Appendix A

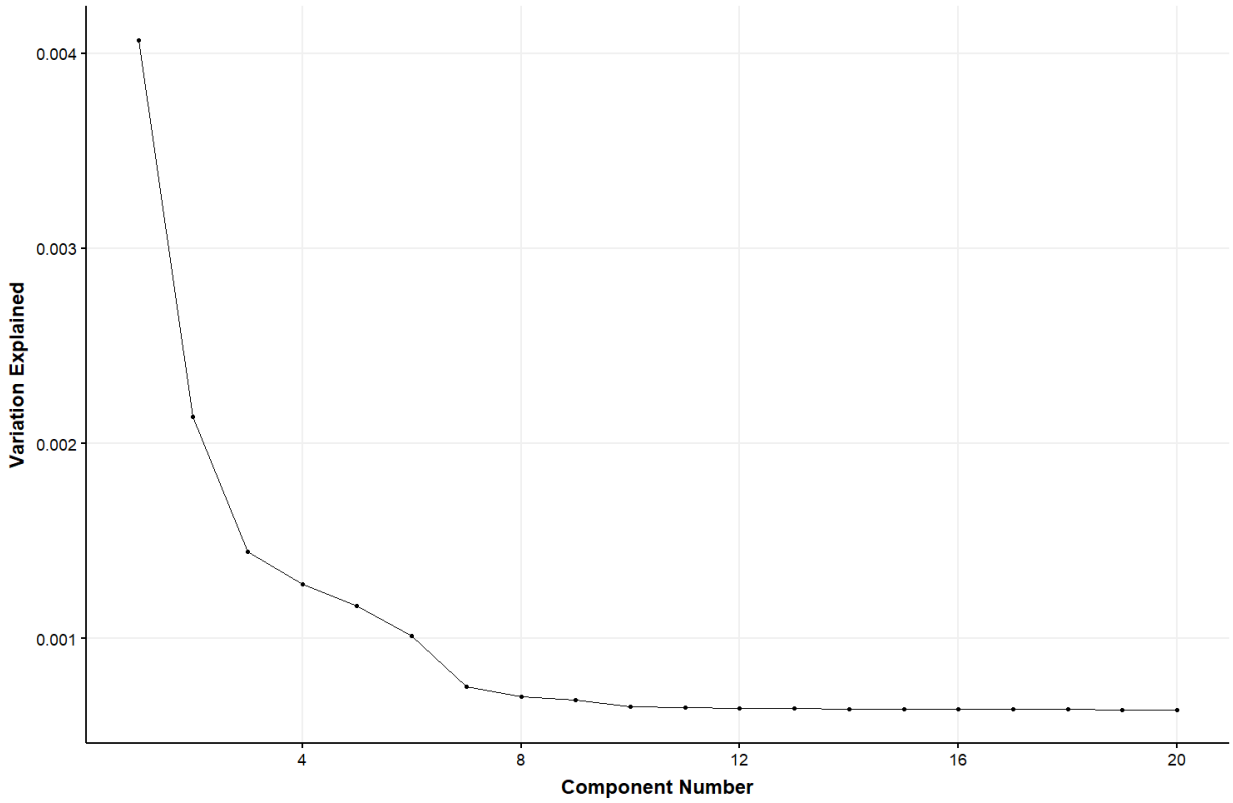
Constructing the *CFTR* Severity Score

Allele 2 Score	1	2	3	4
Allele 1 Score				
1	1			
2	1	2		
3	1	2	3	
4	1	2	4	5

Supplementary Table 1: The *CFTR* severity score is based on a combination of the severity scores from both alleles. For instance, an individual carrying *CFTR* variants G542X (allele score 4) and 3849+10kbC>T (allele score 1) is assigned 1 as his/her *CFTR* severity score.

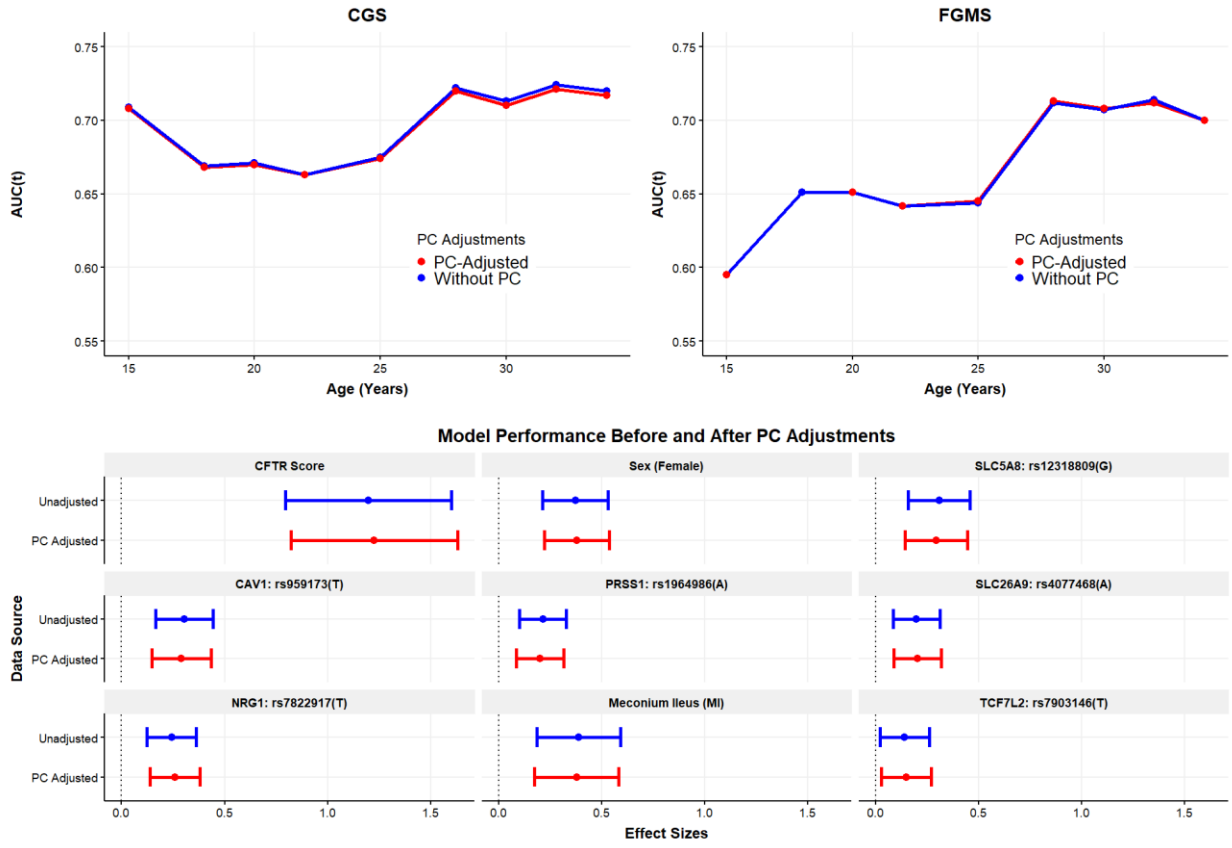
We first scored each *CFTR* variant by severity on a scale of 1 (least severe) to 4 (most severe)[17]. For example, G542X, a variant that results in no functional *CFTR* protein, is assigned a score of 4 and the variant 3849+10kbC>T, which results in insufficient but functional *CFTR* protein, is assigned a score of 1. Since a CF individual carries one *CFTR* variant on each allele, the *CFTR* severity score is based on a combination of the severity scores from both alleles.

Appendix B



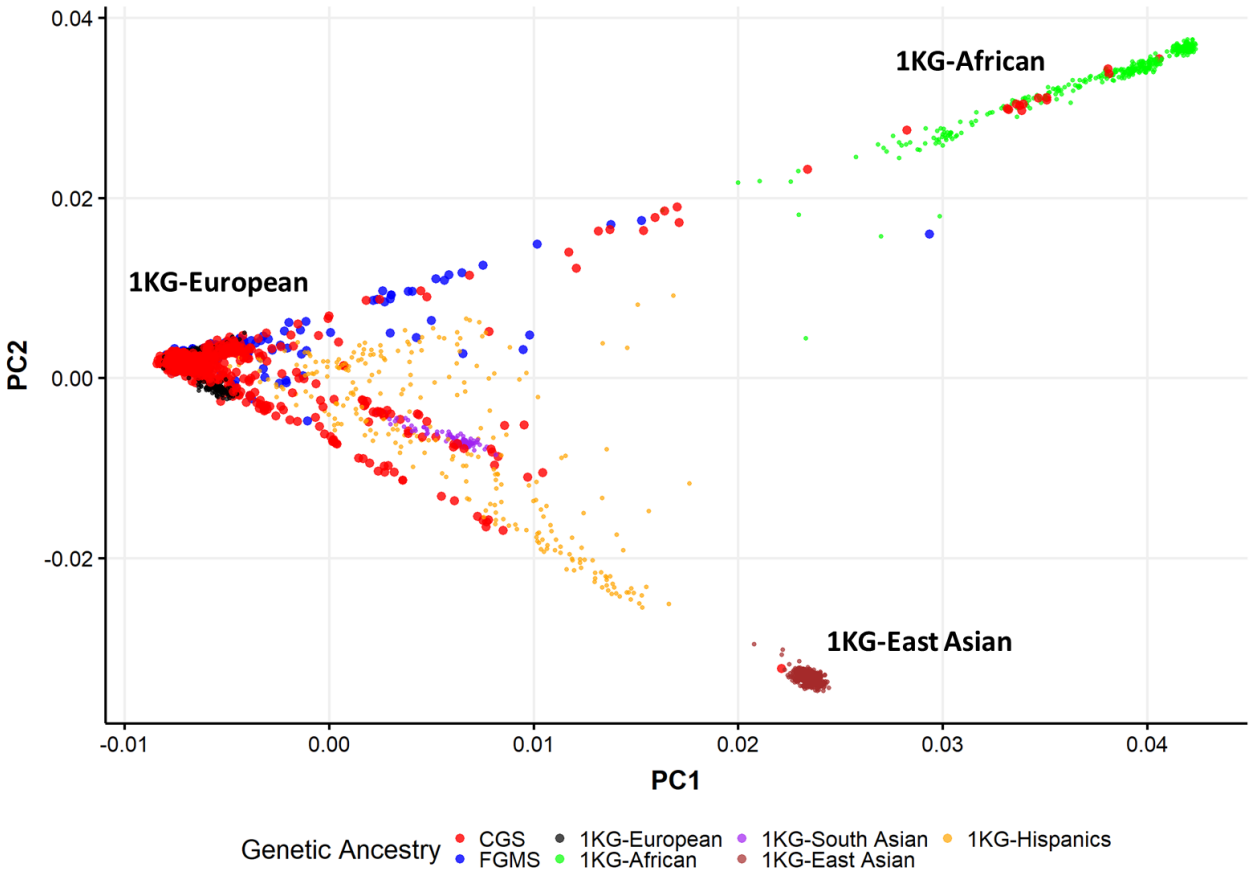
Supplementary Figure 1: Scree plot depicting variation explained by each principal component on the Canadian CF Gene Modifier (CGS) dataset. Ten principal components, deemed statistically significant by the Tracy-Widom test at the 0.01 significance level ($p<0.01$), were incorporated as predictors in feature (stability) selection and model fitting to account for population stratification.

Appendix C



Supplementary Figure 2: Model Performances in the CGS and FGMS studies and Univariate Log Hazard Ratios estimated in the CGS with and without PC adjustment (10 PCs). The lack of difference in model performance with and without adjusting for PCs suggests that there is likely no confounding due to population structure and that the CGS is largely ethnically homogeneous and comparable to the FGMS.

Appendix D

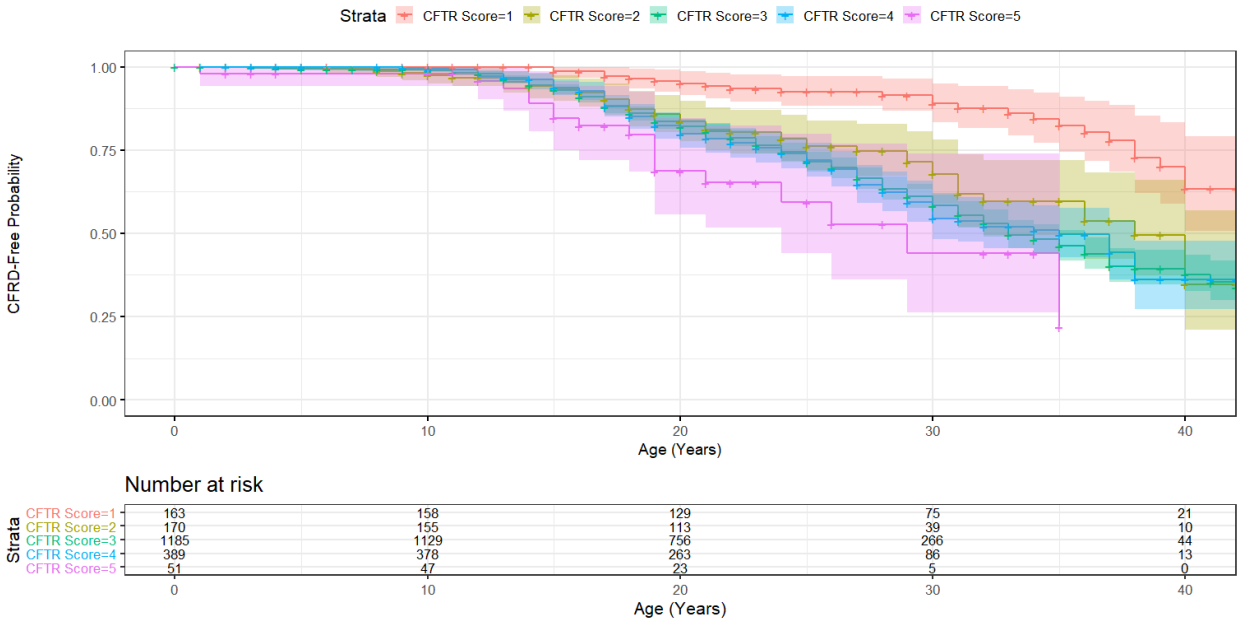


Supplementary Figure 3: The first two genetically determined principal components of CF individuals in the CGS (Red), FGMS (Blue) and reference populations from the 1000 Genomes (1KG). Both the CGS and FGMS are largely homogeneous, of European genetic ancestry and the composition is similar between the two study cohorts.

Ethnicity	Canadian GMS (n=1,958)	French GMS (n=1,003)
Europeans	1853 (94.6%)	978 (97.5%)
Africans	9 (0.5%)	0 (0.0%)
South Asians	11 (0.6%)	0 (0.0%)
Admixed/Hispanics/East Asians/Unknown	85 (4.3%)	25 (2.5%)

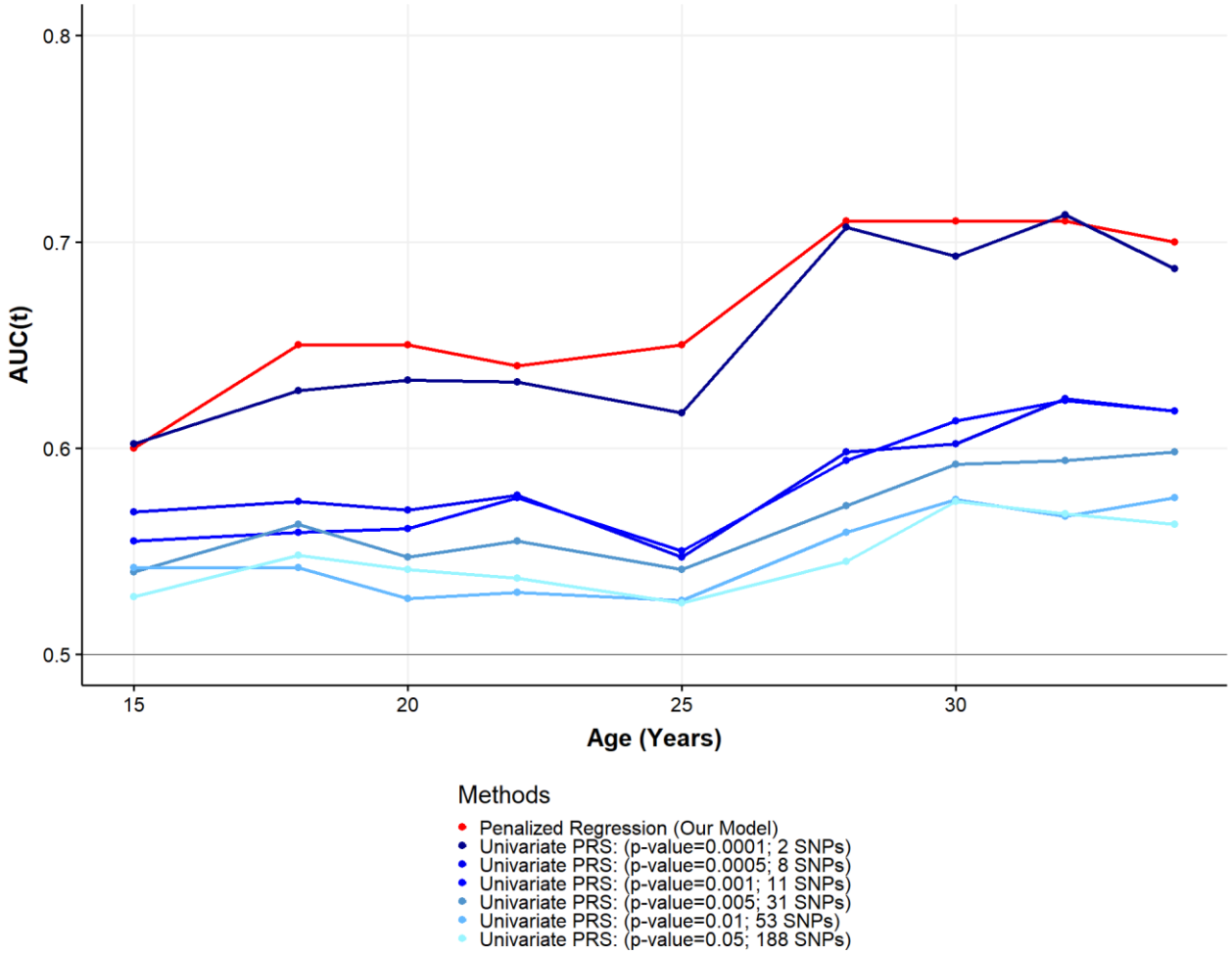
Supplementary Table 2: Inferred genetic ancestry for CF individuals in the CGS and FGMS studies. Both populations are predominantly European. Non-Europeans were defined as >3 S.D away from the center of the 1000 Genomes European cluster.

Appendix E



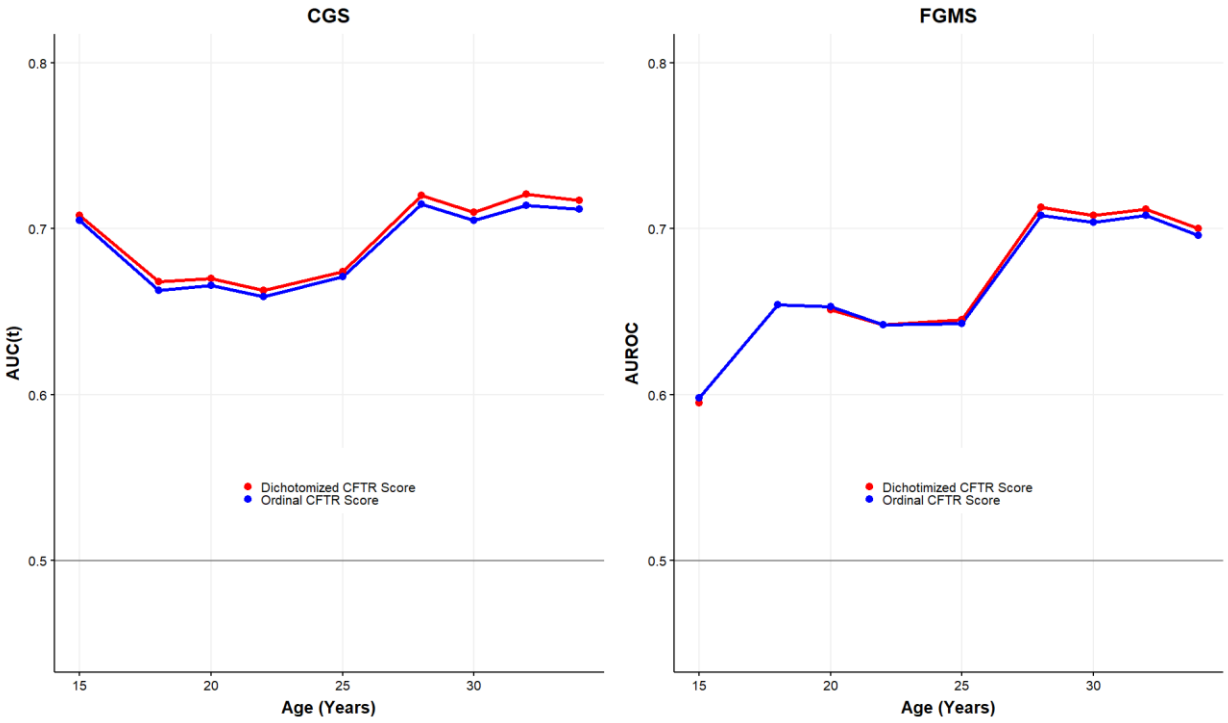
Supplementary Figure 4: CFRD-free probabilities and the 95% confidence intervals at different ages for Canadians with different *CFTR* severity scores. Individuals carrying the least severe *CFTR* variants (red) have the highest CFRD-free probabilities across all ages. In contrast, CFRD-free probabilities for those with *CFTR* severity scores either overlap extensively or cannot be reliably estimated due to the small sample size.

Appendix F



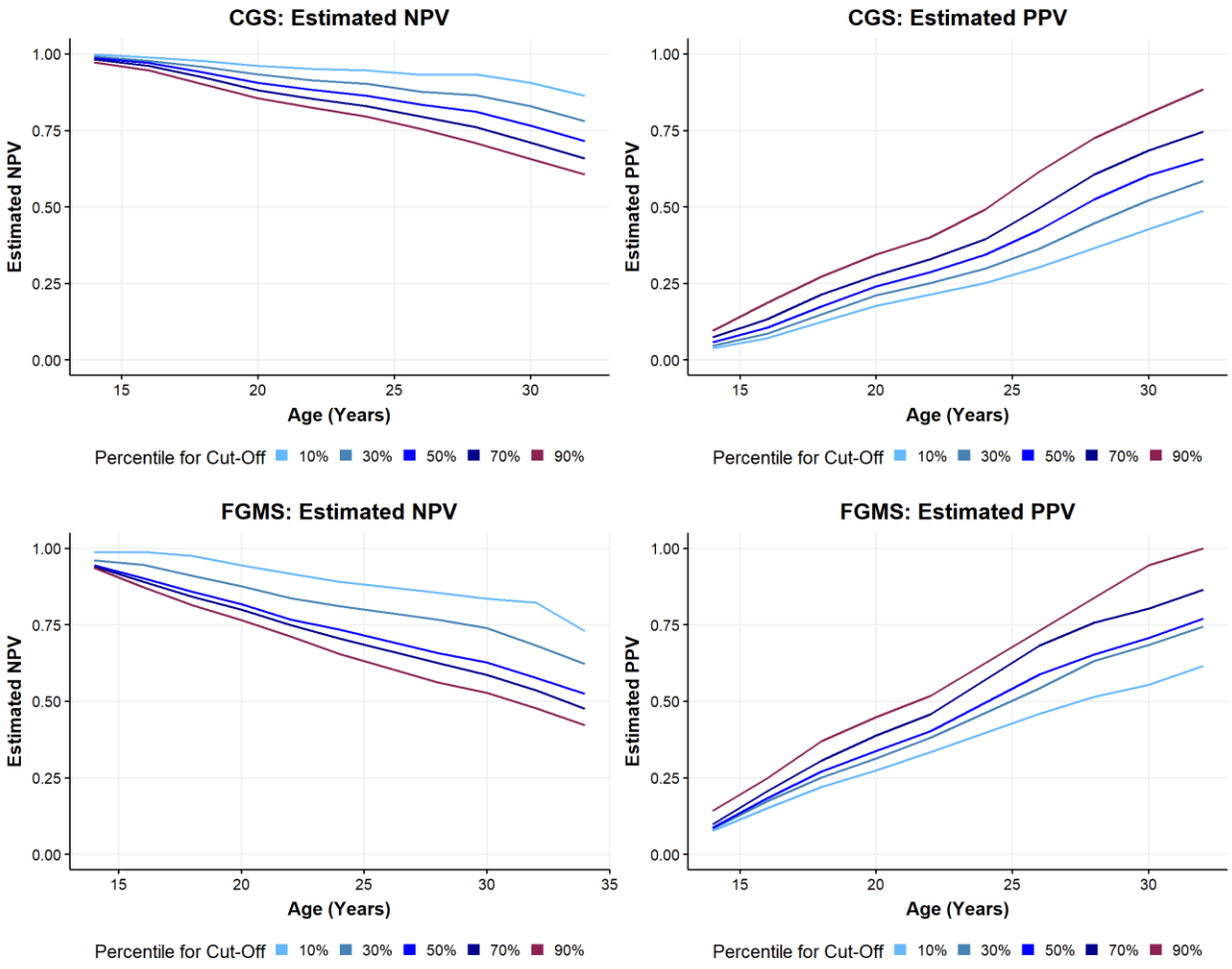
Supplementary Figure 5: Model Performance trained on the CGS and validated in the FGMS comparing a univariate, pruning and thresholding Polygenic Risk Score (PRS) approach to our penalized regression model. The *CFTR* severity score, 10 PCs, and the clinical variables including sex, MI and cohort were included in all models to ensure fair comparisons. Our penalized regression model resulted in the best performance across all univariate PRS models, regardless of the chosen p-value cut-off.

Appendix G



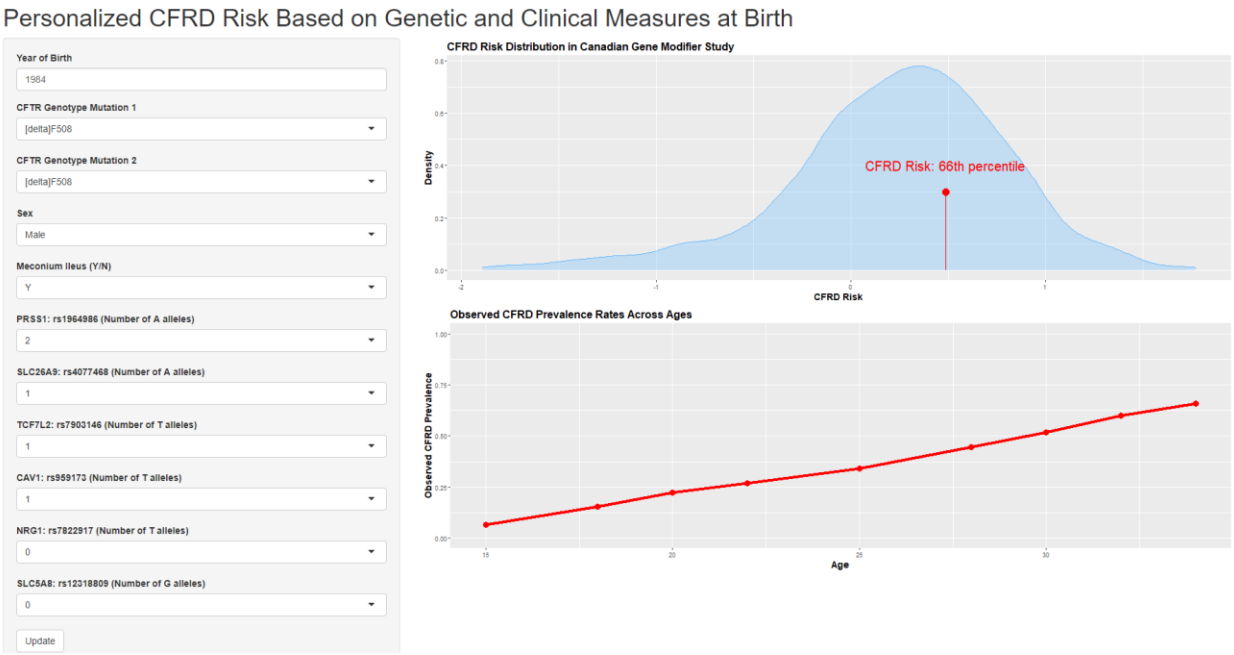
Supplementary Figure 6: Model Performance in the CGS and FGMS comparing the use of a dichotomized *CFTR* severity score versus an ordinal *CFTR* score. Using a dichotomized *CFTR* score can avoid excess uncertainty in the fitted model without loss of predictive accuracy.

Appendix H



Supplementary Figure 7: Estimated positive predictive values (PPV) and negative predictive values (NPV) across ages using different thresholds for defining a CFRD high-risk group. Using the 90% cut-off (purple; PPV) to indicate that only the top 10% individuals are in the high CFRD risk group, we expect CFRD prevalence rates to exceed 50% in this high-risk group in their early-to-mid 20s in both the CGS and FGMS. Using the 10% cut-off (skyblue; NPV plots), indicating that the model assigns the bottom 10% of individuals to be in the low CFRD risk group, we expect >80% in this low-risk group to be free of CFRD into their early-30s.

Appendix I



Supplementary Figure 8: An online application that allows users to enter their genetic and clinical measurements and the individual's estimated CFRD risk, as a function of age, is returned. Each predictor is weighted differently based on the hazard ratios reported in the study (Table 2). The application uses drop-down menus to specify the genetic and clinical measurements required from the user. For instance, the user is required to supply the number of A alleles for SNP rs1964986 in *PRSS1*. The application then outputs the population-level CFRD risk distribution and the estimated percentile of CFRD risk given an individual's entered genetic and clinical measurements. The application also outputs the estimated CFRD prevalence rates across ages to facilitate clinical decision making and encourage increased adherence to OGTT screening. Example output is displayed for an individual with the noted measurement values in the drop-down menu.

Appendix J

Three-Stage Approach to Model CFRD Risk

1) **Hierarchical Clustering:** We performed hierarchical clustering to remove highly correlated SNPs. Since the number of covariates far exceed the limited sample size available in this study, directly applying variable selection techniques such as LASSO can lead to highly unstable models with correlated predictors[22]. Using spearman correlation of 0.8 as a cutoff, only 2,488 of the 3,984 pre-selected SNPs were extracted for downstream analyses.

2) **Variable Selection:** We performed variable selection using 100 iterations of stability selection with component-wise gradient boosting (CWGB). At each iteration, the model is trained on half of the training data and returns different sets of selected variables. Stability selection counts the number of iterations each variable is chosen and ranks variable importance by their selected frequencies. A frequency cutoff was applied to select the most stable predictors that consistently predict CFRD risk in different subsets of the data. In this study, we used a 50% cutoff to select variables that are associated with CFRD in more than half of the subsets.

3) **Re-estimate over-penalized effect sizes:** Penalized regression models such as CWGB with early-stopping rules can over-penalize effect sizes for the chosen variables. Therefore, we re-estimated effect sizes using a Cox Proportional Hazards model (Cox PH) with covariates selected by stability selection while adjusting for the first 10 PCs. A predicted risk score can then be generated for each individual by substituting one's covariate information (without the PCs) in the Cox model's linear predictor.

Appendix K

Research Ethic Board Approving the Study at each Participating Study Site

Canada

Research ethic boards at the following institutions:

- The Hospital for Sick Children
- Alberta Children's Hospital
- BC Children's Hospital
- Children's Hospital of Eastern Ontario
- Children's Hospital of Western Ontario
- Children's Hospital of Winnipeg
- Foothills Medical Centre
- Grand River Hospital
- Kingston Health Sciences Centre
- CRCHUM (Centre de recherche du CHUM (Centre hospitalier de l'Université de Montréal))
- IUCPQ-Université Laval
- IWK Health Centre
- Janeway Children's Health & Rehabilitation Centre
- Queen Elizabeth II Health Sciences Centre
- Royal University Hospital
- St. Mary's General Hospital
- St Michael's Hospital
- St Paul's Hospital
- University of Alberta Hospital

France

- French Ethical Committee (CPP n°2004/15)
- Commission Nationale de L'informatique et des Libertés (n°04.404)