


Research and Applications

Estimating real-world performance of a predictive model: a case-study in predicting mortality

Vincent J. Major , Neil Jethani, and Yindalon Aphinyanaphongs

Department of Population Health, NYU Langone Health, New York, New York, USA

Corresponding Author: Vincent J. Major, MS, Department of Population Health, NYU Langone Health, 227 East 30th St, 6th Floor, New York, NY 10016, USA; Please vincent.major@nyulangone.org

Received 2 December 2019; Revised 5 March 2020; Editorial Decision 12 March 2020; Accepted 19 March 2020

ABSTRACT

Objective: One primary consideration when developing predictive models is downstream effects on future model performance. We conduct experiments to quantify the effects of experimental design choices, namely cohort selection and internal validation methods, on (estimated) real-world model performance.

Materials and Methods: Four years of hospitalizations are used to develop a 1-year mortality prediction model (composite of death or initiation of hospice care). Two common methods to select appropriate patient visits from their encounter history (backwards-from-outcome and forwards-from-admission) are combined with 2 testing cohorts (random and temporal validation). Two models are trained under otherwise identical conditions, and their performances compared. Operating thresholds are selected in each test set and applied to a “real-world” cohort of labeled admissions from another, unused year.

Results: Backwards-from-outcome cohort selection retains 25% of candidate admissions ($n = 23\,579$), whereas forwards-from-admission selection includes many more ($n = 92\,148$). Both selection methods produce similar performances when applied to a random test set. However, when applied to the temporally defined “real-world” set, forwards-from-admission yields higher areas under the ROC and precision recall curves (88.3% and 56.5% vs. 83.2% and 41.6%).

Discussion: A backwards-from-outcome experiment manipulates raw training data, simplifying the experiment. This manipulated data no longer resembles real-world data, resulting in optimistic estimates of test set performance, especially at high precision. In contrast, a forwards-from-admission experiment with a temporally separated test set consistently and conservatively estimates real-world performance.

Conclusion: Experimental design choices impose bias upon selected cohorts. A forwards-from-admission experiment, validated temporally, can conservatively estimate real-world performance.

LAY SUMMARY: The routine care of patients stands to benefit greatly from assistive technologies, including data-driven risk assessment. Already, many different machine learning and artificial intelligence applications are being developed from complex electronic health record data. To overcome challenges that arise from such data, researchers often start with simple experimental approaches to test their work. One key component is how patients (and their healthcare visits) are selected for the study from the pool of all patients seen. Another is how the group of patients used to create the risk estimator differs from the group used to evaluate how well it works. These choices complicate how the experimental setting compares to the real-world application to patients. For example, different selection approaches that depend on each patient’s future outcome can simplify the experiment but are impractical upon implementation as these data are unavailable. We show that this kind

of “backwards” experiment optimistically estimates how well the model performs. Instead, our results advocate for experiments that select patients in a “forwards” manner and “temporal” validation that approximates training on past data and implementing on future data. More robust results help gauge the clinical utility of recent works and aid decision-making before implementation into practice.

Key words: experimental design, data science, machine learning, reproducibility of results, mortality

INTRODUCTION

When building machine learning models with electronic health record (EHR) data, we need to decide how to select patients from the population to both learn and evaluate a predictive model. To be successful, we want our model to perform well beyond the experiment and into the proposed use-case. In order to estimate our model performance, we should validate using an unseen dataset that closely resembles real-world data in order to inform discussions on its potential utility in practice.

First, we must select an experimental cohort to train a predictive model. This selection process is complicated by realities of EHR data: some patients are seen multiple times and others lost to follow-up with uncertain outcomes. Many machine learning works employ experimental simplifications to ameliorate such practical data challenges. Examples include studies that use cross-sectional data, upsample rare events, or select a desirable patient population or phenotype. One thing each of these approaches have in common is the careful selection of which patients to include and when. In a literature review of recent medical informatics works including development of at least 1 predictive model, 4 of 13 employ one of these approaches (Supplementary Table S1). By removing patients, these simplified experiments are helpful to assess the feasibility of new applications, but are expected to give an optimistic impression of performance on flawed real-world data.

Validation is the phase of model development where any experimental simplifications impact performance. But not all validation is equal. Predictive models are typically validated *internally* from the same population,¹ often with subsampling methods² or explicit “holdout” cohorts.^{3,4} After internal validation, some models are further validated *prospectively* from a future time or *externally* from an outside patient population or environment. But, more often than not, model development only includes internal validation results.

Since any application of machine learning in practice involves learning from past patients to predict the future, estimates of model performance should replicate this temporal process. Although prospective validation is considered a higher standard than internal validation,² it cannot be employed during model development as it requires data to be collected live. An approximation of prospective validation can be performed during model development with *temporal* validation, where the most recent experimental data are selected for validation (rather than the typical random sample). Moreover, since the patients included in temporal validation are different, it can also be classified as a weak approximation of external validation.^{2,5} Temporal validation is a compromise between the practicality of random validation and the rigor of prospective (and external) validation as the “test set” distribution resembles that expected upon implementation.

Unfortunately, there remains sparse empirical results describing how different experimental and validation approaches impact model

performance when deployed into the real-world. Instead, existing literature often focuses on the benefits of and need for secondary studies of prospective and/or external validation. Many machine learning works do not translate into such secondary validation studies. For many different methods and applications, the only data that exist are that published with the first derivation work which, in many cases, only includes random validation (Supplementary Tables S1 and S2). With no subsequent temporal, prospective, or external validation to evaluate the bigger picture of generalizability,¹ no one can truly assess the feasibility of such applications without reimplementation on their data.

In this article, we revisit several experimental designs and validation strategies often used when constructing machine learning experiments from retrospective patient data. Within a setting of mortality prediction, we describe how these experimental choices impact the model development data and ultimately affect estimated real-world performance. We hope these results serve as an empirical reference for practitioners and add to the discussion of pragmatic machine learning experiments in healthcare.

End-of-life use-case

End-of-life interventions—those that are more palliative than curative—are often initiated too late in a patient’s life. Guidelines suggest palliative care is appropriate for anyone living with a chronic or serious illness that will eventually cause their death.⁶ However, palliative care is often only initiated in the last weeks or months of life.⁷ One reason for this may be that physicians are poor at estimating prognosis, typically being optimistic,^{8–10} ultimately influencing care decisions.¹¹ Practical systems that identify patients who may benefit from palliative care, or other end-of-life interventions, are needed.^{12,13}

Several recent works have described various machine learning methods to predict mortality as a proxy task for palliative care needs.^{4,14,15} However, the experimental designs employed often have limitations determined by how they apply inclusion criteria and select prediction visits from each patient’s history. Namely, the validation cohorts no longer resemble the real-world cohort expected upon deployment.

OBJECTIVE

The goal of this work is to compare several candidate experimental designs in terms of (1) their impact on model development data, and (2) the resulting predictive performance of a machine learning model. In particular, performance within 2 test cohorts will be compared and contrasted against that of a single, uniform “real-world” cohort that, unlike the development data, is not affected by any experimental choices.

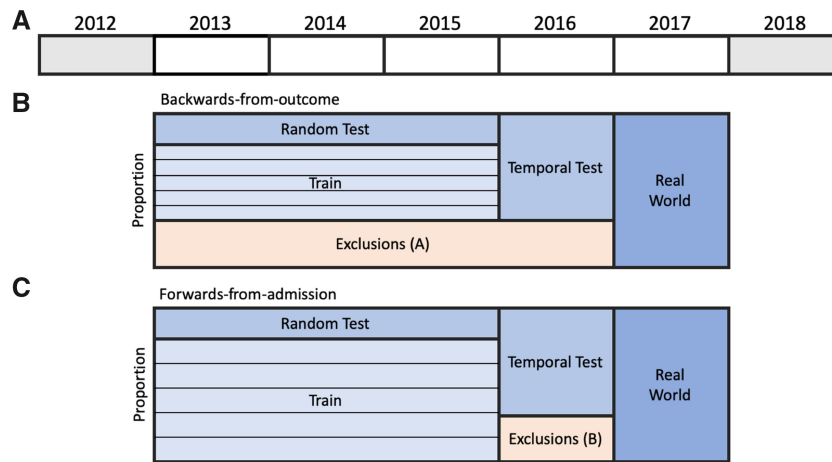


Figure 1. Schematic representation of available data and selected patient cohorts. (A) Seven years of available data leaves 5 calendar years of admissions. (B) Backwards-from-outcome cohort selection that only includes 1 admission per patient. (C) Forwards-from-admission cohort selection that includes readmissions within each set but excludes readmissions into another set. The thin horizontal lines inside the training cohort depict cross-validation folds employed for parameter tuning.

MATERIALS AND METHODS

Data

Prediction task

Mortality prediction is a common task, but challenges in modeling, experimental design, and implementation have limited translation into practice. Mortality modeling works often introduce mortality risk as a proxy for patient appropriateness for palliative care.^{4,14} In these proposed workflows, prediction timing is crucial. Inpatient settings are promising as specialized clinicians are available to initiate meaningful interventions. More specifically, prediction at, or near, admission can aid the care team to approach the entirety of a patient's admission cognizant of their risk.

Machine learning-based approaches to mortality prediction often simplify “time-to-event” analyses into binary outcomes of death within a certain time, commonly 1 year. As an objective for this work, we build on prior work and choose 1-year mortality as a target prediction task.⁴

Patient cohort

Data collected in routine clinical care of 1 urban, tertiary hospital are considered. Data are extracted from the clinical data warehouse where data are available from 2011. Admissions starting January 1, 2013 were considered for this work to allow prior patient data to accrue. At the time of data retrieval (March 2019), we consider admissions up to December 31, 2017, yielding 5 consecutive years of admissions as described in Figure 1A. However, the entire year of 2017 is held out of model development and will be used to simulate deployment.

Outcome data

Mortality data are innately noisy as both the mechanism of death and effect of any life-preserving interventions vary between patients. Moreover, data collection is complicated by the inability to observe deaths beyond our hospital system. Three sources of data are combined to form a composite measure: (1) internal medical center death data, (2) purchased death data (derived from the Social Security Administration's Master Death File), and (3) hospice discharges. None are perfect data capture mechanisms but together provide an

“end-of-life” label where the majority of patient outcomes are affirmed by at least 2 sources.

Data categories

The same broad data categories as prior works^{4,11} are considered as predictors: demographics, encounters, diagnoses, medications, procedures, and laboratory results. Each of these categories, except demographics and encounters, has structured ontology-based vocabularies (ICD-10 diagnoses, RxNorm medications, CPT procedures, and LOINC laboratory tests). Data of this type can be problematic as there are a very large number of highly specific codes where only a few codes are attributed to each encounter. From a modeling perspective, code-based data produce many sparse features which limits the number of feasible methods. Further details on the methods employed to mitigate these data challenges can be found in the [Supplementary Material](#).

Experimental factors to compare

Design of a retrospective experiment requires numerous choices. In this study, we introduce 2 particularly important decisions and compare 2 common approaches within each. Although intuitive arguments are often made, quantitative results from each of the 4 combinations will be presented similarly to enable direct comparisons.

Cohort selection: by outcome or admission?

Due to the complexity of hospital admission data, machine learning practitioners often subset the complete dataset to form a model development cohort using inclusion and exclusion criteria. The result is simplified and sanitized data that mitigate known challenges. Here, 2 contrasting approaches are presented: one that selects (instances) backwards from outcome and another that selects forwards from admission. Further details of both selection methods can be found in [Supplementary Figure S1](#).

Backwards-from-outcome cohort selection refers to a simplifying strategy that starts from a known outcome and works backwards to identify an appropriate prediction instant for each patient, in effect, a retrospective case-control study. This design is also known as right-censoring and is commonly employed in some medical ma-

chine learning applications.¹⁶ Model development is relatively simple with such a design as the raw clinical data is cleansed by excluding 2 particular challenges of the medical domain: (1) uncertain labels, due to patients who are lost to follow-up, and (2) multiple admissions per patient, where the 1-year mortality outcome may flip from negative to positive. The result is model development data that have less label noise.

Avati et al⁴ employ a backwards-from-outcome design to select one prediction instant per patient by working backwards from either (1) their known death, or (2) their last encounter. However, the authors enforce other criteria, namely, requiring 1 year of patient history. With such a design, Avati et al⁴ select a cohort of 221 284 patients from a population of approximately 2 million. These selection criteria are reproduced in this work.

Forwards-from-admission cohort selection refers to a design-oriented forwards from each admission, analogous to recruitment of a prospective cohort study, sometimes referred to as a retrospective cohort study or left-censoring. The result is a cohort followed from their admission date without foresight of their outcome. Naturally, some patients are lost to follow-up within the 1-year observation period, or are subsequently readmitted. One advantage of forwards-from-admission selection is the more intuitive inclusion of readmissions, enabling a patient's risk to evolve over time. While selecting forwards-from-admission is more realistic of deployment of medical applications, it does add experimental complexity which may discourage practitioners.

Test set selection: random sample or temporal separation?

To evaluate how a model performs on unseen patients during model development (i.e. internal validation), a test set must be selected from the complete population before model training. How the test set is separated from the population determines how test set performance estimates generalizability. Here, 2 common methods of selecting a test set are compared:

Random test set. A random test set includes patients randomly sampled from the same population and time period as the training data. In doing so, this test set measures generalizability. This design assumes sampling will ensure that the testing and training distributions are similar, which may not hold in practice. Several recent mortality prediction works employ a random test set in their work.^{4,14,15}

Temporal test set. A temporal test set includes patients from the same population but selects patients that are separated temporally from the training data. In doing so, this type of test set also measures the historic transportability² of the model into the near future, likely more challenging than generalizing. This test set is more closely aligned with the deployment process of training on prior data and deploying on new patients at some future time.

Combining cohort selection and test sets

Both backwards-from-outcome and forwards-from-admission experiments enable random and temporal test sets. As evidenced by a literature review of recent works, most rely on random test sets regardless of their selection method (see [Supplementary Material](#)). The approach to select a test set often represents the practitioner's experimental goals: complex applications often benefit from the relative simplicity of backwards-from-outcome selection with a random test set. However, this simplicity complicates generalization to a real-world application, hindering decision-making. To compare approaches, both random and temporal test sets are utilized leaving 1 training cohort for each selection method, as depicted in [Figure 1](#).

Each individual patient should only exist in 1 cohort to mitigate data "leakage" from training to testing. Backwards-from-outcome selection is simple as it yields 1 prediction per patient, whereas forwards-from-admission selection is more complex. In this work, only new patients in 2016 are recruited to the temporal test set. Excluded admissions are depicted in [Figure 1B, C](#) and detailed in [Supplementary Figure S1](#). Backwards-from-outcome selection removes all but one admission (the nearest to 12 months prior to death or last censor). In contrast, selection forwards-from-admission removes a smaller set of readmissions from the temporal set to mitigate data leakage.

Experimental factors consistent across designs

Feature construction and modeling

We fix the feature construction and modeling for each experimental design, by adapting that of Avati et al⁴ to include lab results and employing a random forest classifier. Further details can be found in the [Supplementary Material](#).

Thresholding criteria and simulating deployment

Typically, only metrics of overall model performance are reported—most commonly area under the receiver operating characteristic (AUROC) and, increasingly, area under the precision recall curve (AUPRC). However, in many applications, an operating threshold is selected within a test set by imposing a criterion on 1 particular measure. Since our outcome of interest, mortality, is rare but very important, we are most concerned with systems operating at a clinically acceptable, typically high, precision (positive predictive value)⁴ while maximizing recall (sensitivity). Together, these metrics can assist thresholding and inform how a model could be incorporated into an existing clinical workflow to recommend an intervention.

Deployment typically requires 1 operating threshold to separate patients at high risk of dying from those at low risk. We simulate this process by selecting a variety of potential operating thresholds by requiring precision to exceed some criterion (namely, the median selected threshold under bootstrapped subsampling conditions). We select 8 precision values spanning the range of realistic choices: 20–90%. Each threshold is "deployed" by predicting in the real-world cohort and identifying patients who exceed the threshold. Notably, the real-world set is identical across experiments as it is not constrained to the same criteria as the development sets and is intended to represent the real-world where:

1. Patients are readmitted,
2. Patients have varying degrees of clinical history available,
3. Patients die within days of admission with no recent hospitalizations, and
4. Some patients are lost to follow-up and thus do not have a reliable outcome.

The first 3 factors are incorporated as sources of realistic noise but the last issue is difficult to overcome without introducing bias. For this cohort, patients lost to follow-up within 1 year are omitted.

As each threshold is applied, the performance criterion selected for within the test set produces a corresponding real-world performance. Direct comparison of model performance across test sets is unfair, as the groups are dissimilar, but the degree of "migration" from test to real-world performance is reported and compared. Moreover, since the real-world cohort is consistent across experiments, the performance can be compared across thresholds, cohort selection method, and employed test sets.

Table 1. Model training and testing cohorts, stratified by outcome, for backwards-from-outcome and forwards-from-admission designs

Design	Outcome	Train	Random test	Temporal test	Subtotal	Real world
Backwards-from-outcome	Survival	11 540	2906	7861	22 307	15 428
	Death	732 (5.96%)	192 (6.20%)	348 (4.24%)	1272 (5.39%)	2440 (13.7%)
Forwards-from-admission	Survival	52 834	13 389	18 474	84 697	15 428
	Death	4913 (8.51%)	1157 (7.95%)	1381 (6.96%)	7451 (8.09%)	2440 (13.7%)

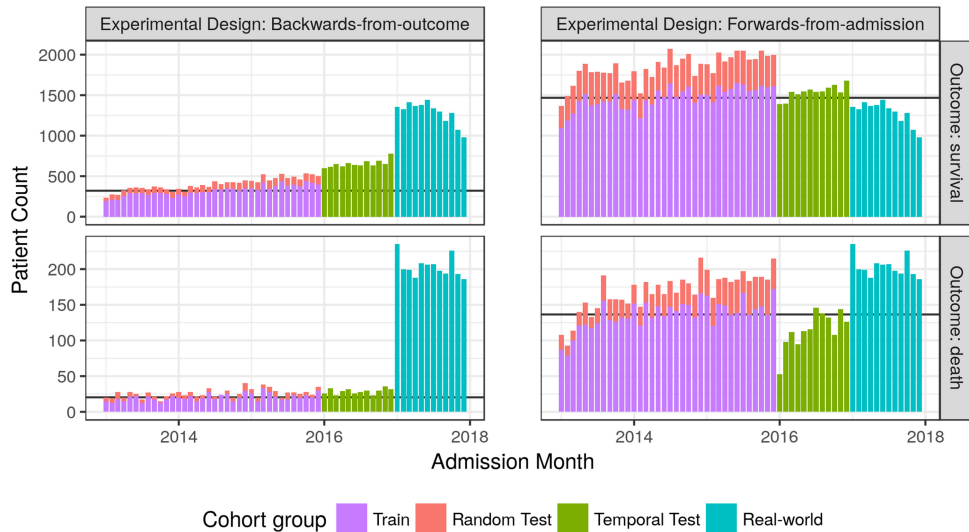


Figure 2. Cohort distributions of groups by calendar month. For reference, the monthly mean of the training set is represented with a horizontal black line.

RESULTS

Patient cohort

The study period spanned 5 calendar years, 2013–2017, and included all adult inpatient admissions to 1 hospital. In this period, 128 328 admissions were recorded covering 87 293 unique patients. In this cohort, the median [IQR] age at admission was 56.0 [35.4, 72.4] years, where 60.6% are female, 9.8% identify as Hispanic, 67.5% as white, 10.1% as African-American, 8.1% as Asian, and 14.2% as other or unknown race. Please refer to [Supplementary Table S3](#) for a more detailed breakdown of patient demographics, comorbidities, and model features. Of these admissions, 16 004 have a known death outcome (6501 unique patients), with a median [IQR] time from admission to death of 138 [22, 493] days, where 10 932 admissions lead to mortality within 365 days, a prevalence of 8.5% (comparable with other large prognostic cohorts¹⁷).

The real-world cohort is separated early from this larger cohort as described in [Supplementary Figure S1](#). Defined as any admission in 2017 ($n=29\ 382$) but further restricted to patients known to have died or censored beyond 365 days results in a real-world cohort of 17 868 admissions. The model development period of 2013–2016 consists of 98 946 candidate admissions for cohort selection.

Cohort selection

Backwards-from-outcome selection

The backwards-from-outcome experiment is very restrictive and discards the large majority of observed admissions yielding a model development cohort of $n=23\ 579$ that is separated into training, random and temporal test sets as described in [Table 1](#) and [Supple-](#)

[mentary Figure S1](#). The exclusions introduced by this design are apparent when plotted alongside the real-world set in [Figure 2](#) (left). This approximate 75% reduction in cohort size is comparable with similar works, namely Avati et al⁴ that reported an approximate 90% reduction (221 284 patients from approximately 2 million).

One primary limitation of backwards-from-outcome selection is the introduction of temporal bias. Deaths occur reasonably uniformly across time, and thus selection should select a similar number of deaths each month. However, when applied to the last encounter of the survival group, recent times are more likely to be selected as patients tend to continue to receive care within the same system until death, recovery or relocation. In the 4 years of model development data described in [Figure 2](#), a noticeable increase in admissions per month is observed in the survival group (top-left) but not the death group (bottom-left). This introduced discrepancy between outcome groups will only grow as the time period extends, challenging the use of such a design in cohorts that span longer-time periods.

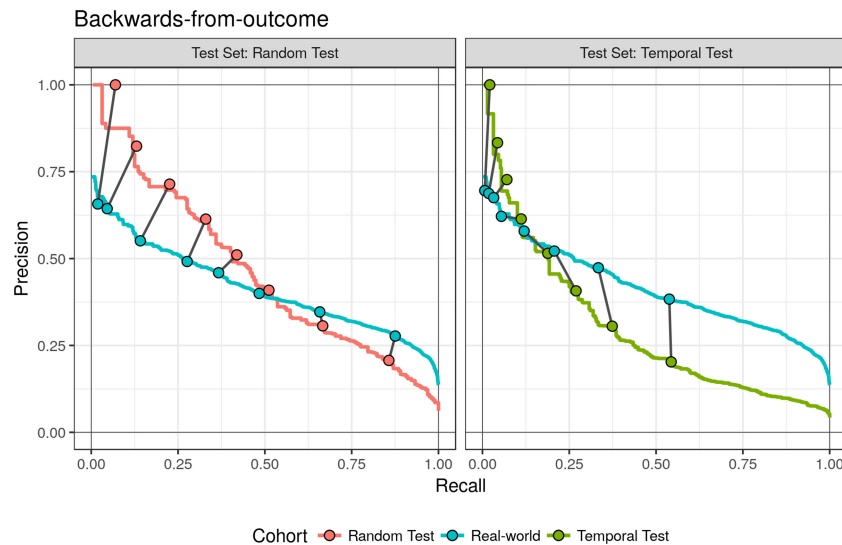
Forwards-from-admission selection

The less strict selection criteria of the forwards-from-admission design results in the inclusion of many thousands more admissions, evident in both [Table 1](#) and [Figure 2](#). The only admissions removed are those excluded upon transition to temporal test, where a noticeable drop is evident in [Figure 2](#) starting January 2016, particularly in the death group (bottom-right). The resulting cohort of 92 148 admissions is almost 4 times larger than the backwards-from-outcome cohort with a larger mortality prevalence, closer to that of the population.

Table 2. Model development and testing cohort performances in terms of AUROC and AUPRC for backwards-from-outcome and forwards-from-admission designs.

Design	Evaluation measure	Train (mean [min, max] cross-validation)	Random test [95% CI]	Temporal test [95% CI]	Real world [95% CI]
Backwards-from-outcome	AUROC (%)	89.9 [89.3, 91.2]	89.5 [87.0, 92.1]	85.4 [83.4, 87.3]	83.2 [82.3, 84.1]
	AUPRC (%)	40.7 [39.1, 43.1]	45.9 [37.9, 55.0]	29.3 [24.7, 34.8]	41.6 [39.6, 44.0]
Forwards-from-admission	AUROC (%)	90.4 [90.0, 90.7]	90.5 [89.7, 91.4]	90.3 [89.5, 91.2]	88.3 [87.6, 89.1]
	AUPRC (%)	48.7 [47.2, 50.5]	45.0 [42.1, 48.7]	46.9 [44.0, 50.3]	56.5 [54.2, 59.1]

Bold values indicate highest real world performance.

**Figure 3.** Precision and recall curves with highlighted operating points describing estimated real-world performance by the backwards-from-outcome experimental design in comparison to the internal and temporal test sets.

Test set performance: random vs. temporal

Backwards-from-outcome selection

Parameter optimization in 5-fold cross-validation within the backwards-from-outcome training set selected the optimal random forest model with 500 trees, depth of 100, and negative sample weight of 0.2. The mean [minimum, maximum] AUROC and AUPRC within cross-validation is 89.9% [89.3, 91.2] and 40.7% [39.1, 43.1], respectively. When retrained and applied to test sets, model performance shifts unpredictably as reported in Table 2, highlighting the differences between the 2 test sets.

Precision recall curves are plotted in Figure 3 for random (left; red) and temporal (right; green) test sets alongside that of the real-world cohort (blue). Performance appears better when evaluating the model on the random test set as compared with the temporal test set. Each operating threshold migrates from the estimated test set performance into a real-world performance, depicted by connected colored dots. Importantly, these performance migrations are not small vertical or horizontal movements, instead, they describe drastic performance shifts in both precision and recall. Interestingly, both random and temporal test set precision recall curves intersect the real-world curve such that the direction of the performance shift depends on the desired precision criteria. These large shifts suggest

that *neither test set provides an accurate estimation of real-world performance when using a backwards-from-outcome design*, likely due to distributional differences introduced by right-censoring.

Forwards-from-admission selection

Parameter optimization in 5-fold cross-validation within the forwards-from-admission training set selected the optimal random forest model with 1000 trees, depth of 200, and negative sample weight of 0.2, similar to the backwards-from-outcome design. The mean [minimum, maximum] AUROC and AUPRC within cross-validation is 90.4% [90.0, 90.7] and 48.7% [47.2, 50.5], respectively, and only marginal reductions in performance are reported when applied to random and temporal test sets, described in Table 2, suggesting the cohorts are similar.

Both the random and temporal test set precision recall curves of Figure 4 describe a drastic drop in performance in the very low recall region—known as the early retrieval problem¹⁸ that is especially common in tasks with imperfect labels. Despite this drop in precision, thresholds selected in both test sets typically result in improved performances when applied to the real-world set. The performance difference is relatively consistent between each test set and the real-world set, but *employing a forwards-from-admission design with a*

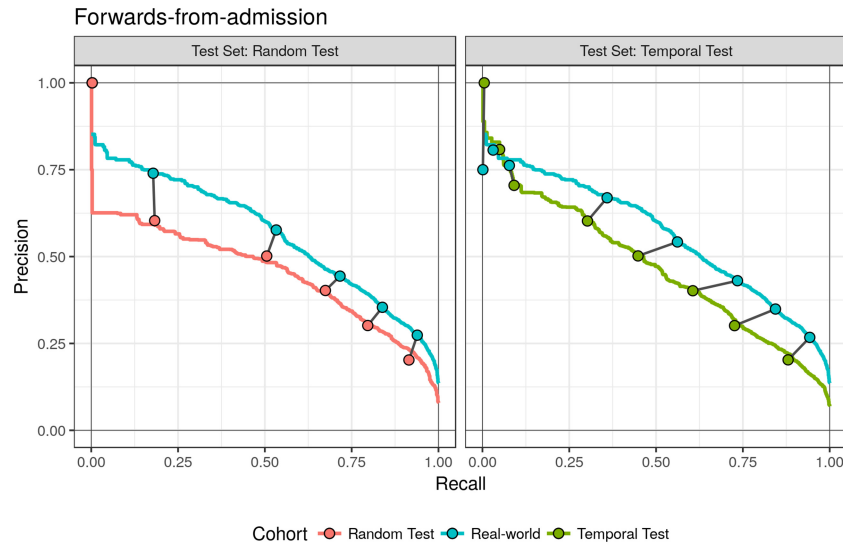


Figure 4. Precision and recall curves with highlighted operating points describing estimated real-world performance by the forwards-from-admission experimental design in comparison to the internal and temporal test sets.

temporal test set provides consistently conservative estimation of deployment—a safe scenario for deployment.

Real-world performance

The backwards-from-outcome design excludes many patients without certain labels which improves the model’s ability to distinguish very high-risk patients from at-risk patients (e.g., recall <0.20). But, the result is a drastically smaller training cohort than the forwards-from-admission alternative. The additional sample size of the forwards-from-admission experiment *improves real-world performance by 5.2% (95% CI 4.6, 5.7) AUROC and 14.8% (95% CI 13.5, 16.5) AUPRC*, described in [Table 2](#), that is *consistently and conservatively estimated with either test set*.

DISCUSSION

The machine learning community advocates for rigorous experimental design but, the medical domain presents challenges that cannot be ignored. Backwards-from-outcome selection and random testing are commonly employed, likely because they strictly address many of these challenges and simplify training and validation. However, doing so cleanses experimental data from the expected distribution upon deployment, as evident in the patient demographics and comorbidities of [Supplementary Table S3](#). In fact, many statisticians and epidemiologists would acknowledge that using forwards-from-admission selection (i.e. a cohort study) with a temporal test set is good practice despite the added complexity and potential for introduced bias—the lessor of the evils. Despite this, medical machine learning works are published with neither of these preferred design components nor any secondary prospective or external validation (as illustrated by a literature review in the [Supplementary Material](#)).

When applying machine learning to healthcare data, experimental choices are crucial, not only for their effect on validation performance but also on the real-world performance. Gold-standard results from prospective or external validation are infeasible in many applications without tremendous technical infrastructure and motivated by estimates of impactful performance. Without transparent reporting of experimental design choices, there is no way to dis-

criminate between potentially disruptive new applications and those reliant on biased, idealistic results. Since some experimental designs cleanse the underlying data beyond resemblance to real-world, prospective data, *these designs should be discouraged* by the community. Their continual use, without grounding in prospective or external results, *only perpetuates unrealistic expectations* of machine learning and artificial intelligence that may fuel disillusionment.

Experimental design considerations

Backwards-from-outcome selection limits one individual to one outcome group even if no appropriate prediction instant can be identified. In reality, patients that die are likely to be hospitalized more than once (55% of patients in our dataset), and in many cases will have long clinical histories with increasing risk. Since the forwards-from-admission design determines an outcome per admission, readmissions are accommodated and the model is penalized for premature or delayed risk estimation along each patient’s trajectory.

However, the forwards-from-admission design may introduce bias upon the temporal test set. To mitigate data leakage, all temporal test patients must have not been recently admitted—otherwise they would have been recruited for training—and therefore may be less acutely ill. A brief drop in admissions is observed in the death group of [Figure 2](#) (bottom-right). Separation of the training and temporal testing cohorts with a buffer time period, typical of prospective validation, may mitigate this introduced bias at the expense of omitted data.

Simulated deployment

The thresholding and simulated deployment presented here is overly simplified. In practice, thresholding would challenge the quality of some “gold-standard” outcome labels, especially those patients predicted at relatively high risk. Since death data are prone to missingness,¹⁹ a process including chart review or follow-up may reconcile some cases. Any improvements may be particularly helpful in the forwards-from-admission random test set as thresholding is currently hindered by poor performance at very low recall.

In this experiment, the real-world cohort begins immediately after the temporal test set. In reality, at least 1 year will pass between development data and implementation, to collect outcomes and develop the model. Implementation will almost certainly be more difficult than this simulated deployment, such that measuring generalization with a sampled test set becomes inadequate. Instead, a temporally separated test set will likely yield more realistic estimates of performance, as it assesses the model's adaptation to the future—historic transportability.²

Limitations

The results presented here consider one dataset. However, it is not unreasonable to expect this pattern of behavior in other datasets. Although we cannot predict future changes, we can be sure they will arrive. Expecting the challenges of deploying in a temporally evolving domain highlights the importance of pragmatic solutions for calibration drift²⁰ that may require retraining or planned recalibration,²¹ but future research is required.

Comparison to previous work is unfortunately limited as only 5 years of data were available in a setting with, potentially, less continuity of care and follow-up than Avati et al⁴ (given the geographic differences between New York City and Santa Clara County). In addition, predictions were restricted to inpatient admissions for practicality, a random forest model was employed for portability and efficiency, and hospice discharges were included as a proxy for death upon observing that hospice patients were often lost to follow-up with no confirmed death date (unpublished).

Using a binary “gold-standard” outcome label is common,²² however, implementation evaluation is plagued with uncertainty. Cohort selection and strict inclusion/exclusion criteria inadvertently separate the 2 classes, but patients identified in deployment will be lost to follow-up. Testing on a dataset that does not represent the expected deployment distribution is likely optimistic.

The extreme reduction in cohort size when selecting backwards-from-outcome results in relatively small cohorts in the random and temporal test sets, especially the death groups of 192 and 348 admissions, respectively. These groups are likely inadequate for reliable thresholding within the precision recall space, especially under subsampling conditions. More testing data may improve the thresholds. However, the sanitization of data by the backwards-from-outcome design will limit the utility of any test set threshold upon deployment.

Future work

We consider a simplified scenario where we plan to evaluate the learned model for use within our institution. Before we comment on its utility elsewhere, spectrum bias²³ and other pillars of transportability² must be addressed. Our first step is generalizing to other hospitals within our system which contain different patient distributions with varying data collection processes that may impede generalization.

CONCLUSION

Evaluation of a predictive model in terms of its utility and feasibility for deployment presumes the reported performance is achievable in the real world. In this work, we evaluated 2 common retrospective experimental designs employed for model development, and selected a range of thresholds across 2 test sets in order to compare their differences in terms of estimated model performance. Cohort selection

in a forwards-from-admission design results in higher real-world performance. Moreover, temporal validation within a test set comprised of more recent data yields more consistent estimates of real-world performance especially at higher precision than a randomly sampled internal validation set. In all cases, migration from test set to real-world performance was observed and should be expected when deploying predictive models into clinical practice.

STUDY APPROVAL

The NYU Grossman School of Medicine IRB has determined that this work is not human subjects research. Accordingly, IRB review was not required and not obtained.

AUTHOR CONTRIBUTIONS

V.J.M. retrieved data, wrote code for data preprocessing and model development, conducted experiments, and assembled results. N.J. also wrote code and conducted experiments. Y.A. advised model development, experimentation, and interpretation of results. All authors wrote, revised, and approved the final version of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGEMENTS

We thank Simon Jones for early discussion and feedback.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

- Goldstein BA, Navar AM, Pencina MJ, et al. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017; 24 (1): 198–208.
- Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999; 130 (6): 515–24.
- Hastie T, Friedman J, Tibshirani R. *The Elements of Statistical Learning*. 2001. doi:10.1007/978-0-387-21606-5
- Avati A, Jung K, Harman S, et al. Improving palliative care with deep learning. *BMC Med Inform Decis Mak* 2018; 18 (S4): 122.
- Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J* 2014; 35 (29): 1925–31.
- Qaseem A, Snow V, Shekelle P., et al.; for the Clinical Efficacy Assessment Subcommittee of the American College of Physicians. Evidence-based interventions to improve the palliative care of pain, dyspnea, and depression at the end of life: a clinical practice guideline from the American College of Physicians. *Ann Intern Med* 2008; 148 (2): 141–6.
- Hui D, Kim SH, Roquemore J, et al. Impact of timing and setting of palliative care referral on quality of end-of-life care in cancer patients. *Cancer* 2014; 120 (11): 1743–9.
- Christakis NA, Lamont EB. Extent and determinants of error in doctors' prognoses in terminally ill patients: prospective cohort study. *BMJ* 2000; 320 (7233): 469–73.

9. Glare P, Virik K, Jones M, *et al.* A systematic review of physicians' survival predictions in terminally ill cancer patients. *BMJ* 2003; 327 (7408): 195–8.
10. Amano K, Maeda I, Shimoyama S, *et al.* The accuracy of physicians' clinical predictions of survival in patients with advanced cancer. *J Pain Symptom Manage* 2015; 50 (2): 139–46.e1.
11. Elfiky A, Pany M, Parikh R, *et al.* A machine learning approach to predicting short-term mortality risk in patients starting chemotherapy. *bioRxiv*. 2017;204081. doi: 10.1101/204081
12. Reinke LF, Vig EK, Tartaglione EV, *et al.* Symptom burden and palliative care needs among high risk veterans with multi-morbidity. *J Pain Symptom Manage* 2019; 57 (5): 880–9.
13. Rajaram A, Morey T, Dosani N, *et al.* Palliative care in the twenty-first century: using advanced analytics to uncloak insights from big data. *J Palliat Med* 2019; 22 (2): 124–5.
14. Wegier P, Koo E, Ansari S, *et al.* mHOMR: a feasibility study of an automated system for identifying inpatients having an elevated risk of 1-year mortality. *BMJ Qual Saf* 2019;28:971–9.
15. Wang L, Sha L, Lakin JR, *et al.* Development and validation of a deep learning algorithm for mortality prediction in selecting patients with dementia for earlier palliative care interventions. *JAMA Netw Open* 2019; 2 (7): e196972.
16. He J, Hu Y, Zhang X, *et al.* Multi-perspective predictive modeling for acute kidney injury in general hospital populations using electronic medical records. *JAMIA Open* 2019; 2 (1): 115–22.
17. van Walraven C, McAlister FA, Bakal JA, *et al.* External validation of the Hospital-patient One-year Mortality Risk (HOMR) model for predicting death within 1 year after hospital admission. *CMAJ* 2015; 187 (10): 725–33.
18. Swamidass SJ, Azencott C-A, Daily K, *et al.* A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics* 2010; 26 (10): 1348–56.
19. Curtis MD, Griffith SD, Tucker M, *et al.* Development and validation of a high-quality composite real-world mortality endpoint. *Health Serv Res* 2018; 53 (6): 4460–76.
20. Davis SE, Lasko TA, Chen G, *et al.* Calibration drift among regression and machine learning models for hospital mortality. In: proceedings of *AMIA Annual Symposium* 2017; 2017: 625–34.
21. Davis SE, Greevy R, Matheny ME. Developing a Testing Procedure to Select Model Updating Methods. In: proceedings of *AMIA Annual Symposium* 2018; 2018.
22. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc* 2018; 25 (10): 1419–28.
23. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *N Engl J Med* 1978; 299 (17): 926–30.