Mining high-dimensional administrative claims data to predict early hospital readmissions

Danning He,1,2 Simon C Mathews,3 Anthony N Kalloo,4 Susan Hutfless4

ABSTRACT

Background Current readmission models use administrative data supplemented with clinical information. However, the majority of these result in poor predictive performance (area under the curve (AUC)<0.70).

Objective To develop an administrative claim-based algorithm to predict 30-day readmission using standardized billing codes and basic admission characteristics available before discharge.

Materials and Methods The algorithm works by exploiting high-dimensional information in administrative claims data and automatically selecting empirical risk factors. We applied the algorithm to index admissions in two types of hospitalized patient: (1) medical patients and (2) patients with chronic pancreatitis (CP). We trained the models on 26 091 medical admissions and 3218 CP admissions from The Johns Hopkins Hospital (a tertiary research medical center) and tested them on 16 194 medical admissions and 706 CP admissions from Johns Hopkins Bayview Medical Center (a hospital that serves a more general patient population), and vice versa.

Performance metrics included AUC, sensitivity, specificity, positive predictive values, negative predictive values, and F-measure.

Results From a pool of up to 5665 International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) diagnoses, 599 ICD-9-CM procedures, and 1815 Current Procedural Terminology codes observed, the algorithm learned a model consisting of 18 attributes from the medical patient cohort and five attributes from the CP cohort. Within-site and across-site validations had an AUC≥0.75 for the medical patient cohort and an AUC≥0.65 for the CP cohort.

Conclusions We have created an algorithm that is widely applicable to various patient cohorts and portable across institutions. The algorithm performed similarly to state-of-the-art readmission models that require clinical data.

BACKGROUND AND SIGNIFICANCE

Administrative claims data are records of transactions that occurred between patients and healthcare providers. They are often the electronic versions of paper healthcare providers submit to payers for outpatient visits and inpatient hospital stays. The core data elements include services, diagnosis, procedures, beneficiary demographics, and admission characteristics. Administrative data are a viable source for research1 because they are readily available, inexpensive, cover large populations, and can be available in near real time in some hospitals (depending on the source). Researchers have used administrative data for a wide range of applications, including predicting hospital readmissions,2–4 assessing quality and patient safety,5–8 improving risk adjustment for hospital mortality9 and morbidity10 identifying cohorts,11–15 and evaluating effectiveness of treatments.16–19 These predictive models can be used to make decisions about reimbursement, compare quality across hospitals, and perform comparative effectiveness research.2–3

Hospitalizations account for almost 31% of healthcare expenditures.20 Yet it is estimated that 20% of Medicare patients discharged from a hospital return within 30 days, and 34% are rehospitalized within 90 days.21 Most readmissions are potentially preventable and are indicators of poor quality of care or poor coordination among healthcare providers.21–23 The Medicare Payment Advisory Commission initiatives24 aim to reduce hospital readmissions through payment policies that penalize for excessive readmission rates. A first step to reducing preventable hospital readmissions is to identify predictors of early readmission and assess risk in individual patients. Reducing readmissions will reduce unnecessary costs and increase the value of the healthcare institution.23

A recent systematic review2 summarized 26 unique readmission prediction models based on administrative data,3 electronic medical records,26 27 or a combination of the two.26 28 Of the 14 models using 30-day readmission as the outcome, most models have poor predictive performance. Among the four models,1 3 26 27 29 with moderate discriminative power (area under the curve (AUC)>0.70), three3 27 29 used clinical details and one26 was based on small samples (700 training samples, 704 validation samples).

OBJECTIVE

In this study, we developed an algorithm to predict 30-day readmission using billing codes and basic administrative information available before discharge. We evaluated the performance using both within-site cross-validation and across-site validation at The Johns Hopkins Hospital (JHH) (a tertiary referral center) and at Johns Hopkins Bayview Medical Center (BMC) (a community hospital). We used two distinct application settings, (1) predicting 30-day outcome in medical patients and (2) predicting 30-day outcome in patients with chronic pancreatitis (CP) (known to be at high risk of readmission).30 The resulting models demonstrated discriminative power and easy portability between the referral and community hospitals. The latter model demonstrated a disease-specific application of this modeling approach, which could be applied to a number of other clinical contexts.

METHODS

Data source The JHH’s Casemix Information Management Department developed the Johns Hopkins Medicine

We used a five-step procedure to achieve automatic feature selection and model learning (figure 2). The latest source code for the algorithm is available at https://my.vanderbilt.edu/readmission/.

1. **Defining input attributes.** Input attributes included age, sex, ethnicity, marital status, length of stay, accumulated number of inpatient admissions in the past 5 years, ICD-9-CM diagnosis and procedure codes, and CPT codes. For billing codes, an attribute takes a value of 1 if that code is observed during the index admission, or 0 otherwise.

2. **Curating potential attributes by prevalence prioritization.** This step excluded the attributes with low prevalence (<1%) from further analyses. Prevalence is the proportion of individuals who have a specific code. Although candidate attributes were of high dimension (table 1), the majority of them occurred in <1% of the population. In hierarchical coding, increasing the granularity decreases the prevalence, suggesting that generic codes may be preferable. However, fine-level information serves as a better proxy of the conditions associated with readmission.

![Figure 1](image)

**Figure 1** Evaluation workflow. The figure shows the model learning on the entire Johns Hopkins Hospital (JHH) cohort and tested on the entire Bayview Medical Center (BMC) cohort. Training on BMC and testing on JHH were performed in a similar fashion (not pictured).
and procedures experienced by the patient, making it easier for healthcare providers to interpret models. We therefore kept all codes at their original level.

3. **Univariate variable selection.** This step identified top ranked attributes. Each attribute’s significance of correlation with 30-day outcome was defined as the Likelihood Ratio Test (LRT) p values of the fitted logistic regression, in which the response factor was 30-day outcome, and the explanatory factor was the tested attribute. Attributes with a significance level above the user-supplied threshold were retained for further analyses.

4. **Multivariate variable selection.** This step performed subset selection for the multivariate model by forward variable selection along a path generated by random ordering of the variables. LRT was conducted to determine whether the addition of a new attribute significantly reduced the unexplained variance. An insignificant attribute was dropped, and the next attribute in the path was added. The process was repeated until all attributes had been tested once. Further details about this approach and discussion of its limitations are provided in online supplementary text and in the source code.

5. **Final model.** We estimated a risk score for each patient as the predicted probability of 30-day readmission conditional on attributes selected from step 4. We performed logistic regression on the selected attributes using the GLM function in the statistical program R.

### Issues related to parameter adjustment

The algorithm required two parameters: the p value for univariate variable selection and the p value for multivariate variable selection, which were chosen via internal fivefold cross-validation. The search of p values was on a log10 scale. The search spaces for the first p values were those that allow 0.5–1.5% of the total observed codes to be included in the logistic regression model, and the search spaces for the second p values were those that allow <0.5% of the total observed codes to be in the final model (table 2).

### Evaluation of performance

Within-site performance was evaluated by fivefold cross-validation. The complete algorithm (step 1 through step 5) was applied to the four training folds, and the final model learned from the fourfolds was validated on the remaining onefold. We reported performance metrics such as AUC, sensitivity, specificity, optimal F-measure, positive predictive values, and negative predictive values at the optimal cutoff (averaged over folds for cross-validation). We plotted receiver operating characteristic

<table>
<thead>
<tr>
<th>Table 2 Parameter adjustment and model size after univariate variable selection and multivariate variable selection</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cohort</strong></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>ME</td>
</tr>
<tr>
<td>JHH</td>
</tr>
<tr>
<td>BMC</td>
</tr>
<tr>
<td>CP</td>
</tr>
<tr>
<td>JHH</td>
</tr>
<tr>
<td>BMC</td>
</tr>
</tbody>
</table>

The models refer to the ones learned from the entire patient cohort that were subject to across-site analyses using the p value cutoffs identified via internal cross-validation. BMC, Bayview Medical Center; CP, patients with chronic pancreatitis; JHH, Johns Hopkins Hospital; ME, all patients admitted to the medical service.
RESULTS

We list basic demographic and admission characteristics for the patient cohorts from JHH and BMC in Table 1. The JHH cohorts had longer lengths of stay (Wilcoxon test p value=1.6E-120), possibly because the patient population included sicker patients seen at a referral center.

Evaluation on medical patients

In the JHH medical patient cohort, of 26,091 indexed admissions between January 2012 and April 2013 (corresponding to 18,974 individuals), 11.5% were unplanned readmissions within 30 days of the prior discharge. From a pool of 5,665 ICD-9-CM diagnoses, 599 ICD-9-CM procedures, 1,815 CPT procedures (which have at least one occurrence in the population), the learned model contained 18 attributes (one ICD-9-CM procedure, 16 CPT procedures, and one administrative attribute). We list the coefficients’ estimates and their p values for the top 10 attributes in Table 3, and provide a full model in online supplementary Table S2. Across-site validation on the BMC cohort had AUC=0.81 (Figure 3A). We provide the model derived from the entire BMC cohort (n=16,194) in online supplementary Table S3. Across-site validation on the JHH cohort had AUC=0.78 (Figure 3B). In general, the algorithm showed similar performance for JHH and BMC. Across-site validations showed similar performance to within-site fivefold cross-validation (AUC=0.75±0.01 on the JHH cohort and AUC=0.79±0.02 on the BMC cohort). Other performance metrics are shown in Table 4.

Of the 19 attributes in the model derived from the JHH cohort, the top predictor was accumulated number of inpatient admissions over the past 5 years (t test p value=3.13E-172). Consistent with previous studies of readmissions,28 29 32 33 the accumulated number of prior admissions was an important predictor of early readmission.

Evaluation on CP patients

In the JHH CP cohort, of 3,218 indexed admissions between January 2007 and April 2013 (corresponding to 1,763 individuals), 15.6% were readmissions within 30 days of the prior discharge. There were 3,696 ICD-9-CM diagnoses, 751 ICD-9-CM procedures, and 1,333 CPT procedures that occurred at least once in the population, of which seven ICD-9-CM diagnoses, four ICD-9-CM procedures, 24 CPT procedures, and one administrative attribute passed univariate variable selection (we list their Spearman correlation coefficients in Figure 4). The final model consisted of five attributes (one ICD-9-CM procedure, three CPT procedures, and one administrative attribute); we list their coefficients’ estimates and p values in Table 5. Testing the model on the BMC cohort produced an AUC=0.65 (Figure 3C). We provide the model derived from the BMC CP cohort (n=706) in online supplementary Table S4. Testing it on the JHH cohort produced an AUC=0.73 (Figure 3D). We list various performance metrics in Table 3.

Of five attributes in the model derived from JHH, the strongest predictor was the accumulated number of inpatient admissions over the past 5 years (t test p value=1.41E-28), the same attribute as that of medical patients. Other predictors were disease specific. The next two predictors were lipase (CPT: 83690, t test p value=1.78E-07), which is related to pancreatic enzyme analysis and radical pancreaticoduodenectomy (ICD-9-CM procedure: 527, LRT p value=6.45E-20), a surgical procedure for total pancreatic resection. The other two attributes were arterial blood gas (CPT: 82803, t test p value=3.56E-01) and surgical pathology (CPT: 88309, t test p value=6.26E-02).

DISCUSSION

Significance of the method

We developed an algorithm to predict hospital readmission based on standardized billing codes and basic administrative attributes that are readily accessible in most billing data. Our data-driven method was inspired by pharmacoepidemiologic studies,19 34 which identified empirical covariates such as diagnoses and medications from administrative data to adjust for confounders. We subsequently evaluated the predictive performance of the algorithm on large patient cohorts at both a tertiary and community hospital. The proposed algorithm has four significant advantages. First, it is fully automatic and can be applied to a wide range of patient cohorts, from medical patients to patients with specific diseases such as CP. Second, a type of branch and bound feature selection was applied by exploiting high-dimensional candidate attributes. No assumptions were made about the potential predictors of interest, with the algorithm automatically selecting top ranked, independent attributes specific to the given patient cohort. Such automated routines are most appropriate for fast exploitation of a large number of attributes to identify candidate risk factors and their surrogates for further detailed analysis, since little is known.

| Category | Code | Description | Estimate | SE | Z value | Pr(>|z|) |
|----------|------|-------------|----------|----|---------|---------|
| CPT      | 36430| Transfusion, blood or blood components | 0.74     | 0.12| 5.99    | 2.16E-09|
| procedure| J1170| Injection, hydromorphone, up to 4 mg | 0.29     | 0.05| 5.60    | 2.12E-08|
|          | 86900| Blood typing: ABO | 0.29     | 0.05| 5.53    | 3.15E-08|
|          | J7507| Tacrolimus, oral, per 1 mg | 0.32     | 0.07| 4.59    | 6.06E-07|
|          | 87086| Culture, bacterial; quantitative colony count, urine | 0.20     | 0.04| 4.70    | 2.61E-06|
|          | 96375| Therapeutic, prophylactic, or diagnostic injection (specify substance or drug); each additional sequential intravenous push of a new substance/drug (list separately in addition to code for primary procedure) | 0.49     | 0.11| 4.34    | 1.40E-05|
|          | 83605| Lactate (lactic acid) | 0.19     | 0.05| 4.05    | 3.36E-04|
|          | 64891| Nerve graft (includes obtaining graft), single strand, hand or foot; more than 4 cm length | -1.09    | 0.30| 3.64    | 2.70E-04|
|          | 83609| Lipase | 0.17     | 0.05| 3.34    | 8.50E-04|
| Admission| adm_acc | 5-year accumulated number of admissions | 0.07     | 0.00| 27.98   | 3.13E-172|

We provide the complete model in online supplementary Table S2. Estimate, estimated coefficients; Pr(>|z|), coefficient t test p values.

CPT, Current Procedural Terminology; JHH, Johns Hopkins Hospital; ME, all patients admitted to the medical service.
about the potential risk factors at this preliminary stage of discovery and there is no a priori reason to prefer one candidate attribute over another. Third, the algorithm requires only basic administrative attributes and billing codes that are readily available and standardized in billing databases across institutions. We were able to develop the model using a referral center cohort, and successfully validate it on a community hospital cohort (and vice versa) because of its easy portability. Although it is possible that additional information that is not stored as part of billing records or within standard electronic medical records may add to the predictability of 30-day readmission, such as functional status or physiological status at the time of discharge, these measures are not universally available. We aimed to create an algorithm that was widely applicable to various settings, and therefore did not include these measures as potential attributes. Fourth, our method demonstrated fairly good discriminatory power to predict the risk of readmission in medical patients (internal validation, AUC=0.75±0.01 and 0.79±0.02; cross-validation, AUC=0.78±0.02 and 0.79±0.02).

Table 4 Model performance

<table>
<thead>
<tr>
<th>Cohort</th>
<th>Validation</th>
<th>F-measure</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>ME</td>
<td>Within-site CV on JHH</td>
<td>0.37±0.01</td>
<td>0.50±0.05</td>
<td>0.84±0.03</td>
<td>0.29±0.01</td>
<td>0.93±0.01</td>
<td>0.75±0.01</td>
</tr>
<tr>
<td></td>
<td>CV on BMC</td>
<td>0.36±0.10</td>
<td>0.50±0.05</td>
<td>0.88±0.02</td>
<td>0.28±0.02</td>
<td>0.95±0.01</td>
<td>0.79±0.02</td>
</tr>
<tr>
<td></td>
<td>Across-site Test on BMC</td>
<td>0.34</td>
<td>0.47</td>
<td>0.88</td>
<td>0.27</td>
<td>0.95</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td>Test on JHH</td>
<td>0.34</td>
<td>0.58</td>
<td>0.76</td>
<td>0.24</td>
<td>0.93</td>
<td>0.78</td>
</tr>
<tr>
<td>CP</td>
<td>Within-site CV on JHH</td>
<td>0.36±0.10</td>
<td>0.50±0.11</td>
<td>0.68±0.20</td>
<td>0.34±0.06</td>
<td>0.84±0.02</td>
<td>0.71±0.05</td>
</tr>
<tr>
<td></td>
<td>CV on BMC</td>
<td>0.28±0.05</td>
<td>0.56±0.19</td>
<td>0.79±0.11</td>
<td>0.20±0.06</td>
<td>0.95±0.02</td>
<td>0.65±0.05</td>
</tr>
<tr>
<td></td>
<td>Across-site Test on BMC</td>
<td>0.19</td>
<td>0.85</td>
<td>0.41</td>
<td>0.11</td>
<td>0.97</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td>Test on JHH</td>
<td>0.38</td>
<td>0.60</td>
<td>0.71</td>
<td>0.27</td>
<td>0.91</td>
<td>0.73</td>
</tr>
</tbody>
</table>

We report sensitivity, specificity, positive predictive values (PPVs), and negative predictive values (NPVs) at the cutoff thresholds that maximize the F-measure. For cross-validation, the table shows the mean±SD of the performance metric over the folds.

AUC, area under the curve; BMC, Bayview Medical Center; CP, patients with chronic pancreatitis; JHH, Johns Hopkins Hospital; ME, all patients admitted to the medical service.

Figure 3 Receiver operating characteristic curves for across-site analyses on medical patient (ME) cohort and chronic pancreatitis (CP) cohort. (A) Training ME cohort on Johns Hopkins Hospital (JHH) and testing on Bayview Medical Center (BMC) (area under the curve (AUC)=0.81). (B) Training ME cohort on BMC and testing on JHH (AUC=0.78). (C) Training CP cohort on JHH and testing on BMC (AUC=0.65). (D) Training CP cohort on BMC and testing on JHH (AUC=0.73). Line color: threshold cutoff.
across-site validation, AUC=0.78 and 0.81). In a recent systematic review\(^2\) that summarized 14 unique prediction models for 30-day readmission, of eight models targeting medical patients, only two had an AUC >0.70\(^,26\) and only one had an AUC >0.75 (this study was based on a much smaller cohort with a training size of 700 patients and validation size of 704 patients; internal validation, AUC=0.77\(^,26\)). The most recently published study\(^{29}\) on 30-day readmission in medical patients had an AUC of 0.71 via internal validation. We had hoped to compare our model with the most recent model, but we were not able to because three laboratory test values required by the model were not available in our administrative data. In summary, our model was created in a fully automatic fashion yet still showed comparable or even better performance when compared with state-of-the-art approaches that involve data not available in standard billing records.

Admittedly, there are many other machine learning techniques serving similar purposes. There are several reasons why we prefer the current approach to many others. (1) Our approach is simpler and more intuitive than other sophisticated machine learning techniques; it is important for non-statisticians such as clinicians and administrative staff to understand the method in order to adopt it. (2) Our choice of approach is also related to the administrative data we used. Unlike clinical data, which contain fine-grained detail at high accuracy, administrative data only provide coarse-grained information about a patient. Given that the documented diagnoses and procedures themselves only serve as proxies for the true conditions experienced by the patients and that our goal is fast exploitation of all attributes to identify candidate risk factors or their surrogates for further detailed clinical investigation, we believe the current approach is sufficient for pivotal discovery. (3) We compared the

---

**Table 5** Final chronic pancreatitis readmission model learned from the JHH cohort

| Category         | Code | Description                                                                 | Estimate | SE  | Pr(>|z|) |
|------------------|------|-----------------------------------------------------------------------------|----------|-----|----------|
| ICD-9-CM procedure | 527  | Radical pancreaticoduodenectomy                                              | −1.301   | 0.649 | 4.50E-02 |
| CPT procedure     | 82803| Gases, blood, any combination of pH, pCO\(_2\), pO\(_2\), CO\(_2\), HCO\(_3\) (including calculated O\(_2\) saturation) | −0.138   | 0.150 | 3.56E-01 |
|                  | 83690| Lipase                                                                      | 0.652    | 0.125 | 1.78E-07 |
|                  | 88309| Level IV—surgical pathology, gross and microscopic (colon biopsy, lymph node biopsy, colorectal polyp) | −0.818   | 0.439 | 6.26E-02 |
| Admission         | adm_acc | 5-year accumulated number of admissions                                     | 0.021    | 0.002 | 1.41E-28 |


---

computation and predictive performance of our approach with the standard stepwise forward approach: our approach has less computation complexity, but similar predictive performance (using AUC as the performance metric). Since we aimed to produce an algorithm applied to high-dimensional data for cross-institutional analyses, speed is of key concern here. (4) The contribution of our study lies in the algorithmic workflow as a whole; further research can definitely be based on our proposed framework, but replacing the simple approach with more sophisticated machine learning techniques.

Significance of the discovery
While the large cohort of medical patients served as a good patient population for evaluating the performance of our algorithm relative to recent readmission prediction approaches, the additional application to a cohort of patients with a specific disease known to be at high risk of readmission may allow targeted interventions to reduce early readmissions. CP was selected for evaluation for several reasons. First, there is no current prediction model available for readmission in CP patients. Second, many patients with this condition need early unplanned readmission to hospital, resulting in significant healthcare costs. Finally, the incidence of CP is increasing in the USA. During hospitalizations, gastroenterologists often guide care. For an intervention to be implemented for patients with CP, the gastroenterologist must agree that these predictors have content validity before they are likely to support a readmission intervention as part of their care. Here, we implemented the algorithm in a CP patient cohort with the aim of identifying individual-level risk factors for preventable readmission. The risk factors, together with their highly correlated proxies (figure 4), differ from those for all patients admitted to the medical service and are specific to CP patients. Our model provides validated predictors for readmission on a going-forward basis creating a foundation for identifying potential targeted interventions. For example, in our CP cohort, this could mean real-time monitoring of specific test utilization associated with disease-specific readmission (eg, lipase and blood gas) through the electronic medical record, while also advocating more aggressive follow-up in a specific sub-population (eg, patients who have had a pancreaticoduodenectomy). The gastroenterology co-investigators are in the process of designing an intervention that will facilitate the adoption of best practice to prevent costly readmission of CP patients identified as high risk on the basis of predictors from our model and their proxies. The intervention may include delaying discharge, scheduling early outpatient follow-up, or phone calls from or home visits by gastroenterology nurses soon after discharge. In contrast, if the patient is predicted to be at low risk of readmission, the clinician might feel comfortable discharging him or her without the need for these interventions to prevent readmission.

Study limitations
Our study has several limitations. First, models based on administrative claims data may contain coding irregularities and lack clinical details. In addition, they generally cannot be used early in the course of an initial hospitalization, limiting the time and interventions available to effectively prevent potentially avoidable readmissions, especially for admissions with a short length of stay. Second, we were not able to identify readmissions to hospitals outside of JHH and BMC because of the lack of a universal identifier shared across sites. Third, our identification of unplanned readmissions was based on clinical judgment, since there is no follow-up indicator variable available in our administrative data. It is likely that our exclusion of planned readmissions is incomplete. Fourth, our identification of a CP patient cohort based solely on ICD-9-CM diagnosis codes may not be accurate enough, and predictive performance may be compromised as a result. Except for certain medical conditions, such as hip fracture and cancer, diagnosis/procedure codes for many other medical conditions, such as peripheral vascular disease (sensitivity=0.58, positive values=0.53), are of limited accuracy. For example, the absence of a chronic disease diagnosis code may be the result of more serious and acute illnesses that push the chronic condition off the list. However, we found the average diagnosis ranking of CP was 4, and our administrative data allow up to 50 diagnosis positions, suggesting that this potential limitation had minimal impact on our cohort identification. Fifth, the predictive performance of CP-specific readmission is lower than that of medical patients. Possible reasons may be related to cohort identification (discussed in the section above) and the small size of the validation cohort at BMC (n=706) (a significant proportion of codes that occurred in the JHH CP cohort never occurred in the BMC CP cohort, whereas this is not the case for the medical patient cohort). Another possible reason is that, whereas the CP cohort is a homogeneous population, the medical patient cohort contains various disease types. Patients with some types of disease were more likely than average to be readmitted, and surrogates of such diseases increase the predictive performance of the model, although these predictors may not be the driving underlying reason for readmission itself. Sixth, we were unable to incorporate clinical laboratory data in our model and as a result were not able to directly compare its performance with other recent approaches. Nevertheless, our model had performance characteristics that were comparable to or better than existing alternatives. Seventh, there are many other more sophisticated machine learning techniques for variable selection than our proposed approach. However, given that the administrative data only provide coarse information on the patients, and the documented diagnoses and procedures themselves only serve as proxies for the true conditions experienced by the patients, we felt that a simple intuitive approach, which would be more acceptable for clinicians and administrative staff, would suffice for our purposes. In fact, a comparison with another variable-selection method reveals similar predictive performances (see online supplementary text).

CONCLUSION
This study presents a branch and bound algorithm for predicting 30-day readmission using administrative claims data. Through exploitation of high-dimensional yet universally available information in claims, the algorithm is widely applicable to various patient cohorts across both tertiary and community-based hospital centers, with good performance characteristics. In summary, the advantage of the algorithm is its ability to maintain performance while maximizing portability and adaptability. Future application of this approach could include the use of nationally available data such as Medicaid and Medicare billing records for cross-institution analyses.

Acknowledgements The authors wish to acknowledge and thank Harold Lehmann for his helpful suggestions on this study, and Eric Vohr for his editorial assistance with the manuscript.

Contributors Study initialization: SH, ANK, DH, and SCM. Study design: DH and SH. Analysis and interpretation of data: DH, SH, and SCM. Drafting of the manuscript: DH, SH, SCM, and ANK. All authors contributed to refinement of the manuscript and approved the final manuscript.

Funding The study was funded from departmental resources.

Competing interests None.

Ethics approval The Johns Hopkins Medicine Institutional Review Board approved this study (IRB#: NA_00082734).

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES


