A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions

Jenna Wiens,[1] John Guttag,[1] Eric Horvitz[2]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[2]Microsoft Research, Redmond, Washington, USA

Correspondence to
Jenna Wiens, Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue 32G-904, Cambridge, MA 02139, USA; jwiens@mit.edu

Received 1 July 2013
Revised 5 January 2014
Accepted 7 January 2014
Published Online First 30 January 2014

ABSTRACT

Background Data-driven risk stratification models built using data from a single hospital often have a paucity of training data. However, leveraging data from other hospitals can be challenging owing to institutional differences with patients and with data coding and capture.

Objective To investigate three approaches to learning hospital-specific predictions about the risk of hospital-associated infection with *Clostridium difficile*, and perform a comparative analysis of the value of different ways of using external data to enhance hospital-specific predictions.

Materials and methods We evaluated each approach on 132 853 admissions from three hospitals, varying in size and location. The first approach was a single-task approach, in which only training data from the *target* hospital (ie, the hospital for which the model was intended) were used. The second used only data from the other two hospitals. The third approach jointly incorporated data from all hospitals while seeking a solution in the target space.

Results The relative performance of the three different approaches was found to be sensitive to the hospital selected as the target. However, incorporating data from all hospitals consistently had the highest performance.

Discussion The results characterize the challenges and opportunities that come with (1) using data or models from collections of hospitals without adapting them to the site at which the model will be used, and (2) using only local data to build models for small institutions or rare events.

Conclusions We show how external data from other hospitals can be successfully and efficiently incorporated into hospital-specific models.

INTRODUCTION

The ability to learn an accurate model for predicting patient outcomes at a specific hospital typically hinges on the amount of training data available. When ample data are available, it is often possible to learn models that can accurately predict even rare or infrequent events. However, institutions with smaller numbers of patients cannot collect enough data to construct useful models. Even larger hospitals may not be able to collect sufficient data about rare events.

According to the American Hospital Association more than half of all hospitals registered in the USA have fewer than 100 beds.[1] An average length of stay of 4.8 days[2] results in fewer than 8000 admissions a year (this estimate assumes 100% capacity and is therefore an upper bound). Given a goal of learning models to predict rare events (eg, an event occurring in <1% of the population), the smaller institutions can collect no more than 80 positive training examples a year. In medicine, where relationships between covariates and outcomes are usually complex, 80 positive training examples is usually too few to learn a predictive model that can be generalized to new cases. Such paucity of data makes it difficult to build hospital-specific models. Global models, developed for general use across multiple hospitals, have been developed for some areas of medicine; statistical models (and heuristic risk scores) have been developed and tested on data accessed from large national registries (eg, the American College of Surgeons National Surgical Quality Improvement Program[3]). However, as suggested by Lee *et al*, [4] these models often perform poorly when applied to specific institutions, because they do not take into account institutional differences.

We investigate an approach to building predictive models that involves augmenting data from individual hospitals with data from other hospitals. Applying data from multiple hospitals to predictions at a single target hospital presents an opportunity for *transfer learning*—the leveraging of evidential relationships in one or more related *source* 'tasks' for making predictions in a *target* task. Here, we shall focus on the target task of predicting which admissions to a specific hospital will result in a positive test result for toxigenic *Clostridium difficile*, a challenging healthcare-associated infection. Labeled datasets from other hospitals make up the source tasks. We consider a set of three hospitals, all belonging to the same hospital network. The data collected at each hospital contain hospital-specific distinctions —for example, the labels used to refer to units and rooms within the hospital. Moreover, the hospitals differ in the types of patients admitted. These differences contribute to differences in sets of observations or *feature spaces* that characterize relationships among observations and outcomes.

We explore three different solutions for building predictive models for a specific institution, in the context of using the base statistical methodology of L2-regularized logistic regression. The methods vary in the training data used and the details of the evidential features considered. The results suggest practical approaches to moving beyond a reliance only on local data to build institution-specific models for small institutions or rare events.

BACKGROUND AND SIGNIFICANCE

Transfer learning tackles the problem of leveraging data from a related *source* task to improve performance on a *target* task. There are different flavors of transfer learning depending on how the source and target tasks differ and the distribution of labeled

To cite: Wiens J, Guttag J, Horvitz E. *J Am Med Inform Assoc* 2014;**21**:699–706.

training data across source and target tasks. See Pan and Yang[5] for a review of transfer learning.

Most of the studies performed in transfer learning have dealt with differences in underlying distributions across tasks but often assume that all of the data lie in the same observation or *feature* space.[4 6–10] In our problem formulation, the outcome of interest, a *C difficile* infection, is the same across all tasks. However, the datasets lie in distinct but overlapping feature spaces.

In previous work, Evgeniou and Pontil[6] generalized regularization-based methods from single-task to multi-task learning. Their proposed solution was a natural extension of existing kernel-based learning methods that builds on ideas from hierarchical Bayesian modeling.[7 8] They assumed a solution for each predicted outcome (task) of the form $\mathbf{w}_t = \mathbf{w}_0 + \mathbf{v}_t$, where $\mathbf{w}_0$ represents a common solution shared among all tasks and $\mathbf{v}_t$ represents the task-specific variation from the common solution. Their method learns both common and task-specific components simultaneously, but assumes that all data are sampled from the same feature space.

Similar transfer learning approaches have been applied successfully to medical data.[4 11] Lee *et al* explored transfer learning for adapting surgical models to individual hospitals.[4] Like us, and others,[12] they hypothesized that models learnt in a straightforward manner from pooled data fail to reflect individual variations across hospitals. Their approach was two-step: using cost-sensitive support vector machines, they first trained a model on the source data and then learnt a model for the target data while regularizing the model parameters towards that of the source model. Their experiments showed a significant improvement over other methods such as ones learnt only from target data or from source data. However, their work assumed that there are no missing data and that all of the data lie in an identical feature space. The omission of hospital-specific features is typical in multicenter studies—for example that of Sugiyama *et al*.[13] In reality, transferring data across hospitals can be messy because many of the observational variables, such as staff, protocol, and locations, are hospital-specific. In 'Experiments and results', we show how important these hospital-specific variables can be.

Researchers have investigated the task of transferring knowledge across different feature spaces in other contexts. Bel *et al*[14] explored the problems that arise when classifying documents in different languages. Common solutions to this problem either translate one document to the target language or map both documents to a language-independent feature space, analogous to either mapping the source data into the target domain or mapping both to a shared representation. Previous work has also proposed approaches to translate auxiliary data into the target space from one medium (eg, text) to another (eg, an image).[15] Such *translated learning* applies to a different scenario than ours; in translated learning there is no explicit correspondence between the source feature space and the target feature space.

Several researchers have investigated transfer learning in the context of linear classifiers.[9 10] Previous work has explored modifications to the support vector machine objective function to include consideration of the loss associated with both target and source data, but exclude the source data from either the constraint or the set of support vectors. In comparison, we consider methods based on L2-regularized logistic regression that do not involve explicit modification of the objective function.

MATERIALS AND METHODS

Data and preprocessing

Our data come from three hospitals belonging to the MedStar Health healthcare system. The institutional review board of the Office of Research Integrity of the Medstar Health Research Institute approved the statistical analysis of retrospective medical records. We refer to the hospitals as hospital A, hospital B, and hospital C. All three hospitals are described below.

- *Hospital A*: the smallest of the three hospitals. It has about 180 beds and sees just over 10 000 admissions a year.
- *Hospital B*: an acute care teaching hospital. It has about 250 beds and 15 000 inpatient visits a year.
- *Hospital C*: a major teaching and research hospital with over 900 beds and more than 40 000 inpatient visits a year.

Hospitals A and B are located in the same city only 10 miles apart, whereas hospital C is located in a different city about 50 miles away. Despite the large differences in size and location, hospital C overlaps with hospitals A and B in many of the services provided. Table 1 describes the population of patients admitted to each hospital over the same 2 years.

We are interested in risk-stratifying patients at the time of admission based on their probability of testing positive for toxigenic *C difficile* during their hospital admission. The availability of a well-calibrated prediction could enable proactive intervention for patients at high risk of becoming infected. If a patient tests positive for toxigenic *C difficile* at any time during his/her admission, the admission is assigned a positive label (negative otherwise). We consider all inpatient visits for the 2 years between April 2011 and April 2013. This results in a total of

Table 1 Descriptive statistics comparing the study population across the three different institutions

	Hospital A (%) (n=21 959)	Hospital B (%) (n=29 315)	Hospital C (%) (n=81 579)
Female gender	62.34	50.29	55.97
Age:			
[0, 2)	14.38	0.00	9.00
[2, 10)	0.75	0.00	0.00
[10, 15)	0.80	0.07	0.00
[15, 25)	7.23	3.77	6.73
[25, 45)	21.27	15.46	19.05
[45, 60)	21.28	30.98	22.77
[60, 70)	13.16	21.19	16.78
[70, 80)	10.79	15.97	13.74
[80 100)	8.11	10.20	9.24
≥100	2.25	2.36	2.67
Hospital admission type:			
Newborn	13.13	0.00	8.74
Term pregnancy	7.53	0.00	8.89
Routine elective	15.87	31.28	17.39
Urgent	7.53	7.84	11.26
Emergency	10.79	15.97	13.74
Hospital service:			
Medicine	51.18	49.15	40.85
Orthopedics	5.61	18.76	1.54
Surgery	7.53	5.97	10.28
Obstetrics	13.97	0.00	10.09
Cardiology	0.00	2.99	11.36
Newborn	13.15	0.00	9.01
Psychiatry	0.00	13.11	3.70
Hemodialysis	3.06	5.32	6.76
Diabetic	24.44	32.73	33.59
Clostridium difficile	0.80	1.08	1.05
Previous visit in past 90 days	5.87	7.43	5.54

Table 2 The amount of available data varies significantly across the three different institutions

Hospital	Admissions (n)	*Clostridium difficile* cases (n)
First year (Apr 2011–Apr 2012)		
A	11 380	82
B	14 675	161
C	39 467	426
Second year (Apr 2012–Apr 2013)		
A	10 579	94
B	14 640	157
C	42 112	428

The outcome we consider occurs in about 1% of the population, resulting in low numbers of positive examples at smaller institutions.

132 853 admissions and 1348 positive cases of *C difficile* (see table 2 for the distribution across hospitals).

Because we are interested in stratifying patients by risk at the time of admission, we consider only data available at admission. For admissions from all three hospitals, we extract observations or *features* pertaining to the categories listed in table 3. We map all features to binary-valued observations and remove variables that do not occur in at least 1% of at least one hospital's population. This preprocessing results in 578 binary features: 256 shared by all three hospitals.

The remaining features are specific to either a single hospital or shared by two hospitals. Figure 1 shows a labeling of the sets of shared and specific features across the different hospitals. Table 3 gives more detail about the types of features present across the three different hospitals.

Risk stratification

To preserve interpretability, we consider learning in the context of linear classifiers. We formulate the problem as follows.

We have N datasets:

$$\mathcal{D}_j = \{(\mathbf{x}_{j_i}, y_{j_i}) | \mathbf{x}_{j_i} \in \mathcal{X}_j, y_{j_i} \in \{-1,1\}\}_{i=1}^{n_j}$$

where $j = 0 \ldots N-1$. $\mathcal{D}_0$ represents the target task and $\mathcal{D}_1, \ldots, \mathcal{D}_{N-1}$ represent the source tasks. n_j represents the number of labeled examples available from each task. The binary classification goal is the same for each task. However, we must contend with different sets of variables, which we refer to as *feature spaces* $\mathcal{X}_0, \ldots, \mathcal{X}_{N-1}$. We assume that there is some overlap between the features spaces for each of the source tasks and the target task under consideration, that is, $\forall 1 \leq i \leq N-1, \mathcal{X}_0 \cap \mathcal{X}_i \neq \emptyset$. Figure 2 depicts the possible intersection among feature spaces for a specific target task when $N = 3$.

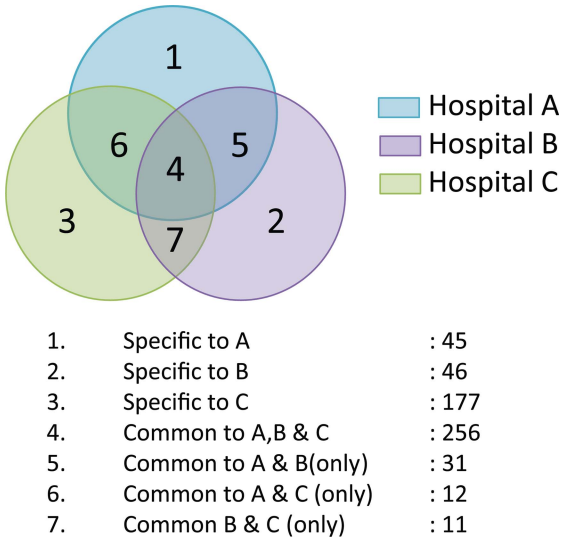

Figure 1 The data for each hospital lie in a different feature space. Here we give the amount of overlap among the different institutions.

In logistic regression, we seek a function $f : \mathbb{R}^d \rightarrow [0,1]$ of the form:

$$f(\mathbf{x}_i) = \frac{1}{1 + \exp^{-(b_0 + \mathbf{w}^T \mathbf{x}_i)}} \quad (1)$$

where $\mathbf{w} \in \mathbb{R}^d$ (and $\mathbf{x} \in \mathbb{R}^d$). Solving for the regression coefficients $\mathbf{w}$ and b_0 is a maximum likelihood estimation problem. To improve generalizability, we consider L2-regularized logistic regression, where λ is a tuning parameter.

$$\min_{\mathbf{w}} \frac{\lambda}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^{n} \log(1 + \exp^{-y_i \mathbf{w}^T \mathbf{x}_i}) \quad (2)$$

Table 3 The variables considered can be grouped into the broad categories given here

Variable type	Set 1 A	Set 2 B	Set 3 C	Set 4 A, B, C	Set 5 A, B	Set 6 A, C	Set 7 B, C
Admission details*	2	2	3	28	0	4	4
Patient demographics†	2	0	0	52	1	0	5
Patient history‡	0	0	0	19	0	0	0
Previous visit statistics (LOS)	0	0	0	20	0	0	0
Medications from previous visit	4	0	5	137	30	8	1
Home medications	0	0	95	0	0	0	0
Attending doctor identification number	27	30	12	0	0	0	0
Location units	10	14	62	0	0	0	1
Totals	45	46	177	256	31	12	11

Each column gives the number of features within a category pertaining to the feature subset (see figure 1).
*Includes complaint, source, hospital service, expected surgery, and month.
†Includes age, marital status, race, sex, city, and financial class.
‡Includes previous diagnoses (ICD9 codes), and history of *Clostridium difficile*.
LOS, length of stay.

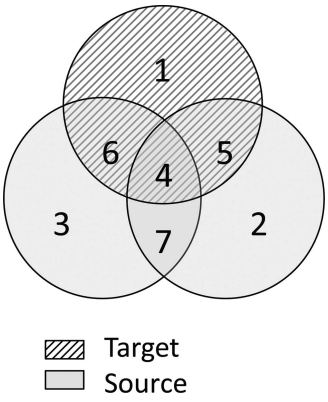

Figure 2 When N=3 the target feature space contains four different sets of features: features common to all tasks, target-specific features, and features shared only between the target and one of the sources.

Note that we add an extra constant dimension to $\mathbf{x}$ and compute the offset b_0 implicitly. The solution to (2) depends on the $\mathbf{X}$ and $\mathbf{y}$ employed in the training. Here we describe three potential solutions:

▶ Target-only

This approach is a single-task approach and uses data only from the target task (ie, $\mathcal{D}_0$). This results in a solution of the form $\mathbf{w} \in \mathcal{X}_0$.This approach can easily overfit if the number of training examples from the target task is small.

▶ Source-only

Given only labeled data from the source tasks, this approach seeks a solution by exploiting the shared features of each of the source tasks and the target tasks. The solution to this approach lies in the union of the intersections of the target feature space and each of the source features spaces (eg, regions 4, 5, and 6 in figure 2). The solution is of the form $[\mathbf{w}_c; \mathbf{v}_1; \mathbf{v}_2; \ldots; \mathbf{v}_{N-1}]$, where w_c represents the common features shared among all tasks (ie, region 4 in figure 2), and $\forall 1 \leq i \leq N-1, \mathbf{v}_i$ represents the features shared only between the target and the source $\mathcal{D}_i$ (ie, regions 5 and 6 in figure 2). We rewrite the objective function in (2) to incorporate data from different sources. Here source data are mapped to the common feature space shared with the target task by removing all source-specific features.

$$\min_{\mathbf{w}_c,\mathbf{v}_1,\ldots,\mathbf{v}_{N-1}} \frac{\lambda_c}{2}\|\mathbf{w}_c\|^2 + \sum_{j=1}^{N-1}\left(\frac{\lambda_j}{2}\|\mathbf{v}_j\|^2 + \sum_{i=1}^{n_j}\log\left(1+\exp^{-y_{l_i}[\mathbf{w}_c;\mathbf{v}_j]^T\mathbf{x}_{l_i}}\right)\right) \quad (3)$$

Because this solution depends on source data only, target-specific features (eg, region 1 in figure 2) will have no effect on the classification of a test patient.

▶ Source+Target

With Source+Target, we extend the solution described above to incorporate the target data, and the target-specific features.

$$\min_{\mathbf{w}_c,\mathbf{v}_0,\mathbf{v}_1,\ldots,\mathbf{v}_{N-1}} \frac{\lambda_c}{2}\|\mathbf{w}_c\|^2+\frac{\lambda_0}{2}\|\mathbf{v}_0\|^2+\sum_{i=1}^{n_0}\log(1+\exp^{-y_{0_i}[\mathbf{w}_c;\mathbf{v}_0;\ldots;\mathbf{v}_{N-1}]^T\mathbf{x}_{0_i}}) + \sum_{j=1}^{N-1}\left(\frac{\lambda_j}{2}\|\mathbf{v}_j\|^2+\sum_{i=1}^{n_j}\log(1+\exp^{-y_{l_i}[\mathbf{w}_c;\mathbf{v}_j]^T\mathbf{x}_{l_i}})\right) \quad (4)$$

This approach assumes a solution of the form $[\mathbf{w}_c; \mathbf{v}_0; \ldots; \mathbf{v}_{N-1}]$ where v_0 pertains to target specific features. The final solution $[\mathbf{w}_c; \mathbf{v}_0; \ldots; \mathbf{v}_{N-1}] \in \mathcal{X}_0$ as in the Target-only, approach but incorporates data from all tasks.

Note that if $\forall j, \lambda = \lambda_c = \lambda_j$ we can rewrite the objective function of (4) as:

$$\min_{\mathbf{w}_t} \frac{\lambda}{2}\|\mathbf{w}_t\|^2 + \sum_{i=1}^{n_0+\ldots+n_{N-1}}\log(1+\exp^{-y_i\mathbf{w}_t^T\mathbf{x}_i}) \quad (5)$$

where $\mathbf{w}_t=[\mathbf{w}_c;\mathbf{v}_0;\ldots;\mathbf{v}_{N-1}], \mathbf{y}=[\mathbf{y}_0;\ldots;\mathbf{y}_{N-1}], \mathbf{X}=[\mathbf{X}_0;\mathbf{X}_1'';\ldots;\mathbf{X}_{N-1}'']$. The target data are used in their original form while the source data undergo two transformations. First, they are mapped to the common feature space $\mathbf{X}_j \rightarrow \mathbf{X}_j'$ (removing source-specific features) and then mapped to the target feature space $\mathbf{X}_j' \rightarrow \mathbf{X}_j''$ (by augmenting with zeros). Transforming the data in this way renders the objective function analogous to (2). These target-specific transformations to the data allow for transfer of knowledge across hospitals.

EXPERIMENTS AND RESULTS

In this section we outline a series of experiments in which we investigate the applicability of each of the learning approaches described in the previous section. In each subsection we analyze different aspects of the problem in order to gain insight into how and when to transfer knowledge across hospitals (see table 4).

Including source data helps

Eventually, we will consider risk stratification for *C difficile* at each of the three hospitals. To start, we consider the task of risk stratification at the smallest hospital (ie, hospital A). So, hospital A is our *target* task and hospital B and hospital C represent the *source* tasks. We split the data for each hospital temporally into data from the first year and data from the second year (see table 2). In all of the experiments, we train on data from the first year and test on data from the second year.

As described in 'Risk stratification', depending on the scenario considered, the training set consists of all or only a subset of the available training data. The dimensionality of the classifier

Table 4 Outline of experiments presented in the remainder of this section

Section	Training data	Test data	Description
Including source data helps	Target-only Source-only Target+Source	Target hospital	Compares the three approaches presented in the previous section
Target-specific features are important	Target-only $\mathcal{D}$=D1 d=D2 d=D3	Target hospital	Measures the importance of the target-specific features to each target task. We keep the training and test data constant but vary the dimensionality of the solution (D1>D2>D3)
More data are not always better	Target-only Source-only	Target hospital	Measures the effect of having a small amount of data from the target task versus twice as much data from the source tasks
Not all transfer is equal	Source-only Source 1 Source 2	Target hospital	Investigates the relative contribution each source (source 1 and source 2) makes to the target task, by considering each source independently

All experiments are repeated three times such that each hospital is considered as the *target* task.

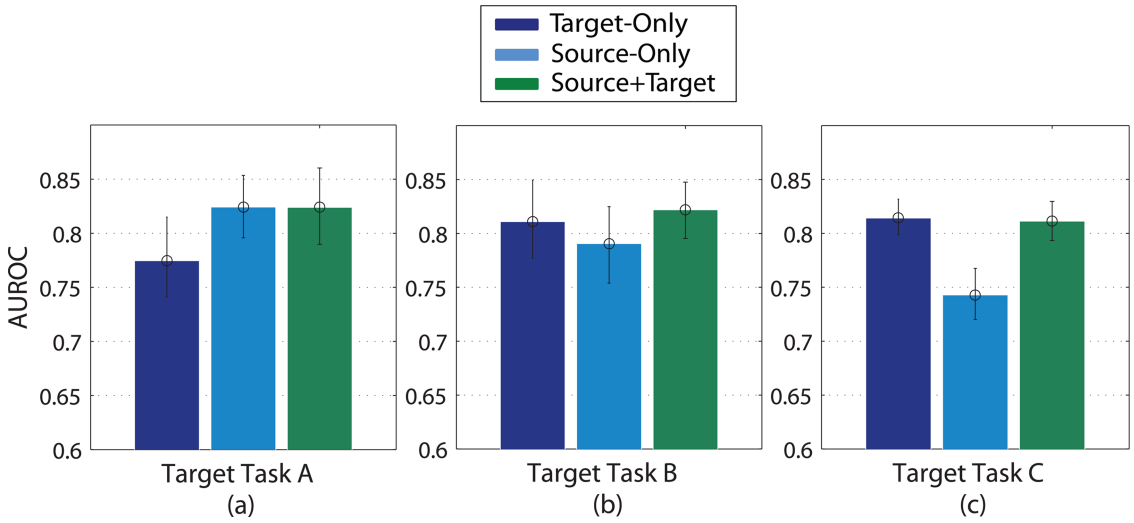

Figure 3 Results of applying all three approaches to each of the target tasks. In each case the source data pertain to data from the other two hospitals. AUROC, area under the receiver operating characteristic (curve).

learnt depends on the origin of the training data. Both Target-only and Source+Target approaches incorporate data from the target task, so the dimensionality is that of the target task. When data from the source tasks only (Source-only) are used, the features that occur only at the target hospital (target-specific) are ignored and the solution lies in a lower dimension (see figure 2).

Using LIBLINEAR,[16] we learn three different risk prediction models based on the approaches described in the 'Risk stratification' section. We apply each classifier to the same hold-out set (data from the second year) from hospital A. We select hyperparameters using five-fold cross-validation on the training set. We set the hyperparameters equal to one another, as in equation (5). Although this assignment is not optimal, it makes training a model more efficient, because otherwise optimization would require a search over three dimensions to find $\lambda_1, \lambda_2,$ and λ_3.

The results of this initial experiment are shown in figure 3A and table 5 (denoted by Target Task '**A**'). The results give the performance on the hold-out set in terms of the area under the receiver operating characteristic (AUROC) curve, the area under the precision recall curve, the breakeven point where precision=recall, and finally the OR (using a cutoff point based on the 95th centile). We calculated the 95% CIs using bootstrapping on the hold-out set. Comparing the performance of three classifiers in figure 3A, we see that the classifier learnt solely on data from the target task (ie, Target-only) performs the worst. When data from hospitals B and C are included in the training set, we see a significant improvement in performance. These results demonstrate how auxiliary data can be used to augment hospital-specific models. Hospital A has only 82 positive training examples, compared with hospitals B and C with a combined 587 positive training examples. These additional positive examples help the model generalize to new data.

In figure 3A, Source-only and Source+Target perform almost identically. This might be because (1) the relatively small amount of added data when training the Source+Target classifier is not enough to have a significant influence on the performance, and/or (2) the target task (hospital A) does not differ significantly from the source tasks (hospitals B and C).

To explore how the amount of available training data from the target task affects the relative performance of the three approaches, we repeat the experiment described above target tasks B and C. The results of these additional experiments are displayed in figures 3B,C and table 5.

Figure 3 shows how the relative performance of the three approaches differs depending on the target task at hand. When ample data from the target task are available (eg, target task C), ignoring the target data can significantly hurt performance. This result highlights the importance of including available target-specific data when training a model.

Target-specific features are important

Across all three hospitals (ie, target tasks), we note that the Source+Target approach performs at least as well as the best classifier. The Source+Target approach jointly incorporates all of the available training data and all of the features relevant to

Table 5 Result of applying the three approaches to each hospital, as described in Methods and Materials

Target task	Approach	No of training examples (positive)	AUROC (95% CI)	AUPR (95% CI)	Breakeven Precision=Recall	OR (95th centile)
A	Target-only	11 380 (82)	0.7746 (0.74 to 0.82)	0.0379 (0.01 to 0.06)	0.0957	6.3886
	Source-only	54 142 (587)	0.8242 (0.80 to 0.85)	0.0656 (0.02 to 0.10)	0.1383	9.8679
	Source+Target	65 522 (669)	0.8239 (0.79 to 0.86)	0.0638 (0.03 to 0.09)	0.1489	9.3806
B	Target-only	14 675 (161)	0.8110 (0.78 to 0.85)	0.0664 (0.04 to 0.09)	0.1274	11.3245
	Source-only	50 847 (508)	0.7907 (0.75 to 0.82)	0.0557 (0.03 to 0.08)	0.1146	10.3604
	Source+Target	65 522 (669)	0.8219 (0.80 to 0.85)	0.0699 (0.04 to 0.10)	0.1656	10.3604
C	Target-only	39 467 (426)	0.8142 (0.80 to 0.83)	0.0526 (0.04 to 0.06)	0.0958	7.9779
	Source-only	26 055 (243)	0.7428 (0.72 to 0.77)	0.0356 (0.03 to 0.04)	0.0818	6.0304
	Source+Target	65 522 (669)	0.8114 (0.79 to 0.83)	0.0518 (0.04 to 0.06)	0.1051	8.9709

AUROC, area under the receiver operating characteristic; AUPR, area under the precision recall curve.

the target task. In the next set of experiments, we measure how the inclusion or exclusion of target-specific features affects classifier performance.

For each task, we learn three different classifiers on the same training data but in different feature spaces. First, we train a Target-only classifier, as in the previous experiments, using the available target training data for each of the target tasks. Next, we learn two additional Target-only classifiers but in a lower dimensionality than with the first classifier. For example, consider target task A, the first classifier (A1) learns a solution using all of the features available to task A (ie, the union of sets 1, 4, 5, and 6 in figure 2), the second classifier (A2) ignores the target-specific features (ie, it uses sets 4, 5, and 6), while the final classifier (A3) considers only features common to all tasks (ie, set 4). In doing so, we control for the amount of training data and any changes in underlying distributions that could influence performance on the hold-out data (eg, relationship between the conditional or marginal distributions of the source and target data). For the three classifiers, the training data and test data are identical except for the set of features considered.

The results of this experiment are shown in figure 4. The trend across all three tasks is the same, fewer features lead to worse performance. The detrimental effect of removing the target specific features is most noticeable for target task C. Hospital C has 177 hospital-specific features not found at the other two hospitals. Ignoring these target-specific features leads to a significant drop in performance from an AUROC of 0.814 (95% CI 0.798 to 0.834) to an AUROC of 0.776 (95% CI 0.755 to 0.795). The removal of the target-specific features at the other two hospitals has less of an impact on performance. For hospitals A and B there are fewer target-specific features (45 and 46 features, respectively) and fewer target-specific training data. This might explain why there is no significant difference between the AUROC achieved by the Source-only and Source +Target approaches for these two target tasks (see table 5).

In a follow-up experiment, we learnt a single classifier by pooling all of the data and searching for a solution in the feature space common to all three tasks (ie, region 4 in figure 1). Applied to the hold-out data from task A, B, and C we achieve an AUROC of 0.8178 (0.78 to 0.86), 0.7947 (0.76 to 0.84), and 0.7664 (0.74 to 0.79), respectively. This straightforward approach, ignores the target-specific features and as we might expect results in a worse performance relative to the Source+Target approach.

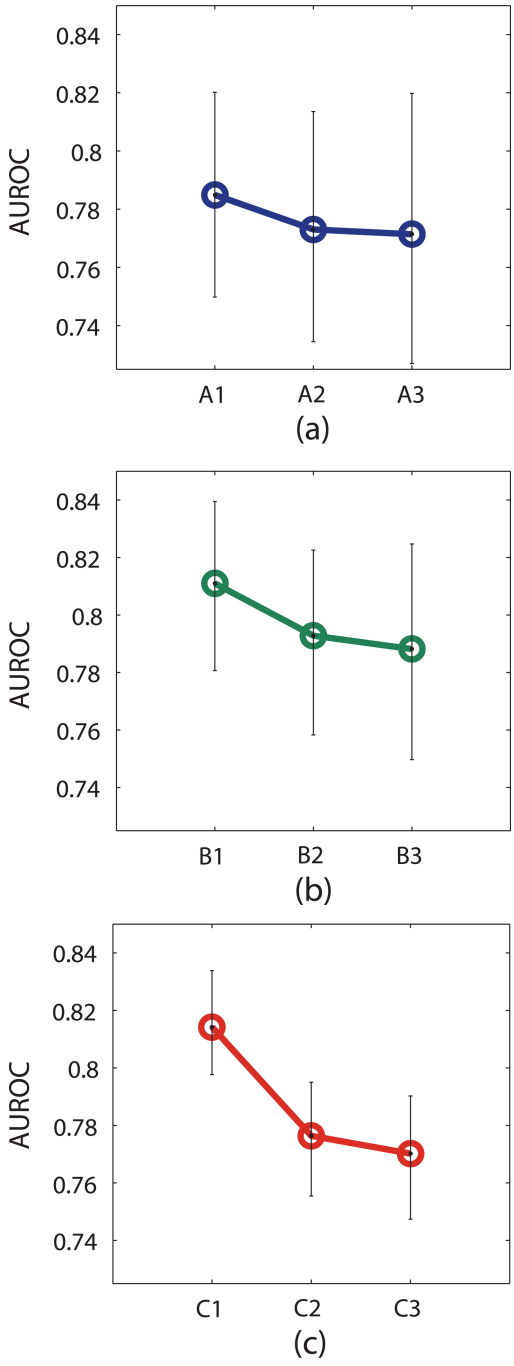

Figure 4 Here the amount of training and test data are kept constant but the dimensionality of the solution varies. AUROC, area under the receiver operating characteristic

More data are not always better

In our next experiment we compare three different models for each hospital (1) a Target-only model using a random sample of 5000 admissions, (2) a Target-only model at each hospital using a random sample of 10 000 admissions, (3) a Source-only model using a random sample of 5000 admissions from each of the two hospitals. The average performance across 10 repetitions is shown in figure 5.

For hospitals B and C having a small amount of data from the target task is better than having *twice* as much data from the source task. However, for hospital A the Source-only approach does better than the Target-only approach despite the same amount of training data. These two approaches seek solutions in different feature spaces. The Target-only approach seeks a solution in a higher dimensionality. This discrepancy in performance could be for a combination of reasons (1) the source data are a better approximation of what will happen during the next year at hospital A and/or (2) the target-specific features for hospital A are not informative.

Not all transfer is equal

When the results for target tasks A and B are compared, the Source-only approach appears to work better for target task A than it does for target task B. The amount of training data used in training a classifier for hospital A is only slightly greater than for hospital B (54 142 vs 50 847). This raises the question of whether the discrepancy in performance is simply owing to the effect of having 6.5% more data or owing to differences in the underlying similarities between the source and target tasks. Data from hospital C are included in the training data for both target tasks, but it might be that data from hospital C transfer more readily to target task A than to target task B.

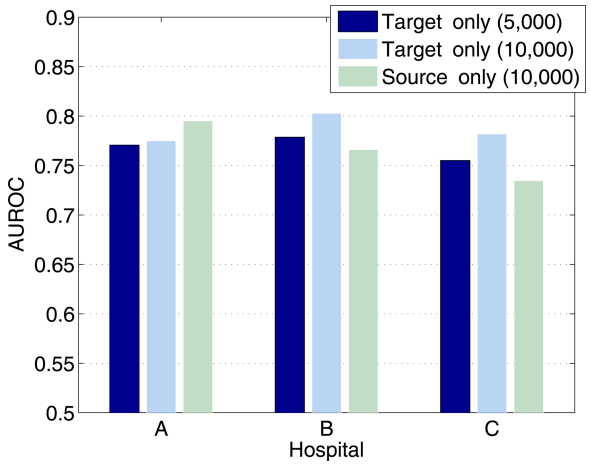

Figure 5 The results of experiments from 'More data are not always better', applied to each target hospital. AUROC, area under the receiver operating characteristic

To investigate this question, we apply the Source-only approach to each of the three target tasks. However, instead of combining data from the two available source hospitals we learn a model for each source independently (while controlling for the amount of training data) and apply it to the target task. The results of this experiment are shown in figure 6. The source of the training data are denoted along the x axis in figure 6. We control for the amount of training data available at each hospital by randomly undersampling data from the larger hospital.

These results suggest that data from hospital C might transfer more readily to hospital A than data from hospital B, even though hospital A and hospital B have more features in common. This observation is supported by the last pair of bars in figure 6: for target task C a classifier trained on data only from hospital A outperforms a classifier trained on data only from hospital B. This suggests that hospital B is the most different of the three hospitals. This might explain why, despite the large amount of training data, the Source-only approach performs relatively poorly for target task B, but performs well for target task A (see figures 3A,B). Additionally, as alluded to earlier, these relationships might explain why the performance

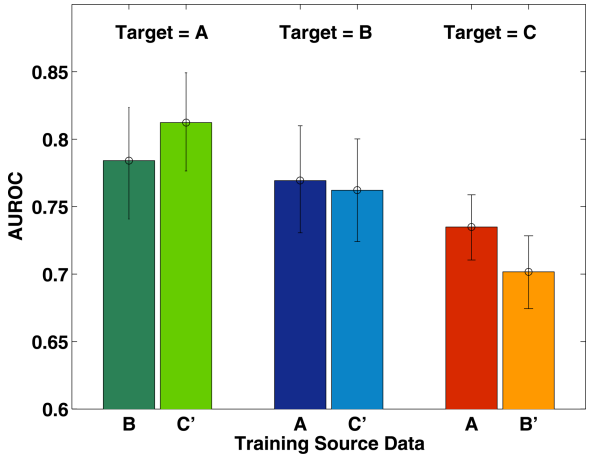

Figure 6 Here only data from a single source task are used to learn a model, which is then applied to the target task. The ' indicates that the amount of data used in training was limited to the amount of data available from the other source. AUROC, area under the receiver operating characteristic

of the Source-only and Source+Target approaches are almost identical for target task A.

DISCUSSION AND CONCLUSION

In the previous section, our experiments were limited to three hospitals (N=3). With over 5000 registered hospitals in the USA alone, larger numbers of N are feasible. Opportunities for scaling raise several important considerations and implications. First, as N increases the number of features common to all hospitals will shrink and therefore the number of hospital-specific features will increase. Limiting models to only the shared feature set (as in the study by Lee *et al*[4]) risks ignoring possibly crucial hospital-specific information. Second, as N increases the variation among hospitals will increase. Even for hospitals within the same network, we found that the transferability of knowledge was neither equal nor symmetric among hospitals. As the variation among tasks increases, it is plausible that including auxiliary data when training a model might actually diminish performance on the target task. Future work is needed to investigate how to best select source data from a large pool of hospital databases. Depending on the task, this could mean selecting the best subset of hospitals, or the best subset of data from each hospital. Third, as N increases the number of hyperparameters increases. Each hyperparameter controls the extent to which data from each hospital contribute to the final model. Procedures for identifying an optimal setting for hyperparameters can quickly become inefficient with increasing N, posing new challenges and opportunities in machine learning.

We note that many of the problems that arise when transferring knowledge across hospitals involve transferring knowledge across time at an individual hospital. Over time, hospital populations, physical plants, tests, protocols and staff change. Furthermore, electronic medical records change in the data collected and the precise meanings of variables. Incorporating past data into current models is an important future direction.

We have presented methods and experiments using data from three hospitals to understand the potential gains and challenges associated with leveraging data from external hospitals in building predictive models for *C difficile* infections. Although there is no global model for the considered prediction task, the inconsistent performance of the Source-only approach across target tasks indicates why national models often perform poorly when applied to specific institutions.[4] Auxiliary data tend to have the greatest impact when the number of target training examples is small, the number of shared features is large and there is significant overlap in the shared feature space. When ample data from the target space are available, our results demonstrate the importance of including target-specific data and target-specific features when training hospital-specific risk stratification models. Our findings highlight the promise of leveraging external data for building models at specific hospitals at which predictions will be used. We believe that further study of techniques that facilitate the incorporation of all available data across hospitals and databases should be a top priority in efforts to construct and harness predictive models in healthcare.

Contributors All authors contributed significantly to this work, meet the criteria for authorship and have seen and approved the final manuscript.

Funding This work was supported in part by the US National Science Foundation, Quanta Computer Inc, Microsoft Research and the National Science and Engineering Research Council of Canada.

Competing interests None.

Ethics approval Office of Research Integrity, MedStar Health Research Institute.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 http://www.ahrq.gov/legacy/qual/hospsurvey12/hosp12tab3-1.htm (accessed Jun 2012).
2 http://www.cdc.gov/nchs/fastats/hospital.htm (accessed Jun 2012).
3 Fink A, Campbell D, Mentzer R, *et al*. The national surgical quality improvement program in non-veterans administration hospitals. *Ann Surg* 2001;263:344–54.
4 Lee G, Rubinfeld I, Syed Z. Adapting surgical models to individual hospitals using transfer learning. *IEEE ICDM 2012 Workshop on Biological Data Mining and its Applications in Healthcare (BioDM)*. 2012.
5 Pan S, Yang Q. A survey on transfer learning. *IEEE Trans on Knowl and Data Eng* 2010;22:1345–59.
6 Evgeniou T, Pontil M. Regularized multi-task learning. *KDD*. 2004.
7 Heskes T. Empirical Bayes for learning to learn. *ICML*. 2000.
8 Bakker B, Heskes T. Task clustering and gating for Bayesian multi-task learning. *J Mach Learn Res* 2003;4:83–99.
9 Liao X, Xue Y, Carin L. Logistic regression with an auxiliary data source. *Proceedings of ICML*. 2005.
10 Wu P, Dietterich TG. Improving SVM accuracy by training on auxiliary data sources. *ICML*. 2004.
11 Widmer C, Leiva-Murilla J, Altun Y, *et al*. Leveraging sequence classification by taxonomy-based multitask learning. *Res Comput Mol Biol* 2009;6044: 522–34.
12 Reis BY, Mandl KD. Integrating syndromic surveillance data across multiple locations: effects on outbreak detection performance. *AMIA Annu Symp Proc* 2003:549–53.
13 Sugiyama H, Kazui H, Shigenobu K, *et al*. Predictors of prolonged hospital stay for the treatment of severe neuropsychiatric symptoms in patients with dementia: a cohort study in multiple hospitals. *Int Psychogeriatr* 2013;25:1365–73.
14 Bel N, Koster C, Villegas M. Cross-lingual text categorization. *Proceedings ECDL*. 2003.
15 Dai W, Chen Y, Xue GR, *et al*. Translated learning: transfer learning across different feature spaces. *NIPS*. 2009.
16 Fan R, Chang K, Hsieh XC, *et al*. LIBLINEAR: a library for large linear classification. *J Mach Learn Res* 2008;9:1871–4.