# Evaluating Model Performance in Medical Datasets Over Time

**Helen Zhou*** HLZHOU@ANDREW.CMU.EDU
*Carnegie Mellon University, United States of America*

**Yuwen Chen*** YUWENC2@ANDREW.CMU.EDU
*Carnegie Mellon University, United States of America*

**Zachary Lipton** ZLIPTON@CMU.EDU
*Carnegie Mellon University, United States of America*

## Abstract

Machine learning (ML) models deployed in healthcare systems must face data drawn from continually evolving environments. However, researchers proposing such models typically evaluate them in a time-agnostic manner, splitting datasets according to patients sampled randomly throughout the entire study time period. This work proposes the Evaluation on Medical Datasets Over Time (EMDOT) framework, which evaluates the performance of a model class across time. Inspired by the concept of backtesting, EMDOT simulates possible training procedures that practitioners might have been able to execute at each point in time and evaluates the resulting models on all future time points. Evaluating both linear and more complex models on six distinct medical data sources (tabular and imaging), we show how depending on the dataset, using all historical data may be ideal in many cases, whereas using a window of the most recent data could be advantageous in others. In datasets where models suffer from sudden degradations in performance, we investigate plausible explanations for these shocks. We release the EMDOT package to help facilitate further works in deployment-oriented evaluation over time.

**Data and Code Availability** We use the following data: (1) the Surveillance, Epidemiology, and End Results (SEER) cancer dataset (**?**), (2) the COVID-19 Case Surveillance Detailed Data provided by the CDC (**?**), (3) the Southwestern Pennsylvania (SWPA) COVID-19 dataset, (4) the MIMIC-IV intensive care database (**?**), (5) the Organ Procurement and Transplantation Network (OPTN) database for liver transplant candidates (**?**), and (6) the MIMIC-CXR-JPG database of chest radiographs (**??**). MIMIC-IV and MIMIC-CXR-JPG (referred to as MIMIC-CXR in this paper) are available on the PhysioNet repository (**?**). Except for the SWPA dataset, all are publicly accessible (after accepting a data usage agreement). Details for accessing each dataset are in Appendices **??**–**??**. The code is publicly available on GitHub.

**Institutional Review Board (IRB)** This research does not require IRB approval.

## 1. Introduction

As medical practices, healthcare systems, and community environments evolve over time, so does the distribution of collected data. Features are deprecated as new ones are introduced, data collection may fluctuate along with hospital policies, and the underlying patient and disease populations may shift.

Amidst this ever-changing environment, models that perform well on one time period cannot be assumed to perform well in perpetuity. In the MIMIC-III critical care dataset, **?** found that a change to the electronic health record (EHR) system in 2008 coincided with sudden degradations in AUROC for mortality prediction. In COVID-19 data from the Centers for Disease Control and Prevention (CDC), **?** noted that the age distribution among cases shifted continually throughout the pandemic, and that these continual shifts confounded estimates of improvements in mortality rate.

We propose an evaluation framework to characterize model performance over time by simulating training procedures that practitioners could have ex-

* These authors contributed equally

ecuted up to each time point, and subsequently deployed in future time points. We argue that standard time-agnostic evaluation is insufficient for selecting deployment-ready models, showing across several datasets that it over-estimates deployment performance. Instead, we advocate for EMDOT as a worthwhile pre-deployment step to help practitioners gain confidence in the robustness of their models to shifts in the data distribution that have occurred in the past and may to some extent repeat in the future.

There is a large body of work that addresses adaptation under various structured forms of distribution shift, including covariate shift (**?????**), label shift (**?????**), missingness shift (**?**), and concept drift (**??**). However, in the real-world medical datasets we analyze, none of these structural assumptions can be guaranteed, and distributional changes in covariates, labels, missingness, etc. could even occur simultaneously. This motivates our empirical work, as it is unclear across a variety of model classes and medical datasets, how existing models might degrade due to naturally occurring changes over time, and whether different training practices might impact on robustness over time.

However intuitive it might seem, evaluation of models over time remains uncommon in standard machine learning for healthcare (ML4H) research. In the proceedings of the Conference on Health, Inference, and Learning (CHIL) 2022, for example, none of the 23 papers performed evaluations which took time into account (see Appendix **??** for similar statistics from CHIL 2021 and the Radiology medical journal). One possible reason for this is lack of access—as noted by **?**, it is common practice to remove timestamps when de-identifying medical datasets for public use. In this work, we identify six sources of medical data containing varying granularities of temporal information per-record, five of which are *publicly available*. We profile the performance of various training strategies and model classes across time, and identify possible sources of distribution shifts within each dataset. Finally, we release the Evaluation on Medical Datasets Over Time (EMDOT) Python package (details in Appendix **??**) to allow researchers to apply EMDOT to their own datasets and test techniques for handling shifts over time.

## 2. Related work

The promise of ML for improving healthcare has been explored in several domains, including cancer survival prediction (**?**), diabetic retinopathy detection (**?**), antimicrobial stewardship (**??**), recognizing diagnoses from electronic health record data (**?**), and mortality prediction in liver transplant candidates (**??**). Typically, these ML models are evaluated on randomly held out patients, and sometimes externally validated on other hospitals or newly collected data. Even with cross-site validations, we cannot be sure how models will perform in the future.

For decades, the medical community has had a history of utilizing (mostly) fixed, simple risk scores to inform patient care (**????**). Risk scores often prioritize ease-of-use, are computed from few variables, verified by domain experts for clear causal connections to outcomes of interest, and validated through use over time and across hospitals. Together, these factors give clinicians confidence that the model will perform reliably for years to come. With increasingly complex models, however, trust and adoption may be hindered by a lack of confidence in robustness to changing environments.

As noted by **?**, ML models often exhibit unexpectedly poor behavior when deployed in real-world domains. A key reason for these failures, they argue, is *under-specification*, where ML pipelines yield many predictors with equivalently strong held-out performance in the training domain, but such predictors can behave very differently in deployment. By testing performance across a variety of distribution shifts that have previously occurred over time, EMDOT could serve as a stress test to help combat underspecification.

Although evaluation over time is far from standard in ML4H literature, changes in performance over time have been noted in prior work. To predict wound-healing, **?** found that when data were split by cutoff time instead of patients, benefits of model averaging and stacking disappeared. **?** found degradation in performance of a model for wait times dependent on how much historical data was trained on. To predict severe COVID-19, **?** found that learned clinical concept features performed more robustly over time than raw features. Closest to our work is **?**, which evaluated AUROC in MIMIC-III critical care data from 2003–2012, comparing training on just 2001–2002; the prior year; and the full history. Using the full history and curated clinical concepts, they bridged a big drop in performance due to changing EHR systems. Whereas **?** considers three models per test year, EMDOT simulates model deployment every year and evaluates across *all future years*.

While we do not consider time series models in this work (instead considering those which treat data as i.i.d.), there are similarities between how training sets are defined in EMDOT and in techniques for evaluating time-series forecasts (**??**). These techniques often roll forward in time, taking either a window of recent data or all historical data as training sets, and evaluate test performance on the next time point. Performance from each time point is then averaged to summarize performance. This type of back-testing technique is common in rapidly evolving, non-stationary applications like finance (**??**), where time series models are constantly updated. In the healthcare domain, however, models may not be so easily updated, with risk scores developed several years ago still being used to this day (**????**). Thus, we track performance not only the immediate year after the training set, but all subsequent years in the dataset. Additionally, instead of collapsing performance from models trained at different time points into summary statistics, which could conceal distribution shifts over time, our framework tracks these granular fluctuations over time, and creates tools to help provide insight into the nature and potential causes of such changes.

## 3. Data

We sought medical datasets that had: (1) a timestamp for each record, (2) interesting prediction task(s), and (3) enough distinct time points to evaluate over. Six data sources satisfied these criteria: SEER cancer data, national CDC COVID-19 data, COVID-19 data from a healthcare provider in Southwestern Pennsylvania (SWPA), MIMIC-IV critical care data, OPTN data from liver transplant candidates, and MIMIC-CXR chest radiographs. All datasets are tabular except for MIMIC-CXR (medical imaging data). All but SWPA are publicly accessible.

Table 1 summarizes the dataset outcomes, time ranges, and number of samples. Figure 1 visualizes data quantity over time. Appendices **??**–**??** include cohort selection diagrams, cohort characteristics, features, heat maps of missingness, preprocessing steps, and additional details. Categorical variables are converged to dummies, and numerical variables are normalized and centered at 0. Missing values in categorical variables are treated as another category, and in numerical variables they are imputed with the mean. In all datasets except MIMIC-CXR (where each sample is a distinct radiograph), each sample corresponds to a distinct patient.

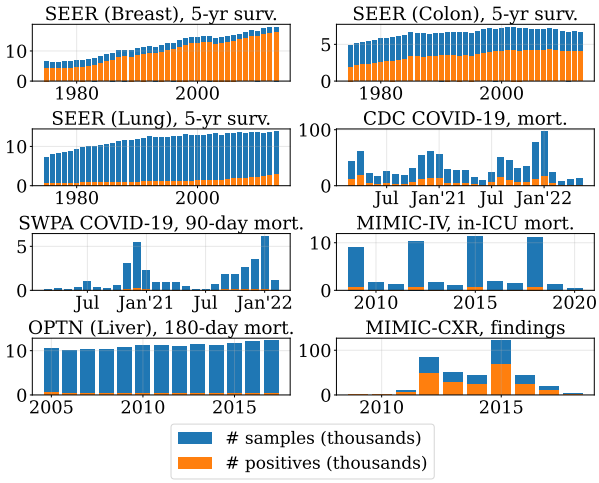


Figure 1: Number of samples and positive[2] outcomes per time point.

### 3.1. SEER Cancer Data

The Surveillance, Epidemiology, and End Results (SEER) Program collects cancer incidence data from registries throughout the U.S. Each case includes demographics, primary tumor site, tumor morphology, stage, diagnosis, first course of treatment, and survival outcomes (collected with follow-up) (**?**). We use the SEER*Stat software (**?**) to define three cohorts of interest: (1) breast cancer, (2) colon cancer, and (3) lung cancer. The outcome is 5-year survival, i.e. whether the patient was confirmed alive five years after the year of diagnosis. The amount of data has mostly increased each year (Figure 1). Performance over time is evaluated *yearly*. See Appendix **??** for more details.

### 3.2. National CDC COVID-19 Data

The COVID-19 Case Surveillance Detailed Data (**?**) is a national dataset provided by the CDC. It has the largest number of samples among the datasets considered, and contains 33 elements, with patient-level data including symptoms, demographics, and state of residence. The cohort consists of all lab-confirmed positive COVID-19 cases that were hospitalized, so the quantity of samples over time has a seasonality reflecting surges in COVID-19 (Figure 1). The outcome of interest is mortality, defined by `death_yn = Yes`

2. In MIMIC-CXR, all labels except "No Finding" are considered positive for the purposes of Figure 1 and Table 1.

Table 1: Summary of datasets used for analysis. For more details, see Appendices **??**–**??**.

| Dataset name | Outcome | Time Range (time point unit) | # samples | # positives |
|---|---|---|---|---|
| SEER (Breast) | 5-year Survival | 1975–2013 (year) | 462,023 | 378,758 |
| SEER (Colon) | 5-year Survival | 1975–2013 (year) | 254,112 | 135,065 |
| SEER (Lung) | 5-year Survival | 1975–2013 (year) | 457,695 | 49,997 |
| CDC COVID-19 | Mortality | Mar 2020–May 2022 (month) | 941,140 | 190,786 |
| SWPA COVID-19 | 90-day Mortality | Mar 2020–Feb 2022 (month) | 35,293 | 1,516 |
| MIMIC-IV | In-ICU Mortality | 2009–2020 (year) | 53,050 | 3,334 |
| OPTN (Liver) | 180-day Mortality | 2005–2017 (year) | 143,709 | 4,635 |
| MIMIC-CXR | 14 diagnostic labels | 2010–2018 (year) | 376,204 | 209,088 |

in the dataset. Performance over time is evaluated on a *monthly* basis. See Appendix **??** for more details.

### 3.3. SWPA COVID-19 Data

The Southwestern Pennsylvania (SWPA) COVID-19 dataset consists of EHR data from patients tested for COVID-19. It is the smallest dataset considered in this paper, and was collected by a major healthcare provider in SWPA. Features include patient demographics, labs, problem histories, medications, inpatient vs. outpatient status, and other information collected in the patient encounter. The cohort consists of COVID-19 patients testing positive for the first time, and not already in the ICU or mechanically ventilated. Similar to the CDC COVID-19 dataset, there is a seasonality to the monthly number of samples that reflects surges in COVID-19 (Figure 1). The outcome of interest is 90-day mortality, derived by comparing the death date and test date. The performance over time is evaluated on a *monthly* basis. See Appendix **??** for more details.

### 3.4. MIMIC-IV Critical Care Data

The Medical Information Mart for Intensive Care (MIMIC)-IV (**?**) database contains EHR data from patients admitted to critical care units from 2008–2019. MIMIC-IV is an update to MIMIC-III, adding time annotations placing each sample into a three-year time range, and removing elements from the old CareVue EHR system (before 2008). We approximate the year of each sample by taking the midpoint of its time range, but note that this causes certain years (2009, 2012, 2015, 2018) to have substantially more samples than others (Figure 1). The cohort is selected by taking the first encounter of all patients in the `icustays` table, and the outcome of interest is in-ICU mortality. Performance over time is evaluated on a *yearly* basis. See Appendix **??** for more details.

### 3.5. OPTN Liver Transplant Data

The Organ Procurement and Transplantation Network (OPTN) database tracks organ donation and transplant events in the U.S. The selected cohort consists of liver transplant candidates on the waiting list. The same pipeline as **?** is used to extract the data, except that the first record is selected for each patient. The outcome of interest is 180-day mortality from when the patient was added to the list. The performance over time is evaluated on a *yearly* basis. More details are in Appendix **??**.

### 3.6. MIMIC-CXR

The MIMIC Chest X-ray (MIMIC-CXR) JPG dataset (**?**) contains chest radiographs in JPG format. Similar to MIMIC-IV, we approximate the year by taking the midpoint of its three-year time range. The selected cohort consists of all radiographs from 2010 to 2018. The outcomes of interest are 14 diagnostic labels: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumonia, Pneumothorax, Pleural Other, Support Devices, and No Finding. Performance over time is evaluated on a *yearly* basis. More details are in Appendix **??**.

## 4. Methods

We tackle the following guiding questions:

1. On each dataset, what would the reported performance of a model be if it were trained using standard time-agnostic splits (**all-period**)?

2. **Simulating** how a practitioner might have trained and deployed models in the past, how would performance have varied **over time**?
3. When might it be better to train on a **recent window** of data versus **all historical** data?
4. What is the comparative performance of different **classes of models** over time?
5. To what extent might we be able to diagnose possible **reasons** for changes in model performance?

### 4.1. All-period Training

We mimic common practice in evaluation by using time-agnostic data splits which randomly place patients from the entire study time range into train, validation, and test sets (details in Appendix **??**), and reporting the test set performance. We refer to training with this type of split as *all-period* training.

### 4.2. EMDOT Evaluation

For more realistic simulation of how practitioners train models and subsequently deploy them on future data, we define the *Evaluation on Medical Datasets Over Time* (EMDOT) framework. At each time point $t$ (termed *simulated deployment date*), an *in-period* subset of data from times $\leq t$ is available for model development. After training a model on this in-period data, one might be interested in both recent in-period performance (at time $t$) and future *out-of-period* performance (at times $> t$).

In-period data is split into train, validation, and test sets (split ratios in Appendix **??**). For MIMIC-CXR, where one patient could have multiple radiographs, the data is split such that there are no overlapping patients between splits. Recent in-period performance is evaluated on held-out test data from the most recent time point. Out-of-period performance is evaluated on all data from each future time point. For example, a model trained up to time 6 is tested on data from 6, 7, 8, etc. (Figure 2). At time 8, the model is considered two time points *stale*. Although this procedure can take $O(T)$ times more computation than all-period training for $T$ time points, we argue that this procedure yields a more realistic view of the type of performance that one might expect models to have over time.

Additionally, practitioners face a tradeoff between using recent data perhaps most reflective of the present and using all available historical data for a larger sample size. Intuitively, the former may be appealing in modern applications with massive datasets, whereas the latter may be necessary in data-scarce applications. We explore these two training regimes, with different definitions of in-period data (Figure 2):

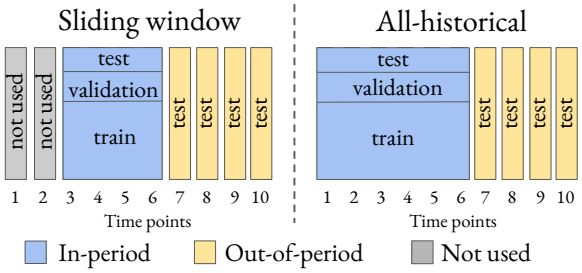


Figure 2: EMDOT training regimes, with a simulated deployment date of $t = 6$.

1. **Sliding window**: The last $W$ time points are considered in-period. In this paper, we use window size $W = 4$ for sufficient positive examples.
2. **All-historical**: Any data prior to the current time point is considered in-period.

To decouple the effect of sample size from that of shifts in the data distribution, comparisons are also performed with all-historical data that is **subsampled** to be the same size as the corresponding training set under the sliding window training regime.

To summarize more formally, let $D_t$ refer to the set of all data points occurring at time $t \in \{1, ..., T\}$, where $T$ is the number of time points that the dataset spans. Each $D_t$ can be partitioned by splitting patients at random into disjoint train, validation, and test sets: $D_t = D_t^{\text{train}} \cup D_t^{\text{val}} \cup D_t^{\text{test}}$. For simulated deployment dates $t^* \in \{W, W+1, ..., T\}$, training, validation, and test sets are defined for the *sliding window* training regime as follows:

- training: $\bigcup_{k=t^*-W+1}^{t^*} D_k^{\text{train}}$
- validation: $\bigcup_{k=t^*-W+1}^{t^*} D_k^{\text{val}}$
- in-period test: $D_{t^*}^{\text{test}}$
- out-of-period test: $D_k$ for $k = t^* + 1, ..., T$

Training, validation, and test sets are defined for the *all-historical* training regime as follows:

- training: $\bigcup_{k=1}^{t^*} D_k^{\text{train}}$
- validation: $\bigcup_{k=1}^{t^*} D_k^{\text{val}}$
- in-period test: $D_{t^*}^{\text{test}}$
- out-of-period test: $D_k$ for $k = t^* + 1, ..., T$

At each simulated deployment date $t^*$, models are trained using the training set, validated using the validation set, and tested on the in-period test set as well as all out-of-period test sets. If a model with simulated deployment date $t^*$ is being evaluated on an out of period test set $D_{t^*+j}$, then the model is $j$ time points *stale*.

### 4.3. Evaluation Metrics

All binary classification tasks are evaluated by AUROC. For multi-label prediction in MIMIC-CXR, each of the 14 diagnostic labels is treated as a separate binary classification task, and a weighted sum of AUROCs is computed, where the weight for a particular label is given by the proportional prevalence of that label among all positive labels. That is, for some class $a$, its weight is $p_a / \sum_x p_x$, where $p_x$ is the number of positives with label $x$. Samples are treated in an i.i.d. manner for training.

### 4.4. Models

Logistic regression (LR), gradient boosted decision trees (GBDT) and feedforward neural networks (MLP) are trained on the tabular datasets. DenseNet-121 is trained on the MIMIC-CXR imaging dataset. Hyperparameters are selected based on in-period validation performance, and the hyperparameter grids are in Appendix **??**.

### 4.5. Detecting Sources of Change

To better understand possible reasons for changing performance, we create *diagnostic plots* to track model performance alongside changes in the data distribution over time.

In tabular datasets, we plot feature importances and average values of the most important features over time. Generating these plots for logistic regression, we define feature importance by the magnitudes of the coefficients, but note that other feature importance techniques could be used for more complex model classes. To avoid overcrowding the plots, we take the union of the top $k$ most important features from each time point is taken, where $k$ is tuned depending on the dataset. We additionally highlight (using a thicker line) categorical features with consistently high prevalence or which experience a large change in prevalence across one time point, and numerical features with high average rank (see Appendix **??** for thresholds for each dataset).

For the imaging dataset, where feature importance is less straightforward, we plot the distribution of pixel intensities over time, along with proportions of each of the 14 diagnostic labels.

By highlighting sudden changes in model performance and the corresponding time periods in all other plots, diagnostic plots can help bring attention to shifts in the distribution of data that coincide with changing model performance.

### 4.6. EMDOT Python Package

We release the EMDOT python package[3] to help practitioners move from standard model evaluation to EMDOT evaluation. See Appendix **??** for a schematic of the EMDOT workflow, and see the GitHub repository for a step-by-step tutorial.

## 5. Results

### 5.1. All-period Training

In standard time-agnostic evaluation, GBDT and MLP achieve the highest average test AUROC on all tabular datasets except MIMIC-IV (Table 2). Note however that LR often has comparable or only slightly lower AUROC than the more complex models. The top 10 coefficients of each LR with all-period training are in Appendices **??**–**??**, and the per-label AUROC of MIMIC-CXR is in Appendix Table **??**. To form a baseline for comparison across time, we also evaluate the all-period models on subsets of the all-period test data that belong to each year (red dotted line in Figure 3), but note that this type of training (on future data) is not feasible in deployment.

### 5.2. EMDOT Evaluation

Figure 3 plots the AUROC of LR for all tabular datasets (and DenseNet-121 for MIMIC-CXR) over time when using the all-historical training regime. Plots for GBDT and MLP are in Appendix **??**, along

3. https://github.com/acmi-lab/EvaluationOverTime

Table 2: Test AUROC from all-period training and time-agnostic evaluation.

| Model | SEER (Breast) | SEER (Colon) | SEER (Lung) | CDC COVID-19 | SWPA COVID-19 | MIMIC-IV | OPTN (Liver) | MIMIC-CXR |
|---|---|---|---|---|---|---|---|---|
| LR | 0.888 | 0.863 | 0.894 | 0.837 | 0.928 | **0.935** | 0.846 | - |
| GBDT | **0.891** | 0.868 | 0.894 | 0.851 | **0.930** | 0.931 | **0.854** | - |
| MLP | **0.891** | **0.869** | **0.898** | **0.852** | 0.928 | 0.898 | 0.847 | - |
| DensetNet | - | - | - | - | - | - | - | **0.860** |

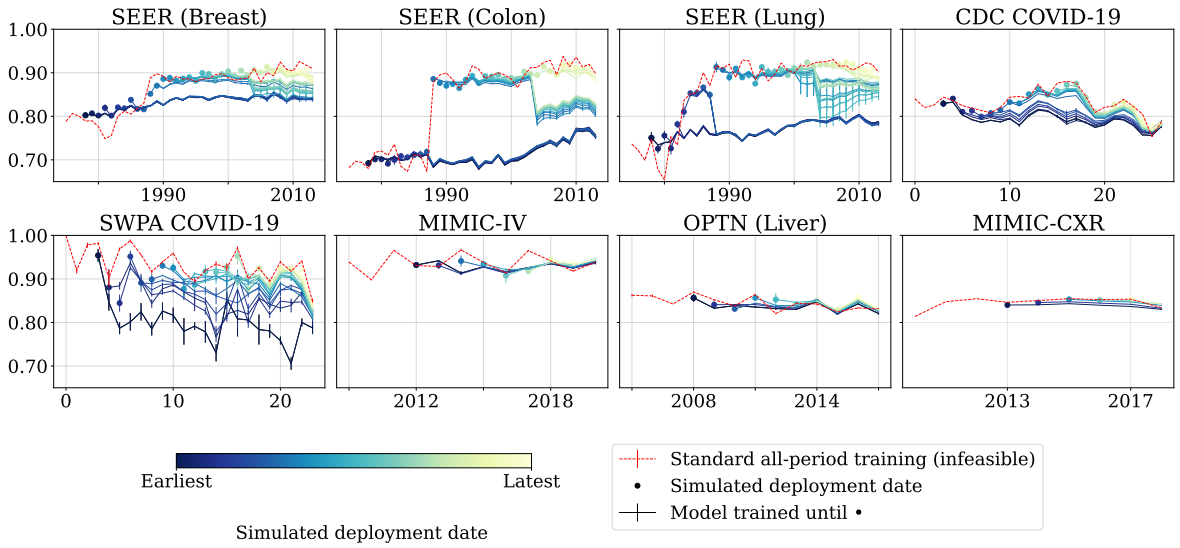


Figure 3: Average test AUROC of logistic regression vs. time. Each solid line gives the performance of a model trained up to a simulated deployment time (marked by a dot), evaluated across future time points. Error bars are ± standard deviation computed over 5 random splits. Red dotted line gives per-timepoint test performance of a model from all-period training (infeasible in reality, as it would involve training on data after the simulated deployment date).

with plots for AUPRC. We mainly discuss AUROC, but note that AUPRC observes similar trends as in AUROC. One difference however is that the baseline AUPRC performance is given by the label prevalence (rather than a constant 0.5, as in AUROC), and so observed trends in label prevalence over time appear to influence trends in AUPRC (Appendix Figure **??**).

For both AUROC and AUPRC, the reported test performance of a model from standard all-period training (red dotted line) mostly sits above the performance of any model that could have realistically been deployed by that date. Thus, all-period training tends to provide an over-optimistic estimate of performance upon deployment.

Across the datasets, a variety of trajectories of model performance are observed over time. In the SEER datasets, the AUROC of freshly trained models increases dramatically near 1988, but several of these models experience a large drop in AUROC around 2003 (Figure 3). Additionally, in-period test AUROCs tend to increase over time. By contrast, in CDC data, in-sample test AUROCs fluctuate up and down, and model performance over time varies more smoothly, appearing to loosely follow the in-sample performance. Models trained after December 2020 have a slight boost in AUROC, coinciding with a surge in cases (and hence sample size, Figure 1), however by January 2022 the in-sample AUROC decreases. In SWPA COVID-19, there is more variation and uncertainty in AUROC early in the pandemic, where sample sizes are small. In December 2020, sample sizes increase, and models seem to be-

come more robust to changes over time. Finally, in the MIMIC-IV, MIMIC-CXR, and OPTN datasets, AUROC appears relatively stable across time.

### 5.3. Training Regime Comparison

As the staleness of training data increases (i.e. as the test date gets further from the simulated deployment date), different training regimes can fare differently depending on the dataset (Figure 4, left).

In SEER (Breast) and SEER (Lung), sliding window is initially comparable to all-historical on fresh (low-staleness) data, but significantly underperforms both all-historical and all-historical (subsampled) when data are 8 to 22 years stale. At larger stalenesses, all training regimes start to become comparable. In CDC COVID-19, sliding window outperforms all-historical regardless of how stale the data is. By contrast, in SWPA COVID-19, which has the least amount of data (Table 1), both sliding window and all-historical (subsampled) underperform all-historical. In SEER (Colon), performance is relatively stable regardless of training regime. In MIMIC-IV, OPTN (Liver), and MIMIC-CXR, sliding window is on average comparable or slightly outperforms all-historical when staleness is 0, but at nonzero stalenesses all-historical outperforms both sliding window and all-historical subsampled.

### 5.4. Model Comparison

In SEER (Breast) and OPTN, GBDT outperforms both LR and MLP across the entire time range (Figure 4, right). In SEER (Colon), SEER (Lung), and CDC COVID-19, both GBDT and MLP initially outperform LR when staleness of the training data is less than 4 years, 4 years, and 7 months, respectively, however both eventually underperform LR as staleness increases further. While there is an uptick in GBDT performance on CDC COVID-19 towards 21-month staleness, we note this data point is derived from less data than other points on the line because the data time range is finite. In the SWPA COVID-19 dataset, LR, MLP, and GBDT appear to perform comparably over time. In the MIMIC-IV dataset, LR performed best to begin with and remained the best.

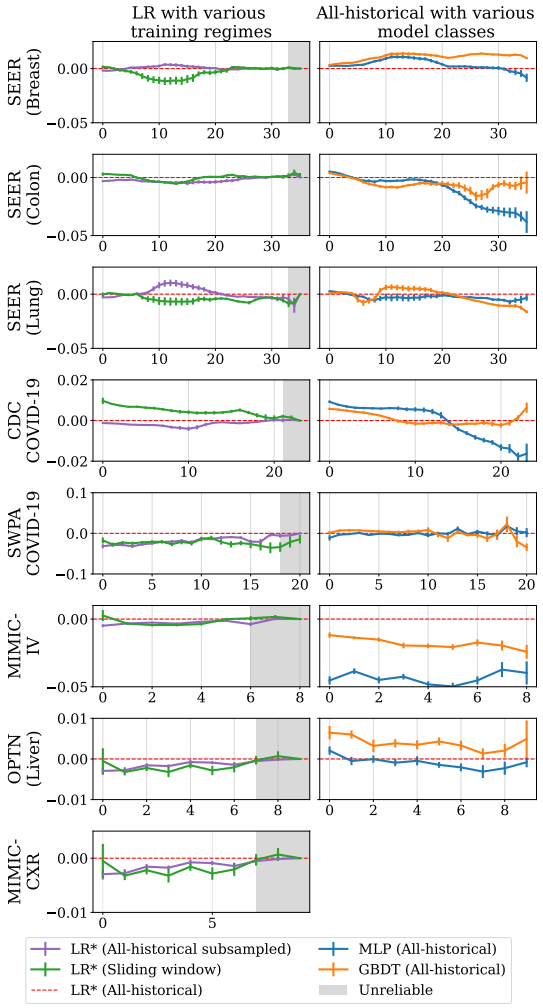


Figure 4: AUROC − $\text{AUROC}_{\text{LR* all-historical}}$ vs. staleness. i.e., AUROC difference relative to a LR* all-historical baseline across varying stalenesses of data,[5] for different training regimes (left) and model classes (right). Error bars are ± std. dev. (*in MIMIC-CXR, DenseNet-121 is used instead of LR)

### 5.5. Detecting Possible Sources of Change

Diagnostic plots for all datasets are in Appendix **??**. Here, we discuss SEER (Lung) (Figure 5) in detail as it has several interesting changes in model performance over time. In 1983, as `EOD 4` features from the

5. Note: at the largest stalenesses, there are fewer simulated deployment dates being averaged over, and they must be early in the dataset. Here, the sliding window and all-historical can be expected to perform similarly (especially when the sliding window is not much larger than or even matches the history). Since this is an artifact of finite time ranges, we gray out stalenesses where at least half of the all-historical data is the first sliding window of data.

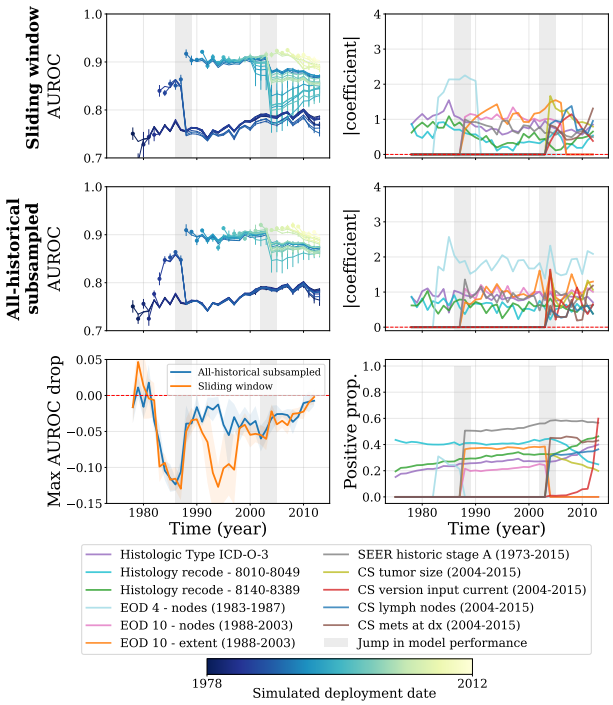


Figure 5: SEER (Lung) diagnostic plots. AUROC vs. time for sliding window (top-left) and all-historical subsampled (mid-left), max. drop in AUROC for each simulated deployment time (low-left), absolute feature coefficients for LR models from sliding window (top-right) and all-historical subsampled (mid-right) and prevalences of important features over time (low-right).

extent of disease coding schema are introduced (Figure 5, bottom right), a sudden jump in AUROC occurs (Figure 5, top and middle left). However, models trained at this time later experience a large AUROC drop (Figure 5, bottom left). By 1988, `EOD 4` is phased out, and `EOD 10` features are introduced. This coincides with another jump in AUROC, sustained until 2003 when the `EOD 10` features are removed. In this dataset, the all-historical training regime seems more robust to changes over time, as all-historical models trained after 1988 avoid the drop that sliding window models undergo once their window excludes pre-1988 data (Figure 5, bottom left).

## 6. Discussion

Reported model performance from standard all-period training tends to be over-optimistic (Figure 3) as models are evaluated on time points already seen in their training set (unrealistic in deployment settings). Thus, AUROCs reported from all-period training do not capture degradation that would have occurred in deployment.

Comparing model classes, in all datasets except MIMIC-IV, GBDT and MLP slightly outperform LR under standard time-agnostic evaluation (Appendix Table 2). However, evaluated across time, LR is often comparable and even outperforms more complex models once enough time passes after the simulated deployment date. For example, MLP achieves the best AUROCs in SEER Breast, Colon, and Lung in standard time-agnostic evaluation (Table 2). However, in evaluation over time, LR had superior performance once some amount of time (30, 5, 4 years respectively) had passed (Figure 4, right). In most datasets GBDT appears more robust over time than MLP, however as the training data becomes more stale it tends to become comparable to LR (in all datasets except OPTN Liver and SEER Breast, GBDT dipped below the performance of LR for several stalenesses). Thus, although complex model classes may appear to outperform simpler linear model classes in standard time-agnostic evaluation, one should consider performance over time when selecting a model class for deployment. As demonstrated by the different relative performances of model classes when evaluated over time versus in a time-agnostic manner, EMDOT can serve as a helpful stress-test to combat under-specification.

Regarding training regimes, we find that with increasing stalenesses, all-historical appears more reliable than sliding window across all datasets except for CDC COVID-19 (Figure 4, left). In SWPA COVID-19, MIMIC-IV, OPTN (Liver), and MIMIC-CXR, the benefit of all-historical data likely comes from the increased sample size, as subsampling all-historical data to be the same size as the corresponding sliding window resulted in comparable performance to sliding window. In the SEER datasets, the effect of sample size is less pronounced, as sliding window and subsampled all-historical are frequently comparable to all-historical. There are certain stalenesses for which sliding window underperforms all-historical, which may be due to the addition and removal of features. If the sliding window model learns

to rely on recently added features which are later removed, this could result in drops in performance whereas an all-historical model which had learned to predict without the presence of such features would be more robust to such changes. On the other hand, in CDC COVID-19 (the setting with the most data and fewest features), subsampled all-historical performs comparably to all-historical, and sliding window outperforms both across all stalenesses (Figure 4, left). This suggests that the performance of LR may have been saturated even when a sub-sample of all-historical data was used, and the benefit of using more recent data outweighs the larger sample size afforded by all-historical. More broadly, in rapidly evolving environments with simple models, few features, and large quantities of data, the sliding window training regime could be advantageous.

The SEER datasets had dramatic changes in data distribution in both 1988 and 2003, when important features were added and/or removed (Figure 5). One possible reason for the robustness of all-historical models in this dataset is that after 2003, when features like EOD 10 were removed, the model could still rely on features that were introduced prior to the use of EOD 10 in 1988. More broadly, we hypothesize that if a model was trained on a mixture of distributions that occurred throughout the past, it may be better equipped to handle shifts to settings similar to those distributions in the future.

While the SEER datasets and COVID-19 datasets displayed several changes in model performance over time, the OPTN and MIMIC datasets had relatively stable behavior. One possible reason for this is that the outcomes or diseases of interest were relatively stable in nature, we did not observe any substantial changes in the distribution of data. Another is that in the MIMIC datasets, a three-year range was given for each sample rather than a specific date. This uncertainty around the date, along with the limited number of date ranges, could result in a smoothing effect on the resulting estimates of performance.

In conclusion, EMDOT not only yields insights into the suitability of different model classes or training regimes for deployment, but also helps one detect distribution shifts that occurred in the past. Understanding such shifts may help practitioners be prepared for shifts of a similar nature in the future. Although the EMDOT framework does require additional computational time than the standard time-agnostic evaluation setup, we argue that the insights that could be gained from this procedure are worthwhile, especially before deployment in high-stakes settings.

**Limitations and Future Work** One possible reservation that users might have about using EMDOT is that it could involve training up to $T$ times as many models as would normally be required (where $T$ is number of timepoints). To help alleviate this concern, in future work we plan to implement parallelization in EMDOT. For noisier estimates of model performance in less time, one could also subsample the dataset. Another interesting extension is exploring performance over time in other data modalities (e.g. time series, natural language, etc.). Depending on the complexity of models used in these modalities, this may require additional computational resources. More broadly, we hope that others may also build upon EMDOT to shine new light on how models and methodologies fare when evaluated with an eye towards deployment.