Psychometric Contributions to Improving the Efficiency and Fidelity of Clinical Assessment and Research
Fokkema, M.

2015

document version
Publisher's PDF, also known as Version of record

Link to publication in VU Research Portal

citation for published version (APA)

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

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.

E-mail address:
vuresearchportal.ub@vu.nl
Summary and Discussion

Summary

In this thesis, I have addressed a broad range of issues in clinical assessment, and presented new methods for improving its efficiency and accuracy. In clinical practice, assessments are often performed for making decisions: for example, assessment may be aimed at deciding which treatment should be provided to a patient. By taking the decision for which assessment is performed into account, the accuracy and efficiency of assessment can often be improved (Cronbach & Gleser, 1965). Improving the efficiency and accuracy of clinical assessment is important, as it reduces the costs of assessment (i.e., shorter assessment length, less incorrect decisions).

In Chapter 1, I discussed and applied curtailment, an algorithm that allows for early stopping of item administration when questionnaires are used for binary classification decisions. Application of (stochastic) curtailment to three mental health questionnaires was found to result in test length reductions of about 25% to 45%, without reducing diagnostic accuracy, compared to administration of the full-length test, and with better accuracy compared to an adaptive test with similar test length, based on item response theory. Compared to other adaptive testing procedures, curtailment offers three advantages: it involves minimal assumptions about the distribution of the data, allows for creation of easy to use look-up tables with stopping criteria, and allows for retaining the same item order in training and application of the test.
Classification decisions in clinical practice may often be based on several tests, instead of a single test. Therefore, in Chapter 2, I presented a new algorithm for assessment length reduction of tests batteries: CART-SC. CART-SC combines a classification tree (CART) with application of stochastic curtailment (SC) in every node of the tree, allowing for reduction of both the number of tests administered within a battery, as well as the number of items administered within tests. I simulated application of CART-SC on an existing dataset of item responses to a questionnaire about symptoms of mood and anxiety disorders, consisting of five subscales. Compared to the more traditional method of selecting subscale scores by means of linear discriminant analysis, and administering these subscales for classifying respondents, CART-SC provided an assessment length reduction of 56%, without reducing diagnostic accuracy.

The CART algorithm is a powerful method for selecting relevant attributes for decision-making in clinical practice, because CART trees provide sequential testing plans. In addition, CART can deal with a large number of potential predictor variables, can be used for subgroup detection, and allows for automatic detection of interactions between predictor variables. These characteristics are shared by all recursive partitioning methods (RPMs). Therefore, RPMs seem preeminently suited for improving the efficacy of clinical assessment, and more appropriate than many of the data-analytic methods traditionally used in clinical research. In Chapters 3 and 4, I further discussed the potential of RPMs for improving the efficacy of clinical assessment.

Clinical decisions are often based on predictions of future behavior. Therefore, in Chapter 3, I explored the use of (ensembled) prediction rules for clinical assessment. I argued that prediction rules may be easier to apply in decision making in clinical practice than the linear main-effects models usually applied in clinical research. As such, prediction rules are a promising tool for increasing the use of actuarial prediction methods in clinical practice. I provided an illustration by application of the RuleFit
algorithm (Friedman & Popescu, 2008) to a dataset on prediction of the course of depressive and anxiety disorders, and compared its accuracy and efficiency with that of logistic regression (LR). I found that the Rule-Fit ensemble provided predictive accuracy similar to that of LR, while requiring less cue evaluation and less computation. In addition, the rules in an ensemble can be visually represented as a simple decision tree, which may further improve usability in practice.

One of the most important and often encountered decisions in clinical practice are treatment decisions. In personalized medicine, this involves prescribing the treatment that has the greatest expected benefit, given a patient’s characteristics. In Chapter 4, I discussed the use of model-based recursive partitioning for detecting predictors of differential treatment efficacy, or treatment-subgroup interactions. Detection of such subgroups may often be performed in clustered datasets, where observations are clustered within research centers, persons (in longitudinal studies), or studies (in meta analysis), for example. I presented a new algorithm called lmertree, that allows for detection of treatment-subgroup interactions by means of model-based recursive partitioning, as well as taking into account the clustered structure of the data by estimation of cluster-specific random effects. In a simulation study, I compared the accuracy of lmertree with that of model-based recursive partitioning without random-effects estimation. The lmertree algorithm performed more accurate, both in recovering true treatment-subgroup interactions, and predicting treatment effect differences. Therefore, lmertree provides a powerful method for detecting treatment-subgroup interactions in clustered datasets.

The accuracy of predictions of treatment effects depends not only on the data-analytic method used, but also on the accuracy with which treatment outcomes are measured. Biased measurement of treatment outcomes will likely introduce bias in the estimation of a predictive model. Therefore, in Chapter 5, I discussed response-shift bias: changes in the measurement
models underlying total scores on self-report inventories. Changes in the structure or parameters of a measurement model can be interpreted as response shifts, which may obfuscate true change (i.e., changes in the underlying construct) when observed scores are used for assessing treatment outcomes (Oort, 2005). As an illustration, I examined the measurement models underlying item responses to the Beck Depression Inventory (Beck et al., 1961), from an influential trial comparing the effects of four treatments for depression. Results show that, compared to before treatment, after treatment, item scores overestimate depressive symptomatology, measurement errors are smaller and there is a stronger association between the underlying constructs. These effects were more apparent in psychotherapy groups, than in pharmacotherapy groups. These changes indicate that response shifts have occurred, and that observed-score comparisons over time yield confounded measures of treatment efficacy.

Discussion

Algorithmic and stochastic models for prediction

In this thesis, I have shown how curtailment, CART, model-based recursive partitioning and decision-rule ensembles may be used to reduce the number of attributes that require evaluation for decision making in clinical practice, without reducing predictive accuracy. An additional advantage of these algorithms is that their results (look-up tables, decision trees, prediction rules) may be more easily applicable by human decision makers, than the results of methods traditionally applied in clinical research (e.g., linear main-effects regression models).

Traditionally, researchers in psychology may be more interested in explanation and in testing of causal theories, than in prediction. This explains why researchers in psychology often favor the use of stochastic data models like (generalized) linear regression, which aim to recover the param-
eters of the stochastic process that generated the data (Shmueli, 2010; Breiman, 2001). However, by fitting such models to the data, researchers (inadvertently) impose a bias on the predictive model, as they involve many assumptions about the joint distribution of the data. Algorithmic models like CART or prediction rule ensembles, on the other hand, can model the association between input and output variables more flexibly, as they do not impose restrictions on the relationships between variables, and assume the data generating process unknown. Although algorithmic models generally aim at optimizing predictive accuracy, they are also powerful explanatory tools, providing valuable information about underlying mechanisms (Shmueli, 2010; Breiman, 2001). As shown in Chapter 3, for example, application of a relatively complex algorithm provided a small set of features relevant for prediction, which has at least as much explanatory power as a linear regression model.

Modern developments like CART and rule-based methods have been criticized by Hand (2006), who noted that the improvements provided by these techniques may be quite small, or even illusory. Hand (2006), in line with for example Holte (1993), concludes that “the marginal gain from complicated models is typically small compared to the predictive power of the simple models” (p. 12). This may seem to contrast with Chapters 2 and 3 of this thesis, where I concluded that CART and related methods are highly flexible and usable tools for prediction in clinical assessment. The results in these chapters, however, indicate that the predictive accuracy of partition-based models involving a small number of predictor variables does not differ much from that of linear models involving a small number of predictor variables. In other words, in terms of predictive accuracy, it may not matter much whether a compensatory or non-compensatory model is used, as long as the relevant predictor variables are selected. However, for efficiency in cue evaluation, it does matter whether a compensatory or non-compensatory model is chosen, which will be further discussed below in the subsection on sequential testing and recursive partitioning.
Response shifts

The algorithms aimed at optimizing predictive accuracy discussed in this thesis, are based on observed scores only. This is a major advantage, as they involve few assumptions about the distribution of the data. However, an obvious threat to predictive accuracy is a lack of reliability and validity of test scores. As illustrated in Chapter 5, treatment outcomes quantified in terms of total scores on self-report inventories may show bias due to response shifts. As a result, these total scores can not be taken as straightforward indicators of treatment effects. This may result in for example wrong conclusions on the efficacy of treatments in an RCT, or bias in the detection of treatment-subgroup interactions. The extent to which response-shift bias may result in biased predictions requires exploration in future research (cf., Millsap, 2007). In addition, although response shifts may adversely impact the quality of treatment decisions, they also provide valuable insight into different aspects of change. Therefore, performing tests of response-shift bias on datasets from clinical trials is important from both quantitative as well as qualitative viewpoints.

Sequential testing and recursive partitioning

Cronbach and Gleser (1965) introduced the sequential testing approach to psychological assessment, in which at every stage of testing, a new test is selected which is most informative for predicting class membership given previous outcomes, until a final classification decision can be made. RPMs take a similar approach to prediction of outcome variables: in every step, a variable and splitting value are selected that are most informative for predicting the value of the outcome variable, in the current partition of the dataset, until the prediction of the value of the outcome variable can no longer be improved.
As shown in Chapters 2 and 3 of this thesis, sequential testing can substantially reduce the number of cues that have to be evaluated in order to arrive at a final decision. RPMs hold great promise for sequential testing, and in this thesis I have shown how RPMs can be applied to a broad range of issues in clinical assessment. However, it should be noted that recursive partitioning methods are greedy algorithms, making a locally optimal choice at every step. Although this approach is an efficient heuristic, it is not guaranteed to find a global optimum in terms of accuracy and efficiency. Alternatively, the use of algorithms for finding globally optimal trees (Grubinger, Zeileis & Pfeiffer, 2011) or rules (Letham, Rudin, McCormick & Madigan, 2012) for creating sequential testing plans may be explored in future research.

Model-based recursive partitioning (Zeileis et al., 2008) offers a very generic framework for subgroup detection and sequential testing. In Chapter 4 of this thesis, I presented the lmertree algorithm, which is a first step to combining model-based recursive partitioning and random-effects estimation. Combining the latter two allows for making better recovery of information in clustered datasets, like for example in full-data meta analysis, which is becoming increasingly important as more data become available. The lmertree algorithm allows for accurate detection of treatment-subgroup interactions for continuous outcomes in clustered datasets, as discussed in Chapter 4. However, the algorithm also has some limitations, which may be addressed in future research, to allow for more flexible modeling of random and fixed-effects parameters. For example, the lmertree algorithm may be extended to allow for analysis of discrete outcomes. Further, stability tests for random-effects parameters are currently not available, but may be developed in future research.