Modeling the Progression of Type-I Diabetes Using Machine Learning

Elizabeth Holden, Manya Chadha, Sarah Nguyen

Introduction

Type 1 diabetes (T1D) is an autoimmune condition in which the immune system progressively destroys pancreatic β-cells, leading to an absolute need for insulin therapy. The presence of islet autoantibodies (AABs) is the strongest predictor of T1D risk, with nearly all individuals who test positive for multiple AABs eventually developing the disease. However, the rate of progression varies significantly among individuals, making early detection and risk stratification crucial for delaying disease onset and reducing severe complications.

Early identification of high-risk individuals is essential for improving clinical outcomes. The ability to predict disease progression enables timely interventions that may delay or even modify the course of T1D. One of the most severe complications associated with delayed diagnosis is diabetic ketoacidosis (DKA), a life-threatening metabolic crisis resulting from prolonged insulin deficiency. More than 30% of individuals present with DKA at diagnosis, which is associated with increased morbidity and long-term complications. However, early screening and regular metabolic monitoring have been shown to reduce the incidence of DKA by a factor of 10, underscoring the need for accurate and accessible risk prediction tools. Additionally, current staging classifications (Stage 1: normal glucose tolerance, Stage 2: abnormal glucose tolerance) fail to capture individual variability in disease progression, making it difficult to determine the optimal frequency and intensity of monitoring for at-risk individuals. A more personalized approach to T1D risk assessment is necessary to optimize clinical decision-making and improve patient outcomes.

To improve risk stratification, Stephan Pribitzer et al. (2022) at the Benaroya Research Institute (BRI) developed a personalized risk calculator for T1D progression, as described in Beyond Stages: Predicting Individual Time-Dependent Risk for Type 1 Diabetes. The calculator employs Bayesian survival modeling using Weibull priors to predict individual disease trajectories using data from TrialNet's Pathway to Prevention study.

The model was designed in three versions (small, medium, large) to accommodate different clinical scenarios and testing availability:

- Small model: Uses AAB status, fasting glucose, hemoglobin A1c (HbA1c), and age.
- Medium model: Adds 2-hour glucose and C-peptide from oral glucose tolerance tests (OGTT).
- Large model: Incorporates additional OGTT-derived predictors.

While these models provide personalized, time-dependent risk estimates, their complexity makes it challenging for clinicians to interpret and update them as new data becomes available. To address these

limitations, our project explores machine learning (ML) approaches to enhance the model's usability, explainability, and predictive performance.

Problem Statement

The Benaroya Research Institute's Bayesian risk calculator represents an important step toward personalized risk assessment. However, its reliance on complex statistical methods poses challenges in model interpretation and adaptability, particularly as more patient data becomes available.

To address these challenges, our project focuses on three key areas:

1. Validating and reproducing the BRI model's results on independent datasets - FR1DA and DPT-1.
2. Exploring machine learning alternatives to improve model interpretability, and usability.
3. Enhancing the user interface to support real-time clinical decision-making.

By improving model accuracy, usability, and explainability, we aim to provide a tool that effectively guides clinicians in monitoring and managing patients at risk of T1D. This tool would fulfill the 2 complementary goals as defined in the paper[1] : to delay the onset of clinical disease and to reduce the morbidity associated with severe hyperglycemia and diabetic ketoacidosis (DKA), often present at the time of clinical diagnosis.

Background Research

The team reviewed literature surrounding explainability frameworks in ML, various forecasting methodologies and architectures and current work in risk forecasting of type 1 diabetes. The goal was to cover a breadth of topics and set a foundation for a proposed solution.

1. **Forecasting Methodology for Diabetes Onset - Random Survival Forests**

The paper "Predicting Time to Diabetes Diagnosis Using Random Survival Forests" by Saha et al. (2024)[1], provides a new approach to predicting time till diagnosis for Type 2 Diabetes using a Random Survival Forest model. This approach is novel because this model integrates survival analysis into the random forest algorithm, so instead of predicting the binary outcome of whether someone has or will get T2D, the model predicts the timeline for diagnosis. This is a similar output of the current models that Benaroya Research is using for predicting T1D timeline to diagnosis, yet they are using Bayesian based statistical models. I am very interested in taking the methods presented in this paper that were used for T2D and applying them to the T1D and the data that we have.

This paper has a big impact on the industry by allowing prediction for time to diagnosis, there can be earlier intervention for patients. This is the same goal that Benaroya Research has for our capstone project for T1D. The goal is to make predictions of time to diagnosis easily accessible in a clinical setting which

[1] Saha, P., Marouf, Y., Pozzebon, H., Guergachi, A., Keshavjee, K., Noaeen, M., & Shakeri, Z. (2024). Predicting time to diabetes diagnosis using random survival forests. medRxiv. https://doi.org/10.1101/2024.02.03.24302304

would allow for more personalized monitoring of patients who will eventually be diagnosed. More frequent testing and monitoring of patients allows for diagnosis and medical intervention before the patient reaches diabetic ketoacidosis (DKA) and is in a critical state.

The research in this paper shows strong evidence of good performance with a concordance index of 0.84. For this study, tuning parameters were chosen for quick training, so with other choices made there is potential for higher accuracy. This study uses biomarkers that are specific to T2D, which is different from a previous study that predicted time to diagnosis with cardiovascular fitness. The dataset we have contains biomarkers specific to T1D so we can apply this methodology to our features. The article mentions that there is room for improvement around dealing with missing data and mentions that many of the patients in their dataset only had a single record. In order to apply this methodology, we would have to look into how many records, and if they are time ordered, the patients in our dataset have. The potential issue or limitation of using this model is that the purpose of this model architecture is to capture longitudinal measurements, so if the dataset does not have many repeated measurements, then there is no longitudinal data to capture.

2. Joint Modelling for survival analysis and repeated measurements

The paper "Joint modelling of repeated measurement and time-to-event data: an introductory tutorial" by Asar et. al (2015)[2] discusses a methodology for simultaneously analyzing two types of data: longitudinal (repeated measurements) and survival (time-to-event) data. This approach is particularly useful when both types of data are collected for the same subjects over time. The data in TrialNet, DPT-1, and FR1DA, fulfills this condition by collecting repeated biomarker measurements and status of T1D over time.

Joint modeling combines two sub models through shared random effects:

- **Longitudinal Submodel (Linear Mixed Effects Model):** Linear mixed-effects models (LME) that capture the trajectory of key biomarkers like Hba1c over time.
- **Survival Submodel (Weibull Distribution):** Weibull parametric survival models that estimate time to T1D diagnosis.
- **Association structures:** Generally involves using either the absolute value or using it in conjunction with the slope of the biomarker being estimated.

These models are linked through shared random effects and specific association structures, enabling the simultaneous analysis of the repeated measurements and survival outcomes. The primary advantage of joint modeling over separate analyses is that it accounts for the measurement error in longitudinal data, which could otherwise lead to biased estimations of the relationship between the longitudinal measures and survival outcomes.

In the paper, the authors use data from the Chronic Renal Insufficiency Standards Implementation Study (CRISIS) to illustrate the method. This dataset includes repeated measurements of kidney function (eGFR) and survival data (time to renal replacement therapy, RRT). By using joint modeling, they

[2] Asar, Özgür et al. “Joint modelling of repeated measurement and time-to-event data: an introductory tutorial.” International journal of epidemiology vol. 44,1 (2015): 334-44. doi:10.1093/ije/dyu262

demonstrate that the association between eGFR and the risk of RRT is better understood by considering both the longitudinal data and survival data together, much like we aim to do with T1D progression.

Datasets & Pipeline

Since our project utilizes proprietary, highly sensitive health research data, access to the training data has been provided through a private repository via Github. The data is shared as a .Rds file, which is a binary file format native to R. The training dataset is moderately large, and has 33821 rows corresponding to patient IDs and 25 columns corresponding to different biological and medical features. It should be noted that there are 6193 unique patient IDs. We are using a python library called “pyreadr” to access the data and a data dictionary provided by the sponsors as reference for column names.

We were provided with three different datasets, located in Table 1. Each dataset contained longitudinal biomarker data, along with the status of T1D at the time of each measurement.

Dataset	Age range	Number of unique patients	Number of observations	Details	Notes
PTP: TrialNet Pathway to Prevention	2 - 45	6193	33,821	Screening study of individuals at risk for T1D, based in the US	Subjects had >=1 islet autoantibody present at 2 separate visits and at least 1 visit with metabolic monitoring including OGTT
FRIDA	2 - 10	420	2,277	T1D screening study based in Germany	NIL
DPT-1	0 - 50 (mostly < 20 years)	274	1,050	Insulin therapy trial based in the US	Relatives of T1D patients at intermediate to high risk for T1D.

Table 1. Dataset Overview

Methodology

We had initially explored both deep and machine learning methods. Some of the deep learning methods we considered were algorithms like TransformerJM, DeepSurv, and DeepHit. However, due to the longitudinal nature of the data and the final goal being to perform inference with one single datapoint, we decided to move forward with the following models:

Modeling approach

Random Survival Forest

Random Survival Forest[3] (RSF) extends the traditional random forest algorithm to handle time-to-event data, making it suitable for predicting the survival function of patients at risk for T1D. In our project, RSF was employed to estimate cumulative incidence via the transformation $1 - S(t)$[4], where $S(t)$ is the survival function.

Model Assumptions and Target Variable

RSF assumes independence of observations, which is crucial when applying it to longitudinal clinical data. The target variable is defined as a time interval,$totaltime = t\_stop - t\_start$, measured in years, representing the duration until T1D onset or censoring. Given RSF's assumption of independent observations, we explored three strategies for incorporating longitudinal data:

1. **Last Observation Approach**: Using only the final observation from each patient.
2. **First Biomarkers with Last T1D Outcome**: Utilizing the first recorded biomarker measurements paired with the final T1D status.
3. **All Observations Approach**: Treating all observations as independent, excluding "base age" and "sex" variables. The variable “base age” was excluded due to it being a static measure that acted as a unique identifier for patients.

Model Performance and Selection

After evaluating various RSF models, we selected the "RSF First Biomarkers and Last T1D status" approach for all dataset sizes (small, medium, large) due to its superior performance in Area Under the Curve (AUC) and competitive Mean Absolute Error (MAE-PO) scores. This model demonstrated the highest LM scores among the candidates while maintaining a relatively low MAE, indicating a good balance between fit and prediction accuracy. For the small and medium datasets the "RSF First Biomarkers and Last T1D status" was clearly the best for balancing AUC and MAE-PO performance, but for the large dataset, the RSF all observations gave slightly better performance for MAE-PO, the feature importance results showed results that did not make sense to the domain experts, so this model was not considered further.

Feature Importance

Feature importance is assessed via permutation importance, providing insights into each predictor's relative contribution. Model evaluation uses the out-of-bag (OOB) error, computed as: $OOB\ Error = 1 - C - index$. The RSF model revealed key predictors for T1D progression. Here are the feature importance plots for the "RSF First Biomarkers and Last T1D status" large, medium, and small datasets:

---

[3]Shakeri, Z., Momtazian, N., Lim, S., Jaakkola, M., Croitoru, M., & Gandhi, M. (2024). Predicting time to diabetes diagnosis with random survival forests. *medRxiv*. https://doi.org/10.1101/2024.02.03.24302304v1

[4] Bradburn, M. J., Clark, T. G., Love, S. B., Altman, D. G. (2003). Survival analysis part I: basic concepts and first approaches. *British Journal of Cancer*, *99*(2), 170–175. https://doi.org/10.1038/sj.bjc.6603218

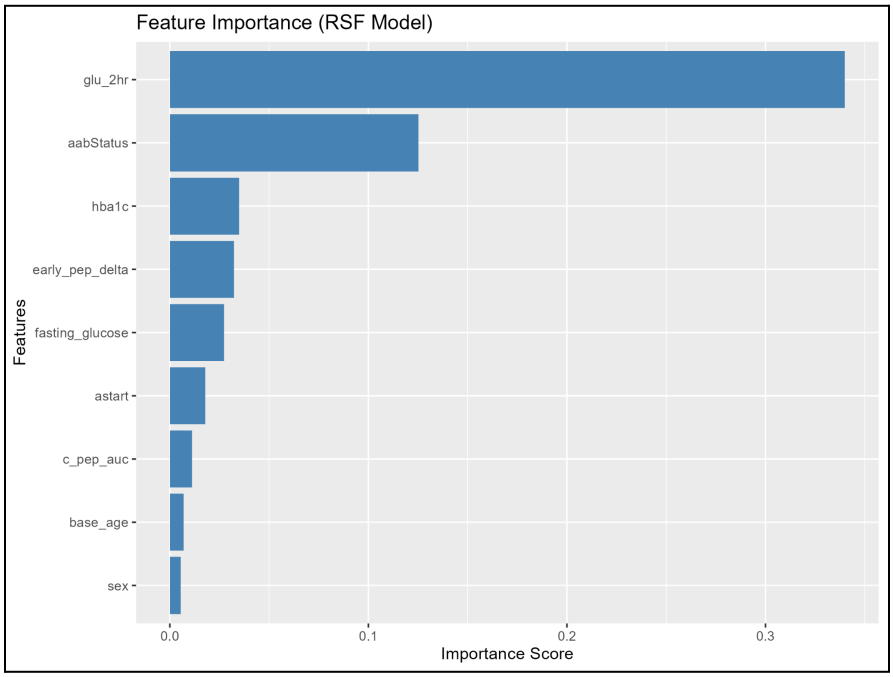


Figure 1: Feature Importance for RSF Large Model

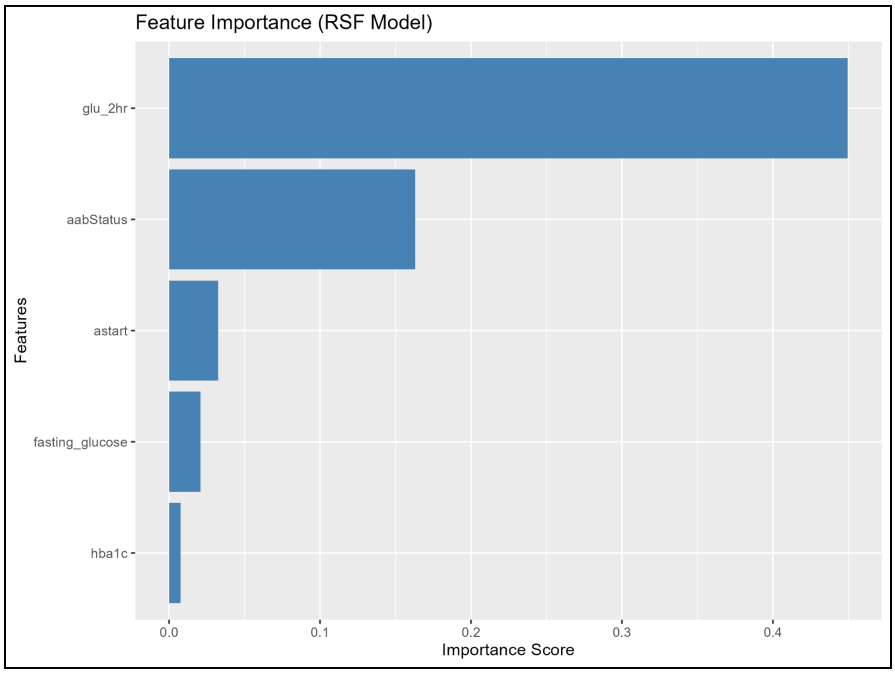


Figure 2: Feature Importance for RSF Medium Model

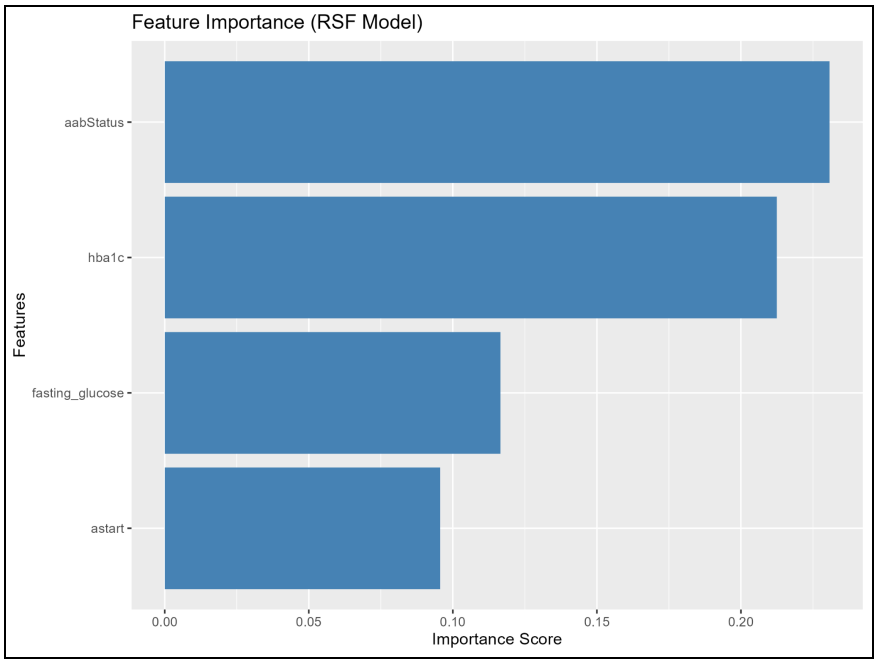


Figure 3: Feature Importance for RSF Small Model

In the large and medium models where the variable “glu_2hr” is available, it is the most important, and aabStatus is the second most important, and in the case of the small model, the most important.

Inference Considerations

During inference, the total time interval is set to 0 and t1d status must be 0. While these are the variables being predicted during inference, they must still be present in the prediction data. The values are not used when making inferences, but for standardization, all were set to 0.

In conclusion, the RSF methodology, particularly the "First Biomarkers and Last T1D status" approach, offers a robust framework for T1D risk prediction. Its non-parametric nature and fast training time allow for a very flexible model. Future work will focus on refining the model through hyperparameter tuning to optimize AUC and MAE-PO metrics, potentially enhancing its predictive performance and clinical utility.

DynForest

DynForest Modeling

The DynForest[5] method extends the Random Survival Forest framework to handle longitudinal data by integrating mixed model techniques, reducing the complexity of the time component and enabling more accurate survival analysis for studies with repeated measurements.

[5] Devaux, A., Helmer, C., Genuer, R., & Proust-Lima, C. (2023). DynForest: A random survival forest methodology to predict an event from longitudinal endogenous covariates. *arXiv*. Retrieved from https://arxiv.org/pdf/2208.05801

Model Overview

DynForest accommodates both fixed (non-time-dependent) and longitudinal (time-dependent repeated measures) features. Fixed features, such as "base age" and “sex” remain constant over time, while longitudinal features like "astart" (age at measurement), and all of the biomarkers coming from lab tests change over time. The model requires explicit specification of these features along with a unique identifier for each patient.

Time Metric and Survival Prediction

The model uses "tstart" as the time metric, representing the elapsed time between measurements. In survival mode, DynForest returns the Cumulative Incidence Function (CIF) derived from predicted survival probabilities. Due to the nature of our prediction tool, at the point of inference the time point t0 is set to 0, meaning that no longitudinal data is used and that only the measurements at time zero are used. Due to this, the value t0 must be set to 0 when predicting. For further work, longitudinal data can be used and the time period used for prediction can be adjusted by updating the t0 parameter.

Model Evaluation and Parameter Tuning

DynForest employs the Integrated Brier Score (IBS) as the Out-of-Bag (OOB) error metric, providing a comprehensive measure of prediction accuracy over time. Experiments were conducted with mtry values of 5 and 3, where mtry represents the number of features selected during bootstrap sampling at each node. There was limited tuning done due to very expensive compute time for this model. In the original paper proposing DynForest, mtry was shown to be the hyperparameter that led to the most performance gains.

Initial Exploration and Methodological Considerations

Initial explorations involved fitting both linear and non-linear lines of best fit to assess trends in features over time and employing linear mixed effects models for all features to capture their longitudinal behavior. The linear lines were determined to be the best fit. For the aabstatus variable, a generalized linear mixed model (GLMM) was applied, and the extracted random effects were incorporated as fixed features to enhance model performance. This did not end up leading to performance improvements in the model, so a linear mixed effects model was fit to the binary variable, aabStatus, for ease.

Feature Importance and Performance

The DynForest models with mtry = 5 and mtry = 3 resulted in very similar values for AUC and MAE-PO. Ultimately, DynForest with mtry = 5 demonstrated slightly better MAE-PO performance and was selected as the best DynForest model. This configuration likely strikes an optimal balance between feature diversity and model complexity, allowing for more robust predictions while avoiding overfitting. The model's ability to capture complex interactions between longitudinal features contributes to its effectiveness in predicting T1D progression.

Feature importance analysis revealed key predictors of T1D progression, which could provide valuable insights for clinical decision-making and patient monitoring strategies. The following is the feature importance plot for DynForest mtry = 5:

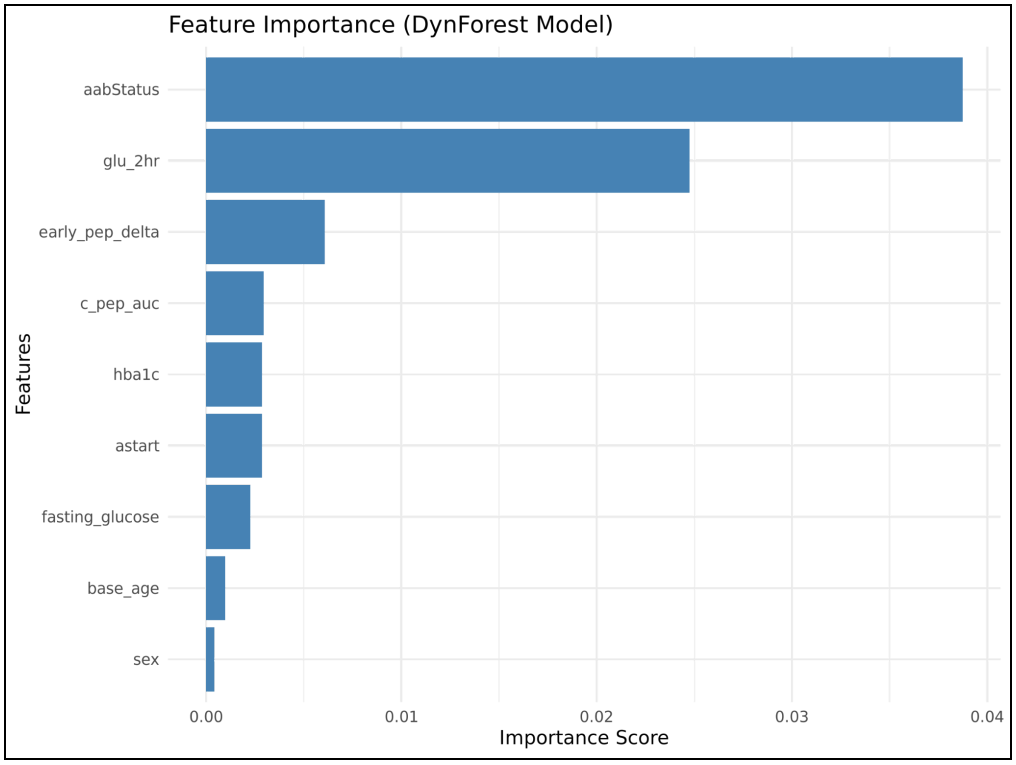


Figure 4: Feature Importance for Dynforest Model

Similarly to the RSF model, “aabStatus” and “glu_2hr” are the most important variables for predicting, however, here “aabStatus” is the most important.

In conclusion, DynForest's ability to handle longitudinal data effectively, coupled with its robust evaluation metrics and tuned hyperparameters, makes it a powerful tool for survival analysis in clinical studies with repeated measures. For future work, adding patients' historical data would enable the full benefits of the DynForest model to be used. Additionally, an experiment to see how prediction accuracy improves as more longitudinal data is used would allow for a more in-depth understanding of the performance of DynForest.

Joint Modeling (JM)

Joint modelling combines the power of traditional survival analysis with ML approaches like linear mixed effects modelling to account for the effect of longitudinal biomarker measurements on the disease progression outcome. It is a highly modifiable and flexible architecture, and several considerations and experiments were conducted to arrive at the most suitable combinations of parameters, features and association structures.

Feature Selection

Within the longitudinal sub-model, it is important to decide what biomarker (feature) needs to be modelled using the rest of the biomarker measurements.

- For the small model, both fasting_glucose and hba1c were modelled separately using the remaining biomarkers as predictors and the fit was evaluated. The best performance was achieved through estimating hba1c using age, aab_status and fasting_glucose as the predictor variables.
- For the medium model, glu_2hr and hba1c were modelled separately using the remaining biomarkers as predictors and the fit was evaluated. The best performance was achieved through estimating hba1c using age, aab_status, fasting_glucose and glu_2hr as the predictor variables.
- For the large model, c_pep_auc, early_pep_delta and hba1c were modelled separately using the remaining biomarkers as predictors and the fit was evaluated. The best performance was achieved through estimating c_pep_auc using age, aab_status, hba1c, fasting_glucose, glu_2hr and early_pep_delta as the predictor variables.

Survival Function Selection

For the survival sub-model, two different survival models were considered:

1. Cox Proportional Hazards Survival (CoxPH): A semi-parametric approach which estimates the effects of covariates on the hazard (risk) of an event occurring, without requiring specification of the baseline hazard function. It assumes that the hazard ratios stay constant over time.
2. Weibull Survival Model: Unlike the semi-parametric Cox proportional hazards model, the Weibull model fully specifies the baseline hazard function using a two-parameter distribution (shape and scale).

After evaluating both approaches, the weibull survival model yielded better results, primarily due it's ability to model increasing hazard rates (shape parameter > 1), constant hazard rates (shape parameter = 1, equivalent to the exponential distribution), and decreasing hazard rates (shape parameter < 1). It is also the survival model chosen for the BRI research project and can be modified to align with their estimated shape and scale parameters as well. The weibull model was thus used as a survival sub-model for all JMs.

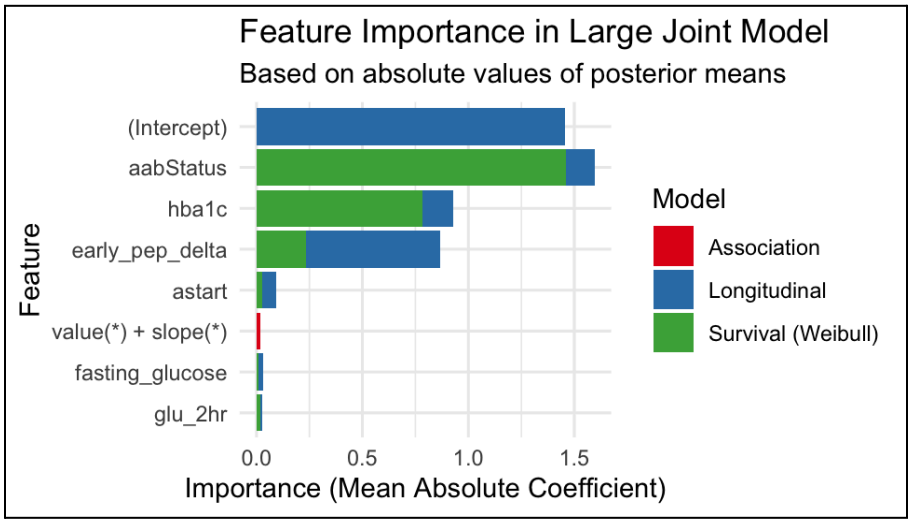
Association Structure

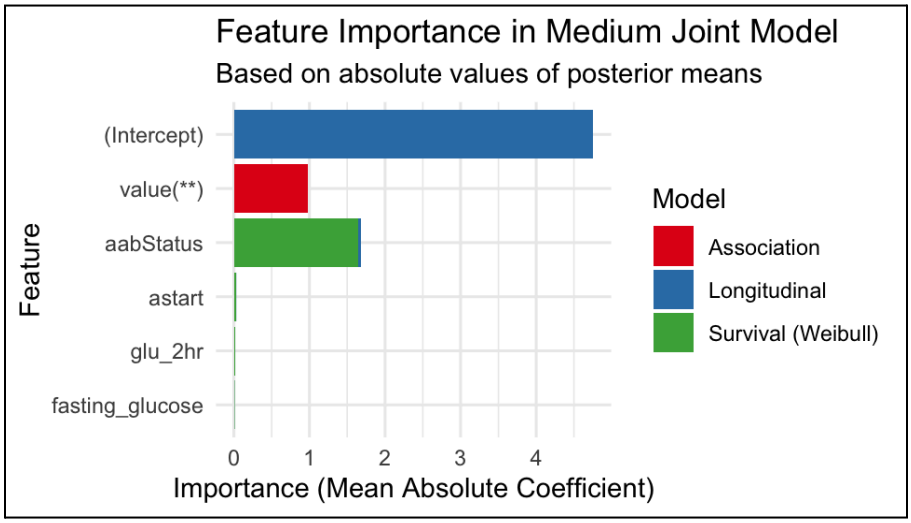
The association structure determines how information from one component (e.g., a biomarker's trajectory) influences the other component (e.g., time-to-event). This linkage is typically established through shared random effects or by directly including features of the longitudinal process in the survival model. These different association structures allow researchers to test various hypotheses about how biomarkers relate to disease progression. For example, in some diseases, the rate of change in a biomarker might be more predictive than its absolute value, while in others, crossing a specific threshold might be what matters most. Several association structures were tested for each implementation and the best ones for each set of variables were chosen.

	Feature chosen	Predictors	Association structure
Small	Hba1c	AAB_status, fasting_glucose, age	Slope(hba1c) + Value(hba1c)
Medium	Hba1c	AAB_status, fasting_glucose, age, glu_2hr	Value(hba1c)
Large	C_pep_auc	AAB_status, Hba1c, fasting_glucose, age, glu_2hr, early_pep_delta	Slope(c_pep_auc) + Value(c_pep_auc)

Table 2. Final JM architecture

Feature Importance





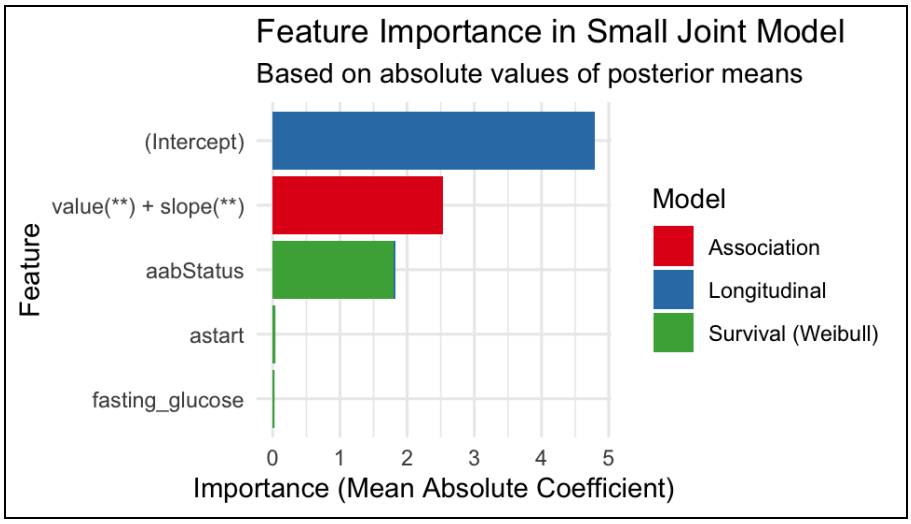


Figure 5. Feature importance graphs for each joint model (* = c_pep_auc and ** = hba1c)

The feature importance graphs for each joint model show that AAB status and Hba1c are the strongest predictors of T1D onset across all models. For the large model, the addition of early_pep_delta contributes significantly to its predictive power, and therefore its overall performance as well.

Model Evaluation

Evaluating survival models presents unique challenges due to the nature of time-to-event data and the presence of censoring. Unlike traditional prediction tasks, survival data often includes incomplete information, as some subjects may not experience the event of interest during the study period or they may drop out of the study. This censoring makes it difficult to directly compare predicted outcomes with actual events for all individuals. Additionally, there is no ground truth to evaluate the survival model on. Furthermore, the performance of survival models can vary across different time points. We tackled this challenge using two primary approaches: Landmarked AUC and MAE-PO, each providing distinct insights into model performance.

Landmarked AUC

We employed landmarked Area Under the Curve (AUC) to evaluate the discriminative ability of our survival models at specific time points. Landmarking, a dynamic prediction approach, allows us to assess model performance at different stages of disease progression, offering a more relevant evaluation of model discrimination. Our analysis used landmark time-dependent Receiver Operating Characteristic (ROC) curves to evaluate the ability of the models to distinguish between patients who will and will not develop T1D within specified time horizons. The landmarking procedure involved going “landmark” time points back from the last observation of each patient and generating predictions for that “landmarked” interval, using past data. This was then compared to the actual T1D status at the time point and AUC was computed. The analysis was performed on multiple validation datasets, including holdout validation data, holdout test data, DPT-1, and FRIDA data. Note that Frida data for the large model did not have data for landmark year 3.

AUC values range from 0.5 (no better than chance) to 1.0 (perfect discrimination). We report both landmark-specific AUCs and mean AUC across landmarks to provide a comprehensive assessment of model performance. This approach allows us to understand how prediction accuracy may change over the course of disease progression and identify optimal time points for risk stratification in clinical practice.

MAE-PO

To address the challenge of evaluating survival models with censored data, we employed the Mean Absolute Error - Pseudo-Observation (MAE-PO) metric which measures how closely predicted T1D onset aligns with actual outcomes even with censored data. This innovative approach, detailed in Qi et al. (2023), provides a more robust and accurate evaluation method, particularly for datasets with high censoring rates[6].

Pseudo-observation Calculation
For each subject i, a pseudo-observation is computed using the formula:

$$e_{\text{pseudo-obs}}(t_i, \mathcal{D}) = N \times \hat{\theta} - (N-1) \times \hat{\theta}^{-i}$$

where N is the total number of subjects, $\hat{\theta}$ is an unbiased estimator for the event time (e.g., mean survival time from Kaplan-Meier estimator), and $\hat{\theta}^{-i}$ is the estimator applied to the dataset without the i-th instance.

To ensure consistency across models, we applied the following considerations:
1. The training dataset used for MAE-PO evaluation matched the specific preprocessed data used for each model's training.
2. For each patient, we used the first observation and predicted the survival curve for 9 years.
3. While RSF and DynForest models do not have horizon times, JM and Benaroya models allow for specified prediction times. To standardize the analysis, we truncated all predicted curves at 9 years, which is the minimum prediction horizon for RSF and DynForest.

By utilizing MAE-PO, we aim to provide a more comprehensive and reliable evaluation of our survival models, particularly in scenarios with high censoring rates. This approach allows for a fairer comparison between different modeling techniques and helps identify the most accurate predictors of T1D progression.

---

[6] Qi et al. (2023). An Effective Meaningful Way to Evaluate Survival Models. *Proceedings of the 40th International Conference on Machine Learning* (PMLR 202), 28295–28318. https://proceedings.mlr.press/v202/qi23b/qi23b.pdf

Results

Benaroya Research Institute Results - Validation

Using the BRI model verification script, validation was performed on all datasets to get the AUC results of the Weibull model. The results are outlined in Table 3.

The BRI AUC values for FR1DA & DPT-1 indicate that the small Benaroya model doesn't generalize well to new patients. The low AUC for FRIDA indicates that the small model performs no better than random guessing. It is important to note that FRIDA had a smaller age distribution than the other two datasets. The BRI large model, however, performs well on external validation data. This indicates that when 2-hour glucose, early peptide delta, and c-peptide AUC are included, the large model is able to determine T1D status more accurately on unseen data than the other BRI models.

Model	Landmark T =1			Landmark T = 3		
	TrialNet	DPT-1	FRIDA	TrialNet	DPT-1	FRIDA
Small	0.73	0.66	0.56	0.71	0.62	0.59
Med.	0.73	0.71	0.77	0.76	0.66	0.60
Large	0.82	0.80	0.76	0.83	0.77	No data

Table 3. BRI AUC Results Per Dataset

The violin plots in Figure 6. illustrate the predicted probability distributions of Type 1 Diabetes (T1D) progression for patients who did and did not develop T1D within the duration of each clinical trial (TrialNet, DPT-1, and FR1DA). The goal of these plots is to evaluate how well the model distinguishes between individuals who ultimately developed T1D (T1D group) and those who did not (No group).

Ideally, for the "No" group, we would expect a higher density of predicted probabilities near zero, indicating that the model correctly assigns low risk to those who did not progress to T1D. Conversely, for the "T1D" group, we would prefer a higher density of predicted probabilities above 0.5, suggesting the model is accurately identifying individuals at higher risk of progression. However, the observed distributions reveal some variability. In the "No" group across all datasets and model sizes (Small vs. Large), there is generally a concentration of predicted probabilities near zero, though some instances exhibit a broader spread, indicating occasional overestimation of risk. In contrast, the "T1D" group does not consistently show a strong density of predicted probabilities above 0.5, suggesting that while the model captures some signal, it may not be optimally calibrated to separate high-risk individuals effectively.

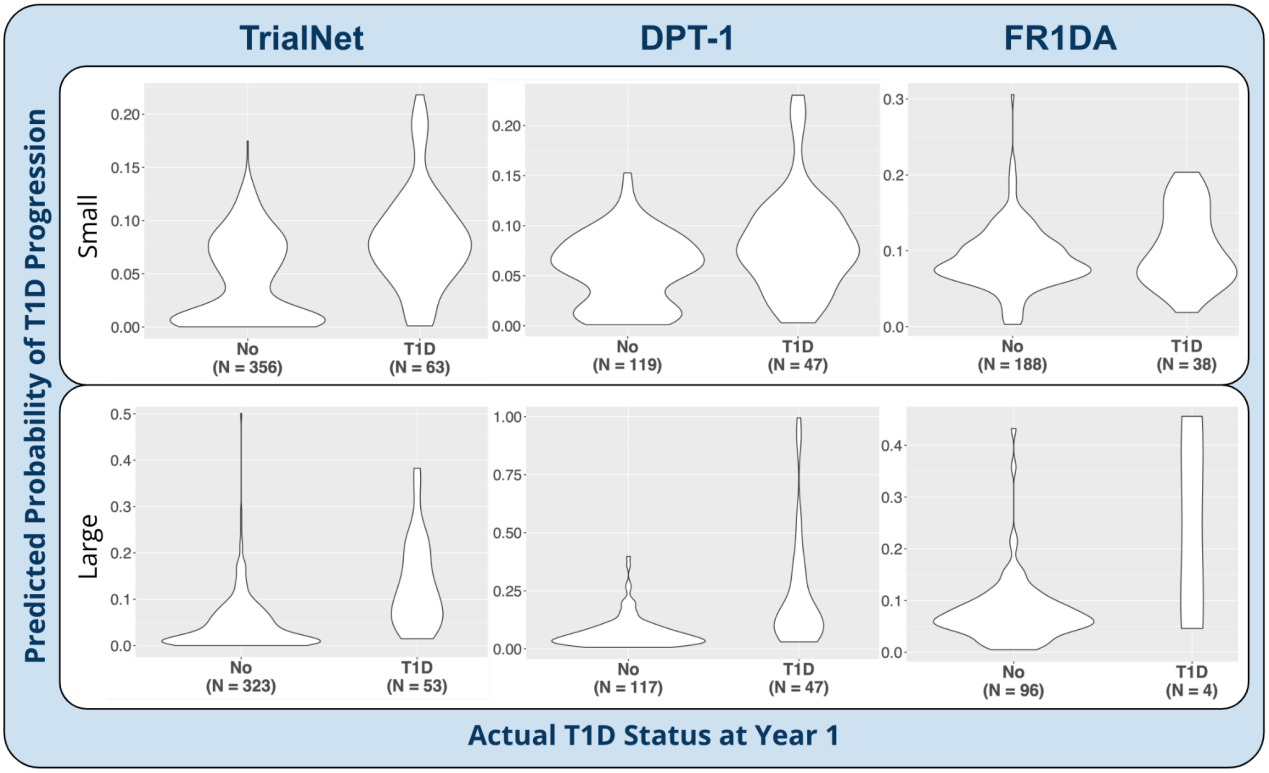


Figure 6: Violin plots for the small and large BRI model

Notably, the larger model variant tends to produce a slightly wider range of predicted probabilities for both groups, with some cases of individuals who developed T1D reaching probabilities closer to 1. However, the spread remains substantial, particularly in FR1DA, where the "T1D" group's predictions are more varied. The small sample size in certain groups, such as the T1D group in FR1DA, may also contribute to inconsistencies in the probability distributions. Overall, while the model shows a general ability to distinguish between the two groups, improvements in calibration and decision thresholds may be needed to enhance its predictive reliability for clinical application. More violin plots for the medium sized model, as well as additional landmark times (LM=3) are available in the Github repo under the folder benaroya_model_verification/plots.

Machine Learning Approaches - Results

Evaluation was performed for all ML models using AUC and MAE-PO. Heatmaps were produced to help illustrate model performance. The heatmaps illustrate the AUC performance of various models across different datasets (TrialNet Validation, TrialNet Test, FRIDA, and DPT-1) and two landmark time points (T=1 and T=3).

Area Under the Curve

AUC helps to assess the discriminative ability of a model, i.e., how well each model distinguishes between individuals who will and will not develop Type 1 Diabetes (T1D).

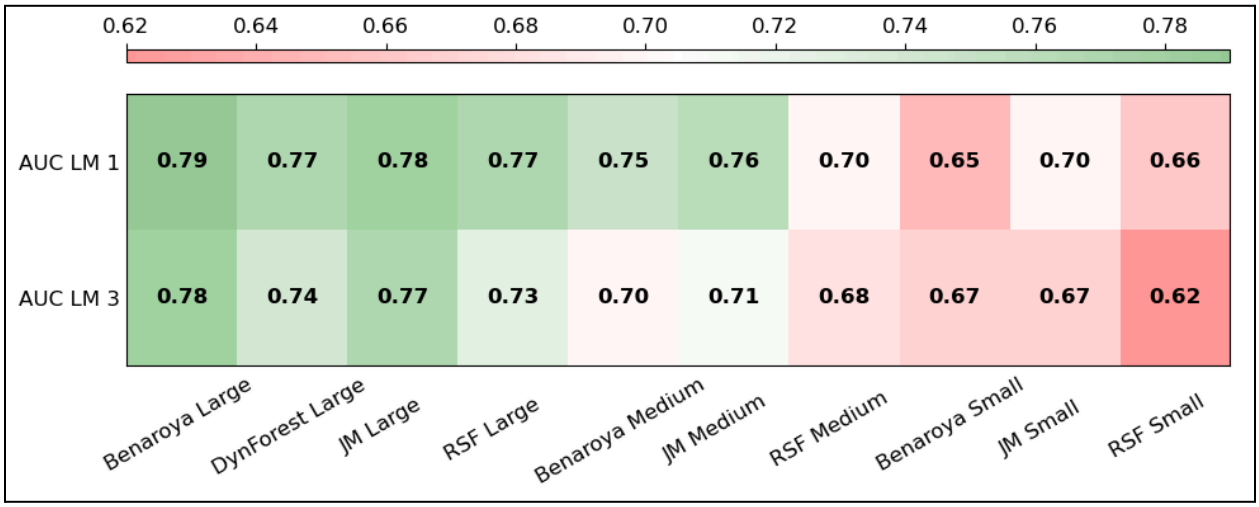


Figure 6. Heatmap of average AUC across all datasets per landmark time

		Landmark Time = 1				Landmark Time = 3			
		TrialNet (Val)	TrialNet (Test)	FR1DA	DPT - 1	TrialNet (Val)	TrialNet (Test)	FR1DA	DPT - 1
Small Variables Used	BRI	0.727	0.678	0.596	0.659	0.713	0.743	0.592	0.616
	RSF	0.735	0.697	0.675	0.628	0.716	0.783	0.410	0.655
	JM	0.724	0.722	0.699	0.656	0.719	0.767	0.519	0.694
Medium Variables Used	BRI	0.726	0.776	0.767	0.712	0.760	0.761	0.601	0.660
	RSF	0.725	0.792	0.736	0.732	0.770	0.758	0.459	0.732
	JM	0.748	0.783	0.781	0.738	0.751	0.790	0.619	0.683
Large Variables Used	BRI	0.823	0.777	0.758	0.797	0.834	0.732	No data	0.772
	RSF	0.794	0.732	0.776	0.774	0.772	0.653	No data	0.695
	JM	0.820	0.774	0.755	0.760	0.824	0.713	No data	0.761
	DynForest	0.790	0.748	0.789	0.746	0.815	0.715	No data	0.716

Table 4. AUC results for all datasets for each model - Best AUCs highlighted

Across all datasets and LM times, the **Original BRI model (L)** and **Joint Model (L)** seem to be the best-performing models. The **worst-performing model** is **RSF Small**, with an **average AUC of 0.64**, with **Benaroya (S)** close behind at **0.66**.

Key Takeaways

- **Larger models (L) generally outperform smaller models (S),** indicating that the added features are able to capture more predictive signals and variability.
- **TrialNet and DPT-1 datasets show a general trend where performance improves from T=1 to T=3,** implying that longer observation periods may enhance prediction accuracy.

When broken down per dataset, the results suggest that **Original BRI model (L) and Joint Model (L) are the most reliable models**, while **RSF (S) struggles across all datasets**, particularly in FRIDA. JM Large is a **strong competitor** across datasets but **does not dominate every dataset**—Dynforest (L) and Original BRI model (L) often achieve the highest AUCs. JM Large remains a **reliable and well-performing model**, particularly for generalization across datasets. These findings highlight the importance of model selection based on dataset characteristics and time-dependent factors.

Best Performing Models

1. **Benaroya Large (0.79, 0.78)**
    - Consistently achieves the highest AUC across both landmark times.
    - Strongest model overall, suggesting that larger, more complex models trained on Benaroya's approach generalize well across datasets.

2. **JM Large (0.78, 0.77)**

    - Very close in performance to Benaroya Large.
    - Maintains strong predictive power across datasets, making it a robust choice.

3. **DynForest Large (0.77, 0.74)**
    - Performs well but slightly weaker at **LM 3**, suggesting it may lose predictive power over longer time frames.

Worst Performing Models

1. **RSF Small (0.66, 0.62)**
    - The lowest-performing model, particularly weak at **LM 3**.
    - Indicates that random survival forests (RSF) struggle compared to other approaches.

2. **Benaroya Small (0.65, 0.67)** & **JM Small (0.70, 0.67)**
    - Both models show lower predictive ability than their larger counterparts.
    - Suggests that model size matters significantly in improving prediction accuracy.

When compared to the earlier results broken down by dataset, the overall trends remain consistent. In the individual dataset breakdowns, **Benaroya Large and JM Large** frequently ranked among the best models, which is reaffirmed by the averaged results. The earlier dataset-specific analysis also highlighted the **poor performance of RSF Small, particularly in FRIDA and DPT-1**, a pattern that persists in the aggregated view. However, one key difference is that in the individual dataset breakdowns, certain models performed exceptionally well in specific datasets but not as well in others. For example, **RSF Medium**

had high AUC in TrialNet Test but underperformed in FRIDA, which becomes less apparent in the averaged results. By averaging across datasets, these dataset-specific variations are smoothed out, making the **Benaroya Large and JM Large models emerge as the most robust choices across all settings**.

Ultimately, these results confirm that **larger, more complex models consistently outperform smaller ones, with RSF-based approaches being the weakest overall.** The averaged AUC values reinforce the earlier dataset-level observations while providing a clearer picture of which models are the most reliable across different clinical trials.

Mean Absolute Error

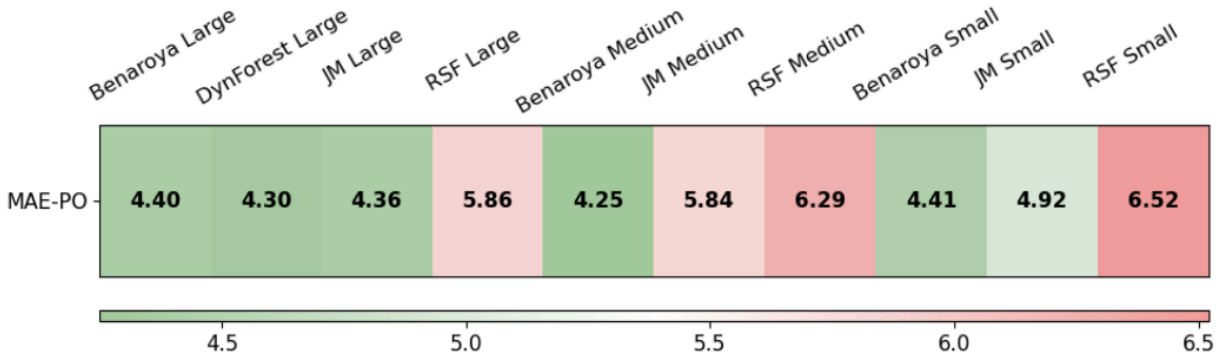


Figure 7. Heatmap of MAE-PO across all datasets and landmark time

Best Performing Model (Lowest MAE-PO)

- **JM Medium** had the lowest MAE-PO at **4.25**, indicating the best performance in terms of prediction accuracy.

Worst Performing Model (Highest MAE-PO)

- **RSF Small** had the highest MAE-PO at **6.52**, making it the worst-performing model in this evaluation.

Benaroya Medium (4.25), DynForest Large (4.30), and JM Large (4.36) also had relatively low MAE-PO values, suggesting they performed well. **RSF Large (5.86), RSF Medium (6.29), and RSF Small (6.52)** had the highest errors, indicating that RSF models struggled the most in this evaluation.

Web Application Overview & Enhancements

Previously, the predictive models for Type 1 Diabetes (T1D) progression were hosted as separate web applications using R Shiny, with each model accessible via its own dedicated webpage. While functional, this structure led to maintenance challenges and inefficiencies. To improve sustainability, usability, and performance, the application has been migrated to Dash in Python, consolidating all models into a single,

unified webpage. The models are now accessed dynamically via a centralized API endpoint, which is built using Plumber.R to ensure compatibility with the existing R-based models.

A key usability improvement involves replacing the age and prediction year slider with a dropdown menu, addressing previous concerns about the slider's intuitiveness. The dropdown selection simplifies user input, ensuring more precise control over the model parameters. For RSF, and DynForest, the forecasted years are fixed at 9.24846 years and 13.97947 years, respectively.

Model Selection: Small, Medium, and Large

The application determines which model variant (Small, Medium, or Large) to use based on the clinical variables provided by the user. The selection process follows a hierarchical approach:

- **Small Model**: Selected when only the **HbA1c level, fasting glucose, autoantibody status, and age** are provided.
- **Medium Model**: Chosen if, in addition to the small model variables, the **2-hour glucose level** is available.
- **Large Model**: Used when all previous variables are provided along with **C-Peptide AUC and Early C-Peptide Change**, offering the most comprehensive prediction.

For **RSF and Dynforest models**, **sex** is also required as an input parameter. If a user selects a model type that requires missing clinical variables, an error message is displayed, prompting them to enter the required data.

How the Web Application Works

The **Dash** application provides a structured user interface for inputting patient details and clinical variables. Once submitted, the selected model's API is called based on the input data. The API returns a **cumulative incidence plot** representing the predicted probability of T1D progression over time, which is dynamically displayed in the application.

Additionally, users can **download a report in PDF format**, containing patient details, selected model parameters, and the generated plot. This feature is enabled only after a patient ID is provided and a prediction has been generated.

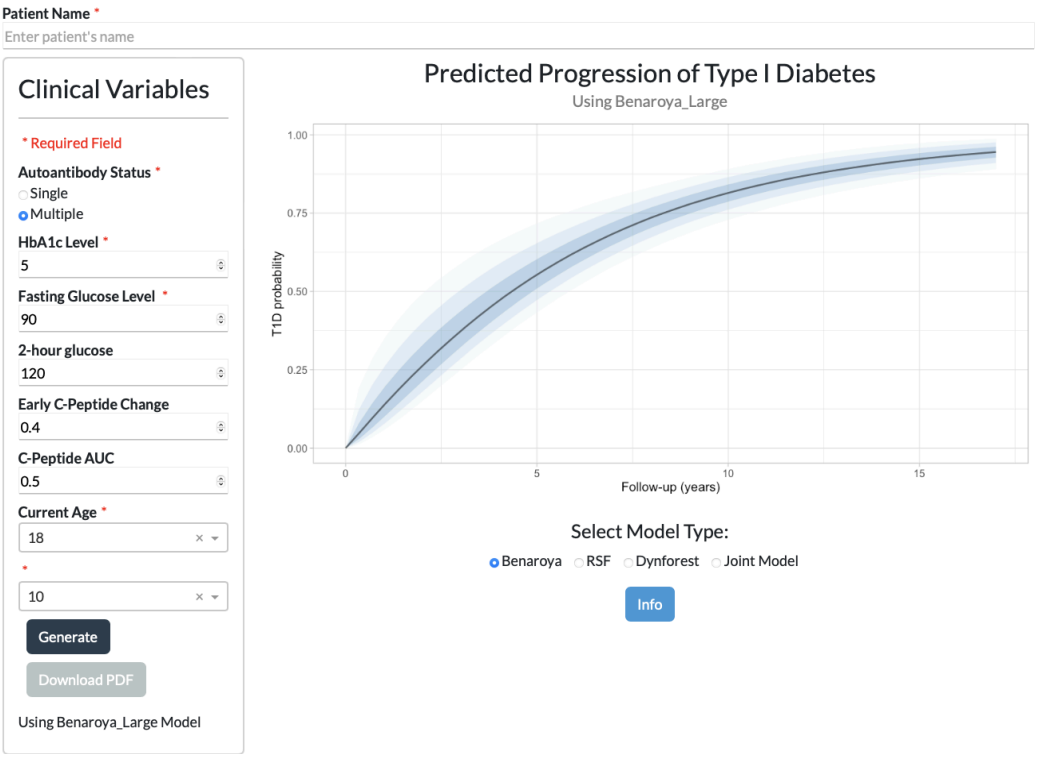


Figure 8. Screenshot of the DashApp using Benaroya's large model

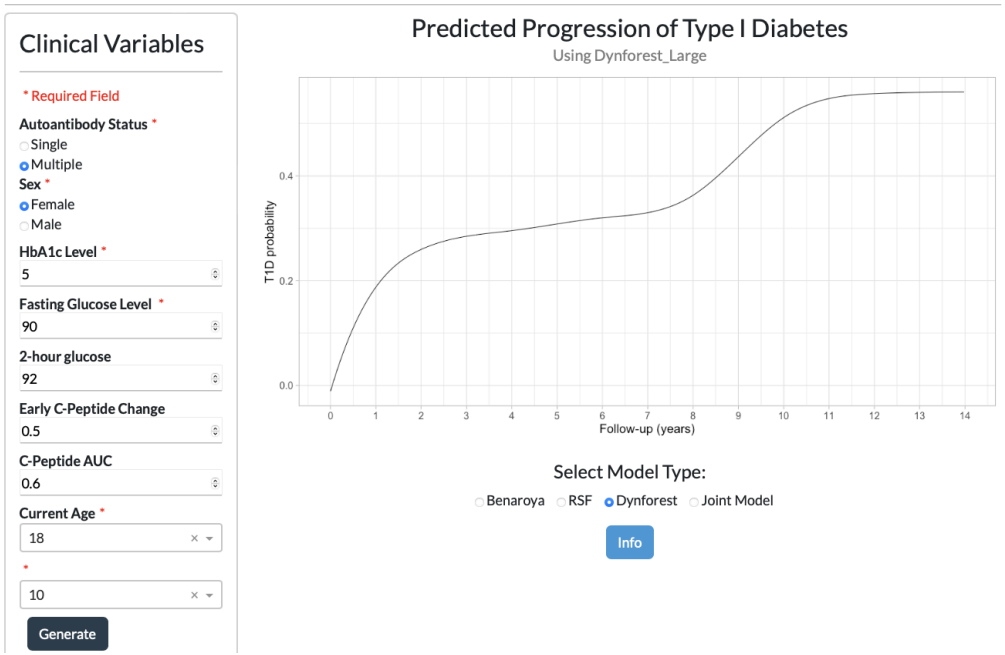


Figure 9. Screenshot of the Dash app, using the Dynforest model

Conclusion

Model Size	Small	Medium	Large
Best AUC	Joint Model, 0.685	Joint Model, 0.735	Benaroya, 0.785
Best MAE	Benaroya, 4.41	Benaroya, 4.25	DynForest, 4.30

Table 5. Best models per variable size

- The best overall models, balancing AUC and MAE-PO, are Benaroya Large and Joint Model Large. Among ML models, joint modeling is the most competitive alternative, offering strong AUC and MAE-PO, making it a viable ML alternative for improved accuracy and explainability.
- An added advantage of the joint model is the ability to dynamically update predictions as historical patient data becomes available through electronic health records.
- The feature importance graphs reiterate AAB Status is the strongest predictor of T1D development, across different models.

Future Work

While significant progress has been made in improving both the predictive models and the web application, several areas remain for future exploration and enhancement.

One key area for further analysis is assessing model performance across different age groups. Given that T1D is typically diagnosed at a younger age, our dataset is naturally skewed toward younger individuals. Creating age distribution plots could help determine whether certain models perform better for specific age ranges and whether adjustments need to be made to improve predictions for older individuals.

Additionally, further exploratory analysis can be conducted by examining which parameters are driving predictions. A useful approach would be to create histograms for individuals with predicted probabilities above and below 0.4, allowing us to analyze the key factors influencing the model's decision-making process. This could provide insights into which biomarkers or clinical variables are most strongly associated with T1D progression, potentially refining feature selection in future iterations.

From a usability perspective, enhancing the web application to recommend the most appropriate model size dynamically based on available variables would improve the user experience. Instead of relying on manual selection, the system could automatically suggest the best-performing model (Small, Medium, or Large) for a given set of inputs, ensuring clinicians and researchers use the most reliable model for their specific case.

Another crucial avenue for improvement is incorporating historical patient data into predictions. Currently, the models make predictions based on a snapshot from a single visit, but longitudinal patient data (multiple visits over time) could significantly enhance accuracy. Both DynForest and Joint Models have the capability to leverage this type of information. Future work should explore how to integrate

historical patient data dynamically, allowing for more personalized and data-driven risk assessments. An initial exploratory analysis of this approach has already been conducted in the project's GitHub repository, focusing on incorporating repeated measures into inference under ML_models/dynforest/longitudinal_experiment.R.

By addressing these future directions, we can further refine model performance, usability, and predictive accuracy, ensuring the models continue to evolve to better serve both researchers and clinicians in forecasting T1D progression.