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

Weilin Meng[1], Xinyuan Zhang[1], Boshu Ru[1], and Yuanfang Guan[1]

[1]Affiliation not available

August 16, 2022

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 high correlation to progression-free survival and 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[2,4,6–9].

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[10], 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 cancer populations[7,11,12]. The study opens a new direction of modeling population-level rwTTD, which has great values for directing post-clinical stage drug administrations. This machine learning scheme will also have meaningful implications to population-based predictions for other problems, as machine learning algorithms have so far been focused on predictions for individual samples.

Results

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 (**Fig. 1a**)[1]. However, existing survival models only predict individual patient's likelihood of survival. As shown below 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 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, Kaplan–Meier curve, to represent the termination ratio of the training set [13]. 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 (**Fig. 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 kinds 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 [14](**Fig. 1b**). We generated a population of total $n$individuals, where the termination rate for each individual is drawn from a population of p ˜ $N(p_{mean}$,?), 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 below. In this simulation experiment, we Let individual feature values correlates to p by:

$$v_{\mathrm{kj}} = p_k \times f_j(1 + \theta \times \epsilon_j)$$

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, ?]. Each feature j is parameterized by noise factor $?_j$, uniformly drawn from [0, ?]. When ? goes up, a larger sampling range will result in less correlation between the feature and the expected future time. The value of the *jth* feature of the *kth* sample,v_{kj} , is further parameterized by ?, which is uniformly distributed sampled between [-0.5, 0.5].

We set the maximal allowed observation date of all individuals to$?_{max}$. Between $[0, ?_{max}$], we create a binomially distributed vector of length $?$ $_k$ ˜B ($?_{max}$, , 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 $[0, ?_{max}$] and define the censoring datet_{censor} . Ift_{term} $>t_{censor}$, the last observation timet_{last}=t_{censor} , and the status is 0 (censored point and no termination date is observed); otherwise, thet_{last}=t_{term} with a status =1 (termination observed and the date is defined).

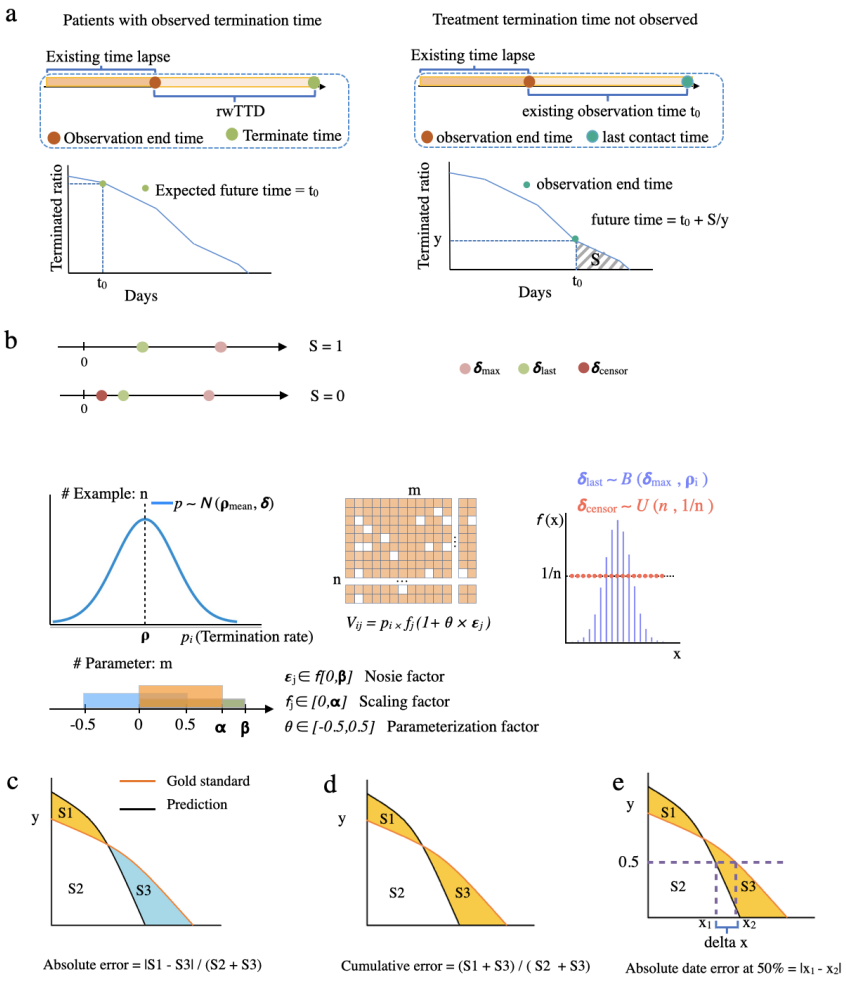

Figure 1. The machine learning and evaluation scheme of rwTTD prediction. a. Calculation of future time in a censored population. b. Simulation of rwTTD data capturing a variety of factors potentially affecting performance. c-e. Three evaluation schemes used in the study: absolute error, cumulative error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

We developed three metrics to evaluate the model performance (**Fig. 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 paper), 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 at 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 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 in 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 (**Fig.**

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 set and corresponding future time are 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 either the absolute error or the cumulative errors using ExtraTreeRegressor by different numbers of maximal dates considered and the absolute error date when 50% of the population is terminated (**Fig. 2d-e**).

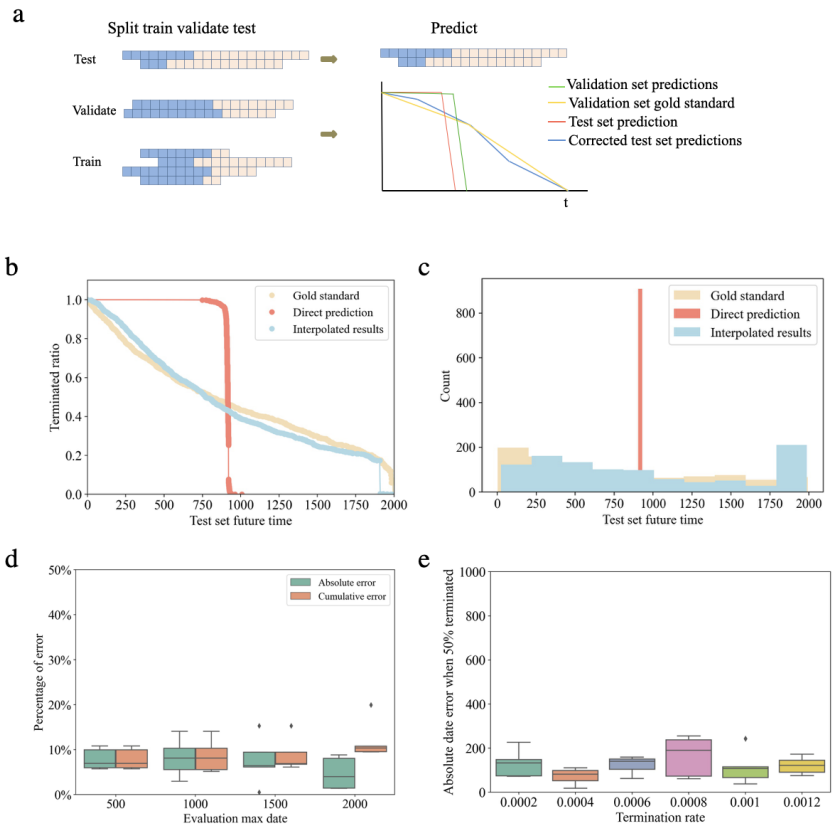

Figure 2 . Interpolation resolves discrepancy between the predicted value distribution and the true distribution of expected future time when using ExtraTreeRegressor as the base learner. a. Using a validation set to interpolate real-world distribution. b. Interpolation resolves the discrepancy between 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.

Performance is robust across different simulated situations

We started with $?_{max}$= 2000,p_{mean} = 0.0008, $?$ = 0.0008. ? = 1, ? = 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 Support Vector Machines (SVM)[15].

With the above starting point, we examined the behaviors of the model. With the increase of mean termination rate of the population, performance stayed strong. (**Fig 3b-c, Fig S1a, Fig 2**). The median error rate at p_{mean} = 0.0008 for cumulative errors are 9.11%, 8.97%, 9.10% for ExtraTreeRegressor, Linear Regression, and SVM, respectively, compared to 7.89%, 8.15%, 7.47%, which are their respective errors at p_{mean}= 0.0012 . Overall, we saw little variance when the termination rate of the population changes.

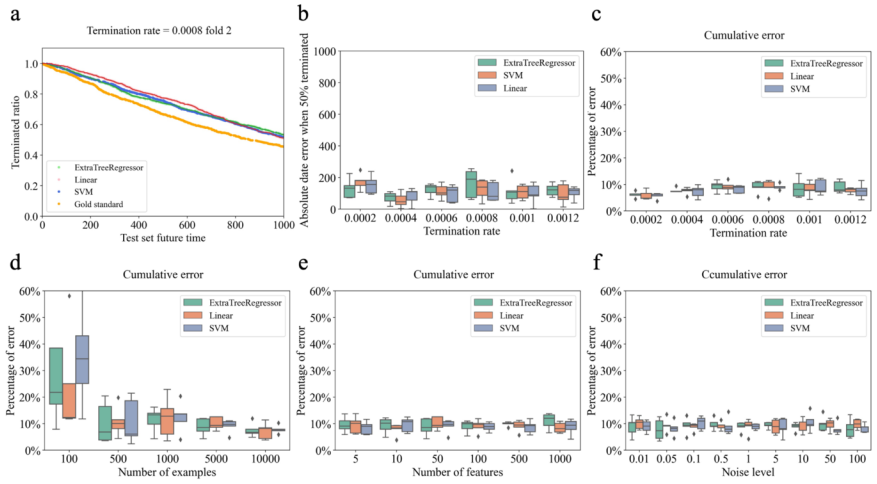

Figure 3 . Performance of rwTTD prediction in homogeneous population during cross-validation. a. Example terminated ratio curve at 0.0008 termination rate. b. Comparison between predicted curve and gold standard curve by different base learners at different termination rates. c. Cumulative error at different 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.

With the increase of examples, there is a steady decrease in the percent of error (**Fig 3d, Fig. S1b, Fig. S3**). This is expected as we have more training examples, the inference of the overall curve is improved. With 100 examples, the median error using cumulative errors are 19.84%, 22.92%, 20.22% for ExtraTreeRegressor, Linear Regression, and SVM respectively. In contrast, with 10,000 examples, the median errors using cumulative error is 6.81%, 7.95%, 6.28% for ExtraTreeRegressor, Linear Regression, and SVM, respectively. We consider this is caused by more stable performance and inference of parameters in models with more training examples. On the other hand, the number of predictive features does not affect performance (**Fig. 3e, Fig. S1c, Fig. S4**). Additionally, with a sufficient number of examples (5000), noise level on individual features does not affect model performance (**Fig. 3f, Fig. S1d ,Fig. S5**). The above results demonstrated the overall robust performance of the model when the patients are derived from the same population.

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 ExtraTreeRegressor, which is a representative machine learning base learner.

The most important factor affecting 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 (**Fig. 4a, Fig. S6-7**). 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 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 0.0008 termination rate and tested with 0.0012 termination rate.

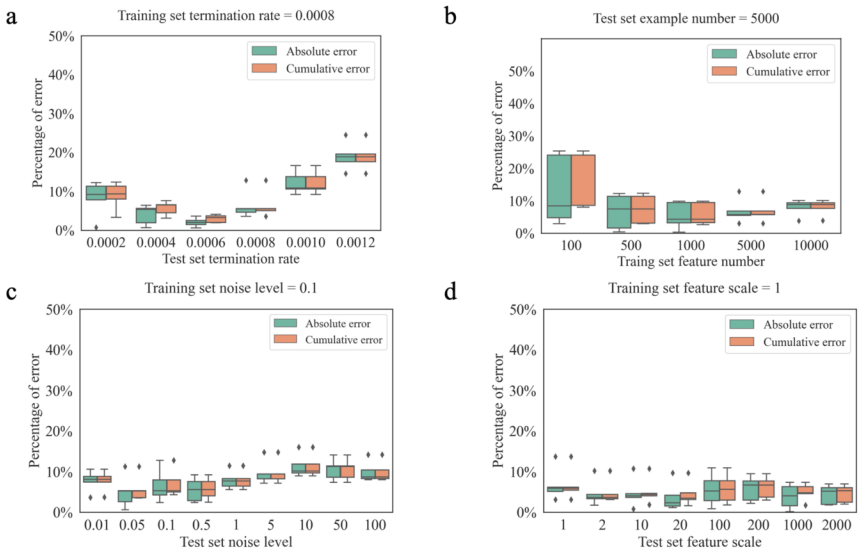

Figure 4. Performance of rwTTD prediction across heterogeneous populations. a. Performance of different test set termination rates, when the training set is at 0.0008 termination rate. b. Performance of different training set examples, when the number of test set examples is fixed at 5000. c. Performance 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.

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 (**Fig. 4b, Fig. S8-9**). Noise level on individual features does not affect overall performance on population-wise rwTTD (**Fig. 4c, Fig. S10-11**). 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) (**Fig. 4d, Fig. S12-13**). The above results point to a stable performance of the model across two distinct populations against a variety of factors.

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 **Methods**). rwTTD, the duration between the first

dosing to the last administration are defined by the following three criteria: a. switch to a different treatment: This is an event point, and rwTTD is defined between the first dosing to the last available administration. b. death: This is also an event point, and rwTTD is defined between the first dosing to the death date. c. With a gap >= 120 days between last known administration and last known activity: This is 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 last administration date or the gap is < 120 days).

We carried out three evaluation experiments (**Fig. S14**). 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**Methods**). 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 sceneria, 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 (**Fig. 5a-c, Fig. S15-17**). We observed an average 14.12% 13.15%, 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%, 34.15% respectively (**Fig. 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 (**Fig. 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 (**Fig. 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.

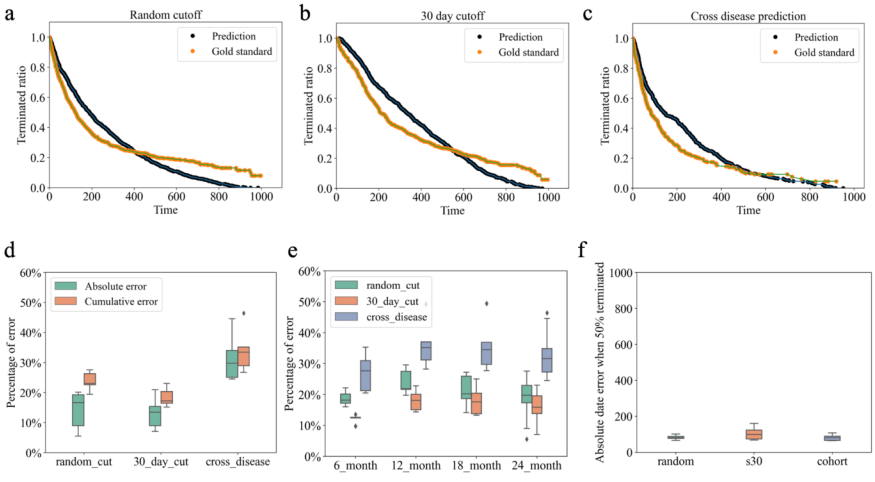

Figure 5 . Performance of rwTTD models in real world lung cancer and advanced head and neck cancer treatment using pembrolizumab. a. Comparison of predicted curve and gold standard curve with random cutoffs in lung cancer (fold 1). b. Comparison of predicted curve and gold standard curve with 30 day 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.

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 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 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 the 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.

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? While we touched 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, Gaussian Progress Regression work with this model? Our approach allows 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.

Methods

Base learner implementation and parameters

For the simulation experiment, we tested three base learners: ExtraTreeRegressor, Linear Regression and Support Vector Machines (SVM). For ExtraTreesRegressor, we used 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. For SVM, we used the SVR (support vector regressor) implemented in sklearn, with C=1.0, and epsilon = 0.2. For Linear Regression we used Ridge penalization with alpha = 1.0.

Selection of cohorts from Flatiron Health database

We used the following criteria to select advanced NSCLC Patients and advanced head and neck patients from Flatiron Health database[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** .

Processing of feature data

We used the following data tables for feature extraction before the cutoff date: 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, we extracted diverse meta-features. We first selected the most frequent 100 concept IDs in each of the above Flatiron data tables, and the last eight points of records are 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 are included. Using these mean and the standard deviation, we generate normalized values for the initial 800 features for each table, and we record the time difference between each record and the previous one. Lastly, we include a binary indicator 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.

Code availability

Code is open upon reasonable request to the corresponding authors.

Competing interests

WM, BR are Merck & Co. employees. XZ is an Ann Arbor Algorithms employee. YG serves as scientific advisor to Merck & Co. on this project.

References

1. Yang, S., Tsiatis, A. A. & Blazing, M. Modeling survival distribution as a function of time to treatment discontinuation: A dynamic treatment regime approach.*Biometrics* **74**, 900–909 (2018).

2. Gong, Y. *et al.* Time to treatment discontinuation (TTD) as a pragmatic endpoint in metastatic non-small cell lung cancer (mNSCLC): A pooled analysis of 8 trials.*Journal of Clinical Oncology* vol. 36 9064–9064 (2018).

3. Walker, M. S., Herms, L. & Miller, P. J. E. Performance of time to discontinuation and time to next treatment as proxy measures of progression-free survival, overall and by treatment group. *Journal of Clinical Oncology* vol. 38 e19135–e19135 (2020).

4. Ramakrishnan, K. *et al.* Real-world time on treatment with immuno-oncology therapy in recurrent/metastatic head and neck squamous cell carcinoma. *Future Oncol.* **17**, 3037–3050 (2021).

5. Velcheti, V. *et al.*Real-World Time on Treatment with First-Line Pembrolizumab Monotherapy for Advanced NSCLC with PD-L1 Expression [?] 50%: 3-Year Follow-Up Data.*Cancers* vol. 14 1041 (2022).

6. Weis, T. M., Hough, S., Reddy, H. G., Daignault-Newton, S. & Kalemkerian, G. P. Real-world comparison of immune checkpoint inhibitors in non-small cell lung cancer following platinum-based chemotherapy. *J. Oncol. Pharm. Pract.***26**, 564–571 (2020).

7. Bauml, J. *et al.*Pembrolizumab for Platinum- and Cetuximab-Refractory Head and Neck Cancer: Results From a Single-Arm, Phase II Study. *J. Clin. Oncol.* **35**, 1542–1549 (2017).

8. Blumenthal, G. M. *et al.* Analysis of time-to-treatment discontinuation of targeted therapy, immunotherapy, and chemotherapy in clinical trials of patients with non-small-cell lung cancer. *Ann. Oncol.* **30**, 830–838 (2019).

9. Stewart, M. *et al.*An Exploratory Analysis of Real-World End Points for Assessing Outcomes Among Immunotherapy-Treated Patients With Advanced Non-Small-Cell Lung Cancer. *JCO Clin Cancer Inform* **3**, 1–15 (2019).

10. Royston, P. & Parmar, M. K. B. Restricted mean survival time: an alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC Medical Research Methodology*

vol. 13 (2013).

11. Lopes, G. *et al.*Pembrolizumab (pembro) versus platinum-based chemotherapy (chemo) as first-line therapy for advanced/metastatic NSCLC with a PD-L1 tumor proportion score (TPS) [?] 1%: Open-label, phase 3 KEYNOTE-042 study.*Journal of Clinical Oncology* vol. 36 LBA4–LBA4 (2018).

12. Argiris, A. & Johnson, J. Faculty Opinions recommendation of Pembrolizumab for Platinum- and Cetuximab-Refractory Head and Neck Cancer: Results From a Single-Arm, Phase II Study. *Faculty Opinions – Post-Publication Peer Review of the Biomedical Literature* (2017) doi:10.3410/f.727429341.793534852.

13. Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Springer Series in Statistics* 319–337 (1992) doi:10.1007/978-1-4612-4380-9_25.

14. Guan, Y. *et al.* A survival model generalized to regression learning algorithms. *Nat Comput Sci* **1**, 433–440 (2021).

15. Suthaharan, S. Support Vector Machine. *Machine Learning Models and Algorithms for Big Data Classification* 207–235 (2016) doi:10.1007/978-1-4899-7641-3_9.

16. Flatiron Health.*Flatiron Health* https://flatiron.com/ (2017).