, Välimäki and colleagues (2008) have developed the Mieli.Net portal, a patient-centered computer-based support system for schizophrenia patients. It aims to support self-management by offering (i) information on treatment, support, and rights; (ii) a channel for peer support; (iii) a tool for counseling; and (iv) interaction with clinicians by means of a question-and-answer column. A prototype was developed and has been evaluated by patients and healthcare staff. Both nurses and patients were able to work with the system (Koivunen et al. 2007, Välimäki et al. 2008, Koivunen et al. 2010). Patients were able to access services and find relevant information (Koivunen et al. 2007), and they report their satisfaction with the system (Kuosmanen et al. 2010).

In the Netherlands, two recent initiatives have been launched aimed at enabling empowerment of schizophrenia patients. The first is “Eigen regie bij schizofrenie” (translation: personal control over schizophrenia), a website to support patients in their self-management (*Eigen Regie Bij Schizofrenie* 2011). It offers tools for scheduling appointments, checking medication, viewing the treatment plan, sharing experiences, and requesting services. Clinicians can use the website to monitor the condition of patients and detect problems early. The second initiative is SamenKeuzes-Maken.nl (translation: making decisions together), a website that is modeled after a program of Deegan and colleagues (2008) that implements the concept of shared decision making (*Samen Keuzes Maken* 2011). It offers information about recovery, videos portraying experienced patients, a questionnaire in preparation for meeting the clinician, and links to informational websites. We note that there is no true sharing of information here, since the patient fills out a separate questionnaire on the website and does not gain access to the assessment results that their clinician has.

2.6.1 Wegweis

In relation to other ontology-based applications in healthcare, our application Wegweis is novel because it is the first application that shows information originally intended for clinicians (assessment results) to schizophrenia patients, and uses an ontology to automate the translation from results to information. This automated translation is an important step in implementing one of the core requirements of shared decision making (i.e., the sharing of medical information) at low operational costs.

While there are other web applications for schizophrenia patients that support shared decision making, they do not support the direct sharing of assessment information. In addition, Wegweis provides an interpretation through applying ontological reasoning, as we will explain in Chapter 4. Wegweis can rank and personalize information for individual patients. This functionality can also be abstracted and applied to existing self-management websites in order to make them more personalized and easier to use for patients.

The question that remains is whether applications such as Wegweis have measurable benefits or other effects for the patient. In the next chapter, we take a closer look at the efficacy of different types of e-health self-management applications for psychotic disorders.

Parts published as:

L. van der Krieke, L. Wunderink, A. Emerencia, P. de Jonge, S. Sytema – “E-Mental Health Self-Management for Psychotic Disorders: State of the Art and Future Perspectives,” *Psychiatric Services*, (65:1), pp. 33-49, 2014.

Chapter 3

E-health self-management for psychotic disorders

The aim of this chapter is to investigate to what extent information technology may support self-management among service users with psychotic disorders. The investigation aimed to answer the following questions: What types of e-mental health self-management interventions have been developed and evaluated? What is the current evidence on clinical outcome and cost-effectiveness of the identified interventions? To what extent are e-mental health self-management interventions oriented toward the service user?

3.1 Introduction

Online therapies (Marks et al. 2007), web-based self-management systems (Proudfoot et al. 2007), and internet forums (Haker et al. 2005, Vayreda and Antaki 2009) are rapidly becoming part of the mental health services repertoire. These “e-mental health” technologies are deemed likely to facilitate self-help processes (Marks et al. 2007, Kenwright et al. 2001); to lessen risk of stigmatization (Marks et al. 2007); to offer faster, easier, and more (cost-) effective access to help (Marks et al. 2007, Kenwright et al. 2001, McGorry et al. 2009, McCrone et al. 2004, Kilbourne 2012); and to provide a more neutral space in which service users can speak more freely (Ainsworth 2002, Marks et al. 2007). As a consequence, e-mental healthcare has the potential to support shared decision making, service user empowerment and self-management (Gerber and Eiser 2001, Grohol 2003, Sanyal 2006, Bos et al. 2008). A review of self-management interventions has shown that computer-based interventions are effective for service users with panic disorders, phobias, and obsessive-compulsive disorders, leading to reduction of symptoms and better quality of life (Barlow et al. 2005). Moreover, most service users seem to appreciate computerized interventions, in particular for enabling them to access services at home whenever they choose (Barlow et al. 2005).

It is, however, unclear to what extent information technology is used to support self-management for people with psychotic disorders. Researchers and practitioners tend to consider psychotic disorders to be less suitable for e-mental health interventions because of the complexity and severity of the disorder (Kersting et al. 2009). Cognitive deficits may limit effective navigation through user interfaces (Rondoni et al. 2007), and delusions may interfere with the use of webcams, sensors, and other devices (Bell et al. 2005). So far, only one review has investigated the use of information and communication technology by service users with psychotic disorders (Välimäki et al. 2012), and it focused on psychoeducation interventions only. Results indicated that there were no differences in effect on compliance and overall functioning between these technology-based psychoeducation interventions and standard care. This finding is important because it might indicate that e-health interventions may be more cost-effective than standard care if e-health can be implemented with little cost.

In this chapter, we explore the state of the art of e-mental healthcare applications for self-management for people with a psychotic disorder. We aimed to answer the following questions: What types of e-health self-management interventions have been developed and evaluated? What is the current evidence on clinical outcome and cost-effectiveness of the identified interventions? To what extent are e-health self-management interventions service user oriented?

3.2 Methods

3.2.1 Search strategy

We conducted a systematic literature search of the following databases, up to July 2012: MEDLINE, PsycINFO, AMED, CINAHL, and the Library, Information Science and Technology database. We used the terms schizophrenia, schizophrenic, schizoid, schizo-affective, schizoaffective, schizophreniform, schizophrenia*, schizophrenic*, schizoid*, schizo-affective*, schizoaffective*, schizophreniform*, schizomaniac, psychosis, psychotic, delusion, delusional, severe mental illness, and severe mental disease. These terms were crossed with computer*, digital, online, Web, Web-technology, Web-based, Internet*, Internet portal, Web technology, technology, computer aided, computer facilitated, information technology, CD-ROM, communication technology, interactive, gaming, multimedia, informatics, cell phone, smartphone, mobile phone, ecological momentary assessment, experience sampling, decision support system, decision aid, serious gaming, edutainment, edugame, telehealth, telepsychiatry, telemedicine, e-health, and e-mental health as free text words and medical subject heading terms.

The search was limited to references in English, German, French, and Dutch. Reference lists of retrieved articles were searched for additional relevant studies. The full search strategies can be obtained on request.

3.2.2 Definitions

E-mental health was defined as the use of information and communication technology to support or improve mental healthcare. To define self-management, we used the description introduced by Barlow and colleagues (2005): "Self-management refers to the individual's ability to manage the symptoms, treatment, physical and psychosocial consequences and life style changes inherent in living with a chronic condition. Efficacious self-management encompasses the ability to monitor one's condition and to affect the cognitive, behavioral and emotional responses necessary to maintain a satisfactory quality of life." As reflected in the definition, self-management is a broad concept involving multiple domains.

3.2.3 Study selection criteria

We included clinical trials as well as observational (feasibility and acceptability) studies because our aim was to provide a comprehensive overview of the interventions developed. In addition, feasibility and acceptability studies offer valuable information for setting future directions for research and development. A study protocol was established before study selection. It was tested on a sample of seven studies and refined accordingly. Articles were included when they described a study focusing on the use of an e-health tool or intervention delivered via a computer, phone or mobile phone, personal digital assistant (PDA), or other device connected to a computer or server, whether Internet based or not for use by persons with schizophrenia or a related psychotic disorder or described a tool or intervention that can help service users with schizophrenia or a related psychotic disorder to manage their illness and well-being and improve their outcomes. Articles had to present original data; that is, reviews were excluded.

Exclusion criteria were studies describing an e-health tool or intervention designed for research or diagnostic purposes only or for use by service users' relatives. Letters, editorials, speeches, posters, comments, book reviews, and theoretical or background articles also were excluded. Furthermore, we excluded articles investigating computer-based cognitive remediation or cognitive enhancement therapy, because good reviews of remediation have already been published (Twamley et al. 2003, McGurk et al. 2007, Grynszpan et al. 2011, Wykes et al. 2011).

In addition, we decided that in case of multiple publications on the same study, the most representative publication (the most recent or complete study or the best

study design) was to be included and described in the Results section, with reference to the related publications.

3.2.4 Data extraction

Studies were identified and selected by three raters independently. Interrater reliability of the selection of studies, calculated as Fleiss' kappa, was .78, which indicates good reliability (Altman 1991). Disagreements between the raters were discussed until consensus was reached. For a flowchart of the retrieval procedure see Figure 3.1. Data were extracted by one reviewer, and a random check was conducted by a second reviewer, which revealed no significant deviations.

3.2.5 Quality assessment

Quality assessment of the clinical trials was conducted by using the Downs and Black scale (Downs and Black 1998), which consists of 27 criteria to evaluate both randomized controlled trials (RCTs) and nonrandomized trials. The Downs and Black scale is considered to address the key quality methodological domains important for assessment in the context of systematic reviews (West et al. 2002), covering reporting, external validity, bias, confounding, and power. In the original version of the scale, studies can obtain a maximum of 32 points. For this study, the original scoring was modified slightly; specifically, the scoring for question 27, dealing with statistical power, was simplified to 1 or 0, as has been done by others (Chudyk et al. 2009, Samoocha et al. 2010). Consequently, the maximum total score that studies could obtain in this review was 28. The score ranges were grouped into the following four quality levels: excellent (score=26–28), good (score=20–25), fair (score=15–19), and poor (score <15) (Chudyk et al. 2009, Samoocha et al. 2010).

Three raters independently conducted the quality assessment. Interrater reliability—calculated with two-way, single-measure mixed intraclass correlations with absolute agreement—was .72, which is good, according to Cicchetti (1994). A quality assessment of acceptability and feasibility studies was not conducted, because there are no validated quality assessment instruments of this kind in this area.

3.2.6 Statistical analysis

To calculate effect sizes of the clinical trials, we used Hedges' g coefficient, which is a standardized mean difference, d , multiplied by a correction factor, J , where $J = 1 - [3/(4df - 1)]$, in which $df = df_{\text{total}} - 2$. Positive values indicated that the intervention condition improved more than the control condition, and we used Cohen's (1988) stratification of effect sizes, where .20 is small, .50 is medium, and

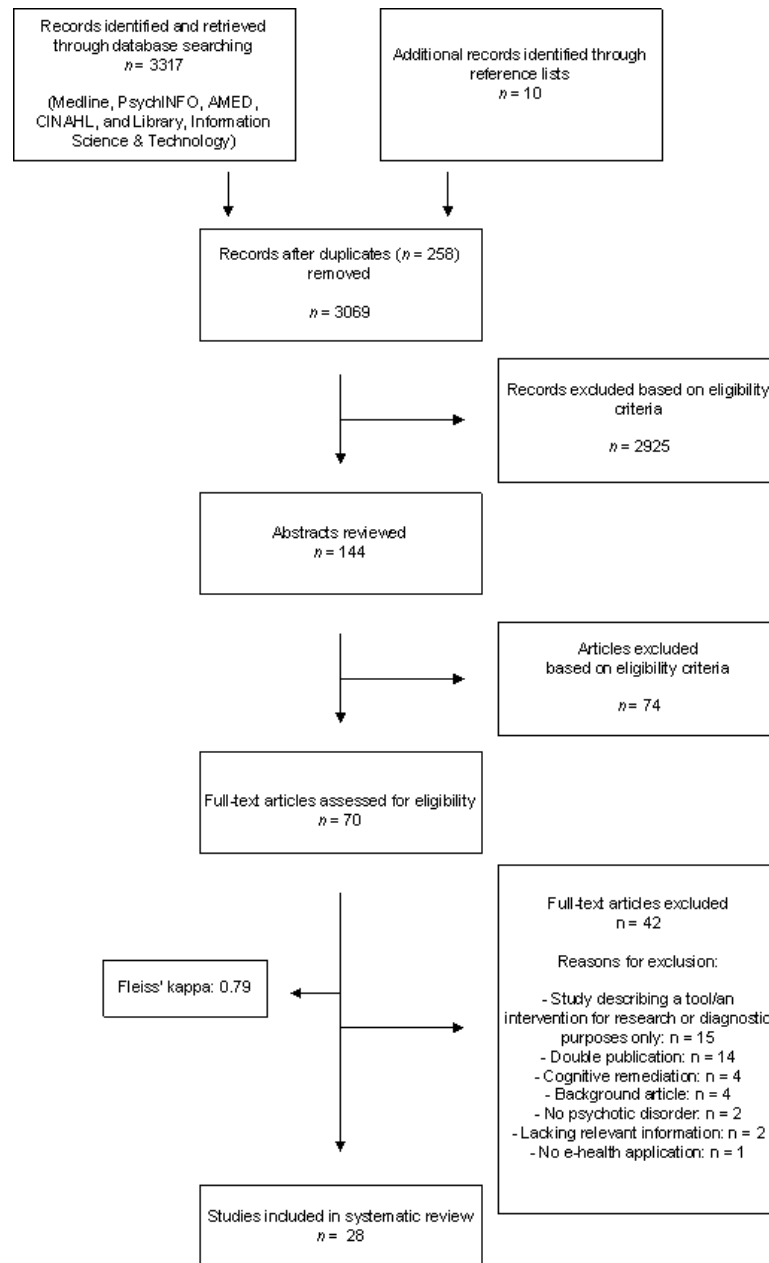


Figure 3.1: Flow diagram of the retrieval procedure.

.80 is large. A meta-analysis was performed when two or more studies could be clustered on the basis of intervention type and when these studies had a similar outcome measure. In case of multiple primary outcome measures, we chose the one that best fit the goal of the intervention type. When multiple control groups were included, we compared the intervention group with the group that received care as usual. In cases where more than one assessment was available, we used the first assessment after the intervention ended. For studies that could not be included in the meta-analysis, we calculated individual effect sizes.

In all cases, the random-effects model was chosen because of anticipated heterogeneity between research designs. All analyses were performed with version 2 of Biostat's comprehensive meta-analysis program.

3.3 Results

The search identified a total of 28 studies meeting the inclusion criteria for the systematic review; 14 studies were clinical trials (11 RCTs and three nonrandomized trials), and 14 were feasibility and acceptability studies. Study characteristics and key results are presented in Van der Krieke et al. (2014). Our quality assessment revealed that four clinical trials were of fair quality and the remaining trials were of good quality. Across all studies, attrition varied from 0% to 50% and was lowest in studies in which convenience sampling was used as the recruitment strategy.

3.3.1 E-mental health self-management interventions and outcome

Although the identified self-management interventions showed substantial variability in form, content, and duration, the studies could be clustered according to the self-management components they focused on, as presented below. Effect sizes of clinical trials, grouped by intervention type, are presented in Figure 3.2. Summary effect sizes could be calculated for three intervention types, namely psychoeducation, medication management, and communication and shared decision making. For the remaining intervention types, the number of included studies was not sufficient to calculate a summary effect size.

Psychoeducation

Most studies focused on psychoeducation. Computer programs (available offline, not via the Internet) examined by Madoff and colleagues (1996), Walker (2006), and Jones and colleagues (2001), as well as the Web portal described by Farrell and colleagues (2004), provide general information about schizophrenia and psychotic disabilities, medication, other treatment options, and various community services,

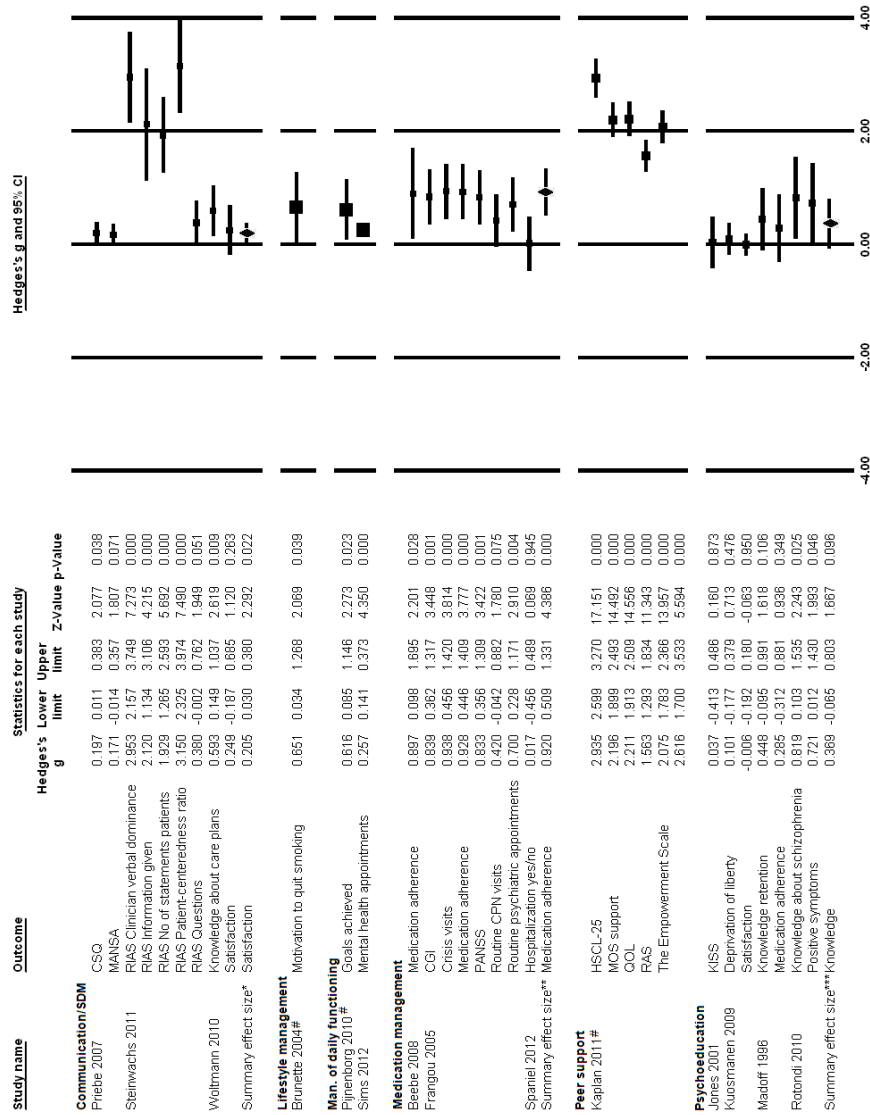


Figure 3.2: Effect sizes of the clinical trials. # Control group is waiting list (control groups of all other studies are care as usual). *Based on Steinwachs et al. (2011), Wolmann et al. (2011). **Based on Beebe et al. (2008), Frangou et al. (2005). ***Based on Jones et al. (2001), Madoff et al. (1996), Rotondi et al. (2010). NB: Size of the central point in hedges' g indicates the sample size.

such as housing, employment services, and rehabilitation services. Two other studies described computer programs that contain additional interactive parts, such as online psychoeducation therapy groups and a channel for peer support (Kuosmanen et al. 2009, Rotondi et al. 2010). An additional study reported results of a so-called “serious game” (Shrimpton and Hurworth 2005), which is a game designed for an educational purpose, thus combining learning with fun. In this case, the game was designed to enhance service users’ understanding of psychosis. In the usage scenario anticipated by the designers, service users could play the game during several sessions at a community mental health center or at home and discuss their gaming experiences afterward with a clinician.

The effect size for e-mental health computerized psychoeducation interventions compared with usual care on the outcome of knowledge was small (Hedges’ $g=.37$; 95% confidence interval [CI]=-0.07 to .80), based on three studies (Madoff et al. 1996, Jones et al. 2001, Rotondi 2010).

Medication management

Four studies investigated an e-health tool or intervention directed at management of medication. In the study by Frangou and colleagues (2005), service users were provided a medication dispenser that recorded their medication adherence. Every time service users opened the box to take a pill, the medication dispenser transmitted this information via a modem to the computer of the research team. When service users took less than 50% of their prescribed medication, the computer sent an e-mail alert to their clinician. The study by Španiel and colleagues (2012) described a mobile phone intervention that aimed to detect early-warning signs of psychotic relapse. Service users in the study were instructed to complete a ten-item Early Warning Signs Questionnaire sent weekly by an automated system to their mobile phones, via short-message system (SMS text message) request. If a certain threshold was exceeded, the service user’s psychiatrist received an e-mail alert recommending contacting the client and increasing the dosage of antipsychotic medication by 20%. In these two studies, the interventions primarily enabled better monitoring of service users by clinicians.

The other two studies focused on medication management by promoting a more active role among service users. Beebe and colleagues (2008) described a nursing telephone intervention to support problem solving. Participating service users received a weekly phone call from a nurse. During this phone call, service users were guided in problem-solving processes for a variety of difficulties identified. Furthermore, they received reminders regarding medication and were provided means to assess the effectiveness of coping efforts. Bickmore and colleagues (2010) examined a computer-based antipsychotic medication adherence system with an avatar agent

installed on a laptop at the service users' homes. After service users powered on the laptop, the avatar started talking to them about their medication use. Service users could respond by clicking a button from a dynamically updated multiple-choice menu. The avatar also taught techniques for self-maintenance (such as using a multi-compartment pill box and a calendar) and encouraged service users to engage in physical activity, such as a 30-minute walk.

E-health medication management interventions compared with care as usual had a large effect on medication adherence (Hedges' $g=.92$; $CI=.51-1.33$). This finding is based on two studies (Frangou et al. 2005, Beebe et al. 2008).

Communication and shared decision making

Six studies were directed toward improved communication between service user and clinician or toward a process of shared decision making. Priebe and colleagues (2007) described a computer program for service users to rate their satisfaction with and need for extra help on eight life domains. The output was interpreted by the clinician and used in a therapy session with the service user. Sherman (1998) reported on an intervention with an electronic application to support service users in creating advance directives. Advance directives are documents containing instructions about what actions should be taken in regard to service users' health in case psychosis renders them incapable of making rational decisions. Service users were provided with an interactive presentation about the purpose, types, and pros and cons of advance directives; they were evaluated to determine whether they had the capacity to master the information; and they were interviewed about topics they would like to include in their directives. Finally, a copy of the advance directives was printed, including a wallet-sized card stating that an advance directive exists and where to access it.

In the study by Deegan and colleagues (2008), service users were provided with an Internet-based computer program that supported them in identifying and formulating their personal values associated with medication use in advance of an appointment with their psychiatrist. If service users needed help using the computer, they received it from a peer. The computer program first explained the concept of recovery and encouraged service users to reflect on their own personal strategies and means of supporting recovery and wellness. Service users completed a survey inquiring about their symptoms, psychosocial functioning, and medication use. In addition, they were asked about a number of common concerns regarding medication use, and finally, they were encouraged to formulate a personal goal before their psychiatric appointment. After service users completed the various steps, the computer generated a report for them as well as for their psychiatrist, for discussion at their next appointment.

Woltmann and colleagues (2011) investigated the feasibility of an application to facilitate shared decision making in care planning. At a computer kiosk in the mental health service facility, clients could use a touch screen to indicate their personal priorities and ideas for healthcare services. On the basis of this information, service users could create their personal care plan. After case managers completed a similar process, the two perspectives were merged electronically and discussed in a meeting in which service user and case manager created a final care plan. Steinwachs and colleagues (2011) reported about YourSchizophreniaCare, a Web-based intervention that helps service users navigate six areas of care (medication, side effects, referrals, family support, employment, and quality of life). Service users answered questions and were given personalized feedback, including videos of actors recommending how to discuss specific topics with clinicians. In the most recent study, van der Krieke and colleagues (2012) assessed the usability of a Web-based support system that gives service users access to the results of their routine outcome monitoring and provides concrete and personalized advice. The system is designed to support service user participation in medical decision making.

E-health communication and shared decision-making interventions compared with care as usual had a small effect on satisfaction (Hedges' $g=.21$; $CI=.03-.38$), a finding based on two studies (Priebe et al. 2007, Woltmann et al. 2011).

Management of daily functioning

Five studies investigated e-health tools and interventions aiming at management of daily functioning. Pijnenborg and colleagues (2010) investigated a mobile phone intervention in which SMS text messages functioned as prompts to remind service users of the goals they had set for themselves when identifying individual needs during a six-week psychoeducation intervention. The goals that service users chose varied from "taking medication," to "relaxing two hours during the afternoon," to "attending a band rehearsal." In a comparable study, Sablier and colleagues (2012) programmed PDAs with prompts to remind service users of their personal schedule of daily activities. Service users could register completed activities and indicate whether they experienced any clinical symptoms. The registered information was sent to the PDA of their caregivers, whose PDA application allowed them to create, modify, and delete date and time of the daily activities of their clients. Sims and colleagues (2012) investigated the effect of SMS text messages as reminders to service users of appointments with their clinician.

Another study, by Ku and colleagues (2007), examined an intervention consisting of conversational training in a virtual environment with avatars. Service users were presented a virtual social situation, displayed on a big screen, in which they had to go through a scenario of greeting others and introducing themselves, starting

the conversation, choosing conversation topics, alternating listening and speaking, and ending the conversation. In the opening scenario, service users approached a group of people sitting around a table, and they had to decide whether or not they could join the group.

Depp and colleagues (2010) described two interventions, one of which is a 24-week telephone-based program aimed at increasing social skills and everyday living. Participants received a 20-minute phone call from a counselor, who discussed various topics, including service users' well-being, emotions, symptoms, specific skills to reinforce previous training, barriers to practicing skills and achieving goals, and reinforcement of achievements. The other intervention Depp and colleagues described was a mobile phone intervention directed at assessment and cognitive-behavioral therapy for three domains, namely auditory hallucinations, medication adherence, and socialization.

Lifestyle management

Two studies could be classified as focusing on lifestyle management. Brunette and colleagues (2011) described a Web-based computer decision support system to encourage service users to quit smoking. The program initially assessed a user's smoking behavior (such as number of cigarettes smoked per day, money spent on tobacco products, and carbon monoxide level) and provided feedback about these measures. Information about the health risks of smoking was presented as an image of the human body with interactive parts. Service users completed exercises that resulted in a summary list of smoking pros and cons, which could be printed out and taken to an appointment with a clinician. Users also were provided an opportunity to discuss matters with a smoking cessation specialist.

Killackey and colleagues (2011) described a running fitness program that is Web based for mobile devices. Two freely available applications can be downloaded to an iPod Touch, namely the Couch-to-5K training application (*The Couch-to-5K Running Plan: C25K Mobile App* 2012) and the Nike+ application (*Nike+ Running App* 2013), which measures running activities through a Nike+ running sensor that is attached to running shoes. Service users participating in the running program are provided with an iPod Touch, and they can track the distance traveled, the duration of each run, and the pace. Furthermore, they have access to a social networking Web site and a Nike+ account, where training progress is displayed.

Peer support

Two studies investigated the use of online peer-support forums for people with a psychotic disorder (Haker et al. 2005, Kaplan et al. 2011). These forums function

as a platform for service users to exchange information and personal experiences with peers, either moderated (Kaplan et al. 2011) or not (Haker et al. 2005). Another study (Gleeson et al. 2012) reported the development of a Web site that integrates therapy modules with a private moderated social networking “cafe.” The e-cafe functions included a personal profile page, a network of friends, a group problem-solving function, and a discussion forum.

Experience sampling monitoring

Myin-Germeys and colleagues (2011) described the development of a PDA-like device called Psymate for monitoring symptoms. The Psymate’s primary focus is self-assessment beyond the clinical setting to aid in the treatment of paranoia, hallucinations, negative symptoms, and other problems.

3.3.2 Cost-effectiveness and user involvement

Only one study included an economic analysis, which showed that costs of e-mental health self-management interventions were higher than expected because of the lack of computers at service users’ homes and the need for transportation to locations with computer facilities (Jones et al. 2001).

Table 3.1 indicates to what extent service users are involved in e-mental health self-management interventions. In almost all interventions described, service users receive feedback on their input, and most interventions or e-health tools are tailored to the individual user. In approximately one-third of the studies, service users were involved in development of the interventions, which were based explicitly on service users’ needs, and the design of the e-health tool could be adapted to their usability needs.

3.4 Discussion

This is the first comprehensive review exploring the area of e-mental healthcare applications for self-management by service users with a psychotic disorder. Results suggest that people with psychotic disorders are able and willing to use e-health services. Whereas two clinical trials required access to the Internet or a mobile phone and some observational studies used a convenience sample, the vast majority of studies had no special requirements for service users’ access to and experience with technological devices. However, attrition rates indicate that this finding should be interpreted with caution. Based on the number of service users enrolled in the study, attrition rates varied from 0% in studies using convenience sampling to 50% in studies with more systematic recruitment strategies. Starting from the total number of

Table 3.1: Types of service user involvement in studies of e-mental health interventions for people with a psychotic illness. Reported items are checked (✓); items that were either not reported or reported in the study as not being included are marked with a dash. NA, not applicable.

| Study | Intervention based on service user needs assessment | Service users involved in development | During intervention service users receive feedback or input | Intervention or system is tailored to the service user | Design adapted to target group |
|------------------------------|---|---------------------------------------|---|--|--------------------------------|
| Beebe et al. (2008) | - | - | ✓ | ✓ | NA |
| Bickmore et al. (2010) | - | - | ✓ | ✓ | ✓ |
| Brunette et al. (2011) | - | - | ✓ | ✓ | ✓ |
| Deegan et al. (2008) | ✓ | ✓ | ✓ | ✓ | ✓ |
| Depp et al. (2010) study 1 | - | - | ✓ | ✓ | - |
| Depp et al. (2010) study 2 | - | ✓ | ✓ | ✓ | ✓ |
| Farrell et al. (2004) | ✓ | ✓ | ✓ | - | ✓ |
| Frangou et al. (2005) | - | - | ✓ | - | - |
| Gleeson et al. (2012) | ✓ | ✓ | ✓ | - | ✓ |
| Haker et al. (2005) | ✓ | - | ✓ | ✓ | - |
| Jones et al. (2001) | - | - | ✓ | ✓ | - |
| Kaplan et al. (2011) | ✓ | - | ✓ | ✓ | - |
| Killackey et al. (2011) | - | - | ✓ | ✓ | - |
| Ku et al. (2007) | - | - | ✓ | - | - |
| Kuosmanen et al. (2009) | ✓ | ✓ | ✓ | ✓ | ✓ |
| Madoff et al. (1996) | - | - | ✓ | - | - |
| Myin-Germeys et al. (2011) | - | - | ✓ | ✓ | ✓ |
| Pijnenborg et al. (2010) | - | ✓ | ✓ | ✓ | - |
| Priebe et al. (2007) | - | - | ✓ | ✓ | - |
| Rotondi et al. (2010) | ✓ | ✓ | ✓ | - | ✓ |
| Sablier et al. (2012) | - | - | - | ✓ | ✓ |
| Sims et al. (2012) | - | - | ✓ | ✓ | - |
| Sherman (1998) | ✓ | ✓ | ✓ | ✓ | - |
| Shrimpton et al. (2008) | - | - | ✓ | ✓ | - |
| Spaniel et al. (2012) | - | - | - | - | - |
| Steinwachs et al. (2011) | ✓ | - | ✓ | ✓ | - |
| Van der Krieke et al. (2012) | ✓ | ✓ | ✓ | ✓ | ✓ |
| Walker et al. (2006) | - | ✓ | ✓ | - | - |
| Woltmann et al. (2011) | - | - | ✓ | ✓ | ✓ |

service users invited, we found that dropout rates varied from 32% to 65%.

3.4.1 Types of e-mental health self-management interventions

Our search found a wide variety of interventions, and this diversity indicates that multiple aspects of self-management are being targeted. A theme that seems to be missing from the existing interventions is that of finding meaning and main-

taining a positive outlook, which service users have indicated is an important component of self-management (Martyn 2002). Future initiatives for self-management interventions may benefit from taking a recovery approach. A logical step may be to transform parts of the illness management and recovery program (Mueser et al. 2002, 2006) into e-mental health interventions.

3.4.2 Evidence base for clinical outcome and cost-effectiveness

The results suggest that e-mental health interventions are at least as effective as standard mental healthcare, according to the effect sizes of individual studies. These studies were predominantly on the right-hand side of the forest plot in Van der Krieke et al. (2014). Summary effect sizes indicate that interventions focusing on medication management and, to a lesser degree, on psychoeducation and on communication and shared decision making are more effective than care as usual or non-technological approaches to mental healthcare. What should be taken into account, however, is that the care-as-usual conditions were not always clearly described. Moreover, in some trials, usual care was compared with usual care plus the intervention, meaning that the technological approaches functioned as a supplement to routine care. In addition, our calculations were based on very few studies.

Although the results need to be interpreted with caution, the fact that none of the studies showed a negative effect seems promising. The results of our study are partly in line with the outcomes reported by Välimäki and colleagues (2012). Their results showed that e-mental health interventions focusing on psychoeducation were as effective as standard care. Furthermore, they reported that technology-based interventions improved medication compliance in the long term. However, the difference in focus and included studies precludes a detailed comparison between our study and that of Välimäki and colleagues (2012).

No conclusions can be drawn about cost-effectiveness of e-mental health self-management interventions, because this aspect barely has been addressed in the studies conducted so far. The one study we found that conducted an economic analysis reported higher costs in the intervention condition because computers were purchased for service users. In some studies, costs were not analyzed, but a reduction of costs seemed very plausible, as in the case of text message reminders that significantly decreased the number of missed appointments with clinicians (Sims et al. 2012).

Lack of evidence can be partly explained by the newness of this field of research. However, some of the usability studies included in our analysis were conducted more than five years ago and have not been followed up by a clinical trial. A reason for this omission may be that e-health projects often entail up-front expenditures

of energy and capital for the design and development of the technological tool, and therefore these projects run the risk of expiring before clinical effectiveness and cost-effectiveness have been investigated. Moreover, conducting RCTs may be particularly challenging in the e-mental health area. Not only are RCTs expensive, but the length of clinical trials may be disproportionate to the rapid developments in the available technology.

Future projects should incorporate clinical and cost-effectiveness analysis in a way that accounts for the dynamic nature of e-mental health interventions. The field may benefit from stepped-wedge research designs or designs that focus on multiple assessments on an individual level. Furthermore, we may need to distinguish between technological interventions that simply computerize existing nondigital methods and innovative interventions. Digital translations of evidence-based nondigital methods are not groundbreaking, but they could be effective in reducing healthcare costs in the short term. Innovative interventions may maximally exploit the opportunities of e-technology, but they may be less likely to reduce costs in the short term.

3.4.3 Orientation of self-management interventions

Service user involvement in e-mental health interventions for self-management appears to be not as self-evident as one might expect. User-centered development is as yet not common practice in this population, and in some interventions the clinical perspective predominates. As a result, e-mental health interventions for self-management do not always contribute to service user empowerment. This is a missed opportunity that developers need to account for.

Future technology will provide means of facilitating more intensive and more accurate monitoring of health and health-related behavior. The development of smart and consumer-priced technological devices enables the move toward an era of personalized medicine and the “quantified self.” Yet, this move can be for better or worse. Schermer (2009) has sketched two possible scenarios: either e-mental health technology will reproduce an outdated paternalistic paradigm of patient-clinician interaction in which compliance and monitoring are the aim (Big Brother scenario), or it will create a new situation that centers on shared decision making and self-management that adds to the autonomy of service users. One way to increase chances for the latter scenario is to involve service users in conceptual and developmental stages of e-mental health interventions.

3.4.4 Limitations

Our review has a number of limitations. The main limitation is the heterogeneity of results, given the broad definition of self-management. First, there was heterogeneity in control groups. Most individuals in the control groups received care as usual—often a nontechnological intervention—but a detailed description of the control condition was lacking in most cases. Furthermore, there was heterogeneity of study quality, and a comprehensive meta-analysis that included all studies was not possible because of heterogeneity of interventions and outcome variables.

Another limitation is that we were not able to systematically assess the quality of the acceptability and feasibility studies. A suitable assessment instrument that was sufficiently flexible and specific to account for the variety in these studies was not available.

Finally, we note that a publication bias is likely to exist in this area of research. Apart from the fact that positive results are more likely to be published than negative results, we suspect that many e-mental health interventions have not been scientifically investigated. The reason for this is that e-mental health approaches are considered not always to be innovative but simply to be easier, more efficient versions of regular approaches that either have already been proven to be evidence based, rendering new research redundant, or are assumed to be effective (comparable with the implementation of consultation by telephone).

This review shows that research into the usability and effectiveness of information and communication technology in self-management interventions for people with psychotic disorders has rapidly increased in the past five years. Our findings indicate that e-health interventions are at least equally effective as standard, non-technology-based care. The greatest potential gain of e-health self-management interventions may be to reduce healthcare costs for service providers as well as service users. To find out whether this assumption is justified, future studies focusing on e-health interventions should include economic analyses.

Parts published as:

A. Emerencia, L. van der Krieke, S. Sytema, N. Petkov, and M. Aiello – “Generating personalized advice for schizophrenia patients,” *Artificial Intelligence in Medicine* (58:1), pp. 23–36, 2013.

A. Emerencia, L. van der Krieke, N. Petkov, M. Aiello – “Assessing Schizophrenia with an Interoperable Architecture,” in *Proceedings of the first International Workshop on Managing Interoperability and Complexity in Health Systems, MIXHS’11*, pp. 79–82, 2011.

Chapter 4

A system for generating personalized advice

In this chapter we present, evaluate, and explain our web application called Wegweis, which can perform an automated explanation and interpretation of ROM (Routine Outcome Monitoring) assessment results. ROM assessments consist of a series of schizophrenia-related questionnaires and lab tests. In the Northern Netherlands, ROM assessments are performed annually for all schizophrenia patients. Wegweis was designed in iterations using feedback from patients and in cooperation with clinicians from all four mental health institutions in the Northern Netherlands (GGZ Drenthe, GGZ Friesland, Lentis, and UCP). Wegweis supports shared decision making by providing patients with their assessment results and an interpretation thereof in the form of personalized advice.

Since not every patient is eager to be confronted with the problems of their illness, Wegweis offers solution-oriented information. In order to make the website attractive for patients, the information is presented in the form of advice, personalized suggestions, helpful tips, and information. The advice consists of information derived from evidence-based research (e.g., the Dutch Multidisciplinary Guideline for Schizophrenia), clinical expertise, and patient experiences. For example, the contents of the advice units range from recommending nearby fitness centers and patient organizations, to providing information about medication side effects and locally available cognitive behavioral therapy modules.

To the best of our knowledge, Wegweis is the first web application that is able to rank information as experienced clinicians do and in a way that is considered helpful by schizophrenia patients, as we show in this chapter. We explain how we designed and implemented an ontology-based approach to reasoning over background knowledge and to determining the applicability and specificity of relevant information for a patient. Ranking information simplifies navigation for a patient, since the most relevant information is likely to be on the first few pages of the re-

sults.

With the availability of Wegweis as a web application, patients can access its information at any time, and without pressure or supervision. Patients should be given access to Wegweis prior to meeting with their clinician. Wegweis encourages patients to bring their own point of view to the discussion, thereby making patient and clinician equal participants in deciding the treatment plan.

The rest of the chapter is organized in the following way: Section 4.1 explains the system design of Wegweis; Section 4.2 explains the user interface; Section 4.3 details the problem ontology; and Section 4.4 presents the algorithm for selecting and ranking advice for a patient. We evaluate the system in the next chapter.

4.1 Wegweis system design

To facilitate its main functionality of generating and showing advice to patients, Wegweis retrieves information from external services and has an interface for experts to manage the advice.

Retrieving information from external services is illustrated in Figure 4.1. This figure shows how Wegweis retrieves patient information and routine outcome monitoring (ROM) data from RoQua, an online questionnaire manager used by mental health institutions in the Northern Netherlands (*RoQua* 2011). RoQua is used by clinicians and interfaces with electronic health records at mental health institutions. Thus, Wegweis interfaces only indirectly with the electronic health records.

Figure 4.1 also shows that patients can view their advice, and that experts can manage the advice units. Patients view advice based on an advice selection and ranking process that uses questionnaire answers, patient information, and a problem ontology. We note that all domain knowledge is isolated in the problem ontology, so the approach used by Wegweis is not necessarily schizophrenia-specific. Wegweis has an interface for experts to manage the advice units. The advice units that we used for our experiments (Section 5.2) are written with an emphasis on keeping the text simple and to the point, and are validated by psychiatrists, psychologists, and patients. The user interface for managing advice units is described in the next section.

Before patients can view their advice, they need to have an account with Wegweis. We created a plug-in for RoQua that allows clinicians to send patients an invitation for Wegweis. Sending an invitation also sends a request to Wegweis to create an account for the patient, and allows Wegweis to retrieve ROM data and patient information for that patient through RoQua. After the invitation is sent, the patient decides whether or not to respond to the invitation. The invitation e-mail links to an account-creation page in Wegweis that is authorized to create an account

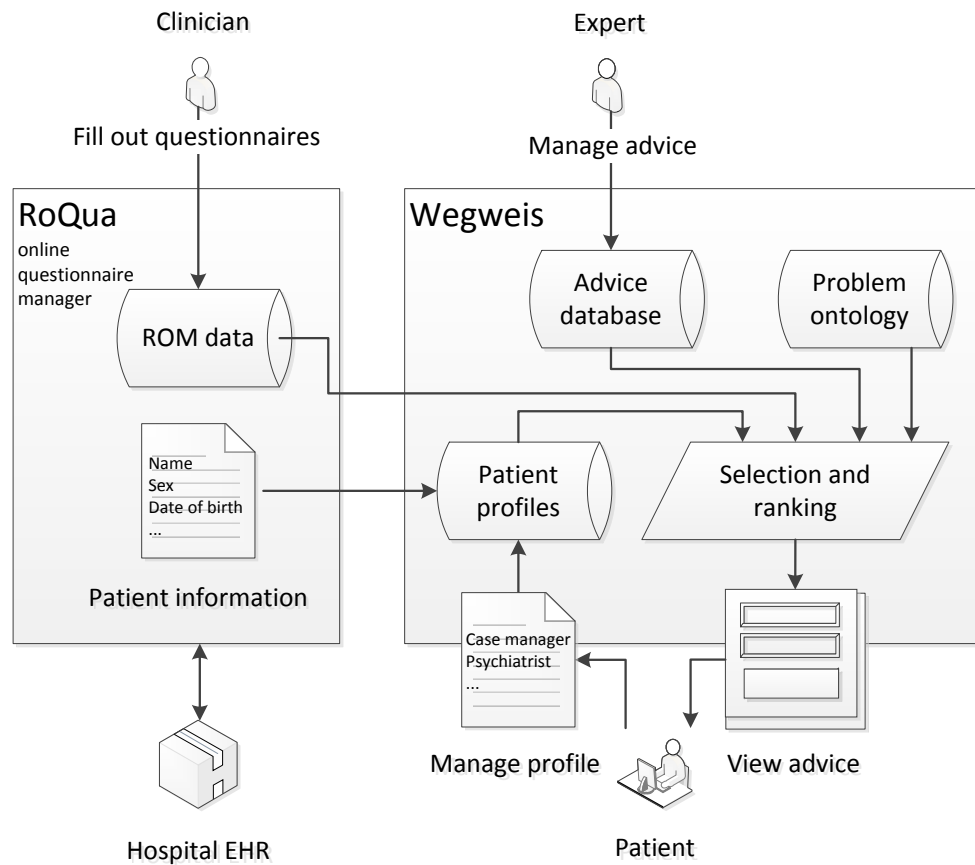


Figure 4.1: Flow of information for selecting and ranking advice in Wegweis.

linked to the information of that particular patient. On the account-creation page, the patient can optionally provide Wegweis with the names of his/her psychiatrist and case manager, which are used to personalize the advice texts. Once the account has been created, the patient is instructed to click on “My Advice” which immediately shows the advice that our system has selected based on the assessment results. In this chapter we explain how our system selects and ranks advice for patients.

4.2 Wegweis user interface

Schizophrenia patients have specific needs regarding the content, structure, and layout of a website (Schrack et al. 2010). They frequently have cognitive problems, such as concentration problems, as a result of the illness and side effects of med-

ication. Rotondi and colleagues (2007) showed that for people with severe mental illnesses, best practices are to keep the navigation simple, to keep words and phrases simple, to avoid having too much text on one page, and to refrain from using flashing or otherwise distracting elements.

We design and implement a way to display advice that respects these limitations. Figure 4.2 shows part of the “My Advice” page, listing the first page of advice for a patient. This page originally contains Dutch text; shown here is a translation. The advice on the page is divided into three sections. We call these sections *advice units*. Each advice unit has a title, in bold, that represents the problem area (e.g., “Is school or work not going so well?”) and two or three solutions, shown in the gray boxes. Note that these solutions are just single lines of text. By clicking these lines, interested readers can open up more information. These expanded contents can again contain collapsed elements. Thus, we gradually show more information to the patient by revealing small chunks of text at a time. This interface was found to be usable by most schizophrenia patients in our usability study (Van der Krieke et al. 2012).

Wegweis employs aspects of *personalization* to appeal to patients. Personalization in web applications can be defined as any action that tailors the web experience to a particular user or set of users (Mobasher et al. 2000). Wegweis implements two levels of personalization in the process of generating advice for patients. First, the selection of advice units and the order in which they are presented depends on the ROM data of a patient, and is therefore personalized. This process of selecting and ranking advice units is part of the main contribution of this chapter, and is explained and evaluated in Sections 4.4 and 5.2. Second, the contents of the advice units can be made to appear more personal by including certain variables. These variables are evaluated at run-time in the context of the patient. For example, when we use the variable `case_manager` or `psychiatrist` in the advice contents, the patients see the actual name of their practitioner instead. This second level of personalization is implemented by simply locating all occurrences of variables and replacing them with the corresponding information from patient profiles.

4.3 Problem ontology

The advice ranking and selection process in Wegweis is based on questionnaire items (i.e., the questions of a questionnaire), which are handled individually. This individual treatment contrasts with the common interpretation of schizophrenia questionnaires. Commonly, schizophrenia questionnaires are interpreted through mean or summation scores of multiple items (Wing et al. 1998, Priebe et al. 1999). We chose to handle each item individually to keep information loss at a minimum,

Advice

Is school or work not going so well?

✎ Are you taking college or university courses and are in need

✎ Is it not going so well at work?

Problems sleeping or relaxing

✎ Do you have problems falling asleep or do you often wake up

✎ Are you very tired or do you sleep a lot?

✎ Do you find it hard to relax?

Do you have physical complaints?

✎ Some physical complaints can be side effects of medication

✎ Try discussing your complaints with your psychiatrist.


 Print advice

Figure 4.2: Part of the “My Advice” page in Wegweis.

on the assumption that each item identifies a distinct problem. Hence, we use the terms “questionnaire item” and “problem” interchangeably.

Our approach for the individual treatment of questionnaire items involves (i) identifying a schizophrenia-related problem for each item and (ii) interpreting the answer as a measurement of the severity of that problem for a patient. This two-step process transforms a filled-out questionnaire into a list of problems and severities. The second step in this process (i.e., interpreting a questionnaire answer as a problem severity) is detailed in the next section, where we show how the list of problems and severities selects and ranks the advice units for patients. The first step (i.e., associating questionnaire items with schizophrenia-related problems) and the problem ontology used therein are explained in the remainder of this section.

Recognizing questionnaire items as individual problems creates 97 problem variables for the four questionnaires that we consider (16 for MANSA (Priebe et al. 1999), 12 for HoNOS (Wing et al. 1998), 24 for CANSAS-P (Trauer et al. 2008), and 45 for OQ-45 (Lambert and Finch 1999)), some of which we found to be very similar. For example, item 11 of the OQ-45 questionnaire is associated with the problem called `AlcoholAbuse`, while item 3 of the HoNOS questionnaire is associated with the problem called `AlcoholOrDrugAbuse`. Since these two problems are semantically similar, it is likely that an advice unit that applies to one of them also applies to the other. Associating an advice unit with problems would be tedious if we had to determine applicability for all problems of all questionnaires manually.

In order to take advantage of the similarities that exist among the problems identified, we created a *problem ontology*, which imposes a hierarchy on the problems and allows us to identify groups of problems with similar semantics. In contrast to the traditional approach of interpreting schizophrenia-related questionnaires (which considers the summation of the severities of a group of related questionnaire items), our approach considers the maximum severity. Thus, in our approach, any individual problem that is severe enough can trigger advice. Hence, we can tailor the advice for a patient, based on individual problems.

The problem ontology decouples the questionnaire items from the advice units and thereby simplifies the process of associating an advice unit with problems. The decoupling is due to the fact that we associate questionnaire items and advice units with problem concepts rather than with each other. The simplification in advice unit association is due to the knowledge stored in the ontology that allows us to associate an advice unit with those problems that represent groups of semantically similar problems, rather than having to determine all applicable problems manually.

In our ontology, the schizophrenia-related problems are the only *concepts* and their hierarchy is the only *relationship*. This relationship, called the *is a* relationship, is a partial order (i.e., relations are reflexive, antisymmetric, and transitive) that de-

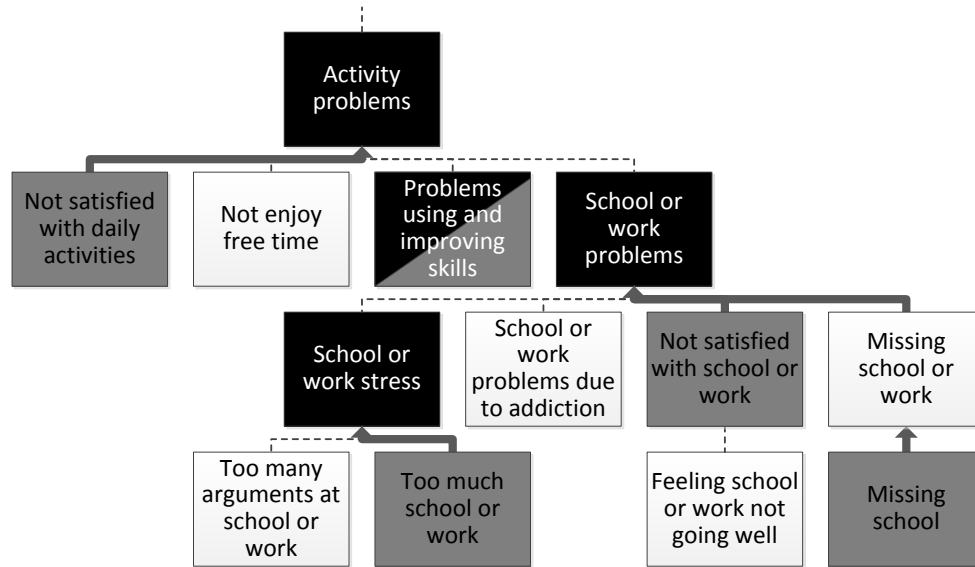


Figure 4.3: Part of the ontology.

notes specificity. Essentially, the inferred relationships form a tree with root node *Problems* that branches out into increasingly specific problems. Thus, every child node is a more specific problem concept of its parent node. For example, in our ontology, the node *Fatigue* has the following ancestors (listed in reverse hierarchical order): *NegativeSymptoms*, *PsychoticProblems*, *PsychicProblems*, and *Problems*. From the properties of our ontology, we deduce that the applicable advice for an active problem concept (i.e., a problem affecting the patient) consists of the advice associated with the problem concept or with any of its ancestors.

In our approach, the ontology is traversed in reverse hierarchical order to find advice in cases where an active problem concept is not associated with any advice units. This process is illustrated in Figure 4.3. This figure shows part of the ontology as a tree with problem concepts as nodes and the *is a* relationship as edges. Furthermore, in this figure, nodes with a black background are associated with advice units, nodes with a gray background are active nodes (i.e., associated with a questionnaire item that was answered above a certain threshold), and nodes with a white background are inactive and can be ignored. We make no distinction between leaf nodes and other nodes, i.e., any node can be associated with advice units, with questionnaire items, or with both. The arrows in Figure 4.3 indicate the paths from active nodes to their first ancestor that is associated with advice and show how advice for certain questionnaire problems is found higher up in the ontology. For

example, advice that is associated with the `School or work problems` node is triggered with the maximum problem severity of the questionnaire items associated with the `Not satisfied with school or work` and `Missing school` nodes. We cover the algorithm for selecting and ranking advice units in more detail in the next section.

We opted for creating a new ontology rather than using an existing one, because we found that existing ontologies did not cover some of the problem concepts that we identified (e.g., problems typically associated not with the patient but with their surroundings). Our idea was that the problem ontology should represent the full spectrum of problems that can affect a schizophrenia patient. The recommended approach for using ontologies in healthcare applications is to use an existing medical ontology such as SNOMED-CT (Stearns et al. 2001). However, we found that existing medical ontologies have no equivalent for some of the identified problem concepts. This is because some of the identified problem concepts are not medical in nature or not associated with the patient. For example, item 2 of the MANSA questionnaire asks whether the patient is satisfied with his/her residence, which in our ontology is associated with the `NotSatisfiedWithResidence` problem concept. This concept has no equivalent in existing medical ontologies, since the problem is not medical in nature and (arguably) not associated with the patient but with his/her residence.

The primary argument for using an existing ontology is to facilitate interoperability (i.e., exchanging data with other systems), which can still be achieved with our approach. In our case, interoperability refers to the importing and exporting of patient summaries. With our custom ontology, we can still achieve interoperability by associating (a subset of) the problem concepts with a standardized ontology, such as SNOMED-CT, in an ontology mapping. With such an ontology mapping, we can use the same algorithms that we designed for finding the most relevant advice to find the most relevant concepts that exist in a standardized ontology, thus allowing for interoperability with other systems that use the same ontology.

We constructed the problem ontology for Wegweis with the help of a psychiatrist and a psychologist. These professionals identified relationships among problem concepts and indicated groups of problems, to which the same advice would apply. We incorporated their assessments into the structure of the problem ontology. This ontology (including the associations with advice units and questionnaire items) was validated by ROM experts and clinicians. They stated that they had studied the ontology and did not find any abnormalities. Furthermore, they noted that the reasoning applied in the hierarchy was sound and made intuitive sense.

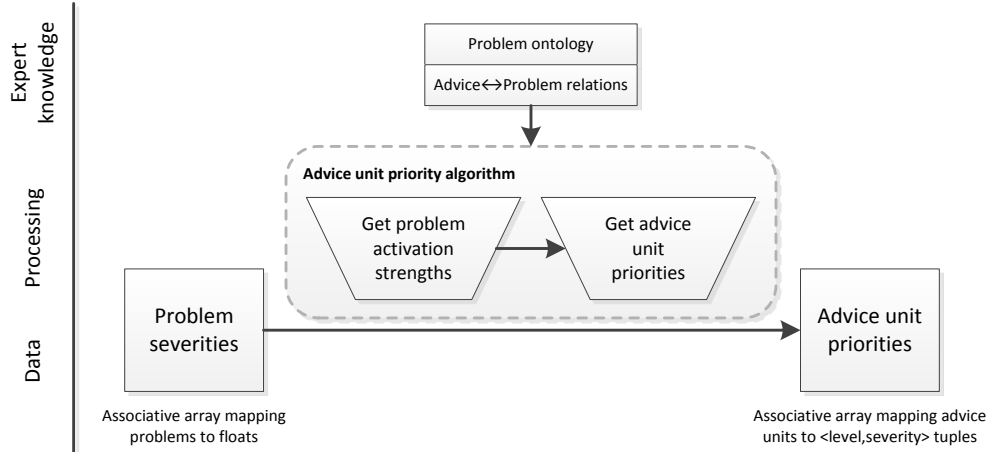


Figure 4.4: An overview of our approach for using problem severities to rank advice units.

4.4 Selecting and ranking advice

Since having too much text on one page can overwhelm the patient (Schrank et al. 2010), Wegweis shows only three advice units per page. Therefore, the order in which these advice units are listed is very important. We let the order of advice units be determined by the inferred severity of the problems associated with them. We use no exclusion criteria for advice, since we consider leaving out key advice more harmful than giving too much advice. In our experiments, we assessed the validity of our approach (see Section 5.2). We first introduced the algorithms for implementing our approach in Emerencia et al. (2011), without an evaluation. Everything about these algorithms, including the design, terminology, and implementation, was done by us.

4.4.1 An algorithmic overview

Figure 4.4 gives an overview of our approach for transforming the answers of a patient for a certain questionnaire into a sorted list of advice units. The *problem severities* shown in the overview are the result of a preprocessing step in which the raw questionnaire answers are normalized. Thus, after the preprocessing step, we have the problem severities for the problem concepts that are associated with the questionnaire items of the filled-out questionnaire. For these problem concepts and for all their ancestors in the ontology, we calculate a similar metric that we call the *activation strength*, which combines problem severity with specificity, as we will explain

in this section. Finally, we convert a list of problem concepts and their activation strengths into a list of advice units and their priorities. We define the *priority* of an advice unit as the maximum activation strength of the problems that are associated with the advice unit. The result is a list of applicable advice units and their priorities. These priorities are then used to sort the applicable advice, and this sorted list of advice units then forms the contents of the “My Advice” pages such as the one shown in Figure 4.2. The remainder of this section describes the above steps in more detail, with the help of pseudocode and a sample run case.

In the preprocessing step of our approach, we convert questionnaire answers into problem severities. We define the term *problem severity* to denote the normalized questionnaire answer such that 0 and 1 denote the least and most severe answer option, respectively, and values for intermediate strata follow from linear interpolation at equidistant intervals. For example, most items of the MANSA questionnaire are rated on a seven-point satisfaction scale, from 1 = “Couldn’t be worse” to 7 = “Couldn’t be better”. Thus, the problem severity corresponding to answer 1 is 1, since it denotes the most severe condition, and analogously the problem severity corresponding to answer 7 is 0. Likewise, an item answered with 2 = “Displeased” translates to a problem severity of ≈ 0.833 . Translating questionnaire answers into problem severities in this way is possible because we found that the schizophrenia questionnaires that we considered had the same structure. In this structure, the questionnaire items relate to some problem or condition, and the answers are an indication of how much the problem affects the patient and are expressed on a rating scale with a certain number of strata. These linear rating scales allow for a straightforward normalization to unit range.

The core of our approach, shown in Figure 4.4, is our *advice unit priority algorithm*, a two-step process that converts problem severities into advice unit priorities. As we explained earlier, the problem severities map problems (associated with questionnaire items) to severities (the normalized questionnaire answers). Our algorithm consists of two steps: (i) calculating the activation strengths and (ii) using the activation strengths to calculate the advice unit priorities. We describe these steps next.

4.4.2 Calculating the activation strengths

In the first step of our *advice unit priority algorithm*, we convert problem severities into activation strengths. We define *activation strengths* as $\langle \text{level}, \text{severity} \rangle$ tuples that are ordered lexicographically by highest level first and by highest severity second. For example, the following list of activation strengths appears sorted in order: $\langle 0, 0.33 \rangle, \langle -1, 0.83 \rangle, \langle -1, 0.44 \rangle$. The activation strength for a problem p is calculated as the maximum augmented activation strength of p and its descendants, where the

augmentation for a descendant q of p consists of decreasing the specificity for every advice unit that applies to q but not to p . For example, imagine that we want to calculate the activation strength of the School or work problems node in Figure 4.3, with the following nodes being active: Missing school with problem severity 0.25, Not satisfied with school or work with problem severity 0.50, and Too much school or work with problem severity 0.75. Now, the activation strengths of these nodes from the point of view of the School or work problems node are $\langle 0, 0.25 \rangle$ for Missing school, $\langle 0, 0.50 \rangle$ for Not satisfied with school or work, and $\langle -1, 0.75 \rangle$ for Too much school or work. The Too much school or work node has a lower level, since there is an advice unit (associated with the School or work stress node) that applies to the Too much school or work node but not to the School or work problems node. Thus, the activation strength of the School or work problems node is $\langle 0, 0.50 \rangle$, which is the maximum augmented activation strength of itself and its descendants, since the tuples are ordered lexicographically by highest level first and by highest severity second.

A description in pseudocode for this step is the GETPROBLEMACTIVATION-STRENGTHS algorithm shown in Algorithm 4.1. This algorithm starts by initializing P to be the set of all problem concepts in the ontology and T to be a mapping of problems to activation strengths, which are initialized as tuples of problem severities with level 0 for the nodes associated with active questionnaire items. In the algorithm, T and A hold intermediate results, while B is eventually returned. The outer loop traverses over all nodes in P by selecting the leaf nodes of P in every iteration and removing them from P afterward. In the inner loop, $T[p]$ is set to the maximum T value of p and its descendants, and if this value is not null, then it is copied to $B[p]$. When all leaf nodes in an iteration have been considered, T and A are updated to account for advice given in the iteration.

The algorithm makes use of the GETLEAFNODES function, which is shown in Algorithm 4.2. This function returns the subset of relative leaf nodes within a given set of nodes P . The relative leaf nodes are the nodes that have no descendant nodes that are in the set P . This definition has a straightforward description in pseudocode. In the pseudocode in Algorithm 4.2, the algorithm iterates over all problems in P and returns those problems whose sets of descendants, according to the ontology, have no elements in common with P .

After each iteration of the outer loop body of GETPROBLEMACTIVATION-STRENGTHS, the levels of the activation strengths are updated by the UPDATEPROBLEMLEVELS algorithm. In the pseudocode of the UPDATEPROBLEMLEVELS algorithm (Algorithm 4.3), we first set U to be the set of all advice units that are associated with active nodes in N . Then, for each advice unit, the algorithm tries to

GETPROBLEMACTIVATIONSTRENGTHS(V)

Input: associative array V mapping problems to problem severities (floats).

Data: ontology functions `all_problems` and `descendants`.

Output: associative array mapping problems to $\langle level, severity \rangle$ tuples, for all triggered problems.

```

 $P \leftarrow \text{all\_problems}()$ 
 $B \leftarrow$  empty associative array
 $T \leftarrow$  empty associative array
 $A \leftarrow$  empty associative array
for each problem  $p \in V.\text{keys}$ 
  do  $T[p] \leftarrow \langle 0, V[p] \rangle$ 
while  $P$  is not empty
   $N \leftarrow \text{GETLEAFNODES}(P)$ 
  for each problem  $p \in N$ 
    for each problem  $q \in \text{descendants}(p)$ 
      do if  $T[q]$ 
      then  $T[p] \leftarrow \max(T[p], T[q])$ 
    do
      if  $T[p]$ 
      then  $B[p] \leftarrow T[p]$ 
      remove  $p$  from  $P$ 
   $T, A \leftarrow \text{UPDATEPROBLEMLEVELS}(N, T, A)$ 
return ( $B$ )

```

Algorithm 4.1: The GetProblemActivationStrengths algorithm.

decrease the level of all problems that the advice unit applies to (i.e., all problems that are associated with the advice unit and all descendants of those problems). Some bookkeeping is done in A to ensure that one advice unit does not decrease the level of a node more than once (which could occur over the span of multiple iterations).

4.4.3 Calculating the advice unit priorities

In the second step of our *advice unit priority algorithm*, we convert activation strengths into advice unit priorities. The advice unit priorities map advice units to $\langle level, severity \rangle$ tuples which, like the activation strengths, are ordered lexicograph-

GETLEAFNODES(P)

Input: set of problems P .

Data: ontology function `descendants`.

Output: the subset of problems that are relative leaf nodes.

$L \leftarrow$ empty set

for each problem $p \in P$

do $\left\{ \begin{array}{l} \text{if } (\text{descendants}(p) \cap P) \text{ is empty} \\ \text{then add } p \text{ to } L \end{array} \right.$

return (L)

Algorithm 4.2: The GetLeafNodes algorithm.

ically by highest level first and by highest severity second. In fact, we define the *priority* of an advice unit as the maximum activation strength of the problems that are associated with the advice unit. The algorithm GETADVICEUNITPRIORITIES, shown in Algorithm 4.4, shows a straightforward description of this definition and returns a mapping of advice units to priorities. These advice units are all the applicable advice units for the patient, based on the questionnaire answers provided, and the priorities are used to order the advice units.

From the algorithms used for our *advice unit priority algorithm*, we deduce that our approach ranks specific advice before generic advice and aims to diversify the top results (i.e., not letting the three advice units on the first page of advice all correspond to the same problem). For every advice unit associated with a problem in N , the UPDATEPROBLEMLEVELS algorithm decreases the level of the activation strengths of all problems that the advice unit applies to. Decreasing the levels of the activation strengths causes the affected problem nodes to have lower activation strengths for triggering advice in later iterations. We assume that the advice selected in later iterations is more generic, since it is associated with problem nodes that are more generic (because we traverse leaf nodes first, and leaf nodes are the most specific nodes according to the hierarchy of the ontology). Thus, by lowering the activation strengths of selected nodes after each iteration, our approach awards the highest rank to the most specific advice for a problem. Moreover, any advice triggered by the same problem in a later iteration is ranked lower than all specific advice (i.e., advice units triggered with an activation strength with level 0), regardless of severity.

Thus far, we assumed that there was one single filled-out questionnaire; how-

UPDATEPROBLEMLEVELS(N, T, A)

Input: set of problems N , associative array T mapping problems to $\langle level, severity \rangle$ tuples, associative array A mapping problems to lists of advice units.

Data: ontology function descendants,
function problems_associated_with,
function advice_associated_with.

Output: updated T and A , where the mappings have been updated to reflect advice given by N .

$U \leftarrow$ empty set

for each problem $p \in N$

do if $T[p]$

then { **for each** advice unit $a \in$ advice_associated_with(p)
 do add a to U }

for each advice unit $u \in U$

for each problem $p \in$ problems_associated_with(u)

for each problem $q \in (\{p\} \cup \text{descendants}(p))$

do if $T[q]$ **and not** $u \in A[q]$

then { $\langle l, s \rangle \leftarrow T[q]$
 $T[q] \leftarrow \langle l - 1, s \rangle$
 $A[q] \leftarrow A[q] \cup \{u\}$ }

return (T, A)

Algorithm 4.3: The UpdateProblemLevels algorithm.

ever, our approach also works for multiple filled-out questionnaires. The only additional complication is that there is a possibility that items of different questionnaires point to the same problem concept in the ontology. If this is the case, we take the (normalized) average of those answers as the problem severity for that problem.

4.4.4 An example run

We now illustrate the operation in pseudocode of our *advice unit priority algorithm* by calculating advice priorities in an example scenario shown in Figure 4.5. The figure shows a subset of the nodes from Figure 4.3, with the addition of an advice unit associated with the School or work stress node. In Figure 4.5, as in Figure 4.3, nodes with a black background are associated with advice units, nodes

GETADVICEUNITPRIORITIES(B)

Input: associative array B mapping problems to $\langle level, severity \rangle$ tuples (i.e., `GETPROBLEMACTIVATIONSTRENGTHS()`).

Data: function `advice_associated_with`.

Output: associative array mapping advice units to $\langle level, severity \rangle$ tuples.

$R \leftarrow$ empty associative array

for each problem $p \in B.keys$

do { **for each** advice unit $a \in \text{advice_associated_with}(p)$

do $R[a] \leftarrow \max(R[a], B[p])$

return (R)

Algorithm 4.4: The GetAdviceUnitPriorities algorithm.

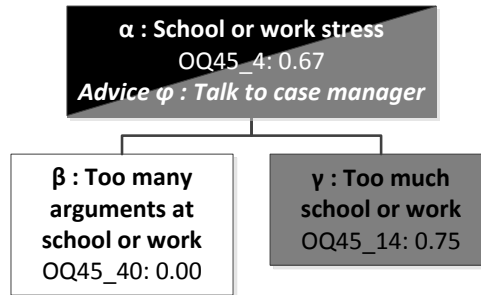


Figure 4.5: An example scenario with three nodes.

with a gray background are active nodes (i.e., associated with a questionnaire item that was answered above a certain threshold), and nodes with a white background are inactive and can be ignored. In this sample run, we refer to the three nodes in Figure 4.5 as α , β , and γ . Each of these nodes is associated with an item of the OQ-45 questionnaire, but only two nodes are considered active. We consider nodes as active only if they have a problem severity above a certain threshold (here we used 0.5). We explain our motivation for using this particular threshold in more detail in the next section. For now, it is sufficient to know that we consider nodes α and γ (with problem severities 0.67 and 0.75, respectively) as active and node β as inactive. Furthermore, note that node α is the only node associated with an advice unit (φ : “Talk to case manager”).

The function `GETPROBLEMACTIVATIONSTRENGTHS` (from Algorithm 4.1) is

called with $V = \{\alpha \Rightarrow 0.67, \gamma \Rightarrow 0.75\}$. The node β is not included in V because it is not considered active. The variable P is initialized to $P = \{\alpha, \beta, \gamma\}$ because it is simply a list of all nodes in the ontology. The variables B , T , and A are initialized to empty associative arrays. The first for-loop sets $T = \{\alpha \Rightarrow \langle 0, 0.67 \rangle, \gamma \Rightarrow \langle 0, 0.75 \rangle\}$.

In the first iteration of the while-loop, we find as leaf nodes $N = \{\beta, \gamma\}$. Since neither of these nodes has descendants, T remains unchanged in the first inner loop. B becomes $\{\gamma \Rightarrow \langle 0, 0.75 \rangle\}$. Note that β is not included in B because β was not included in V . Variables T and A remain unchanged after the call to `UPDATEPROBLEMLEVELS` (from Algorithm 4.3), since none of the nodes in N are associated with advice units.

In the second iteration of the while-loop in `GETPROBLEMACTIVATIONSTRENGTHS`, by having removed β and γ from P , we now find $N = \{\alpha\}$, and T becomes $\{\alpha \Rightarrow \langle 0, 0.75 \rangle, \gamma \Rightarrow \langle 0, 0.75 \rangle\}$, since γ is a descendant of α . These are also the values returned by B . After the second iteration, `UPDATEPROBLEMLEVELS` sets A to $\{\alpha \Rightarrow \varphi, \gamma \Rightarrow \varphi\}$ and T to $\{\alpha \Rightarrow \langle -1, 0.75 \rangle, \gamma \Rightarrow \langle -1, 0.75 \rangle\}$, signifying that an advice unit φ was given that applies to these problems. These values for T would normally be used in future iterations; however, in this example, there are no future iterations, since there are no nodes left in P .

The second step in our approach in Figure 4.4 is to call the function `GETADVICEUNITPRIORITIES` (from Algorithm 4.4) with $B = \{\alpha \Rightarrow \langle 0, 0.75 \rangle, \gamma \Rightarrow \langle 0, 0.75 \rangle\}$. Since the only node associated with an advice unit in our example is node α , and since this node is included in B , we find that this results in $R = \{\varphi \Rightarrow \langle 0, 0.75 \rangle\}$.

Thus, for this sample scenario we find that the list of selected advice units consists of a single advice unit φ triggered with priority $\langle 0, 0.75 \rangle$. The level 0 signifies that the advice unit is the most specific advice unit for a certain problem (`School` or `work stress`, i.e., node α , for which the strength is calculated as the maximum of it and its descendants that are not covered by a more specific advice unit) and that it should be sorted by severity among other level 0 advice units, that is, before any advice units triggered with level -1 or lower. In the next chapter, we validate and test our approach against the opinions of clinicians and patients.

4.5 Implementation

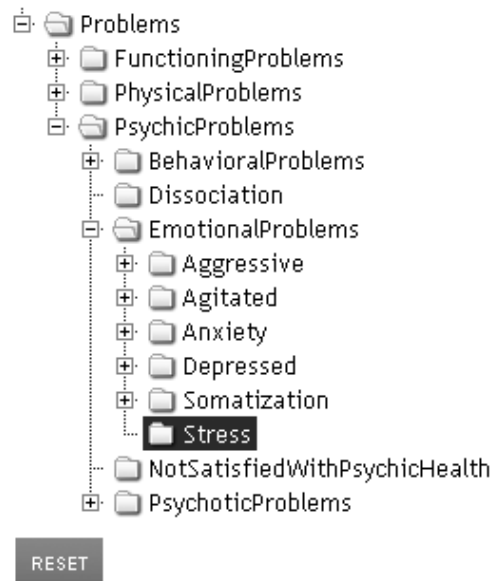
Wegweis is implemented in Ruby on Rails (*Ruby on Rails* 2013), an open source web application framework. It uses a MySQL (*MySQL* 2013) database for storage. In Figure 4.1, RoQua interfaces with the EHRs using HL7, a communications standard used in healthcare applications (Dolin et al. 2006). The communication between RoQua and Wegweis uses JSON (Crockford 2006) over HTTPS. The communication between RoQua and Wegweis is restricted on both ends by IP and a 256-bit shared

secret.

While the interface for managing advice units in Wegweis (shown in Figure 4.6) is based on an existing CMS framework called BrowserCMS (*BrowserCMS* 2011), we implemented additional functionality to facilitate writing advice units. Figure 4.6 shows how the problems that are associated with an advice unit (i.e., the problems that can trigger an advice unit) are selected from a tree view. The advice contents are written in the Liquid templating language (*Liquid Templating Language* 2011). We chose a lightweight templating language, since it allows people without a technical background to easily create HTML content. We extended the Liquid syntax to allow for customized variables (`case_manager` and `psychiatrist`) and scopes (collapsed text, tips, warnings, quotes, and notes). The advice units can embed audio clips, video fragments, as well as other advice units (e.g., when reusing common texts). We also added a live preview with syntax checking for the advice contents, to avoid common errors. Advice units can be added on-the-fly and changes propagated immediately. The advice pages load without noticeable delay, because intermediate stages of the advice unit selection process are cached and embedded content is loaded asynchronously. The implementational details of the staged caching process fall outside the scope of this chapter.

We implemented the problem ontology using Protégé (Gennari et al. 2003) in OWL, the Web Ontology Language (McGuinness and Van Harmelen 2004). Expressed in OWL terminology, the problem concepts are `Classes` and the relationships are defined using `SubClassOf` axioms. The inferred hierarchical structure of the ontology is the result of running the HermiT 1.2.2 Reasoner on the ontology in Protégé. The inferred ontology is exported to an OWL file that is parsed by Wegweis. In addition to the problem concepts and their hierarchy, the ontology also stores the associations between questionnaire items and problem concepts, but it does not store the associations between advice units and problem concepts. Our reasoning for this design is that both the problem concepts and the questionnaire items make sense to domain experts (i.e., they make sense outside the context of Wegweis), while advice units are objects specific to Wegweis. The associations between advice units and ontology concepts are stored in the database of Wegweis. Wegweis identifies ontology concepts by their name and continuously monitors the OWL files to avoid inconsistencies. For example, if a problem concept was removed from the problem ontology, then any advice unit associated with this problem concept should be updated to reflect that it can no longer be activated by said problem concept. In contrast, the associations between questionnaire items and ontology concepts are part of the ontology and are modeled in OWL as `AnnotationAssertion` axioms with questionnaire items represented as `Literals` (e.g., `Mansa_1`, `HoNOS_5`). Our ontology is available online (*Wegweis Ontology* 2011).

Advice problems: click a problem to toggle selection.
 All selected problems are always visible (expanded).
 Questionnaire bindings help.



Liquid content

tip. For more information, ask your {{case_manager}}.

Syntax help.

Preview



For more information, ask your case manager **Kees Jansen**.

SAVE AND PUBLISH

SAVE

Figure 4.6: The expert interface for adding an advice unit.

4.6 Discussion

We have presented the development and design of Wegweis, a patient-centered web application driven by an ontology-based approach that uses ROM assessment results to select and rank advice for schizophrenia patients. The system has minimal impact on the way clinicians work, because it integrates with an existing questionnaire manager. Adding support for a questionnaire in Wegweis is simplified by the fact that questionnaires are decoupled from advice by virtue of the problem ontology. Background knowledge, embedded in the structure of the ontology, is used to infer advice when no exact match is found, which adds to the robustness of the system.

We believe that Wegweis can be a helpful addition in improving patient care. The improvement is due to two reasons. First, an automated explanation and interpretation of assessment results empowers the patient because it allows patients to prepare for discussing their treatment plan without requiring any help. Second, where clinicians may forget to mention or choose to ignore certain alternatives, an automated approach presents the patient with all the options it knows about and leaves the decision up to the patient. We conclude that a system such as Wegweis can work as a useful adjunct to the care of schizophrenia patients in the form of a *second perspective*: unbiased advice that is ordered in a way that has high similarity to what a clinician would discuss, given the same questionnaire data.

The approach we used for selecting and ranking advice can be used to enhance self-management websites for other chronic illnesses as well. Since all domain knowledge is stored in the ontology, the approach lends itself to providing personalized advice in other areas of healthcare. However, an advice system relies heavily on the domain-specific problem ontology and on the advice contents. Moreover, its performance is very dependent on the specific questionnaires. Thus, porting the approach to other areas of healthcare would not be a trivial task. A new ontology would have to be built, based on disease-specific questionnaires and terms, and a new body of advice contents would have to be collected and validated by experts.

Parts published as:

A. Emerencia, L. van der Krieke, S. Sytema, N. Petkov, and M. Aiello – “Generating personalized advice for schizophrenia patients,” *Artificial Intelligence in Medicine* (58:1), pp. 23–36, 2013.

L. van der Krieke, A. Emerencia, M. Aiello, and S. Sytema – “Usability Evaluation of a Web-Based Support System for People With a Schizophrenia Diagnosis,” *Journal of Medical Internet Research* (14:1), pp. e24, 2012.

Chapter 5

Evaluation of Wegweis

Routine Outcome Monitoring (ROM) is a systematic way of assessing service users’ health conditions for the purpose of improving their care. ROM consists of various measures used to assess a service user’s physical, psychological, and social condition. While ROM is becoming increasingly important in the mental healthcare sector, one of its weaknesses is that it is not always sufficiently service user-oriented. First, clinicians tend to concentrate on those ROM results that provide information about clinical symptoms and functioning, whereas it has been suggested that a service user-oriented approach needs to focus on personal recovery. Second, service users have limited access to ROM results and they are often not equipped to interpret them. These problems need to be addressed, as access to resources and the opportunity to share decision making has been indicated as a prerequisite for service users to become a more equal partner in communication with their clinicians. Furthermore, shared decision making has been shown to improve the therapeutic alliance and to lead to better care.

5.1 Usability Evaluation

Our aim is to build a web-based support system which makes ROM results more accessible to service users and to provide them with more concrete and personalized information about their functioning (e.g., symptoms, housing, social contacts) that they can use to discuss treatment options with their clinician. In this study, we report on the usability of the web-based support system for service users with schizophrenia.

First, we developed a prototype of a web-based support system in a multidisciplinary project team, including end-users, as described in the previous chapter. We then conducted a usability study of the support system consisting of (1) a heuristic

evaluation, (2) a qualitative evaluation and (3) a quantitative evaluation.

Fifteen service users with a schizophrenia diagnosis and four information and communication technology (ICT) experts participated in the study. The results show that people with a schizophrenia diagnosis were able to use the support system easily. Furthermore, the content of the advice generated by the support system was considered meaningful and supportive.

This study shows that the support system prototype has valuable potential to improve the ROM practice and it is worthwhile to further develop it into a more mature system. Furthermore, the results add to prior research into web applications for people with psychotic disorders, in that it shows that this group of end users can work with web-based and computer-based systems, despite the cognitive problems these people experience.

Although there is no universal definition, ROM can be described as the use of standard instruments to systematically and continuously assess aspects of mental health service users' health for the purpose of better aiding their care (Trauer 2010). The format of ROM varies between countries, but it usually consists of several quantitative measures used to assess a service user's physical, psychological, and social condition. ROM is carried out for service users with a single diagnosis and short-term problems as well as for people with a severe mental illness. This latter group includes service users diagnosed with schizophrenia.

The effects of ROM on mental healthcare have had mixed success. On the one hand, research shows that the use of outcome measures, combined with adequate feedback, helps clinicians to recognize and anticipate problems in individual treatment processes and to provide better care as a result (Lambert et al. 2001, 2005, Whipple and Lambert 2011). On the other hand, ROM is not always used in a way that empowers service users and improves shared decision making between service user and clinician (Lakeman 2004, Guthrie et al. 2008). One problem is that clinicians tend to concentrate on those ROM results that provide information about clinical symptoms and functioning. However, service user-oriented approaches promote a focus on personal recovery, which reflects the importance of finding meaning and giving value to personal experiences (Lakeman 2004). A second problem is that service users have limited access to ROM results and they are often not equipped to interpret them (Guthrie et al. 2008, Happell 2008). These problems need to be addressed, as research has shown that access to resources and the opportunity to share decision making has been indicated as a prerequisite for service users to become a more equal partner in communication with their clinicians (GGZ Nederland 2009, Deegan 1997). Furthermore, shared decision making has been shown to improve the therapeutic alliance, and to lead to better care and treatment (Mahone et al. 2011, Frank and Gunderson 1990).

Since 2007, ROM assessments have been a regular element in care for people with psychotic disorders in the northern provinces of the Netherlands. The ROM protocol (called PHAMOUS), which is specifically developed for psychotic disorders, consists of a physical investigation (e.g., weight, height, waist measurement, and glucose levels), multiple interviews and questionnaires concerning psychiatric and psychosocial issues, and service user satisfaction (*PHAMOUS. Pharmacotherapy monitoring and outcome survey* 2011). All service users with schizophrenia who receive care from any mental healthcare organization involved take part in ROM assessment at least once a year. After completion of the assessment, the parameters of the ROM assessment are uploaded into a central database by clinicians and research nurses via a link in the patient's electronic file. Currently, the ROM-results are only reported to clinicians. Clinicians are supposed to discuss the results with their patients so that they can mutually decide whether the course of treatment needs readjustment (Makkink and Kits 2011). However, a large percentage of service users do not receive adequate feedback concerning their ROM-results, as clinicians are not yet accustomed to discussing ROM results with service users (Schaefer et al. 2011).

In an attempt to improve ROM practice and to increase potential for service user empowerment, we developed a prototype of a web-based support system that provides service users diagnosed with schizophrenia with personalized advice, based on their ROM results. By means of this support system, the current problems with ROM practice may be partly tackled. The personalized advice provides users with accessible information about their ROM results, which may enable them to participate in shared decision making, and pave the way to better care. Prior research has shown that people with psychotic disorders can work with web-based and computer-based systems, despite the severity of their symptoms, e.g. (Schrang et al. 2010, Kuosmanen et al. 2010, Jones et al. 2001, Rotondi et al. 2010, Bickmore et al. 2010). Findings are, however, inconsistent as to the amount of support service users need in working with computers (e.g., (Kuosmanen et al. 2010) versus (Bickmore et al. 2010)).

In the present study, we extended the existing research by investigating the usability of a web-based support system for ROM. We examined whether our support system can make ROM-results more accessible to service users and provide them with more concrete information that they can use to discuss their personal goals with their clinician. The aim of this section is to provide a brief overview of the web-based system and to report on its usability from the perspective of service users with schizophrenia.

5.1.1 Methods

Implementation

The prototype of the web-based support system is called Wegweis, which is a Dutch abbreviation that stands for web environment for empowerment and individual advice. The Wegweis support system offers users advice about various topics related to psychiatric treatment, rehabilitation, and personal recovery. This advice is based on the service user's ROM assessment results, as conducted in the northern provinces of the Netherlands. The support system is a website, which can be accessed by entering a username and a password. The system is to be used by service users at home or in a clinical setting (e.g., a community hospital).

When building the prototype, we focused on two important and widely used ROM measures, namely the clinician-rated Health of the Nation Outcome Scale (HoNOS) (Wing et al. 1998), which measures health and social functioning, and the service user-rated Manchester Short Assessment of Quality of Life (MANSA) (Priebe et al. 1999), which measures quality of life. Based on item scores of these measures and using innovative algorithms combined with ontological reasoning, the system identifies specific healthcare problems for each individual service user and provides relevant and tailored advice (Chapter 4). The algorithms are innovative because they break with conventional case-based reasoning approaches in that they decouple symptoms from outcomes, allowing the outcomes to be dynamic. The content of the advice consists of information derived from evidence-based research (e.g., the Dutch Multidisciplinary Guideline for Schizophrenia), clinical expertise, and service user experiences.

When, for example, the ROM results indicate that a service user is experiencing physical problems, the system offers advice indicating that physical problems can be a side effect of medication, referring to the Dutch Multidisciplinary guideline for schizophrenia. Furthermore, the advice suggests that side effects may be resolved by adjustment of the medication. Service users are also referred to their psychiatrist – by name – for more information. When service users appear to experience problems with personal safety, they are provided information about and linked to the local patient counselor. They also have the opportunity to read about experiences of other service users. In another example, service users who are troubled by hearing voices are provided a video showing someone suffering from the same condition and offering information about treatment options. More information about the advice can be found in Van der Krieke et al. (2011). The algorithm for advice selection, as well as a brief overview of system design and architecture are presented in Chapter 4.

The prototype is created with open source software, using the Ruby on Rails

Web-framework. The website uses secure connections for all traffic. Service users can access their ROM-results by logging in with a username and password, which they can create by clicking on the link sent to them in an invitation email. Failed login attempts are logged by the system. ROM-results can only be accessed via patient accounts.

Development of the prototype

The prototype of the web-based support system was developed by a multidisciplinary team of computer, social, and medical scientists in close collaboration with a group of service users with a schizophrenia spectrum disorder. The content and functionality of the first prototype was based on a needs assessment (unpublished material) conducted in 2009, consisting of semi-structured interviews with service users, relatives of service users, nurses, psychologists, psychiatrists, and people involved in e-mental health services for people with a psychiatric disability.

We put particular focus on the design of the support system's user interface, as it has been suggested that people with schizophrenia have special needs with regard to web design (Rotondi et al. 2007). This is supported by the theory that the quality of a user interface is partly determined by the extent to which users are able to create a so-called mental model of the website. A mental model can be described as a representation of a person's thought processes regarding the functionality and structure of the website, and the flow of information therein. Therefore, it is important for designers to match as closely as possible the user interface with this mental model (Cooper and Reimann 2003). Finding a good match can be particularly challenging. This is especially the case when dealing with people with schizophrenia, who experience cognitive problems such as concentration, memory, and information processing difficulties (Rotondi et al. 2007). As a result, their mental models may differ from those of other users.

A few studies have investigated the challenges in web design for people with a schizophrenia diagnosis. Results from these studies suggest that users with schizophrenia experience difficulties with stimulus overflow, large amounts of text or information, interpretation of two-word labels, and remembering previous steps in the navigation process (Schränk et al. 2010, Kuosmanen et al. 2010, Rotondi et al. 2007, Välimäki et al. 2008). Furthermore, some of them experience paranoia when using computers and Internet (Schränk et al. 2010).

In conjunction with the general guidelines as described in *User Interfaces for all* (a handbook for user interface design) (Stephanidis 2001) and taking into account the findings from prior research, we set out some specific rules for the design of the support system's interface. The most important of these specific rules were the following: no use of unexpected pop-ups, transparency of procedures (i.e., clear

information about what happens when users click a button, what purposes their personal information is used for and who it is available to, etc), use of concrete descriptions (including using the name of a service user's psychiatrist, instead of the general designation 'your psychiatrist'), limited amount of text on one screen with an option to increase/decrease the amount of information, use of video material in addition to text, limited number of bright colors and avoiding jargon or difficult terms.

Participants

Service users were recruited from four mental healthcare organizations in the Netherlands through snowball sampling. Snowball sampling involves asking a key informant or study participant whether they can suggest a person who fits the study criteria and asking them to introduce this person to the researcher (Hennink et al. 2011, pp. 81–107). In our case, study participants were recruited by 5 clinicians and fellow study participants. The study was conducted in March and April 2011. The inclusion criteria were (1) having a diagnosis of schizophrenia or a related psychotic disorder (e.g., schizo-affective disorder, schizophreniform disorder, schizotypal disorder), (2) being between 18 and 65 years old and (3) being fluent in Dutch. There were no exclusion criteria.

Sixteen service users were asked to participate and a total of 15 service users, 10 male and 5 female, agreed to participate in the study. The age of the participating service users ranged from 23 to 61 years, with a mean age of 42. The duration of illness for 13 of these service users was known and ranged from 3 to 25 years, with a mean duration of 13 years. All service users received care in an outpatient setting except for one, who was committed in a forensic setting. In order to provide participants with some time to consider their participation, they were informed about the purpose and content of the testing by either a clinician or one of the experimenters at least a week prior to testing. Directly before the usability testing was to start, written informed consent was obtained. After completing the study, participants received a gift voucher of 15 euros.

Four Information and Communication Technology ICT experts participated in the study. They fulfilled the role of evaluator in a heuristic evaluation process, as described below. All ICT experts were employed at the UMCG and experienced in developing ICT applications for mental healthcare organizations.

Usability Testing

Usability can be defined as the ease with which users can use a particular tool or object to achieve a specific goal. Nielsen distinguishes five main quality components

of usability (Nielsen 1993): (1) *learnability*: how easy is it for users to accomplish basic tasks the first time they encounter the design; (2) *efficiency*: once users have learned the design, how quickly can they perform tasks; (3) *memorability*: when users return to the design after a period of not using it, how easily can they re-establish proficiency; (4) *errors*: how many errors do users make, how severe are these errors, and how easily can they recover from the errors; and (5) *satisfaction*: how pleasant is it to use the design.

Usability can be assessed by usability testing. There are three testing categories: heuristic evaluation, qualitative evaluation, and quantitative evaluation. These categories are described in the following sections.

Heuristic Evaluation

We started the usability testing by conducting a heuristic evaluation. This is a research method for detecting usability problems with the interface early in the testing process (Nielsen 1993). Heuristic evaluation is conducted by evaluators and takes place prior to the testing by end-users (in our case service users). Problems detected by the evaluators are dealt with immediately so they do not influence the rest of the testing process.

Heuristic evaluation is usually conducted by more than one evaluator because it is difficult for one person to detect all usability problems. We appointed four ICT experts to fulfill the role of the evaluator, as this falls into the range of the optimal number (Nielsen and Landauer 1993). The process of heuristic evaluation used in this study is based on Nielsen's recommendations (Nielsen 1994). The evaluators were given a brief introduction to the background and rationale of the web application under review, then given instructions on how to conduct the heuristic evaluation. One of the most important instructions was that they were not allowed to communicate with each other during the testing process. Then, the evaluators sat at the computer and went through the user interface according to a scenario written by the experimenters. The scenario included using log-in procedures, username and password retrieval processes, font size modification, completing questions, going through advice units, printing information, searching for advice by means of key words, and providing feedback about the website. The evaluators inspected the interface independently, assessing the various elements based on a list of ten recognized usability principles ("heuristics") translated into a series of questions (see Table 5.1). Their findings were put in a template developed by the experimenters.

The data in the four completed templates was assembled in one document and its content was analyzed, meaning that the data was categorized according to Nielsen's usability topics (see also Table 5.1). Finally, a list of usability violations was created and sorted according to frequency and priority. A debriefing meeting was

organized with evaluators and the experimenters, during which the results of the heuristic evaluation were discussed during a brainstorm session. Decisions were made as to which usability issues were considered most urgent and how these issues could best be solved.

Qualitative Evaluation

After completion of the heuristic evaluation, we conducted a qualitative evaluation. In this process, end-users fulfilled the role of the evaluator. The participants were invited to sit at a computer. We then asked them to use the web application following a scenario written by the experimenters (the same scenario as used in the heuristic evaluation). Users were encouraged to work through the scenario step by step, starting with the log-in procedures. We decided not to ask participants to think aloud, as we suspected that this might affect their way of working substantially.

Two-thirds of the end-user participants carried out the testing at our research center. During the testing, one of the experimenters observed the users' actions via a beamer projection on a screen, while making notes. One-third of the users conducted the testing at home on their own computer and were joined by an experimenter who observed from a distance. When users finished the testing, they were asked to verbally describe their first impression of the support system.

As the main aim of this part of the testing was to find out how users interact with the web system, the research method used in this qualitative evaluation was (non-participant) observation (Hennink et al. 2011, pp. 169–200). One experimenter was present during the testing session and made notes (using paper and pencil) which indicated how participants worked their way through the scenario. The sessions were not audiotaped, as observation was the main evaluation method and we assumed that participants might not feel at ease with audiotaping. The verbal information provided by service users was analyzed by identifying positive and negative feedback items.

Quantitative Evaluation

After the qualitative evaluation was completed, a quantitative evaluation was conducted. End-user participants were asked to fill out a short questionnaire, consisting of 5 questions measured on a 5-point Likert scale. They were asked about their computer and Internet use. This questionnaire was derived from another European study testing a web application developed for a comparable group of end-users (Kuosmanen et al. 2010). Furthermore, participants completed a Satisfaction Questionnaire, measuring their satisfaction with various aspects of the web application concerning layout, structure, user-friendliness and content. This questionnaire consisted of 13 statements to be subsequently rated on a 7-point Likert scale, ranging

Table 5.1: Assessment Criteria for Heuristic Evaluation.

| Usability principle | Question |
|--|--|
| 1. Visibility of system status | Are there any incidents where the website is unresponsive or slow? |
| 2. Match between system and the real world | Are there any words/sentences used on the website that do not match the language used by the intended group of users? |
| 3. User control and freedom | Are there any instances where important changes made by users cannot be easily undone? |
| 4. Consistency and standards | Are there any inconsistencies concerning language use or functionality? |
| 5. Error prevention | Are there any instances where users can easily make mistakes? Before executing an action, are users asked for confirmation where needed? |
| 6. Recognition rather than recall | Are there any pages where the content or structure is unclear or insufficiently explained? |
| 7. Flexibility and efficiency of use | Are there any frequently used functionalities on the website that are not accessible fast enough? |
| 8. Aesthetic and minimalist design | Are there any instances in which the website offers too much information, whereby the user can lose track of the situation? |
| 9. Help users recognize, diagnose, and recover from errors | Are there any error alerts which are not clear to users, which do not identify the problem correctly or do not provide a solution? |
| 10. Help and documentation | Is there enough help or documentation available? |

from completely disagree (1) to completely agree (7). The Satisfaction Questionnaire was specifically designed for this study by the research group. Descriptive analysis

(mean, standard deviation) of the quantitative data was conducted with SPSS 16.0 statistical software for Windows (SPSS Inc., Chicago, IL, USA).

5.1.2 Results

The results of the usability tests are a combination of the three categories of testing mentioned above, namely heuristic evaluation, qualitative evaluation, and quantitative evaluation.

Heuristic Evaluation

All ICT experts evaluating the website were able to complete the scenario written by the experimenters. No major problems were reported with regards to language, undoing changes, structure or content of the pages, accessibility of functionality and clarity of error messages (i.e., usability principles 2, 3, 4, 6, 7 and 9). However, there were some instances in which the website was unresponsive or slow. Furthermore, at times the website seemed to offer too much information at once, and three situations occurred whereby users were not clearly directed to the right page. The most obvious problem reported was that the Disclaimer page was empty and that there was no existing Help section or Frequently Asked Questions section.

During the debriefing meeting, all problems were discussed and decisions were made on how to solve problems most effectively. All problems were solved prior to the qualitative and quantitative testing with service users, except for the missing Frequently Asked Questions section, which was composed after the usability testing with service users.

Qualitative Evaluation

All end-user participants were able to complete the scenario, although three of them needed some hints in order to continue to the next step. For instance, one participant had difficulty finding out how to adjust his personal profile, and the experimenter had to explain how he could access the profile. Although the participants were not asked to think aloud during the evaluation, most of them did so spontaneously. One of the difficulties expressed was that some buttons were hard to find or that their function was not entirely clear. One example is the 'Feedback' button. This button was located at the left part of the web page, situated vertically and separately from the Navigation Bar. Three participants could not immediately locate it and two did not know what to use it for. Furthermore, several participants suggested that the website could be made more attractive by using more color, more images and videos, and more links. However, others indicated they were happy with the layout and found the website to be nice and simple.

With reference to the content of the website, participants expressed that they recognized many issues that people suffering from schizophrenia are faced with and believed that the website could be a useful instrument in supporting people in their personal recovery process. In addition, while reading the advice, various service users came up with relevant information that they thought should be added to the advice. A few other participants, however, stated that the information about illness symptoms and medication should be more extensive. In addition, one participant suggested creating a possibility for online communication between clinicians and service users within the system.

Quantitative Evaluation

The participating end-users reported to be well experienced in using computers and the Internet, to have good computer and Internet skills (see Table 5.2), and to have a positive attitude towards technology (see Table 5.3). There was one participant who reported to have almost never used the Internet. He appears not to have access to the Internet due to the fact that he was a forensic service user admitted into a penitentiary where Internet use was not allowed.

The mean score of satisfaction with the web-based support system prototype was 73.60 (the maximum being 90) with a standard deviation of 6.64. Ratings of the individual statements are presented in Table 5.4. As this table shows, the most disagreement amongst the participants concerned the question of whether or not the website was boring. This is in line with the results of the qualitative analysis, which showed that some participants found the website nice and quiet, whereas others suggested that it could be improved by using more color, images, and so on.

Table 5.2: Service Users' Computer/Internet Use and Skills

| | Almost never | Less than once a month | Monthly | Every week | Every day |
|--------------|--------------|---------------------------|---------|------------|-----------|
| Computer use | 0 | 0 | 0 | 1 | 14 |
| Internet use | 1 | 0 | 0 | 1 | 13 |

5.1.3 Discussion

In this study, we investigated the usability of the first prototype of a web-based support system for people diagnosed with schizophrenia. The heuristic evaluation with ICT experts revealed some minor problems; the most important ones of which were (i) the processing of information being slow and unresponsive; (ii) too much information being displayed at once; (iii) an empty Disclaimer page; and (iv) no ex-

Table 5.3: Service Users' Attitude Towards Computers

| | Very bad | Bad | Not bad, not good | Good | Very good |
|----------------------------|---------------|----------|-------------------|----------|---------------|
| Computer skills | 1 | 0 | 5 | 8 | 1 |
| Internet skills | 1 | 0 | 4 | 9 | 1 |
| | Very negative | Negative | Neutral | Positive | Very positive |
| Attitude towards computers | 0 | 0 | 0 | 11 | 4 |

isting Help section. The first three problems were solved before testing with service users. During qualitative testing, our group of end-users reported some difficulties with, among other things, the location and function of the 'Feedback' button and with understanding how to adjust one's personal profile. In addition, several suggestions were made to make the interface more attractive. These results indicate that the end-users involved in this study, varying in age, sex, and duration of illness, were able to use the support system easily. Furthermore, the content of the advice generated by the support system was judged to be meaningful and supportive. We can therefore conclude that, overall, the support prototype has valuable potential for improving the ROM practice and that it is worthwhile to develop it further into a more mature system.

Related work

Our preliminary results are in line with previous research, which shows that people with psychotic disorders can work with web-based and computer-based systems (Schrack et al. 2010, Kuosmanen et al. 2010, Jones et al. 2001, Rotondi et al. 2010, Bickmore et al. 2010), but there are some differences between our research and that of others that we need to address.

Whilst designing the interface, we followed some specific rules based on existing literature in the field and for this group of end-users as well as applying general rules of interface design. However, we did not comply with all recommendations presented in the literature as feedback from individual service users during the design process, which took place prior to the usability testing (not described in this chapter), suggested it might not be necessary. For instance, we decided to use a bright background color (yellow) for the web pages, and we used arrow heads and drop down menus instead of pop-ups, which was advised against by Rotondi et al. (2010). However, these deviations did not result in any usability violations.

This may be explained by the fact that there appears to be a difference between

Table 5.4: Results of the satisfaction questionnaire

| | Mean (<i>sd</i>) | Percentage (%) of service users who agreed (score 6) or completely agreed (score 7) with the statement and (<i>N</i>) |
|---|--------------------|---|
| I can easily find my way on the website. | 5.73 (0.88) | 80 (12) |
| I am satisfied with the language used on the website. | 6.13 (0.35) | 100 (15) |
| The website is boring. | 3.13 (1.55) | 7 (1) |
| I am satisfied with the font used on the website. | 5.87 (0.83) | 93 (14) |
| The color of the website was appealing. | 5.33 (1.35) | 67 (10) |
| The website does not contain distracting elements. | 5.8 (1.21) | 80 (12) |
| The advice provides me with meaningful information. | 5.67 (0.72) | 80 (12) |
| The amount of information in the advice is too much. | 2.87 (1.55) | 7 (1) |
| The advice can help me reflect on what I want. | 5.73 (1.16) | 80 (12) |
| I can imagine myself discussing the advice with my clinician in the future. | 5.67 (1.11) | 80 (12) |
| I can imagine the advice being helpful to others. | 6.27 (0.46) | 100 (15) |
| I think I will use the website in the future. | 5.53 (0.83) | 60 (9) |
| I would recommend the website to others. | 5.87 (0.64) | 86 (13) |

basic principles for user interface design and concrete applications thereof. Each basic principle can be translated into various concrete applications. If the principle is to avoid an abundance of information, this can be achieved by either limiting the amount of text on one page, or by ordering the information in a surveyable way. Both forms can be effective, depending on, among other things, users' individual preferences. Furthermore, as the functionality of Internet browsers develops very

quickly and new innovations emerge, some earlier problems with the user interface may be no longer relevant. For instance, Rotondi et al. (2010) discourage the use of an absolute font size that cannot be enlarged. Given the flexibility of modern-day browsers, however, this is hardly an issue anymore, as font sizes can be adjusted rather easily.

Another issue to be addressed is the context for which the support system is developed. As mentioned before, our system is intended for independent use by service users at their home or on a hospital ward. This is in line with the study by Bickmore et al. (2010), who developed a computer-based medication adherence system with relational agents for service users with schizophrenia, to be used at home and without assistance or interpretation from clinicians. Results of their pilot evaluation study ($N = 16$) show that independent use of the computer system was acceptable for all but one of the study participants, who were recruited at an outpatient clinic. However, these results seem to contradict with the findings of Kuosmanen et al. (2010), who reported that service users with psychotic symptoms needed support from nurses in using their web system. This difference in findings could be explained by symptom severity of service users, as the study by Kuosmanen et al. (2010) was conducted in a locked-door setting, while the one by Bickmore et al. (2010) and our study primarily involved service users staying at home.

The results of our study add to previous studies in that usability tests suggest that there need not be insurmountable barriers in independent use of web-based systems for people with psychotic disorders. However, we need to investigate the system in a real world setting in order to draw broader conclusions. In future research, the most important question will be not so much whether or not service users with psychotic symptoms can independently work with web systems, but rather, under what conditions they can successfully work with them. These conditions may depend upon the service users' circumstances, such as receiving care in an inpatient or outpatient setting, severity of specific symptoms (e.g., paranoid ideas), and, of course, the level of computer experience. In addition, they might also be related to the web-system, such as the content and the complexity of the system's functionality.

Limitations

Our study should be viewed with consideration of certain limitations that we encountered. First, our sample of service users was small and we used a method of snowball sampling, which is a form of convenience sampling. One disadvantage of convenience sampling is that one runs the risk of compiling a non-representative study sample. In our case, the study sample was quite diverse in age, sex, and duration of illness, which favors the sample's representativeness.

In contrast, what appears to be less favorable for the sample's representativeness is the fact that the service users recruited for this study might have had a particular interest in working with computers and websites, which could have affected our results. This could be the case given that the service users concerned were reported to be quite skilled in using the computer and Internet. However, we need to take into account that the Netherlands is one of the countries with the highest Internet penetration rates. In March 2011, 88.3% of the Dutch population had Internet access, while the world wide average is only 30.2% (*Internet World Stats. Top 58 countries with highest penetration rates 2011*). This suggests that skillful computer and Internet use is not uncommon in the Netherlands. Understandably, there are differences between the level of computer and Internet skills of the general Dutch population and people with mental disorders. However, we believe that the representativeness of our sample on this point does not necessarily invalidate our conclusions.

Second, the presence of an experimenter during the testing session may have affected the behavior of service users conducting the testing. Although the experimenter encouraged participants to mention both strong and weak features of the web application, they might have felt reluctant to be critical.

Third, the support system was not tested in the context of a full ROM assessment, but as a somewhat isolated part thereof. Therefore, at the moment, we cannot gain a comprehensive view of the system's functioning in its full setting. This issue needs to be addressed in future research in a clinical evaluation, followed by an examination of its effectiveness in a randomized controlled trial, in order to determine whether or not the present system can genuinely contribute to improving ROM practice.

5.2 Evaluation involving patients and clinicians

We evaluate the utility of our system in two experiments, both based on results of the MANSA questionnaire (Priebe et al. 1999). The first experiment compares the identification of important problems vis-à-vis the opinions of clinicians, and the second experiment compares the selection of relevant advice topics vis-à-vis the opinions of patients.

For our first experiment, given a set of filled-out questionnaires, we tested how closely our method which is based on problem severities corresponds, in terms of identifying important problems, to the opinions of clinicians who give patients advice on a day-to-day basis. The goal is to determine whether clinicians are primarily steered by the type of problem (i.e., some problems are considered more important than others) or by the severity of the problem, our system being based on the latter assumption.

For our second experiment, we measure the effects of using a *severity threshold* to truncate the list of advice units for a patient by letting patients evaluate the perceived relevance of selected advice topics. Additionally, this experiment allows us to draw conclusions about whether the system is considered helpful and relevant by the patients.

We chose to use the MANSA questionnaire for our experiments because: (i) it is part of the standard ROM protocol; (ii) it is a relatively short questionnaire, yet it identifies a variety of problems; and (iii) it can be filled out by the patients themselves. In the following section, we introduce some concepts common to both experiments.

5.2.1 Evaluation measurements

In the evaluation of the results of our experiments, we used measurements of precision, recall, and their harmonic mean (also called the *F-measure*). In both experiments, for each filled-out questionnaire, we compared two selections, one made by the system and one made by the expert. We established the selection made by the expert as a *ground truth*, allowing the relevance of the selection made by the system to be expressed in terms of precision, recall, and harmonic mean. The *precision* is the fraction of items selected by the system that are also selected by the expert, while *recall* is the fraction of items selected by the expert that are also selected by the system.

We applied these measurements in both experiments, but we applied them to different concepts. The selections made by the system and experts consist of items (called “topics” in the formulas below), which are problem areas for our first experiment and advice units for our second experiment. Likewise, the term “expert” refers to the clinicians for our first experiment and to the patient for our second experiment. Furthermore, the *selections* are the topics considered most relevant.

We calculated the precision, recall, and harmonic mean using a cut-off to consider only the first n topics ($n = 1, 2, 3$). The first three topics form a good evaluation criterion for our experiments, since Wegweis shows only three advice units on the first page of advice for a patient. In the following definitions, let T_n^e denote the set of the n most relevant topics according to the expert, and let T_n^s denote the set of the n most relevant topics according to the system. We formulate P_n (i.e., precision at n) as follows (Van Rijsbergen 1979).

$$P_n = \frac{t \in \{T_{\mathcal{S}}^e \cap T_n^s\}}{t \in T_n^s}$$

Here, t denotes the number of topics. Thus, precision at n is the fraction of the n most relevant topics identified by the system that are also identified as relevant by

the expert. Likewise, we define R_n (i.e., recall at n) as follows (Van Rijsbergen 1979).

$$R_n = \frac{t \in \{T_n^e \cap T_{\mathcal{O}}^s\}}{t \in T_n^e}$$

Thus, recall is the fraction of the n most relevant topics identified by the expert that are also identified as relevant by the system. Finally, we define F_n (i.e., the harmonic mean of precision and recall at n) as follows.

$$F_n = 2 \cdot \frac{P_n \cdot R_n}{P_n + R_n}.$$

In our experiments, we evaluated the effects of applying a *severity threshold* to limit the number of results returned. If we were to simply return all results, that is, marking as relevant every problem that did not have a perfect answer, the patient would be overwhelmed by the amount of advice and would receive a lot of advice for issues that he/she would not consider to be a problem (e.g., MANSA items answered with 6 = “Pleased”). Thus, since we base our relevance selection solely on problem severity, we needed to use a severity threshold to limit the amount of results returned. The MANSA questionnaire consists of 16 items, 4 of which are binary items (i.e., answered using “Yes” or “No”) and the other 12 are rated on a seven-point satisfaction scale (ranging from 1 = “Couldn’t be worse” to 7 = “Couldn’t be better”). Since the most complex answer type in the MANSA questionnaire is a seven-point rating scale, there are six possible thresholds. To find the best threshold, we evaluated these described measurements for all threshold values on our test set. The results listed “with thresholding” correspond to the optimal threshold value (which ignores answers in the 5-7 range).

In cases where there is no unique ordering (e.g., because multiple problems have the same severity), we take the average over all possible permutations that satisfy the criterion of being sorted according to severity. This guarantees that the ordering depends solely on severities, even when these are equal, without introducing an arbitrary bias.

5.2.2 Clinicians and problem severities

As our first experiment, we test how a system based on problem severities corresponds to the opinion of clinicians, with respect to identifying important problems in the MANSA questionnaire. We executed this experiment twice, with different sets of samples, and the results presented in this section pertain to the two sets combined. In the first execution, we selected five samples (i.e., filled-out MANSA questionnaires) with several severe problems and asked five clinicians (2 psychiatrists and 3 nurse practitioners) to give a list of problem areas in descending order

of importance, which they would discuss with the patient, for each sample. We then compared these 25 results to those of Wegweis. In the second execution, we repeated this experiment with 3 clinicians and 30 samples. Contrary to the first set of samples, this second set was chosen fully at random, that is, the samples did not necessarily have any severe problems. In point of fact, five of the samples in this set actually did not have any severe problems. The executions amounted to a total of 35 samples, which were evaluated by clinicians in 115 lists, which we then compared with the results of Wegweis. The samples that we used in this experiment were selected from a data set (which we acquired through RoQua) of MANSA questionnaires filled out by schizophrenia patients.

Five of the samples that we used in the second execution for this experiment did not include any severe problems and so were excluded from this test. The reason for this was that we cannot use samples without severe problems to prove or disprove our assumption that clinicians select severe problems. Moreover, with severity thresholding applied, our approach only gives results for a sample when it contains severe problems. From our data set of 2601 samples from 1379 patients, 291 samples (11.19%) had no severe problems. We simply accepted the fact that our approach did not apply to the 11.19% of schizophrenia patients who had no severe problems, which we justify by arguing that we do not need to give advice if there is no need for it.

An impression of the distribution of answers of schizophrenia patients for this questionnaire is given in Figure 5.1. This figure shows 2601 filled-out MANSA questionnaires from 1379 schizophrenia patients in the Northern Netherlands as *heat maps*. A heat map is a two-dimensional plot in which the values of a variable are embedded through color intensities or gray levels. In Figure 5.1, the gray level denotes the sample frequency, such that the average gray level of each row is the same, that is, dark squares denote popular choices. The figure shows three heat maps, one for each answer type of the MANSA. The severity of the responses increases from left to right, with the two smaller heat maps representing the yes/no and no/yes items. The braces give an indication of the spread of the answers for an item, and are placed at one standard deviation from the mean on either side. The `nil` column indicates missing or blank values, which are ignored. This figure shows that even though the questionnaire has only 16 questions, many distinct combinations of answers exist, and identifying the important problems is not a trivial task.

We established the *ground truth* in this experiment by averaging over the rankings given by the clinicians. For each sample, this resulted in a single ordered list of problem areas. However, these lists could include outliers (e.g., topics that were selected by only one clinician) that should be discarded. For this purpose, we restricted the maximum length of the list of topics selected by the clinicians to the

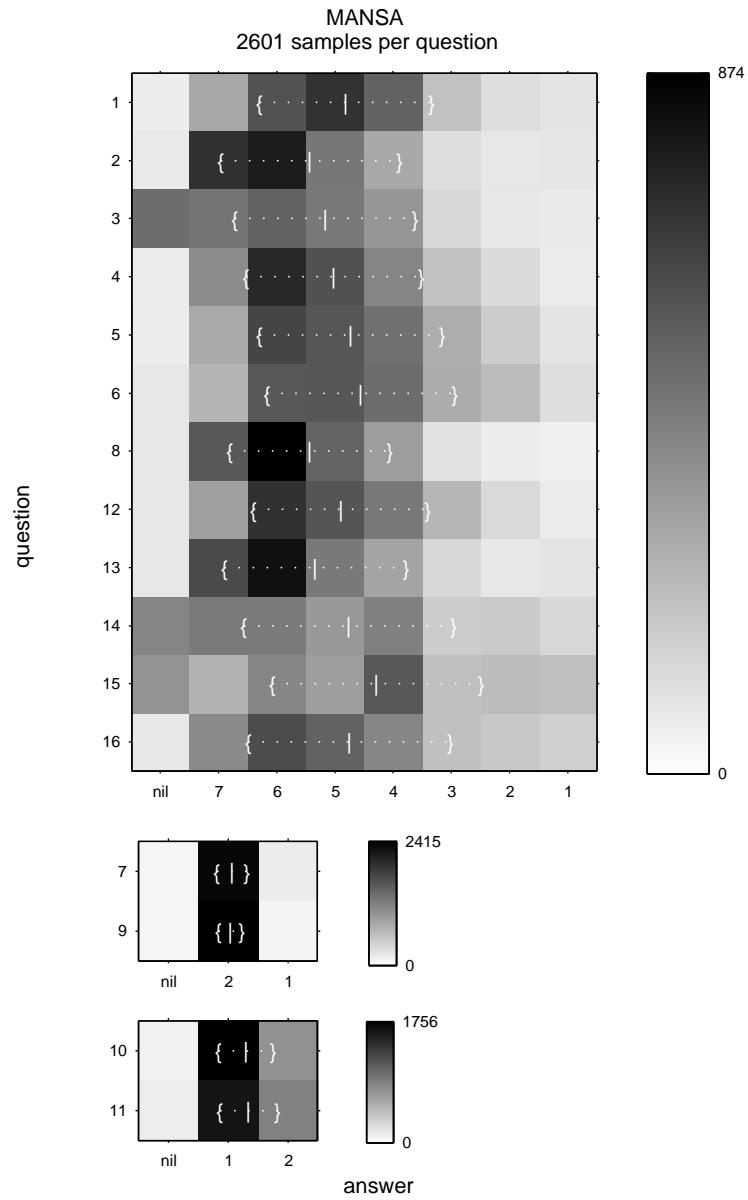


Figure 5.1: Heat map showing answers from schizophrenia patients in 2601 MANSA questionnaires.

Table 5.5: Comparing the system (with thresholding) to the opinion of the clinicians.

| n | Precision@ n | Recall@ n | F-measure@ n |
|-----|----------------|-------------|----------------|
| 1 | 0.983 | 1.000 | 0.992 |
| 2 | 0.957 | 1.000 | 0.978 |
| 3 | 0.943 | 0.944 | 0.944 |

Table 5.6: A breakdown per topic for $n = \infty$, comparing the system (with thresholding) to the opinion of the clinicians.

| Topic | Only clinicians | Only system | Both |
|------------------|-----------------|-------------|------------|
| Sex | 0.0% (0) | 66.7% (12) | 33.3% (6) |
| Physical health | 0.0% (0) | 38.5% (5) | 61.5% (8) |
| Daily activities | 30.8% (4) | 7.7% (1) | 61.5% (8) |
| Life | 8.3% (1) | 25.0% (3) | 66.7% (8) |
| Security | 18.8% (3) | 12.5% (2) | 68.8% (11) |
| Finances | 0.0% (0) | 28.6% (4) | 71.4% (10) |
| Housing | 5.3% (1) | 10.5% (2) | 84.2% (16) |
| Psychic health | 11.8% (2) | 0.0% (0) | 88.2% (15) |
| Relationships | 0.0% (0) | 7.7% (2) | 92.3% (24) |
| Accused of crime | 0.0% (0) | 0.0% (0) | 100.0% (2) |

number of severe problems in the sample. Our reason for basing the cut-off on the number of severe problems is that we are interested in the problems that are considered relevant by clinicians in spite of other problems that are more severe. For example, if a sample indicates three severe problems, and we consider the first three problems selected by the clinicians as relevant, then any difference with the selection of the system is an indication of non-severe problems that clinicians consider more relevant than certain severe problems.

We compared the selections of the clinicians to the selections of the system with thresholding, and the result is shown in Table 5.5. This table shows measurements of precision, recall, and F-measure for $n = 1, 2, 3$. From Table 5.5 we note that with severity thresholding we retain perfect recall values for $n = 1$ and $n = 2$. Thus, we find that in our experiments, the two most important topics according to a clinician are always severe problems. Moreover, for the first three results, our approach based on problem severities complies with clinicians evaluations on average 94% of the time.

While Table 5.5 shows the similarity between system and clinicians for the first three results, for a comparison of the full selections (i.e., for $n = \infty$), we refer to Table 5.6. This table gives a breakdown per topic of the selections made by system and clinicians. The “Only clinicians” column shows the topics that were non-severe problems yet were included by clinicians, the “Only system” column shows the problems that were severe yet were excluded by clinicians, and the “Both” column shows topics that were included by both. On average, we find that 7.3% of selected topics were non-severe problems yet were included by clinicians, and 20.7% were severe problems yet were excluded by clinicians. Thus, for the full selections, our approach corresponds 72.0% of the time with the clinicians, but as we saw in Table 5.5, this percentage is higher (94%) for the first three results.

5.2.3 Patients and advice relevance

For our second experiment, we evaluated to what extent the advice units selected by Wegweis for a patient were considered relevant by that patient. In this experiment, we let patients fill out a MANSA questionnaire and had them evaluate the advice selected by the system, based on those questionnaire answers. We performed this particular experiment for two reasons. First, this experiment allows us to evaluate the effect, with respect to patient satisfaction, of limiting the number of selected advice units by applying a severity threshold. We evaluated this effect by presenting the patients with all the applicable advice units, letting them make their own selection of relevant advice, and then comparing that selection to the selection of the system after applying the severity threshold. Second, this experiment evaluated our advice selection and the ranking algorithms that were explained in Section 4.4. These algorithms are used because the connection between questionnaire items and advice units is not necessarily direct but can be inferred through the problem ontology. Thus, the advice selection for a patient can, for instance, contain very generic advice for very specific problems. Therefore, the assumption to be tested is that the overall selection of advice is still deemed relevant by the patient.

In this experiment, the *ground truth* is the opinion of the patient who filled out the questionnaire, and the results are averaged over all patients. For this experiment, we asked 13 patients (for information on the selection procedure for patients, we refer to our usability study (Van der Krieke et al. 2012)) to fill out the MANSA questionnaire. These filled-out questionnaires were then processed by Wegweis to calculate the full set of applicable advice units (i.e., without thresholding) for each patient. The patients were then asked to select from their set those advice units that they considered relevant to their personal situation and to list them in order of relevance. We told the patients to evaluate the relevance of the topics of the advice

Table 5.7: Comparing the system (with and without thresholding) to the opinion of the patients.

| n | Precision@ n | Recall@ n | F-measure@ n |
|----------------------|----------------|-------------|----------------|
| Without thresholding | | | |
| 1 | 0.652 | 1.000 | 0.790 |
| 2 | 0.617 | 1.000 | 0.763 |
| 3 | 0.665 | 1.000 | 0.798 |
| ∞ | 0.361 | 1.000 | 0.530 |
| With thresholding | | | |
| 1 | 0.652 | 0.846 | 0.737 |
| 2 | 0.643 | 0.808 | 0.716 |
| 3 | 0.702 | 0.815 | 0.754 |
| ∞ | 0.574 | 0.756 | 0.653 |

units (i.e., the advice titles) and not the relevance of the advice contents. The advice contents were not evaluated in this chapter, because they were independent of our approach for inferring, selecting, and ranking advice. To clarify, we want the user to evaluate whether the advice addresses problems that are important to them, not whether or not they agree with the contents of the advice. While the advice contents are part of the Wegweis system, they are variable and may change. The fixed part of our approach that we want to evaluate here is our algorithms for selecting and ranking advice based on the ontology.

The results of comparing the selections of the patients to the selections of the system (both with and without thresholding) are shown in Table 5.7. This table shows measurements of precision, recall, and F-measure for $n = 1, 2, 3, \infty$. The thresholding used for the bottom half of the table is the same thresholding we used in our first experiment, that is, it implies that the system ignores non-severe problems. The perfect (1.000) values for recall in the top half of Table 5.7 are explained by the fact that the system does not omit any advice unless a threshold is used.

In Table 5.7, we find that for increasing values of n , the measurements do not show a steady decrease but show fluctuation. This fluctuation is due to the fact that the measurements for different values of n are based on different amounts of samples, because some samples have only one or two relevant advice units. For example, when the number of relevant advice units for a sample according to the system (or the patient) is two, then this sample is included in the average for $n = 2$ but not in the average for $n = 3$. Despite these fluctuations, we can derive that, for our advice system based on severities, on average two of the three advice units

on the first page of advice are considered relevant by the patient (0.702 precision at $n = 3$).

Table 5.7 also shows that applying a severity threshold results in a higher F-measure when comparing all relevant advice. The rows with $n = \infty$ in Table 5.7 correspond to the standard definitions for precision, recall, and F-measure. These rows show that the precision increases when applying a severity threshold. More specifically, when applying a threshold, 57.4% of the advice given is considered relevant by patients, up from 36.1%. This increase in precision comes coupled with a decrease in recall from 100% to 75.6%, which indicates that only 75.6% of the advice units considered relevant by the patients link to severe problems. However, the combined effect of thresholding remains positive. This effect is shown by the increase of F-measure (from 0.530 to 0.653). These findings suggest that, according to the patients, the use of the severity threshold improves the quality of the advice returned by the system. A breakdown into individual advice topics was omitted from this chapter, since it did not identify any significant trends.

The values of Table 5.7 are relatively low, which indicates that, for patients, the problem severity is not the only criterion for determining the relevance of an advice unit. For example, in our experiment, there were multiple patients with severe problems who marked only non-severe advice units as relevant. In a dismissed alternative approach, we applied global relevance learning to identify popular advice units for patients. However, we found that global relevances did not improve the results. This outcome suggests that the relevant advice selection of patients is highly patient-specific.

We performed a second run of the experiment by inviting 14 more patients (none of which participated in the first run) to use and evaluate our system, to comment on its utility, and to report any abnormalities. Their responses were consistent with our earlier observations. Eight patients responded to our invitation, five of whom had severe problems. For these five patients, of the first three advice units selected by the system with thresholding, 46.7% was found relevant. A possible explanation as to why this number is lower is because, for this run, we used questionnaire data from the most recent assessment of the patients, which was outdated in some cases. For example, one patient remarked that the advice addressed problems that he had reported six months earlier but which had been resolved since then, and thus the associated advice was no longer relevant. In a typical setting, where Wegweis is used as soon as the assessment results are in, the relevance is likely to be higher.

5.2.4 Discussion

The results of our current study show that for the task of identifying the most important problems from a filled-out MANSA questionnaire, an approach based on problem severities can be an adequate approximation of the way clinicians prioritize information for a patient. For the three most important problems, our approach corresponded to the opinion of clinicians in 94% of tested cases, and for all problems, our approach corresponded in 72%. The differences appear to be restricted to a subset of the topics. For example, in Table 5.6, we find that frequently occurring problems such as housing, psychic health, and relationships were identified by the system and clinicians roughly equally often. However, sexual problems, finances, and physical health are issues that clinicians sometimes choose to omit, even when these problems are severe. In contrast, clinicians sometimes discuss daily activities without these being a severe problem. The possible bias for this topic was explained by one of the clinicians, who remarked that when there is nothing else to discuss, they would ask the patient what their plans were for the upcoming week, which is a discussion topic that would be classified under daily activities in our experiments. Another clinician remarked that they would ask the patient if they had any other problems or topics that they wanted to discuss. While not modeled in the results, this interaction roughly equates to the search function on the Wegweis website.

However, we found that patients do not prioritize information in the same way as clinicians do (i.e., using only problem severities). While problem severities have some significance for patients, patients, in their relevance selections, may consider other factors which are unknown to us. In spite of this fact, our experiments show that patients still consider most advice given by the system to be relevant and perceive a quality improvement when a severity threshold is used. The fact that the severity threshold had a positive effect was explained during our feedback sessions by patients, who stated that they did not appreciate being given advice for problems where they had answered 6 = “Pleased” instead of 7 = “Couldn’t be better.” Our experiments also tested the use of the problem ontology to infer generic advice for specific problems, since 5 of the 16 MANSA items had no directly associated advice in the problem ontology at the time of testing. Inferring advice through the ontology did not lead to any logically unexpected advice, according to the patients. Feedback from patients concerning the relevance of advice was related mostly to the contents of the advice rather than to the reason that the advice was given. For example, one patient noted that he talked about physical problems with his physician and not his psychiatrist.

Related work

Prior studies have noted the importance of ethical imperatives such as shared decision making (Drake and Deegan 2009). Shared decision making requires the sharing of medical information between patient and clinician. In the current treatment of schizophrenia patients, the clinician decides which information is shared. We believe that information sharing and shared decision making as a whole can be facilitated by automated ways of interpreting and explaining medical data in forms that are accessible and understandable for patients.

The results of this study are consistent with those of other studies that demonstrated the utility of self-management applications in healthcare (Proudfoot 2004). Furthermore, our experiments have not yielded any evidence to support the traditional belief that there is danger in giving schizophrenia patients direct access to their medical information. On the contrary, our experiments are consistent with the more recent belief that patients benefit from shared decision making (Godolphin 2009).

Limitations

The results need to be interpreted with caution as they are based on small sample sizes. Moreover, our approach only applies for samples that have at least one severe problem, otherwise no advice is shown. Furthermore, the experiment with clinicians is not an entirely accurate scenario in some cases, since in practice clinicians take the patient history into account when giving advice. Whether or not this would shift the results significantly and whether the patient would benefit more from biased or unbiased advice are topics of debate.

Parts published as:

A. Emerencia, L. van der Krieke, E. H. Bos, P. de Jonge, N. Petkov, and M. Aiello – “Automating vector autoregression on electronic patient diary data,” submitted.

Chapter 6

Automating vector autoregression

With the advances in portable consumer electronics, i.e., phones and tablets with internet access, the medical field has started using electronic patient diaries as an important means of collecting medical data. Electronic patient diary data is data entered by patients in a (web) application. The patient fills out a questionnaire using the application, and the results of the questionnaire are used as data points. Participating patients are asked to fill out the questionnaire either daily or at multiple times per day, at set intervals. Electronic patient diary data (also known as Ecological Momentary Assessments or Experience Sampling Method data) can accurately reflect the momentary state of various aspects of a patient. Analysis of this data can reveal how the symptoms, emotions, and activity of an individual evolve over time, how they can be predicted, and which factors contribute to the symptoms, allowing for effective treatment.

A recent development in the medical field is to analyze electronic patient diary data using vector autoregression (VAR). Vector autoregression has its origins in the field of Econometrics (Sargent 1979) and is typically used in analyzing and forecasting financial models (Anderson 1979, Burbidge and Harrison 1984, Litterman 1986, Primiceri 2005). VAR has recently been applied in the medical field to find cause-and-effect relations between symptoms using electronic patient diary data (Wild et al. 2010, Oorschot et al. 2012). The use of VAR techniques in medicine are in line with the upcoming person-centered paradigm called for in clinical practice and research (Tennen and Affleck 1996, Conner et al. 2007, Molenaar and Campbell 2009). For example, in psychosomatic research, VAR models can be used to determine, for individual patients, whether inactivity predicts depressive symptoms or whether depressive symptoms predict inactivity. Using VAR results, clinicians can thus derive whether a patient would benefit more from certain medication or from physical exercise.

The application of VAR models to analyze electronic patient diary data is not yet common practice. The main reason is that the construction of VAR models is a time-consuming and complex process that requires statistical expertise. Figure 6.1

shows the different steps in the manual VAR modeling process. In this figure, the steps are listed in the center, a description per step is shown on the right side, and an abstract example is shown on the left. Manual VAR analysis typically includes preprocessing, following an iterative procedure to find a valid model, and determining optimal constraints for that model (Lütkepohl 2005, pp. 6–7). The manual VAR modeling process can take a statistician several hours up to several days, for a single patient. Current available software solutions for automated vector autoregression such as PcGive (Hendry and Krolzig 2001) are a step in the direction of automation but still rely heavily on the expertise of the user in configuring the program correctly, and they do not automate some of the key operations that a statistician might perform when working manually.

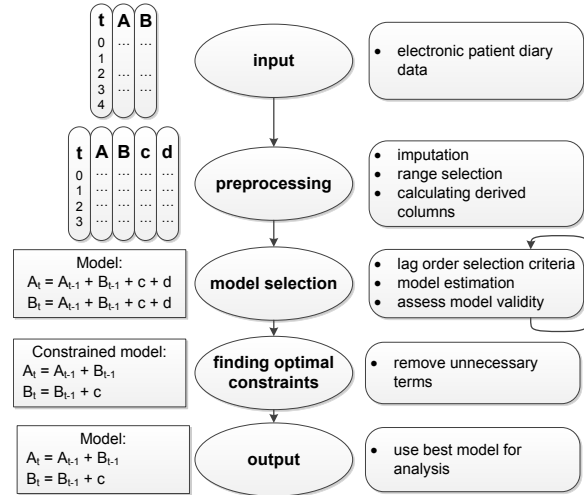


Figure 6.1: The different steps of a manual VAR analysis.

To simplify and speed up the VAR modeling process in a way that closely resembles how statisticians work, we developed Autovar. Autovar automates the process of finding optimal VAR models. Autovar is an open-source package written in the statistical programming language R (*The R Project for Statistical Computing* 2013) and has a web application front-end. Autovar finds and evaluates hundreds of potential models in seconds, selects those that are considered valid as determined by an array of tests, and optimizes the discovered valid models by placing individual constraints. Autovar returns every discovered valid model, along with summary statistics, including *Granger causality* graphs (used for analyzing cause-and-effect relations between time series variables (Granger 1969)), to provide comprehensive and robust insight into the possible model space of a set of time series variables.

We modeled the approach of Autovar after how a statistician selects and finds VAR models. We identified key decision points in the modeling process, e.g., which statistical tests to perform at which time and how the results should be interpreted, adhering to best-practice guidelines, and encapsulated this knowledge in the program flow of our implementation.

In this thesis, we introduce our approach for automating vector autoregression, and we explain the design and implementation of Autovar. We compare the performance of Autovar against VAR models manually constructed by experts, and we compare its features against those of other software used for automating vector autoregression.

In this chapter, we provide a brief introduction to vector autoregression and explain our approach for automating vector autoregression. We evaluate our approach in the next chapter.

6.1 Vector autoregression

Time series data describes the measurements of a set of variables at successive points in time spaced by regular intervals. A VAR model can be specified as a set of equations that express linear dependencies among multiple time series variables (Lütkepohl 2005, pp. 4–5). Here we explain vector autoregression using a model with two variables, adapted from Rosmalen et al. (2012). In the formulas below, Act and Dep refer to measurements of the two variables modeled in this example, activity and depression.

$$\begin{aligned} Act_t &= \alpha_0 + \sum_{i=1}^p \alpha_i Act_{t-i} + \sum_{i=1}^p \beta_i Dep_{t-i} + \zeta X_t + \epsilon_{1,t} \\ Dep_t &= \beta_0 + \sum_{i=1}^p \gamma_i Act_{t-i} + \sum_{i=1}^p \delta_i Dep_{t-i} + \eta X_t + \epsilon_{2,t} \end{aligned} \quad (6.1)$$

A k -variable VAR model consists of k equations (in the above example, $k = 2$). An *endogenous variable* is a variable whose values are predicted by the VAR model. Thus, each of the k equations predicts the values of an endogenous variable in the model. The equations are parameterized by t , the index (or *time points*) of the time series data. The term p is the *lag order* of the system. A VAR equation predicts the value of an endogenous variable Y at time index t , based on previous values from all endogenous variables in the system, including Y itself, of up to p measurements before t . It is not hard to see that if we have n data points, we can predict $n - p$ values at most. Furthermore, in the following, we assume that there are no missing values in the time series data. The error terms ϵ are the *residuals* of the VAR model. These terms are strictly not part of the VAR equations. They merely denote the difference between the predicted values for the endogenous variables (e.g., Act'_t) and their actual values (Act_t), such that for the first formula, $\epsilon_{1,t} = Act_t - Act'_t$. As

Figure 6.2 illustrates, for n data points, we have $n - p$ residuals per variable. The lag order p is 2 because the model uses values of at most 2 measurements before t .

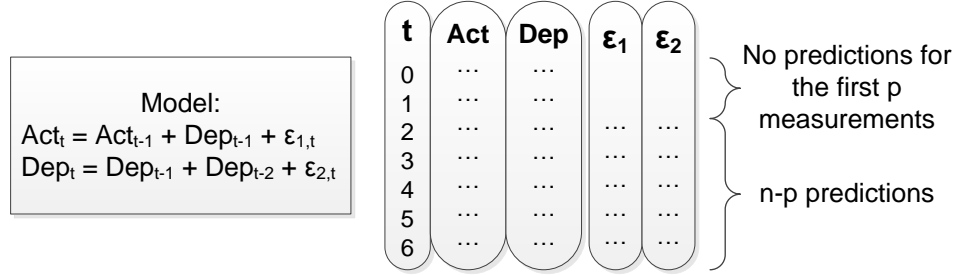


Figure 6.2: When the lag order $p = 2$ and the number of measurements $n = 7$, the number of predictions and residuals in a VAR model is 5.

The formulas in a VAR model may also include variables that are not endogenous in the system. Such variables are called *exogenous variables*. In equation (6.1), X_t is an exogenous variable. We do not consider the exogenous variables to have lagged effects, and thus we only include their *contemporaneous* values in our formulas, i.e., the values at the current time t .

A characteristic of VAR is that the contemporaneous effects of endogenous variables are not part of the model specification (Lütkepohl 2005). In other words, when a prediction for an endogenous variable at time t is based on an endogenous variable at time q , then $q < t$. This facilitates deriving Granger causalities between the endogenous variables.

In equation (6.1), the regression coefficients are the terms α_i , β_i , γ_i , δ_i , ζ , and η . A term is *constrained* or *restricted* when its regression coefficient is set to 0. Constraints are used to remove terms that do not contribute significantly to the prediction accuracy of the model. In our approach, each formula may have a distinct set of constraints. For example, some terms may be constrained in the predictions for Act_t that are unconstrained in predictions for Dep_t . We discuss the approach for setting constraints in more detail in Section 6.6.

6.2 Autovar overview

In Autovar we mimic the way in which a statistician would manually perform VAR model selection (Figure 6.1). There are different manual approaches to VAR model selection. In our approach, we adhere to best practices such as those described in, e.g., Lütkepohl (2005). For example, our approach incorporates elements to favor simple models that explain more of the data.

There are a number of ways in which the approach of Autovar differs from statisticians working manually. Whenever a statistician would make a decision that cannot objectively be classified as correct, in Autovar we choose to exhaustively try all available options. For example, instead of using lag order selection criteria to determine which lag order to use, in Autovar we consider models from every lag order up to a specified maximum.

Following multiple execution paths instead of choosing one naturally leads to a situation wherein multiple models are under consideration. This is the main distinction between not only Autovar and the manual approach, but also between Autovar and other approaches to automated model selection (Hendry and Krolzig 2001, Perez-Amaral et al. 2003), which return one best model. Our approach does not discard any valid model found but ranks the returned models by model fit instead.

The different steps in the approach of Autovar are shown in Figure 6.3. Autovar takes as input the time series data and some parameters. This input is used to determine an initial set of *model configurations*, which are specifications for creating a model. We then construct the VAR models based on their model configurations and assess their validity. If a model proves to be invalid, we may choose to modify some of its properties and reassess several modified variations of the model. If a model was found to be valid, it is added to the results. For every valid model, we also include a constrained version in the results. Finally, we rank the valid constrained and unconstrained models by how well they fit the data and present these models to the user, along with some summary statistics.

The main difference between the approach of the statisticians that we introduced in Figure 6.1 and our approach is that we always consider multiple models, regardless of how well one model performs. There is still an aspect of an iterative approach, expressed by the possibility of adding additional model configurations to consider. The number of model configurations to be evaluated depends on the properties of the data set and on the parameters specified by the user. This process is explained in Section 6.3.

Internally, Autovar is driven by an iterative procedure that maintains a queue of potential models (in the form of model configurations) to be evaluated. We also keep track of which model configurations have already been evaluated to prevent evaluating a model configuration more than once. We evaluate a model using a number of statistical tests, and when a model fails one or more of these tests, we consider the model to be invalid. The aspect of model validity in our approach is discussed in Section 6.4.

Invalid models are discarded. However, Autovar may modify certain properties of the invalid model and requeue these offspring models for assessment. The different scenarios of model invalidity and the subsequent actions to be performed are

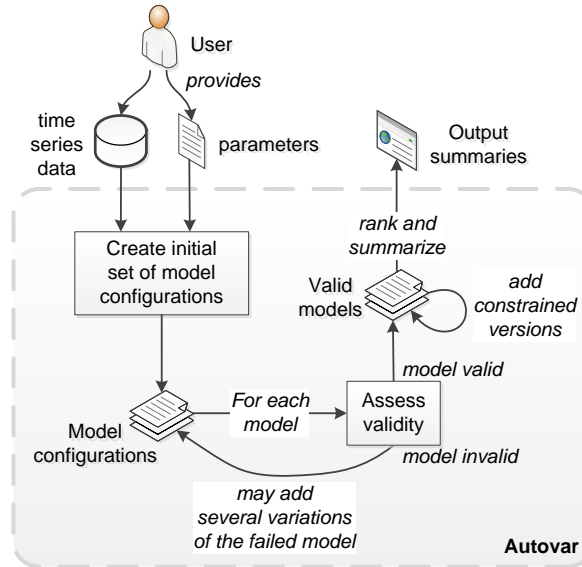


Figure 6.3: The flow of information in Autovar.

discussed in Section 6.5.

The models that pass the validity tests may still include unnecessary terms. Removing those terms may improve the model fit. We developed and implemented a novel approach for finding constraints that produces better results than can feasibly be achieved without automation. Our approach for constraining a VAR model is explained in Section 6.6.

The main algorithm for determining the initial model configurations, assessing their validity, and requeueing modified configurations is explained in detail using pseudocode in Section 6.7.

The implementation of Autovar accepts time series data in certain formats and requires a set of parameters. In the returned results, the models are ranked by how well they fit the data, in terms of their AIC (Akaike Information Criterion (Akaike 1974)) or BIC (Bayesian Information Criterion (Schwarz 1978)) score. Since we return multiple models, we also show summary statistics to provide insight into the properties of the data set and to guide the user in selecting a model. The summary statistics include a graphical Granger causality (Granger 1969) summary and a graph of the contemporaneous correlations. Chapter 7 further details the implementational specifics of Autovar and explains the input and output specifications, along with examples from the web application front-end.

6.3 Model configurations

Let a model configuration be defined as a set of parameters that specifies the terms to be included in the formulas of a VAR model, and as such, as a unique specification for a VAR model. Model configurations have a limited number of parameters that all have a limited number of values. Let the model configuration space define the combinatorial space of all possible models that Autovar can return. When searching for valid models, we limit the search to certain parts of this space, with other parts being invalidated by statistical reasoning or tests performed on the data set.

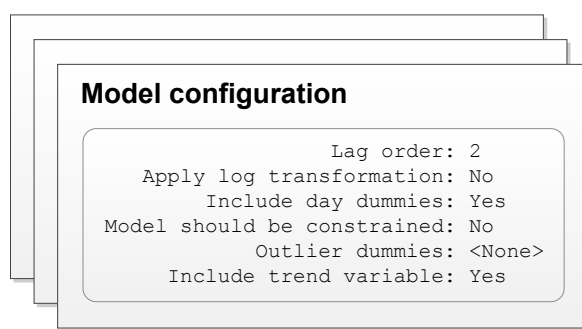


Figure 6.4: An example model configuration.

Figure 6.4 shows the six parameters that we use in model configurations. In the next sections, we explain these parameters in detail.

6.3.1 Trend variable inclusion

When a time series linearly increases or decreases with time t , it is considered stationary around a trend (Nelson and Plosser 1982). Autovar employs the Phillips-Perron test (Phillips and Perron 1988) to determine whether a trend variable should be included. Throughout this chapter, we use the canonical 5% level (Stigler 2008) (corresponding to a p -value ≤ 0.05) as criterion for determining statistical significance.

We run the Phillips-Perron test for each of the endogenous variables. We add a trend to all VAR equations of the model if for one or more of them the Phillips-Perron test is significant ($p \leq 0.05$) and the trend itself is significant. Autovar runs the Phillips-Perron test individually for each endogenous variable, including all lags in the model, and reruns the tests when a model with a different lag order or when a model for the log-transformed data set is under consideration. Thus, the Phillips-Perron results are always specifically calculated for each model configuration.

We only consider linear trends, which follow the definition of an exogenous variable $X_t = t$ for integer t with $1 \leq t \leq n$, n being the number of observations in the data set. In particular, we do not consider the case where the series may have a unit root (a stochastic trend), which may imply that we have to take the first differences of the series as a trend. Support for stochastic trends could be added to facilitate modeling more complex types of data, but for electronic patient diary data linear trends proved sufficient.

6.3.2 Dummy variables for weekdays

Time series with multiple measurements per day may exhibit cyclicity because events at the same time of day may correlate. For example, Figure 6.5 shows a patient with increased depressive symptoms in the evenings. Likewise, time series data may show weekly cyclicity.

Seasonal dummy variables are exogenous variables that are added to a VAR model to account for cyclicity in the series. Seasonal dummy variables are called *dummy variables* because they are zero everywhere except for on specific time points, where their value is 1 (Lütkepohl 2005, pp. 585).

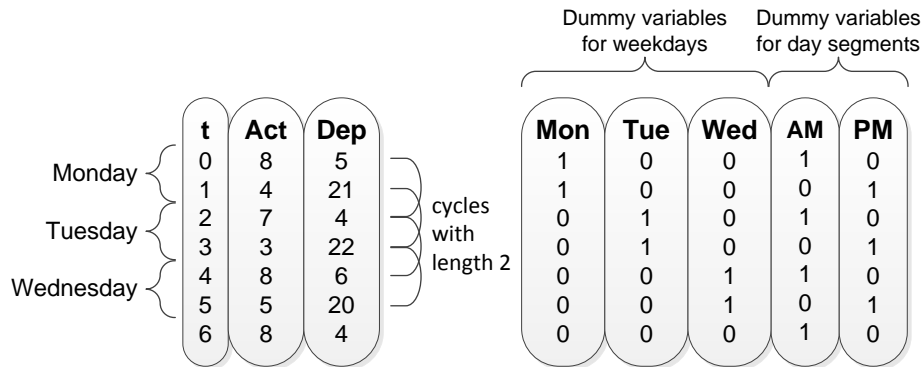


Figure 6.5: An example showing cyclicity associated with day segments.

In Autovar, we consider two types of seasonal dummy variables, those for day segments and those for weekdays. Formally, for weekday dummy variables S_c , we have for n observations that $S_c = i_0, i_1, i_2, \dots, i_{n-1}$, with $i_a = 1$ for all a with $\text{MOD}(a, 7) = c$ and 0 otherwise, where $0 \leq c < 7$ is the index of the day in the week. Figure 6.5 shows how cyclicity may be associated with seasonal dummy variables.

To the best of our knowledge, there is no reliable test to indicate whether any weekly cyclicity present would warrant the inclusion of weekday dummy variables in the models. Hence, Autovar explores both options for all otherwise distinct ini-

tial model configurations. To reduce the complexity of our approach, we choose to always include dummy variables for day segments in unrestricted models, and thus their inclusion is not seen as part of the model configurations.

6.3.3 The lag order

Recall from Section 6.1 that the lag order (or *lag length*) of a VAR model is defined by the highest lag used anywhere in the model. Adding more lags may invalidate a previously valid model, while any lag length on itself may result in a valid model. Statisticians working manually cannot feasibly search for valid models in all applicable lag lengths. They often choose to limit their search scope to the lag lengths recommended by certain lag order selection criteria (Lütkepohl 2005, pp. 135), which are functions that report the lag lengths most appropriate in terms of a combination of goodness of fit and parsimony.

We found that testing only the lags recommended by lag order selection criteria in practice frequently results in a significant number of valid models being overlooked. The reason is that in models with higher lag lengths, the LR-test (Huelsenbeck and Crandall 1997) often prefers the highest lag, while the AIC (Akaike 1974), HQIC (Hannan and Quinn 1979), and BIC (Schwarz 1978) often prefer the lowest lag. This is due to the fact that the latter criteria use a penalty for the number of estimated parameters in the model. If some of the effects are significant on the higher lags while intermediate lags are non-informative, criteria that use a penalty for the number of estimated parameters dismiss the higher-lag option. Nevertheless, a higher-lag model may have a better fit if its intermediate lags were to be constrained. In our approach, we circumvent this problem by choosing to search for VAR models for all lag lengths up to a specified maximum.

6.3.4 Log-transforming the data

We define a *log-transformed model* as a model for the (natural) log-transformed data set. If a log transformation is applied, it is applied to all endogenous variables in the model. A log transformation has a moderating effect on outliers and can thus result in finding valid models for lag lengths where there are no valid models without log transformation.

Statisticians working manually may choose to model log-transformed data only if they fail to find valid models without log transformation. However, to minimize information loss, in Autovar, we explore both options for all otherwise distinct initial model configurations.

Since log-transformed models are strictly models of a different data set, we cannot directly compare their model fit with those of models without log transfor-

mation. For a fair comparison, in Autovar we adjust the calculation of the log-likelihood for log-transformed models to negate the effect of the log transformation on the data (the net effect of this adjustment is to subtract from the log-likelihood the sum of the log-transformed data).

6.4 Model validity

Figure 6.6 shows a schematic overview for assessing the validity of a VAR model. While the properties and assumptions that define VAR model validity are widely recognized (Lütkepohl 2005, pp. 157/212), the specific tests used to evaluate those assumptions may vary. This is due to the fact that the assumptions can be evaluated by different tests and that certain tests are only applicable when the number of measurements is below or above a certain limit.

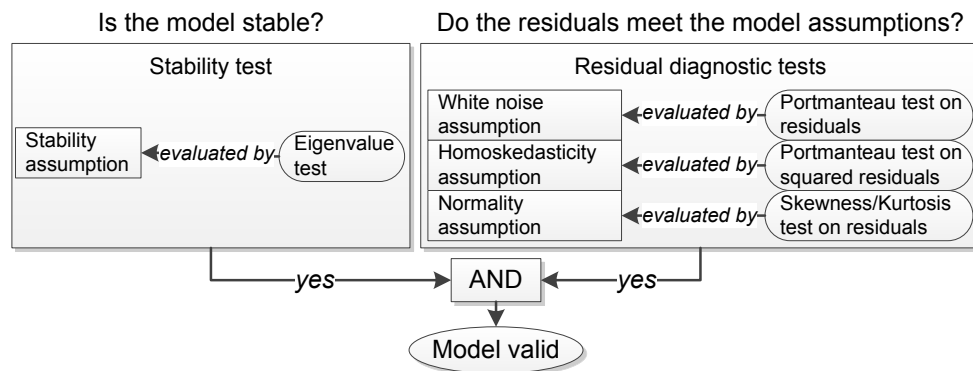


Figure 6.6: Decision chart for assessing VAR model validity as implemented in Autovar. Shown are the properties of valid models, the assumptions whose conjunction defines those properties, and the tests that evaluate those assumptions.

Electronic patient diary data sets typically span between a few weeks and a few months, which is a level of variation that can be covered without having to change test functions. In practice, we found that statisticians use the same set of tests for each electronic patient diary data set. We use this exact set of tests in Autovar (shown in Figure 6.6), automating their evaluation and interpretation.

We use four diagnostic tests in our approach. One test evaluates the model stability (Figure 6.6, left). The other three tests (the *residual diagnostic tests*) evaluate whether the residuals meet the model assumptions (Figure 6.6, right). We consider a model *valid* when it passes all four tests.

6.4.1 Stability test

A VAR model is stable when all eigenvalues of its companion coefficient matrix lie inside the unit circle (Hamilton 1994, Lütkepohl 2005), and this assessment is called the eigenvalue test.

6.4.2 Residual diagnostic tests

The *white noise assumption* states that the residuals of a valid VAR model have serial independency (Box et al. 1976, Diebold 1998). In Autovar, this assumption is evaluated using the Portmanteau test of Ljung-Box (Ljung and Box 1978) on the residuals (Lütkepohl 2005, pp. 169).

The *homoskedasticity assumption* requires that the residuals of a valid VAR model are homoskedastic, i.e., that the variance is stable over time (White 1980). To evaluate this assumption, we perform the Portmanteau test on the squares of the residuals (Granger and Andersen 1978).

The *normality assumption* is evaluated using a Skewness-Kurtosis test (Jarque and Bera 1980, Lütkepohl 2005, pp. 174).

6.5 Handling invalid models

When any of the four tests fail, the model is marked as invalid and will not be included in the list of results. Any remaining tests are still performed if there are equations that passed all other tests so far.

The actions performed when a model fails one of the tests depend on which property is being invalidated, and are described next. The result is typically that one or more variations of the model configuration are queued for assessment.

6.5.1 When the model is not stable

Trend inclusion in Autovar is determined by the Phillips-Perron test for the initial model configurations. However, if the stability test for a model fails, we toggle the trend inclusion setting (meaning if there was a trend we remove it, and otherwise we add a trend) and queue the modified model configuration for assessment. This step is modeled after the iterative approach of statisticians working manually. If the modified model still fails the stability test, the model configuration is discarded.

6.5.2 When the model fails residual diagnostic tests

When the residuals do not meet the model assumptions, depending on which test failed, a statistician working manually may choose to add more lags or to log-

transform the data set. Since Autovar already considers all relevant lag lengths and log-transformed models, such a step is not needed.

Another strategy used by statisticians to solve assumption violation problems is to include special dummy variables in the model that allow residual outliers to be tuned individually (Belsley et al. 2004). As a result, residuals have fewer outliers and a higher chance of passing the homoskedasticity and normality tests.

We mimicked this process in Autovar. We designed a relaxation procedure that creates dummy variables based on outliers of the residuals of a model that failed the residual diagnostic tests. When we include these dummy variables in the failed model, the resulting model has an increased chance of passing the residual diagnostic tests. In the following, let *masking* an outlier denote including its index in a dummy variable that is 0 everywhere except on the time point of the outlier value.

When any of the three tests (shown in Figure 6.6) evaluating the residuals fails, we may queue one or several variations of the model for assessment, each with dummy variables to mask distinct sets of outliers in the residuals of the variables failing one or more tests. When the equation still fails in the new model, we queue a model with increasingly more points masked in outlier dummy variables, and perform up to three iterations of this procedure per VAR equation or until the equation passes the tests.

The reason for using multiple iterations of masking outliers is that choosing one particular threshold for masking outliers may not perform well on different data sets. Our procedure is modeled after the manual approach of statisticians, who plot the residuals and try to add dummy variables for any extreme value. A common substitute for this method is the “factor times standard deviation (std) threshold” approach that we use here. Cousineau et al. (2010) provide motivation for using specific thresholds. In some fields, it is common to use a threshold (or *factor*) 3.5, while in other fields 3.0 or 2.5 is more commonly used. Thus, in Autovar we simply iterate over these three factors until we find a valid model. We start with fewer outliers (3.5) and add more outliers only if the tests for an equation keep failing (3.0 and 2.5). For example, when a certain VAR equation still fails the tests when $3.5 \times \text{std}$ residual outliers of that variable are placed in dummy variables, we queue a new model with $3.0 \times \text{std}$ residual outliers of that variable in dummy variables. In order to favor models that explain more of the data, outliers are masked in dummy variables only if doing so is necessary to establish model validity.

The iterations are tracked individually per VAR equation, and Autovar considers all possibilities for finding optimal VAR models. For example, consider a VAR model of two variables, A and B , with both equations failing the white noise assumption. We then queue three new models, one with $3.5 \times \text{std}$ outliers of A in dummy variables, one with $3.5 \times \text{std}$ outliers of B in dummy variables, and one

that includes both sets of dummy variables. Since A and B may have outliers in common, including this third model is not redundant because it is not guaranteed that it is reachable from the other two models, meaning that we may otherwise not consider this model.

6.6 Constraining valid models

In the VAR model-fitting process, individual terms can be *constrained* (or *restricted*) per equation, effectively removing them. The goal of setting constraints is to obtain a model with better fit as measured by the AIC (Akaike Information Criterion (Akaike 1974)) or BIC (Bayesian Information Criterion (Schwarz 1978)). These criteria include a penalty that scales with the number of estimated coefficients in the model. Thus, removing insignificant terms often improves model fit. Autovar has the option to optimize either for lower AIC scores or for lower BIC scores (with lower scores indicating a better model fit), hence in the following we write *AIC/BIC* to denote whichever information criterion was chosen.

Searching for optimal constraints is a computationally expensive process as there are many distinct constraint configurations. For example, consider a VAR model with three endogenous variables, lag order 6, three measurements per day (two dummy variables), weekday dummies (six dummy variables), and a trend variable. Each VAR equation in this model has $3 \cdot 6 + 2 + 6 + 1 = 27$ terms that could potentially be constrained (not counting any outlier dummy variables), or $2^{27} - 1$ distinct constraint configurations. Additional complication stems from the fact that each placed constraint requires a full recalculation and re-evaluation of the VAR model as the residual diagnostics and statistical significance of other terms may have changed drastically.

Since statisticians working manually cannot feasibly test millions of constraint configurations, several greedy approaches are used in practice (Lütkepohl 2005, pp. 206). These algorithms have a time complexity of $O(n)$ or $O(n^2)$, with n the number of terms in the equations. For example, in a Sequential Elimination of Regressors Strategy (Lütkepohl 2005, pp. 211), the term with the highest p -value (i.e., the least significant term) is constrained in an iterative procedure that is ran until the AIC/BIC score no longer decreases. The validity of the model is assessed afterward. This approach uses no intermediate validity testing. The approach is based on the assumption that terms that do not contribute significantly to the model may be removed as long as the model fit improves as a result.

The described Sequential Elimination of Regressors Strategy does not assert nor guarantee validity of the resulting model. Thus, there is no good estimation of how many models it needs to be run on in order to get good results. A commonly used

approach is therefore to run it on all evaluated models. This works well for statisticians working manually, who consider a small number of models. In Autovar, we found that performing a constraint search for each model under consideration, where each step requires a full re-estimation and re-evaluation of the VAR model, has a significant impact on the running time (up to several minutes per data set). In addition, we found that performing a constraint search only for models that are already valid without constraints often results in finding the exact same set of valid constrained models.

In Autovar, we choose to constrain only the most promising models, i.e., those valid without constraints. Thus, we potentially overlook models that would become valid when certain terms in the equation were to be constrained. However, we found this to be a rare occurrence. A possible explanation is that the evaluated model configurations have considerable overlap, i.e., some of the unconstrained models could be considered as constrained versions of others.

While the approach used for setting constraints in Autovar is similar to the Sequential Elimination of Regressors Strategy described earlier, we developed and implemented improvements that result in lower AIC/BIC scores. First, because models are initially valid, we may impose the assertion that the resulting constrained models should always be valid as well. We follow a greedy approach and constrain the term with the highest p -value as long as the resulting model remains valid and the AIC/BIC score does not increase. Like other greedy approaches, ours is not guaranteed to always find the best constraints. Second, when constraining the term with the highest p -value is not possible (either because it invalidates the model or because it increases AIC/BIC scores), we continue with the term with the second-highest p -value and so on. This step causes the constraint-setting algorithm to have quadratic time complexity. However, it does frequently result in better constraints (we refer to Section 7.2.1 for a comparison) and guarantees model validity since the initial models are valid and validity is asserted in every step.

6.7 Algorithm for model selection

We now present the main procedure for selecting valid models in Autovar. The `GetValidModels` function (Algorithm 6.1) returns an unordered list of valid VAR models and their configurations, given a data set and other input parameters. The parameter options P specify the minimum and maximum lag order to consider. If zero-order lag models should be included, minimum lag order $P.min_lag$ is 0, otherwise it is 1.

In the first step of the algorithm, we initialize the model configuration queue Q to contain an initial set of model configurations based on the data set D and given

```

GETVALIDMODELS( $D, P$ )

Input: data set  $D$ , parameters  $P$  (min. lag and max. lag).
Data: functions evaluate_var_model, stability_test,
         portmanteau_tests, and skewness_kurtosis_test.
Output: list of  $\langle \text{configuration}, \text{model} \rangle$  tuples, representing the valid models
         found.

 $Q \leftarrow \text{INITIALMODELCONFIGURATIONS}(D, P)$ 
 $R \leftarrow$  empty list
 $S \leftarrow$  empty set
while  $Q$  is not empty
     $M \leftarrow Q.\text{pop}()$ 
     $B \leftarrow \text{evaluate\_var\_model}(D, M)$ 
     $A \leftarrow \text{TRUE}$ ,  $T \leftarrow \text{FALSE}$ 
     $O \leftarrow$  empty set
    if stability_test( $B$ ) fails
        then  $A \leftarrow \text{FALSE}$ ,  $T \leftarrow \text{TRUE}$ 
    if portmanteau_tests( $B$ ) fails
         $A \leftarrow \text{FALSE}$ 
        then { for each variable  $V$  that failed
                do insert  $V$  in  $O$ 
            }
    if skewness_kurtosis_test( $B$ ) fails
         $A \leftarrow \text{FALSE}$ 
        then { for each variable  $V$  that failed
                do insert  $V$  in  $O$ 
            }
    if  $A$ 
        do {
            add  $\langle M, B \rangle$  to  $R$ 
            if not  $M.\text{restrict}$ 
                then {  $N \leftarrow \text{copy}(M)$ 
                        then {  $N.\text{restrict} \leftarrow \text{TRUE}$ 
                                add  $N$  to  $Q$ 
                            }
                    }
            if  $T$ 
                 $N \leftarrow \text{copy}(M)$ 
                then {  $N.\text{trend} \leftarrow \neg N.\text{trend}$ 
                        if  $N \notin S$ 
                            then insert  $M$  in  $S$ , add  $N$  to  $Q$ 
                    }
            for each set  $W \in \{\mathcal{P}(O) - \emptyset\}$ 
                 $N \leftarrow \text{copy}(M)$ 
                for each variable  $V \in W$ 
                    do if not  $N.\text{outliers}.V = 3$ 
                        then  $N.\text{outliers}.V++$ 
                if  $N \notin S$ 
                    then insert  $N$  in  $S$ , add  $N$  to  $Q$ 
        }
    return ( $R$ )

```

Algorithm 6.1: The GetValidModels algorithm.

INITIALMODELCONFIGURATIONS(D, P)

Input: data set D , parameters P (variable names, max. lag, etc.).

Data: function `phillips_perron`.

Output: queue of tuples of model parameters.

```

 $Q \leftarrow$  empty queue
for each  $l \in [P.min\_lag, P.max\_lag]$  do
  for each  $t \in \{FALSE, TRUE\}$  do
    for each  $d \in \{FALSE, TRUE\}$  do
       $\langle lag = l,$ 
         $apply\_log\_transform = t,$ 
         $include\_day\_dummies = d,$ 
      add  $restrict = FALSE,$ 
         $outliers = NULL,$ 
         $trend = phillips\_perron(D, t, l) \rangle$ 
      to  $Q$ 
return ( $Q$ )

```

Algorithm 6.2: The InitialModelConfigurations algorithm.

parameters P . These initial model configurations are returned by the InitialModelConfigurations function shown in Algorithm 6.2. This algorithm returns a queue of initial model configurations for the given parameters. It contains model configurations of lags up to the given maximum lag, with and without weekday dummy variables (if applicable), and with and without log transformation. For each model configuration, the `trend` parameter, which signifies the inclusion of a trend variable in the model, is set according to the Phillips-Perron test as explained in Section 6.3.1. Furthermore, dummy variables for day segments are included in each model (Section 6.3.2).

Returning to Algorithm 6.1, we initialize R , our return variable, and S , a set to keep track of the model configurations that have been tested so far. We use this set to ensure that we do not evaluate models more than once. We loop through the main body as long as there are model configurations to be tested. We evaluate each model configuration M popped from the queue Q to create a model B .

We proceed to introduce two state flags in the loop body. The variable A is true as long as we consider the model B to be valid. The variable T becomes true when the stability test fails. The set O keeps track of the names of the variables that failed

at least one of the residual diagnostic tests.

We first test the stability of the model B using the eigenvalue test. If the model fails the test, we set A to false to denote that the model is invalid. We also set T to true to consider toggling the trend inclusion later on.

The function `portmanteau_tests` runs the Portmanteau test on the residuals (white noise assumption) and on the squares of the residuals (homoskedasticity assumption). Each variable V that fails either of these tests is added to the set O . Furthermore, if any variable fails either of the two tests, we set A to false to denote that the model is invalid.

The function `skewness_kurtosis_test` evaluates the skewness and kurtosis of the model. The model is invalidated (A set to false) if the residuals of any VAR equation show significant skewness or kurtosis. The offending variables are inserted in the set O .

After running the tests, we check whether A is still true to determine if the model passed all tests. If the model passed all tests, we consider it to be valid and add it to the return variable R in a tuple with its model configuration. In addition, if the model was unrestricted, we queue a copy of the model configuration with the `restrict` flag set to true to denote that this is a valid model configuration for which we should try to find constraints. Constraints are set as part of the functionality of the `evaluate_var_model` function, according to the approach explained in Section 6.6. Moreover, recall that constrained models remain valid and thus T will never be true and O will always be empty for restricted models.

Next, we check if T is true. Recall that T is true if and only if the stability test failed. In this case, we toggle the inclusion of the trend variable in the model configuration and add the new model configuration N to the queue Q . To ensure that we only toggle the inclusion once, we first check whether N is not in the processed set S . If it is not in this set, we add the original model M to this set S . Note that it is not necessary to add N to this set.

The final for-each statement is for queueing model configurations with more outliers masked in dummy variables for variables that failed at least one of the residual diagnostic tests. Recall from Section 6.5.2 that we consider all combinations for decreasing the outlier threshold by 0.5 for each failing variable. This number of combinations is $2^f - 1$, with f the number of failing variables and is signified by the powerset of O minus the empty set. Also recall that we use three levels for thresholding outliers into dummy variables, maintained separately per variable. These levels are used in the `evaluate_var_model` function to add outlier dummy variables.

Not shown in the code are several intermediate checks for duplicates to ensure that, e.g., created dummy variables for outliers are never empty and that constrained models do not degenerate to lower order models that already exist.

6.8 Discussion

We have presented Autovar, our automated approach for finding valid vector autoregressive models for electronic patient diary data. Autovar can be described as an exhaustive approach that finds all valid models within a parameter space that is restricted by statistical tests and logic.

Autovar was modeled after the way in which statisticians work manually, while adhering to best-practice guidelines for finding valid models. Autovar incorporates improvements over any manual approach by virtue of its constraint-finding method, which uses backtracking to find better constraints. Autovar contrasts with other approaches for automated model selection in that it returns all valid models found instead of one best model.

The approach for automated model selection described in this chapter is not limited to electronic patient diary data. Any time series data (i.e., any set of features measured at periodic intervals) of 2-3 features that contains linear trends at most, can be analyzed efficiently with Autovar. Autovar can easily be used and adjusted for other purposes because it is an open source package written in an open source language.

Parts published as:

A. Emerencia, L. van der Krieke, E. Bos, P. de Jonge, N. Petkov, and M. Aiello – “Automating vector autoregression on electronic patient diary data,” submitted.

Chapter 7

Evaluation of Autovar

The previous chapter described our approach for automating vector autoregression, Autovar. What remains is an evaluation of its performance and functionality. For this purpose, in Section 7.2, we compare the results of Autovar on actual data sets versus those of experts working with the statistical software STATA. We compare the valid models found based on model fit. In addition, we provide a formal evaluation of performance aspects of our approach where we consider aspects of time complexity, memory complexity, and scalability. Finally, in Section 7.3, we compare the functionality of Autovar to that of the most used commercial software available today. We conclude the chapter in Section 7.4.

We first take a closer look at the implementational aspects of Autovar and its web application front-end.

7.1 Implementation

We developed Autovar as a package in the open-source statistical programming language R (*The R Project for Statistical Computing* 2013). The source of Autovar is publicly available on GitHub (*Autovar: GitHub repository* 2013).

7.1.1 Imported, modified, or implemented functions

Autovar makes use of other open source packages. The model evaluation uses the `VAR` function from the `vars` package (Pfaff 2008) to construct the VAR models. Reading the STATA and SPSS file input uses the `foreign` package (*foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase, ...* 2013). For the implementation of the Phillips-Perron test, we use the `pp` function from the `urca` package (*urca: Unit root and cointegration tests for time series data* 2013). We use the `vars::roots` function (Pfaff 2008) for the stability test.

The web application front-end is a single HTML page, stylized with Bootstrap (*Twitter Bootstrap* 2013). The back-end is an Apache server (*The Apache Soft-*

ware Foundation 2013) running OpenCPU (*OpenCPU: Scientific computing in the cloud* 2013) to provide a RESTful interface for executing R code from a web-application. The web application uses the `knitr` (*knitr: A general-purpose package for dynamic report generation in R* 2013) and `markdown` (*markdown: Markdown rendering for R* 2013) packages to render output from R in HTML form. The `ggplot2` (Wickham 2009) package is used to display graphs. A live version of the web application can be accessed from <http://autovar.nl>.

In addition to building the Autovar framework, we implemented or adapted several statistical functions because previous implementations either did not exist, were faulty, or were otherwise unusable for our purposes.

We implemented the Portmanteau test, modeled after the approach from Ljung-Box (Ljung and Box 1978), as part of Autovar ourselves in order to obtain results separately per VAR equation. Our approach relies on the individual assessment of the VAR equations to identify the residuals for which to mask additional outliers in dummy variables (Section 6.5.2). The implementation that we wrote in Autovar resembles the `wntestq` function from STATA, in that we calculate the Portmanteau test statistic for each individual equation. There was an existing implementation of the Portmanteau test available in the `vars` package, as a function called `vars::serial.test` (Pfaff 2008), but it returns results for the model as a whole rather than individual results per equation.

We implemented two Skewness-Kurtosis tests as part of Autovar, the Jarque-Bera test (comparable to `jbtest` in STATA) and the Skewness-Kurtosis test (`sktest` in STATA). The `vars` package in R does include an implementation of the Jarque-Bera test by means of the `jb` function, but this function does not suit our needs for two reasons. First, the `vars::jb` function does not separate the skewness and kurtosis values per variable but only returns one set of values for all VAR equations. Second, the `vars::jb` function, as it is available to us (version 1.5-0), is not compatible with constrained models. We thus implemented a Jarque-Bera test function `jb_test` as part of Autovar following the approach described in Jarque and Bera (1980). For handling smaller sample sizes, we implemented another such test, called the Skewness-Kurtosis test (D’Agostino et al. 1990, Royston 1992). This function is now the default Skewness-Kurtosis test in Autovar, but the Jarque-Bera test remains available as an option.

7.1.2 Input data and parameters

The minimum required parameters for Autovar to run are the name of an input file and the names of the endogenous variables. Autovar accepts STATA (.dta) or SPSS (.sav) input files. The rows in a data file should correspond to consecutive

measurements at equidistant time intervals (i.e., Autovar currently does not support irregular time series). The columns should represent the different variables measured. The user specifies which columns of the input file should be included as endogenous variables. The web application interface for this process is shown in Figure 7.1. In this appendix, we show how Autovar can be used on the data set 45 `Stre Musc` from Table 7.1.

The maximum lag length can be specified manually for optimal results (Figure 7.1). There is a default value of 3, but the maximum lag length should typically be chosen based on theoretical and practical considerations (e.g., as a multiple of the sampling frequency and taking the number of observations into account). Autovar has the option to extend the search space to include *zero-order* lag models. Zero-order lag models are effectively lag-1 models with all lag-1 terms constrained in all equations. Thus, if a zero-order lag model passes the validity tests, it means that each endogenous variable can be accurately approximated by a constant, unaffected by time or previous measurements.

7.1.3 Exogenous variables

In our approach, we toggle trend inclusion for models that fail the eigenvalue stability test. It serves to note that for the electronic patient diary data sets we tested on, the model stability test has not failed once, and thus all our models adhere to the Phillips-Perron test recommendations with regards to trend inclusion. Figure 7.2 shows that in the web application of Autovar the inclusion of trend variables for any model can optionally be disabled.

Columns for the seasonal dummy variables are generated by Autovar and hence do not need to be present in the data set. The user only needs to specify the date of the first measurement, the sampling frequency (the number of measurements per day), and the offset (specified as the part of day of the first measurement). Under the assumption that the data set does not contain any missing records, Autovar then constructs weekday dummy variables, and if there is more than one measurement per day, Autovar also constructs dummy variables for the different day segments. If there is no timestamp data available for the supplied data set, or when it contains missing values, Autovar runs without creating seasonal dummy variables.

For every full set of seasonal dummy variables, Autovar includes all but one. The reasoning is that the presence of the omitted variable can be derived from the others. Hence, introducing this linear dependency does not contribute to the expressive power of the model. For example, in the case of weekdays, we include six dummy variables for six of the weekdays since we know the seventh is one if and only if all the others are zero. A similar construction holds for the dummy variables for day

Autovar

[Other version](#) | [Source](#) | [Contact](#)

Input

[Exogenous Variables](#)

[Advanced Settings](#)

Data set

Choose File

Dataset45Jdaggem.dta

Select .sav or .dta file containing the data set.

VAR columns

"Stre", "Musc"

ID

Date

Stre

Musc

Dysp

Abdo

Sunday

Monday

Tuesday

Wednesday

Select the variables to run VAR on (click while holding down CTRL or ⌘). Select the original variables only, not the log-transformed variables.

Max. lag

3

The maximum lag order to consider (typically a multiple of the number of measurements per day).

Lag-0 models

☒ Include models at lag 0

Models at lag 0 are actually models at lag 1 with all lag-1 parameters constrained.

Run

Figure 7.1: Part of the user interface of the web application front-end of Autovar.

Autovar

[Other version](#) | [Source](#) | [Contact](#)[Input](#)

Exogenous Variables

[Advanced Settings](#)Trend column ☒ Include trend variables

Check to create and include a trend variable (named 'index') for models that need a trend according to the Phillips-Perron test.

☐ Include squared trend

For those models that include a trend variable, also include the square of this trend.

Timestamps ☒ Set timestamps

Check to create and include dummy variables for weekdays. If the date of the first measurement is known, then check this box and fill out the date below. This causes Autovar to evaluate every model twice: once with and once without dummy variables for weekdays. Additionally, if the number of measurements per day is ≥ 1 , then dummy variables for day parts are added to all models.

Date of first measurement

2005-03-03

Measurements per day

1

Daypart of first measurement

1

If the first measurement in the data set should not be treated as the first, but as the second or third (etc.) measurement of that day, then specify this here. For example, if "Measurements per day" is set to 3, and "Daypart of first measurement" is set to 2, then the first two measurements in the data set will be tagged with Afternoon and Evening, and the third measurement in the data set will be tagged with Morning of the next day.

Figure 7.2: Part of the user interface of the web application front-end of Autovar, showing settings of the Exogenous Variables tab.

segments. For example, in Figure 6.5, the PM column is the inverse of the AM column, and can thus be removed without the model losing any expressive power.

For calculating the seasonal dummy variables in particular, Autovar assumes that the data set represents a sequence of measurements with a constant amount of time between consecutive measurements. To account for missing values, Autovar currently has a very limited, basic imputation scheme using linear interpolation applied to the five closest surrounding points (taking the mean for numerical data, and the mode for nonnumeric data). More sophisticated methods for imputation (such as expectation maximization imputation (Dempster et al. 1977)) are currently not implemented.

For masking residual outliers in dummy variables, the default iteration limit in Autovar is 2 (masking residual outliers at $3 \times \text{std.}$), but can be set to any integer between 0 (meaning no outliers are masked) and 3 (masking outliers at $2.5 \times \text{std.}$) inclusive. A setting of 0 signifies that the models should not use outlier dummy variables at all. For the third iteration, we opted to include outliers of the squared residuals also. This iteration is only used if we specifically choose to (because it is not the default setting), which is only when we were unable to find any valid models using up to two iterations.

While all equations in the unrestricted VAR model include the same set of outlier dummy variables, their regression coefficients (ζ and η in (6.1)) are likely to differ. In addition, Autovar has several options for distributing the indices over a different number of outlier dummy variables. Instead of creating one dummy variable for all outliers, the default setting in Autovar is to split up the indices per endogenous variable. If further fine-tuning is needed, Autovar also has the option to create one dummy variable for each outlier.

The reason for the default setting of combining the outliers into a single variable per equation is that we found that in many cases the effect of better configurability on the AIC/BIC scores is relatively small compared to the effect of reducing the number of exogenous variables in the equations by compacting outlier dummies into single variables. Partitioning the outliers into individual variables in theory should allow for better configurability of the model but may incur a slight performance hit due to the increase in the number of terms of the VAR equation.

7.1.4 Web application output

The web application functions as a user interface wrapped around the functionality of the Autovar R-package. It is designed to perform VAR analysis quickly and exposes the most commonly used features of Autovar. Its output contains summaries and details for the valid models found, and thus can convey a comprehensive

understanding of the time series data.



Figure 7.3: Part of the output shown by the web application front-end of Autovar, illustrating how data sets are loaded and how the time series data is visualized.

The selected options in the user interface (e.g., Figures 7.1 and 7.2) are converted into function calls interpreted by the Autovar package. The output shows these snippets of R code (in gray boxes) interspersed with their resulting output text and figures.

When the user clicks the “Run” button on the web application, Autovar performs a number of function calls and shows the generated output. First, the data set is loaded and a trend variable is added (Figure 7.3). Next, timestamps are set, creating dummy variables for day segments and weekdays. Then, plots are shown to display the endogenous variables graphically. Finally, Autovar calls the main procedure for finding valid models, and shows a graphical summary of contemporaneous correlations found in the valid models (Figure 7.4), a graphical summary of Granger causalities found in valid models (Figure 7.5), a summary of properties of the valid model configurations, and the full list of valid model configurations found, sorted by AIC/BIC score. For the best log-transformed model and the best model without log transformation, Autovar also shows a more detailed description. This description includes coefficients, standard errors, and p -values for the terms as well as the output of the validity tests.

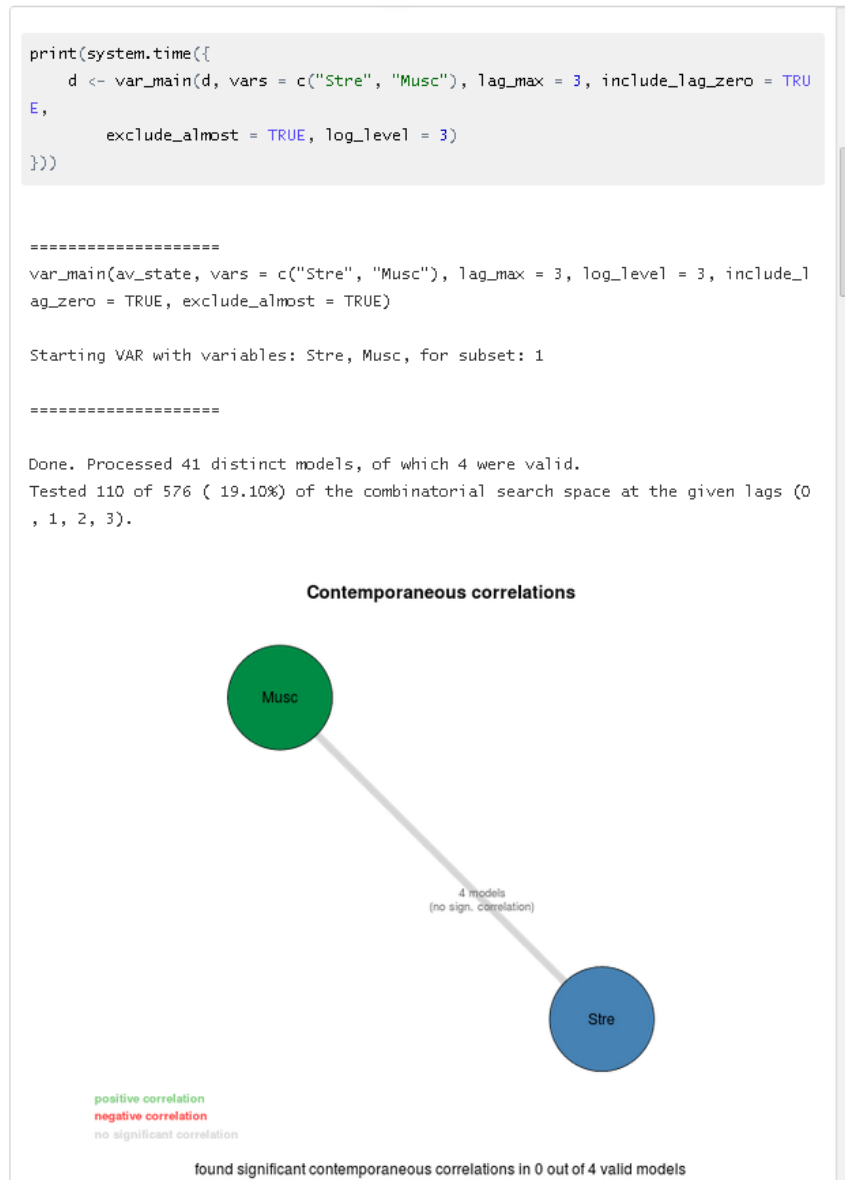


Figure 7.4: Part of the output shown by the web application front-end of Autovar, illustrating how the main VAR procedure is called and showing the Contemporaneous correlations summary graph.

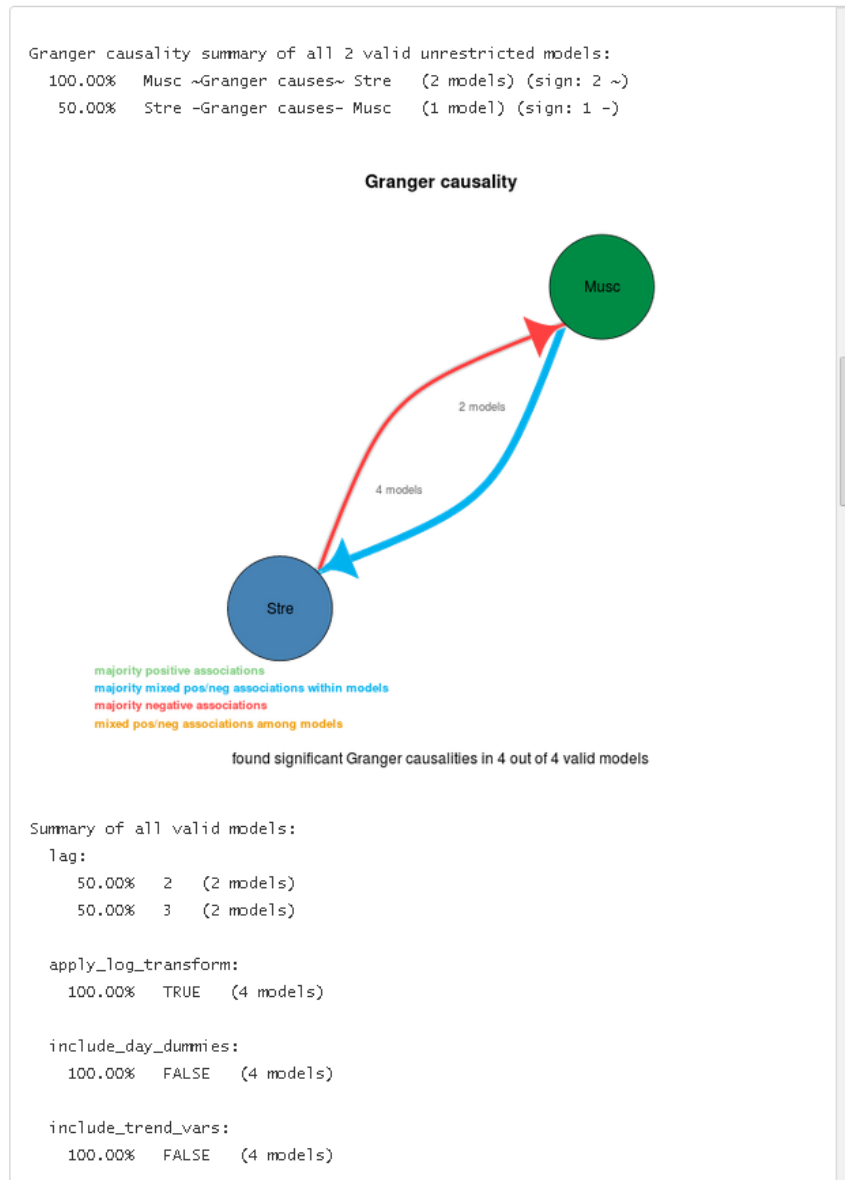


Figure 7.5: Part of the output shown by the web application front-end of Autovar, illustrating the Granger causality summary graph and summary statistics of valid models.

7.2 Evaluation

Here, we evaluate the practical and theoretical performance of our approach.

7.2.1 Comparison with manual analysis

We compare Autovar to experts working manually with respect to the model fit of valid models found.

Data set

The data set consists of a sample of 20 patients with multiple, persistent Functional Somatic Symptoms (FSS). Electronic diaries were used to collect the times series data on stress and FSS. The data were collected between January 2004 and February 2006. The data were preprocessed to yield one measurement per day, resulting in an average of 86 measurements per patient (max. 100, std. 6.58).

The patients helped to identify their three most severe, applicable, or frequent symptoms from the following list: muscle pain (*Musc*), joint pain (*Join*), back pain (*Back*), headache (*Head*), abdominal pain (*Abdo*), pelvic pain (*Pelv*), bowel symptoms (*Bowe*), dyspepsia (*Dysp*), nausea (*Naus*), tight throat (*Tigh*), chest pain (*Ches*), weakness (*Weak*), numbness (*Numb*), and palpitations. This data set was collected by Burton et al., who provide a full description of how each symptom was measured (Burton et al. 2009).

Setup

For each patient, three bivariate data sets were constructed, each one using Stress (*Stre*) as one of the endogenous variables and one of the three FSS symptoms selected by the patients as the other. Missing data was previously imputed for each individual data set using the Expectation Maximization function in SPSS 20 (*IBM SPSS software* 2013). Neither approach uses dummy variables for day segments since there is only one measurement per day.

Manual analysis

The manual approach we are comparing to was performed by van Gils et al. (2014) using STATA 11. We believe that comparing their models against those of Autovar is fair because both approaches use the same diagnostic tests to assess the validity of models.

The manual approach first includes both a linear trend variable and weekday dummy variables, and then removes those that are not statistically significant. The lag order of the model is determined by majority voting of several lag length se-

lection criteria. Specific measures were taken to improve the model, depending on which assumptions of the model were violated according to the diagnostic tests. Residual autocorrelation was solved by including higher lags. Heteroskedasticity and skewness were solved by using a log transformation on the endogenous variables. If the non-normality merely stemmed from a few outliers, then dummy variables masking outliers at $3 \times \text{std}$ of the residuals were used. Statistically insignificant terms were pruned from the estimated models in descending order of p as long as the BIC score did not increase. No diagnostic tests were performed at intermediate steps when placing constraints.

Autovar analysis

Autovar used the same parameters for every patient data set. The maximum lag length was set to 3 (which is our default value if we do not know anything about the data) and zero-lag models were included. Like the manual approach, constraints were chosen to optimize for low BIC scores. Each data set was timestamped, allowing Autovar to derive and include dummy variables for weekdays if needed. All other settings were left at their default value. If a run returned no valid models, Autovar was called a second time, with identical parameters except with maximum lag at 7 instead of 3 and the lowest factor for masking outliers at $2.5 \times \text{std}$ instead of 3 and including outliers of the squared data set.

In Autovar, the ranking of models by model fit is based on adjusted AIC/BIC scores that compare log-transformed and non-logtransformed models fairly. However, since the AIC/BIC scores of the manual approach do not include this adjustment, to avoid confusion, we show only the unadjusted AIC/BIC scores in the results, and we compare only the AIC/BIC scores of data sets where both approaches have either log-transformed or non-logtransformed models.

Comparison

Table 7.1 shows a comparison of the best models found between Autovar and the manual approach. Note that Autovar always returns multiple models, but this table only shows the results of the best model of each approach. The rows are the data sets. The left column identifies the data set. The number identifies a patient. For each patient, three data sets are analyzed, each having two endogenous variables, stress and one other FSS symptom indicated by the patient. The remaining columns show the details of the model of Autovar with the lowest BIC score (columns 2–7) and the final model obtained in the manual approach (columns 8–13). The Exogenous variables column denotes which exogenous variables are used in the selected models. The variable `Nr` denotes the linear trend variable. The variables `Mon`, `Tues`,

Wed, Thurs, Fri, Sat, and Sun denote dummy variables for the respective weekdays (note that no model uses all seven of these). Individual numbers denote time points included in exogenous dummy variables for residual outliers. Per row, the better (lower) AIC and BIC scores are printed in boldface. Since the models were optimized for lower BIC scores, the comparison of AIC scores is less meaningful. In cases where the approaches differ with respect to applying a log transformation, we chose not to compare the models (meaning neither is printed in boldface).

Both approaches use the same diagnostic tests. Table 7.1 has a “pass all tests” column denoting if a model passes all diagnostic tests. Since models returned by Autovar always pass all diagnostic tests, if the value in this column is “No,” it means Autovar returned no models and the rest of the row is left empty. In the manual approach, if the experts found a model for which they considered the violation of the assumptions not severe enough as determined by manual inspection of the histograms of the residuals, they proceeded to use that model for their analysis. The “pass all tests” column uses boldface to denote that the model passes all diagnostic tests when the model of the other approach did not.

We note here that although the `VAR` function in R and the `var` function in STATA use different optimizations for determining the coefficients in a VAR model, the solutions are usually quite similar and their behavior is indistinguishable. In particular, we found that the results of the validity tests for the tested data sets are transferable. Thus, for each model that was found to be valid in R, we can construct a model in STATA with the same parameters that passes the validity tests in STATA.

Discussion

For the data sets used in this experiment, we find that Autovar outperforms experts working manually on average with respect to the BIC scores and the number of valid models found (Table 7.1). Autovar found a model that passes all diagnostic tests for 57 of the 60 data sets (95%) compared to 27 (45%) for the manual approach.

There were 18 data sets (30%) where the best model found by the approaches differed with respect to applying a log transformation. Of the remaining 42 data sets, there are 34 instances (81%) where Autovar had a lower (better) BIC score than the manual approach, and 8 instances (19%) where the manual approach had the lower BIC score.

For the 27 data sets for which both approaches found a valid model, there are 3 cases (11%) where Autovar favors a log-transformed model while the manual approach favors a model without log transformation. Cases where a valid model of the manual approach favored a log transformation while Autovar did not, did not occur. For the remaining 24 data sets where both approaches used the same log transformation setting, Autovar had the lower BIC score 22 times (91.7%) compared

Table 7.1: Comparison of best models found by Autovar vs. manual analysis

| Data set | Autovar | | | | | Manual | | | | |
|--------------|----------------------------------|-----------------------|---|-----------------|-----------------|----------------------------------|-----------------------|--|------------------|------------------|
| | pass lag all or- tests der | log trans- form | Exogenous variables | AIC | BIC | pass lag all or- tests der | log trans- form | Exogenous variables | AIC | BIC |
| 33 Stre Bowe | Yes 3 | No | Nr, Mon, Tues, Fri, 68, 83 | 1447.064 | 1475.797 | Yes 1 | No | 41, 68, 83 | 1475.591 | 1497.361 |
| 33 Stre Musc | Yes 3 | No | Mon, Tues, Wed, Fri, 41, 68 | 1393.013 | 1424.141 | Yes 2 | No | 41, 68 | 1421.313 | 1450.193 |
| 33 Stre Naus | Yes 2 | Yes | Mon, Tues | 219.017 | 238.271 | Yes 1 | Yes | - | 227.4712 | 251.6596 |
| 35 Stre Musc | Yes 3 | No | Tues, Thurs, 36, 42 | 1275.214 | 1296.764 | Yes 1 | No | 36, 42 | 1313.727 | 1325.821 |
| 35 Stre Head | Yes 3 | No | 36, 42 | 1375.835 | 1397.385 | No 1 | No | 36, 42 | 1410.244 | 1429.595 |
| 35 Stre Bowe | Yes 1 | No | Nr, 11, 36, 42 | 1253.775 | 1275.545 | No 1 | No | 36, 42 | 1264.472 | 1288.66 |
| 36 Stre Bowe | Yes 1 | No | Nr, 50 | 1263.114 | 1277.77 | Yes 1 | No | 50 | 1266.984 | 1286.525 |
| 36 Stre Join | Yes 2 | No | Tues, 50 | 1192.406 | 1209.422 | Yes 1 | No | 3, 50 | 1204.809 | 1229.236 |
| 36 Stre Head | Yes 3 | No | Tues, 50 | 1124.304 | 1138.817 | Yes 1 | No | 2, 50 | 1163.438 | 1187.865 |
| 38 Stre Musc | Yes 3 | No | - | 1179.635 | 1191.668 | Yes 2 | No | Sun, Mon, Thurs, Fri | 1162.923 | 1196.786 |
| 38 Stre Pelv | Yes 6 | No | Mon, Fri, 81 | 1111.614 | 1140.047 | Yes 4+11 | No | Mon, 81 | 1039.029 | 1078.198 |
| 38 Stre Dysp | Yes 2 | No | - | 1250.792 | 1262.886 | Yes 1+11 | No | Nr, Mon, Sat | 1089.812 | 1112.852 |
| 40 Stre Musc | Yes 3 | No | Nr, 33, 40, 47 | 1165.019 | 1191.358 | Yes 3 | No | Nr, 33, 40, Mon | 1164.511 | 1198.033 |
| 40 Stre Dysp | Yes 2 | Yes | Nr, 16 | 59.907 | 76.754 | No 2 | No | Nr, Mon, 8, 33, 64 | 1086.229 | 1112.703 |
| 40 Stre Tigh | Yes 1 | Yes | 2, 26, Mon, Tues, Fri | -75.466 | -51.277 | No 3 | No | Mon, Fri, 17 | 1207.781 | 1238.909 |
| 42 Stre Musc | Yes 3 | Yes | Nr, Mon, Tues, Wed, Thurs, Fri | 254.43 | 288.124 | No 2 | No | Nr, Sat, Sun, 84 | 1393.725 | 1427.589 |
| 42 Stre Dysp | Yes 3 | Yes | Sun, Mon, Tues, Wed, Thurs, Fri | 364.27 | 393.151 | No 1 | No | Sat, Sun, 6, 72, 78, 84 | 1268.017 | 1299.618 |
| 42 Stre Head | Yes 3 | Yes | Mon, Tues, Wed, Thurs, Fri | 354.263 | 385.55 | Yes 3 | No | Sun, Mon, Fri, Sat, 12, 84 | 1445.85 | 1491.578 |
| 44 Stre Bowe | Yes 3 | Yes | Nr | 291.685 | 303.718 | No 5 | No | Sat, Sun, 15, 21 | 1387.277 | 1437.3 |
| 44 Stre Join | Yes 3 | Yes | 35 | 293.354 | 315.014 | Yes 2 | Yes | - | 313.2308 | 337.4192 |
| 44 Stre Head | Yes 3 | Yes | - | 340.557 | 354.997 | No 1 | No | 6, 15, 21, 69 | 1486.348 | 1503.363 |
| 45 Stre Musc | Yes 3 | No | Yes 6, 61 | 277.265 | 298.704 | Yes 3 | Yes | 6, 61 | 279.6366 | 310.603 |
| 45 Stre Abdo | Yes 3 | Yes | Nr | 390.291 | 402.201 | Yes 1 | Yes | Nr | 401.9835 | 416.4239 |
| 45 Stre Dysp | Yes 1 | Yes | Nr | 384.989 | 397.023 | Yes 1 | Yes | Nr | 388.1983 | 407.4521 |
| 46 Stre Join | Yes 5 | Yes | Tues, 38, 43, 46, 61 | 281.12 | 297.323 | No 1 | Yes | 2, 5, 7, 43, 48, 62 | 272.1127 | 280.6212 |
| 46 Stre Abdo | No | - | - | - | - | No 1 | Yes | Sun, 2, 5, 7, 43, 48, 62 | 255.2248 | 265.8604 |
| 46 Stre Ches | Yes 7 | Yes | Sun, Mon, Tues, Wed, Thurs, 22, 43 | 185.978 | 220.409 | No 1 | Yes | Nr, 2, 5, 7, 43, 48, 62 | 195.6016 | 216.873 |
| 48 Stre Join | Yes 2 | No | - | 1397.809 | 1415.307 | Yes 2 | No | - | 1398.623 | 1418.621 |
| 48 Stre Musc | Yes 2 | No | - | 1400.564 | 1415.563 | Yes 2 | No | - | 1400.425 | 1417.924 |
| 48 Stre Abdo | Yes 2 | No | Fri | 1429.777 | 1447.276 | Yes 2 | No | - | 1431.71 | 1454.208 |
| 49 Stre Bowe | Yes 6 | No | Nr, Wed, Fri, 35 | 1385.754 | 1423.461 | No 1 | Yes | Nr, Sun, Mon, Tues, Thurs, Sat, 53, 58, 70 | 133.0106 | 171.7121 |
| 49 Stre Musc | Yes 3 | No | Nr, Fri, 35 | 1411.453 | 1430.609 | No 1 | No | Nr, 35 | 1446.143 | 1467.913 |
| 49 Stre Join | Yes 3 | No | Nr, Mon, Tues, Thurs, Fri, 35 | 1390.72 | 1417.059 | No 2 | No | Nr, Mon, 35 | 1409.505 | 1440.792 |
| 52 Stre Join | No | - | - | - | - | No 1 | Yes | Nr, Sun, Mon, Tues, Sat, 10, 26, 53 | 228.2698 | 260.1763 |
| 52 Stre Pelv | No | - | - | - | - | No 2 | Yes | Nr, Sat, 10, 26 | 202.4142 | 229.2833 |
| 52 Stre Naus | Yes 2 | Yes | Nr, Sun, Mon, Tues, Wed, Fri, 10, 26 | 279.807 | 313.671 | Yes 2 | Yes | Nr, Mon, Thurs, Fri, Sat, 10, 26, 53 | 269.3661 | 315.7765 |
| 53 Stre Naus | Yes 6 | Yes | Mon, Tues | 528.09 | 554.154 | No 2 | Yes | - | 563.0236 | 582.3743 |
| 53 Stre Musc | Yes 5 | Yes | - | 84.514 | 98.806 | No 4 | No | 39, 55, 80 | 1282.093 | 1308.432 |
| 53 Stre Numb | Yes 7 | Yes | - | 528.841 | 542.981 | No 3 | Yes | Nr | 579.7007 | 618.2082 |
| 54 Stre Abdo | Yes 7 | No | Nr, Sun, 18, 27, 37, 42 | 1321.805 | 1352.274 | No 1 | Yes | Nr, Sat, 21, 27, 42, 53, 71 | 157.7765 | 181.9649 |
| 54 Stre Musc | Yes 2 | No | Tues | 1427.58 | 1439.613 | Yes 1 | No | Nr, Tues, Sat | 1453.685 | 1480.292 |
| 54 Stre Tigh | Yes 2 | No | - | 1409.593 | 1424.033 | Yes 2 | No | Sat | 1407 | 1426.254 |
| 56 Stre Join | Yes 3 | Yes | Thurs | -18.881 | -4.368 | No 1 | No | Nr, Thurs, 35, 43 | 1351.284 | 1380.596 |
| 56 Stre Head | Yes 7 | No | Nr, Wed, Thurs, 43 | 1301.93 | 1342.21 | No 4 | No | Nr, Mon, Thurs, 35, 46 | 1342.746 | 1393.287 |
| 56 Stre Weak | Yes 3 | Yes | Mon, Tues, Thurs, 8, 59 | 19.58 | 41.349 | No 1 | Yes | Thurs, 8 | 33.43997 | 52.98118 |
| 57 Stre Musc | Yes 1 | Yes | Nr, Thurs, 38, 50, 90 | -180.172 | -160.263 | No 6 | Yes | Nr, 38, 50, 90 | -208.7957 | -184.4875 |
| 57 Stre Bowe | Yes 5 | Yes | Nr, 38, 50, 67, 90 | 181.821 | 198.837 | No 1 | Yes | Nr, 38 | 234.6111 | 254.5202 |
| 57 Stre Weak | Yes 2 | Yes | Nr, 38, 50, 90 | -63.58 | -46.239 | No 1 | No | Nr, 38, 90 | 1301.346 | 1326.232 |
| 58 Stre Bowe | Yes 3 | No | Nr | 1490.777 | 1508.038 | No 2 | No | Nr, Mon, Tues, Thurs, Fri, Sat | 1489.363 | 1531.478 |
| 58 Stre Join | Yes 2 | No | Nr, Sun, Mon, Tues, Wed, Thurs, Fri | 1435.548 | 1475.186 | No 1 | Yes | Nr, Sat | -26.95512 | -9.53467 |
| 58 Stre Back | Yes 2 | No | 5 | 1422.227 | 1439.488 | No 1 | No | Nr, Sat, 2, 5 | 1438.138 | 1463.024 |
| 60 Stre Abdo | Yes 3 | Yes | Nr | 112.373 | 131.627 | Yes 7 | No | Nr, Tues, Sat, 29 | 796.3394 | 855.2571 |
| 60 Stre Tigh | Yes 1 | Yes | Nr, 85 | 16.559 | 33.575 | No 1 | No | Nr, Tues, Fri, 15 | 915.4404 | 942.1793 |
| 60 Stre Head | Yes 1 | Yes | Nr | 179.448 | 194.033 | No 1 | No | Nr, Tues, Fri | 1014.066 | 1038.374 |
| 63 Stre Musc | Yes 5 | Yes | Sun, Tues, 14, 20, 47, 70, 74, 77, 85, 86 | 21.426 | 50.737 | No 3 | Yes | Sun, 14, 20 | 54.97506 | 99.56713 |
| 63 Stre Abdo | Yes 3 | Yes | Sun | 487.984 | 505.325 | Yes 1 | Yes | Sun | 505.9073 | 523.406 |
| 63 Stre Head | Yes 3 | Yes | Sun | 490.451 | 510.269 | Yes 3 | Yes | Sun, Fri | 487.6599 | 519.8653 |
| 64 Stre Abdo | Yes 2 | Yes | - | 458.686 | 471.611 | Yes 2 | Yes | - | 464.2884 | 487.5531 |
| 64 Stre Musc | Yes 1 | Yes | Mon, Wed | 204.924 | 223.09 | Yes 1 | No | Tues, Fri, Sat, 8, 26, 42 | 1647.414 | 1675.961 |
| 64 Stre Ches | Yes 2 | Yes | Thurs, 45, 57 | 373.265 | 391.288 | No 2 | No | Nr, 8, 26 | 1716.424 | 1752.614 |

to 2 times (8.3%) for the manual approach. In both instances where the manual approach had the lower BIC score, this was due to using a high lag (11) that is outside the search range of Autovar.

There are 14 instances where the manual approach reaches a lower AIC score than Autovar. However, this result is not unexpected because both approaches optimize for low BIC scores, thus having a lower AIC score but a higher BIC score is the result of setting suboptimal constraints.

Surprisingly, in all 5 instances (18.5% of 27) where the lag order, log-transform, and exogenous variables are identical for both approaches, Autovar still reached a lower BIC score because of a difference in the constraints used. In these cases, Autovar has one or two different constraints that result in a slightly lower BIC score. These results suggest that the added complexity of our constraint-finding method in practice may frequently result in better constraints. Another surprising result is that the built-in preference of Autovar for favoring models with fewer masked outliers did not result in significantly higher AIC/BIC scores on average.

While not shown in the results, we note that in 21 out of 27 cases (77.8%) where both approaches find a valid model, Autovar also found a model at the same lag order and with the same log-transform setting as the manual model (with the only differences being in the exogenous variables and the constraints). One of these cases (42 *Stre Head*) was the only tested case where setting a constraint that invalidates the model would result in a valid model (with a lower BIC score than the solution of Autovar) by adding more constraints. This finding supports our implicit assumption that such constraint combinations occur infrequently in practice. Reasons for Autovar not finding certain models are due to the manual approach using higher lags or different outliers (i.e., there is one instance where a mistake was made in calculating the set of outliers in the manual approach which resulted in a valid model). The number of valid models missed because of constraining only valid models is not reflected in these results, as both Autovar and the experts applied constraints only to valid models.

7.2.2 Performance

Next, we consider aspects of time complexity, memory complexity, and scalability of our approach.

Time complexity

The minimum number of models evaluated by Autovar is $O(4l)$, where l is the number of lags to consider, i.e., $\text{max_lag} - \text{min_lag} + 1$. The factor $4 = 2^2$ follows from considering at most 2 options for applying a log transformation and 2 options

for including weekday dummy variables. In the worst case, if the stability test fails for all initial models, we need to evaluate twice this number of models. In addition, for each of the stable models, we may need to evaluate an additional set of models depending on the outcome of the residual diagnostic tests. Thus, the total number of models that is evaluated is $O(4l + 4l4^k)$, with k the number of endogenous variables, since we may need to consider all possible subsets of outliers for up to 3 iterations. Since we are estimating a VAR model in every step, which is a costly operation, k cannot be too large. Adding one endogenous variable to the system will cause Autovar to take about four times as long to evaluate all models. We have tested Autovar with $k = 2$ and $k = 3$, and it typically runs between 1 and 3 seconds for $k = 2$ and up to a minute for $k = 3$, measured as single-threaded run time on an i7 PC at 3.5GHz. We have not tested Autovar with $k \geq 4$.

The maximum number of valid models returned by Autovar is $O(8l4^k)$. The derivation of this bound follows the reasoning above, and taking into account that for every valid model we also return a constrained version. The different iterations of outliers are often mutually exclusive, so the full 4^k subsets of models will rarely, if ever, all be estimated. In practice, we of course find that the number of valid models returned by Autovar is far lower. For example, for the data sets shown in Table 7.1, where $k = 2$ and $l = 8$, the average number of valid models returned by Autovar per data set is 8.07 with a standard deviation of 5.17 and a maximum of 27.

A significant portion of the running time is spent on finding constraints for the valid models found. Following the above reasoning we find that an upper bound on the number of models to be restricted is $O(4l4^k)$. Recall from Section 6.6 that the constraint-setting procedure has $O(n^2)$, with n the number of terms in the equations. Since there are k equations, the number of terms in the equations is k times the number of terms in one equation. In the unconstrained models, each of the k endogenous variables appears with all its l lags in each equation. It follows that the total number of terms in an unconstrained model is $O(lk^2)$. With an $O(n^2)$ complexity for setting constraints, in the worst case we perform $O(l^2k^4)$ full VAR model estimations for every valid unconstrained model. To put these numbers in perspective, for, e.g., a model with $k = 3$, $l = 6$, and having found 3 valid unconstrained models, we spend around half the running time on constraining the 3 valid models found and the other half on assessing the validity of all models under consideration.

Memory complexity

Our approach requires the implementation to retain a list of all model configurations in memory. We need to distinguish between 2 options for applying a log transformation, 2 options for including weekday dummy variables, 2 options for applying restrictions, and 2 options for trend variable inclusion. In addition, we

need to encode the lag order of the system and the iterations for masking outliers for the k endogenous variables.

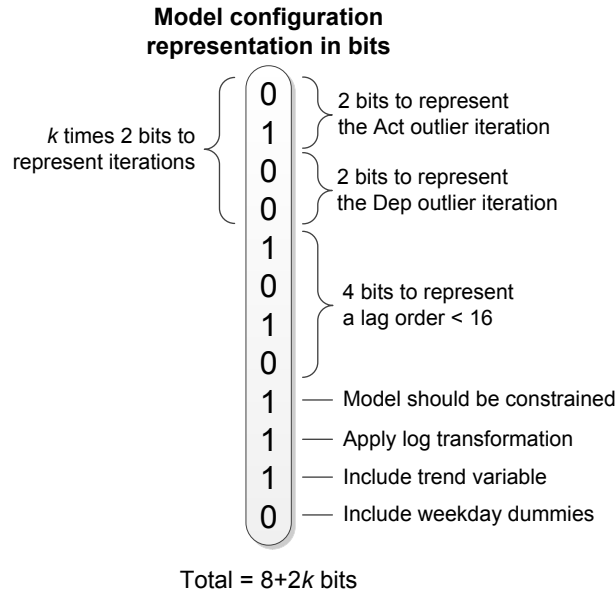


Figure 7.6: Encoding a model configuration as an integer number. A total of $2k + 8$ bits (with k the number of endogenous variables) is required to distinguish between all possible model configurations.

Figure 7.6 shows how model configurations can be represented as integer numbers. The iterations for masking outliers for the different equations can be encoded as a 2 bit number because the iterations range from 0 to 3, inclusive. For a system with two variables, we find that $2 \cdot k + 8 = 2 \cdot 2 + 8 = 12$ bits are needed to represent each possible model configuration. If we encode model configurations as numbers indexing into a Boolean array, this array would need to have a size of $2^{12} = 4096$. If we assume that 1 byte of memory is used per element in a Boolean array, when $k = 2$, retaining the “processed” state of all model configurations requires 4KB of memory. However, to accommodate debugging, our implementation in R is less space efficient.

To generate its output, our approach also needs to retain the valid VAR model estimations in memory. From the time complexity analysis we know that our approach finds $O(8l4^k)$ valid models. The size of the estimated models is implementation-dependent and varies in practice, but includes at least the coefficients of the terms of the formula. On the assumption that the storage size for a model estimation grows linearly in relation to the number of coefficients in the

model, the memory size for a model estimation scales with $O(lk^2)$.

Scalability

For finding and outputting models for all 60 data sets of Section 7.2.1 on an i7 PC at 3.5GHz, Autovar required around 25 minutes single-threaded execution time in total. This does not include the approximate 10 minutes that the authors needed to write an R script to process all data sets in sequence using Autovar. In comparison, the analysis of the experts working manually required several working days.

While not exploited in the current implementation of Autovar, our approach for constructing and evaluating VAR models (Algorithm 6.1) allows for parallelization. The conditions are that all access to queue Q and result list R must be synchronized by mutual exclusion. If each initial model configuration and the variations thereof were to be executed in parallel (requiring at least $4l$ processors), then assessing the validity of all models takes $O(1 + 4^k)$ time. If we may assume that $k \leq 3$, assessing the validity of all models can be performed in constant time, with a constant factor of at most 65 VAR model estimations per processor. However, reducing the complexity of or introducing parallelization to the constraint-setting procedure is more difficult and remains a bottleneck in our approach. Even if all valid models were constrained on different processors, each processor would still have to perform $O(l^2k^4)$ full VAR model estimations.

7.3 Related Work

The findings of the current study are consistent with those of Hendry and Krolzig (2001), who found that automatic modeling techniques can perform on a competitive level with experts working manually. However, previous work warns for an approach based on “data mining” for models as it could potentially lead to random models passing tests by chance (Owen 2003). This issue applies to Autovar as well. However, the relatively low number of models that Autovar evaluates on average combined with the low probability of a random model passing all three tests render it unlikely that any random models passed the tests for the data sets we tested on. Autovar performs three tests at a 0.05 significance level, and if we were to assume that all three tests are independent, then there is a probability of $0.05^3 = 0.0125\%$ of a model randomly passing all three tests. That translates into evaluating 8000 models on average before we expect to see one random model passing all tests. For the data sets of Table 7.1, the maximum number of distinct models we tested for any particular data set was 237 (with an average of 63.8). However, if we assume a worst-case scenario in which two of the three tests are fully statistically dependent, the probability of a model passing all tests randomly becomes

$0.05^2 = 0.25\%$ or 1 in 400 models, which makes the event more probable. This is one of the reasons why Autovar returns not one best model, but all valid models found, along with summary statistics to show the user which model configuration settings are common among the valid models. Returning multiple valid models instead of just one is one of the main distinctions between Autovar and other approaches to automated model selection. We consider it to be one of its main contributions because a list of all valid models found for a data set grants more insight into the properties of the valid models than a single model does. For example, if we want to determine whether a certain Granger causality is present in a data set, an approach that returns a single model could only base its answer on the relations found in that model, while Autovar can average over all valid models found and answer in the form of a probability.

7.3.1 PcGive

Here we present a comparison of the functionality of Autovar to that of PcGive. To the best of our knowledge, PcGive (previously PcGets (Owen 2003)) is currently the only other software that can perform fully automated VAR model fitting. RETINA (Perez-Amaral et al. 2003) is another known implementation for automated model selection but is not suited for vector autoregression. Other software exists for modeling vector autoregression, e.g., Eviews (Vogelvang 2005), Mathematica (*ARProcess in Mathematica* 2013), Matlab (*Vector Autoregressive Models in Matlab* 2013), TSP (*The VAR function in TSP* 2013), GAUSS (*GAUSS: Time Series MT* 2013), gretl (Baiocchi and Distaso 2003, Rosenblad 2008), SHAZAM (White and McRae 1987, *SHAZAM features* 2013), R (Pfaff 2008) (also available in sage and S-PLUS (Venables et al. 1994)), LIMDEP and NLOGIT (Hilbe 2006), Stata (Baun 2006, *STATA: Data Analysis and Statistical Software* 2013), RATS (Doan 2010), and Microfit (Pesaran and Pesaran 2010), but these programs do not feature automated model selection. There are, however, frameworks that provide a theoretical basis for an automated approach to model selection. Pesaran and Timmermann (2000) describe a non-sequential approach with specific-to-general aspects (Owen 2003), and Phillips provides the basis for a Bayesian framework for automated model selection (Phillips 1996).

Our original goal was to compare the performance of PcGive to Autovar, but a fair comparison proved impossible. This is because Autovar and PcGive use different tests to assert the validity of the models, thus, e.g., models that are considered valid in PcGive fail tests in Autovar (and vice versa). As such, comparing AIC/BIC scores of winning models between the two programs is unfair because a winning model in PcGive may also have been found in Autovar, yet have been discarded

because it failed one of its validity tests.

Table 7.2: Comparing the functionality of Autovar and PcGive

| | Autovar | PcGive 14 |
|------------------------------|---|---|
| Approach | Exhaustive search restricted by statistical tests. | General-to-specific modeling strategy. |
| Model-selection results | Multiple valid models. | A single best model. |
| Additional results | Granger causality summary, Contemporaneous correlation summary, model configuration summary statistics, plots of input variables, test results. | Test results for the model returned, plots of input variables, forecasts, simulation and impulse response, dynamic analysis, cointegration tests. |
| Max. lag setting | Yes | Yes (set per variable) |
| Zero-order lag models | Yes | Yes |
| Outlier detection | Large residuals. | Large residuals, impulse indicator saturation, or step indicator saturation. |
| Automatic outlier variables | Yes | Yes (with linear combinations) |
| Automatic weekday variables | Yes | No |
| Automatic day-segments vars. | Yes | No |
| Automatic trend inclusion | Yes (by Phillips-Perron test) | No |
| Automatic log-transforms | Yes | No |
| Automatic constraints | Yes (equation-specific) | Yes |
| Portmanteau test | Yes | Yes |
| Homoskedasticity test | Yes | Yes |
| Normality test | Yes | Yes |
| Chow test | No | Yes |
| Stability test | Yes | No |
| Validity test inclusion | Not configurable | Configurable |
| Automatic data imputation | Very limited | No |
| Scripting support | Yes (R script) | Yes (OxMetrics batch language) |
| Modeling non-VAR systems | Not supported | Supported |
| Data input formats supported | STATA, SPSS | STATA, Excel, *.csv |

Instead, we compare the programs based on their functionality, as shown in Table 7.2. This table compares features and functionality (left column) of Autovar (middle column) to those of PcGive (right column). We base our comparison on PcGive 14, which was released in June 2013.

7.3.2 Comparison

From Table 7.2, we see that PcGive is a more extensive software suite. It supports not only VAR modeling but various other statistical models as well. Furthermore, it not only finds models but can also apply them, for example in forecasts and impulse response simulations. Autovar, on the other hand, is easier to use and incorporates more automation. It features automatic creation and inclusion of seasonal dummy variables for weekdays and day segments, of trend variables, of log transformations of the data, and of constraints specific per VAR equation. These aspects of automation make Autovar easier to use because in most cases a user can just access the web application, upload a data set, select the VAR columns and click “Run.” With respect to configurability, PcGive favors an approach of extensive configurability that relies on the expertise of the user in specifying the proper settings, while Autovar prefers an approach of automatically trying to determine which settings to use for a data set, having embedded the expertise in its algorithms for finding models. Another important distinction is that Autovar discards any models that fail any of the tests

while PcGive always finds and returns a best model, even when it is not valid.

7.4 Discussion

With the recent developments of widespread portable consumer electronics devices being used as a means of data collection in healthcare, we investigated whether a fully automated approach to vector autoregression is possible that does not require statistical expertise to operate, while still closely resembling the logic and decision-making of statisticians working manually. The existing alternative follows a general-to-specific (Hendry and Krolzig 2001) approach that is different from the approach implemented in Autovar, and it does not automate some of the key operations that a statistician might perform when working manually (e.g., log-transforming a data set or including dummy variables for weekdays). Autovar leverages the power of automation to consider more potential models and to improve on the manual process by developing a novel way for finding better constraints. Autovar serves as a proof of concept, and in this chapter we compared its performance against experts working manually, and its features against those of commercially available software (PcGive).

The results need to be interpreted with caution because the performance does not necessarily generalize to other manual analyses or data sets. Autovar needs to undergo simulation studies and statistical evaluation in order to assess the properties of the approach and to determine whether the approach is useful outside the context of patient diary data. Also note that, for patient diary data in particular, VAR analysis may not be accurate when measurements are obtained at unequal intervals. Autovar currently has no functionality to preprocess the data to account for unequal intervals, and only very limited support for imputing missing values. With regards to the comparisons performed in the current study, we note that AIC/BIC scores are not the only measure of fit for ranking models. For example, the model with the best predictions is not necessarily the model that has the best fit on the current data (Lütkepohl 2005, pp. 62). Moreover, in practice, a model that does not pass all validity tests can still be useful if it is reasonably close to passing those tests. These considerations are often taken into account by human experts. Because Autovar discards any models that fail any of the tests, its performance depends on the particular set of validity tests chosen, and since this set is not configurable, the flexibility of the approach is heavily limited.

The total volume of electronic medical data around the world is increasing rapidly, allowing for new and exciting applications to change the way we think about care. We set out to find answers to the questions of which aspects of care that involve knowledge sharing can be automated, and how this automation can be performed. Our work focused on automating two aspects of care that traditionally require human supervision. These aspects are generating personalized advice for schizophrenia patients and finding the best vector autoregression model for electronic patient diary data.

8.1 Summary

Wegweis has set the trend by providing schizophrenia patients with direct access to and automated recommendations based on their assessment results. Our findings suggest that an approach based on problem severities is suitable for identifying important problem areas from schizophrenia-related questionnaires, and that such an approach can be considered helpful and relevant by patients in selecting and ranking advice.

Our findings have important implications for the development of systems that automate the translation and interpretation of assessment results for patients with chronic illnesses. If such systems can be shown to work for schizophrenia patients, who impose numerous restrictions on the user interface, then these systems are likely to work for patients with other chronic illnesses too. In those branches of healthcare, this paves the way for automated solutions that support the sharing of information between patient and clinician as an integral part of shared decision making.

The present results are significant because they demonstrate the efficacy of an intuitive way to prioritize information in the same way as a clinician would. However, our approach does not explain the relevance selection of the patients very well, leaving room for improvement.

Our second project, Autovar, automates the interpretation of time series data. The most important implication of Autovar is in making vector autoregression fea-

sible on a large scale. Current manual approaches may require days for analyzing a single data set, i.e., they function on a small scale only. Likewise, other automated approaches work only on a small scale because their operation still requires a background in statistics. This is because applying, and determining the applicability of, certain actions, such as log-transforming the data, including a trend, or creating seasonal dummy variables, is not covered by automation in other automated approaches. Scaling any of the current alternatives, including any manual approach, to process multiple data sets in parallel would require employing multiple statisticians, which is expensive. Autovar, on the other hand, can perform the same tasks in minutes and does not require statistical expertise because its operation can be fully automated with trivial efforts (e.g., a line of R code to call Autovar with a filename). Thus, Autovar can work on a large scale at merely the cost of hardware.

Autovar is a demonstration of an exhaustive approach for VAR model selection that is relatively safe to use. The requirement is that there is enough logic implemented to restrict the search space for models to the extent where the possibility of random models passing tests by chance is virtually nil. Under this assumption, performing large scale VAR model analysis without a background in statistics appears feasible, and a widespread application of fast and easy automated VAR analysis in healthcare could benefit more patients.

8.2 Future work and open issues

Our research has raised many questions in need of further investigation. Specific to Wegweis, more experiments are needed to determine how questionnaires other than the MANSA would score in the experiments. Another issue worth investigating is the extent to which clinicians take the patient history into account when identifying important problems, and how this can be modeled.

Another unaddressed question is how to make the advice rankings match the patient opinions more closely. An approach that takes previous assessments into account may help to construct a more complete image of a patient and would allow for reasoning over changes in the condition of a patient over time. While we are aware that some work has been started in this area (*Eigen Regie Bij Schizofrenie* 2011), we believe that these efforts could benefit from the added robustness of an ontology-based approach.

An open issue is the question of how the knowledge base should be maintained, i.e., how the advice should be kept up-to-date. Any system with a sufficiently large and knowledgeable user base could perhaps be self-moderating and self-sustaining. In smaller settings, the responsibility for maintenance lies with the care organizations.

The utility of Autovar can be improved by adding ways to use the models directly. For example, *forecasts* and *impulse response functions* allow us to predict how a system would react to an introduced shock. For the model with activity and depression, if we find that inactivity Granger causes depression, by using impulse response analysis, we could determine the number of days during which depressive symptoms are relieved as a result of one hour of exercise. A physician can use this information to inform a patient of the exact duration and frequency of physical exercise for it to have an optimal effect.

We showed that Autovar allows for easy parallelization, reducing the complexity for determining all valid models from $O(4l + 4l4^k)$ to $O(1 + 4^k)$ on $4l$ processors, with k endogenous variables and lag length l . If Autovar were to be used on a larger scale, support for parallel computation on multiple cores or in a cluster would need to be added. The R language originally is single-threaded, but support for parallel computing is added through certain packages (e.g., through the `parallel` package that was added in 2.14.0).

8.3 Outlook

In order to keep healthcare accessible and affordable in the coming decades, we believe that routine aspects of care that are derivative of electronic medical data will need to be fully automated. These aspects encompass a substantial part of care, in particular for patients with chronic illnesses. We have seen that in the past, the main role of computer applications and artificial intelligence in medicine has been to support the clinician, and that the focus is currently shifting toward applications that support the patient.

In this thesis, we have demonstrated the efficacy of systems for automated information processing and interpretation for both patients and clinicians. Potential benefits of automating aspects of care include wider availability, lower healthcare costs, a more consistent level of care between different organizations, fewer medical errors, better informed patients, and less time of healthcare professionals spent on doing routine analysis.

On the patient side, efforts should be directed toward patient-centered self-management applications that provide advice and explain test results and treatment options without requiring human intervention. For the type of information that houses potential risks, ontologies and rule sets should be employed to ensure that generated advice is in accordance with treatment policies. For automated systems to provide information to patients directly, the ontologies, rule sets, and advice contents need to be verified by experts.

On the side of the clinicians, procedures in treatment protocols that involve the

(statistical) analysis of electronic medical data should be analyzed, explicated, and automated to a level where human expertise is no longer required for day-to-day operations. Here, confidence intervals should rule out nonsensical outcomes. The extracted knowledge can be conveyed to both clinicians and patients.

Before coming to rely solely on automated reports and recommendations, we envision a transitional period wherein such systems are developed, trained, and used as a *second opinion*. For example, advice systems can complement the recommendations of clinicians, and systems for automated data analysis can help statisticians evaluate their own conclusions. For both types of systems, training can occur either explicitly or implicitly using aspects of machine learning.

The global development in automating aspects of healthcare can be sped up significantly by prioritizing interoperability. For hospitals in particular, standardization of electronic medical data seems key. All automated systems use electronic medical data as input. If this data adheres to globally accepted standards, then automated systems become usable worldwide.

Bibliography

- Adlassnig, K., Combi, C., Das, A., Keravnou, E. and Pozzi, G.: 2006, Temporal representation and reasoning in medicine: Research directions and challenges, *Artificial Intelligence in Medicine* **38**(2), 101–113.
- Ainsworth, M.: 2002, *My life as an e-patient*, e-Therapy. Case studies, guiding principles, and the clinical potential of the internet, W.W. Norton & Company, New York/London, pp. 194–215.
- Akaike, H.: 1974, A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**(6), 716–723.
- Altman, D.: 1991, *Practical statistics for medical research*, Chapman & Hall, London.
- American Psychiatric Association: 2000, *Diagnostic and statistical manual of mental disorders: DSM-IV*, American Psychiatric Publishing, Inc.
- Anderson, P. A.: 1979, Help for the regional economic forecaster: Vector autoregression, *Federal Reserve Bank of Minneapolis Quarterly Review* **3**(3), 2–7.
- Andry, F., Freeman, L., Gillson, J., Kienitz, J., Lee, M., Naval, G. and Nicholson, D.: 2008, Highly-Interactive and User-Friendly Web Application for People with Diabetes, *IEEE International Conference on Communication Systems (HEALTHCOM 2008)*, pp. 118–120.
- ARProcess in Mathematica*: 2013, <http://reference.wolfram.com/mathematica/ref/ARProcess.html>. (Accessed: 11 December 2013).
- Arsand, E. and Demiris, G.: 2008, User-centered methods for designing patient-centric self-help tools, *Informatics for health & social care* **33**(3), 158–169.

- Augusto, J.: 2005, Temporal reasoning for decision support in medicine, *Artificial Intelligence in Medicine* **33**(1), 1–24.
- Augusto, J. C. and McCullagh, P.: 2007, Ambient intelligence: Concepts and applications, *Computer Science and Information Systems/ComSIS* **4**(1), 1–26.
- Auramo, Y. and Juhola, M.: 1996, Modifying an expert system construction to pattern recognition solution, *Artificial Intelligence in Medicine* **8**(1), 15–21.
- Autovar: *GitHub repository*: 2013, <https://github.com/roqua/autovar>. (Accessed: 14 October 2013).
- Baiocchi, G. and Distaso, W.: 2003, GRETl: Econometric software for the GNU generation, *Journal of Applied Econometrics* **18**(1), 105–110.
- Barlow, J. H., Ellard, D. R., Hainsworth, J. M., Jones, F. R. and Fisher, A.: 2005, A review of self-management interventions for panic disorders, phobias and obsessive-compulsive disorders, *Acta Psychiatrica Scandinavica* **111**(4), 272–285.
- Barry, M. J. and Edgman-Levitan, S.: 2012, Shared decision making – the pinnacle of patient-centered care, *New England Journal of Medicine* **366**(9), 780–781.
- Baun, C.: 2006, *An introduction to modern econometrics using Stata*, Stata Press.
- Beebe, L. H., Smith, K., Crye, C., Addonizio, C., Strunk, D. J., Martin, W. and Poche, J.: 2008, Telenursing intervention increases psychiatric medication adherence in schizophrenia outpatients, *Journal of the American Psychiatric Nurses Association* **14**(3), 217–224.
- Bell, V., Grech, E., Maiden, C., Halligan, P. W. and Ellis, H. D.: 2005, ‘internet delusions’: a case series and theoretical integration, *Psychopathology* **38**(3), 144–150.
- Belsley, D. A., Kuh, E. and Welsch, R. E.: 2004, *Regression diagnostics: Identifying influential data and sources of collinearity*, Vol. 546 of *Wiley Series in Probability and Statistics*, John Wiley & Sons.
- Bender, D. and Sartipi, K.: 2013, HL7 FHIR: An agile and RESTful approach to healthcare information exchange, *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, IEEE, pp. 326–331.
- Bichindaritz, I. and Marling, C.: 2006, Case-based reasoning in the health sciences: What’s next?, *Artificial Intelligence in Medicine* **36**(2), 127–135.
- Bichindaritz, I. and Montani, S.: 2011, Guest editorial: Advances in case-based reasoning in the health sciences, *Artificial Intelligence in Medicine* **51**(2), 75–79.

- Bickmore, T. W., Puskar, K., Schlenk, E. A., Pfeifer, L. M. and Sereika, S. M.: 2010, Maintaining reality: relational agents for antipsychotic medication adherence, *Interacting with Computers* **22**(4), 276–288.
- Blobel, B.: 2006, Advanced and secure architectural EHR approaches, *International Journal of Medical Informatics* **75**(3–4), 185–190.
- Blobel, B.: 2007, Comparing approaches for advanced e-health security infrastructures, *International Journal of Medical Informatics* **76**(5–6), 454–459.
- Blobel, B., Nordberg, R., Davis, J. and Pharow, P.: 2006, Modelling privilege management and access control, *International Journal of Medical Informatics* **75**(8), 597–623.
- Blobel, B. and Pharow, P.: 2009, Analysis and evaluation of EHR approaches, *Methods of Information in Medicine* **48**(2), 162–169.
- Bond, A., Hacking, A., Milosevic, Z. and Zander, A.: 2013, Specifying and building interoperable ehealth systems: ODP benefits and lessons learned, *Computer Standards & Interfaces* **35**(3), 313–328.
- Bos, L., Marsh, A., Carroll, D., Gupta, S. and Rees, M.: 2008, Patient 2.0 Empowerment, *Proceedings of the 2008 International Conference on Semantic Web & Web Services SWWS*, pp. 164–167.
- Box, G. E., Jenkins, G. M. and Reinsel, G. C.: 1976, *Time series analysis: forecasting and control*, Holden-Day, San Francisco, CA.
- Brailer, D. J.: 2005, Interoperability: the key to the future health care system, *Health Affairs* **24**, W5.
- Brey, P.: 2005, Freedom and privacy in ambient intelligence, *Ethics and Information Technology* **7**(3), 157–166.
- Brown, A. and Wehl, B.: 2011, An update on google health and google powermeter, <http://googleblog.blogspot.nl/2011/06/update-on-google-health-and-google.html>. (Accessed: 18 June 2013).
- BrowserCMS: 2011, <http://www.browsercms.org>. (Accessed: 18 November 2012).
- Brunette, M. F., Ferron, J. C., McHugo, G. J., Davis, K. E., Devitt, T. S., Wilkness, S. M. and Drake, R. E.: 2011, An electronic decision support system to motivate people with severe mental illnesses to quit smoking, *Psychiatric services (Washington, D.C.)* **62**(4), 360–366.

- Bull, L., Bernadó-Mansilla, E. and Holmes, J.: 2008, Learning Classifier Systems in Data Mining: An Introduction, *Computational Intelligence (SCI)* **125**, 1–15.
- Buranarach, M., Supnithi, T., Chalortham, N., Khunthong, V., Varasai, P. and Kawtrakul, A.: 2009, A semantic web framework to support knowledge management in chronic disease healthcare, in F. Sartori, M. Sicilia and N. Manouselis (eds), *Metadata and Semantic Research*, Vol. 46 of *Communications in Computer and Information Science*, Springer Berlin, Heidelberg, pp. 164–170.
- Burbidge, J. and Harrison, A.: 1984, Testing for the effects of oil-price rises using vector autoregressions, *International Economic Review* **25**(2), 459–484.
- Burton, C., Weller, D. and Sharpe, M.: 2009, Functional somatic symptoms and psychological states: an electronic diary study, *Psychosomatic medicine* **71**(1), 77–83.
- Charles, C., Gafni, A. and Whelan, T.: 1997, Shared decision-making in the medical encounter: what does it mean? (or it takes at least two to tango), *Social Science & Medicine* **44**(5), 681–692.
- Chudyk, A. M., Jutai, J. W., Petrella, R. J. and Speechley, M.: 2009, Systematic review of hip fracture rehabilitation practices in the elderly, *Archives of Physical Medicine and Rehabilitation* **90**(2), 246–262.
- Cicchetti, D. V.: 1994, Guidelines, criteria, and rules of the thumb for evaluating normed and standardized assessment instruments in psychology, *Psychological Assessment* **4**(284), 290.
- Cohen, J.: 1988, *Statistical power analysis for the behavioral sciences*, 2nd edn, Lawrence Earlbaum Associates, Hillsdale, NJ.
- Coiera, E.: 2003, *The Guide to Health Informatics*, Arnold, London.
- Combi, C., Cucchi, G. and Pincioli, F.: 1997, Applying object-oriented technologies in modeling and querying temporally oriented clinical databases dealing with temporal granularity and indeterminacy, *IEEE Transactions on Information Technology in Biomedicine* **1**(2), 100–127.
- Conner, T. S., Barrett, L. F. and Tugade, M. M.: 2007, Idiographic personality: The theory and practice of experience sampling, in R. W. Tennen, Howard Robins, R. C. Fraley and R. F. Krueger (eds), *Handbook of research methods in personality psychology*, Guilford Press, New York, NY, pp. 79–96.
- Cook, D. J., Augusto, J. C. and Jakkula, V. R.: 2009, Ambient intelligence: Technologies, applications, and opportunities, *Pervasive and Mobile Computing* **5**(4), 277–298.

- Cooper, A. and Reimann, R.: 2003, Implementation models and mental models, in A. Cooper and R. Reihmann (eds), *About Face 2.0: the essentials of user interface design*, Wiley Publishing, Indianapolis, pp. 21–32.
- Cooper, G.: 1993, Probabilistic and decision-theoretic systems in medicine, *Artificial intelligence in medicine* **5**, 289–292.
- Cousineau, D. and Chartier, S.: 2010, Outliers detection and treatment: a review, *International Journal of Psychological Research* **3**(1), 58–67.
- Crockford, D.: 2006, The application/json media type for javascript object notation (json), Available from: <https://tools.ietf.org/html/rfc4627>. (Accessed: 18 November 2012).
- Crutzen, C. K.: 2007, Ambient intelligence, between heaven and hell. a transformative critical room?, *Gender Designs IT*, Springer, pp. 65–78.
- D’Agostino, R. B., Belanger, A. and D’Agostino Jr, R. B.: 1990, A suggestion for using powerful and informative tests of normality, *The American Statistician* **44**(4), 316–321.
- Deegan, P. and Drake, R.: 2006, Shared decision making and medication management in the recovery process, *Psychiatric Services* **57**(11), 1636–1639.
- Deegan, P. E.: 1997, Recovery and empowerment for people with psychiatric disabilities, *Social work in health care* **25**(3), 11–24.
- Deegan, P., Rapp, C., Holter, M. and Riefer, M.: 2008, Best practices: a program to support shared decision making in an outpatient psychiatric medication clinic, *Psychiatric Services* **59**(6), 603–605.
- Dempster, A. P., Laird, N. M. and Rubin, D. B.: 1977, Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 1–38.
- Depp, C. A., Mausbach, B., Granholm, E., Cardenas, V., Ben-Zeev, D., Patterson, T. L., Lebowitz, B. D. and Jeste, D. V.: 2010, Mobile interventions for severe mental illness: design and preliminary data from three approaches, *Journal of Nervous & Mental Disease* **198**(10), 715–721.
- Diebold, F. X.: 1998, *Elements of forecasting*, South-Western.
- Dietterich, T., Domingos, P., Getoor, L., Muggleton, S. and Tadepalli, P.: 2008, Structured machine learning: the next ten years, *Machine Learning* **73**(1), 3–23.
- Doan, T. A.: 2010, *Rats, Version 8: User’s Guide*, Estima.

- Dolin, R., Alschuler, L., Boyer, S., Beebe, C., Behlen, F., Biron, P. and Shvo, A.: 2006, HL7 clinical document architecture, release 2, *Journal of the American Medical Informatics Association* **13**(1), 30–39.
- Downs, S. H. and Black, N.: 1998, The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions, *Journal of epidemiology and community health* **52**(6), 377–384.
- Drake, R. and Deegan, P.: 2009, Shared decision making is an ethical imperative, *Psychiatric Services* **60**(8), 1007.
- Duncan, E., Best, C. and Hagen, S.: 2008, Shared decision making interventions for people with mental health conditions, *Cochrane Database of Systematic Reviews* **3**.
- Eigen Regie Bij Schizofrenie*: 2011, <http://www.eigen-regie.nl>. (Accessed: 18 November 2012).
- Emerencia, A., van der Krieke, L., Bos, E., de Jonge, P., Petkov, N. and Aiello, M.: 2014, Automating vector autoregression on electronic patient diary data, (Submitted).
- Emerencia, A., van der Krieke, L., Petkov, N. and Aiello, M.: 2011, Assessing schizophrenia with an interoperable architecture, in M.-M. Bouamrane and C. Tao (eds), *Proceedings of the first international workshop on Managing interoperability and complexity in health systems*, ACM, New York, NY, pp. 79–82.
- Emerencia, A., Van der Krieke, L., Sytema, S., Petkov, N. and Aiello, M.: 2013, Generating personalized advice for schizophrenia patients, *Artificial intelligence in medicine* **58**(1), 23–36.
- Farrell, S. P., Mahone, I. H. and Guilbaud, P.: 2004, Web technology for persons with serious mental illness. SO: Arch Psychiatr Nurs. 2004 Aug;18(4):121-5.
- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R.: 1996, *Advances in knowledge discovery and data mining*, MIT press.
- Ferranti, J., Musser, R., Kawamoto, K. and Hammond, W.: 2006, The clinical document architecture and the continuity of care record: a critical analysis, *Journal of the American Medical Informatics Association* **13**(3), 245–252.
- Fisher, M. D., Gabbay, D. M. and Vila, L.: 2005, *Handbook of temporal reasoning in artificial intelligence*, Vol. 1, Elsevier.

- foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, dBase, ...*: 2013, <http://cran.r-project.org/web/packages/foreign/index.html>. (Accessed: 10 September 2013).
- Frangou, S., Sachpazidis, I., Stassinakis, A. and Sakas, G.: 2005, Telemonitoring of medication adherence in patients with schizophrenia, *Telemedicine journal and e-health : the official journal of the American Telemedicine Association* **11**(6), 675–683.
- Frank, A. F. and Gunderson, J. G.: 1990, The role of the therapeutic alliance in the treatment of schizophrenia: relationship to course and outcome, *Archives of general psychiatry* **47**(3), 228–236.
- Friedewald, M., Vildjiounaite, E., Punie, Y. and Wright, D.: 2007, Privacy, identity and security in ambient intelligence: a scenario analysis, *Telematics and Informatics* **24**(1), 15–29.
- Fullwood, C., Kennedy, A., Rogers, A., Eden, M., Gardner, C., Protheroe, J. and Reeves, D.: 2013, Patients experiences of shared decision making in primary care practices in the United Kingdom, *Medical Decision Making* **33**(1), 26–36.
- GAUSS: *Time Series MT*: 2013, <http://www.aptech.com/products/ gauss-applications/time-series-mt/>. (Accessed: 11 December 2013).
- Gené Badia, J., Grau, I., Sánchez, E. and Bernardo, M.: 2009, Forumclínic: the virtual community for chronic patients, *Journal of Health Innovation and Integrated Care* **1**(1), 1–6.
- Gennari, J. H., Musen, M. A., Fergerson, R. W., Grosso, W. E., Crubézy, M., Eriksson, H., Noy, N. F. and Tu, S. W.: 2003, The evolution of protégé: an environment for knowledge-based systems development, *International Journal of Human-Computer Studies* **58**(1), 89–123.
- Gerber, B. S. and Eiser, A. R.: 2001, The patient physician relationship in the internet age: future prospects and the research agenda, *Journal of medical Internet research* **3**(2), E15.
- GGZ Nederland: 2009, Naar herstel en gelijkwaardig burgerschap. visie op de (langdurende) zorg aan mensen met ernstige psychische aandoeningen, <http://www.ggznederland.nl/scrivo/asset.php?id=305955>. (Accessed: 9 December 2011).

- Gleeson, J. F., Alvarez-Jimenez, M. and Lederman, R.: 2012, Moderated online social therapy for recovery from early psychosis, *Psychiatric services (Washington, D.C.)* **63**(7), 719.
- Godolphin, W.: 2009, Shared decision-making, *Healthcare Quarterly* **12**, e186–e190.
- Gorman, J. M. and Braber, M. D.: 2008, Semantic Web Sparks Evolution of Health 2.0 A Road Map to Consumer-Centric Healthcare, *SWW3923*, number Spring.
- Granger, C. W.: 1969, Investigating causal relations by econometric models and cross-spectral methods, *Econometrica: Journal of the Econometric Society* **37**(3), 424–438.
- Granger, C. W. J. and Andersen, A. P.: 1978, *An introduction to bilinear time series models*, Vandenhoeck und Ruprecht Göttingen.
- Grohol, J. M.: 2003, *The road online to empowered clients and empowered providers*, Telepsychiatry and e-Mental Health, first edn, Royal Society of Medicine Press Ltd, London, pp. 337–348.
- Grynszpan, O., Perbal, S., Pelissolo, A., Fossati, P., Jouvent, R., Dubal, S. and Perez-Diaz, F.: 2011, Efficacy and specificity of computer-assisted cognitive remediation in schizophrenia: a meta-analytical study, *Psychological medicine* **41**(1), 163–173.
- Guthrie, D., McIntosh, M., Callaly, T., Trauer, T. and Coombs, T.: 2008, Consumer attitudes towards the use of routine outcome measures in a public mental health service: a consumer-driven study, *International journal of mental health nursing* **17**(2), 92–97.
- Haker, H., Lauber, C. and Rossler, W.: 2005, Internet forums: a self-help approach for individuals with schizophrenia?, *Acta Psychiatrica Scandinavica* **112**(6), 474–477.
- Hamilton, J. D.: 1994, *Time series analysis*, Vol. 2, Cambridge Univ Press.
- Hannan, E. J. and Quinn, B. G.: 1979, The determination of the order of an autoregression, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 190–195.
- Happell, B.: 2008, Meaningful information or a bureaucratic exercise? Exploring the value of routine outcome measurement in mental health, *Issues in Mental Health Nursing* **29**(10), 1098–1114.

- Hare, K., Whitworth, B. and Deek, F.: 2006, A New Approach to Clinical IT Resistance: The Need for Information Technology Confidentially and Mobility, *HIC 2006 and HINZ 2006*, Health Informatics Society of Australia, pp. 440–443.
- Häyrinen, K., Saranto, K. and Nykänen, P.: 2008, Definition, structure, content, use and impacts of electronic health records: a review of the research literature, *International journal of medical informatics* **77**(5), 291–304.
- Hendry, D. F. and Krolzig, H.-M.: 2001, *Automatic econometric model selection using PcGets 1.0*, Timberlake Consultants.
- Hennink, M., Hutter, I. and Bailey, A.: 2011, *Qualitative Research Methods*, Sage Publications Ltd, London.
- Hilbe, J. M.: 2006, A review of LIMDEP 9.0 and NLOGIT 4.0, *The American Statistician* **60**(2), 187–202.
- Hoenders, H. R., Bos, E. H., de Jong, J. T. and de Jonge, P.: 2012, Temporal dynamics of symptom and treatment variables in a lifestyle-oriented approach to anxiety disorder: a single-subject time-series analysis, *Psychotherapy and psychosomatics* **81**(4), 253–255.
- Horn, W.: 2000, Artificial intelligence in medicine and medical decision making Europe, *Artificial Intelligence in Medicine* **20**, 1–3.
- Horn, W.: 2001, AI in medicine on its way from knowledge-intensive to data-intensive systems, *Artificial Intelligence in Medicine* **23**(1), 5–12.
- Huelsenbeck, J. P. and Crandall, K. A.: 1997, Phylogeny estimation and hypothesis testing using maximum likelihood, *Annual Review of Ecology and Systematics* pp. 437–466.
- IBM SPSS software: 2013, <http://www.ibm.com/software/analytics/spss/>. (Accessed: 11 December 2013).
- Igras, E.: 2007, eHealth Business Opportunities, *Technical Report* 403.
- Internet World Stats. Top 58 countries with highest penetration rates: 2011, <http://www.internetworldstats.com/stats9.htm>. (Accessed: 20 October 2011).
- Jarque, C. M. and Bera, A. K.: 1980, Efficient tests for normality, homoscedasticity and serial independence of regression residuals, *Economics Letters* **6**(3), 255–259.

- Jensen, P. B., Jensen, L. J. and Brunak, S.: 2012, Mining electronic health records: towards better research applications and clinical care, *Nature Reviews Genetics* **13**(6), 395–405.
- Jones, R. B., Atkinson, J. M., Coia, D. A., Paterson, L., Morton, A. R., McKenna, K., Craig, N., Morrison, J. and Gilmour, W. H.: 2001, Randomised trial of personalised computer based information for patients with schizophrenia, *BMJ (Clinical research ed.)* **322**(7290), 835–840.
- Kaplan, K., Salzer, M. S., Solomon, P., Brusilovskiy, E. and Cousounis, P.: 2011, Internet peer support for individuals with psychiatric disabilities: A randomized controlled trial, *Social science & medicine (1982)* **72**(1), 54–62.
- Kenwright, M., Liness, S. and Marks, I.: 2001, Reducing demands on clinicians by offering computer-aided self-help for phobia/panic. feasibility study, *The British journal of psychiatry : the journal of mental science* **179**, 456–459.
- Kersting, A., Schlicht, S. and Kroker, K.: 2009, Internet therapy: Opportunities and boundaries, *Der Nervenarzt* **80**(7), 797–804.
- Kilbourne, A. M.: 2012, E-health and the transformation of mental health care, *Psychiatric services (Washington, D.C.)* **63**(11), 1059.
- Killackey, E., Anda, A. L., Gibbs, M., Alvarez-Jimenez, M., Thompson, A., Sun, P. and Baksheev, G. N.: 2011, Using internet enabled mobile devices and social networking technologies to promote exercise as an intervention for young first episode psychosis patients, *BMC psychiatry* **11**, 80.
- knitr: A general-purpose package for dynamic report generation in R: 2013, <http://cran.r-project.org/web/packages/knitr/>. (Accessed: 14 October 2013).
- Kohane, I. S., Greenspun, P., Fackler, J., Cimino, C. and Szolovits, P.: 1996, Building national electronic medical record systems via the world wide web, *Journal of the American Medical Informatics Association* **3**(3), 191–207.
- Koivunen, M., Välimäki, M., Patel, A., Knapp, M., Hätönen, H., Kuosmanen, L., Pitkänen, A., Anttila, M. and Katajisto, J.: 2010, Effects of the implementation of the web-based patient support system on staffs attitudes towards computers and IT use: a randomised controlled trial, *Scandinavian Journal of Caring Sciences* **24**(3), 592–599.
- Koivunen, M., Välimäki, M., Pitkänen, A. and Kuosmanen, L.: 2007, A preliminary usability evaluation of web-based portal application for patients with schizophrenia, *Journal of Psychiatric and Mental Health Nursing* **14**(5), 462–469.

- Kolodner, J. and Kolodner, R.: 1987, Using experience in clinical problem solving: Introduction and framework, *IEEE Transactions on Systems, Man, and Cybernetics* **17**(3), 420–431.
- Kononenko, I.: 2001, Machine learning for medical diagnosis: history, state of the art and perspective, *Artificial Intelligence in Medicine* **23**(1), 89–109.
- Krummenacher, R., Simperl, E., Cerizza, D., Della Valle, E., Nixon, L. J. B. and Foxvog, D.: 2009, Enabling the European Patient Summary through triplespaces, *Computer methods and programs in biomedicine* **95**(2 Suppl), S33–S43.
- Ku, J., Han, K., Lee, H. R., Jang, H. J., Kim, K. U., Park, S. H., Kim, J. J., Kim, C. H., Kim, I. Y. and Kim, S. I.: 2007, Vr-based conversation training program for patients with schizophrenia: a preliminary clinical trial, *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society* **10**(4), 567–574.
- Kuosmanen, L., Jakobsson, T., Hyttinen, J., Koivunen, M. and Välimäki, M.: 2010, Usability evaluation of a web-based patient information system for individuals with severe mental health problems, *Journal of Advanced Nursing* **66**(12), 2701–2710.
- Kuosmanen, L., Välimäki, M., Joffe, G., Pitkänen, A., Hätönen, H., Patel, A. and Knapp, M.: 2009, The effectiveness of technology-based patient education on self-reported deprivation of liberty among people with severe mental illness: A randomized controlled trial, *Nordic journal of psychiatry* pp. 1–7.
- Kuriyama, D., Izumi, S., Itabashi, G., Kimura, S., Ebihara, Y., Takahashi, K. and Kato, Y.: 2007, Design and Implementation of a Health Management Support System Using Ontology, in N. Chotikakamtorn (ed.), *Proceedings of the International Conference on Engineering, Applied Sciences, and Technology*, IEEE, Thailand, pp. 746–749.
- Lakeman, R.: 2004, Standardized routine outcome measurement: pot holes in the road to recovery, *International journal of mental health nursing* **13**(4), 210–215.
- Lambert, M. and Finch, A.: 1999, The outcome questionnaire, in M. E. Maruish (ed.), *The use of psychological testing for treatment planning and outcomes assessment (2nd ed.)*, Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 831–869.
- Lambert, M. J., Hansen, N. B. and Finch, A. E.: 2001, Patient-focused research: using patient outcome data to enhance treatment effects, *Journal of consulting and clinical psychology* **69**(2), 159–172.

- Lambert, M. J., Harmon, C., Slade, K., Whipple, J. L. and Hawkins, E. J.: 2005, Providing feedback to psychotherapists on their patients' progress: clinical results and practice suggestions, *Journal of clinical psychology* **61**(2), 165–174.
- Lavrac, N.: 1999, Selected techniques for data mining in medicine, *Artificial intelligence in medicine* **16**(1), 3–23.
- Lavrac, N., Keravnou, E. and Zupan, B.: 2000, Intelligent data analysis in medicine, *Encyclopedia of computer science and technology* **42**(9), 113–157.
- Leong, T.-Y., Kaiser, K. and Miksch, S.: 2007, Free and Open Source Enabling Technologies for Patient-Centric, Guideline-Based Clinical Decision Support: A Survey, *IMIA Yearbook of Med. Inf.* (April), 74–86.
- Liquid Templating Language*: 2011, <http://liquidmarkup.org>. (Accessed: 18 November 2012).
- Litterman, R. B.: 1986, Forecasting with bayesian vector autoregressions—five years of experience, *Journal of Business & Economic Statistics* **4**(1), 25–38.
- Ljung, G. M. and Box, G. E.: 1978, On a measure of lack of fit in time series models, *Biometrika* **65**(2), 297–303.
- Lucas, P., van Der Gaag, L. and Abu-Hanna, A.: 2004, Bayesian networks in biomedicine and health-care, *Artificial Intelligence in Medicine* **30**(3), 201–214.
- Lütkepohl, H.: 2005, *New introduction to multiple time series analysis*, Cambridge Univ Press.
- Madoff, S. A., Pristach, C. A., Smith, C. M. and Pristach, E. A.: 1996, Computerized medication instruction for psychiatric inpatients admitted for acute care, *M.D.computing : computers in medical practice* **13**(5), 427–31, 441.
- Mahone, I. H., Farrell, S., Hinton, I., Johnson, R., Moody, D., Rifkin, K., Moore, K., Becker, M. and Barker, M. R.: 2011, Shared decision making in mental health treatment: qualitative findings from stakeholder focus groups, *Archives of psychiatric nursing* **25**(6), e27–e36.
- Maier, E., Reimer, U., Schär, S. and Zimmermann, P.: 2010, SEMPER: a web-based support system for patient self-management, in T. Owens (ed.), *Proceedings of the 23rd Bled eConference*, number 17, AIS Electronic Library, Atlanta, pp. 196–209.
- Makkink, S. and Kits, L.: 2011, Herstellen doe je zelf, in V. Hees, P. V. der Vlist and N. Mulder (eds), *Van weten naar meten. ROM in de GGZ*, Boom, Amsterdam, pp. 97–108.

- markdown: Markdown rendering for R*: 2013, <http://cran.r-project.org/web/packages/markdown/>. (Accessed: 14 October 2013).
- Marks, I. M., Cavanagh, K. and Gega, L.: 2007, *Hands-on help: Computer-aided psychotherapy*, Psychology Press, New York.
- Martyn, D.: 2002, The experiences and views of self-management of people with a schizophrenia diagnosis, *Technical report*, Rethink.
- McCrone, P., Knapp, M., Proudfoot, J., Ryden, C., Cavanagh, K., Shapiro, D. A., Ilson, S., Gray, J. A., Goldberg, D., Mann, A., Marks, I., Everitt, B. and Tylee, A.: 2004, Cost-effectiveness of computerised cognitive-behavioural therapy for anxiety and depression in primary care: randomised controlled trial, *The British journal of psychiatry : the journal of mental science* **185**, 55–62.
- McGorry, P. D., Yung, A. R., Pantelis, C. and Hickie, I. B.: 2009, A clinical trials agenda for testing interventions in earlier stages of psychotic disorders, *The Medical journal of Australia* **190**(4 Suppl), S33–6.
- McGuinness, D. and Van Harmelen, F.: 2004, OWL web ontology language overview (W3C recommendation), <http://www.w3.org/TR/owl-features/>. (Accessed: 18 November 2012).
- McGurk, S. R., Twamley, E. W., Sitzler, D. I., McHugo, G. J. and Mueser, K. T.: 2007, A meta-analysis of cognitive remediation in schizophrenia, *The American Journal of Psychiatry* **164**(12), 1791–1802.
- METU-SRDC: 2007, RIDE: A Roadmap for Interoperability of eHealth Systems in Support of COM 356 with Special Emphasis on Semantic Interoperability, *Technical report*.
- Miller, R. H. and Sim, I.: 2004, Physicians use of electronic medical records: barriers and solutions, *Health affairs* **23**(2), 116–126.
- Minsky, M. and Seymour, P.: 1969, *Perceptrons*, MIT press.
- Mobasher, B., Cooley, R. and Srivastava, J.: 2000, Automatic personalization based on web usage mining, *Communications of the ACM* **43**(8), 142–151.
- Molenaar, P. C. and Campbell, C. G.: 2009, The new person-specific paradigm in psychology, *Current Directions in Psychological Science* **18**(2), 112–117.
- Montani, S.: 2011, How to use contextual knowledge in medical case-based reasoning systems: A survey on very recent trends, *Artificial intelligence in medicine* **51**(2), 125–131.

- Mueser, K. T., Corrigan, P. W., Hilton, D. W., Tanzman, B., Schaub, A., Gingerich, S., Essock, S. M., Tarrier, N., Morey, B., Vogel-Scibilia, S. and Herz, M. I.: 2002, Illness management and recovery: a review of the research, *Psychiatric services (Washington, D.C.)* **53**(10), 1272–1284.
- Mueser, K. T., Meyer, P. S., Penn, D. L., Clancy, R., Clancy, D. M. and Salyers, M. P.: 2006, The illness management and recovery program: rationale, development, and preliminary findings, *Schizophrenia bulletin* **32 Suppl 1**, S32–43.
- Musen, M.: 1999, Stanford Medical Informatics: uncommon research, common goals, *MD COMPUTING* **16**, 47–55.
- Myin-Germeys, I., Birchwood, M. and Kwapil, T.: 2011, From environment to therapy in psychosis: a real-world momentary assessment approach, *Schizophrenia bulletin* **37**(2), 244–247.
- MySQL: 2013, <http://www.mysql.com>. (Accessed: 2 July 2013).
- Nelson, C. R. and Plosser, C. R.: 1982, Trends and random walks in macroeconomic time series: some evidence and implications, *Journal of monetary economics* **10**(2), 139–162.
- Nielsen, J.: 1993, What is usability?, *Usability engineering*, Morgan Kaufmann, pp. 23–48.
- Nielsen, J.: 1994, Heuristic evaluation, in R. L. Mack and J. Nielsen (eds), *Usability inspection methods*, Wiley & Sons, pp. 25–62.
- Nielsen, J. and Landauer, T. K.: 1993, A mathematical model of the finding of usability problems, *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, ACM, pp. 206–213.
- Nike+ Running App: 2013, http://nikeplus.nike.com/plus/products/gps_app/. (Accessed: 4 June 2013).
- Oorschot, M., Lataster, T., Thewissen, V., Wichers, M. and Myin-Germeys, I.: 2012, Mobile assessment in schizophrenia: a data-driven momentary approach, *Schizophrenia bulletin* **38**(3), 405–413.
- OpenCPU: Scientific computing in the cloud: 2013, <https://public.opencpu.org/>. (Accessed: 14 October 2013).
- Opler, L. A., Ramirez, P. M. and Mougios, V. M.: 2002, Outcome measurement in serious mental illness, *Outcome measurement in psychiatry: a critical review*. Washington, DC: American Psychiatric Pub pp. 139–154.

- Owen, P. D.: 2003, General-to-specific modelling using PcGets, *Journal of Economic Surveys* **17**(4), 609–628.
- Paganelli, F.: 2007, An ontology-based context model for home health monitoring and alerting in chronic patient care networks, in L. O’Conner (ed.), *Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops*, Vol. 2, IEEE Press, pp. 838–845.
- Pantazi, S. V., Arocha, J. F. and Moehr, J. R.: 2004, Case-based medical informatics, *BMC Medical Informatics and Decision Making* **4**(1).
- Patel, C., Gomadam, K., Khan, S. and Garg, V.: 2010, TrialX: Using semantic technologies to match patients to relevant clinical trials based on their Personal Health Records, *Web Semantics: Science, Services and Agents on the World Wide Web* **8**(4), 342–347.
- Pena-Reyes, C. and Sipper, M.: 2000, Evolutionary computation in medicine: an overview, *Artificial Intelligence in Medicine* **19**(1), 1–23.
- Perez-Amaral, T., Gallo, G. M. and White, H.: 2003, A flexible tool for model building: the relevant transformation of the inputs network approach (RETINA), *Oxford Bulletin of Economics and Statistics* **65**(s1), 821–838.
- Perner, P.: 2006, GUEST EDITORIAL: Intelligent data analysis in medicine-Recent advances, *Artificial Intelligence in Medicine* **37**(1), 1–5.
- Pesaran, B. and Pesaran, M. H.: 2010, *Time Series Econometrics Using Microfit 5.0: A User’s Manual*, Oxford University Press, Inc.
- Pesaran, M. H. and Timmermann, A.: 2000, A recursive modelling approach to predicting UK stock returns, *The Economic Journal* **110**(460), 159–191.
- Pfaff, B.: 2008, VAR, SVAR and SVEC models: Implementation within R package vars, *Journal of Statistical Software* **27**(4), 1–32.
- PHAMOUS. *Pharmacotherapy monitoring and outcome survey*: 2011, <http://www.phamous.eu/home.html>. (Accessed: 8 December 2011).
- Phillips, P. C.: 1996, Econometric model determination, *Econometrica: Journal of the Econometric Society* **64**(4), 763–812.
- Phillips, P. C. and Perron, P.: 1988, Testing for a unit root in time series regression, *Biometrika* **75**(2), 335–346.

- Pijnenborg, G. H., Withaar, F. K., Brouwer, W. H., Timmerman, M. E., van den Bosch, R. J. and Evans, J. J.: 2010, The efficacy of SMS text messages to compensate for the effects of cognitive impairments in schizophrenia, *The British journal of clinical psychology / the British Psychological Society* **49**(Pt 2), 259–274.
- Plastiras, P., O’Sullivan, D. and Weller, P.: 2014, An ontology-driven information model for interoperability of personal and electronic health records, *eTELEMED 2014, The Sixth International Conference on eHealth, Telemedicine, and Social Medicine*, pp. 130–133.
- Priebe, S., Huxley, P., Knight, S. and Evans, S.: 1999, Application and results of the manchester short assessment of quality of life (MANSA), *International Journal of Social Psychiatry* **45**(1), 7–12.
- Priebe, S., McCabe, R., Bullenkamp, J., Hansson, L., Lauber, C., Martinez-Leal, R., Rossler, W., Salize, H., Svensson, B., Torres-Gonzales, F., van den Brink, R., Wiersma, D. and Wright, D. J.: 2007, Structured patient-clinician communication and 1-year outcome in community mental healthcare: cluster randomised controlled trial, *The British journal of psychiatry : the journal of mental science* **191**, 420–426.
- Primiceri, G. E.: 2005, Time varying structural vector autoregressions and monetary policy, *The Review of Economic Studies* **72**(3), 821–852.
- Proudfoot, J.: 2004, Computer-based treatment for anxiety and depression: is it feasible? is it effective?, *Neuroscience & Biobehavioral Reviews* **28**(3), 353–363.
- Proudfoot, J., Parker, G., Hyett, M., Manicavasagar, V., Smith, M., Grdovic, S. and Greenfield, L.: 2007, Next generation of self-management education: Web-based bipolar disorder program, *The Australian and New Zealand Journal of Psychiatry* **41**(11), 903–909.
- Ramos, C.: 2007, Ambient intelligence—a state of the art from artificial intelligence perspective, *Lecture Notes in Computer Science* **4874**, 285–295.
- Ramos, C., Augusto, J. C. and Shapiro, D.: 2008, Ambient intelligence—the next step for artificial intelligence, *Intelligent Systems, IEEE* **23**(2), 15–18.
- Riper, H.: 2007, *E-mental health: High tech, high touch, high trust*, Trimbos Instituut, Utrecht.
- RoQua: 2011, <http://www.roqua.nl>. (Accessed: 18 November 2012).
- Rosenblad, A.: 2008, gretl 1.7.3, *Journal of Statistical Software* **25**(s01).

- Rosmalen, J. G., Wenting, A. M., Roest, A. M., de Jonge, P. and Bos, E. H.: 2012, Revealing causal heterogeneity using time series analysis of ambulatory assessments: application to the association between depression and physical activity after myocardial infarction, *Psychosomatic Medicine* **74**(4), 377–386.
- Rotondi, A. J.: 2010, *Schizophrenia*, Using technology to support evidence-based behavioral health practices: A clinician's guide, Routledge/Taylor & Francis Group, New York, NY US, pp. 69–89.
- Rotondi, A. J., Anderson, C. M., Haas, G. L., Eack, S. M., Spring, M. B., Ganguli, R., Newhill, C. and Rosenstock, J.: 2010, Web-based psychoeducational intervention for persons with schizophrenia and their supporters: one-year outcomes, *Psychiatric services (Washington, D.C.)* **61**(11), 1099–1105.
- Rotondi, A., Sinkule, J., Haas, G., Spring, M., Litschge, C., Newhill, C., Ganguli, R. and Anderson, C.: 2007, Designing websites for persons with cognitive deficits: Design and usability of a psychoeducational intervention for persons with severe mental illness, *Psychological Services* **4**(3), 202–224.
- Royston, P.: 1992, Comment on sg3. 4 and an improved D'Agostino test, *Stata Technical Bulletin* **1**(3).
- Ruby on Rails*: 2013, <http://rubyonrails.org>. (Accessed: 2 July 2013).
- Rumelhart, D. E., Hintont, G. E. and Williams, R. J.: 1986, Learning representations by back-propagating errors, *Nature* **323**(6088), 533–536.
- Sablier, J., Stip, E., Jacquet, P., Giroux, S., Pigot, H. and Franck, N.: 2012, Ecological assessments of activities of daily living and personal experiences with mobus, an assistive technology for cognition: A pilot study in schizophrenia, *Assistive Technology* **24**(2), 67–77.
- Safran, C., Bloomrosen, M., Hammond, W. E., Labkoff, S., Markel-Fox, S., Tang, P. C. and Detmer, D. E.: 2007, Toward a national framework for the secondary use of health data: an american medical informatics association white paper, *Journal of the American Medical Informatics Association* **14**(1), 1–9.
- Samen Keuzes Maken*: 2011, <http://www.samenkeuzesmaken.nl>. (Accessed: 18 November 2012).
- Samoocha, D., Bruinvels, D. J., Elbers, N. A., Anema, J. R. and van der Beek, A. J.: 2010, Effectiveness of web-based interventions on patient empowerment: a systematic review and meta-analysis, *Journal of medical Internet research* **12**(2), e23.

- Sánchez, C., Rueda, A. and Romero, E.: 2007, A granular prototype for telemedicine based on hl7 information model, *Segundo Congreso Colombiano de Computación. Universidad Javeriana*.
- Sanyal, I.: 2006, Empowering the impaired through the appropriate use of information technology and internet, *Studies in health technology and informatics* **121**, 15–21.
- Sargent, T. J.: 1979, Estimating vector autoregressions using methods not based on explicit economic theories, *Federal Reserve Bank of Minneapolis Quarterly Review* **3**(3), 8–15.
- Sauermann, S., Frohner, M., Urbauer, P., Forjan, M., Pohn, B., Drauschke, A., Mense, A. et al.: 2013, The adolescence of electronic health records: Status and perspectives for large scale implementation, *Acta Informatica Pragensia* **2**(1), 30–38.
- Schaefer, B., Nijssen, Y. and Weeghel, J. V.: 2011, Van meten naar oplossingsgericht werken, in S. V. hees, P. V. der Vlist and N. Mulder (eds), *Van weten naar meten. ROM in de GGZ*, Boom, Amsterdam, pp. 89–96.
- Scher, D. L.: 2012, How patient-centric care differs from patient-centered care, <http://davidleescher.com/2012/03/03/how-patient-centric-care-differs-from-patient-centered-care-2/>. (Accessed: 5 May 2014).
- Schermer, M.: 2009, Telecare and self-management: opportunity to change the paradigm?, *Journal of medical ethics* **35**(11), 688–691.
- Schrank, B., Sibitz, I., Unger, A. and Amering, M.: 2010, How patients with schizophrenia use the internet: qualitative study, *Journal of medical Internet research* **12**(5), e70.
- Schwarz, G.: 1978, Estimating the dimension of a model, *The annals of statistics* **6**(2), 461–464.
- SHAZAM features: 2013, <http://econometrics.com/features/>. (Accessed: 11 December 2013).
- Sherman, P. S.: 1998, Computer-assisted creation of psychiatric advance directives, *Community mental health journal* **34**(4), 351–362.
- Shortliffe, E.: 1976, *Computer-based medical consultations, MYCIN*, Elsevier Publishing Company.

- Shrimpton, B. and Hurworth, R.: 2005, Adventures in evaluation: Reviewing a CD-ROM based adventure game designed for young people recovering from psychosis, *Journal of Educational Multimedia and Hypermedia* **14**(3), 273–290.
- Sims, H., Sanghara, H., Hayes, D., Wandiembe, S., Finch, M., Jakobsen, H., Tsakanikos, E., Okocha, C. I. and Kravariti, E.: 2012, Text message reminders of appointments: a pilot intervention at four community mental health clinics in london, *Psychiatric services (Washington, D.C.)* **63**(2), 161–168.
- Smolensky, P.: 1987, Connectionist AI, symbolic AI, and the brain, *Artificial Intelligence Review* **1**(2), 95–109.
- Soto, G. and Spertus, J.: 2007, EPOCH and ePRISM: a Web-based translational framework for bridging outcomes research and clinical practice, *Computers in Cardiology* pp. 205–208.
- Spaniel, F., Hrdlicka, J., Novak, T., Kozeny, J., Hoschl, C., Mohr, P. and Motlova, L. B.: 2012, Effectiveness of the information technology-aided program of relapse prevention in schizophrenia (ITAREPS): A randomized, controlled, double-blind study, *Journal of psychiatric practice* **18**(4), 269–280.
- Stacey, M. and McGregor, C.: 2007, Temporal abstraction in intelligent clinical data analysis: A survey, *Artificial Intelligence in Medicine* **39**(1), 1–24.
- STATA: *Data Analysis and Statistical Software*: 2013, <http://www.stata.com>. (Accessed: 11 December 2013).
- Stearns, M., Price, C., Spackman, K. and Wang, A.: 2001, SNOMED clinical terms: overview of the development process and project status, in S. Bakken (ed.), *Proceedings of the American Medical Informatics Association Symposium*, Hanley & Belfus Inc., Philadelphia, PA, pp. 662–666.
- Steinwachs, D. M., Roter, D. L., Skinner, E. A., Lehman, A. F., Fahey, M., Cullen, B., Everett, A. S. and Gallucci, G.: 2011, A web-based program to empower patients who have schizophrenia to discuss quality of care with mental health providers, *Psychiatric services (Washington, D.C.)* **62**(11), 1296–1302.
- Stephanidis, C.: 2001, *User interfaces for all: concepts, methods, and tools*, Lawrence Erlbaum Associates.
- Stigler, S.: 2008, Fisher and the 5% level, *Chance* **21**(4), 12–12.
- Tennen, H. and Affleck, G.: 1996, Daily processes in coping with chronic pain: Methods and analytic strategies, in M. Zeidner and N. S. Endler (eds), *Handbook of coping: Theory, research, applications*, John Wiley & Sons, Oxford, England, pp. 151–177.

- The Apache Software Foundation*: 2013, <http://www.apache.org/>. (Accessed: 14 October 2013).
- The Couch-to-5K Running Plan: C25K Mobile App*: 2012, http://www.coolrunning.com/engine/2/2_3/181.shtml. (Accessed: 4 June 2013).
- The R Project for Statistical Computing*: 2013, <http://www.r-project.org>. (Accessed: 14 October 2013).
- The VAR function in TSP*: 2013, <http://www.nber.org/tsp/tsphelp/var.htm>. (Accessed: 11 December 2013).
- Thorp, J.: 2010, Europe's e-health initiatives, *J AHIMA* **81**, 56–8.
- Trauer, T.: 2010, Introduction, in T. Trauer (ed.), *Outcome measurement in mental health: theory and practice*, first edn, Cambridge University Press, Cambridge, pp. 1–14.
- Trauer, T., Tobias, G. and Slade, M.: 2008, Development and evaluation of a patient-rated version of the camberwell assessment of need short appraisal schedule (CANSAS-P), *Community Mental Health Journal* **44**, 113–124.
- Twamley, E. W., Jeste, D. V. and Bellack, A. S.: 2003, A review of cognitive training in schizophrenia, *Schizophrenia bulletin* **29**(2), 359–382.
- Twitter Bootstrap*: 2013, <http://getbootstrap.com>. (Accessed: 14 October 2013).
- urca: Unit root and cointegration tests for time series data*: 2013, <http://cran.r-project.org/web/packages/urca/index.html>. (Accessed: 10 September 2013).
- Välimäki, M., Anttila, M., Hätönen, H., Koivunen, M., Jakobsson, T., Pitkänen, A., Herrala, J. and Kuosmanen, L.: 2008, Design and development process of patient-centered computer-based support system for patients with schizophrenia spectrum psychosis, *Informatics for Health and Social Care* **33**(2), 113–123.
- Välimäki, M., Hätönen, H., Lahti, M., Kuosmanen, L. and Adams, C. E.: 2012, Information and communication technology in patient education and support for people with schizophrenia, *Cochrane database of systematic reviews (Online)* **10**, CD007198.
- Van der Krieke, L., Emerencia, A. C., Aiello, M. and Sytema, S.: 2012, Usability evaluation of a web-based support system for people with a schizophrenia diagnosis, *Journal of medical Internet research* **14**(1), e24.

- Van der Krieke, L., Emerencia, A. and Sytema, S.: 2011, An online portal on outcomes for dutch service users, *Psychiatric Services* **62**(7), 803.
- Van der Krieke, L., Wunderink, L., Emerencia, A. C., de Jonge, P. and Sytema, S.: 2014, E-mental health self-management for psychotic disorders: State of the art and future perspectives, *Psychiatric Services* **65**(1), 33–49.
- Van Gils, A., Burton, C., Bos, E., Janssens, K., Schoevers, R. and Rosmalen, J.: 2014, Individual variation in temporal relationships between stress and functional somatic symptoms, *Journal of Psychosomatic Research* **77**(1), 34–39.
- Van Os, J. and Kapur, S.: 2009, Schizophrenia, *The Lancet* **374**(9690), 635–645.
- Van Os, J. and Sham, P.: 2003, Gene-environment interactions, in R. Murray, P. Jones, E. Susser, J. van Os and M. Cannon (eds), *The Epidemiology of Schizophrenia*, Cambridge University Press, Cambridge, UK, pp. 235–254.
- Van Rijsbergen, C.: 1979, *Information retrieval*, 2nd edn, Butterworth & Co Ltd., London, chapter 7, pp. 111–143.
- Vayreda, A. and Antaki, C.: 2009, Social support and unsolicited advice in a bipolar disorder online forum, *Qualitative health research* **19**(7), 931–942.
- Vector Autoregressive Models in Matlab*: 2013, <http://www.mathworks.com/help/econ/var-models.html>. (Accessed: 11 December 2013).
- Venables, W. N., Ripley, B. D. and Venables, W.: 1994, *Modern applied statistics with S-PLUS*, Vol. 250, Springer-verlag New York.
- Vogelvang, B.: 2005, *Econometrics: Theory and Applications with EViews*, Pearson Education.
- Vogt, J. and Wittwer, D.: 2007, *Open Standards for Data Exchange in Healthcare Systems*, Seminar thesis in e-health, University of Fribourg.
- Walker, H.: 2006, Computer-based education for patients with psychosis, *Nursing standard (Royal College of Nursing (Great Britain) : 1987)* **20**(30), 49–56.
- Walker, J., Pan, E., Johnston, D., Adler-Milstein, J., Bates, D. W. and Middleton, B.: 2005, The value of health care information exchange and interoperability, *Health Affairs* **24**, W5.
- Wegweis Ontology*: 2011, Available from: <http://www.wegweis.nl/ontologies/problems.owl>. (Accessed: 18 November 2012).
- Werbos, P. J.: 1994, *The roots of backpropagation: from ordered derivatives to neural networks and political forecasting*, Wiley-Interscience.

- West, S., King, V., Carey, T. S., Lohr, K. N., McKoy, N., Sutton, S. F. and Lux, L.: 2002, Systems to rate the strength of scientific evidence, *Technical Report 47*, Agency for Healthcare Research and Quality. U.S. Department of Health and Human Services.
- Whipple, J. L. and Lambert, M. J.: 2011, Outcome measures for practice, *Annual review of clinical psychology* **7**, 87–111.
- White, H.: 1980, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica: Journal of the Econometric Society* pp. 817–838.
- White, K. J. and McRae, R. N.: 1987, SHAZAM: A general computer program for econometric methods (version 5), *The American Statistician* **41**(1), 80.
- Wickham, H.: 2009, *ggplot2: elegant graphics for data analysis*, Springer Publishing Company, Incorporated.
- Wild, B., Eichler, M., Friederich, H.-C., Hartmann, M., Zipfel, S. and Herzog, W.: 2010, A graphical vector autoregressive modelling approach to the analysis of electronic diary data, *BMC medical research methodology* **10**(1), 28.
- Wing, J., Beevor, A., Curtis, R., Park, S., Hadden, S. and Burns, A.: 1998, Health of the nation outcome scales (HoNOS): Research and development, *The British Journal of Psychiatry* **172**(1), 11–18.
- Woltmann, E., Wilkniss, S., Teachout, A., McHugo, G. and Drake, R.: 2011, Trial of an electronic decision support system to facilitate shared decision making in community mental health, *Psychiatric Services* **62**(1), 54–60.
- Wykes, T., Huddy, V., Cellard, C., McGurk, S. R. and Czobor, P.: 2011, A meta-analysis of cognitive remediation for schizophrenia: methodology and effect sizes, *American Journal of Psychiatry* **168**(5), 472–485.
- Yuan, C. Z., Isa, D. and Blanchfield, P.: 2008, A hybrid data mining and case-based reasoning user modeling system (HDCU) for monitoring and predicting of blood sugar level, *2008 International Conference on Computer Science and Software Engineering*, Vol. 1, IEEE, pp. 653–656.
- Zhou, L. and Hripcsak, G.: 2007, Temporal reasoning with medical data—a review with emphasis on medical natural language processing, *Journal of biomedical informatics* **40**(2), 183–202.
- Zupan, B., Holmes, J. and Bellazzi, R.: 2006, Knowledge-based data analysis and interpretation, *Artificial Intelligence in Medicine* **37**(3), 163–165.

Samenvatting

De hoeveelheid medische gegevens die digitaal is opgeslagen groeit wereldwijd in een snel tempo. Deze groei biedt de mogelijkheid aan nieuwe en interessante applicaties om ons besef van zorg te veranderen. We stellen ons de vraag welke onderdelen van zorg waarin kennis wordt gedeeld te automatiseren zijn en hoe deze automatisering te bewerkstelligen is. Ons werk richt zich op het automatiseren van twee aspecten van zorg die traditioneel gezien menselijke supervisie vereisen. Namelijk het genereren van persoonlijk advies voor schizofreniepatiënten en het vinden van het beste vectorautoregressiemodel voor digitale patiëntendagboeken.

In Hoofdstuk 2 schetsen we een beeld van de geschiedenis van de toepassing van kunstmatige intelligentie in de medische wereld. We geven een chronologisch overzicht van de applicaties en ontwikkelingen die relevant zijn voor ons onderzoek. Kunstmatige intelligentie binnen de medische wereld heeft traditioneel gezien altijd gediend ter ondersteuning van de behandelaar (hieronder vallen bijvoorbeeld applicaties ten behoeve van diagnose en beslissingsondersteuning). In het laatste decennium zien we steeds meer applicaties die ofwel de belangen van de patiënt centraal stellen, dan wel gebruikt worden door de patiënt zelf.

Hoofdstuk 3 gaat dieper in op de specifieke toepassing van computerapplicaties ten behoeve van self-management bij psychotische aandoeningen. Aan de hand van een systematisch literatuuronderzoek proberen we de initiatieven die op dit gebied al zijn ondernomen in kaart te brengen en proberen we te achterhalen of de effectiviteit van dit soort systemen is bewezen. Alhoewel de data niet altijd toereikend is om statistisch beproefde conclusies te trekken, ontdekken we in de afgelopen jaren een toename in het aantal e-health interventies. Ook kunnen we met enige zekerheid stellen dat e-health self-management interventies tenminste evenveel effect hebben als zorg die niet van dergelijke technologische hulpmiddelen is voorzien. Dit lijkt een zwakke conclusie, maar kan in bepaalde gevallen tot een mogelijke kostenbe-

sparing leiden (alhoewel ook hier de data te kort schoot om betrouwbare conclusies te trekken).

In Hoofdstuk 4 beschrijven we het ontwerp, de gebruikersinterface en de algoritmes van Wegweis. Wegweis is een web-applicatie waar mensen met schizofrenie geautomatiseerd persoonlijk advies kunnen krijgen. We hebben hiervoor een eigen ontologie opgesteld waarin het gehele problemspectrum van een schizofreniepatiënt kan worden ondergebracht. Deze concepten hebben we hiërarchisch gestructureerd, waardoor we problemen van generiek naar specifiek kunnen indelen. Vervolgens hebben we de items uit het elektronisch patiëntendossier en de items uit onze adviesdatabase gekoppeld aan concepten in de ontologie. Robuuste algoritmes maken vervolgens een gesorteerde selectie van relevante adviezen voor een patiënt. Aan de hand van de gebruikersprofielen kunnen we de informatie in de adviezen van persoonlijke details voorzien.

Hoofdstuk 5 beschrijft een drietal experimenten waarin we de praktische werking van Wegweis hebben onderzocht, vergeleken, en geëvalueerd. Ten eerste hebben we de gebruiksvriendelijkheid onderzocht in een heuristische, kwalitatieve en kwantitatieve evaluatie. Vervolgens hebben we de functionele aspecten van ons systeem getest in een tweetal experimenten met patiënten en klinici. In het experiment met de klinici vergeleken we het identificeren van belangrijke problemen door het systeem met de meningen van klinici. In het experiment met de patiënten vergeleken we het selecteren van relevante adviesonderwerpen met de meningen van patiënten. De resultaten van de experimenten tonen aan dat voor de taak van het identificeren van de belangrijkste problemen uit een MANSA vragenlijst, een aanpak gebaseerd op probleemsterktes een gepaste benadering is voor de manier waarop klinici de informatie over een patiënt prioriteren. Daarnaast vonden we dat patiënten informatie niet op dezelfde manier prioriteren als klinici, maar dat de patiënten wel een meerderheid van het geselecteerde advies relevant vonden.

In Hoofdstuk 6 beschrijven we het ontwerp, de algoritmes, en de logica achter Autovar. Autovar is een applicatie die het proces van het bepalen van een vectorautoregressiemodel voor een bepaalde dataset automatiseert. Autovar onderscheidt zich door een aanpak die gebaseerd is op de manier waarop statistici werken. De aanpak is toegespitst op het specifieke geval van het vinden van vectorautoregressiemodellen voor data uit elektronische patiëntendagboeken. Vervolgens hebben we in deze aanpak verbeteringen aangebracht, zoals een algoritme dat betere constraints kan zetten. Het resultaat is een aanpak die meer aspecten kan automatiseren dan de alternatieven, en daarnaast ook meerdere modellen oplevert, gesorteerd op fitheid.

Hoofdstuk 7 beschrijft de implementatie en evaluatie van Autovar. We bespreken de technische details zoals gebruikte software en geïmplementeerde functies, en

illustreren de werking, invoer, opties, en uitvoer van de webapplicatie. In de evaluatie van Autovar vergelijken we het aantal gevonden modellen en de fitheid van de gevonden modellen op 60 data sets met die van experts werkend met STATA. De belangrijkste conclusie is dat Autovar sneller werkt, meer modellen vindt, en modellen met betere fitheid vindt dan de experts. Wel dient de tijdscomplexiteit in de gaten gehouden te worden. Ook hebben we Autovar vergeleken met de meest-gebruikte commerciële software voor het geautomatiseerd vinden van statistische modellen, PcGive. Alhoewel we de modellen niet direct konden vergelijken, viel bij een vergelijking van functionaliteit op dat Autovar simpeler is in het gebruik omdat het meer stappen automatiseert waar PcGive de expertise van de gebruiker vereist.

We hebben een aantal belangrijke stappen gezet op weg naar het geautomatiseerd verwerken van digitale medische gegevens. Ten behoeve van de patiënt hebben we aangetoond dat het geven van advies aan schizofreniepatiënten, een taak die voorzichtigheid en nauwkeurigheid vereist, zeer goed te automatiseren is. Ten behoeve van de clinicus hebben we aangetoond dat tevens complexe processen, zoals het vinden van het optimale vectorautoregressiemodel, te automatiseren zijn. Om de zorg toegankelijk en betaalbaar te houden in de komende decennia, verwachten we dat het noodzakelijk is om routineoperaties gebaseerd op digitale medische gegevens grotendeels te automatiseren. Echter, voordat we zover zijn, verwachten we een overgangperiode waarin dit soort systemen worden ontwikkeld, getraind, en gebruikt als *second opinion*.

