Observational Study

Medicine®

OPEN

Template matching for benchmarking hospital performance in the veterans affairs healthcare system

Brenda M. Vincent, MS[a,*], Wyndy L. Wiitala, PhD[a], Kaitlyn A. Luginbill, MPH[a], Daniel J. Molling, MS[a], Timothy P. Hofer, MD[a,b], Andrew M. Ryan, PhD[c], Hallie C. Prescott, MD[a,b]

Abstract

Comparing hospital performance in a health system is traditionally done with multilevel regression models that adjust for differences in hospitals' patient case-mix. In contrast, "template matching" compares outcomes of similar patients at different hospitals but has been used only in limited patient settings.

Our objective was to test a basic template matching approach in the nationwide Veterans Affairs healthcare system (VA), compared with a more standard regression approach.

We performed various simulations using observational data from VA electronic health records whereby we randomly assigned patients to "pseudo hospitals," eliminating true hospital level effects. We randomly selected a representative template of 240 patients and matched 240 patients on demographic and physiological factors from each pseudo hospital to the template. We varied hospital performance for different simulations such that some pseudo hospitals negatively impacted patient mortality.

Electronic health record data of 460,213 hospitalizations at 111 VA hospitals across the United States in 2015.

We assessed 30-day mortality at each pseudo hospital and identified lowest quintile hospitals by template matching and regression. The regression model adjusted for predicted 30-day mortality (as a measure of illness severity).

Regression identified the lowest quintile hospitals with 100% accuracy compared with 80.3% to 82.0% for template matching when systematic differences in 30-day mortality existed.

The current standard practice of risk-adjusted regression incorporating patient-level illness severity was better able to identify lower-performing hospitals than the simplistic template matching algorithm.

Abbreviations: APACHE = acute physiology and chronic health evaluation, IPEC = inpatient evaluation center, IQR = interquartile range, SAIL = strategic analytics for improvement and learning, SMR = standardized mortality ratio, VA = veterans affairs.

Keywords: health care research, outcomes research, quality of care

1. Introduction

Hospital performance measurement seeks to identify and remediate low-quality care. However, because patients are not randomly allocated to hospitals, cross-hospital comparisons are limited by differences in patient case-mix and illness severity. The US Center for Medicare and Medicaid Services, the Veterans Affairs (VA), and other programs sponsors typically adjust for differences in patient case-mix using regression models for indirect standardization.[1–4] However, clinicians often consider these models: unfair because of enduring concerns about residual confounding; unclear because clinicians have limited expertise in interpreting and applying the findings from complex multi-level risk-adjusted regression models; and unhelpful because the models do not reveal why, or in which patients, outcomes differ across hospitals. As a result, the National Academy of Medicine has called for greater transparency and interpretability of hospital benchmarking systems.[5,6]

Template matching, a form of direct standardization, has been proposed to have greater fairness and transparency than conventional regression approaches used for case-mix adjustment.[7] In template matching, hospitals are compared on a set of patients with similar characteristics, thereby making benchmarking more credible.[7] A set of hospitalization profiles are identified (e.g., an 80-year-old nursing home resident with a history of heart failure who is hospitalized with acute renal failure and has a

Editor: Phil Phan.

This work was supported by grant IIR [IIR 17–219] from the U.S. Department of Veterans Affairs Health Services Research & Development Service. This paper does not necessarily reflect of the position or policy of the US government or Department of Veterans Affairs.

Preliminary results from this project were presented in an oral talk at the 2018 Joint Statistical Meetings on August 3, 2018 in Vancouver, Canada.

The authors declare no conflicts of interest.

Supplemental Digital Content is available for this article.

[a] Center for Clinical Management Research, Veterans Affairs Ann Arbor Healthcare System, [b] Department of Internal Medicine and Institute for Healthcare Policy and Innovation, [c] Department of Health Management and Policy, School of Public Health, University of Michigan, Ann Arbor, Michigan.

[*] Correspondence: Brenda M. Vincent, 2800 Plymouth Rd, North Campus Research Complex, Ann Arbor, MI 48109 (e-mail: Brenda.vincent@va.gov).

Medicine (2019) 98:20(e15644)

Received: 5 February 2019 / Received in final form: 26 March 2019 / Accepted: 16 April 2019

http://dx.doi.org/10.1097/MD.0000000000015644

creatinine of 2.5). Each hospitalization profile specifies all key risk attributes at the time of admission; several hundred hospitalization profiles aggregated into a template set. Hospitals are then compared directly with the outcomes of their own patients who best match this standard set of hospitalization profiles. Compared with regression models, template matching could be considered fairer (as hospitals are compared on a similar set of patients), clearer (no further risk-adjustment is required), and more helpful (hospitals can examine the specific hospitalizations to see where their outcomes lag).

However, template matching has been evaluated in only limited patient and hospital populations, focusing on select common surgical procedures or medical diagnoses.[7–9] It is unclear how template matching performs for assessing overall hospital quality in a real-world health care system. Furthermore, it is unknown whether template matching results similar benchmarking assessments than regression assessments incorporating patient's physiology (the current standard within the VA), and if not, which benchmarking approach is more accurate.

Thus, in this study, we sought to test a basic template matching approach for benchmarking hospital performance in the nationwide VA healthcare system, compared with the standard regression approaches currently in use. In contrast to prior work, we incorporated all hospitalizations (as opposed to sampling specific diagnoses or medical procedures) and used a simpler matching algorithm, as this approach would be easier to implement in practice. To assess the accuracy of template matching versus regression benchmarking assessments, we used simulation, in which the true hospital effect could be known with certainty.

2. Methods

2.1. Context

The VA healthcare system has been a leader in developing and implementing performance measurement methods. The VA was among the first healthcare systems to have a universal electronic patient record,[10] and to operationally measure and report risk-adjusted mortality using a risk-adjustment model that performs as well as APACHE IV (Acute Physiology and Chronic Health Evaluation score, Cerner Corporation).[11] The VA uses comprehensive performance assessment methods—Strategic Analytics for Improvement and Learning (SAIL)[12]—which incorporates a variety of metrics including risk-adjusted 30-day mortality. Hospitals are compared with their overall performance using a star rating system (from 1 to 5 stars). In this study, we focus on all-cause risk-adjusted 30-day mortality, and assign "stars" based on the quintiles in which each hospital falls for this metric.

2.2. Data source and cohort

We used clinical data from the 2015 VA Inpatient Evaluation Center (IPEC) dataset, which contains all hospitalizations in the VA healthcare system. We excluded hospitalizations for a primary psychiatric diagnosis and limited our sample of hospitals to those with at least 960 hospitalizations per year (to allow for a 4:1 matching ratio in the template matching procedure, which is described below). For each hospitalization, we generated a predicted 30-day mortality using variables contained in the VA's illness severity score[13] (age, sex, race, admission source, surgical indicator, top 20 admission diagnosis categories, 29 comorbid conditions, and 11 laboratory values drawn in the first 24 hours of hospitalization: sodium, blood urea nitrogen, glomerular filtration rate, glucose, albumin, bilirubin, white blood cell count, hematocrit, pH, $PaCO_2$, and PaO_2), as we have done in prior analyses.[14,15] Patients with missing laboratory measures were assumed to have a normal laboratory value. The c-statistic for the predicted mortality model was 0.856.

2.3. Measurement scenarios

We considered 8 simulation scenarios for assessing template matching versus conventional regression (Table 1). These scenarios are characterized by variation (or uniformity) in patient case-mix, hospital case volume, and hospital performance (modeled as systematic differences in mortality rate, as described below). First, we considered a baseline scenario in which there were no systematic differences in patient case-mix, case volume, or performance across hospitals. We simulated this scenario by allocating the patient hospitalizations in the dataset to "pseudo-hospitals" by random assignment to eliminate any true hospital level effects that may exist. Therefore, any true variations in 30-day mortality were due to patient-level factors and patient case-mix (and random variation) rather than hospital-level care.

In subsequent scenarios, we varied patient case-mix, hospital case volume, and hospital risk-adjusted mortality rates—both separately, and in combination. For case-mix variation, we caused illness severity or age to vary across pseudo-hospitals. We achieved this case-mix variation in the simulations by weighting

Table 1

Eight measurement scenarios, characterized by hospital case-mix, case-volume, and mortality.

Scenario	Hospital case-mix	Hospital case-volume	Hospital performance (30-day mortality)
1	Uniform	Uniform	Uniform
2a	Varies—by age	Uniform	Uniform
2b	Varies—by predicted mortality	Uniform	Uniform
3	Uniform	Uniform	Varies
4a	Varies—by age	Uniform	Varies
4b	Varies—by predicted mortality	Uniform	Varies
5	Uniform	Varies	Uniform
6a	Varies—by age	Varies	Uniform
6b	Varies—by predicted mortality	Varies	Uniform
7	Uniform	Varies	Varies
8a	Varies—by age	Varies	Varies
8b	Varies—by predicted mortality	Varies	Varies

the random allocation of hospitalizations to pseudo-hospitals by age (scenarios 2a, 4a, 6a, and 8a), and separately, by predicted 30-day mortality as a measure of illness severity (scenarios 2b, 4b, 6b, and 8b) (online Supplement, Appendix 1, http://links.lww.com/MD/C994). We chose this rather simplistic implementation of case-mix variability (weighting the allocation by just a single variable) over more complex approaches to maximize the transparency and interpretation of the simulations.

For scenarios in which we specify that hospital risk-adjusted mortality varies (scenarios 3, 4a, 4b, 7, 8a, and 8b), we did this by randomly generated a "pseudo-outcome." Conceptually, we randomly selected one-fifth of the pseudo-hospitals to be in the lowest-quintile (higher 30-day mortality rate), one-fifth of the pseudo-hospitals to be in the top quintile (lower 30-day mortality rate), and the remaining pseudo-hospitals as median hospitals. The pseudo-outcome was generated by adding a hospital level component of variance, generated from the geometric distribution, to the patients' real predicted mortality such that the top and lowest quintile pseudo-hospitals had a standardized mortality ratio (SMR) of approximately 2.0 and 0.5, respectively (Online Supplement, Appendix 2, http://links.lww.com/MD/C994).

2.4. Template matching procedure: selecting the optimal template

We selected a template size of 240 hospitalization profiles, since we estimated that this would provide 80% power to detect a SMR of ≥1.75 (Online Supplement, Appendix 3, http://links.lww.com/MD/C994). We sampled 240 hospitalizations from the entire population (at random, without replacement) 500 times to generate 500 potential templates. Following the approach of Silber et al,[7] we assessed the fit of each potential template to the overall population and selected the template that most resembled the patient population. We determined this by selecting the template that minimized the Mahalanobis distance,[16] a commonly used multivariate measure of the sum of squares of the difference of the means (in units of the standard deviation) with an adjustment for the correlation between variables.

2.5. Template matching procedure: matching hospitals to the template

Next, we matched hospitalizations at each hospital to the template based on 70 variables (Online Supplement, Appendix 4, http://links.lww.com/MD/C994) including demographic characteristics (age, sex, ethnicity), hospitalization characteristics (admission source, indicator for major surgery within 24 hours of admission), values for 11 laboratories collected within 24 hours of admission, predicted 30-day mortality, 29 Elixhauser comorbidities, admitting diagnosis and laboratory values at admission. We elected to include physiologic data (laboratories, predicted risk of death) since such data improve prediction of hospital mortality over administrative data alone.[17] Furthermore, incorporation of physiological data in template matching reduces variation in illness severity across matched hospital sample and results in different hospital rankings.[9]

We used SAS PSMATCH procedure[18] with an optimal matching algorithm, equal weighting of each matching variable, and no calipers (to ensure a full 240 hospitalizations would be matched to the template for each pseudo-hospital). This matching approach is simpler than the approach described by Silber et al[7] which used fine balance and multiple integer processing to ensure covariate balance across the matched cohorts. However, if effective, this simpler approach would be much easier to implement for benchmarking within the VA system.

2.6. Comparing template matching and regression in simulations

For each measurement scenario, we completed 1000 simulations. In each simulation, we: identified a template by sampling without replacement from the overall pool of hospitalizations; allocated patients to the pseudo-hospitals; matched hospitalizations from each pseudo-hospital to the template using the simple matching algorithm; assessed the performance of each pseudo-hospital using template matching and a regression model adjusting for predicted 30-day mortality; and compared the template matching versus regression performance assessments. Specifically, we were interested in the extent to which template matching and regression agreed on the classification of hospitals as being in the lowest quintile (to align with VA's star rating system). Furthermore, for measurement scenarios where there was a systematic difference in 30-day mortality between pseudo-hospitals (scenarios 3, 4a, 4b, 7, 8a, 8b), we assessed the extent to which template matching and regression correctly identified lowest-quintile pseudo-hospitals. All analyses were done in SAS version 9.4 (SAS Institute Inc., Cary, NC) and a *P*-value threshold of .05 was used to indicate statistical significance. The SAS code for these simulations is available at: https://github.com/CCMRcodes/TemplateMatching.

2.7. Comparing template matching and regression in real data

Additionally, we compared the performance of template matching and regression using 2015 IPEC data (preserving patients' true hospital assignment). The best template was selected from 500 potential templates and hospitalizations were matched from each hospital using the same techniques as in our simulations. We assessed the performance of each hospital using template matching and regression. We were interested in how often hospitals classified as in bottom quintile for 30-day mortality by regression were also classified as bottom quintile by template matching.

2.8. Assessing the selected templates and matching quality

We performed a post hoc assessment of the fit of the selected template to the population. Because we used a multivariate measure when selecting the template, there was no guarantee that the individual variables would be representative of the population. We assessed the distributions of 3 illustrative variables—30-day mortality (outcome), age, and proportion with admission through the emergency department—across all 500 potential templates with the population rates and highlight the selected template highlighted. We illustrate these distributions over 20 simulation iterations.

In a second post-hoc assessment, we evaluated the match quality across one iteration of the baseline scenario. We use the standardized mean difference[19] for each covariate, which is calculated as the difference in the means of the template and the matched groups divided by the standard deviation (template and matched pooled) of the covariate. If the standardized mean

difference <0.25, then a covariate is considered balanced. We use a box-and-whisker plot to display the distributions of the standardized mean difference across the 111 pseudo hospitals for each variable.

2.9. Post-match adjustment

We performed an additional post-hoc analysis where we used a post-match adjustment of variables due to poor performance of template matching in our initial simulations. We considered the most complex scenario (8b). A hierarchical logistic model adjusted for 30-day predicted mortality using patients matched to the template from our simulations. We included a random intercept identifying the template patient to which the observed patient was matched. Using this method, the outcomes of patients resembling each patient in the template was compared across each of the pseudo-hospitals, adjusting for differences in illness severity. This differs from our initial simulations, where all patients in the template are compared across pseudo-hospitals. This technique more closely aligns with what was done by Silber et al.[7]

3. Results

In 2015, there were 470,263 hospitalizations at 129 hospitals with a non-psychiatric principal diagnosis. After excluding hospitalizations that occurred at hospitals with fewer than 960 hospitalizations that calendar year, there were 460,213 hospitalizations at 111 hospitals available for the analysis. Only 2.1% of hospitalizations were at hospitals with <960 hospitalizations. Patient and hospitalization characteristics are provided in Table 2.

Table 2

Descriptive characteristics of patients and hospitalizations.

Hospitalizations, N	460,213
30-day mortality, N (%)	22,935 (5.0%)
Predicted mortality, median (IQR)	0.020 (0.008, 0.051)
Age, median (IQR)	66 (60,70)
Male, N (%)	436,643 (94.9%)
Race, N (%)	
White	333,097 (72.4%)
Black	95,134 (20.7%)
Hispanic	27,042 (5.9%)
Other	4940 (1.1%)
Top 5 diagnosis categories, N (%)	
CHF	24,157 (5.3%)
Sepsis	16,821 (3.7%)
Alcohol-related disorders	15,178 (3.3%)
Dysrhythmia	15,117 (3.3%)
Pneumonia	14,250 (3.1%)
Laboratory values, median (IQR)	
Albumin	3.9 (3.3, 4.3)
White blood cell count	8.2 (6.6, 11.2)
Creatinine	1.1 (0.8, 1.5)
Comorbidities, N (%)	
COPD	93,688 (20.4)
Liver disease	31,341 (6.8%)
Metastatic cancer	18,860 (4.1%)
CHF	68,347 (14.9%)

Predicted mortality ranges from 0 to 1.
IQR = Interquartile range.

3.1. Simulation results

Simulation results are presented in Table 3. For the baseline measurement scenario, where patients were allocated to pseudo-hospitals at random (i.e., no systematic difference in hospital case-mix, case-volume, or mortality), there was weak correlation between a pseudo-hospital's quintile ranking by template matching versus conventional regression, $\rho = 0.192$ (95% confidence intervals [CI]: 0.187, 0.198). Of hospitals ranked as bottom quintile by regression, 29.1% were also ranked as bottom quintile for template matching (whereas in the null scenario we would expect 20% and in a perfect agreement scenario we would expect 100%).

For the scenario with case-mix variation by age (2a), there was weak correlation between a pseudo-hospital's lowest-quintile classification by template matching versus regression, $\rho = 0.212$ (95% CI: 0.207, 0.218). Likewise, as in the baseline scenario, hospitals classified as bottom quintile by regression were also classified as bottom quintile by template matching 28.3% of the

Table 3

Simulation results.

	Pseudo-hospital variation in					Percent classified correctly by	
	Case-Mix	Case-Vol.	Mortality	Correlation between quintile ranking by TM and regression (ρ and 95% CI)	Percent agreement between TM and regression for bottom quintile (*P* and 95% CI)	TM	Reg.
1				0.192 (0.187, 0.198)	29.1 (28.5, 29.7)	N/A	N/A
2a	✓			0.212 (0.207, 0.218)	28.3 (27.7, 28.9)	N/A	N/A
2b	✓			0.415 (0.410, 0.420)	37.4 (36.8, 38.1)	N/A	N/A
3			✓	0.736 (0.733, 0.738)	81.5 (81.0, 82.1)	81.5	100
4a	✓		✓	0.738 (0.735, 0.740)	81.8 (81.3, 82.3)	81.8	100
4b	✓		✓	0.733 (0.730, 0.735)	80.3 (79.8, 80.9)	80.3	100
5		✓		0.225 (0.220, 0.230)	29.8 (29.2, 30.4)	N/A	N/A
6a	✓	✓		0.242 (0.236, 0.247)	30.2 (29.6, 30.8)	N/A	N/A
6b	✓	✓		0.355 (0.350, 0.360)	36.0 (35.4, 36.7)	N/A	N/A
7		✓	✓	0.745 (0.742, 0.748)	82.0 (81.5, 82.5)	82.0	100
8a	✓	✓	✓	0.737 (0.734, 0.740)	80.5 (80.0, 81.0)	80.5	100
8b	✓	✓	✓	0.738 (0.735, 0.741)	80.8 (80.2, 81.3)	80.8	100

TM = template matching, Reg = regression.

time—suggesting that variation by age does not induce additional bias into the performance assessment. However, for the scenario with case-mix variation by predicted mortality (2b), there was moderate correlation between a pseudo-hospital's classification by template matching versus regression, $\rho=0.415$ (95% CI: 0.410, 0.420). Template matching agreed with regression classifications of bottom-quintile pseudo-hospitals 37.4% of the time—suggesting that variation by predicted mortality introduced additional bias into the performance assessments.

For the scenario with systematic variation in pseudo-hospital mortality (3) (i.e., there was a true lowest quintile of pseudo-hospitals), there was strong correlation between the template matching and regression classifications, $\rho=0.736$ (95% CI: 0.733, 0.738). Template matching agreed with regression classifications of bottom quintile pseudo-hospitals 81.5% of the time. Across the 1000 simulations, regression correctly classified pseudo-hospitals as lowest quintile with 100% accuracy, while template matching classified 81.5% of pseudo-hospitals correctly.

When pseudo-hospital mortality and case-mix by age both varied (4a), results were similar to the scenario with case-mix by age variation only (2a). Template matching and regression assessments were strongly correlated ($\rho=0.738$ [95% CI: 0.735, 0.740]). Regression identified the lowest-quintile pseudo-hospitals with 100% accuracy; and template matching classified 81.8% of bottom quintile pseudo-hospitals correctly. When pseudo-hospital mortality and case-mix by predicted mortality both varied (4b), results were like case-mix by predicted mortality variation only (2b): template matching and regression assessments were strongly correlated $\rho=0.733$ (95% CI: 0.730, 0.735), regression classified bottom-quintile pseudo-hospitals with 100% accuracy compared with 80.3% for template matching.

When pseudo-hospital case volume varied (5), the correlation between template matching and regression assessments was weak ($\rho=0.225$ [95% CI: 0.220, 0.230]), but slightly stronger than the correlation in the baseline scenario ($\rho=0.192$ [95% CI: 0187, 0.198]). Likewise, template matching agreed with regression classifications of bottom-quintile pseudo-hospitals 29.8% of the time—slightly more than the 29.1% in the baseline scenario.

When case-mix by age and pseudo-hospital case volume varied (6a), the 2 methods identified the same pseudo-hospitals as lowest quintile 30.2% of the time. Template matching and regression adjustments were weakly correlated ($\rho=0.242$ [95% CI: 0.236,0.247]) slightly higher than with case-mix by age alone (2a) ($\rho=0.212$ [95% CI: 0.207, 0.218]) or pseudo-hospital volume alone (5) ($\rho=0.225$ [95% CI: 0.220, 0.230]). However, when both case-mix by illness severity and pseudo-hospital volume varied (6b), the correlation was weaker ($\rho=0.335$ [95% CI: 0.350, 0.360]) than when case-mix by illness severity only varied (2b) ($\rho=0.415$ [95% CI: 0.410, 0.240]), suggesting potential confounding due to pseudo-hospital case volume. Template matching classifications agreed for 36.0% of pseudo-hospitals classified as bottom quintile by regression.

In the scenario when pseudo-hospital case volume and mortality varied (7), the correlation between template matching and regression classifications was strong $\rho=0.745$ (95% CI: 0.742, 0.748) with 82.0% agreement of lowest-quintile pseudo-hospitals between both methods. Again, regression classified the lowest quintiles pseudo-hospitals accurately 100% of the time.

When pseudo-hospital case-mix by age, pseudo-hospital case volume, and pseudo-hospital mortality varied (8a), the correlation between the 2 methods was slightly weaker than when case-mix did not vary (7) ($\rho=0.737$ [95% CI: 0.734, 0.740]), with an 80.5% agreement and 100% accuracy for regression. When pseudo-hospital case-mix varied instead by predicted morality (8b), the correlations and agreements were nearly identical ($\rho=0.738$ [95% CI: 0.735, 0.741]), and 80.8%. This suggests that case-mix slightly biased the template matching identification when there were true differences in mortality. We display the complete results of each scenario in the supplement (Online Supplement, Appendix 5, http://links.lww.com/MD/C994).

3.2. Comparison of template matching and regression with post-match adjustment

In our post hoc analysis where we performed a post-match adjustment for predicted mortality on the matched patients in scenario 8b (case volume, case-mix, and performance variation), the template matching classification agreed with regression classification of bottom quintile pseudo-hospitals 87.1% of the time, compared with 81.8% without the post-match adjustment. The correlation between template matching and regression was $\rho=0.746$ (95% CI: 0.744, 0.749) compared with $\rho=0.738$ (95% CI: 0.736, 0.741), suggesting that the post-match adjustment improved case-mix variation that occurred at random.

3.3. Comparison of template matching and regression in real data

When we compared template matching to regression using the real hospital data, there was moderate correlation between the 2 methods, with $\rho=0.420$ (95% CI: 0.250, 0.562). Of the 22 hospitals classified as lowest quintile by regression, template matching identified 8 (36.4%) as also lowest-quintile (Fig. 1A). After performing a post-match adjustment for predicted mortality, the correlation and agreement increased slightly $\rho=0.482$ (95% CI: 0.325, 0.613), and template matching identified 10 (45.5%) of the 22 bottom quintile hospitals classified by regression (Fig. 1B). Figure 1A depicts the 111 VA hospitals risk-adjusted 30-day mortality and 95% confidence intervals (CIs) from the regression model, arranged from lowest to highest rates. Hospitals ranked in the lowest and highest quintiles by template matching are highlighted. Figure 1B depicts these results after performing a post-match adjustment for predicted mortality in template matching.

3.4. Assessment of the selected templates and matching quality

In our post hoc examination of templates, we found that the selected template—while optimal based on the multivariate distance measure—had mean values outside the interquartile range (IQR) for several variables. In Fig. 2, we display the box plots of 3 demonstrative variables for the 500 potential templates and highlight the selected template in the first 20 iterations. For each of the 3 variables, there were several iterations where the selected template had a value outside of the IQR. For example, in iteration 1, the selected template closely resembled the VA population on 30-day mortality rate and age but had an emergency department admission rate that was in the lowest quartile compared with the all potential templates. Iteration 5 had 30-day mortality and emergency department admission rates among the 25% highest compared with all potential templates and age among the 25% lowest.

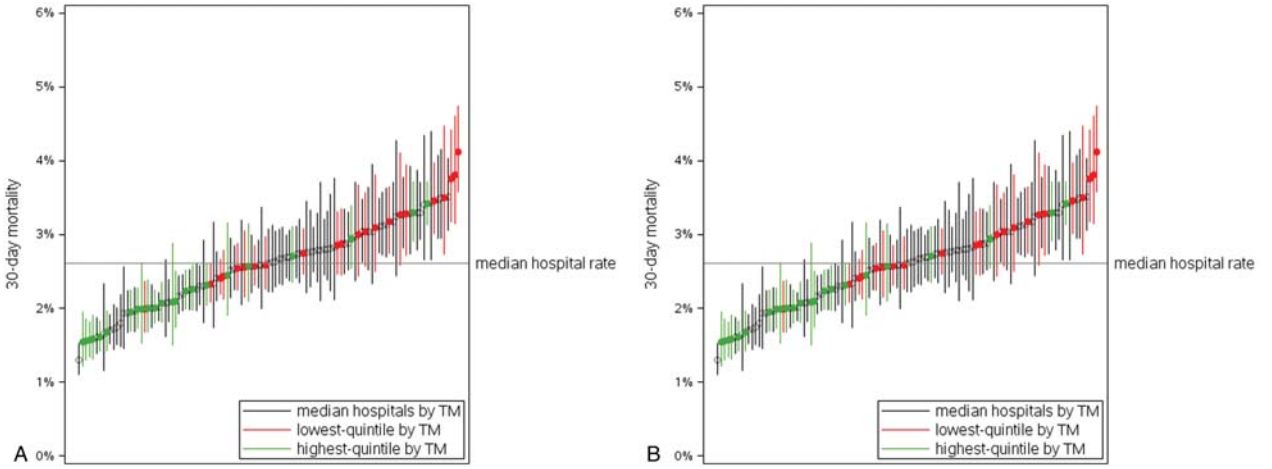

Figure 1. A and B: Caterpillar plots showing regression versus TM rankings in 2015 real data (original, post-match adjustment). TM = template matching.

In our second post hoc assessment of balance, we found that across the 111 pseudo-hospitals, the median standardized mean difference was 0 for most of the covariates. However, there was a wide range across pseudo-hospitals. One of the comorbidities was imbalanced at one pseudo hospital and white blood cell was imbalanced at 2 pseudo-hospitals. The remaining covariates had a standardized mean difference <0.25. We display the box-and-whisker plots of the distribution of standardized mean difference

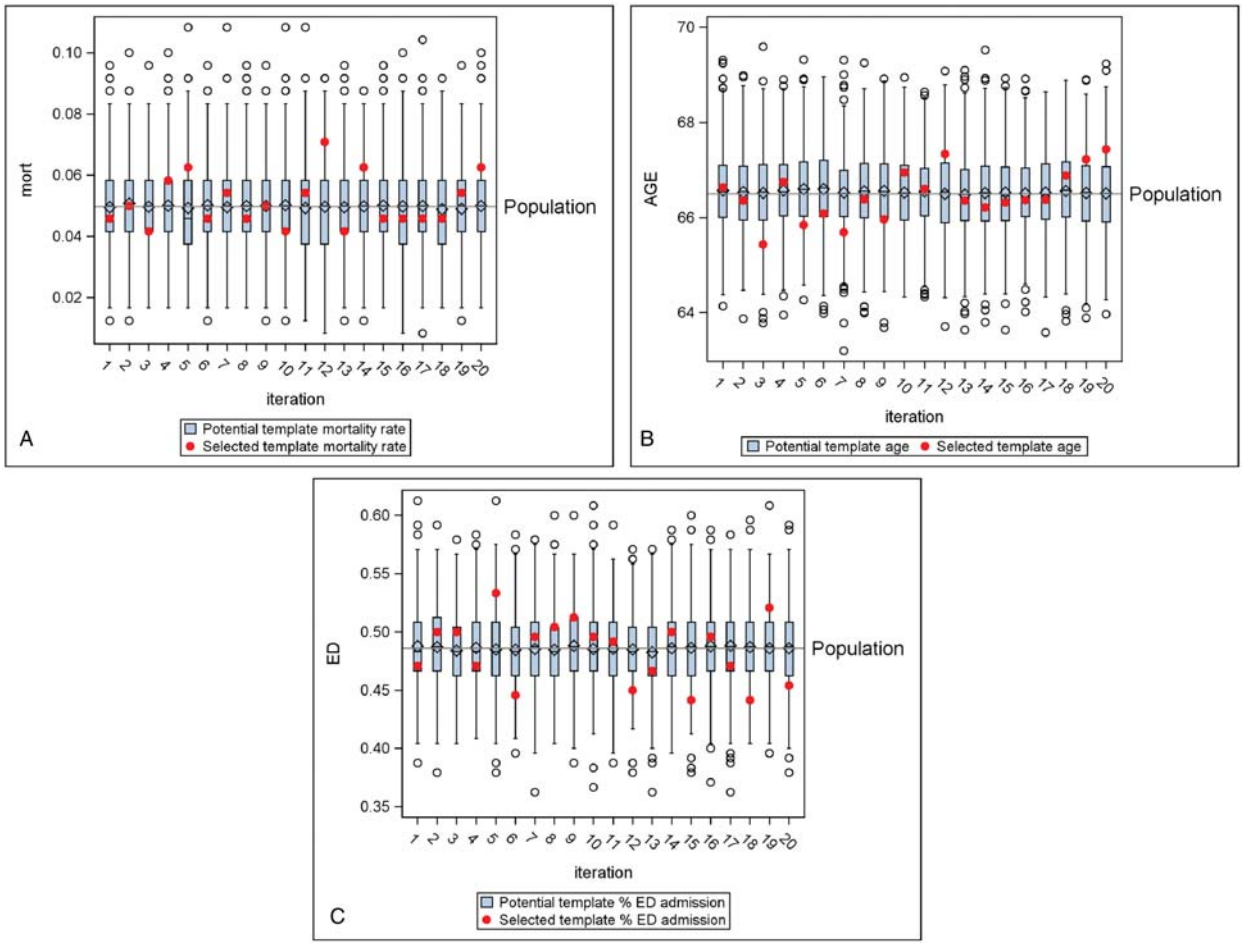

Figure 2. A–C: Plots showing variable distributions in the template (30-day mortality, age, emergency department admission).

across the 111 pseudo-hospitals in the supplement (Online Supplement, Appendix 6, http://links.lww.com/MD/C994).

4. Discussion

In this study, we have implemented a simplified template matching algorithm in a diverse healthcare system. We find that, with this simplified approach, template matching was inferior to the current regression approach used with the VA. When there were systematic differences in mortality, regression identified the lowest quintile hospitals with 100% accuracy, compared with 80.3% to 82.0% for our template matching implementation. However, as described below, we think this may be a limitation of our specific implementation rather than the approach in general. There were several limitations of our template matching method that may have resulted in lower performance relative to regression.

First, our template was not perfectly representative of the overall hospitalized population. We selected our template from a pool of 500 potential templates based on Mahalanobis distance to the overall population means, similar to Silber et al.[7] However, in contrast to prior studies,[7,9] we did not set additional requirements to ensure that the template contained particular proportions of diagnoses or patient demographics. In our post hoc analyses, we found that the selected template was not representative on all variables used in the matching process, although the specific variables that were unbalanced varied from iteration to iteration across the simulations. Furthermore, since the Mahalanobis distance weighs each variable equally, important variables (e.g., predicted mortality) were just as likely to be unrepresentative as less important variables.

Second, we created scenarios that likely did not capture the complexity of the multivariate relationships between the variables included in our algorithm, as evidenced by regression's 100% classification accuracy. Third, the number of profiles in our template may have been too small (i.e., under-powered) to identify the lowest quintile hospitals accurately, although a larger template would limit the ability to find matches at smaller-volume hospitals. Third, we included hospitalization with rare diagnoses, potentially impacting our match quality. Fourth, we included all hospitals in the assessment, even those with a lower quality match to the template. When we performed a post-match adjustment on the most important matching variable (predicted mortality), the performance of template matching improved only modestly. We elected to include all hospitals since benchmarking only a subset would have limited utility for real-world implementation; nonetheless, this likely impacted our results. Fifth, the simplistic algorithm that we used for matching, including many variables with equal weighting, may not be sufficient to ensure good balance on the factors that are most predictive of 30-day mortality.

There are a variety of ways the template matching procedure could be modified to improve its accuracy, and also the relevance of the benchmarking assessments. First, we will likely need to refine the inclusion and exclusion of hospitalizations in our template matching procedure. Prior studies of template matching have focused on common surgical procedures or common medical diagnoses (e.g., asthma, congestive heart failure, pneumonia).[8,9,20] We were interested in using template matching for overall hospital assessment, but still may have been too liberal with our hospital inclusions. In future work, we will consider excluding hospitalizations for uncommon diagnoses (i.e., hospitalizations for organ transplantation, or those with principal diagnoses occurring in $<1/300$ hospitalizations). On the other hand, we excluded hospitalizations for a primary psychiatric diagnosis from this study due to their low probability of mortality. However, as suicide prevention and mental health services are a priority in the VA, we plan include them in future work. Secondly, a larger template ($n=300$) may be needed to increase power. Third, to improve matching on the most important variables, it may be better to select the optimal template based on a subset of variables of highest importance, or alternatively, ensure that proportions of particular diagnoses or demographic groups match the overall population via a stratified sampling approach. Finally, we used a simple matching procedure because it is more computationally efficient, simpler to explain to end-users, and, if successful, would be much easier to implement within the VA computing infrastructure. However, given our results, we believe that more sophisticated matching algorithms will be necessary (e.g., the RCBalance, design-match, or MIPMATCH packages in R). Near-exact matching on important variables (e.g., operative/nonoperative status and decile of predicted mortality) can also be used to improve match quality.

Another potential explanation for the better performance of regression relative to template matching is the strength of the VA's predicted mortality model. In contrast to benchmarking by Center for Medicare and Medicaid services, the VA incorporates physiological data into risk-adjustment. The predicted mortality model typically achieves a c-statistic around 0.85,[13,14] indicating a strong ability to account for differences in case-mix across hospitals. It is possible that template matching may perform similarly or better than regression in scenarios where only administrative data are available for benchmarking. However, we elected to include physiologic data as this is the current standard practice within in the VA system. Likewise, we included all hospitalizations in the regression assessment (but only a subset for the template matching assessment) because our primary goal was to understand how template matching performs relative to current benchmarking practices within the VA.

5. Conclusion

We evaluated a simple template matching algorithm for benchmarking hospital performance in the VA healthcare system. However, the current standard practice of risk-adjusted regression incorporating patient-level physiological data was better able to identify lower-performing hospitals than this simplistic template matching algorithm. Thus, our current algorithm needs additional refinement before it could be used for hospital profiling by the VA.

Author contributions

Conception and design: Brenda M. Vincent and Hallie C. Prescott; Data analyses: Brenda M. Vincent; Interpretation of Data: all authors; Drafting of Manuscript: Brenda M. Vincent; Revision the manuscript for important intellectual content: all authors; approval of the manuscript for submission: all authors.

Conceptualization: Brenda M. Vincent, Wyndy L. Wiitala, Kaitlyn A. Luginbill, Timothy P. Hofer, Andrew M. Ryan, Hallie C. Prescott.

Formal analysis: Brenda M. Vincent, Andrew M. Ryan, Hallie C. Prescott.

Funding acquisition: Brenda M. Vincent, Kaitlyn A. Luginbill, Hallie C. Prescott.
Investigation: Brenda M. Vincent, Wyndy L. Wiitala, Timothy P. Hofer, Andrew M. Ryan, Hallie C. Prescott.
Methodology: Brenda M. Vincent, Wyndy L. Wiitala, Daniel J. Molling, Timothy P. Hofer, Andrew M. Ryan, Hallie C. Prescott.
Project administration: Kaitlyn A. Luginbill.
Resources: Brenda M. Vincent, Kaitlyn A. Luginbill, Timothy P. Hofer, Andrew M. Ryan, Hallie C. Prescott.
Software: Brenda M. Vincent.
Supervision: Hallie C. Prescott.
Validation: Brenda M. Vincent, Hallie C. Prescott.
Visualization: Brenda M. Vincent, Hallie C. Prescott.
Writing – original draft: Brenda M. Vincent, Hallie C .Prescott.
Writing – review & editing: Brenda M. Vincent, Wyndy L. Wiitala, Kaitlyn A. Luginbill, Daniel J. Molling, Timothy P. Hofer, Andrew M. Ryan, Hallie C. Prescott.
Brenda M. Vincent orcid: 0000-0003-4208-1059.
Hallie C. Prescott orcid: 0000-0002-8442-6724.

References

[1] Paul E, Bailey M, Pilcher D. Risk prediction of hospital mortality for adult patients admitted to Australian and New Zealand intensive care units: development and validation of the Australian and New Zealand Risk of Death model. J Crit Care 2013;28:935–41.
[2] Iezzoni LI. The risks of risk adjustment. JAMA 1997;278:1600–7.
[3] Centre ICNAaR. Online Reports; 2018. Accessed May 10, 2017.
[4] Services CfMM. Outcome Measures; 2017. Available at: https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/HospitalQualityInits/OutcomeMeasures.html. Accessed December 6, 2018.
[5] Pronovost PJ, Austin JM, Cassel CK, et al. Fostering Transparency in Outcomes, Quality, Safety, and Costs: A Vital Direction for Health and Health Care; 2016. Available at: https://nam.edu/fostering-transparency-in-outcomes-quality-safety-and-costs-a-vital-direction-for-health-and-health-care/. Accessed May 11, 2017.
[6] Austin JM, McGlynn EA, Pronovost PJ. Fostering transparency in outcomes, quality, safety, and costs. JAMA 2016;316:1661–2.
[7] Silber JH, Rosenbaum PR, Ross RN, et al. Template matching for auditing hospital cost and quality. Health Serv Res 2014;49:1446–74.
[8] Silber JH, Rosenbaum PR, Wang W, et al. Auditing practice style variation in pediatric inpatient asthma care. JAMA Pediatr 2016;170:878–86.
[9] Hu W, Chan CW, Zubizarreta JR, et al. Incorporating longitudinal comorbidity and acute physiology data in template matching for assessing hospital quality: an exploratory study in an integrated health care delivery system. Med Care 2018;56:448–54.
[10] Fihn SD, Francis J, Clancy C, et al. Insights from advanced analytics at the Veterans Health Administration. Health Aff (Millwood) 2014;33:1203–11.
[11] Cerner. APACHE Outcomes; 2018. Available at: https://apacheoutcomes.cernerworks.com/criticaloutcomes-home/. Accessed December 17, 2018.
[12] Learning SAfIa. Strategic Analytics for Improvement and Learning (SAIL) Value Model Measure; 2017. Available at: https://www.va.gov/QUALITYOFCARE/measure-up/SAIL_definitions.asp. Accessed May 10, 2017.
[13] Render ML, Deddens J, Freyberg R, et al. Veterans Affairs intensive care unit risk adjustment model: validation, updating, recalibration. Crit Care Med 2008;36:1031–42.
[14] Prescott HC, Kepreos KM, Wiitala WL, et al. Temporal changes in the influence of hospitals and regional healthcare networks on severe sepsis mortality. Crit Care Med 2015;43:1368–74.
[15] Prescott HC. Variation in postsepsis readmission patterns: a cohort study of veterans affairs beneficiaries. Ann Am Thorac Soc 2017;14:230–7.
[16] Mahalanobis PC. On the generalised distance in statistics. Natl Instit Sci 1936;2:49–55.
[17] Escobar GJ, Gardner MN, Greene JD, et al. Risk-adjusting hospital mortality using a comprehensive electronic record in an integrated health care delivery system. Med Care 2013;51:446–53.
[18] Inc SI. SAS/STAT. 14.2 User's Guide. Cary, NC: SAS Institute Inc; 2016.
[19] Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychol Methods 2010;15:234–49.
[20] Silber JH, Rosenbaum PR, Ross RN, et al. A hospital-specific template for benchmarking its cost and quality. Health Serv Res 2014;49:1475–97.