

# Common sampling and modeling approaches to analyzing readmission risk that ignore clustering produce misleading results

**Huaqing Zhao**

Lewis Katz School of Medicine at Temple University

**Samuel Tanner**

Lewis Katz School of Medicine at Temple University <https://orcid.org/0000-0002-1204-1550>

**Sherita H. Golden**

Johns Hopkins University School of Medicine

**Susan Fisher**

Lewis Katz School of Medicine at Temple University

**Daniel J. Rubin** (✉ [Daniel.Rubin@tuhs.temple.edu](mailto:Daniel.Rubin@tuhs.temple.edu))

<https://orcid.org/0000-0002-6871-6246>

---

## Research article

**Keywords:** logistic models, patient readmission, predictive modeling, sampling strategies, clustering

**Posted Date:** September 14th, 2020

**DOI:** <https://doi.org/10.21203/rs.3.rs-27550/v2>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published on November 25th, 2020. See the published version at <https://doi.org/10.1186/s12874-020-01162-0>.

# Abstract

**Background:** There is little consensus on how to sample hospitalizations and analyze multiple variables to model readmission risk. The purpose of this study was to compare readmission rates and the accuracy of predictive models based on different sampling and multivariable modeling approaches.

**Methods:** We conducted a retrospective cohort study of 17 284 adult diabetes patients with 44 203 discharges from an urban academic medical center between 1/1/2004 and 12/31/2012. Models for all-cause 30-day readmission were developed by four strategies: logistic regression using the first discharge per patient (LR-first), logistic regression using all discharges (LR-all), generalized estimating equations (GEE) using all discharges, and cluster-weighted (CW) GEE using all discharges. Multiple sets of models were developed and internally validated across a range of sample sizes.

**Results:** The readmission rate was 10.2% among first discharges and 20.3% among all discharges, revealing that sampling only first discharges underestimates a population's readmission rate. Number of discharges was highly correlated with number of readmissions ( $r=0.87$ ,  $P<0.001$ ). Accounting for clustering with GEE and CWGEE yielded more conservative estimates of model performance than LR-all. LR-first produced falsely optimistic Brier scores. Model performance was unstable below samples of 6 000-8 000 discharges and stable in larger samples. GEE and CWGEE performed better in larger samples than in smaller samples.

**Conclusions:** Hospital readmission risk models should be based on all discharges as opposed to just the first discharge per patient and utilize methods that account for clustered data.

## Background

Recent interest in value-based care has focused attention on quality metrics in healthcare delivery. One metric, the all-cause emergency (unplanned) 30-day hospital readmission rate, has received considerable attention because it may be related to poor care, and hospitals in the US with excess readmission rates are subject to financial penalties<sup>1,2</sup>. To better understand and ultimately prevent readmissions, numerous studies have examined readmission risk factors and/or developed multivariable predictive models<sup>3-5</sup>.

Despite this growing body of literature, there is little consensus on how to sample hospitalizations and analyze multiple variables to model readmission risk. It is possible that different sampling strategies may yield different readmission rates. For example, the 30-day readmission rate for patients diagnosed with diabetes reported in the literature ranges from 10.0%, in which only the first hospitalization per patient was included as an observation, to 20.4%, in which all hospitalizations were considered as observations<sup>6,7</sup>. While some variation is expected due to differences in populations across studies, sampling strategy may independently affect observed readmission rates.

In addition to variation in sampling methods, there are also multiple approaches to multivariable modeling. The most common approach is logistic regression, which treats each observation (hospital

discharge) as independent <sup>4,5</sup>. While this model is the least computationally demanding and the most familiar, the assumption of independence may not be valid for multiple hospitalizations of an individual patient during a given study period. The readmission risk of a patient at the time of discharge #1 is likely related to the readmission risk of the same patient at the time of discharge #2. In this case, the patient may be considered as a cluster of 2 hospital discharges. Simulation studies show that analytical approaches on clustered data that do not take into account the effects of clustering often yield erroneous results <sup>8,9</sup>.

One method that accounts for longitudinal correlations of data within clusters is generalized estimating equations (GEE) <sup>10</sup>. In the GEE approach, the intra-cluster correlation is modeled to determine the weight that should be assigned to data from each cluster <sup>11</sup>. If the outcome is independent of cluster size (i.e., cluster size is uninformative), then this approach is valid. However, if the outcome is related to cluster size, as is likely with readmission risk and number of discharges, then GEE may generate misleading parameter estimates. For example, people with poor dental health are likely to have fewer teeth than those with good dental

health because factors that lead to poor dental health also lead to tooth loss. Therefore, in a study investigating risk factors for tooth disease, the number of teeth per person (cluster size) would be informative <sup>11</sup>. Cluster-weighted GEE (CWGEE) has been proposed as a valid approach to analyzing clustered data with an informative cluster size <sup>11</sup>. While some examples of the GEE approach exist in the readmission literature <sup>7,12,13</sup>, to our knowledge, CWGEE has not been used in this context.

Herein, we compare readmission rates and predictive model accuracy of different sampling and multivariable modeling approaches in a dataset of diabetes patients previously used to develop a readmission risk prediction model <sup>7</sup>.

## Methods

A cohort of 17 284 patients discharged from an urban academic medical center (Boston Medical Center) between 1/1/2004 and 12/31/2012 were selected, and 44 203 discharges from this cohort comprised the complete dataset. Inclusion criteria for index discharges were diabetes defined by an International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) code of 250.xx associated with hospital discharge or the presence of a diabetes-specific medication on the pre-admission medication list. Index discharges were excluded for patient age <18 years, discharge by transfer to another hospital, discharge from an obstetric service (indicating pregnancy), inpatient death, outpatient death within 30 days of discharge, or incomplete data. Readmission documented within 8 hours of a discharge was merged with the index admission to avoid counting in-hospital transfer as a readmission.

The primary outcome was all-cause readmission within 30 days of discharge. The same 46 variables previously used to develop a readmission risk prediction model were evaluated as predictors of the

primary outcome to construct and validate all prediction models (see Table, Supplemental Digital Content 1, which presents patient characteristics on the variables analyzed) <sup>7</sup>.

Five different measures of performance were assessed for each model.

1) Diagnostic discrimination (*C* statistic): the area under the receiver operating characteristic curve (AUC), for which higher values represent better discrimination <sup>14</sup>. Discrimination is the ability of a model to distinguish high-risk individuals from low-risk individuals <sup>15</sup>. The *C* statistic is the most commonly used performance measure of generalized linear regression models.<sup>16</sup>

2) Correlation: the correlation between the observed outcome (readmission) and the value predicted by the model <sup>17</sup>. Unlike the *C* statistic, correlation represents a summary measure of the predictive power of a generalized linear model.

3) Coefficient of discrimination (*D*): the absolute value of the difference between model successes (the mean predicted probability of readmission,  $p^{\wedge}$ , for readmitted patients) and model failures ( $p^{\wedge}$  for non-readmitted patients) <sup>18</sup>. This is a measure of overall model performance with a more intuitive interpretation for binary outcomes than the more familiar coefficient of determination ( $R^2$ ).

4) Brier score: the mean squared deviation between the predicted probability of readmission and the observed readmission rate. An overall score that captures both calibration and discrimination aspects, the Brier score can range from 0 for a perfect model to 0.25 for a noninformative model with a 50% incidence of the outcome. When the outcome incidence is lower, the maximum score for a noninformative model is lower <sup>16,19</sup>.

5) Scaled Brier score: the Brier score scaled by its maximum score ( $Brier_{max}$ ) according to the equation  $1 - Brier\ score / Brier_{max}$  <sup>16,20</sup>. Unlike the Brier score, the scaled Brier score is not dependent on the incidence of the outcome. For the scaled Brier score, a higher score represents greater accuracy.  $Brier_{max}$  is defined as  $mean(p) * (1 - mean(p))$  where  $mean(p)$  is the average probability of a positive outcome. The scaled Brier score is similar to Pearson's  $R^2$  statistic.<sup>16</sup> The Brier score and scaled Brier score were chosen as measures to highlight potential differences seen when the incidence of readmission varies due to the sampling methods described below.

Sampling was performed by two methodologies. The first method included only the first index discharge per patient during the study period (first discharges). The second method included all index discharges per patient (all discharges), regardless of whether the hospitalization was a readmission relative to a prior discharge. The study sample was then divided randomly into a training sample and a validation sample <sup>15</sup>. The training sample, which comprised 60% of the patients in the study cohort, was used to develop the statistical prediction models. The validation sample contained the remaining 40% of the patients and was used to evaluate the performance of the prediction models.

Characteristics of the study population were described and compared between the training and validation samples. Categorical variables were presented as number (%) while continuous variables were presented as mean (standard deviation) or median (interquartile range). For the first discharge per patient dataset, the validation sample was compared to the training sample by Chi-square tests for categorical variables and two sample t-tests or Wilcoxon rank-sum tests for continuous variables. For the all discharges dataset, the validation sample was compared to the training sample by univariate generalized linear model for all variables. When analyzing all discharges, only current and prior observations available at the time of each index discharge were used for modeling.

The models can be described in mathematical terms as follows. Suppose the  $i^{\text{th}}$  patient has  $n_i$  observations where  $i=1, 2, \dots, N$  and  $j^{\text{th}}$  discharge where  $j=1, 2, \dots, n_i$ . Suppose  $X_{ij}$  is the 46-vector of covariates and  $Y_{ij}$  is the vector of discharges where  $Y_{ij}=1$  if  $i^{\text{th}}$  subject at  $j^{\text{th}}$  discharge readmitted within 30 days and  $Y_{ij}=0$  otherwise.  $X_{ij}$  can be constant over time such as gender or time-varying such as age. We further define a general class of models that specify the potential relation between readmission  $Y$  and covariates  $X$  as  $f(E(Y | X)) = h$ , where  $f(\cdot)$  is a link function, such as the logit function, that determines the relationship between  $Y$  and  $X$ ;  $E(Y | X)$  denotes the conditional mean of  $Y$  given  $X$ ; and  $h$  is a function of covariates, usually a linear function such that  $h = \beta + X$  where  $\beta$  and  $\beta$  are log odds ratios. We fit the readmission prevalence model assuming a logit link to estimate the effect of covariates as  $\text{logit}(Y|X) = \beta + X$ . Then the probability of readmission within 30 days is  $P((Y=1)=\exp(\beta + X)/[1+\exp(\beta + X)]$ . This probability is estimated and used to compare performance of four statistical approaches described below.

The four approaches used to predict readmission were: 1) logistic regression using the first discharges, 2) logistic regression using all discharges, 3) GEE logistic regression with an exchangeable correlation structure using all discharges, and 4) CWGEE logistic regression with an exchangeable correlation structure using all discharges. CWGEE logistic regression is an extension of GEE logistic regression that accounts for cluster size when the outcome among observations in a cluster is dependent on the cluster size (i.e., when cluster size is informative)<sup>11</sup>. For each approach, univariate analyses were performed for all variables to determine those associated with 30-day readmission ( $P<0.1$ ). Multivariable models with best subset selection were performed to determine the adjusted associations of the variables with all-cause 30-day readmission<sup>21,22</sup>. Variables associated with 30-day readmission at the  $P<0.05$  level in the multivariable models were retained.

To examine the effects of sample size on model performance, we conducted resampling studies across a range of sample sizes from 2 000 to 17 000 patients by intervals of 1 000. We randomly sampled each subset from the complete cohort of 17 284 patients. Each subset was then randomly divided 60% for model development and 40% for validation. For each dataset, the models were developed and compared as described above. Changes in model performance measures over sample size are displayed by line charts and compared by analysis of covariance. Lastly, we examined the relationship between the number of discharges per patient (cluster size) and the number of readmissions per patient by Pearson

correlation. In addition, correlations between the number of discharges per patient and predicted readmission rates were assessed. All statistical analyses were conducted using SAS 9.4 (SAS Institute, Cary, NC) and Stata 14.0 (StataCorp, College Station, TX). Institutional Review Board approval was obtained from Boston Medical Center and Temple University.

## Results

There were 17 284 patients with 44 203 discharges, of which 9 034 (20.4%) were associated with 30-day readmission for any cause. The study cohort is well-distributed across middle to older adult age and sex (See Supplementary Table 1, Additional File 1). A majority of the patients are unmarried, English-speaking, insured by Medicare or Medicaid, lived within 5 miles of the hospital, educated at a high-school level or less, disabled, retired, or unemployed, and overweight or obese. This is an ethnically diverse sample, with 37.4% black, 14.7% Hispanic, and 39.7% white. The distribution of characteristics is similar between training and validation sets, with only 3 variables showing statistically significant differences despite absolute differences of <2% (race/ethnicity [ $P=0.03$ ], serum sodium [ $P=0.01$ ], and COPD or asthma [ $P=0.049$ ]). Likewise, the distribution of characteristics is similar between training and validation sets of all discharges, with only 2 variables showing statistically significant differences despite absolute differences of <2% (pre-admission sulfonylurea use [ $P=0.032$ ], and serum sodium [ $P=0.029$ ], See Supplementary Table 2, Additional File 2).

The readmission rate is 10.2% in the sample of first discharges and 20.3% in the sample of all discharges. The AUCs are comparable among logistic regression with first discharges, logistic regression using all discharges, and GEE using all discharges, ranging from 0.803-0.821 (See Supplementary Table 3, Additional File 3, which compares performance of the four modeling methods in the whole cohort). CWGEE using all discharges yielded a lower AUC than logistic regression with first discharges and logistic regression with all discharges. Logistic regression with all discharges resulted in the highest coefficient of discrimination, followed by the GEE, first discharge and WGEE models. Analyzing all discharges both with and without GEE produced the highest predictive power by correlation, while the first discharge showed the least predictive power. Analysis of the first discharges yielded the smallest (best) Brier score while the other three methods yielded comparably higher (worse) Brier scores. Analysis of the first discharges generated the smallest (worst) scaled Brier score, which is borderline different from the largest (best) scaled Brier score produced by logistic regression with all discharges.

The observed rates of readmission were stable with increasing sample size from 2 000 to 17 000 patients for the first discharges and all discharges samples. With sample sizes less than 6 000, the AUCs, coefficients of discrimination, correlations and Brier scores were variable (Figures 1-4). These measures of model performance were relatively stable at sample sizes of 6 000 to 17 000. The scaled Brier scores were relatively stable at sample sizes of 8 000 to 17 000. The AUCs were comparable among the four approaches, with the first discharges models generally having the greatest AUC and CWGEE having the smallest AUC over the sample size range (Figure 1). The coefficients of discrimination were greatest for the analysis of all discharges using logistic regression and generally smallest with CWGEE and first

discharges (Figure 2). The correlations were highest for logistic regression with all discharges and lowest for the first discharges (Figure 3). The first discharges analysis yielded the lowest (best) Brier score by a substantial margin compared to the other 3 methods, which were comparable across all samples sizes (Figure 4). In contrast, the first discharges approach generally produced the lowest (worst) scaled Brier score and all discharges logistic regression yielded the highest (best) across samples sizes (Figure 5).

There were no statistically significant readmission rate changes over sample size for both the first discharges ( $P=0.48$ ) and all discharges ( $P=0.24$ ). The slopes for AUC, coefficients of discrimination, correlations, Brier score, and scaled Brier score were flat for both first discharges and all discharges logistic regression (See Supplementary Table 4, Additional File 4, which compares change over sample size across the model performance measures). However, the slopes for these measures were significantly different from zero for GEE and CWGEE. As sample size increases, the AUC, coefficients of discrimination, correlation measures, and scaled Brier scores increase for GEE and CWGEE, indicating more predictive accuracy. Similarly, the Brier scores decrease for GEE and CWGEE as the sample size increases, indicating greater predictive accuracy.

The distribution of discharges per patient is presented in Supplementary Table 5, Additional File 5. There are 919 (13.3%) patients with one readmission, 686 (9.9%) with 2 or more readmissions. As expected, the number of discharges is positively correlated with the number of readmissions ( $r=0.87$ ,  $P<0.001$ ). The correlation coefficients between number of discharges and predicted readmission rates for all discharges logistic regression, GEE and CWGEE are 0.44, 0.38 and 0.31, respectively. Furthermore, the risk of readmission increases as the number of discharges increases ( $r=0.45$ ,  $P<0.001$ , Figure 6).

## Discussion

This study explored the impact of different sampling and multivariable modeling approaches on readmission risk prediction. We found that sampling only the first discharge per patient substantially underestimates the readmission rate relative to all discharges and is associated with misleading measures of model performance, particularly the Brier score. As expected, the number of readmissions per patient is highly correlated with the number of discharges per patient. In resampling studies across a range of sample sizes, most measures of model performance are unstable below a sample size of 6 000 and stabilize at sample sizes of 6 000 or greater. We also found that the accuracy of GEE and CWGEE is better in larger samples than in smaller samples. Lastly, we showed that using modeling approaches that account for clustering (GEE) and cluster size (CWGEE) yield generally more conservative measures of model performance than logistic regression.

Multiple methods for sampling hospitalizations and analyzing risk factors are used in studies that model predictors of readmission risk. Sampling only the first discharge per patient is a commonly used approach<sup>23-33</sup>. This approach eliminates clusters of multiple hospitalizations per patient, enabling the valid analysis of such data by logistic regression. One disadvantage of sampling the first discharges, however, is that informative data from subsequent discharges are excluded. We show that models based

on all discharges perform better on most measures of model performance than models based on first discharges. A second disadvantage of sampling only the first discharges is that the observed readmission rate is substantially lower than the readmission rate observed for all discharges during the study. In our study, the readmission rate in the sample with all discharges was nearly double the readmission rate in the first discharges sample. This occurs because excluding repeat discharges per patient minimizes the influence of patients with multiple discharges, who are at higher risk of readmission.

Such variation in observed readmission rates by different sampling methods is reflected in the literature. For example, among studies modeling readmission risk factors of patients with diabetes, we identified 14 English-language papers on PubMed published before 2019 that reported models of all-cause 30-day readmission risk<sup>7,12,23-29,34-38</sup>. Among the 7 studies that sampled all discharges in their respective datasets<sup>7,12,34-38</sup>, the mean readmission rate was 18.9% with a range of 16.0 – 21.5%. Among the 7 studies that sampled only first discharges<sup>23-29</sup>, the mean readmission rate was 13.5% with a range of 10.0 – 17.1%. Given that in clinical practice a provider may be treating a patient experiencing their first hospitalization or one of many hospitalizations, we believe it is more generalizable to sample all discharges for analysis so the entire experience of the study population is captured. Although studies that sample only first discharges can provide insights into readmission risk factors, reports of readmission rates in these samples should not be broadly interpreted to represent the readmission rate for the target population. We believe this is an important methodological insight that has not been previously published.

As with approaches to sampling hospitalizations, multiple methods of modeling readmission risk are used. The most common multivariable modeling approach by far is logistic regression<sup>5,39</sup>. Interestingly, logistic regression is utilized in some studies that sample all discharges without acknowledging the probable violation of the independence assumption<sup>13,35,38,40</sup>. In addition, some studies of readmission risk factors that utilize logistic regression do not report how hospitalizations were sampled<sup>41-43</sup>. The most often utilized statistical modeling approach to account for correlations within clustered hospitalization data is GEE<sup>7,12,34,44</sup>. We are unaware of any readmission risk models that employ CWGEE.

While some studies have compared readmission risk models generated by different methods<sup>45,46</sup>, few have explored broader methodologic issues involved with analyzing readmissions. One study found that the LACE+ model, which was derived in a patient-level sample based on one randomly selected hospitalization per patient, performed worse in the parent sample that included all hospitalizations of the patients<sup>47</sup>. The authors concluded that frequent hospital utilizers probably have characteristics that were not adequately captured in the patient-level model and that capturing these characteristics may improve readmission models. Another study showed that readmission risk model performance varied by restricting samples to different reasons for readmission, that different types of data (visit history and laboratory results) contributed more predictive value than other types of data, and that limiting the cohort

to patients whose index admission and readmission diagnoses matched was associated with worse model performance compared with a cohort that did not match admission and readmission diagnoses<sup>48</sup>. Lastly, simulation studies outside the readmission literature show that analytical approaches on clustered data that do not take into account the effects of clustering often yield erroneous results<sup>8,9</sup>.

Our study has some limitations. First, 30-day readmissions that may have occurred at other hospitals were not captured. However, in our cohort the 30-day readmission rate was 20.4%, which is one of the highest readmission rates reported among patients with diabetes. Therefore, it seems unlikely that a significant number of patients were readmitted elsewhere. Second, no external validation was conducted because the data were drawn from a single center. Although it is possible that the specific models generated in our study population would yield different results in other populations, the types of administrative and clinical data analyzed are likely to be generalizable to most hospitals. Therefore we believe the general concepts revealed by our findings are broadly applicable to analyses of readmissions in hospital cohorts across other settings and other chronic conditions than diabetes. Finally, we did not assess calibration by the commonly used Hosmer-Lemeshow goodness-of-fit test<sup>49</sup>. The standard approach with this test compares observed and predicted outcomes by decile of predicted probability. However, this test can give different results depending on the number of groups used<sup>49</sup>. In addition, performance of the test varies by sample size<sup>49,50</sup>, and 4 of the 5 measures of model performance we used incorporate calibration.

These limitations are balanced by a number of strengths. We analyzed a large enough cohort to examine performance of different modeling approaches across a broad range of sample sizes. Furthermore, this was a population well-characterized on 46 different variables used for modeling procedures. In addition, our analyses on the effects of different sampling and modeling approaches on model performance are novel in the field of readmission risk prediction.

## Conclusions

This study demonstrates the impact of different methodological approaches to analyzing hospital readmission data. All discharges in a cohort of hospitalized patients should be analyzed because sampling only the first discharge per patient produces low population estimates of readmission rates and misleading results with Brier scores. Studies should therefore transparently report how discharges are sampled. In addition, the number of readmissions per patient is highly correlated with the number of discharges per patient. Not surprisingly, modeling methods that account for clustering and cluster size yield different, generally more conservative, estimates of model performance. Researchers should be aware of the pitfalls associated with these measures of model performance and modeling procedures. Future studies of readmission risk may generate more valid results if they include all discharges and utilize modeling methods that account for clustered data.

## Abbreviations

Cluster-weighted generalized estimating equations (CWGEE); Generalized estimating equations (GEE); International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM); Logistic regression using the first discharge per patient (LR-first); Logistic regression using all discharges (LR-all); Area under the receiver operating characteristic curve (AUC)

## **Declarations**

### **Ethics approval and consent to participate:**

Institutional Review Board approval was obtained from Boston Medical Center (H-31845) and Temple University (20658). A waiver of informed consent was provided for this retrospective analysis of a limited data set.

### **Consent for publication:**

Not applicable.

### **Availability of data and materials:**

The data that support the findings of this study are available from Boston Medical Center but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of Boston Medical Center.

### **Competing interests:**

The authors declare no conflict of interest.

### **Funding:**

This work was supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health (Grant K23DK102963 to D.J.R.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

### **Authors' contributions:**

HZ conducted the data analysis, interpretation of results, and was a major contributor to the writing of the manuscript. ST conducted the literature review and was a major contributor to the writing of the manuscript. SHG and SF helped planned the study and critically reviewed the manuscript for important intellectual content. DJR planned and supervised the study and was a major contributor to the writing of the manuscript. All authors read and approved the final manuscript.

### **Acknowledgements:**

An abstract of this work was presented as a poster at the 78<sup>th</sup> Scientific Sessions of the American Diabetes Association in June, 2018, Orlando, Florida.

### Authors' information (optional):

Not applicable.

## References

1. Benbassat J, Taragin M. Hospital readmissions as a measure of quality of health care: advantages and limitations. *Arch Intern Med*. 2000;160(8):1074-1081.
2. Kocher RP, Adashi EY. Hospital readmissions and the affordable care act: Paying for coordinated quality care. *JAMA*. 2011;306(16):1794-1795.
3. Leppin AL, Brito JP, Mair FS, et al. Preventing 30-Day Hospital Readmissions: A Systematic Review and Meta-analysis of Randomized Trials. *JAMA Internal Medicine*. 2014;174(7):1095-1107.
4. Kansagara D, Englander H, Salanitro A, et al. Risk Prediction Models for Hospital Readmission: A Systematic Review. *JAMA*. 2011;306(15):1688-1698.
5. Zhou H, Della PR, Roberts P, Goh L, Dhaliwal SS. Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review. *BMJ Open*. 2016;6(6):e011060.
6. Eby E, Hardwick C, Yu M, et al. Predictors of 30 day hospital readmission in patients with type 2 diabetes: a retrospective, case-control, database study. *Curr Med Res Opin*. 2015;31(1):107-114.
7. Rubin DJ, Handorf EA, Golden SH, Nelson DB, McDonnell ME, Zhao H. Development and Validation of a Novel Tool to Predict Hospital Readmission Risk Among Patients With Diabetes. *Endocr Pract*. 2016;22(10):1204-1215.
8. Bouwmeester W, Twisk JW, Kappen TH, van Klei WA, Moons KG, Vergouwe Y. Prediction models for clustered data: comparison of a random intercept and standard regression model. *BMC Med Res Methodol*. 2013;13:19.
9. Galbraith S, Daniel JA, Vissel B. A Study of Clustered Data and Approaches to Its Analysis. *The Journal of Neuroscience*. 2010;30(32):10601-10608.
10. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika*. 1986;73(1):13-22.
11. Williamson JM, Datta S, Satten GA. Marginal analyses of clustered data when cluster size is informative. *Biometrics*. 2003;59(1):36-42.
12. Rubin DJ, Recco D, Turchin A, Zhao H, Golden SH. EXTERNAL VALIDATION OF THE DIABETES EARLY RE-ADMISSION RISK INDICATOR (DERRI()). *Endocr Pract*. 2018;24(6):527-541.
13. Donzé J, Aujesky D, Williams D, Schnipper JL. Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model. *JAMA Internal Medicine*. 2013;173(8):632-638.

14. Chambless LE, Diao G. Estimation of time-dependent area under the ROC curve for long-term risk prediction. *Stat Med*. 2006;25(20):3474-3486.
15. Altman DG, Vergouwe Y, Royston P, Moons KGM. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338:1432-1435.
16. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. 2010;21(1):128-138.
17. Zheng B, Agresti A. Summarizing the predictive power of a generalized linear model. *Stat Med*. 2000;19(13):1771-1781.
18. Tjur T. Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination. *The American Statistician*. 2009;63(4):366-372.
19. BRIER GW. VERIFICATION OF FORECASTS EXPRESSED IN TERMS OF PROBABILITY. *Monthly Weather Review*. 1950;78(1):1-3.
20. Wu Y-C, Lee W-C. Alternative Performance Measures for Prediction Models. *PLOS ONE*. 2014;9(3):e91249.
21. Furnival GM, Wilson RW, Jr. Regressions by Leaps and Bounds. *Technometrics*. 1974;16(4):499-511.
22. Hanley JA, Negassa A, Edwardes MD, Forrester JE. Statistical Analysis of Correlated Data Using Generalized Estimating Equations: An Orientation. *American journal of epidemiology*. 2003;157(4):364-375.
23. Bennett KJ, Probst JC, Vyavaharkar M, Glover SH. Lower rehospitalization rates among rural Medicare beneficiaries with diabetes. *J Rural Health*. 2012;28(3):227-234.
24. Collins J, Abbass IM, Harvey R, et al. Predictors of all-cause-30-day-readmission among Medicare patients with type 2 diabetes. *Curr Med Res Opin*. 2017;33(8):1517-1523.
25. Eby E, Hardwick C, Yu M, et al. Predictors of 30 day hospital readmission in patients with type 2 diabetes: a retrospective, case–control, database study. *Current Medical Research and Opinion*. 2015;31(1):107-114.
26. Healy SJ, Black D, Harris C, Lorenz A, Dungan KM. Inpatient diabetes education is associated with less frequent hospital readmission among patients with poor glycemic control. *Diabetes Care*. 2013;36(10):2960-2967.
27. King C, Atwood S, Lozada M, et al. Identifying risk factors for 30-day readmission events among American Indian patients with diabetes in the Four Corners region of the southwest from 2009 to 2016. *PLoS One*. 2018;13(8):e0195476.
28. Png ME, Yoong J, Chen C, et al. Risk factors and direct medical cost of early versus late unplanned readmissions among diabetes patients at a tertiary hospital in Singapore. *Curr Med Res Opin*. 2018:1-10.
29. Raval AD, Zhou S, Wei W, Bhattacharjee S, Miao R, Sambamoorthi U. 30-Day Readmission Among Elderly Medicare Beneficiaries with Type 2 Diabetes. *Population Health Management*. 2015;18(4):256-264.

30. Billings J, Dixon J, Mijanovich T, Wennberg D. Case finding for patients at risk of readmission to hospital: development of algorithm to identify high risk patients. *BMJ*. 2006;333(7563):327.
31. Howell S, Coory M, Martin J, Duckett S. Using routine inpatient data to identify patients at risk of hospital readmission. *BMC Health Serv Res*. 2009;9:96.
32. Krumholz HM, Wang K, Lin Z, et al. Hospital-Readmission Risk – Isolating Hospital Effects from Patient Effects. *New England Journal of Medicine*. 2017;377(11):1055-1064.
33. Silverstein MD, Qin H, Mercer SQ, Fong J, Haydar Z. Risk factors for 30-day hospital readmission in patients  $\geq 65$  years of age. *Proc (Bayl Univ Med Cent)*. 2008;21(4):363-372.
34. Albrecht JS, Hirshon JM, Goldberg R, et al. Serious Mental Illness and Acute Hospital Readmission in Diabetic Patients. *American Journal of Medical Quality*. 2012;27(6):503-508.
35. Enomoto LM, Shrestha DP, Rosenthal MB, Hollenbeak CS, Gabbay RA. Risk factors associated with 30-day readmission and length of stay in patients with type 2 diabetes. *J Diabetes Complications*. 2017;31(1):122-127.
36. Holscher CM, Hicks CW, Canner JK, et al. Unplanned 30-day readmission in patients with diabetic foot wounds treated in a multidisciplinary setting. *J Vasc Surg*. 2018;67(3):876-886.
37. Ostling S, Wyckoff J, Ciarkowski SL, et al. The relationship between diabetes mellitus and 30-day readmission rates. *Clinical Diabetes and Endocrinology*. 2017;3(1):3.
38. Robbins JM, Webb DA. Diagnosing diabetes and preventing rehospitalizations: the urban diabetes study. *Med Care*. 2006;44(3):292-296.
39. Kansagara D, Englander H, Salanitro A, et al. Risk prediction models for hospital readmission: a systematic review. *JAMA : the journal of the American Medical Association*. 2011;306(15):1688-1698.
40. Emons MF, Bae JP, Hoogwerf BJ, Kindermann SL, Taylor RJ, Nathanson BH. Risk factors for 30-day readmission following hypoglycemia-related emergency room and inpatient admissions. *BMJ Open Diabetes Research & Care*. 2016;4(1).
41. Caughey GE, Pratt NL, Barratt JD, Shakib S, Kemp-Casey AR, Roughead EE. Understanding 30-day re-admission after hospitalisation of older patients for diabetes: identifying those at greatest risk. *Med J Aust*. 2017;206(4):170-175.
42. Najafian A, Selvarajah S, Schneider EB, et al. Thirty-day readmission after lower extremity bypass in diabetic patients. *The Journal of surgical research*. 2016;200(1):356-364.
43. Escobar GJ, Ragins A, Scheirer P, Liu V, Robles J, Kipnis P. Nonelective Rehospitalizations and Postdischarge Mortality: Predictive Models Suitable for Use in Real Time. *Med Care*. 2015;53(11):916-923.
44. Karunakaran A, Zhao H, Rubin DJ. Pre-discharge and Post-discharge Risk Factors for Hospital Readmission Among Patients With Diabetes. *Med Care*. 2018;56(7):634-642.
45. Welchowski T, Schmid M. A framework for parameter estimation and model selection in kernel deep stacking networks. *Artificial Intelligence in Medicine*. 2016;70:31-40.

46. Povalej Brzan P, Obradovic Z, Stiglic G. Contribution of temporal data to predictive performance in 30-day readmission of morbidly obese patients. *PeerJ*. 2017;5:e3230.
47. van Walraven C, Wong J, Forster AJ, Hawken S. Predicting post-discharge death or readmission: deterioration of model performance in population having multiple admissions per patient. *Journal of Evaluation in Clinical Practice*. 2013;19(6):1012-1018.
48. Walsh C, Hripcsak G. The effects of data sources, cohort selection, and outcome definition on a predictive model of risk of thirty-day hospital readmissions. *Journal of biomedical informatics*. 2014;52:418-426.
49. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness-of-fit tests for the logistic regression model. *Stat Med*. 1997;16(9):965-980.
50. Kramer AA, Zimmerman JE. Assessing the calibration of mortality benchmarks in critical care: The Hosmer-Lemeshow test revisited. *Crit Care Med*. 2007;35(9):2052-2056.

## Additional Files

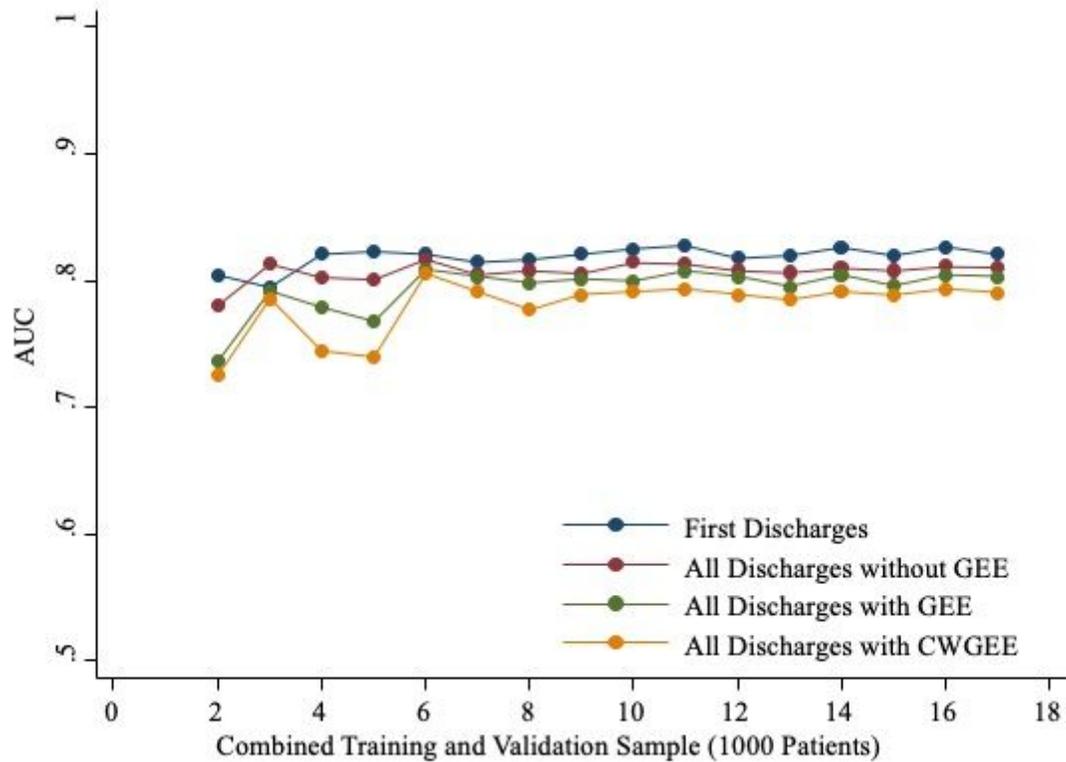
Additional file 1\_Table 1.docx: Supplementary Table 1. Characteristics of Patients With Diabetes at the Time of the First Hospital Discharge by Training and Validation Samples, Boston, Massachusetts, 2004-2012.

Additional file 2\_Table 2.docx: Supplementary Table 2 - Characteristics of hospitalized patients with diabetes for all discharges by training and validation samples

Additional file 3\_Table 3.docx: Supplementary Table 3. Performance of Logistic Regression Using First Discharges (n=6913), All Discharges (n=17801), All Discharges With GEE, and All Discharges With CWGEE to Predict All-Cause 30-Day Readmission Among Adults With Diabetes, Boston, Massachusetts, 2004-2012.

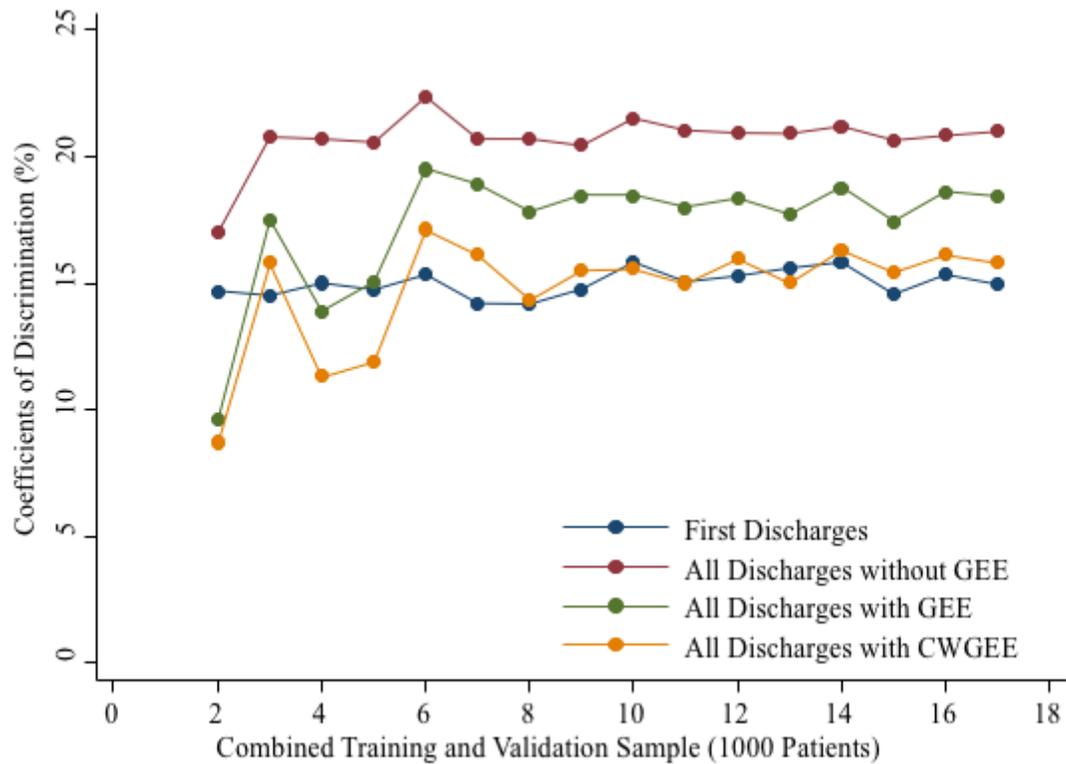
Additional file 4\_Table 4.docx: Supplementary Table 4. Change Over Sample Size of Logistic Regression Model Performance Measures Using First Discharges (n=6913), All Discharges (n=17801), All Discharges With GEE, and All Discharges With WGEE to Predict All-Cause 30-Day Readmission in Validation Sample of Adults With Diabetes, Boston, Massachusetts, 2004-2012.

## Figures



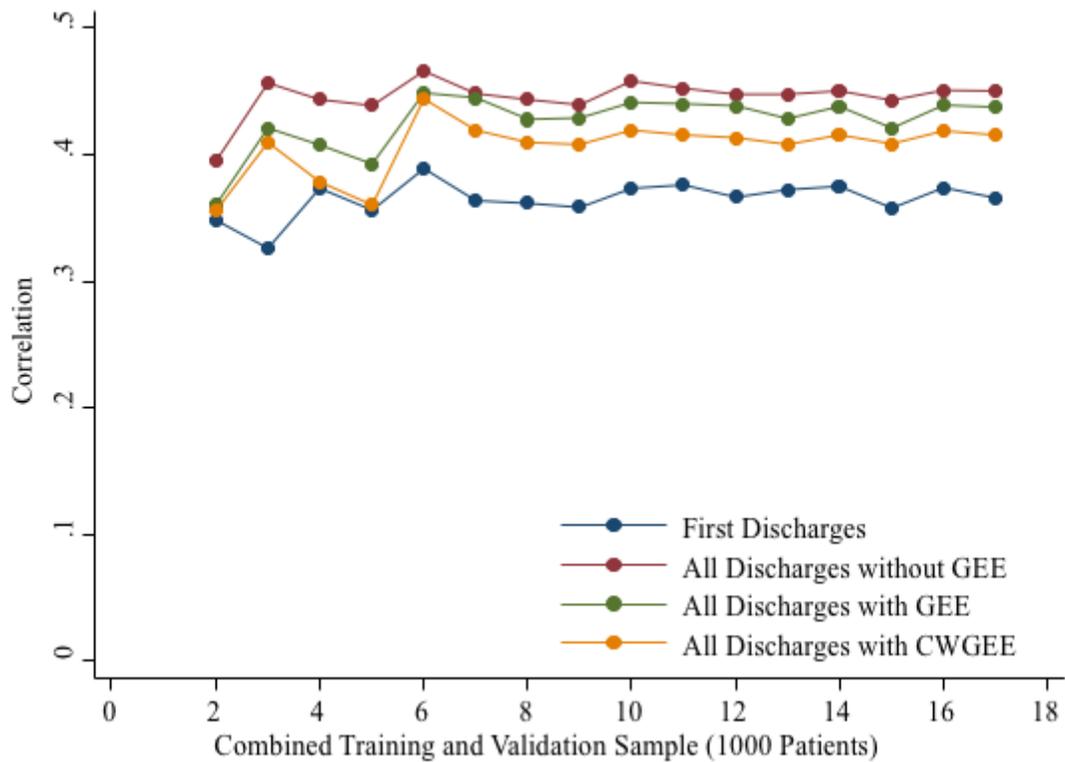
**Figure 1**

AUC of models tested in validation samples of adult patients with diabetes. The x-axis shows the size of the initial cohort from which training and validation samples were drawn. Boston, Massachusetts, 2004-2012. GEE = generalized estimating equations. CWGEE = cluster-weighted generalized estimating equations. AUC = Receiver operating characteristic area under the curve



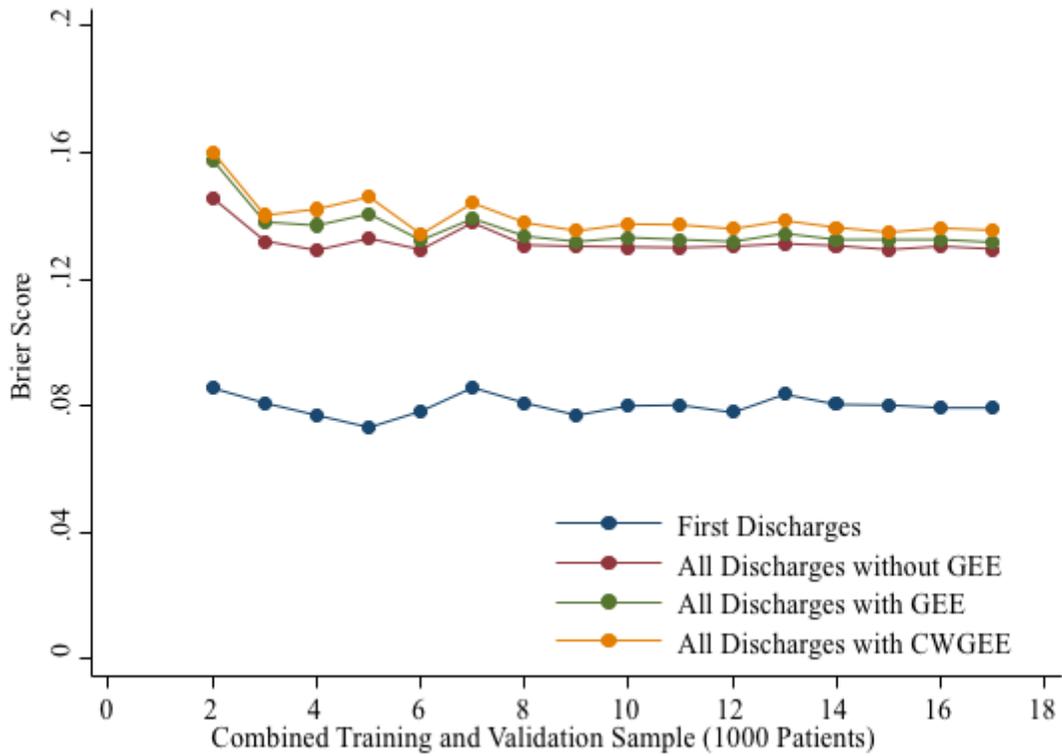
**Figure 2**

Coefficients of discrimination of models tested in validation samples of adult patients with diabetes. The x-axis shows the size of the initial cohort from which training and validation samples were drawn. Boston, Massachusetts, 2004-2012. GEE = generalized estimating equations. CWGEE = cluster-weighted generalized estimating equations.



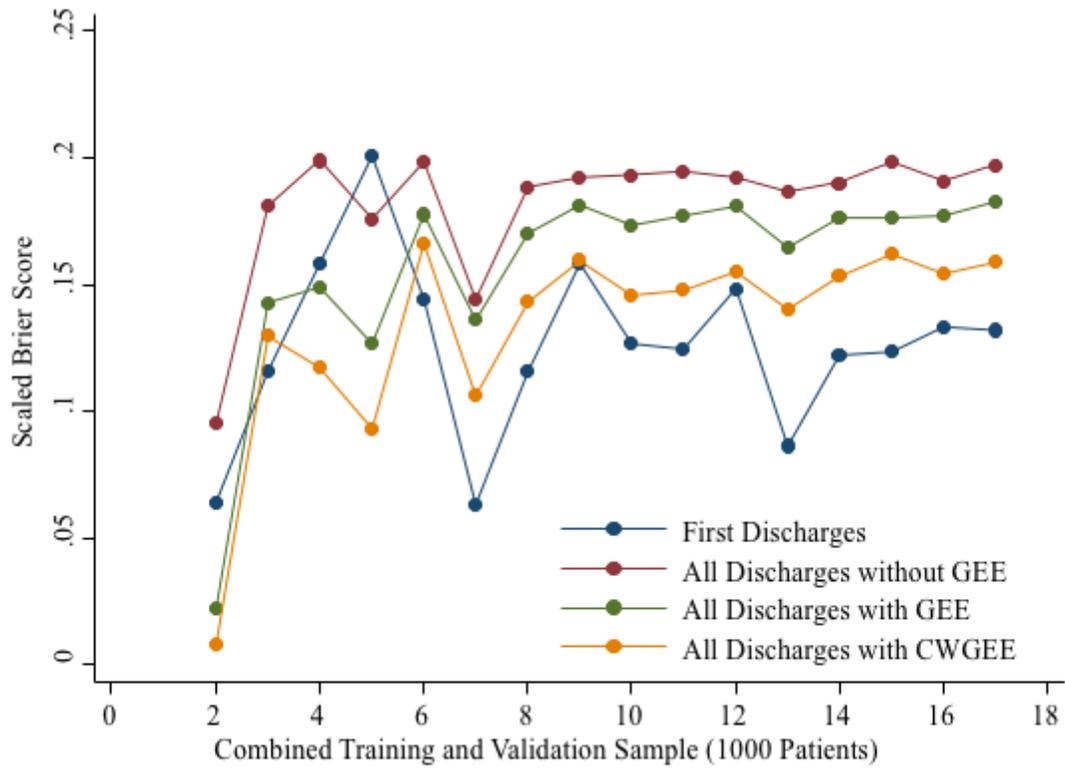
**Figure 3**

Correlation measures of models tested in validation samples of adult patients with diabetes. The x-axis shows the size of the initial cohort from which training and validation samples were drawn. Boston, Massachusetts, 2004-2012. GEE = generalized estimating equations. CWGEE = cluster-weighted generalized estimating equations.



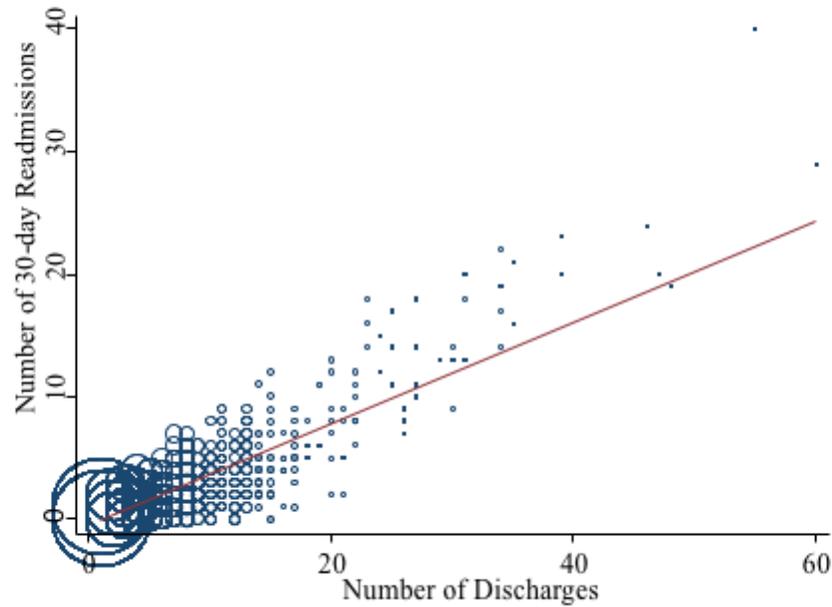
**Figure 4**

Brier scores of models tested in validation samples of adult patients with diabetes. The x-axis shows the size of the initial cohort from which training and validation samples were drawn. Boston, Massachusetts, 2004-2012. GEE = generalized estimating equations. CWGEE = cluster-weighted generalized estimating equations.



**Figure 5**

Scaled Brier scores of models tested in validation samples of adult patients with diabetes. The x-axis shows the size of the initial cohort from which training and validation samples were drawn. Boston, Massachusetts, 2004-2012. GEE = generalized estimating equations. CWGEE = cluster-weighted generalized estimating equations.



**Figure 6**

Correlation of number of 30-day readmissions with number of discharges per patient. The size of each circle corresponds to the number patients with a given number of discharges and readmission, with the smallest circle representing 1 patient and the largest representing 3,852 patients.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1Table1.docx](#)
- [Additionalfile2Table2.docx](#)
- [Additionalfile3Table3.docx](#)
- [Additionalfile4Table4.docx](#)
- [Additionalfile5Table5.pdf](#)