HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2016 December 01.

Published in final edited form as:

J Biomed Inform. 2015 December ; 58: 168–174. doi:10.1016/j.jbi.2015.10.006.

Implications of non-stationarity on predictive modeling using EHRs

Kenneth Jung[1] and **Nigam H. Shah**[2]

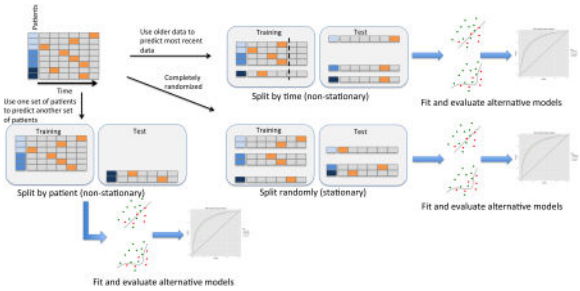
[1]Program in Biomedical Informatics, Stanford University, Stanford CA

[2]Center for Biomedical Informatics Research, Stanford University, Stanford CA

Abstract

The rapidly increasing volume of clinical information captured in Electronic Health Records (EHRs) has led to the application of increasingly sophisticated models for purposes such as disease subtype discovery and predictive modeling. However, increasing adoption of EHRs implies that in the near future, much of the data available for such purposes will be from a time period during which both the practice of medicine and the clinical use of EHRs are in flux due to historic changes in both technology and incentives. In this work, we explore the implications of this phenomenon, called *non-stationarity*, on predictive modeling. We focus on the problem of predicting delayed wound healing using data available in the EHR during the first week of care in outpatient wound care centers, using a large dataset covering over 150,000 individual wounds and 59,958 patients seen over a period of four years. We manipulate the degree of non-stationarity seen by the model development process by changing the way data is split into training and test sets. We demonstrate that non-stationarity can lead to quite different conclusions regarding the relative merits of different models with respect to predictive power and calibration of their posterior probabilities. Under the non-stationarity exhibited in this dataset, the performance advantage of complex methods such as stacking relative to the best simple classifier disappears. Ignoring non-stationarity can thus lead to sub-optimal model selection in this task.

Graphical abstract



Corresponding author: Kenneth Jung, kjung@stanford.edu, tel: (415) 640-5289.

1. Introduction

The rapid adoption of electronic health records (EHRs) is a key enabler of the learning healthcare system [1-5]. One effect of EHR adoption is to vastly increase the amount of data available for tasks such as predictive modeling of clinical outcomes. This increase in data enables developers of such models to employ increasingly sophisticated models to improve performance without overfitting. For example, recent work has applied tensor factorization to discover latent disease subtypes [6]. Such models are a far cry from the logistic regression models that have long been a mainstay of clinical research, and have the potential to transform clinical care. However, the observational nature of EHR derived data raises several practical issues in the development of such models [7-9]. EHR data may be incorrect and incomplete, and the majority of such data is collected primarily for billing purposes. Furthermore, some medical interventions could lower the risk of a particular outcome of interest and the popularity of these medical interventions can change over time as practices change. These factors can affect models that treat labels as unchanging truths. Failure to take these issues into account in the development and deployment of these models could lead to high profile failures that could ultimately delay the learning healthcare system [10, 11].

In this paper, we note that EHRs typically have repeated observations of a constantly evolving set of patients. Furthermore, we note that the health care system in the United States is currently, and for the foreseeable future, in a state of flux, with new systems being adopted and clinical practice evolving at a rapid pace as incentives change. Indeed, we note that this situation is in fact an explicit goal of the learning health care system [1, 5]. In the spirit of Walsh et al [12], which examined the effect of data source, cohort selection and prediction target on the performance of a logistic regression model of hospital readmissions, we explore the effects these changes have on predictive modeling using EHR data.

We focus on the development of a predictive model for delayed wound healing using a dataset previously described in Jung et al [13], which described the development of a predictive model for delayed wound healing and its potential clinical utility. The dataset consists of wound and patient data collected over the course of care at outpatient wound care centers operated by Healogics Inc between 2009 and 2013. In this setting, patients are seen on a weekly basis to monitor the progress of wound healing and adjust care as appropriate. Quantitative and categorical descriptions of wounds are entered into an EHR during each such assessment. The objective of the model is to predict whether or not a given wound will be an outlier with respect to how long it takes to fully heal, given only information collected during the first and second wound assessments. The threshold for delayed wound healing was set to fifteen weeks based on the observations of clinical experts at Healogics. Given accurate prognostic information, it is possible to triage patients for additional care such as additional monitoring and at-home care, hyperbaric oxygen therapy (HBOT), and negative pressure wound therapy (NPWT). Thus an accurate prediction has the potential to change the course of clinical treatment.

Non-stationarity is broadly defined as occurring when the data generating process being modeled changes over time. In this study, the data generating process is the routine care of wounds, as captured by patient and wound information recorded in the EHR. This

information is presumed to be informative about delayed wound healing, and so we fit predictive models that use the EHR data to predict that outcome. Changes in the wound care process over time may render the evaluation and use of these models problematic because the joint distribution of covariates and delayed wound healing giving rise to the training data may not be the same as that giving rise to test data used to evaluate the models, and to future data.

Research into classification under non-stationarity has focused on two tasks – detecting non-stationarity (referred to as *anomaly detection* or *change detection*), and learning under non-stationarity. Moren-Torres et al [14] provides an overview of non-stationarity and related issues, while Hoens et al [15] summarizes current methods for dealing with non-stationarity. In brief, these methods modify the dataset or the model in response to new data such that more weight is given to the most recent data. However, these methods have for the most part been used only in domains such as online fraud detection; to the best of our knowledge, they have never been applied to clinical risk prediction. In this study, we aim to characterize the implications of non-stationarity on the development of predictive models of the sort commonly encountered in clinical informatics.

To that end, we present experiments evaluating the impact of non-stationarity on discriminative power (how well models distinguish between cases and non-cases) and on model calibration (how closely the posterior probabilities of delayed wound healing output by models match observed frequencies of delayed wound healing). We approximate different degrees of stability of the data generating process by changing the way that the data is split into training and test sets. We then examine how such change impacts model selection. To that end, we consider the use of increasingly sophisticated models, starting from regularized logistic regression, progressing through non-linear models capable of automatically modeling interactions between predictors, and ending with ensemble methods that combine the predictions of many base models. Finally, we examine the impact of non-stationarity on engineered, domain specific features.

We demonstrate that in a setting that approximates a stationary data distribution, methods such as stacking can provide significant boosts to predictive power relative to the best base models. However, this performance gain disappears when the data distribution is non-stationary. In both cases, however, there is consistent benefit from using engineered, domain specific features. We find that using non-linear models that capture feature interactions automatically is useful in this dataset but that the benefit from such models is reduced under non-stationarity. Our findings emphasize the importance of matching the model development process with the intended use of the model. If the model is intended for use on future patients, it is critical to take non-stationarity into account to obtain a reliable estimate of model performance.

2 Materials and Methods

Our goal is to investigate the impact of non-stationarity on a predictive model for delayed wound healing, defined in this study as whether or not a given wound will take longer than 15 weeks to heal using information routinely collected during the first week of care. We

approach this by fitting a series of increasingly complex models—with and without domain specific features—to different training and test splits of the data. We observed that the dataset exhibits substantial non-stationarity. We can, however, control the degree of non-stationarity seen by the models by changing the way we split the data. This process is summarized in Figure 1 and explained further in Section 2.2. We evaluate the models for discriminative power and calibration under these different conditions. In the remainder of this section, we provide details about the dataset, feature construction, model development and evaluation.

2.1 Dataset

The dataset is comprised of 1,182,751 time-stamped wound assessments performed at 68 Healogics outpatient wound care centers distributed over 26 states. These wound assessments represented 180,716 unique wounds. Each wound assessment consists of both quantitative information regarding a specific wound, such as length, width, depth and area, in addition to categorical descriptors such as wound type, anatomical location, presence/absence of erythema and ICD9 codes associated with the assessment. Each assessment is also associated with unique wound and patient keys, allowing us to associate each wound with basic demographic information such as age, sex, and insurance status along with its outcome. Wound assessments were performed approximately weekly, and the dataset spans 2009 through 2013. A total of 59,958 patients are represented, and there are no restrictions on patients or wound types. Supplementary Materials Table 1 provides additional demographic details about the dataset, broken down by wound center.

We removed any wounds that were unresolved by the end of the study period unless the wound was already past the 15-week threshold for delayed healing. We also removed wounds with negative or very large values for quantitative features (> 99.9th percentile) or with clearly erroneous demographic information such as negative age. This left us with 150,277 unique wounds for use in training and testing our models. The basic features for our models are the data for each wound that is available at the time of the first wound assessment.

We performed additional pre-processing of the dataset as follows. First, ICD9 codes were aggregated to 3 digit codes. Second, wound types and locations were collapsed into 40 and 37 values from 103 and 216 values, respectively, in order to account for variation in how these variables were recorded in different wound care centers and to aggregate values that were judged to be clinically equivalent (upon manual review by KJ) for the purposes of the predictive model. For instance, the locations ‘Arm – Elbow’ and ‘Elbow’ were both mapped to the single location ‘Elbow’, and ‘Foot – 2nd Toe’ and ‘Foot – 3rd Toe’ were collapsed to ‘Toe’. Third, insurance information was collapsed into four categories - uninsured, private, Medicaid, and Medicare.

In this study, delayed wound healing is defined as taking 15 or more weeks to heal; this threshold was chosen based on the advice of domain experts from Healogics. 11.7% of wounds met this criterion in the final dataset.

The dataset has several characteristics that we believe make it especially illuminating with respect to developing and potentially deploying predictive models in practice. First, the dataset consists of longitudinal observations of multiple wounds from many patients, with no restrictions on the patient population or wound types. Second, the data was collected from 67 geographically distributed wound care centers. Finally, Healogics expanded its operations by both opening new would care centers and taking over the operations of other specialized wound care systems during the study period. Healogics was also increasing adoption of its EHR, and it was apparent from preliminary examination of the data that there were significant changes in EHR use over the study period.

2.2 Training and test splits

We control the degree of non-stationarity seen by the models by changing the way the dataset is split into training and test sets. A naive strategy is to randomly assign wounds into the training and test sets independently of each other, ignoring both time and the fact that a single patient may have many wounds in the dataset. This strategy mirrors the assumption that the training and test data are from the same distribution, and may be appropriate in settings in which this is known or strongly suspected to be true. A second strategy is to split patients into training and test sets, and then assign wounds to training and tests sets according to this patient assignment. This strategy respects the grouping of wounds by patient, but ignores time. It approximates the population of patients undergoing a large change in the future. It may also yield a pessimistic estimate of model performance because it is estimated on completely new patients; in practice, a wound healing prediction model would see a mix of new and previously seen patients. Finally, we simulate a prospective setting by assigning wounds from the end of the study period to the test set. This procedure ignores the grouping of wounds by patient, but respects the fact that we wish to make predictions about the future, and not past events whose outcomes are unknown. Under this setting, the model development process “sees” the non-stationarity in the data. In all cases, data was split 4:1 into training and test sets. In the case of the simulated prospective setting, we performed this split by ordering the wound assessments in time, and selecting the most recent fifth of the data for the test set.

2.3 Model choice

We evaluated the following models: L1 regularized linear regression (LASSO) [16], random forest (RF) [17, 18], neural networks (NN) [19, 20], and gradient boosted trees [21-23]. Each of these models were fit to the training set for each of the splits, with their respective hyperparameters tuned by cross validation on the training data (LASSO) or on a held out subset of the training data. Following hyperparameter tuning, the models were refit to the entire training set.

We also evaluated the use of two model ensemble methods – model averaging and stacking – that combine the outputs of the above base models [24, 25]. Model averaging simply averages the predictions of the base models. We averaged the predictions of the four base models trained on the full training sets. In our stacking experiments, we further split the training sets into two parts – one for training base models (train-base), and another for training the stacked model (train-stack). We fit the base models to the train-stack dataset.

The stacked model was a gradient boosted tree model trained on the train-stack data augmented with the posterior probabilities output by the base models on that data.

2.4 Model fitting

Here we provide details on the training of the models. Unless otherwise specified, all development was performend in R version 3.02. For the LASSO, we used the R package glmnet [16]. Hyperparameter optimization was performed automatically on the training data by 10-fold cross validation. We used the R package randomForest [17] to train the random forest model. 2500 trees were fit, and all were used during evaluation. Gradient boosted tree models were trained using the R package gbm [23] with Bernoulli loss, an interaction depth of 6, shrinkage of 0.0025. The number of trees was determined by fitting 15,000 trees to 90% of the training set; models were then fit to the full training set using the number of trees that resulted in the best loss on the 90% training set. Our neural net model was developed in Python using the pylearn2 framework [20]. We used a three hidden layer neural net with maxout activations in the hidden layers. Each hidden layer consisted of 25 maxout units, each with 5 linear regions. The nets were regularized with dropout of 0.5. The initial learning rate was 0.1, decreasing exponentially with each epoch. Momentum was also used, starting at 0.5 and increasing linearly such that it would have reached 0.99 at 500 epochs. Each hidden unit's weights were column norm constrained to 5. Training was performed as follows. First, the training set was further split 4:1 into training and tuning sets. A net was then trained on versions of the resulting training set that had the minority class (delayed wound healing) up-sampled to 50% of the data for 20 epochs, at which time the net was learning rapidly. Training was then continued on the unbalanced versions of the training data for 150 epochs, and then further trained on the full (unbalanced) training set until the loss on the tuning data was equal to the loss achieved on the training set previously.

Our stacked model consisted of a gradient boosted tree model that used the original set of features, plus the outputs of the four base models described above, as inputs. In order to train the stacking model, we first split the training set 4:1 into a training set for training the base models (train-base) and another, smaller training set for training the stacked model (train-stack). After training the base models on train-base, we obtained the base model outputs for the train-stack dataset and the test set. These outputs were concatenated to the feature vectors in each of these datasets. The stacked model was trained on the train-stack dataset and evaluated on the test set.

2.4 Feature engineering

We constructed additional features for each wound that measure the change in quantitative variables such as wound dimensions between the first and second assessments, and features that summarize the total wound burden of each patient at the time of the first assessment, such as number of wounds and total surface area. These features are intended to allow models to accurately capture the fact that wounds that decrease in size dramatically during the first week of recovery are healing well. In all, a total of 833 features were calculated for each wound, the bulk of which were binary indicators for the levels of categorical variables (31 quantitative versus 802 binary).

2.5 Model evaluation

A useful prognostic model for delayed wound healing should meet two criteria. First, it should be able to discriminate between positive versus negative cases of the delayed wound healing. Second, it should output well-calibrated posterior probabilities of delayed wound healing, i.e., the observed frequency of delayed wound healing should be close to the predicted probability. We evaluated the test set performance of the various models under each all combinations of training-test splits and feature sets. Discriminative power on the test data was evaluated using the Area under the Receiver-Operator Characteristic (ROC) Curve (AUC), which summarizes the sensitivity-specificity trade off across the range of possible thresholds for calling an example a positive case. We obtained 95% confidence intervals for the AUCs by bootstrapping on the test set (N = 30056, 29804, and 30028 for the random, patient, and prospective splits, respectively).

Model calibration was evaluated by Brier reliability on the test data. Brier reliability is a measure of calibration for probabilistic predictions that are stratified into discrete groups, with lower values indicating better agreement between the predictions in each stratum and the observed frequency in those strata[26]. It is equivalent to the mean squared deviation between the predicted probability and the observed frequency within each stratum, weighted by the number of observations in each stratum. We formed strata by dividing the interval [0,1] into ten equal sized bins. In other words, all samples with predicted probabilities between 0 and 0.1 fall into one bin, between 0.1 and 0.2 in the next, etc. We also evaluate calibration visually, by plotting reliability diagrams: for each bin, the mean predicted probability is plotted against the observed fraction of delayed wound healing [27]. Well-calibrated models will thus have its reliability diagram near the diagonal line.

3. Results

3.1 Non-stationarity in the dataset

Our aim is to study the impact of non-stationarity on predictive modeling of delayed wound healing using EHR data. We first established the dataset naturally exhibits non-stationarity in both the covariates and the prevalence of the outcome of interest. For instance, prior to 2013 there were many instances of wounds whose type was coded as ‘Diabetic Wound, Lower Extremity’. From 2013 onwards, this wound type almost entirely disappeared, and was superseded by more specific categories such as ‘Neuropathic diabetic wound’ and ‘Neuro-ischemic diabetic wound’. Furthermore, the prevalence of delayed wound healing was lower in 2013 (i.e., the most recent data available for this study) than previously (10.4% versus 13.2% respectively). Finally, a gradient boosted tree model trained to predict whether a case was from 2013 or prior to 2013 achieved an AUC of 0.986 in a hold out test set.

3.2 Discrimination

Our measure of discriminative power is AUC in the test set (Figure 2), with 95% confidence intervals obtained by bootstrapping for 10,000 iterations. If the data generating process is stationary, i.e., we split the dataset into training and test sets without using patient and temporal information, we find that the non-linear models – random forest, neural networks, and gradient boosted trees – have a significant edge over the linear model (AUCs of 0.861,

0.854, and 0.861 versus 0.827 respectively). Furthermore, combining the outputs of these base models by model averaging and stacking provide additional benefit over these non-linear models (AUCs of 0.872 and 0.876 respectively). These results suggest that one should use the most complex model, stacked gradient boosted trees.

However, under the most realistic training/test split, i.e., a split in which we trained on older data and evaluated on the most recent data, we see that the relative performance of the models is almost the same, with much less difference in performance between the base models, and no benefit from model averaging and stacking (AUCs of 0.879 and 0.880, respectively) relative to the best base classifier, gradient boosted trees (AUC of 0.881). However, we also find that gradient boosted trees still significantly outperform the other base models (AUCs ranging from 0.862-0.867). We note that the absolute performance is somewhat higher in the prospective split setting than the random split setting, but do not generalize from this except to note that it underscores the impact of non-stationarity on estimates of model performance.

When training and test data is split by patient, we find that gradient boosted trees remain the best performing base model but with a smaller advantage over the other models (AUC of 0.843 for gradient boosted trees versus 0.819-0.826 for the other base models). Model averaging and stacking provide no additional benefit over gradient boosted trees (AUC 0.846 for stacking versus 0.843 for gradient boosted trees).

We are also interested in the interaction between stationarity and the use of domain specific features such as those that encode the initial rate of wound closure. Without using these features, we observe the same trends described above, with gradient boosted trees being the best performing base model under all training/test split procedures, but with a smaller edge over the other models. Table 1 summarizes the discrimination performance.

3.3 Calibration

Well-calibrated posterior probabilities are desirable for prognostic models because they enable clinicians and patients to make decisions that are informed by accurate estimates of the probabilities of the outcomes of interest, which is especially important when performing cost-benefit analysis of treatment options. We evaluated model calibration by calculating Brier reliability in the test set (Table 2). Here, the general trend is that under all conditions, the linear and stacked models are the best calibrated. Figure 3 shows reliability diagrams for the models under the various conditions. In Table 2 and Figure 3a, we see that under conditions of stationarity, we observe that the lasso, neural net, and stacked models all have very good calibration, with the mean posterior probability of delayed wound healing matching the observed frequency of delayed healing in each of the bins. The RF and gradient boosted tree models, on the other hand, seem to have relatively poor calibration. However, the situation changes under patient and prospective splits. Under conditions of non-stationarity, the gradient boosted tree and stacked models offer the best calibration.

4. Discussion

The increasing abundance of EHR derived clinical data enables researchers to apply increasingly sophisticated models to clinical problems. Real world processes generating the EHR data are highly non-stationary, driven by factors such as rapidly evolving financial incentives, clinical practice, as well as adoption of and adaptation to new technology. Therefore, we investigated what these changes imply for researchers in biomedical informatics, especially those engaged in predictive modeling.

The development and use of predictive models must often balance competing objectives. For instance, model parsimony and interpretability, exemplified in this study by regularized logistic regression, must often be weighed against the substantial performance gains afforded by more complex models such as random forests and gradient boosted trees. When evaluating these tradeoffs, we typically rely on estimates of performance measured in hold out tests sets or obtained through cross validation. But in the presence of non-stationarity, i.e., when the joint distribution of covariates and the outcome of interest is changing over time, such estimates may be misleading. Consider a scenario in which recent data is substantially different from older data. A naïve split of data into training and test sets that ignored time would randomize this difference away, and test set performance will reflect accuracy on a mixture of old and new data. Thus, this test set estimate of model accuracy may be misleading.

We have presented a case study in which a large dataset was used to predict delayed wound healing in outpatient wound care centers. This dataset empirically exhibits significant non-stationarity. By changing the way in which we split the dataset into training and test sets we were able to evaluate the impact of this non-stationarity on the models. We found that under a stable data distribution, a regularized logistic regression model achieved an AUC of 0.827 while the best and worst non-linear classifiers achieved AUCs of 0.876 and 0.854, respectively. Thus, there was substantial benefit from complex models over the baseline of a linear classifier. We might conclude from these results that the extra complexity of the stacked model, which uses the original set of covariates in addition to the outputs of several base classifiers to arrive at a prediction, is worthwhile due to its substantially higher accuracy and excellent calibration.

However, this benefit was greatly diminished when we trained the models on older data and evaluated them on the most recent data. Under that condition, the regularized logistic regression model achieved an AUC of 0.867, while the best worst non-linear models achieved AUCs of 0.881 and 0.862, respectively. Thus, not only did the simplest model outperform at least two of the more complex, non-linear models, it substantially closed the gap to the best performing model. Furthermore, the advantage that the stacked model over the other models disappeared in this condition – the best performing model was a single base classifier, gradient boosted trees. We argue that this “prospective evaluation” scenario provides the most relevant evaluation of model performance. If we assume that the most recent data most closely reflects future data, then it appears that we should use gradient boosted trees instead of the stacked model if our primary concern is accuracy, or that we can

opt to use a much simpler linear model without nearly as much of a loss of accuracy as the naïve, random split suggested. Our experience in this study suggests the following lessons.

First, when one is uncertain about the possibility or presence of non-stationarity, it may be best to stick with simpler models. Conversely, when one is confident that the data distribution is stable and will remain so for some time, more complex models, including methods such as model averaging and stacking, may provide substantial gains in discriminative power and calibration. We emphasize, however, that we do not endorse a particular model as most resistant to non-stationarity. In the present study, gradient boosted trees, the stacked model and model averaging all outperform the baseline of a linear classifier in all conditions, and in some settings this performance benefit may outweigh the relative interpretability and presumed robustness of simpler models. However, we cannot conclude that these results generalize to different tasks. Rather, we can only beware of non-stationarity during model development to guard against unwelcome surprises in actual use.

Second, given the choice between spending resources on increasingly sophisticated models and the engineering of domain-specific features, it may be advisable to focus on the latter. In this dataset, the benefit from using features that encode the initial rate of wound closure are substantial, and consistent across all conditions, implying that the utility of these features is stable even when the data distribution is non-stationary. Of course it is possible that particular features may not be so robust in other prediction problems.

Finally, we argue that the most important lesson to draw from this study is that one ought to be very clear about the intended purpose of the model being developed, and use a model development process that reflects that purpose. For instance, if one is developing a model whose purpose is cohort selection for a retrospective study, e.g., a classifier for finding patients who had a specific condition in the past, then making predictions on unknown cases from the past is a useful task, and it may make sense to train and evaluate models without regard to time. In the case of our wound healing prediction model, however, it does no good to evaluate by predicting delayed wound healing on past cases. Rather, we are interested in making predictions of outcomes for cases arising in the future, and for both previously seen patients (i.e., someone who who has previously been treated and now has a new wound) and completely new patients (i.e., someone who has a wound for the first time). The model development process should reflect this use by evaluating on test data that is from a later time than the training and validation data.

5. Conclusions

Non-stationarity in the data generating process can have a substantial impact on the development of predictive models. We have evaluated this effect in the context of models of delayed wound healing using a dataset that exhibits significant natural non-stationarity. Model development was carried out under different degrees of non-stationarity simulated by using different train-test split procedures. The relative merits of a range of models of varying complexity were dependent on the presence or absence of non-stationarity. Under conditions approximating stationarity, the most complex model, a stacked model that used the original covariates in addition to the outputs of four base classifiers, achieved a significantly higher

AUC than all other models, and also showed excellent calibration. However, in the non-stationary setting, this advantage disappeared and the best model was a single base classifier, gradient boosted trees. Furthermore, the gap between the best model and the simplest model was substantially reduced. Thus, it is imperative that model development reflects the intended use of the model. If one is principally interested in a classifier for retrospective data, then a random split and thus the stacked model would be very attractive. However, if one is most interested in using the model on future data, then it appears that a substantially simpler model is best, and furthermore that a very simple model, L1 regularized logistic regression, does almost as well. We caution that these conclusions may not generalize to all other tasks.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Conflict of Interest Declaration: This study was funded by a research grant from Healogics Inc. NHS and KJ acknowledge additional funding from NIH grant U54 HG004028 for the National Center for Biomedical Ontology, NLM grant R01 LM011369, NIGMS grant R01 GM101430, and the Smith Stanford Graduate Fellowship. NS is scientific advisor and co-founder of Kyron Inc, and is an advisor to Apixio Inc.

References

1. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. Science translational medicine. 2010 Nov 10.2(57):57cm29.
2. Weng C, Appelbaum P, Hripcsak G, Kronish I, Busacca L, Davidson KW, et al. Using EHRs to integrate research with patient care: promises and challenges. Journal of the American Medical Informatics Association : JAMIA. 2012 Sep-Oct;19(5):684–7. [PubMed: 22542813]
3. Shah NH, Tenenbaum JD. The coming age of data-driven medicine: translational bioinformatics' next frontier. Journal of the American Medical Informatics Association : JAMIA. 2012 Jun; 19(e1):e2–4. [PubMed: 22718035]
4. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff (Millwood). 2014 Jul; 33(7):1123–31. [PubMed: 25006137]
5. Embi PJ, Payne PR. Evidence generating medicine: redefining the research-practice relationship to complete the evidence cycle. Medical care. 2013 Aug; 51(8 Suppl 3):S87–91. [PubMed: 23793052]
6. Ho JC, Ghosh J, Steinhubl SR, Stewart WF, Denny JC, Malin BA, et al. Limestone: high-throughput candidate phenotype generation via tensor factorization. Journal of biomedical informatics. 2014 Dec.52:199–211. [PubMed: 25038555]
7. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Medical care. 2013 Aug; 51(8 Suppl 3):S30–7. [PubMed: 23774517]
8. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton G. Bias associated with mining electronic health records. Journal of biomedical discovery and collaboration. 2011; 6:48–52. [PubMed: 21647858]
9. Paxton CNMA, Saria S. Developing predictive models using electronic medical records: challenges and pitfalls. AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2013; 2013:1109–15. [PubMed: 24551396]
10. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google flu trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PloS one. 2011; 6(8):e23610. [PubMed: 21886802]

11. Boland MR, Hripcsak G, Shen Y, Chung WK, Weng C. Defining a comprehensive verotype using electronic health records for personalized medicine. Journal of the American Medical Informatics Association : JAMIA. 2013 Dec; 20(e2):e232–8. [PubMed: 24001516]
12. 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 Dec.52:418–26. [PubMed: 25182868]
13. Jung K, Sheppard D, Covington S, Sen CK, Januszyk M, Kirsner RS, et al. Rapid identification of slow healing wounds. Annals of Surgery. 2015 In review.
14. Moreno-Torres JG, Raeder T, Alaiz-Rodriguez R, Chawla NV, Herrera F. A unifying view on dataset shift in classification. Pattern Recognition. 2011; 45:521–30.
15. Hoens TR, Polikar R, Chawla NV. Learning from streaming data with concept drift and imbalance: an overview. Progress in Artificial Intelligence. 2012; 1:89–101.
16. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010; 33(1):1–22. [PubMed: 20808728]
17. Liaw A, Wiener M. Classification and Regression by random Forest. R News. 2002; 2(3):18–22.
18. Breiman L. Random Forests. machine Learning. 2001; 45(1):5–32.
19. Goodfellow IJ, Warde-Farley D, Mirza M, Courville A, Bengio Y. Maxout Networks. Journal of Machine Learning Research. 2013; 28(3):1319–27.
20. Goodfellow, IJ.; Warde-Farley, D.; Lamblin, P.; Dumoulin, V.; Mirza, M.; Pascanu, R., et al. arXiv preprint arXiv. 2013. Pylearn2: a machine learning research library; p. 13084214
21. Friedman J. Stochastic Gradient Boosting. Computational Statistics & Data Analysis. 2002; 38(4): 367–78.
22. Friedman JH. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 2001; 29(5):1189–232. 2001/10. en.
23. Ridgeway, G. gbm: Generalized Boosted Regression Models. R Package Ver.2.1. 2007. http://cran.r-project.org/web/packages/gbm/
24. Wolpert DH. Stacked generalization. Neural Networks. 1992; 5(2):241–59.
25. Ting KM, Witten IH. Issues in stacked generalization. Journal of Artificial Intelligence Research. 1999; 10:271–89.
26. Stephenson DB, Coelho CAS, Jolliffe IT. Two Extra Components in the Brier Score Decomposition. Weather and Forecasting. 2008; 23(4):752–7. 2008/08/01.
27. DeGroot M, Fienberg S. The comparison and evaluation of forecasters. Statistician. 1982; 32

Highlights

- We study the effect of non-stationarity on predictive models of delayed wound healing.
- We apply different techniques for splitting our dataset into training and test sets to simulate different types of non-stationarity.
- Estimates of model performance change under different types of non-stationarity.
- Researchers should account for the possibility of non-stationarity during model development.

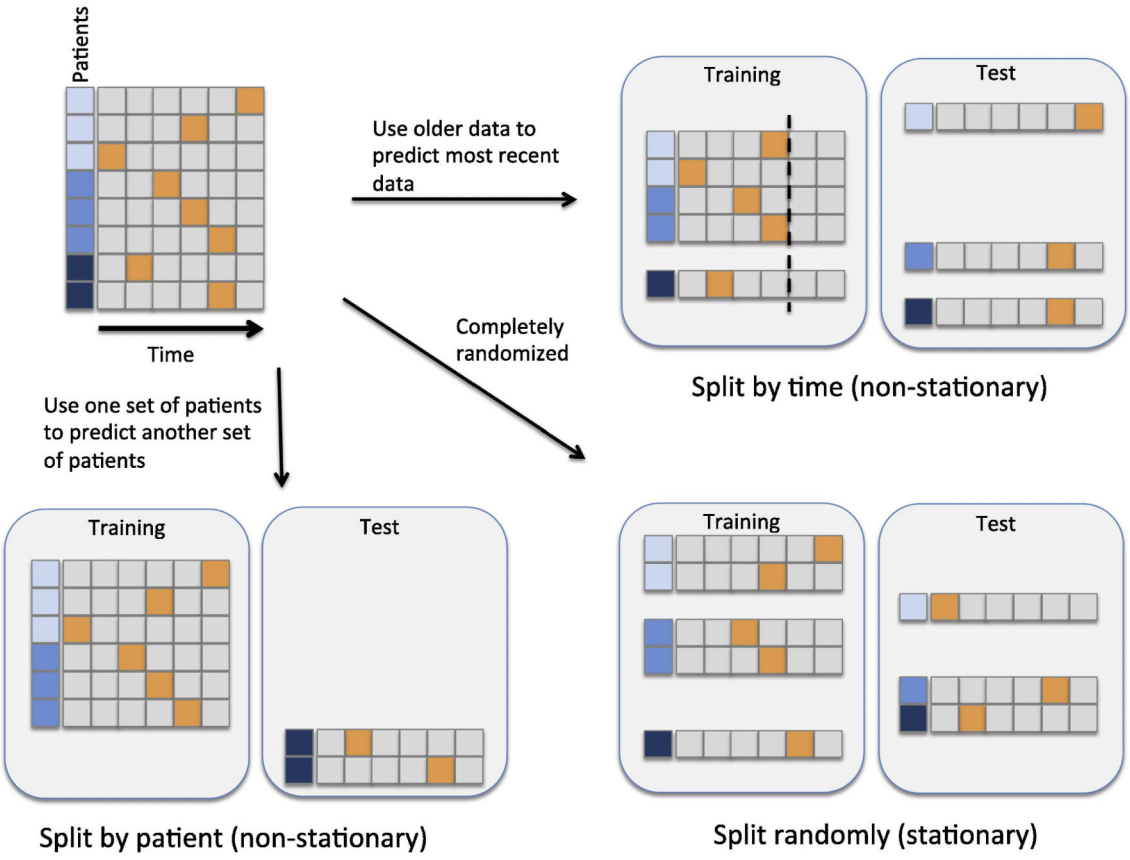


Figure 1.

Approximating different degrees of stationarity using training and test splits. Each row represents a wound. The first column represents the patient for a given wound, and is color-coded such that different shades indicate different patients. The time of the first wound assessment is indicated by the orange cell in the grey columns. In the random split, wounds are randomized into training and test directly. In the patient split, patients are randomized and wounds from a given patient go together. In the prospective split, we assign wounds from the end of the study period (shown by the dashed line) to the test set.

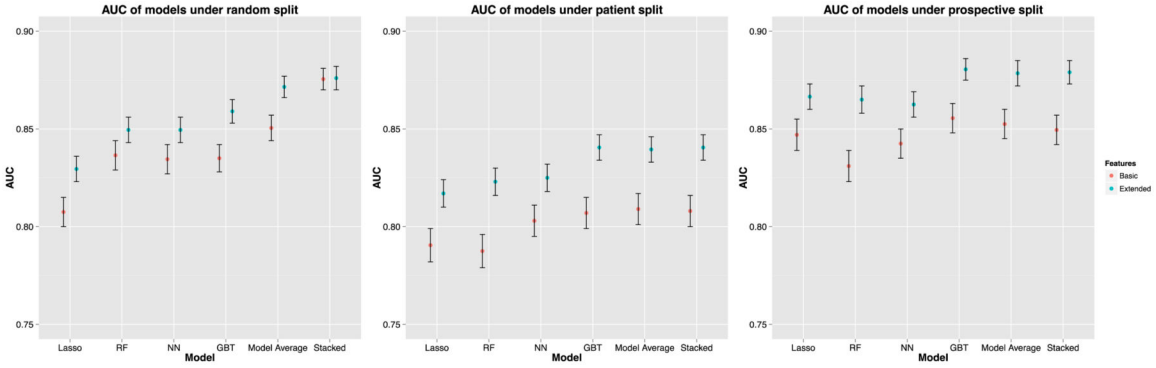


Figure 2.

AUC of models under different conditions – mean of 10,000 bootstrap estimates with 95% confidence intervals. These plots show the relative performance of different models under different degrees of non-stationarity. The models are abbreviated as Lasso for L1 regularized logistic regression, RF for random forest, NN for neural nets, GBT for gradient boosted trees. Each model is trained and evaluated twice under each non-stationarity condition, using either the basic feature set (Basic) or an extended feature set (Extended) that includes features such as the total wound burden of the patient and the rate of wound healing observed over the first week of care.

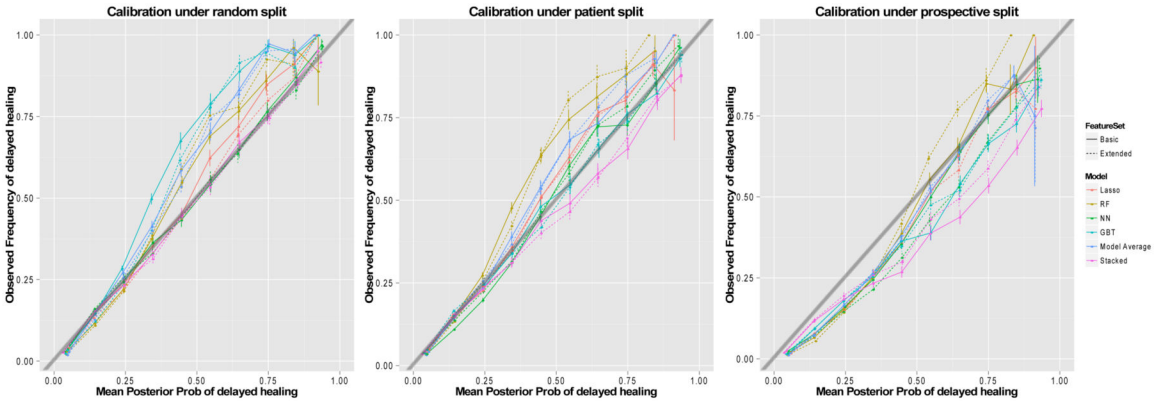


Figure 3.

Calibration of models under different conditions. These plots show calibration curves of different models under different degrees of non-stationarity. The models are abbreviated as Lasso for L1 regularized logistic regression, RF for random forest, NN for neural nets, GBT for gradient boosted trees. Each model is trained and evaluated twice under each non-stationarity condition, using either the basic feature set (Basic) or an extended feature set (Extended) that includes features such as the total wound burden of the patient and the rate of wound healing observed over the first week of care.

Table 1

Model discrimination measured by AUC.

Model	Random split		Prospective split		Patient split	
	With engineered features	No engineered features	With engineered features	No engineered features	With engineered features	No engineered features
Lasso	0.827	0.807	0.867	0.847	0.819	0.791
RF	0.861	0.837	0.866	0.831	0.825	0.788
NN	0.854	0.835	0.862	0.842	0.826	0.807
GBT	0.861	0.834	0.881	0.856	0.843	0.807
Mean	0.872	0.850	0.879	0.853	0.840	0.809
Stacked	0.876	0.875	0.880	0.850	0.841	0.808

Table 2

Model calibration measured by Brier reliability.

Model	Random split		Prospective split		Patient split	
	With engineered features	No engineered features	With engineered features	No engineered features	With engineered features	No engineered features
Lasso	0.000299	0.000256	0.00288	0.00352	0.000372	0.000339
RF	0.00205	0.0011	0.00510	0.00363	0.00163	0.00152
NN	0.000151	0.000136	0.00459	0.00459	0.000194	0.000732
GBT	0.00199	0.00227	0.00215	0.00278	0.000127	0.0000537
Mean	0.00175	0.00161	0.00318	0.00318	0.000764	0.000602
Stacked	0.0000793	0.000075	0.00224	0.00410	0.00035	0.000234