Psychological Medicine (2012), **42**, 41–50. © Cambridge University Press 2011
doi:10.1017/S0033291711000997

ORIGINAL ARTICLE

Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model

R. H. Perlis[1,2]*, D. V. Iosifescu[1,3], V. M. Castro[4], S. N. Murphy[5], V. S. Gainer[4], J. Minnier[6], T. Cai[6], S. Goryachev[4], Q. Zeng[7], P. J. Gallagher[2], M. Fava[1], J. B. Weilburg[1], S. E. Churchill[8], I. S. Kohane[9] and J. W. Smoller[2]

[1] *Depression Clinic and Research Program, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA*
[2] *Psychiatric and Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA*
[3] *Mood and Anxiety Disorders Program, Department of Psychiatry, Mount Sinai Hospital, New York, NY, USA*
[4] *Partners Research Computing, Partners HealthCare System, Boston, MA, USA*
[5] *Department of Neurology, Massachusetts General Hospital, Boston, MA, USA*
[6] *Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA*
[7] *Department of Radiology, Brigham & Women's Hospital, Boston, MA, USA*
[8] *Information Systems, Partners HealthCare System, Boston, MA, USA*
[9] *Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA*

Background. Electronic medical records (EMR) provide a unique opportunity for efficient, large-scale clinical investigation in psychiatry. However, such studies will require development of tools to define treatment outcome.

Method. Natural language processing (NLP) was applied to classify notes from 127 504 patients with a billing diagnosis of major depressive disorder, drawn from out-patient psychiatry practices affiliated with multiple, large New England hospitals. Classifications were compared with results using billing data (ICD-9 codes) alone and to a clinical gold standard based on chart review by a panel of senior clinicians. These cross-sectional classifications were then used to define longitudinal treatment outcomes, which were compared with a clinician-rated gold standard.

Results. Models incorporating NLP were superior to those relying on billing data alone for classifying current mood state (area under receiver operating characteristic curve of 0.85–0.88 *v.* 0.54–0.55). When these cross-sectional visits were integrated to define longitudinal outcomes and incorporate treatment data, 15% of the cohort remitted with a single antidepressant treatment, while 13% were identified as failing to remit despite at least two antidepressant trials. Non-remitting patients were more likely to be non-Caucasian ($p < 0.001$).

Conclusions. The application of bioinformatics tools such as NLP should enable accurate and efficient determination of longitudinal outcomes, enabling existing EMR data to be applied to clinical research, including biomarker investigations. Continued development will be required to better address moderators of outcome such as adherence and co-morbidity.

Received 18 January 2011; Revised 10 May 2011; Accepted 14 May 2011; First published online 20 June 2011

Key words: Antidepressant, classification, machine learning, natural language processing, remission, treatment resistant depression.

Introduction

The analysis of electronic medical records (EMR) has been proposed as an efficient means of characterizing outcomes and rapidly identifying subpopulations within disorders in very large patient populations (Simon & Perlis, 2010). In addition to allowing collection of effectiveness outcomes or pharmacovigilance studies, such a tool could rapidly identify subgroups for biomarkers studies or participation in targeted clinical trials. This approach has advantages in ecological validity, as by definition it reflects clinical practice. It also offers far greater efficiency and feasibility than traditional clinical trials, as the data have already been collected and coded.

On the other hand, billing data typically offer little precision regarding diagnosis or outcome, particularly for psychiatric disorders. To overcome these limitations, computational methods have been developed to extract clinical data from narrative notes in the EMR.

* Address for correspondence: Dr R. H. Perlis, Simches Research Building, 185 Cambridge St, 6th Floor, Boston, MA 02114, USA
(Email: rperlis@partners.org)

Natural language processing (NLP) represents an automated method of chart review by processing text into meaningful concepts based on a set of rules. Outside of medicine, the recent success of a computer contestant on a television game show represents perhaps the most prominent recent example of a NLP application (Ferruci *et al*. 2010). NLP has been applied to a limited number of biomedical settings – mandatory reporting of notifiable diseases (Effler *et al*. 1999; Klompas *et al*. 2008; Lazarus *et al*. 2009), definition of co-morbid conditions (Meystre & Haug, 2006*a*, *b*; Solti *et al*. 2008) and medications (Turchin *et al*. 2006; Levin *et al*. 2007) and identification of adverse events (Bates *et al*. 2003; Penz *et al*. 2007).

To our knowledge, these approaches have received little attention in psychiatric disorders. In particular, given the complexity of phenotypic assessment in these illnesses, a crucial but unresolved question is how well outcomes may be defined based solely on EMR data. If such an approach could be validated, extremely efficient descriptive studies could be conducted and a means of facilitating future prospective studies established. To explore the potential utility of NLP, we examined outcomes of antidepressant treatment in major depressive disorder (MDD). Specifically, we attempted to develop, compare and validate alternative methods of characterizing two key outcomes in the treatment of MDD episodes, symptomatic remission (Rush *et al*. 2003*a*) and treatment resistance. Treatment resistant depression (TRD), typically defined as a failure to respond to at least two adequate trials of medication or other somatic therapies (Fava & Rush, 2006; Rush *et al*. 2006), contributes substantially to the disability and associated costs of MDD (Gibson *et al*. 2010) and may also be associated with elevated risk for suicide (Papakostas *et al*. 2003). The ability to identify individuals at greater risk for TRD might allow clinicians to risk stratify patients and treat or triage them more appropriately.

A particular challenge in defining outcomes is the need to integrate information across multiple visits – that is, evaluating a single assessment may be insufficient to establish an individual's treatment course for a disorder where symptoms may fluctuate over time. Therefore, we first compared ICD-9 codes to a gold standard based upon consensus for clinical status at each visit based upon review by a panel of experienced clinicians. Then, we developed a novel and broadly applicable tool using NLP to classify cross-sectional clinical status using narrative notes and compared it to the gold standard and to ICD-9 codes alone. Finally, we extended this cross-sectional data to define longitudinal outcomes and again validated these outcomes against those generated by consensus of clinical expert reviewers.

Method

Data source

The Partners HealthCare EMR incorporates sociodemographic data, billing codes, laboratory results, problem lists, medications, vital signs and narrative notes from Massachusetts General Hospital (MGH) and Brigham and Women's Hospital (BWH), as well as community and specialty hospitals that are part of the Partners HealthCare system in Boston (MA, USA). Altogether these records comprise about three million unique patients.

Patients with at least one diagnosis of MDD (ICD-9 296.2x, 296.3x) in the billing data or out-patient medical record at MGH or BWH were selected from the EMR for inclusion in a dataset (referred to as a data 'mart'). The data mart consists of all electronic records (psychiatric and non-psychiatric) from 127 504 patients using the i2b2 Workbench software (i2b2 v. 1.4; USA) (Murphy *et al*. 2007). The i2b2 system is a scalable computational framework for managing human health data and the Workbench facilitates analysis and visualization of such data. Billing data were available for all public and private payors. The Partners Institutional Review Board approved all aspects of this study and the usual safeguards for human subjects' data were applied, including data encryption and password protection and elimination of patient identifiers from derived datasets.

Identifying depressed and well patient visits

From the MDD data mart, 5198 patients with at least one billing code indicating a diagnosis of MDD and a psychiatric narrative note were selected for inclusion in the study (Fig. 1). Patients with billing codes for bipolar disorder, schizophrenia or dementia/delirium were excluded, as were those with other depressive disorders, such as dysthymia.

To determine the 'clinical gold standard' for patient status, a panel of three experienced board-certified clinical psychiatrists (J.W.S., D.V.I., R.H.P.) reviewed 724 randomly selected out-patient provider narratives and arrived at a consensus about the clinical status of the patient at the time of the visit. This status was assigned based upon the reported clinical status – that is, based upon clinician characterization of the patient's current mood state. Where this report was ambiguous or absent, DSM-IV mood state criteria were applied – that is, the clinical raters examined presence or absence of individual depression criteria and degree of severity, if present. Raters explicitly did not consider symptoms of co-morbidity such as anxiety or pain – thus, it was possible for subjects to be classified as remitted even with persistence of syndromal anxiety.

https://doi.org/10.1017/S0033291711000997 Published online by Cambridge University Press

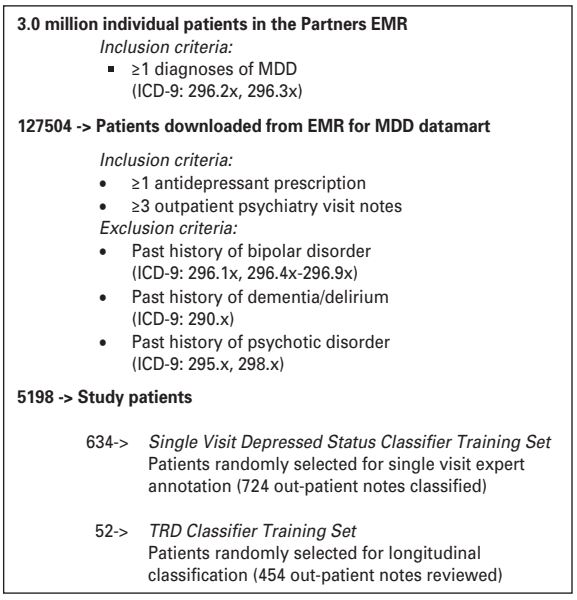


Fig. 1. Consort diagram. EMR, electronic medical records; MDD, major depressive disorder; TRD, treatment resistant depression.

Each note was classified as well (euthymic/remitted), defined as absence or virtual absence of depressive symptoms, depressed, defined as likely to meet criteria for a current major depressive episode or intermediate/subthreshold. The definitions for these states were drawn from prior task force reports on terminology (Frank *et al*. 1991; Rush *et al*. 2006). Raters voted individually but were required to achieve consensus for each note. The confidence level for each assignment was further rated as good confidence, fair confidence or low confidence, recognizing that the quality of notes precluded accurate characterization in some cases. During the classification process, the clinicians also identified words or phrases that were likely to be useful for classification. These terms were subsequently extracted from each narrative note with NLP using the HiTex platform (USA) (Zeng *et al*. 2006).The platform identifies terms using regular expressions (flexible matching) and applies negation and context algorithms to filter inappropriate matches. The presence or absence of a term then becomes a feature of each note, which can be utilized in classification algorithms. Fig. 1 provides a schematic of the study selection procedure for identifying and classifying patient groups.

Logistic regression model

We used the clinician-reviewed classifications to train models to predict the probability of being depressed or well (at the single visit level) based on a logistic regression classifier with adaptive least absolute shrinkage and selection operator (LASSO) procedure. We found that optimal fit was provided by fitting two separate models: depressed *versus* other; well *versus* other. The adaptive LASSO procedure simultaneously identifies important features and provides stable estimates of the model parameters (Zou, 2006). It is often applied in high-dimensionality datasets to select the more useful subset of features for modeling, because it will shrink the coefficients of other features (covariates) to zero (for a review and comparison to alternative approaches, see Bunea *et al*. (2011)). The optimal penalty parameter was determined based on Bayes' Information Criterion. ICD-9 depression billing codes include a digit intended to indicate current severity (e.g. 296.3x), but we anticipated that such digits might not be used consistently in claims data. We therefore developed and compared three different sets of models using: (i) billing codes only; (ii) narrative terms only (NLP); (iii) all available data (billing codes + NLP).

For each clinical state, we selected the threshold probability value for classifying patients as being in the state by setting the specificity level at 95%. Importantly, this rigorous threshold was selected to minimize the false positive rate, as might be optimal for biomarkers studies of extreme phenotypes, for example. Patients whose predicted probability exceeds the threshold value for either state were classified as being depressed or well, denoted by D+ and W+, respectively. We allowed a third state representing intermediate/subthreshold status, recognizing the prevalence of subthreshold depression in clinical practice, capturing those classified as neither D+ nor W+. The sensitivity, precision and area under the receiver operating characteristic curve (AUC) were estimated for D+ and W+ to compare prediction performance based on all three models, compared to the gold standard established by the clinician ratings. At this phase of investigation, data on treatment, if any, were not considered.

A small subset of clinicians in the Partners system routinely ask patients to complete a validated self-report measure of depression severity, the 16-item Quick Inventory of Depressive Symptomatology-Self-Rated (QIDS-SR; Trivedi *et al*. 2004), at every visit. This subset includes clinicians within a specialized major depression treatment program. The availability of these scores provided an opportunity to further examine the cross-sectional classifications in an exploratory fashion: the scores were extracted from the narrative notes and compared between visit classifications (well, depressed, intermediate). Because of their relative paucity in the dataset, these scores were not used directly to train the classification algorithms.

https://doi.org/10.1017/S0033291711000997 Published online by Cambridge University Press

TRD longitudinal classification

Using the single visit classifications, we developed a rule-based algorithm to classify patients as TRD (case) or treatment-responsive (control) based upon standard definitions of outcome (Rush *et al*. 2006) and treatment-resistance (Rush *et al*. 2003*a*). The algorithm was defined by a panel of experienced clinicians to maximize face validity within the limitations of a sparse database.

TRD was defined as meeting all of the following criteria: two or more D+ visits within a 12-month period following an initial antidepressant prescription; no visits classified as W+; a majority of all visits classified as D+; exposure to at least two antidepressants during this period. Patients with at least two consecutive D+ visits following an antidepressant prescription were classified as TRD. Treatment-responsive was defined as two or more W+ visits within a 12-month period following initial antidepressant prescription, no visits classified as D+ and exposure to only one antidepressant during this period. Observations preceding antidepressant prescription were not considered.

As the intention was to identify more extreme phenotypes for future study, patients who otherwise met criteria for responsiveness but received multiple types of different antidepressants during this period were excluded from the responsive group, since requiring multiple antidepressants would typically represent failure of monotherapy. Thus, the treatment-responsive group might be further characterized as 'single treatment responsive'.

In order to validate the rule, a board-certified psychiatrist (R.H.P.), blinded to the rule classifications, reviewed all of the notes for a random sample of 55 patients and assigned them a classification of either TRD or treatment-responsive using the same approach as in cross-sectional analysis and after reviewing standard outcome definitions noted above.

Finally, after deriving these longitudinal phenotypes, we compared patient demographics, visit frequency and medication prescriptions in the derived longitudinal TRD, treatment-responsive and intermediate/partially responsive groups. Co-morbid conditions were also assessed using the previously validated Age-adjusted Charlson Comorbidity Index (Charlson *et al*. 1987, 1994).

Statistical methods

To assess the overall concordance of the single visit algorithms with the training data and to estimate the threshold value for D+ and W+, we used three-fold cross-validation repeated 50 times to correct for potential over-fitting bias. Bootstrapping was used to assess the standard error and obtain confidence intervals (CI) for the accuracy estimates. TRD, responsive and intermediate/partially responsive group demographics were compared using analysis of variance and χ^2 test. Visit frequency, co-morbidity score and medication prescriptions were compared using the Kruskal–Wallis non-parametric test.

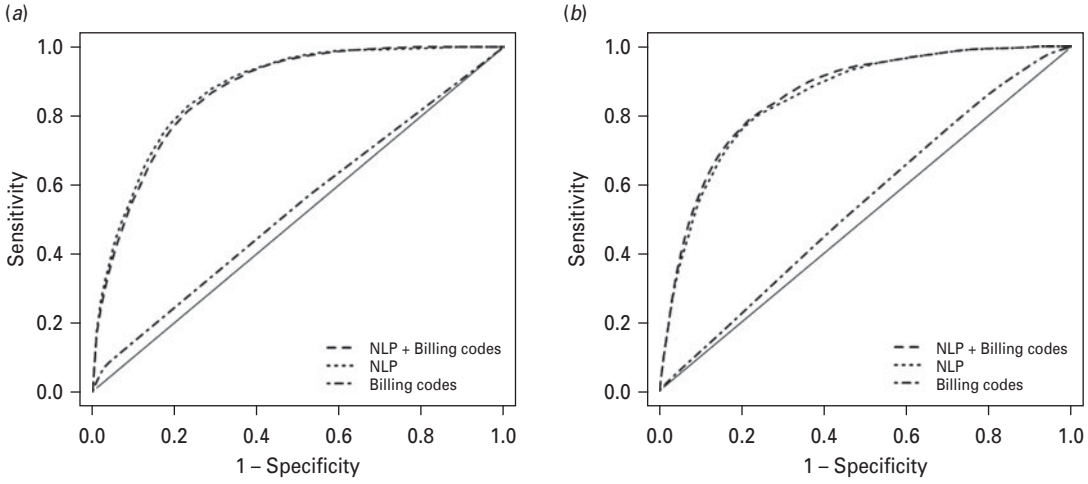
Results

After manual review and classification of 724 narrative notes, 34 NLP terms were identified from the clinician annotations as potentially useful for predicting cross-sectional (visitwise) clinical status. The adaptive LASSO procedure was then used to build three sets of models in a training dataset: one with billing data only (that is, the ICD-9 296.3x or 296.2x severity codes), one with NLP terms and one with both. The combination model selected 23 of the NLP terms and one billing code for the depressed classification and 15 NLP terms and three billing codes for the well classification. Supplementary Fig. S1 depicts the model selection for prediction of single-visit clinical status using NLP and billing codes, while the resulting models are shown in Supplementary Table S1. The initial terms positively associated in the depressed model included 'depressed' and 'mood anxious' and those positively associated in the well model were 'euthymic affect', 'stable' and 'much better'. Some unexpected terms, such as 'energy', also associated positively with depressed status, likely because clinicians described neurovegetative symptoms in both depressed and well patients. Two single visit classifiers were developed, to categorize visits as 'depressed', 'well' or 'intermediate'.

Receiver operating characteristic curves for these sets of models are shown in Fig. 2. Models incorporating NLP were markedly more accurate than those incorporating billing data alone: for prediction of 'depressed', AUC was 0.88 *v.* 0.54, while for prediction of 'well', AUC was 0.85 *v.* 0.55. Models with and without billing codes performed similarly. Fig. 2 also indicates sensitivity for each model, with specificity constrained to be 0.95. For the full (NLP + billing) models, when 'wellness' or 'depression' were classified with a 5% false positive rate, sensitivity was 0.39, i.e. 39% of 'well' or 'depressed' visits were identified.

To further characterize the performance of the mood state classifiers, we examined the classification of notes, which included a validated self-report measure of depressive symptom severity, the QIDS-SR (Trivedi *et al*. 2004). Such notes were available for only a subset of individuals from the full cohort (~20%). Where these measures were available for multiple visits for a single patient, one visit was randomly

https://doi.org/10.1017/S0033291711000997 Published online by Cambridge University Press



(c)

Clinical Status	Model	Specificity	Sensitivity	Precision	AUC
Depressed	Billing codes	0.95	0.09 (0.03)	0.57 (0.14)	0.54 (0.02)
Depressed	NLP	0.95	0.42 (0.05)	0.78 (0.02)	0.88 (0.02)
Depressed	NLP + Billing codes	0.95	0.39 (0.06)	0.78 (0.02)	0.87 (0.02)
Well	Billing codes	0.95	0.06 (0.02)	0.26 (0.27)	0.55 (0.03)
Well	NLP	0.95	0.37 (0.06)	0.86 (0.02)	0.85 (0.02)
Well	NLP + Billing codes	0.95	0.39 (0.07)	0.85 (0.02)	0.86 (0.02)

Fig. 2. Receiver operating characteristic curves for single visit models: (*a*) depressed; (*b*) well; (*c*) cross-validation results for single visit models, specificity fixed at 0.95 (with bootstrapped standard errors shown in parentheses). NLP, natural language processing; AUC, area under the receiver operating characteristic curve.

selected. For the 874 notes classified as depressed, mean QIDS-SR was 12.6 (95% CI 12.0–13.3); for the 479 notes classified as remitted, mean QIDS-SR was 7.9 (95% CI 7.2–8.6) and for the 1606 notes classified as intermediate, mean QIDS-SR was 13.3 (95% CI 12.8–13.7).

Single-visit classifications based on NLP + Billing codes models for both D+ and W+ were then used to construct longitudinal outcomes. Supplementary Fig. S2 shows an example of visualization of longitudinal course using the i2b2 Workbench (Murphy *et al*. 2007). In all, 840 of 5198 patients (16%) met criteria for a period of remission and 574 patients (11%) for TRD; the remaining 3784 (73%) were intermediate/partially-responsive or required multiple treatment trials. Concordance with the clinician gold standard was 0.764.

Table 1 provides group demographic, visit frequency and medication prescription comparisons of each of the study cohorts. Notably, the TRD group had significantly greater proportions of non-Caucasian patients ($p<0.001$) and patients covered by public insurance plans such as Medicare and Medicaid ($p<0.001$). As expected, TRD patients had a significantly greater frequency of depressed visits, number of antidepressant prescription refills and different types of antidepressants prescribed ($p<0.001$). There was no difference between age-adjusted comorbidity index scores between the TRD and responsive groups ($p=0.245$) but both groups had significantly higher scores than patients with partial response ($p<0.001$).

Discussion

Our results demonstrate the feasibility as well as the challenges of assessing clinical outcomes in EMR using NLP of clinicians' narrative notes. Using a simple set of empirically defined terms, which are readily extracted from free text, 23% of narrative notes could be accurately classified as depressed, 22% as euthymic and the remainder as intermediate or subthreshold. We emphasize that a large number of patients and notes remain in this third group by design: criteria were selected *a priori* to maximize specificity for the two outcome categories (TRD and single-treatment responder) anticipating their use in future biomarkers studies. Selection of more liberal thresholds would of course greatly increase the

https://doi.org/10.1017/S0033291711000997 Published online by Cambridge University Press

Table 1. *Demographics and clinical features of case/control groups*

	Treatment-resistant (TRD)	Treatment-responsive (Control)	Intermediate/partially-responsive	*p* value
n	574	840	3784	
Demographics				
Age, mean years ± S.D.	49.6 ± 14.0	51.2 ± 17.2	49.4 ± 15.5	0.004[a]
Gender, % Female	72	72	72	0.709[b]
Race/ethnicity[d], % of group				
White	65	76	75	<0.001[b]
African American	8	5	7	
Hispanic	24	14	14	
Asian	1	1	2	
Other	3	4	3	
Insurance, % of group				
Public	47	44	40	<0.001[b]
Private	45	50	52	
Other/Unknown	8	6	8	
Co-morbidities				
Age-adjusted Charlson co-morbidity index	5 (2–9)	5 (2–9)	4 (1–8)	<0.001[c]
Visits, median (IQR)				
All psych visits	14 (7–25)	10 (4–20)	4 (2–10)	<0.001[c]
Depressed visits	5 (3–10)	0 (0–1)	1 (0–2)	<0.001[c]
Well visits	0 (0–1)	4 (2–8)	0 (0–2)	<0.001[c]
Medications, median (IQR)				
Unique antidepressants	4 (3–6)	1 (1–1)	2 (1–4)	<0.001[c]
Number of refills per antidepressant	14 (7–25)	7 (3–16)	5 (2–11)	<0.001[c]

TRD, Treatment resistant depression; IQR, interquartile range.

[a] Analysis of variance.

[b] Pearson's χ^2.

[c] Kruskal–Wallis test.

[d] Race and ethnicity are collected using a single field in the electronic medical record, so subjects who identify as Hispanic are not further characterized.

proportion of subjects classified to the extreme groups, and might be desirable for other types of investigations such as effectiveness studies seeking to characterize TRD risk.

The intermediate group also reflects the limitations both of the diagnostic system and clinical documentation. That is, many patients will experience only partial improvement and this may not be well captured in the narrative text. Of note, even for those individuals classified as euthymic based on the narrative note, mean QIDS-SR is in the mildly depressed range. One contributor to this discordance might be the specific guidance given to the raters to not score anxiety or other symptoms, while patients might score anxiety symptoms as (for example) agitation or poor concentration – a challenge any time a self-report and clinician-rated assessment are compared. Given the relative paucity and lack of systematic administration of QIDS-SR, these exploratory analyses should be interpreted with caution. This finding underscores the prevalence of residual mood symptoms in clinical practice, as well as the potential utility of using self-report measures in this context (Nierenberg *et al*. 2010).

The superiority of using clinician- or even patient-reported measures to determine symptom severity should be apparent, which might lead one to question the utility of NLP-based approaches. Indeed, these results should highlight the limitations of the narrative text as well as the potential utility of standardized assessments (and their inclusion in EMR systems). On the other hand, progress toward this goal has been remarkably slow even in academic mental health systems and, once implemented, it will be many years until large datasets with these measures accumulate.

https://doi.org/10.1017/S0033291711000997 Published online by Cambridge University Press

During this transition, the value of using existing large datasets, with millions of patients and years of data collection, should also be clear.

Our report is one of the first to examine large-scale use of NLP approaches for classification in psychiatry, although this application was suggested two decades ago (for a review, see Garfield *et al*. 1992). One previous study described a pilot effort to classify suicide notes according to intention (Pestian *et al*. 2008). Outside of psychiatry, modern NLP techniques have demonstrated success in such areas as detecting disease requiring notification of public health officials (Effler *et al*. 1999; Klompas *et al*. 2008; Lazarus *et al*. 2009) and identifying unexpected adverse events (Bates *et al*. 2003; Penz *et al*. 2007), as well as determining co-morbid medical conditions (Meystre & Haug, 2006*a*,*b*; Solti *et al*. 2008) and medications (Turchin *et al*. 2006; Levin *et al*. 2007). With growing interest in the use of large clinical databases for conducting effectiveness research, the development of the toolset necessary to define outcomes in psychiatry may be critical.

Our findings strongly suggest that billing data alone, including ICD-9 codes used for billing, is unlikely to be adequate for establishing outcomes. This likely reflects clinicians' lack of concern for accuracy in such codes, as they do not impact reimbursement and are often used primarily to reflect the diagnosis of the patient and not current clinical status. Indeed, prior reports suggest that such codes may not reliably distinguish individuals by diagnosis, as was illustrated in a cohort of mood disorder patients undergoing electroconvulsive therapy (Jakobsen *et al*. 2008).

We note several *caveats* in interpreting our work. First, the portability of these classification models remains to be determined. Different healthcare systems may have different standards or formats for narrative notes, which would be expected to influence the performance of our classifiers. However, we emphasize that MGH and BWH, the two major hospitals within the Partners Health Care system, include two distinct departments of psychiatry with different medical record systems and approaches to documentation, which should improve portability to other systems. The vast majority of clinical notes derive not from the in-patient units, but from affiliated out-patient clinics in the region, most of which are not primarily academic in orientation.

Second, as we have noted, these classifiers should not be construed as a substitute for systematic and quantitative assessment. Manual review of notes identified a remarkable disparity in quality and nature of documentation and consequent ambiguity in description of clinical states. For example, a common notation was 'depression is stable', which might refer to a patient who continues to be depressed (as in unchanged), or one whose illness is successfully managed (as in remaining in remission). Likewise, it was not uncommon to encounter documentation of details of recent stressors or events, in the absence of mood symptoms. As more health care systems move to EMR, there is a unique opportunity to better quantify outcomes. For example, the 16-item patient-rated QIDS-SR has been shown to be highly correlated with clinician-rated measures and sensitive to treatment effects (Rush *et al*. 2003*b*); another well-validated alternative is the PHQ-9 (Kroenke *et al*. 2001). Their incorporation in EMR systems would greatly improve their capacity to support future outcome studies. At minimum, EMR systems that utilize templates could require clinicians to record a clinical status [for example, using the 7-point Clinical Global Impression scale (Guy, 1976), or even recording remission status].

Third, in defining longitudinal outcomes, multiple assumptions are required about treatment status. As the Partners HealthCare system is not a 'closed' one, there is documentation of a prescription being given but not of it being filled or re-filled. Therefore, there is some risk for misclassification in both directions. Individuals labeled 'responsive' may have remitted in spite of not adhering to treatment, as might be expected given the sizeable rates of placebo response in MDD (Fournier *et al*. 2010). Conversely, individuals labeled as having TRD may actually be non-adherent, or partially adherent, or receive inadequate medication dosage or duration, a phenomenon sometimes referred to as 'pseudoresistance'. This limitation underscores the value of integrating clinical data with pharmacy billing data whenever possible. A related challenge is determining tolerability; some individuals classified as resistant may actually be intolerant to multiple medications and thus unable to achieve therapeutic doses necessary for symptomatic improvement. Whether tolerability can itself be accurately determined with NLP approaches merits further investigation. Incorporating tolerability data is further complicated by its partial correlation with efficacy: individuals may be more likely to tolerate medications that they perceive as being helpful to them, and vice versa. In addition to adherence and tolerability, psychiatric and medical co-morbidity are also important moderators of treatment response to which NLP approaches may be applicable.

It should be emphasized that TRD was selected for this study precisely because it is a difficult problem for NLP. Many outcomes within psychiatry should be substantially easier to define, particularly those such as hospitalization, which are likely to be available from billing data. Given the chronicity of many psychiatric disorders, however, the ability to

https://doi.org/10.1017/S0033291711000997 Published online by Cambridge University Press

parse less 'hard' outcomes such as remission among out-patients will clearly be important in facilitating future studies.

Classification based upon narrative notes provides an opportunity to take advantage of existing EMR systems for highly efficient clinical investigation. In the Partners HealthCare system, there are ~4 years of psychiatry out-patient notes, which, even in the absence of detailed rating scales, yield some perspective on clinical outcomes on a very large scale. With appropriate protection of patients' privacy, this resource could be applied to efficiently identify risk factors for treatment resistance. It can facilitate investigations of effectiveness, for example, by comparing outcomes across different clinics or payor types within a health care system to highlight potential disparities. (We note the importance of considering confounding in these sorts of population-based investigations, and also the well-established methodologies for addressing these concerns.) Finally, it might allow for efficient recruitment of specific clinical populations; for example, investigations of novel interventions specifically for patients with TRD, or pharmacogenomic investigations of TRD. By comparison, in the largest TRD study to date, >4000 patients were enrolled in order to yield fewer than 100 patients per arm in the most treatment-resistant phase (Trivedi *et al*. 2006). If personalized medicine is to become a reality in psychiatry, multiple large datasets will be required to build and validate models for treatment outcome. Our results suggest that applying NLP tools to existing EMR data may help accelerate this process.

Note

Supplementary material accompanies this paper on the Journal's website (http://journals.cambridge.org/psm).

Acknowledgements

The project described was supported by Award # U54LM008748 from the National Library of Medicine (to ISK) and R01MH086026 and R01MH085542 from the National Institute of Mental Health (to R.H.P. and J.W.S., respectively). The content is solely the responsibility of the authors and does not necessarily represent the official view of the National Library of Medicine or the National Institutes of Health.

Declaration of Interest

Roy Perlis has received consulting fees from Proteus Biomedical, Concordant Rater Systems, Genomind, and RID Ventures.

Dan Iosifescu has received grant support from Aspect Medical Systems, Forest Laboratories, Janssen Pharmaceuticals, NARSAD, and NIH. He has received speaker honoraria from Eli & Co., Pfizer, Inc., Forest Laboratories, and Reed Medical Education.

Maurizio Fava – Lifetime Disclosures

Research Support: Abbott Laboratories; Alkermes, Inc.; Aspect Medical Systems; AstraZeneca; BioResearch; BrainCells Inc.; Bristol-Myers Squibb; Cephalon, Inc.; Clinical Trials Solutions, LLC; Covidien; Eli Lilly and Company; EnVivo Pharmaceuticals, Inc.; Forest Pharmaceuticals, Inc.; Ganeden Biotech, Inc.; GlaxoSmithKline; Johnson & Johnson Pharmaceutical Research & Development; Lichtwer Pharma GmbH; Lorex Pharmaceuticals; Novartis AG; Organon Pharmaceuticals; PamLab, LLC.; Pfizer Inc.; Pharmavite® LLC; Roche; RTC Logic, LLC; Sanofi-Aventis US LLC; Shire; Solvay Pharmaceuticals, Inc.; Synthelabo; Wyeth-Ayerst Laboratories.

Advisory/Consulting: Abbott Laboratories; Affectis Pharmaceuticals AG; Amarin Pharma Inc.; Aspect Medical Systems; AstraZeneca; Auspex Pharmaceuticals; Bayer AG; Best Practice Project Management, Inc.; BioMarin Pharmaceuticals, Inc.; Biovail Corporation; BrainCells Inc; Bristol-Myers Squibb; CeNeRx BioPharma; Cephalon, Inc.; Clinical Trials Solutions, LLC; CNS Response, Inc.; Compellis Pharmaceuticals; Cypress Pharmaceutical, Inc.; Dov Pharmaceuticals, Inc.; Eisai Inc.; Eli Lilly and Company; EPIX Pharmaceuticals, Inc.; Euthymics Bioscience, Inc.; Fabre-Kramer Pharmaceuticals, Inc.; Forest Pharmaceuticals, Inc.; GenOmind, LLC; GlaxoSmithKline; Gruenthal GmbH; Janssen Pharmaceutica; Jazz Pharmaceuticals, Inc.; Johnson & Johnson Pharmaceutical Research & Development, LLC.; Knoll Pharmaceuticals Corp.; Labopharm Inc.; Lorex Pharmaceuticals; Lundbeck Inc.; MedAvante, Inc.; Merck & Co., Inc.; Methylation Sciences; Neuronetics, Inc.; Novartis AG; Nutrition 21; Organon Pharmaceuticals; PamLab, LLC.; Pfizer Inc.; PharmaStar; Pharmavite® LLC.; Precision Human Biolaboratory; Prexa Pharmaceuticals, Inc.; PsychoGenics; Psylin Neurosciences, Inc.; Ridge Diagnostics, Inc.; Roche; RCT Logic, LLC; Sanofi-Aventis US LLC.; Sepracor Inc.; Schering-Plough Corporation; Solvay Pharmaceuticals, Inc.; Somaxon Pharmaceuticals, Inc.; Somerset Pharmaceuticals, Inc.; Synthelabo; Takeda Pharmaceutical Company Limited; Tetragenex Pharmaceuticals, Inc.; TransForm Pharmaceuticals, Inc.; Transcept Pharmaceuticals, Inc.; Vanda Pharmaceuticals, Inc.; Wyeth-Ayerst Laboratories.

Speaking/Publishing: Adamed, Co.; Advanced Meeting Partners; American Psychiatric Association; American Society of Clinical Psychopharmacology; AstraZeneca; Belvoir Media Group; Boehringer

Ingelheim GmbH; Bristol-Myers Squibb; Cephalon, Inc.; Eli Lilly and Company; Forest Pharmaceuticals, Inc.; GlaxoSmithKline; Imedex, LLC; MGH Psychiatry Academy/Primedia; MGH Psychiatry Academy/Reed Elsevier; Novartis AG; Organon Pharmaceuticals; Pfizer Inc.; PharmaStar; United BioSource, Corp.; Wyeth-Ayerst Laboratories.

Equity Holdings: Compellis.

Royalty/patent, other income: Patent for SPCD and patent application for a combination of azapirones and bupropion in MDD, copyright royalties for the MGH CPFQ, SFI, ATRQ, DESS, and SAFER.

References

Bates DW, Evans RS, Murff H, Stetson PD, Pizziferri L, Hripcsak G (2003). Detecting adverse events using information technology. *Journal of the American Medical Informatics Association* **10**, 115–128.

Bunea F, She Y, Ombao H, Gongvatana A, Devlin K, Cohen R (2011). Penalized least squares regression methods and applications to neuroimaging. *Neuroimage* **55**, 1519–1527.

Charlson M, Szatrowski TP, Peterson J, Gold J (1994). Validation of a combined comorbidity index. *Journal of Clinical Epidemiology* **47**, 1245–1251.

Charlson ME, Pompei P, Ales KL, MacKenzie CR (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of Chronic Diseases* **40**, 373–383.

Effler P, Ching-Lee M, Bogard A, Ieong MC, Nekomoto T, Jernigan D (1999). Statewide system of electronic notifiable disease reporting from clinical laboratories: comparing automated reporting with conventional methods. *Journal of the American Medical Association* **282**, 1845–1850.

Fava M, Rush AJ (2006). Current status of augmentation and combination treatments for major depressive disorder: a literature review and a proposal for a novel approach to improve practice. *Psychotherapy and Psychosomatics* **75**, 139–153.

Ferruci D, Brown E, Chu-Carroll J, Fan J, Gondek D, Kalyanpur A, Lally A, Murdock JW, Nyberg, E, Prager J, Schlaefer N, Welty C (2010). Building Watson: an overview of the DeepQA project. *Artificial Intelligence* 59–79.

Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC, Fawcett J (2010). Antidepressant drug effects and depression severity: a patient-level meta-analysis. *Journal of the American Medical Association* **303**, 47–53.

Frank E, Prien RF, Jarrett RB, Keller MB, Kupfer DJ, Lavori PW, Rush AJ, Weissman MM (1991). Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. *Archives of General Psychiatry* **48**, 851–855.

Garfield DA, Rapp C, Evens M (1992). Natural language processing in psychiatry. Artificial intelligence technology and psychopathology. *Journal of Nervous and Mental Disease* **180**, 227–237.

Gibson TB, Jing Y, Smith Carls G, Kim E, Bagalman JE, Burton WN, Tran QV, Pikalov A, Goetzel RZ (2010). Cost burden of treatment resistance in patients with depression. *American Journal of Managed Care* **16**, 370–377.

Guy W (1976). *ECDEU Assessment Manual for Psychopharmacology: US Dept Health Education and Welfare publication* (*ADM*), 76–338, pp. 218–222. National Institute of Mental Health: Rockville, MD.

Jakobsen K, Hansen T, Dam H, Larsen E, Gether U, Werge T (2008). Reliability of clinical ICD-10 diagnoses among electroconvulsive therapy patients with chronic affective disorders. *European Journal of Psychiatry* **22**, 167–172.

Klompas M, Haney G, Church D, Lazarus R, Hou X, Platt R (2008). Automated identification of acute hepatitis B using electronic medical record data to facilitate public health surveillance. *PLoS One* **3**, e2626.

Kroenke K, Spitzer RL, Williams JB (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine* **16**, 606–613.

Lazarus R, Klompas M, Campion FX, McNabb SJ, Hou X, Daniel J, Haney G, DeMaria A, Lenert L, Platt R (2009). Electronic support for public health: validated case finding and reporting for notifiable diseases using electronic medical data. *Journal of the American Medical Informatics Association* **16**, 18–24.

Levin MA, Krol M, Doshi AM, Reich DL (2007). Extraction and mapping of drug names from free text to a standardized nomenclature. *AMIA Annual Symposium Proceedings*, pp. 438–442.

Meystre S, Haug P (2006*a*). Improving the sensitivity of the problem list in an intensive care unit by using natural language processing. *AMIA Annual Symposium Proceedings*, pp. 554–558.

Meystre S, Haug PJ (2006*b*). Natural language processing to extract medical problems from electronic clinical documents: performance evaluation. *Journal of Biomedical Informatics* **39**, 589–599.

Murphy SN, Mendis ME, Hackett K, Kuttan R, Pan W, Phillips L, Gainer VS, Berkowicz D, Glaser J, Kohane IS, Chueh H (2007). Architecture of the open-source clinical research chart from informatics for integrating biology and the bedside. *AMIA Annual Symposium Proceedings*, pp. 548–552.

Nierenberg AA, Husain MM, Trivedi MH, Fava M, Warden D, Wisniewski SR, Miyahara S, Rush AJ (2010). Residual symptoms after remission of major depressive disorder with citalopram and risk of relapse: a STAR*D report. *Psychological Medicine* **40**, 41–50.

Papakostas GI, Petersen T, Pava J, Masson E, Worthington JJ, 3rd, Alpert JE, Fava M, Nierenberg AA (2003). Hopelessness and suicidal ideation in outpatients with treatment-resistant depression: prevalence and impact on treatment outcome. *Journal of Nervous and Mental Disease* **191**, 444–449.

Penz JF, Wilcox AB, Hurdle JF (2007). Automated identification of adverse events related to central venous catheters. *Journal of Biomedical Informatics* **40**, 174–182.

Pestian JP, Matykiewicz P, Grupp-Phelan J, Lavanier SA, Combs J, Kowatch R (2008). Using natural language

processing to classify suicide notes. *Annual Symposium Proceedings* of the American Medical Informatics Association, 6 November 2008. Abstract 1091.

Rush AJ, Kraemer HC, Sackeim HA, Fava M, Trivedi MH, Frank E, Ninan PT, Thase ME, Gelenberg AJ, Kupfer DJ, Regier DA, Rosenbaum JF, Ray O, Schatzberg AF (2006). Report by the ACNP Task Force on response and remission in major depressive disorder. *Neuropsychopharmacology* **31**, 1841–1853.

Rush AJ, Thase ME, Dube S (2003*a*). Research issues in the study of difficult-to-treat depression. *Biological Psychiatry* **53**, 743–753.

Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC, Ninan PT, Kornstein S, Manber R, Thase ME, Kocsis JH, Keller MB (2003*b*). The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biological Psychiatry* **54**, 573–583.

Simon GE, Perlis RH (2010). Personalized medicine for depression: can we match patients with treatments? *American Journal of Psychiatry* **167**, 1445–1455.

Solti I, Aaronson B, Fletcher G, Solti M, Gennari JH, Cooper M, Payne T (2008). Building an automated problem list based on natural language processing: lessons learned in the early phase of development. *AMIA Annual Symposium Proceedings* 687–691.

Trivedi MH, Fava M, Wisniewski SR, Thase ME, Quitkin F, Warden D, Ritz L, Nierenberg AA, Lebowitz BD, Biggs MM, Luther JF, Shores-Wilson K, Rush AJ (2006). Medication augmentation after the failure of SSRIs for depression. *New England Journal of Medicine* **354**, 1243–1252.

Trivedi MH, Rush AJ, Ibrahim HM, Carmody TJ, Biggs MM, Suppes T, Crismon ML, Shores-Wilson K, Toprac MG, Dennehy EB, Witte B, Kashner TM (2004). The Inventory of Depressive Symptomatology, Clinician Rating (IDS-C) and Self-Report (IDS-SR), and the Quick Inventory of Depressive Symptomatology, Clinician Rating (QIDS-C) and Self-Report (QIDS-SR) in public sector patients with mood disorders: a psychometric evaluation. *Psychological Medicine* **34**, 73–82.

Turchin A, Morin L, Semere LG, Kashyap V, Palchuk MB, Shubina M, Chang F, Li Q (2006). Comparative evaluation of accuracy of extraction of medication information from narrative physician notes by commercial and academic natural language processing software packages. *AMIA Annual Symposium Proceedings* 789–793.

Zeng QT, Goryachev S, Weiss S, Sordo M, Murphy SN, Lazarus R (2006). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Medical Informatics and Decision Making* **6**, 30.

Zou H (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.