Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder

R. C. Kessler1*, H. M. van Loo2, K. J. Wardenaar2, R. M. Bossarte3, L. A. Brenner4, D. D Ebert1,5, P. de Jonge2, A. A. Nierenberg6, A. J. Rosellini1, N. A. Sampson1, R. A. Schoevers2, M. A. Wilcox7 and A. M. Zaslavsky1

1 Department of Health Care Policy, Harvard Medical School, Boston, MA, USA
2 Interdisciplinary Center Psychopathology and Emotion Regulation (ICPE), University of Groningen, University Medical Center Groningen, Groningen, The Netherlands
3 Department of Veterans Affairs, Office of Public Health, Washington, DC, USA
4 VISN 19 Mental Illness Research Education and Clinical Center, University of Colorado, Anschutz Medical Campus, Anschutz, CO, USA
5 Department of Psychology, Clinical Psychology and Psychotherapy, Friedrich-Alexander University Nuremberg-Erlangen, Erlangen, Germany
6 Department of Psychiatry and Depression Clinical and Research Program, Harvard Medical School and Massachusetts General Hospital, Boston, MA, USA
7 Department of Epidemiology, Janssen Research and Development, Titusville, NJ, USA

Backgrounds. Clinicians need guidance to address the heterogeneity of treatment responses of patients with major depressive disorder (MDD). While prediction schemes based on symptom clustering and biomarkers have so far not yielded results of sufficient strength to inform clinical decision-making, prediction schemes based on big data predictive analytic models might be more practically useful.

Method. We review evidence suggesting that prediction equations based on symptoms and other easily-assessed clinical features found in previous research to predict MDD treatment outcomes might provide a foundation for developing predictive analytic clinical decision support models that could help clinicians select optimal (personalised) MDD treatments. These methods could also be useful in targeting patient subsamples for more expensive biomarker assessments.

Results. Approximately two dozen baseline variables obtained from medical records or patient reports have been found repeatedly in MDD treatment trials to predict overall treatment outcomes (i.e., intervention v. control) or differential treatment outcomes (i.e., intervention A v. intervention B). Similar evidence has been found in observational studies of MDD persistence-severity. However, no treatment studies have yet attempted to develop treatment outcome equations using the full set of these predictors. Promising preliminary empirical results coupled with recent developments in statistical methodology suggest that models could be developed to provide useful clinical decision support in personalised treatment selection. These tools could also provide a strong foundation to increase statistical power in focused studies of biomarkers and MDD heterogeneity of treatment response in subsequent controlled trials.

Conclusions. Coordinated efforts are needed to develop a protocol for systematically collecting information about established predictors of heterogeneity of MDD treatment response in large observational treatment studies, applying and refining these models in subsequent pragmatic trials, carrying out pooled secondary analyses to extract the maximum amount of information from these coordinated studies, and using this information to focus future discovery efforts in the segment of the patient population in which continued uncertainty about treatment response exists.

Introduction

Patients with major depressive disorder (MDD) vary substantially in treatment response and illness course. This heterogeneity of treatment effects (HTE) complicates clinical decision-making. Clinicians have consistently identified the absence of dealing with this variation as a critical gap in personalising MDD treatment (Altschuler et al. 2001; Perlis, 2007; Hetrick et al. 2011; Kuiper et al. 2013). Researchers have tried to address this gap by searching for depression subtypes defined by presumed causes (e.g., postnatal depression) (Cooper & Murray, 1995; Cooper et al. 2007), clinical presentations (e.g., atypical or melancholic depression (Fink et al. 2007; Uher et al. 2011))

* Address for correspondence: Dr R. C. Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, USA.
(Email: Kessler@hcp.med.harvard.edu)
or empirically derived symptom profiles (e.g., based on cluster analysis (Andreasen & Grove, 1982), factor analysis (Romera et al. 2008) and latent class analysis (Lamers et al. 2012)) in hopes of predicting differential treatment response, but results have been disappointing (Baumeister & Gordon, 2012; van Loo et al. 2012). More recent efforts have searched for genetic, neuroendocrine, electrophysiological and brain imaging biomarkers of treatment response (Pizzagalli, 2011; Souslova et al. 2013; Breitenstein et al. 2014; Perlis, 2014), but have failed so far to yield results of sufficient strength to inform clinical decision-making (Simon & Perlis, 2010). Guidelines for MDD treatment selection consequently continue to be based on simple clinical observations about overall MDD severity (National Institute for Health and Clinical Excellence (NICE), 2009; American Psychiatric Association, 2010).

Another promising approach for studying predictors of differential treatment response has received much less attention: to use supervised machine learning methods to develop multivariate prediction equations of treatment outcomes based on symptoms and other easily assessed clinical features that have been found in previous research to predict MDD treatment outcomes (Strobl et al. 2009; Zhang & Singer, 2010; van der Laan & Rose, 2011; James et al. 2013). Although such methods have been used in this way in other areas of medicine (Chang et al. 2012; Chao et al. 2012), applications to MDD have so far been based on samples too small to realise the potential of the methods (Andresen et al. 2008a; Rabinoff et al. 2011; Riedel et al. 2011; Nelson et al. 2012; Jain et al. 2013). Yet, promising preliminary results exist in clinical (Moos & Cronkite, 1999; Perlis, 2013) and community epidemiological (Angst et al. 2011; van Loo et al. 2014) studies designed to predict MDD persistence-severity. In addition, innovative statistical methods for building such models exist but have not yet been applied to MDD (Kent et al. 2010; van der Laan & Gruber, 2010; Diaz Munoz & van der Laan, 2011; Willke et al. 2012; Burke et al. 2014; Neugebauer et al. 2014). We review these developments in the current report.

**Self-report predictors of heterogeneity of MDD treatment effects**

We reviewed the literature on self-reported predictors (i.e., assessed by survey or questionnaire; non-biomarker) of MDD treatment response beginning with a PubMed search using the search string: depress* AND predict* AND (‘treatment outcome’ OR ‘treatment response’ OR ‘course’) AND (‘self-report’ OR ‘survey’ OR ‘questionnaire’). Abstracts were then reviewed and articles read in full if the abstract indicated that: (i) participants underwent treatment for depression (randomised controlled trials, uncontrolled treatment trials, observational studies in which participants were in a treatment during the follow-up period); and (ii) associations were examined between baseline self-report constructs and MDD treatment outcomes. We also accessed and read any studies that were cited in these papers to have examined baseline self-reported predictors of MDD treatment outcomes. All examined associations were recorded on a spreadsheet. If both bivariate and multivariate models were estimated, we recorded the results from the multivariate models.

Replicated significant associations were found between roughly two dozen baseline self-reported constructs and subsequent MDD treatment outcomes (Table 1). It is noteworthy, though, that the typical study reviewed considered only a handful of these modifiers and no single study included all modifiers. Analyses considering only a small number of modifiers are unlikely to provide reliable clinical guidance due to the existence of many HTE predictors, while more complex subgroup analyses are precluded by the small size of MDD treatment trials (Simon & Perlis, 2010; Cuijpers et al. 2012).

**Prior attempts to develop models of heterogeneity of MDD treatment effects**

While, as noted above, no prior study of MDD HTE has included all the predictors in Table 1, encouraging preliminary results nonetheless exist. The first effort along these lines was that of Perlis (Perlis, 2013), who carried out a secondary analysis of the STAR*D dataset, where MDD treatment response was predicted with an area under the receiver operating characteristic curve (AUC) of 0.71 using a simple logistic regression equation containing a small number of easily accessible patient self-report measures (socio-demographics, depressive symptoms, comorbidity and prior MDD history). An AUC of 0.71 is similar to the levels of prediction accuracy found in a number of widely used risk prediction models in other areas of medicine (Anothaiatstawee et al. 2012; Siontis et al. 2012; Echouffo-Tcheugui & Kengne, 2013). However, this analysis focused on overall treatment response rather than differential response across multiple treatments.

In comparison, Kraemer (2013) developed an approach to estimate MDD HTE in a treatment trial comparing the relative effectiveness of exactly two treatment types. The three-step approach began by estimating a conventional modifier model for each potential modifier one at a time (i.e., including predictor variables for treatment type (a dummy variable), the modifier and an interaction term between treatment and the modifier). The second step then consisted of
Table 1. Baseline constructs associated with poor overall depression treatment response and/or differential treatment responses in two or more studies

<table>
<thead>
<tr>
<th>Baseline construct</th>
<th>References demonstrating direct effect on (poor) overall course/treatment response</th>
<th>References demonstrating interaction of baseline construct with treatment type (Treatment A v. Treatment B differential outcome)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Socio-demographics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than full-time employment</td>
<td>Jarrett et al. (2013), Rush et al. (2008), Trivedi et al. (2006)</td>
<td>Fournier et al. (2009) (CT &gt; SSRI)</td>
</tr>
<tr>
<td>Low income</td>
<td>Cohen et al. (2009), Cohen et al. (2006), Trivedi et al. (2006)</td>
<td></td>
</tr>
<tr>
<td>Low education</td>
<td>Agid &amp; Lerer (2003), Marquett et al. (2013), Perlis (2013), Szadoczky et al. (2004)</td>
<td></td>
</tr>
<tr>
<td>Unmarried</td>
<td>Frank et al. (2011), Rush et al. (2008), Trivedi et al. (2005)</td>
<td></td>
</tr>
<tr>
<td>Married/cohabitating</td>
<td></td>
<td>Fournier et al. (2009) (CT &gt; SSRI); Barber &amp; Muenz (1996) (CT &gt; IPT)</td>
</tr>
<tr>
<td>Depression history, triggers and symptoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High baseline severity</td>
<td>Agid &amp; Lerer (2003), Barber &amp; Muenz (1996), Bernecker et al. (2014), Hollon et al. (2014), Jarrett et al. (2013), Perls (2013), Rush et al. (2008), Sotsky et al. (1991), Thase et al. (1997), Thase et al. (1994), Vuorilehto et al. (2009)</td>
<td>Luty et al. (2007) (CBT &gt; IPT); Dimidjian et al. (2006) (BA &gt; CT); Elkin et al. (1989) (IPT &gt; CBT); Hollon et al. (2014) (CT &gt; MED &gt; MED); Thase et al. (1997) (IPT &gt; MED &gt; IPT); Thase et al. (2007) (SNRI &gt; SSRI); Kennedy et al. (2006) (Second generation SSRI &gt; Conventional SSRI)</td>
</tr>
<tr>
<td>High psychomotor symptoms/ low energy</td>
<td>Frank et al. (2011), Perlis (2013)</td>
<td></td>
</tr>
<tr>
<td>High anhedonia/ melancholic features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atypical depression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic/double depression/ longer episodes</td>
<td>Fournier et al. (2009), Lowe et al. (2005), sung et al. (2012), Trivedi et al. (2005)</td>
<td>Cuijpers et al. (2012) (PSY + MED &gt; MED, PSY + MED &gt; PSY, MED &gt; PSY)</td>
</tr>
<tr>
<td>Endogenous depression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current/past suicidality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greater number prior/recurrent episodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earlier age of onset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comorbid disorders/symptoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personality disorder/ symptoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychotic disorder/ symptoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depression history, triggers and symptoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High baseline severity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High psychomotor symptoms/low energy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High anhedonia/ melancholic features</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Atypical depression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chronic/double depression/ longer episodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Endogenous depression</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Current/past suicidality</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greater number prior/recurrent episodes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Earlier age of onset</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comorbid disorders/symptoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personality disorder/ symptoms</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Psychotic disorder/ symptoms</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Continued
estimating a multivariate model to create regression weights for all modifiers judged to be important in the first step. These modifiers were then combined in a third step into a single composite HTE measure for each patient by summing the products $b_m \times M_{im}$, where $b_m$ is the slope of the treatment outcome on modifier $M$ and $M_{im}$ is the score of respondent $i$ on modifier $M$. ‘Importance’ of individual modifiers in the first step was defined by the standardised correlation between modifier scores and differences in treatment effect across the two treatment types, where the latter association was estimated in a person-pair dataset for all $n_1 \times n_2$ pairs of patients in either Treatment A ($n_1$) or Treatment B ($n_2$).

Table 1. Continued

<table>
<thead>
<tr>
<th>Baseline construct*</th>
<th>References demonstrating direct effect on (poor) overall course/treatment response</th>
<th>References demonstrating interaction of baseline construct with treatment type (Treatment A v. Treatment B differential outcome)b</th>
</tr>
</thead>
<tbody>
<tr>
<td>High somatic anxiety Stress and adversity</td>
<td>Feske et al. (1998)</td>
<td>Frank et al. (2011) (IPT &gt; SSRI)</td>
</tr>
<tr>
<td>Childhood maltreatment/trauma High stress Poor physical functioning Poor social support/relationships</td>
<td>Carter et al. (2011), Johnstone et al. (2009), Nanni et al. (2012)</td>
<td>Nemeroff et al. (2003) (CBT &gt; SSRI)</td>
</tr>
<tr>
<td></td>
<td>Dew et al. (1997)</td>
<td>Fournier et al. (2009) (CT &gt; SSRI)</td>
</tr>
<tr>
<td></td>
<td>Lowe et al. (2005), Trivedi et al. (2006)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Denton et al. (2010), Dew et al. (1997), Marquett et al. (2013)</td>
<td></td>
</tr>
<tr>
<td>Personality traits Non-secure attachment style High negative affect/neuroticism Low openness to experience</td>
<td>Constantino et al. (2013)</td>
<td>McBride et al. (2006) (CBT &gt; IPT)</td>
</tr>
<tr>
<td></td>
<td>Joyce et al. (2007), Szadoczky et al. (2004), Vrieze et al. (2014)</td>
<td>Bagby et al. (2008) (SSRI &gt; CBT)</td>
</tr>
<tr>
<td></td>
<td>Bagby et al. (2008), Marquett et al. (2013)</td>
<td></td>
</tr>
</tbody>
</table>

IPT, interpersonal psychotherapy; CT, cognitive therapy; CBT, cognitive-behavioural therapy; BA, behavioural activation; PA, psychoanalysis; PSY, psychotherapy (non-specific); SSRI, selective serotonin reuptake inhibitor; SNRI, serotonin–norepinephrine reuptake inhibitors; TCA, tricyclic antidepressant; AAD, atypical antidepressant (e.g., bupropion); MAOI, monoamine oxidase inhibitor; MED, pharmacotherapy (non-specific); +, combined treatment.

*Predictor and outcome measures varied by study, and only constructs with statistically significant ($p < 0.05$) associations with depression treatment outcome (overall or differential response) in two or more studies are presented here.

bDifferential treatment response depending on the baseline construct is shown in parentheses. Treatment type is operationalised based on broad classes of psychotherapy (e.g., any PSY, IPT, CT, CBT, BA, PA) and pharmacotherapy (e.g., any MED, SSRI, SNRI, TCA, AAD, MAOI). The treatment associated with the better response (among patients with the baseline construct) is listed before the >. In other words, $X > Y$ means that treatment $X$ is favoured relative to treatment $Y$ if the (row) construct is present at baseline.
In an illustration of this approach applied to a small treatment trial in which patients were randomised to receive either interpersonal therapy (IPT) or a selective serotonin reuptake inhibitor (SSRI) and 32 potential treatment effect modifiers were assessed at baseline, Kraemer showed that even though the overall effect size was approximately 0 (i.e., patients responded equivalently in the aggregate to the two treatments), the effect size was 0.50 favouring IPT over SSRI in the segment of the sample in which a composite HTE score (made up of 8 of the 32 original baseline measures) favoured IPT (representing 44% of the patients in this particular trial) and 0.48 favouring SSRI over IPT in the remainder of the sample. (See also Wallace et al. 2013 for a more substantive presentation of the same results.)

While this illustration makes it clear that baseline information could be of great value in guiding clinical decision-making about MDD treatment selection, it is also important to point out that the Kraemer approach to defining individual-level HTE is limited in that it provides no practical way to estimate an optimal clinical decision support model for choosing among the wider range of treatments available for MDD (e.g., IPT, cognitive therapy (CT), behavioural activation, cognitive-behavioural therapy, or some other type of psychotherapy; SSRI, serotonin-norepinephrine reuptake inhibitor (SNRI) or some other type of pharmacotherapy; any combination of a particular psychotherapy with a particular pharmacotherapy). Nor does the Kraemer approach allow for the estimation of stable models that make use of the large number of potential modifiers, some of which might be highly inter-correlated, in ways that consider the possible existence of complex non-linear and/or non-additive multivariate associations (e.g., three-way interactions) with response to particular types of treatment.

DeRubeis et al. (2014) proposed an approach to MDD HTE estimation very similar to the Kraemer approach in that it began by estimating a conventional modifier model for each potential modifier one at a time. However, it differed from the Kraemer approach in that subsequent steps of model-building that combined important modifiers (where ‘important’ was defined initially as significant at the 0.20 level when modifiers were considered one at a time, at the 0.10 level when included in subsequent within-domain multivariate models, and at the 0.05 level when included in final cross-domain models) were carried out at the person level rather than, as in the Kraemer approach, at the person-pair level. This person-level analysis allowed DeRubeis to generate a predicted treatment outcome score for each patient based on the final model separately for the actual type of treatment received as well as based on the counter-factual assumption that the patient had received another type of treatment. Individual-level comparison of these two predicted scores then allowed DeRubeis to determine the preferred treatment for each patient.

In an application of this approach to a small treatment trial in which patients were randomised to receive either antidepressant medication or cognitive behaviour therapy (CBT), interactions of treatment type with 38 potential baseline modifiers (as detailed in Fournier et al. 2009) were estimated initially one at a time and then in sequential multivariate models to arrive at a final model that included nine significant (0.05 level in third-step models) predictors either having interactions with type of treatment (five predictors) or associated with treatment outcome equivalently for both types of treatment (four predictors). Roughly 60% of patients had predicted outcome scores based on the model that differed between the two types of treatment by an amount considered clinically significant (three points on the Hamilton Rating Scale for Depression). In the aggregate, patients in this 60% of the sample who were randomised to the type of treatment to which they were predicted by the model to have better response had a treatment effect size 0.58 greater than that of patients who were randomised to the other treatment.

This research team subsequently applied the same method to another small treatment trial that randomised patients with MDD either to IPT or CT (Huibers et al. 2015). A total of 43 potential baseline modifiers were available and the final model included 13 of them (eight with significant interactions and five others that were associated with treatment outcomes equivalently across the two treatments). In the aggregate, patients who were randomised to the type of treatment to which they were predicted by the model to have better response had a treatment effect size 0.51 greater than that of patients who were randomised to the other treatment.

As with the Kraemer study, the two studies by DeRubeis and colleagues illustrate the potential value of using baseline information to help clinicians select personalised MDD treatments. As with the Kraemer study, though, it is quite likely that the method used by DeRubeis and colleagues would lead to model overfitting; that is, to a situation in which application of the models in independent patient datasets would lead to prediction accuracy being lower, perhaps substantially so, than in the sample in which the models were built. As we discuss later, machine learning methods are designed to address this problem of overfitting. The ad hoc stepwise model-building procedures used by Kraemer and DeRubeis are far inferior to methods designed explicitly to maximise prediction accuracy in independent samples.
It is noteworthy in this regard that the last step of the DeRubeis approach used the leave-one-out (LOO) method to impute individual-level predicted outcome scores. DeRubeis and colleagues asserted that this method addressed the problem of overfitting, but this assertion is incorrect. Overfitting almost certainly occurred at the level of variable selection, where stepwise analysis was used to create models with interactions involving 5 of 38 (the CBT v. SSRI trial) and 8 of 43 (the IPT v. CT trial) initially examined predictors in very small treatment samples (n = 154 in the CBT v. SSRI trial; n = 134 in the IPT v. CT trial). This means that any attempt to use the coefficients in these models to predict differential treatment response in a new sample of patients would almost certainly yield less positive effects than those suggested by the results of studies. The use of the LOO method to estimate the likely strength of the model in an independent sample is an invalid approach when LOO is applied after selection of the final model predictors. At the level of variable selection, furthermore, use of LOO is widely recognised to be suboptimal compared with other types of cross-validation due to the fact that it has high variance (Hastie et al. 2009).

Is there a better way?

Best practices recommendations for HTE analysis call for a different approach. In the simple case of a single treatment v. control evaluation, these recommendation call for a three-step approach: (i) estimate the joint effects of baseline predictors in multivariate prediction equations applied either to an independent sample of people with the disorder (Kent et al. 2010) or, if the clinical trial sample is large enough, to the control group of the trial in which the predictors are being studied (Burke et al. 2014); (ii) apply the predicted probabilities of treatment outcomes from these equations to both intervention and control patients; and (iii) plot treatment outcomes separately in the intervention and control groups to examine differences in absolute risk reduction (ARR) as a function of these predicted probabilities. Patients with high predicted probabilities of recovery will have low ARR because they will recover even without treatment. Patients with low probabilities of recovery might also have low ARR due to available treatments being ineffective in these difficult cases. Depending on the proportions of patients at these tails of the distribution, the trial might be negative overall even though ARR is significant among patients with intermediate predicted probabilities of recovery. This approach to multivariate HTE analysis has proven useful in guiding personalised treatment planning in other areas of medicine (Hayward et al. 2006; Dorresteijn et al. 2011) even though the prediction equations have largely focused exclusively on overall treatment response rather than differential treatment response.

An expansion of this approach to HTE involving multiple types of treatment would either require estimation: (i) of a separate model for each type of treatment v. controls; (ii) of a separate within-treatment model for each type of treatment; or (iii) of a pooled model across active comparator treatments that allowed for interactions of dummy variables for treatment type with baseline variables. In the ideal case, these models would be estimated using modern machine learning methods rather than the ad hoc methods used by Kraemer and DeRubeis in order to reduce the problem of overfitting and maximise out-of-sample performance when applied to independent patient samples (Ritchie, 2005; Upstill-Goddard et al. 2013).

The results of such models could be applied in subsequent patient samples by comparing estimated treatment outcomes for each patient separately for each treatment option to arrive at an estimate of the optimal treatment for each patient. In order to do this, though, the logic would require a large enough trial to obtain stable coefficient estimates within treatment-specific subsamples and the application of the coefficients from that trial to subsequent trials. This is infeasible in the case of MDD treatment trials, though, because MDD treatment trials are too small to support such an analysis. Another problem is that MDD treatment trials do not use a stable set of baseline measures of the sort outlined in Table 1. The problem of small sample size is largely responsible for the fact that Kraemer, in the approach described above, carried out the analysis using patient-pair data, as the approach needed to conserve degrees of freedom in a sample that consisted of only n = 291 patients randomised between two conditions and the analysis examined the modifying effects of 32 baseline predictors. And the problem of inconsistency in baseline measure is largely responsible for the fact that no efforts have been made to pool results across a large number of MDD treatment trials to estimate complex MDD HTE models.

This problem of small sample size has been addressed in other areas of medicine either by developing interactive HTE models based on very large trials (e.g., the Use of Statins in Prevention trial, which randomised 17802 initially healthy men and women to statins or placebo for 10 years to evaluate the effect of early statin use in preventing cardiac events (Dorresteijn et al. 2011) or by using previously developed external risk scores based on prediction equations developed either in large observational samples or in pooled samples that combine data across the multiple observational studies and/or clinical trials (Perel et al. 2006; Prieto-Merino & Pocock, 2012).
The latter would be the more practical approach for MDD HTE, possibly beginning with large observational samples of patients beginning MDD treatment, administering self-report surveys of constructs found previously to predict MDD HTE, and following these patients through treatment to assess treatment outcomes. These data would then be analysed using statistical methods recently developed to estimate comparative treatment effectiveness in observational studies (van der Laan & Gruber, 2010; Neugebauer et al. 2014). Inspection of between-patient differences in predicted outcomes pooled across all treatments could be used to study individual differences in overall treatment outcomes, while inspection of between-patient differences in treatments associated with highest predicted probabilities of recovery could be used to study individual differences in differential treatment outcomes. Internal cross-validation could be used to evaluate in-sample performance. If clinically meaningful individual differences were documented in this way, the same approach could be used in subsequent MDD clinical trials to support HTE analyses.

The above approach would need to begin with large thoughtfully constructed (Madigan et al. 2014) observational samples because the sample sizes of even the largest MDD clinical trials would be much too small to provide stable estimates of predicted HTE (Madigan et al. 2014). Although HTE estimates are biased in observational studies if treatment assignment is informatively non-random, statistical methods exist to adjust for this bias (Picciotto et al. 2014). This is true even for non-random variation in dynamic treatment assignment (e.g., due to side effects or lack of early treatment response) (Suarez et al. 2008; Liu et al. 2014) and for unmeasured determinants (Lin & Chen, 2014; Tchetgen Tchetgen, 2014). Consistent with these observations, a recent Cochrane review concluded that treatment effect size estimates based on well-analysed observational studies are very similar to those based on randomised controlled trials (Anglemyer et al. 2014).

**Preliminary results**

Although we are aware of no existing efforts to develop a multivariate model of MDD treatment response in a clinical trial sample along the lines suggested above, a potentially useful model can be found in a series of studies designed to examine multivariate predictors of long-term depression persistence-severity in secondary analyses of the 1990–1992 National Comorbidity Survey (NCS; Kessler et al. 1994), 2001–2003 NCS follow-up survey (NCS-2; Kessler et al. 2003), 2001–2003 NCS Replication (NCS-R; Kessler et al. 2004), and WHO World Mental Health (WMH) surveys (Demyttenaere et al. 2004). We briefly review the results of these studies in this section of the paper and then discuss prospects for extending the methods used to examine of MDD HTE.

The NCS and NCS-R were nationally representative community epidemiological surveys of common mental disorders in the USA. The NCS-2 was a follow-up survey of NCS 10–12 years after baseline. The WMH surveys were national or regional surveys based on NCS-R in 15 other countries. Initial exploratory analyses based on unsupervised clustering found patterns suggesting that significant associations existed between retrospective reports about incident episode symptoms and subsequent illness course in the NCS-R data. These results were sufficiently promising that subsequent supervised machine learning analyses were carried out to maximise the prediction of MDD persistence-severity from retrospectively reported information on incident episode symptoms in the much larger WMH series, where there were 8261 respondents with lifetime DSM-IV MDD (van Loo et al. 2014; Wardenaar et al. 2014).

Two machine learning algorithms (ensemble recursive partitioning, penalised regression) were used to examine associations of the outcomes with predictors that consisted of retrospectively reported parental history of depression, temporally primary comorbid disorders, and characteristics of incident MDD episodes. The outcomes were two measures of retrospectively reported subsequent MDD persistence (number of years with episodes and with episodes lasting most days throughout the year) and two measures of subsequent MDD severity (hospitalisation; work disability). K-means cluster analysis of the four predicted values found three risk strata that parsimoniously characterised multivariate associations. The high-risk cluster (32.4% of cases) accounted for 56.6–72.9% of high persistence-severity, with area under the receiver operating characteristic curve (AUC) of 0.63–0.70.

As these WMH results were retrospective, a validation study was subsequently undertaken in the NCS/NCS-2 panel. Predicted outcome scores were generated from information collected in the baseline survey scored using model coefficients estimated in the WMH analysis. Associations of these predicted values with outcomes over the intervening 10–12 years were then examined using reports obtained in the NCS-2 follow-up survey. These prospective associations were comparable to the retrospective associations found in WMH (Kessler et al. 2016). Importantly, meaningful discrimination was found both at the upper and lower ends of the predicted outcome distributions. For example, the respondents classified at baseline as being in the top quintile of risk accounted...
for 55.8% of all suicide attempts over the subsequent 10–12 years, while the respondents in the lowest baseline risk quintile accounted for only 1.5% of subsequent suicide attempts.

It is instructive to compare these NCS/NCS-2 results to those of other prospective studies that used baseline measures to predict MDD persistence-severity over 10+ years in samples of initially depressed patients (Moos & Cronkite, 1999; Mueller et al. 1999; Klein et al. 2008; Cronkite et al. 2013) or community residents (Mattisson et al. 2007; Bradvik et al. 2008; Eaton et al. 2008; Fichter et al. 2010; Angst et al. 2011). Although these studies were all quite small (n = 87–424) and none reported AUC, AUC could be computed post hoc from two of them. The first study was a 50-year follow-up of the 19I respondents in the Lundby community study with baseline MDD, 20 of whom subsequently died by suicide (Bradvik et al. 2008). A composite measure of baseline depression severity predicted subsequent suicide with AUC = 0.69 compared with AUC = 0.70 for the most comparable NCS-2 outcome (attempted suicide). The second study followed 313 depressed outpatients 1, 4 and 10 years after baseline and defined chronic depression as either (i) meeting full criteria for MDD at any 2 follow-ups or (ii) meeting full criteria at the 10-year follow-up and partial criteria at both earlier assessments (Moos & Cronkite, 1999). Twenty baseline predictors (depressive symptoms, self-concept, social function and coping) predicted chronicity with AUC = 0.70 compared to AUC = 0.66 for the most comparable NCS-2 outcome (high persistence of episodes). In making these comparisons, it is important to remember that the AUCs in these other studies were not validated in independent samples.

It is also noteworthy that the predictors in the WMH and NCS/NCS-2 studies as well as in the above studies were much less comprehensive than those in Table 1. This means that the estimates of prediction strength in these studies are likely to be lower bounds. A preliminary expansion of the NCS/NCS-2 analysis to go beyond the incident episode predictors considered so far include all the predictors in the baseline survey listed in Table 1. While still only a subset of all the predictors in Table 1, AUC increased to more than 0.80 for each NCS-2 outcomes when the predictors were expanded in this way.

**Implications for developing models of heterogeneity of MDD treatment effects**

Given the above results, one potentially useful next step in studying MDD HTE would be to develop a self-report questionnaire based on the predictors in Table 1, administer that questionnaire to large observational samples of patients at the beginning of MDD treatment, monitor treatment types and responses, and analyse these data to generate predicted MDD treatment outcome scores that could be used as the basis of HTE analyses in subsequent clinical trials. If many different researchers carrying out prospective observational studies and controlled MDD treatment trials used a consistent questionnaire of this type, results could be pooled to predict HTE. There is precedent for this kind of pooling of observational and controlled studies to study consistency of estimated treatment effects and the roles of observational study confounding, compositional differences and variation in treatments in accounting for between-study discrepancies (Prentice et al. 2006; Toh & Manson, 2013).

Another possible extension would be to carry out subsequent pragmatic trials (Lurie & Morgan, 2013) in the same treatment systems where prior observational studies were carried out by randomising participating clinicians either to receive or not to receive actuarial information about optimal treatments based on HTE models for individual patients based on administration of questionnaires prior to initiation of treatment. The treatment outcomes of the patients included in this randomisation could then be tracked to evaluate the effects of making this personalised clinical decision support tool available to clinicians. These predictions could also be used to determine which patients should be targeted for randomisation to interventions involving expensive biomarkers (Uher et al. 2010; Williams et al. 2011; Dunlop et al. 2012; Kennedy et al. 2012; Wallace et al. 2013) that would only be needed if the actuarial model based on self-report questionnaire data yielded equivocal results (Van Staa et al. 2012). It would also be valuable in this context to evaluate the incremental value of promising biomarkers in improving prediction beyond the level achieved less expensively using only self-report data (Li & Lu, 2010; Steyerberg et al. 2014).

**Conclusions**

Significant associations exist between numerous self-report measures and subsequent MDD outcomes. These associations have been documented both in the controlled treatment trials, where the outcomes were measures of treatment response, and in observational studies, where the outcomes were more general measures of MDD persistence-severity. Although no large-scale prospective study has been carried out to evaluate the joint effects of all these predictors at once, the preliminary results reviewed above make a good case that the resulting multivariate equations might be of clinical value in predicting both absolute and differential treatment response. The use of recent advances in machine learning methods to detect
interactions could be valuable in refining these equations, while the use of recent advances in statistical methods to make causal inferences from observational data could help reduce bias in estimating HTE due to non-random treatment assignment and informative loss to follow-up. These equations could then be used both to generate individual-level predicted outcome scores to support the investigation of MDD HTE in subsequent controlled treatment trials and provide useful decision support for clinicians attempting to optimise the treatment of their depressed patients.

Acknowledgements

None.

Financial Support

The NCS data collection was supported by the National Institute of Mental Health (NIMH; R01MH46376). The NCS-2 data collection was supported by the National Institute on Drug Abuse (NIDA; R01DA012058). Dr de Jonge was supported by a VICI grant (no. 91812607) from the Netherlands organisation for Scientific research (NWO-ZonMW). The NCS surveys were carried out in conjunction with the World Health Organization World Mental Health (WMH) Survey Initiative, which was supported by the National Institute of Mental Health (R01MH070884), the John D. and Catherine T. MacArthur Foundation, the Pfizer Foundation, the US Public Health Service (R13MH066849, R01MH069864 and R01DA016558), the Fogarty International Center (FIRCA R03TW006481), the Pan American Health Organization, Eli Lilly and Company, Ortho-McNeil Pharmaceutical, Inc., GlaxoSmithKline, and Bristol-Myers Squibb. Dr de Jonge was supported by a VICI grant (no. 91812607) from the Netherlands organization for Scientific research (NWO-ZonMW). Preparation of this report was supported by Janssen Pharmaceuticals.

Conflict of Interest

In the past 3 years, Dr Kessler has been a consultant for Abbott Laboratories, American Psychiatric Association, Appliance Computing Inc. (MindSite), Basilea, Brain Cells, Inc., Brandeis University, Bristol Myers Squibb, Clintara, Corcept, Dey Pharmaceuticals, Dainippon Sumitomo (now Sunovion), Eli Lilly and Company, EpiQ, L.P./Mylan Inc., Forest, Genaissance, Genentech, GlaxoSmithKline, Hoffman LaRoche, Infomedic, Lundbeck, Janssen Pharmaceuticals, Jazz Pharmaceuticals, Medavante, Merck, Methylation Sciences, Naurex, Novartis, PamLabs, Pfizer, PGx Health, Ridge Diagnostics Shire, Schering-Plough, Somerset, Sunovion, Takeda Pharmaceuticals, Targacept, and Teva; consulted through the MGH Clinical Trials Network and Institute (CTNI) for Astra Zeneca, Brain Cells, Inc, Dainippon Sumitomo/Sepracor, Johnson and Johnson, Labopharm, Merck, Methylation Science, Novartis, PGx Health, Shire, Schering-Plough, Targacept and Takeda/Lundbeck Pharmaceuticals; had grant/research support from the American Foundation for Suicide Prevention, AHRQ, Brain and Behavior Research Foundation, Bristol-Myers Squibb, Cederroth, Cephalon, Cyberonics, Elan, Eli Lilly, Forest, GlaxoSmithKline, Janssen Pharmaceuticals, Lichtwer Pharma, Marriott Foundation, Mylan, NIMH, PamLabs, PCORI, Pfizer Pharmaceuticals, Shire, Stanley Foundation, Takeda, and Wyeth-Ayerst; received honoraria from Belvoir Publishing, University of Texas Southwestern Dallas, Brandeis University, Bristol-Myers Squibb, Hillside Hospital, American Drug Utilization Review, American Society for Clinical Psychopharmacology, Baystate Medical Center, Columbia University, CRICO, Dartmouth Medical School, Health New England, Harold Grinspoon Charitable Foundation, IMEDEX, International Society for Bipolar Disorder, Israel Society for Biological Psychiatry, Johns Hopkins University, Mf Consulting, New York State, Medscape, MBL Publishing, MGH Psychiatry Academy, National Association of Continuing Education, Physicians Postgraduate Press, SUNY Buffalo, University of Wisconsin, University of Pisa, University of Michigan, University of Miami, University of Wisconsin at Madison, APSARD, ISBD, Scimed, Slack Publishing and Wolters Kluwer Publishing; owns stock in Appliance Computing, Inc. (MindSite); Brain Cells, Inc., Medavante; and owns the following copyrights: Clinical Positive Affect Scale and the MGH Structured Clinical Interview for the Montgomery Asberg Depression Scale exclusively licensed to the MGH Clinical Trials Network and Institute (CTNI). Dr Wilcox is an employee of Janssen Pharmaceuticals. The remaining authors report nothing to disclose. The remaining authors report nothing to disclose. The funders/sponsors were not involved in any aspect of the design and conduct of the study other than for the participation of Dr Wilcox as a scientific collaborator.

Ethical Standard

The authors assert that all procedures contributing to this work comply with the ethical standards of the
relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Disclaimer
The views, opinions and/or findings contained in this article are those of the author(s) and should not be construed as an official Department of Veterans Affairs position, policy, or decision unless so designated by other documentation, nor should they be construed to represent the views of any of the sponsoring organizations, agencies, or US Government.

Additional Information
A complete list of NCS and NCS-2 publications can be found at http://www.hcp.med.harvard.edu/ncs. A complete list of WMH publications can be found at http://www.hcp.med.harvard.edu/wmh

References


Anglemeyer A, Horvath HT, Bero L (2014). Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. The Cochrane Database of Systematic Reviews 4, MR000034.


Chang YJ, Chen LJ, Chung KP, Lai MS (2012). Risk groups defined by Recursive Partitioning Analysis of patients with colorectal adenocarcinoma treated with colorectal resection. BMC Medical Research Methodology 12, 2.

Chao ST, Koyfman SA, Woody N, Angelov L, Soeder SL, Reddy CA, Rybicki LA, Djeimil T, Sub JH (2012). Recursive partitioning analysis index is predictive for overall survival in patients undergoing spine stereotactic


with or without personality disorder. The British Journal of Psychiatry: The Journal of Mental Science 192, 124–129.


Prentice RL, Langer RD, Stefanick ML, Howard BV, Pettinger M, Anderson GL, Barad D, Curb JD, Kotchen J,


