OPEN

Citation: Transl Psychiatry (2016) **6**, e921; doi:10.1038/tp.2015.182

www.nature.com/tp

ORIGINAL ARTICLE

Predicting early psychiatric readmission with natural language processing of narrative discharge summaries

A Rumshisky[1,2], M Ghassemi[1], T Naumann[1], P Szolovits[1], VM Castro[3,4,5,6], TH McCoy[3,4,5] and RH Perlis[3,4,5]

The ability to predict psychiatric readmission would facilitate the development of interventions to reduce this risk, a major driver of psychiatric health-care costs. The symptoms or characteristics of illness course necessary to develop reliable predictors are not available in coded billing data, but may be present in narrative electronic health record (EHR) discharge summaries. We identified a cohort of individuals admitted to a psychiatric inpatient unit between 1994 and 2012 with a principal diagnosis of major depressive disorder, and extracted inpatient psychiatric discharge narrative notes. Using these data, we trained a 75-topic Latent Dirichlet Allocation (LDA) model, a form of natural language processing, which identifies groups of words associated with topics discussed in a document collection. The cohort was randomly split to derive a training (70%) and testing (30%) data set, and we trained separate support vector machine models for baseline clinical features alone, baseline features plus common individual words and the above plus topics identified from the 75-topic LDA model. Of 4687 patients with inpatient discharge summaries, 470 were readmitted within 30 days. The 75-topic LDA model included topics linked to psychiatric symptoms (suicide, severe depression, anxiety, trauma, eating/weight and panic) and major depressive disorder comorbidities (infection, postpartum, brain tumor, diarrhea and pulmonary disease). By including LDA topics, prediction of readmission, as measured by area under receiver-operating characteristic curves in the testing data set, was improved from baseline (area under the curve 0.618) to baseline+1000 words (0.682) to baseline+75 topics (0.784). Inclusion of topics derived from narrative notes allows more accurate discrimination of individuals at high risk for psychiatric readmission in this cohort. Topic modeling and related approaches offer the potential to improve prediction using EHRs, if generalizability can be established in other clinical cohorts.

Translational Psychiatry (2016) **6,** e921; doi:10.1038/tp.2015.182; published online 18 October 2016

INTRODUCTION

Early hospital readmission has been identified as a preventable driver of health-care costs and a key quality metric for Medicare.[1] Among psychiatric patients, with disorders often characterized by chronicity and high rates of recurrence, readmission is a substantial concern; the post-hospitalization period is recognized as a high-risk period for outcomes such as suicide and relapse into substance use.

A challenge in reducing readmission risk is the difficulty in identifying individuals at greatest risk, who might benefit from personalized interventions.[2] Few studies in psychiatry have attempted to develop clinically actionable prediction rules,[3,4] despite the widespread use of prediction methodologies in other areas of medicine. A particular challenge in developing prediction models in psychiatry is the paucity of data regarding symptoms or severity provided by diagnostic codes available in electronic health records (EHRs) or claims data sets. We have recently shown, for example, that ICD9 severity codes in major depressive disorder are no more reliable than chance in characterizing actual severity.[5]

The narrative notes in EHRs often include substantial clinical detail, including details of presentation and prior course, which may otherwise be unavailable. In general, efforts to parse narrative notes in EHR data focus on an individual diagnosis or symptom. Typical strategies involve using information extraction systems to identify the terms manually preselected by clinicians, sometimes supplemented by additional vocabulary based on these terms.[5–9]

Approaches in which each term must be selected by experts do not scale well if the goal is to characterize populations in terms of overall psychopathology, rather than aspects of a single diagnosis or outcome. As an alternative, more robust and generalizable term-agnostic approaches to automatic clinical concept extraction have also met with some success in other medical domains.[10–12] However, the resolution of these methods for psychopathology *per se* appears to be limited, as they still rely on standard diagnostic categories.

Another approach to deriving information from narrative notes relies on a form of natural language processing known as topic modeling. This strategy relies on word co-occurrence patterns in order to learn the latent set of topics discussed in the text. We have previously demonstrated that this approach can extract meaningful concepts from a set of intensive care unit notes, which were then applied to generate highly accurate classifiers to predict mortality;[13,14] another recent report utilized an alternative approach to discriminate veterans at risk for suicide based on narrative notes.[15] Here we hypothesized that topic modeling could also be applied to identify topics from psychiatric discharge notes, and that these topics will usefully improve prediction of hospital readmission compared with diagnosis or single terms extracted from the note text.

[1]MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA; [2]Department of Computer Science, University of Massachusetts Lowell, Lowell, MA, USA; [3]Center for Experimental Drugs and Diagnostics, Massachusetts General Hospital, Boston, MA, USA; [4]Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA; [5]Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA, USA and [6]Partners Research Information Systems and Computing, Partners HealthCare System, Boston, MA, USA. Correspondence: Dr R Perlis, Department of Psychiatry, Massachusetts General Hospital, Simches Research Building, 185 Cambridge Street, 6th Floor, Boston, MA 02114, USA.
E-mail: rperlis@partners.org
Received 1 May 2015; revised 14 August 2015; accepted 6 September 2015

MATERIALS AND METHODS

Cohort derivation

The Partners HealthCare EHR includes sociodemographic data, billing codes, laboratory results, problem lists, medications, vital signs, procedure reports and discharge notes from the Massachusetts General Hospital and the Brigham and Women's Hospital, as well as from community and specialty hospitals. We used The Informatics for Integrating Biology and the Bedside (i2b2) software (https://www.i2b2.org; i2b2 v1.6, Boston, MA, USA)[16] to access and manipulate this EHR data. For this study, we selected all patients admitted to an inpatient unit of the Massachusetts General Hospital with a principal diagnosis of major depressive disorder (MDD; ICD9 296.2× and 296.3×) between January 1994 and December 2012. Only individuals age 18 or older were included; no other exclusion criteria were applied except for data-cleaning considerations. The Partners Institutional Review Board approved all aspects of this study.

Outcome definition

Primary analysis examined 30- day psychiatric readmission, corresponding to a standard health system quality metric.

Word extraction and topic model derivation

Text was extracted from all clinical notes by removing all non-letter characters (for example, digits and punctuation), and then using white-space to split text into words. Stop words were removed using the Python sci-kit learn[17] software, which incorporates the Glasgow stop word list (http://ir.dcs.gla.ac.uk/resources/).

We used the term frequency–inverse document frequency (TF–IDF) scores[18] to identify the 1000 most informative words in each patient's notes, and we then limited our overall vocabulary to this set. Note that as 1000 most informative words were selected for each patient separately and some of these words were the same for different patients, the total number of words used in topic modeling was 66 429. As notes were not deidentified, this helped to remove the words that were unique to a single patient, such as surnames. Patients were excluded if their notes had fewer than 100 nonstop words in total or if their admission and death dates were not consistent (for example, if they were marked as being 'readmitted' after their date of death).

A Latent Dirichlet Allocation[19] (LDA) model was trained on the full corpus, and then topics were generated for each individual note. The LDA model produces a distribution over words for each topic, which can be used to score each note for membership in different topics. Our initial experiments using a training data set derived from 70% of the clinical cohort found no statistically significant difference in held-out prediction accuracy across a range of 25–75 topics. Below, we show the results for 75 topics (Supplementary Table 2).

Following the suggested strategy,[20] we set hyperparameters on the Dirichlet priors for the topic distributions $\alpha = 50/k$ and the topic-word distributions $\beta = 200/\text{number\_of\_words}$, where k is the number of topics. Topic distributions were sampled using a Markov Chain Monte Carlo chain with 2500 iterations. This topic-modeling step resulted in a k-dimensional vector of topic proportions for each note.

Supervised learning and model evaluation

The clinical cohort was randomly split 70%/30% into training and test data, maintaining the proportion of positive class (subjects readmitted for MDD) in each. Test and training data were also balanced based on age, gender, use of public insurance and age-adjusted Charlson comorbidity index.[21,22]

We trained a linear two-class support vector machine (SVM) model to predict MDD-related readmissions within 30 days. A separate SVM model was trained for each of the following feature configurations: (1) baseline clinical features, including age, gender, use of public health insurance and Charlson comorbidity index;[21,22] (2) baseline features+top-N bag-of-words features using N most informative words from each patient's record, with N ranging between 1 and 1000; and (3) baseline+topic features derived from the k-topic LDA model, with $k = 75$.

The most informative words for the second configuration type (baseline +top-N words) were selected separately for each patient, that is, the bag-of-words features included the union of the top-N words with highest TF–IDF scores from each patient record. As a result, the top-1 word configuration used 3013 unique words, top-10 configuration used 18 173 unique words and top-100 word configuration used 63 438 unique words. Top-1000 configuration used nearly all the words included in the vocabulary (66 429/66 451).

Features were all scaled to a range of 0–1 before training. The loss parameters for the SVM were selected using threefold cross-validation on the training data to determine the optimal values with area under the curve as an objective. The learned parameters were then used to construct a model for the entire training set and to make predictions on the test data. The same set of test data was consistently held out, and never used for model selection or parameter tuning. We experimented with randomly sub-sampling the negative class in the training set to produce a minimum 70%/30% ratio between the negative and positive classes, but this failed to affect the results and is not addressed further.

Test set distributions were never modified to reflect the reality of class imbalance during prediction, and reported performance reflects those distributions. Topic features were derived from the LDA model trained on the complete data set including both training and test subsets. As deriving topics does not incorporate any knowledge of future readmission, the inclusion of the testing set does not lead to overfitting or inflated estimates of discrimination.

To illustrate discrimination of the SVM models, we generated two separate unstratified Cox regression models for two feature sets: the baseline features; and the baseline features plus 75 topics. Both models used the time to psychiatric readmission as output.

After models were built for each feature set, the corresponding Kaplan–Meier curves for those models were plotted. These are shown in Figures 1 and 2. In both figures, a single model was built for the features chosen. We then used the output from the linear SVM classifier to generate a 'group' prediction. Note that the Cox regression models themselves ignore the group assignment variable; the variable was only used to select different subsets of the patient population.

RESULTS

We identified 4687 inpatient discharge summaries for unique patients admitted with the primary diagnosis of MDD in the billing data, with 470 of those patients readmitted within 30 days with a psychiatric diagnosis, 2977 readmitted with a nonpsychiatric diagnosis and 1240 not readmitted. Clinical features of the full cohort are summarized in Table 1. Supplementary Table 1 shows univariate associations with readmission based on logistic regression models applied to the training data set.

The topics derived by a 75-topic LDA model trained on the full data set were annotated by one of the authors (RHP) based on manual inspection of the individual terms loading on that topic. Table 2 summarizes the top 10 topics identified and the clinician

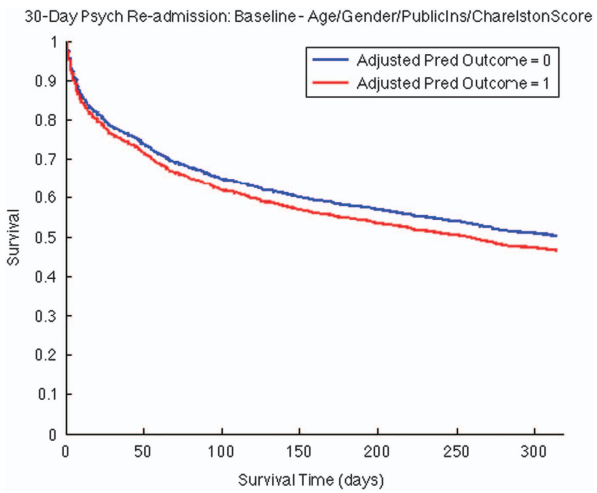

Figure 1. Kaplan–Meier survival curve for time to psychiatric hospital readmission, for a model built using baseline sociodemographic and clinical variables only. Patients are plotted separately for two groups identified by the support vector machine model as: (1) likely psychiatric readmissions in red; and (2) unlikely psychiatric readmissions in blue.

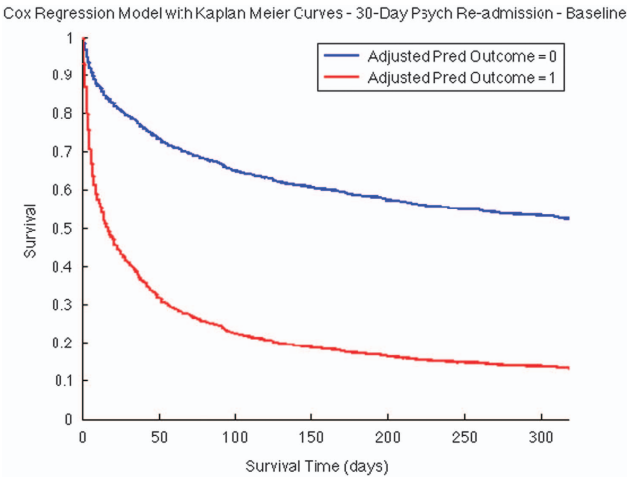

Figure 2. Kaplan–Meier survival curve for time to psychiatric hospital readmission, for a model built using the baseline variables and 75 topics. Patients are plotted separately for two groups identified by the support vector machine model as: (1) likely psychiatric readmissions in red; and (2) unlikely psychiatric readmissions in blue.

Table 1. Socioeconomic and clinical features of the psychiatric hospital readmission cohort

Features	*MDD admission cohort (N = 4687)*
Age at admission, mean (s.d.)	49.5 (17.6)
Gender, % female	64.4
Race, % Caucasian	76.1
Insurance, % public	61.9
Year of discharge, median (IQR)	2005 (2000–2009)
Charlson Index, median (IQR)	3 (0–6)
30-day all-cause readmission, %	22.0

Abbreviations: IQR, interquartile range; MDD, major depressive disorder.

annotation; words particularly informative for annotation are italicized. The identified topics included many linked to psychiatric symptoms (suicide, severe depression, anxiety, trauma, eating/weight and panic) and MDD comorbidities (infection, postpartum, brain tumor, diarrhea and pulmonary disease).

For each of the configurations described above, we saw improving performance in the testing data set as measured by the receiver-operating characteristic curves, from area under the curve 0.618 for baseline to 0.682 for baseline+1000 words to 0.784 for baseline+75 topics, respectively (Table 3). Note that merely discarding less-informative words does not improve performance, as shown by very similar performance of baseline+top-100 and baseline+top-1000 words configurations. Figures 1 and 2 show Kaplan–Meier survival curves for the baseline model alone, and the full model incorporating baseline+75 topics. Table 3 also presents sensitivity and specificity in the testing data set at a default cut-point; of note, the baseline model is highly sensitive but nonspecific, whereas the full model is somewhat less sensitive with substantially increased specificity.

DISCUSSION

In this investigation of 4687 individuals admitted to a psychiatric inpatient unit with a principal diagnosis of MDD, we developed statistical models to predict psychiatric readmission within 30 days.

Table 3. Comparison of models with and without inclusion of LDA topics

Configuration	*AUC*	*Sensitivity*	*Specificity*
Baseline = age/gender/insurance/ Charlson	0.618	0.979	0.104
Baseline+top-1 words	0.654	—	—
Baseline+top-10 words	0.676	—	—
Baseline+top-100 words	0.682	—	—
Baseline+top-1000 words	0.682	0.213	0.945
Baseline+75 topics (no words)	0.784	0.752	0.634

Abbreviations: AUC, area under the curve; LDA, Latent Dirichlet Allocation.

These predictions substantially exceed those expected by chance and may provide an initial means of quantifying risk. More generally, we demonstrate the utility of applying topic modeling to psychiatric narrative notes to improve risk prediction accuracy.

Notably, the specific topics identified represent both psychiatric and clinical illness features not otherwise captured in coded data, lending this approach some face validity. Greater psychiatric comorbidity such as substance use or eating disorders may increase readmission risk, along with general medical illness such as infection or dementia. Not surprisingly, markers of greater depression severity, previously shown to be poorly captured in ICD9 codes,[23] may also improve prediction of readmission. Beyond prediction, it is also possible that identifying particularly high-risk topics will facilitate the development of interventions to reduce readmission risk in particular subgroups.

Outside of psychiatry, topic modeling has been demonstrated to improve prediction of mortality in intensive care unit populations.[13,14,24] Another natural language-processing approach has recently been demonstrated to improve prediction of suicide risk in a small case–control study of narrative notes from military veterans.[15] However, to our knowledge, topic modeling has not previously been applied for predicting psychiatric hospital readmissions.

We emphasize several important limitations in interpreting our results. First, while significantly better than chance, these predictions may not be sufficient by themselves to target interventions, particularly costly ones. Clinical assessment is particularly important in this regard, and the increasing emphasis on symptom quantification may provide another means of improving prediction based on EHRs.[25] Another valuable comparison would consider clinician estimates of risk at time of hospital discharge, and the extent to which the variance explained by the present model overlaps with or complements clinician estimate. The goal of this initial study was rather to establish a baseline, using only artifacts of routine clinical care, which may be improved with more detailed assessments. Although discrimination may not be sufficient for clinical application (that is, sensitivity of 75% but specificity of 63%), the optimal threshold for this trade-off will depend on the nature of the intervention contemplated in those individuals characterized as high risk.

Further, as we were unable to identify a similar clinical cohort available for collaboration, we could not examine the generalizability and portability of our models. This challenge highlights the importance of establishing deidentified clinical cohorts to allow new research and clinical tools relying on natural language processing to be validated, as has been done for obesity, for example.[26] Such validation would be a necessary next step before attempting to disseminate risk stratification models beyond this New England health system. We would expect, however, that topic models would generalize better than (for example) rules-based classifiers that rely on individual terms.

Table 2. Example topics for MDD patients readmitted with a psychiatric diagnosis within 30 days

Terms	*Topic annotation*
*patient *alcohol withdrawal* depression *drinking* end ativan *etoh drinks* medications clinic inpatient diagnosis days hospital *< substance use treatment program name > use abuse* problem number	Alcohol
*mg daily discharge *anxiety klonopin seroquel clonazepam* admission wellbutrin given md lexapro date b signed night low admitted sustained hospitalization	Anxiety
**ideation suicidal mood decreased* hallucinations history *depressed depression thought* psychiatric energy denied sleep auditory appetite homicidal symptoms increased speech *thoughts*	Suicidality
**ect depression* treatment treatments dr mg course *< ECT physician name >* symptoms received medications prior improved decreased medication md trials tsh continued qhs	ECT
**weight eating* admission discharge hospital intake *loss* date hospitalization day dr week physical months *prozac food* increased md did *anorexia*	Anorexia
**seizure seizures* intact *eeg neurology* normal *temporal dilantin* head *bilaterally events activity* weakness sensation disorder tongue *neurologist brain* loss *tegretol*	Seizure
**therapist mother* program *father* disorder age school parents brother *abuse* treatment *relationship* outpatient college behavior partial plan currently *group personality*	Psychotherapy
*psychiatry *suicide overdose attempt transferred* depression *transfer* level *tylenol* hospital service *unit* normal floor screen *tox* room admission medical general	Overdose
**baby delivery bleeding vaginal breast feeding cesarean* weight ibuprofen *maternal newborn* available p fever *pregnancy* sex estimated *danger* gp	Postpartum
**psychotic thought* features *paranoid psychosis paranoia* symptoms psychiatric dose continued treatment mental cognitive memory *risperidone* people th somewhat interview affect	Psychosis

Abbreviation: MDD, major depressive disorder; ECT, electroconvulsive therapy.

Finally, as the health system examined here is not a closed system—that is, it is possible that individuals could be readmitted to another hospital and go undetected—some misclassification is to be expected. In the state of Massachusetts, the vast majority of individuals requiring readmission are (per payer and regulatory requirements) readmitted to the index hospital. Such misclassification would tend to decrease the discrimination of predictive models, so the estimates provided here are likely conservative.

Despite these caveats, these results suggest the potential power of narrative notes to identify meaningful predictors of outcome not available in coded data alone. They indicate that it may be possible to develop models to identify psychiatric patients at high risk for hospital readmission, a necessary first step in developing interventions to address this risk.

CONFLICT OF INTEREST

RHP has served on advisory boards or provided consulting to AssureRx, Genomind, Healthrageous, Pamlab, Perfect Health, Pfizer, Psybrain and RIDVentures. The remaining authors declare no conflict of interest.

REFERENCES

1 Anonymous. CMS Readmissions Reduction Program 2014. Available at http://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/Readmissions-Reduction-Program.html; accessed on 30 March 2015.
2 National Research Council (US) Committee on A Framework for Developing a New Taxonomy of Disease. *Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease*. National Academic Press: Washington DC, USA, 2011.
3 Reis BY, Kohane IS, Mandl KD. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. *BMJ* 2009; **339**: b3677.
4 Perlis RH. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. *Biol Psychiatry* 2013; **74**: 7–14.
5 Perlis R, Iosifescu D, Castro V, Murphy S, Gainer V, Minnier J *et al.* Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med* 2012; **42**: 41–50.
6 Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL *et al.* Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care* 2007; **13**: 281–288.
7 Liao KP, Cai T, Gainer V, Goryachev S, Zeng-treitler Q, Raychaudhuri S *et al.* Electronic medical records for discovery research in rheumatoid arthritis. *Arthritis Care Res* 2010; **62**: 1120–1127.
8 Sohn S, Kocher JPA, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. *J Am Med Inform Assoc* 2011; **18**: i144–i149.
9 Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. *AMIA Annu Symp Proc* 2012; **2012**: 1244–1253.
10 Elkin PL, Froehling D, Wahner-Roedler D, Trusko B, Welsh G, Ma H *et al.* NLP-based identification of pneumonia cases from free-text radiological reports. *AMIA Annu Symp Proc* 2008; **6**: 172–176.
11 Murff HJ, FitzHenry F, Matheny ME, Gentry N, Kotter KL, Crimin K *et al.* Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA* 2011; **306**: 848--855.
12 Garla V, Re VL III, Dorey-Stein Z, Kidwai F, Scotch M, Womack J *et al.* The Yale cTAKES extensions for document classification: architecture and application. *J Am Med Inform Assoc* 2011; **18**: 614--620.
13 Ghassemi M, Naumann T, Joshi R, Rumshisky A. Topic Models for Mortality Modeling in Intensive Care Units. *Proceedings of the ICML Workshop on Machine Learning for Clinical Data Analysis*; Edinburgh, Scotland, 2012.
14 Ghassemi M, Naumann T, Doshi-Velez F, Brimmer N, Joshi R, Rumshisky A *et al.* (eds). Unfolding Physiological State: Mortality Modelling in Intensive Care Units. *Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD2014)*; ACM: New York, NY, USA, 2014, pp 75–84.
15 Poulin C, Shiner B, Thompson P, Vepstas L, Young-Xu Y, Goertzel B *et al.* Predicting the risk of suicide by analyzing the text of clinical notes. *PLoS One* 2014; **9**: e85733.
16 Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S *et al.* Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J Am Med Inform Assoc* 2010; **17**: 124–130.
17 Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O *et al.* Scikit-learn: machine learning in Python. *J Mach Learn Res* 2011; **12**: 2825–2830.
18 Salton G, Wong A, Yang CS. A vector space model for automatic indexing. *Commun ACM* 1975; **18**: 613--620.
19 Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *J Mach Learn Res* 2003; **3**: 993–1022.
20 Griffiths TSM. Finding scientific topics. *Proc Natl Acad Sci USA* 2004; **101**: 5228–5235.
21 Charlson M, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. *J Clin Epidemiol* 1994; **47**: 1245–1251.
22 Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chronic Dis* 1987; **40**: 373–383.

23 Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J *et al.* Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med* 2012; **42**: 41–50.
24 Lehman LSM, Long W, Lee J, Mark R. Risk stratification of ICU patients using topic models inferred from unstructured progress notes. *AMIA Annu Symp Proc* 2012; **2012**: 505–511.
25 Simon GE, Perlis RH. Personalized medicine for depression: can we match patients with treatments? *Am J Psychiatry* 2010; **167**: 1445–1455.
26 Uzuner O. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc* 2009; **16**: 561–570.

Supplementary Information accompanies the paper on the Translational Psychiatry website (http://www.nature.com/tp)