

A Machine Learning Approach to Real-World Time to Treatment Discontinuation Prediction

Weilin Meng, Xinyuan Zhang, Boshu Ru, and Yuanfang Guan*

Real-world time to treatment discontinuation (rwTTD) is an important endpoint measurement of drug efficacy evaluated using real-world observational data. rwTTD, represented as a set of metrics calculated from a population-wise curve, cannot be predicted by existing machine learning approaches. Herein, a methodology that enables predicting rwTTD is developed. First, the robust performance of the model in predicting rwTTD across populations of similar or distinct properties with simulated data using a variety of commonly used base learners in machine learning is demonstrated. Then, the robust performance of the approach both within-cohort and cross-disease using real-world observational data of pembrolizumab for advanced lung cancer and head neck cancer is demonstrated. This study establishes a generic pipeline for real-world time on treatment prediction, which can be extended to any base machine learners and drugs. Currently, there is no existing machine learning approach established for predicting population-wise rwTTD, despite that it is an essential metric to report real-world drug efficacy. Therefore, we believe our study opens a new investigation area of rwTTD prediction, and provides an innovative approach to probe this problem and other problems involving population-wise predictions. An interactive preprint version of the article can be found at: <https://doi.org/10.22541/au.166065465.59798123/v1>.

1. Introduction

Real-world time on treatment (rwToT), also known as real-world time to treatment discontinuation (rwTTD), is defined as the length of time observed in real-world data (as distinct from controlled clinical trials) from initiation of a medication to discontinuation of that medication.^[1,2] The ending of the treatment can be caused by adverse events, deaths, switches of treatment, and loss of follow-up. Because time to treatment discontinuation can be readily obtained from electronic medical records, this effectiveness endpoint is convenient to evaluate the efficacy of a drug that is already approved for public use.^[3] It is often used as a surrogate effectiveness endpoint, showing a high correlation to progression-free survival and a moderate-to-high correlation to overall survival.^[4,5] As rwTTD is an important metric for drug effectiveness, it is routinely reported during the post-clinical trial phase.^[2,4,6–8]


Calculation of rwTTD in patient population is often equivalent to constructing a Kaplan–Meier (KM) curve, with each point representing the proportion of patients that are still on treatment at a specific time point.^[1] Either the entire curve, or mean rwTTD, restricted mean,^[9] or the time point at which a specific portion of the patients (e.g., 50%) dropping treatment is of interest. Currently, there is no existing machine learning scheme established to predict such a curve, or the midpoint, as the vast majority of the machine learning models have been focused on predicting individuals' behavior rather than population-level behavior. Such a machine learning scheme, if established, has many meaningful clinical applications. For instance, given observed clinical parameters and outcomes in clinical trials, how do we derive expected time-to-treatment in the real world? Given the rwTTD for a drug on one patient population, how can we predict the rwTTD when applying this drug to another population (e.g., for a different disease)?

This study establishes a machine learning framework to infer population-wise rwTTD. We showed that population-wise curve prediction differs substantially from aggregating all individuals' results. Our framework models the population-wise curve and is generic to diverse base learners for predicting rwTTD. We demonstrated the effectiveness of this framework based on both simulated data and real-world electronic medical records (EMR) data for pembrolizumab-treated

W. Meng, B. Ru
Center for Observational and Real-World Evidence (CORE)
Merck & Co., Inc.
Rahway, NJ 07065, USA

X. Zhang^[†]
Ann Arbor Algorithms Inc.
Ann Arbor, MI 48104, USA

Y. Guan
Department of Computational Medicine and Bioinformatics
University of Michigan
Ann Arbor, MI 48109, USA
E-mail: gyuanfan@umich.edu

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/aisy.202200254>.

^[†]Present address: AnnHua Chinese School, Ann Arbor, MI 48104, USA

© 2023 The Authors. Advanced Intelligent Systems published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/aisy.202200254

cancer populations.^[6,10,11] The study opens a new direction of modeling population-level rwTTD, which has great value for directing post-clinical stage drug administrations. This machine learning scheme will also have meaningful implications for population-based predictions for other problems, as machine learning algorithms have so far been focused on predictions for individual samples.

2. Results

2.1. A Machine Learning Framework for Predicting Population-Wise rwTTD

Termination of a specific treatment can be considered as survival data, where an observed termination of treatment is an event point and otherwise the patient is censored (Figure 1a).^[1] However, existing survival models only predict individual patient's likelihood of survival. As shown in the following shortly, the aggregation of individuals does not represent the profile of a population. Therefore, we designed an approach that predicts the termination curve of a population.

We started with producing the gold standard (expected future time) for each individual in the training population. This expected future time is defined as the time expected until the treatment is terminated from the point at which we are going to make the predictions. Prior to this point, all observed clinical data are available for making predictions. Two cases can be considered here. In the first case, if we know the termination time of the treatment (an "event" data point), the patient's future time is defined as the time between the end of the observation window, from which we collect feature data used to make a prediction, and the drug termination time. In the second case, if the termination time of the treatment is unknown for a patient (a "censored" data point), we infer the expected future time from the survival curve derived from the training population. In this case, we use a popular method, KM curve, to represent the termination ratio of the training set.^[12] The expected future time is then composed of two parts. The first part is the existing time-lapse, i.e., from the end of the observation time window to the last contact time point, because we know without uncertainty that the patient continued drug treatment until the last contact time point. The second part is the expected time after the last contact time point, which is calculated as the integral of the curve beyond the last contact time point divided by the terminated ratio at the last contact time point (Figure 1a). Adding the first and second part together results in the expected future time for the censored individuals. This approach generates the gold standard for predicting the expected future time for each individual into which any kind of base learners can be built. Later, we will explain how a nested training scheme can extrapolate and aggregate the predictions from individuals to infer the terminated ratio curve for a population.

We simulated drug termination data of a population following a survival study^[13] (Figure 1b). We generated a population of total n individuals, where the termination rate for each individual is drawn from a population of $p = N(p_{\text{mean}}, \sigma)$, and we force the minimal termination rate to be zero. We hypothesize that the probability that a patient terminates the treatment (p) on a single day is driven by a series of (m in total) predictive features f .

These features, in reality, can be demographic information, clinical measurements, or any claim data, as will be shown with the real-world drug treatment experiment in the following. In this simulation experiment, we let individual feature values correlates to p by

$$v_{kj} = p_k \times f_j(1 + \theta \times e_j) \quad (1)$$

where v_{kj} is the value of feature j for patient k . p_k is the termination rate of Patient k . f_j represents the scaling factor of a particular feature, uniformly drawn between $[0, \alpha]$. Each feature j is parameterized by noise factor e_j , uniformly drawn from $[0, \beta]$. When goes up, a larger sampling range will result in less correlation between the feature and the expected future time. The value of the j th feature of the k th sample, v_{kj} , is further parameterized by θ , which is uniformly distributed sampled between $[-0.5, 0.5]$.

We set the maximal allowed observation date of all individuals to δ_{max} . Between $[0, \delta_{\text{max}}]$, we create a binomially distributed vector of length $\delta_k = B(\delta_{\text{max}}, p_k)$ for each individual k . Thus, the higher the p_k , the more likely the individual is to be terminated with the uncertainty defined by the binomial distribution. In this binomially sampled sequence, the first appearance of 1 decides the termination date t_{term} . Next, for each individual, we uniformly sampled between $[0, \delta_{\text{max}}]$ and define the censoring date t_{censor} . If $t_{\text{term}} > t_{\text{censor}}$, the last observation time $t_{\text{last}} = t_{\text{censor}}$, and the status is 0 (censored point and no termination date is observed); otherwise, the $t_{\text{last}} = t_{\text{term}}$ with a status = 1 (termination observed and the date is defined).

We developed three metrics to evaluate the model performance (Figure 1c–e). For the first metric, "absolute error," we calculated the accumulated values of the predicted curve and the gold standard curve from day 0 to a specific date (1000 days, if not otherwise specified in this article), and then divided the total difference by the total number of days. Thus, if the predicted curve is higher than the gold standard curve in the first half, but lower in the later half, the errors could be canceled out by using this metric. For the second metric, "cumulative error," we accumulated the absolute error on each day from day 0 to a specific date, and then divided the total error by the total number of days. Then, no matter positive error or negative error, the absolute errors will aggregate. For the third metric, "Absolute date error at 50% terminated," we calculated when 50% of the patients are terminated (reaching 0.5 on the y -axis on the termination curve), what is the absolute difference in days between the gold standard curve and the predicted curve. The three metrics capture the important aspects of drug administration.

Of note, models can only generate predictions for each individual's expected future time in the test set when trained with a machine learning classifier. When we aggregate the predictions, the resulting curve is closely centered at the average expected future time and substantially deviates from the true distribution (Figure 2a–c). This is due to the innate properties of most machine learning algorithms. When minimizing the squared errors or another similar loss function, the prediction values tend to center around the mean.

To combat such an effect, we further divided the training set into the train set, from which the model parameters are derived, and the validation set, from which the distribution of the prediction value is obtained. The prediction value from the validation

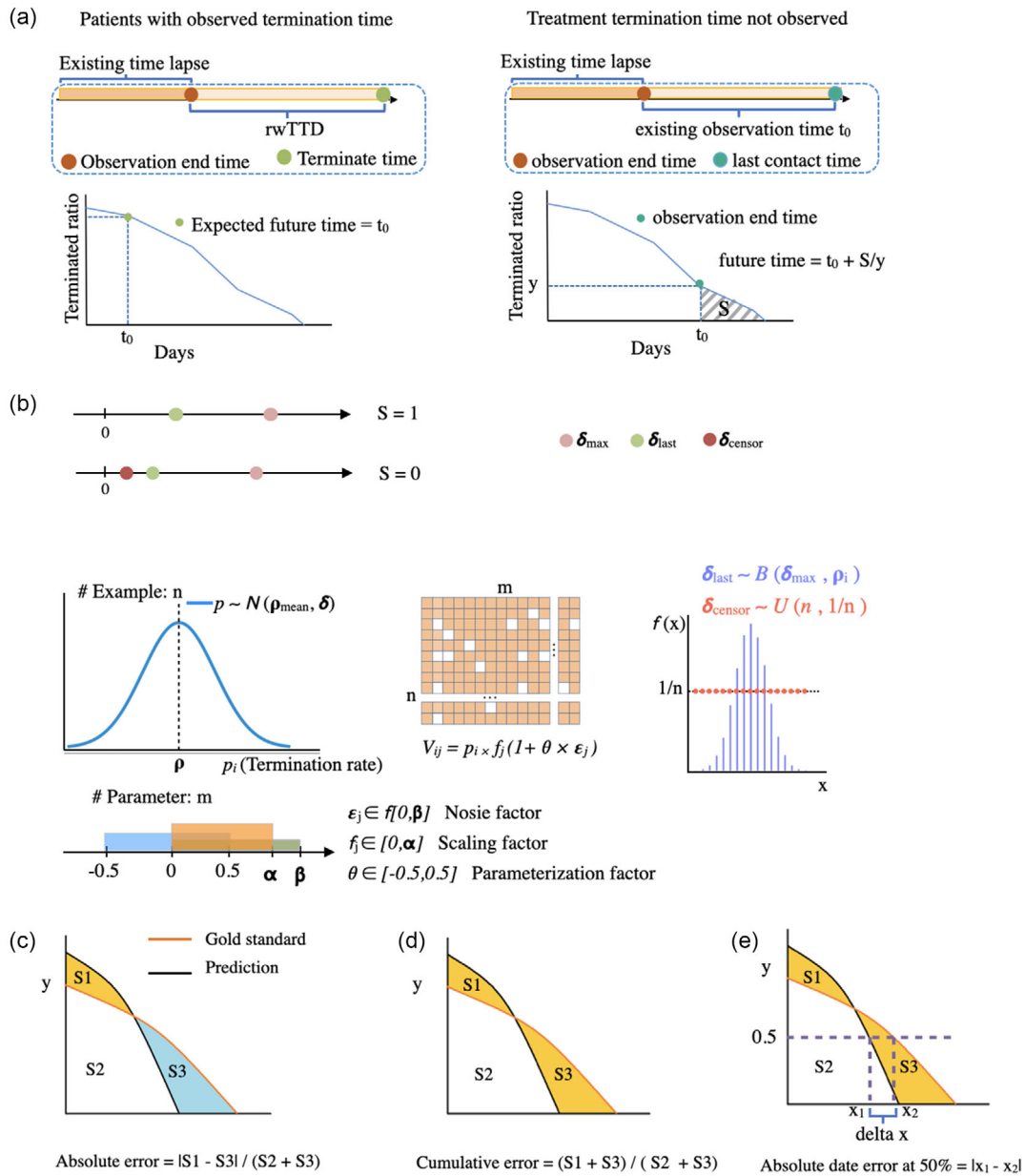


Figure 1. The machine learning and evaluation scheme of real-world time to treatment discontinuation (rwTTD) prediction. a) Calculation of future time in a censored population. b) Simulation of rwTTD data capturing a variety of factors potentially affecting performance. c–e) Three evaluation schemes used in the study: absolute error, cumulative error, and absolute number of error days when 50% of the population is terminated.

set and the corresponding future time is used as a reference to interpolate the prediction results of the test set. In this study, we used first-order interpolation and extrapolation if the test set prediction values go beyond the range of the validation set. By interpolation, we generated a distribution resembling the observed future time distribution of the test set. To further illustrate the functions of the three metrics we used in this study, we showed the illustrations of the percentage of errors using ExtraTreeRegressor by different numbers of maximal dates considered and the absolute error date when 50% of the population is terminated (Figure 2d,e).

2.2. Performance is Robust Across Different Simulated Situations

We started with $\delta_{\max} = 2000$, $p_{\text{mean}} = 0.0008$, $\sigma = 0.0008$. $\beta = 1$, $\alpha = 100$, $n = 5000$, $m = 100$. This created a dataset with 5000 patients and 100 clinical features. Unless otherwise specified for testing model robustness, these are the base parameters we used. We built in three commonly used algorithms for testing: ExtraTreeRegressor, linear regression, and SVM.^[14]

With the above starting point, we examined the behaviors of the model. With the increase in mean termination rate of the population, performance stayed strong. (Figure 3b,c, S1a,

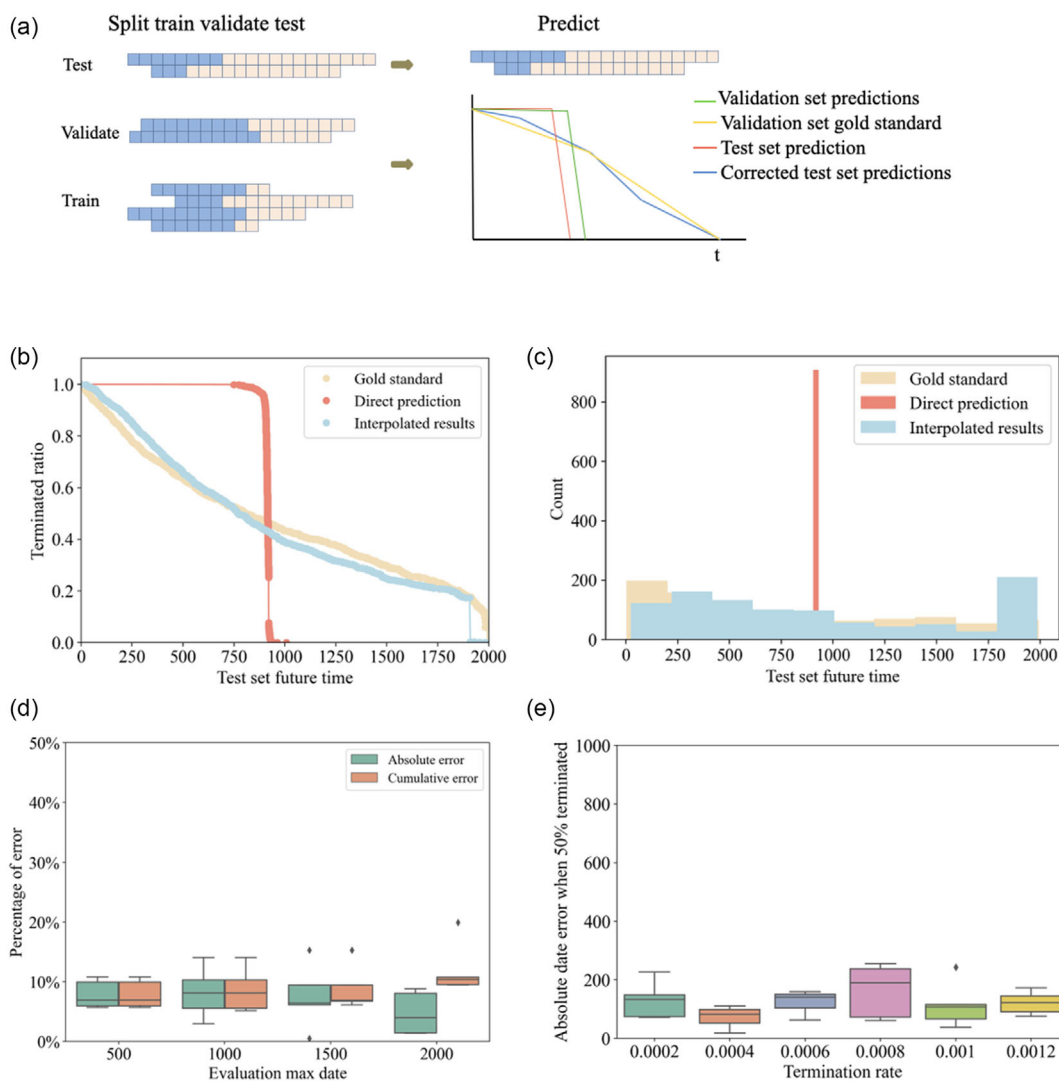


Figure 2. Interpolation resolves the discrepancy between the predicted value distribution and the true distribution of expected future time when using ExtraTreeRegressor as the base learner. a) Using a validation set to interpolate real-world distribution. b) Interpolation resolves the discrepancy between predicted values and the gold standard rwTTD curve. c) Comparison between the distribution of prediction values and gold standard rwTTD future time. d) Histogram of error rates at different evaluation maximal dates. e) Absolute error dates when 50% of the population is terminated.

Supporting Information, Figure 2, Table S2, Supporting Information). The error rate is around 10% and overall, we saw little variance when the termination rate of the population changes.

With the increase of examples, there is a steady decrease in the percent of error (Figure 3d, S1b, S3, Supporting Information). This is expected as we have more training examples, and the inference of the overall curve is improved. With 100 examples, the median error using cumulative errors is around 20% for all base learners (Table S3, Supporting Information). In contrast, with 10 000 examples, the median error using cumulative error is around 10%. We consider this is caused by more stable performance and inference of parameters in models with more training examples. In contrast, the number of predictive features does not affect performance (Figure 3e, S1c, S4, Supporting Information). Additionally, with a sufficient number of examples

(5000), the noise level on individual features does not affect model performance (Figure 3f, S1d, S5, Supporting Information). The above results demonstrated the overall robust performance of the model when the patients are derived from the same population.

2.3. Cross-Validation Across Two Distinct Populations Shows Strong Performance

We further examined the performance by simulating two distinct populations and examined the ability of model extrapolation across different cohorts. Both populations were simulated by the same approach as described in the previous section. Then, we focused on each of the parameters and changed this parameter through a grid search. In this case, we used

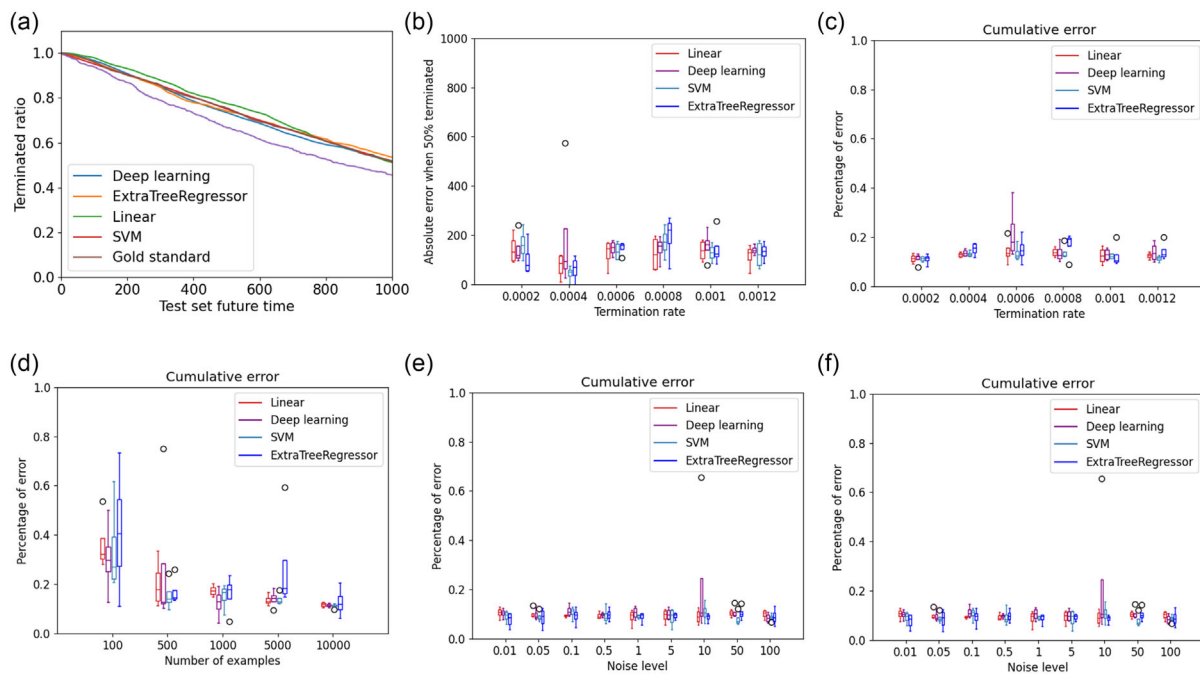


Figure 3. Performance of rwTTD prediction in homogeneous population during cross-validation. a) Example terminated ratio curve at 0.0008 termination rate. b) Comparison between predicted curve and gold standard curve by different base learners at different termination rates. c) Cumulative error at different termination rates. d) Cumulative error with different numbers of training examples. e) Cumulative error with different numbers of predictive features. f) Cumulative error with different feature noise levels.

ExtraTreeRegressor, which is a representative machine learning base learner.

The most important factor affecting the results we observed was the termination rates. When fixing the training set termination rate, the best performance is achieved when the test population is most similar to the training set, and deviates gradually when the two termination rates differ (Figure 4a, S6, S7, Supporting Information). For example, when the training set average termination rate is 0.0008, the model achieved an error rate of 5.464% for both metrics when the test set termination rate is also 0.0008. The error rate becomes higher at both tails when the test set termination error differs from the training set termination error: when the test set termination rate is 0.0002, the model achieved an error rate of 9.18% for absolute error and 9.29% for cumulative error. When the test set termination rate is 0.0012, the model achieved an error rate of 18.82% for both absolute error and cumulative error. This observation is expected, as if the termination rates of the two populations differ too much, and corresponding feature distributions (derived from the termination rate) do not overlap between the two populations, then it would be challenging to predict the patterns. Nevertheless, the error is much lower than directly using the training curve, for which we would expect a 50% error when trained with a 0.0008 termination rate and tested with a 0.0012 termination rate.

The other factors affected little on the performance. When the training set and test set were drawn from the same population, when increasing the number of training examples, the performance steadily improves, while the number of testing examples mainly affects the breadth of the performance (Figure 4b, S8, S9,

Supporting Information). Noise level on individual features does not affect overall performance on population-wise rwTTD (Figure 4c, S10, S11, Supporting Information). We then altered the scaling factor of the features. This alteration would result in feature values distributed at different scales, and thus addressing record disparities across cohorts. As expected, when the training and testing feature scales are similar, the model showed relatively low errors. As the two distributions deviate, the percentage of error increases. However, even when the training set feature scale is 1, and the test set feature scale is 1000, the overall population error was moderate (0.13481 for both metrics) (Figure 4d, S12, S13, Supporting Information). The above results point to a stable performance of the model across two distinct populations against a variety of factors.

2.4. Predicting Population-Level rwTTD for Lung Cancer and Advanced Head and Neck Cancer Treatment Using Pembrolizumab

We tested the above algorithm in the context of lung cancer treatment and head and neck cancer treatment using pembrolizumab (for cohort selection please see Experimental Section). rwTTD, the duration between the first dosing to the last administration is defined by the following three criteria: 1) switch to a different treatment: This is an event point, and rwTTD is defined between the first dosing to the last available administration. 2) death: This is also an event point, and rwTTD is defined between the first dosing to the death date. 3) With a gap ≥ 120 days between the last known administration and last known activity: This is

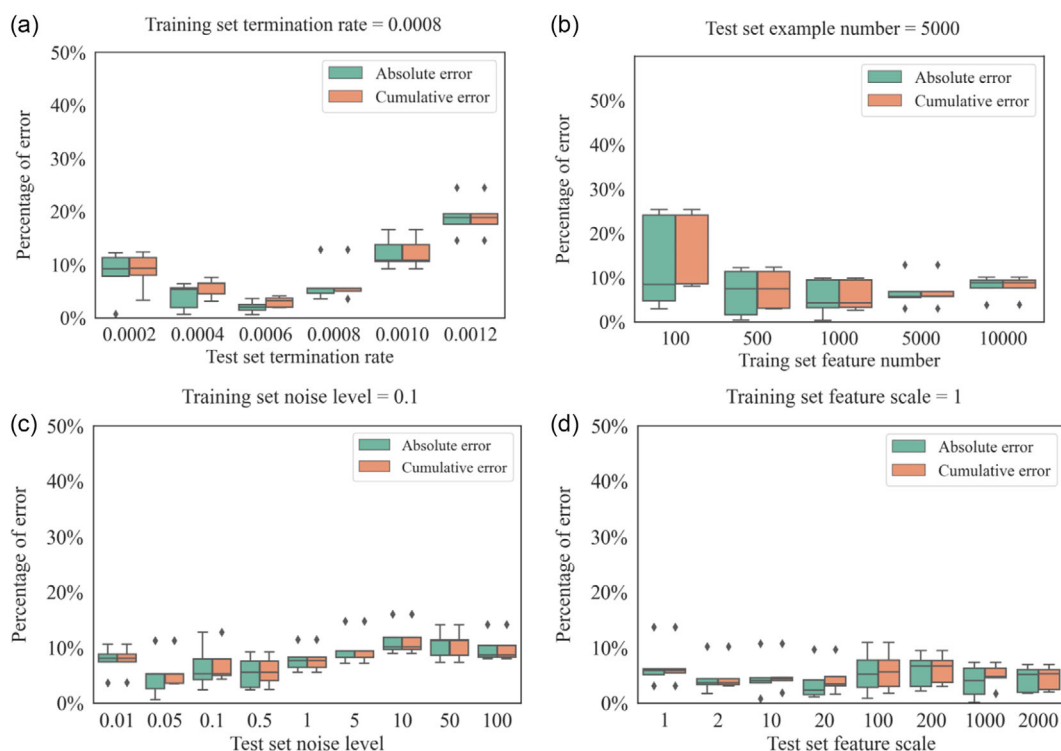


Figure 4. Performance of rwTTD prediction across heterogeneous populations. a) Performance of different test set termination rates, when the training set is at 0.0008 termination rate. b) Performance of different training set examples, when the number of test set examples is fixed at 5000. c) Performance of different test set noise levels, when the training set noise level is 0.1. d) Performance of different test set feature scales when the training set feature scale is 1.

an event point, and rwTTD is defined between the first dose to the last known available administration. If none of the above happens, the data point is considered as censored (no data after the last administration date or the gap is <120 days).

We carried out three evaluation experiments (Figure S14, Supporting Information). The first two experiments used advanced lung cancer data and examined the performance of prediction rwTTD in this homogeneous population. In the first experiment, we randomly selected the cutoff time between the first dose time and the last contact time point (let it be censoring time or termination time), and uniformly and randomly selected a time in between as the cutoff time. All information prior to the cutoff date (observation window) is used to extract feature data (see Experimental Section). The time between the cutoff time and the last contact time point is the time used to calculate the rwTTD curve. Here we are evaluating the ability of predicting rwTTD given a random length of observations. In the second experiment, the cutoff date is consistently 30 days after the first dose. Thus, we are evaluating how well we can predict given 30 days of observation data. The third experiment was trained with lung cancer data with a random cutoff and tested with head and neck cancer. Under these three scenarios, we evaluated the performance of predicting the rwTTD curve.

Overall, we found strong performance for rwTTD in both homogeneous population and cross-disease prediction tasks (Figure 5a–c, S15–S17, Supporting Information). There is a small discrepancy between the predicted curve, which is steep, and the observed curve. This is caused by the loss of recorded

death or end-of-treatment dates which we are unable to recover. We observed an average of 14.12%, 13.15%, and 31.59% percent absolute error rate for random cutoff cross-validation, 30-day cutoff cross-validation, and cross-disease prediction, respectively. The cumulative error rates are 23.78%, 18.43%, and 34.15%, respectively (Figure 5d). Of note, cross-disease errors are expected to be higher as the patient populations are distinct and can respond to the drug differently. We further examined the performance at 6, 12, 18, and 24 months, and error rates remained stable within this range (Figure 5e). In particular, we observed a very low average 50% terminated ratio date prediction, for only 82.90, 105.33, 81.90 for random cutoff cross-validation, 30 day cutoff cross-validation, and cross-disease, respectively (Figure 5f). These results support strong performance in real-world data even when the model is delivered to data derived from a different population but share certain similarities in the EMR data that was collected.

3. Discussion

In this study, we developed a strategy to incorporate machine learning into predicting real-world time-on-treatment curves. To this end, we generalized the problem into predicting the expected future time on treatment and then stratified the distribution of the predicted time. We showed strong performance of this approach in predicting rwTTD across a variety of influencing factors using simulated data. We showed its flexibility to be

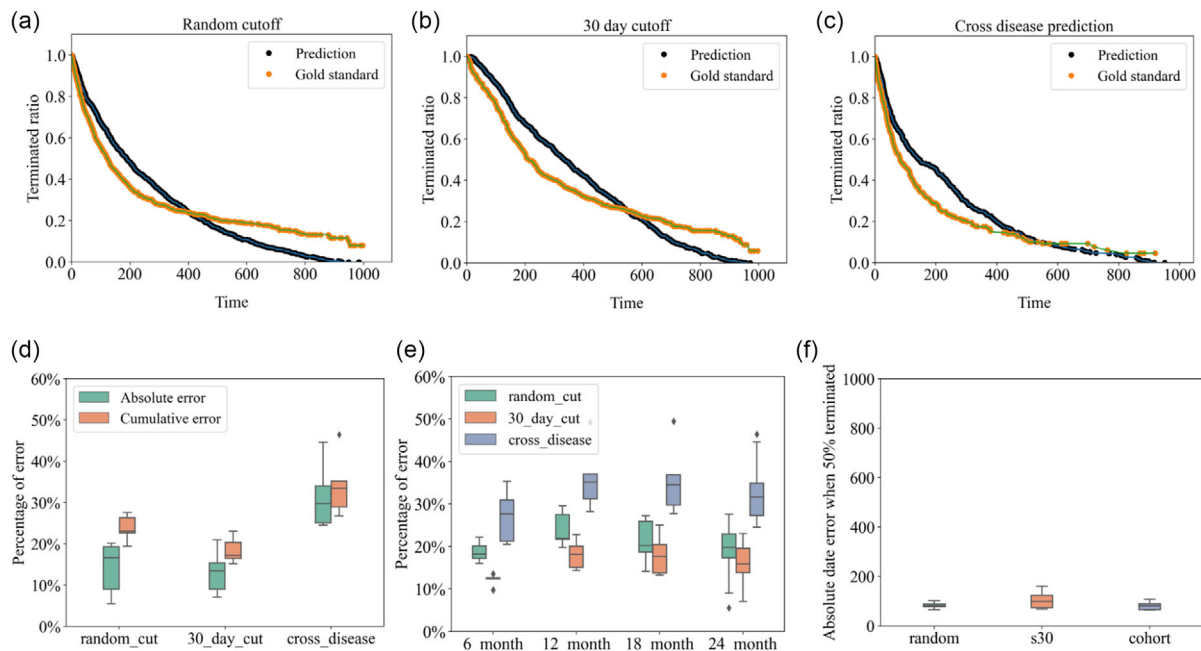


Figure 5. Performance of rwTTD models in real-world lung cancer and advanced head and neck cancer treatment using pembrolizumab. a) Comparison of the predicted curve and the gold standard curve with random cutoffs in lung cancer (fold 1). b) Comparison of the predicted curve and the gold standard curve with 30-day cutoff after treatment starts in lung cancer (fold 1). c) Training with lung cancer data and testing with head and neck data (fold 1). d) Percentage error up to 1000 days for random cutting, 30-day cut cross-validation and cross-disease predictions. e) Percentage error at 6, 12, 18, and 24 months, respectively. f) Absolute date error when 50% of the patients are terminated.

applied to any machine learning base classifiers. We then showed its robustness when trained and tested on different populations. Lastly, we demonstrated its robust performance using real-world lung cancer and head and neck cancer data treated with pembrolizumab.

Although rwTTD is a critical metric in monitoring the efficacy of a treatment in real-world patient populations, no study has yet attempted to establish machine learning models to predict rwTTD. The key obstacle is that rather than predicting individual scores, we are required to predict a curve. This notion and strategy is new, and will spur the field of curve prediction in many other research fields. Of note, we demonstrated that the aggregation of individuals does not reflect the overall profile of the population, which is an important rationale behind the approach we presented in this study.

There are several potential limitations of the study. First, we observed a deviation in the predicted population profiles versus the real-world profiles for the pembrolizumab studies. This is likely caused by incomplete records in the clinical trial database, where some terminations of the treatment were not recorded. Second, time on treatment is often used as a surrogate to real-world treatment failure, as termination of cancer treatment drugs is often caused by death or switch to other drugs. While this is a widely accepted approximation, this is entirely not true for all cases, because drugs can be terminated by successful treatment and/or clinicians' choices as well. Deviations from this assumption could therefore lead to inaccuracy of the population-wise curve modeling.

This study opens the possibility of many follow-up directions. For example, can such models be applied to clinical trial data, and

using the generated model to predict real-world populations? Can models be well generalized from one demographic group to another? Can this approach be extended to combinatorial drug prediction when only single drug observations are available? While we touched on these aspects using simulated data and real-world pembrolizumab data, it will be of interest to test in other diseases and drugs as well. How does the interpolation function affect the performance of the model? How do other base learners such as deep learning, and Gaussian process regression work with this model? Our approach allows the incorporation of any supervised base learner which can be tested in future studies concerning other diseases and therapeutics. Finally, this study opens the possibility of population-wise predictions, which is distinguished from individual-wise prediction. This will have enormous applications in the future in all research areas whose current focus is on individual predictions.

4. Experimental Section

Base Learner Implementation and Parameters: For the simulation experiment, four base learners were tested: ExtraTreeRegressor, linear regression, support vector machine (SVM), and deep learning. For ExtraTreesRegressor, 1000 trees with a maximal depth of 3, squared error as the criterion of split, minimal number of examples as 2 in a split, and minimal number of examples in a node as 1 were used. For SVM, support vector regressor (SVR) implemented in sklearn, with $C = 1.0$, and $\epsilon = 0.2$, was used. For linear regression, ridge penalization with $\alpha = 1.0$ was used. For deep learning, four dense layers with sizes [256, 128, 64, 1] were used to progressively extract information from training data. Mean squared error was used as the loss function, initial learning rate of 0.001, and decay = $1e-6$ with Adam optimizer.

Selection of Cohorts from Flatiron Health Database: We used the following criteria to select advanced NSCLC Patients and advanced head and neck patients from nationwide de-identified electronic health record-derived Flatiron Health database. The Flatiron Health database is a longitudinal database, comprising de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction. During the study period, the de-identified data originated from approximately 280* US cancer clinics (≈ 800 sites of care)^[15,16] 1) The patient should be ≥ 18 years of age at advanced diagnosis. 2) There should be some kind of activity (in drug administration or visit table) within 90 days of the advanced diagnosis. 3) The patient should have at least 1 record of systemic anti-cancer drugs. 4) Exclude drug records that are part of clinical trials. This resulted in 4,784 NSCLC patients and 422 advanced head and neck cancer patients included in this study. The demographic profiles for these patients are described in Table S1, Supporting Information.

Processing of Feature Data: The following data tables were used for feature extraction before the cutoff date for predicting future time: ECOG, enhanced biomarkers, demographics, diagnosis code, visit code, telemedicine code, medication administration code, insurance, lab results, medication order, vitals, and practice.

Feature data can be largely separated into two categories. One set is static data, which does not change over the observation time course, including age, gender, race, etc. The other set is dynamic data, including lab, medication, visit, vitals, diagnosis, etc., which are collected before the cutoff date. For this set of data, diverse meta-features were used. First, the most frequent 100 concept IDs in each of the above Flatiron data tables were selected, and the last eight points of records were binarized (if not originally a continuous value) to generate 800 features, with 1 representing the appearance of the concept ID at that data point, and 0 otherwise. Additionally, if the concept ID represents a real-valued feature, the mean value and the standard deviation of each selected concept ID before the cutoff time were included. Using these mean and the standard deviation, normalized values for the initial 800 features for each table were generated, and the time difference between each record and the previous one was recorded. Lastly, a binary indicator was included for each original feature whether it comes from a missing record (8 values for each Flatiron data table) or an existing record. This matrix will be flattened into a single feature vector, concatenated with the static features, and input into lightGBM.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

Correction added on April 21st, 2023 after online publication: The copyright year was corrected.

Conflict of Interest

WM, BR are Merck Sharp & Dohme LLC, a subsidiary of Merck & Co., Inc., Rahway, NJ, USA employees. YG serves as scientific advisor to Merck & Co., Inc., Rahway, NJ, USA on this project.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions. Code is available at <https://github.com/Merck/Real-world-Time-to-Treatment-Discontinuation-Prediction-Algorithm>.

Keywords

drug efficacy, machine learning, rwToT, time-series prediction

Received: August 9, 2022

Revised: October 12, 2022

Published online: January 31, 2023

- [1] S. Yang, A. A. Tsiatis, M. Blazing, *Biometrics* **2018**, *74*, 900.
- [2] Y. Gong, K. L. Kehl, G. R. Oxnard, S. Khozin, P. S. Mishra-Kalyani, G. M. Blumenthal, *J. Clin. Oncol.* **2018**, *36*, 9064.
- [3] M. S. Walker, L. Herms, P. J. E. Miller, *J. Clin. Oncol.* **2020**, *38*, e19135.
- [4] K. Ramakrishnan, Z. Liu, S. Baxi, S. Chandwani, S. Joo, D. Chirovsky, *Future Oncol.* **2021**, *17*, 3037.
- [5] V. Velcheti, X. Hu, Y. Li, H. El-Osta, M. C. Pietanza, T. Burke, *Cancers* **2022**, *14*, 1041.
- [6] J. Bauml, T. Y. Seiwert, D. G. Pfister, F. Worden, S. V. Liu, J. Gilbert, N. F. Saba, J. Weiss, L. Wirth, A. Sukari, H. Kang, M. K. Gibson, E. Massarelli, S. Powell, A. Meister, X. Shu, J. D. Cheng, R. Haddad, *J. Clin. Oncol.* **2017**, *35*, 1542.
- [7] G. M. Blumenthal, Y. Gong, K. Kehl, P. Mishra-Kalyani, K. B. Goldberg, S. Khozin, P. G. Kluetz, G. R. Oxnard, R. Pazdur, *Ann. Oncol.* **2019**, *30*, 830.
- [8] M. Stewart, A. D. Norden, N. Dreyer, H. J. Henk, A. P. Abernethy, E. Chrischilles, L. Kushi, A. S. Mansfield, S. Khozin, E. Sharon, S. Arunajadai, R. Carnahan, J. B. Christian, R. A. Miksad, L. C. Sakoda, A. Z. Torres, E. Valice, J. Allen, *JCO Clin. Cancer Inform.* **2019**, *3*, 1.
- [9] P. Royston, M. K. B. Parmar, *BMC Med. Res. Methodol.* **2013**, *13*, 33.
- [10] G. Lopes, Y.-L. Wu, I. Kudaba, D. Kowalski, B. C. Cho, G. Castro, V. Srimuninimit, I. Bondarenko, K. Kubota, G. M. Lubiniecki, J. Zhang, D. A. Kush, T. Mok, *J. Clin. Oncol.* **2018**, *36*, LBA4.
- [11] A. Argiris, J. Johnson, Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature **2017**, <https://doi.org/10.3410/f.727429341.793534852>.
- [12] E. L. Kaplan, P. Meier, *Springer Series in Statistics*, Springer Series in Statistics **1992**, pp. 319–337, https://doi.org/10.1007/978-1-4612-4380-9_25.
- [13] Y. Guan, H. Li, D. Yi, D. Zhang, C. Yin, K. Li, P. Zhang, *Nat Comput Sci* **2021**, *1*, 433.
- [14] S. Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification*, Springer, US **2016**, pp. 207–235, https://doi.org/10.1007/978-1-4899-7641-3_9.
- [15] X. Ma, L. Long, S. Moon, B. J. Adamson, S. S. Baxi, (*preprint*) *Medrxiv.*, <https://www.medrxiv.org/content/10.1101/2020.03.16.20037143v2>, submitted Jan **2020**.
- [16] B. Birnbaum, N. Nussbaum, K. Seidl-Rathkopf, M. Agrawal, M. Estevez, E. Estola, J. Haimson, L. He, P. Larson, P. Richardson, (*preprint*) *arXiv:2001.09765*, <https://arxiv.org/abs/2001.09765>, submitted Jan **2020**.