Online Data Supplement

A Risk Prediction Model for Mortality Among Smokers in the COPCGene® Study

Appendix A. Candidate predictors and model selection

Variables considered in initial models included the following: demographic (age, height, weight, BMI, race), behavioral (current smoking, pack-years, years since quit), comorbidities [separate composite scores for CVD, cancer, bones and joints, stroke or Transient Ischemic Attack (TIA)], disease indicators (diabetes, high cholesterol, macular degeneration, gastroesophageal reflux, stomach ulcers and asthma), medication uses, lung (FEV_1, FVC, FEV_1/FVC), CT [Pi10, sAWT, wall area percent, Perc15 (15^{th} percentile of the density histogram) on inspiratory and expiratory scans, volume-adjusted lung density, total lung volume (TLV), PRM variables (functional small airway disease and emphysema), functional residual capacity (FRC)], visual CT variables (emphysema, paraseptal emphysema, bronchial wall thickening), FVC/TLV (using TLV from CT), exacerbations, 6MWD, symptom scores (St. George Respiratory Questionnaire [SGRQ] variables, mMRC dyspnea score), chronic bronchitis. More detail about the quantitative and visual emphysema variables can be found in references 27, 31, 33, 45.

Preliminary analyses were performed using logistic regression to help manage variable selection using stepwise selection methods. Some model selection was then performed directly using the parametric survival models, with input from clinicians to ensure that at least one predictor from each major category was represented in the model. The adequacy of the final set of predictors was assessed by reviewing goodness-of-statistics such as time-dependent area under the curve statistics, diagnostic plots, and examining predictor impact using p-values corresponding to test statistics. The final set of predictors was also reviewed by clinicians for usefulness, plausibility and completeness.

Severe exacerbations and parametric response mapping (PRM) emphysema and gas trapping measures were significant in multivariable models but were not included in final models since many subjects did not have these measures at the time of analysis. 'Years since quit' was tested for addition but was insignificant in the model that already included the other two smoking variables (and other covariates), so was excluded. FEV_1/FVC was included in initial models but excluded from final ones as it also contributed very little beyond the other predictors.

Interaction terms between final predictors were also assessed for inclusion in the model by using stepwise selection methods in logistic regression models (irrespective of time to event), using all 2-way interactions among the final set of 12 predictors. The interaction terms with largest impact were 6MWD*current smoking status for men, and BMI*age for women; given their relatively small contribution to the model as a whole and for ease of interpretation, interactions were not included in final models. Given that CT and exercise variables may be more difficult to obtain in many health centers, another set of models were fit that removed the distance walked, emphysema and airway wall thickening predictors.

Appendix B. Determining point scores from the fitted survival model, with examples

Risk estimates can be determined for any time within 10 years using Equation (1) from the manuscript text; the required inputs are time *t*, and estimates of the linear predictor $\boldsymbol{x}_i'\boldsymbol{\beta}$, and the scale parameter σ. An alternative to computing the linear predictor directly is to approximate it using the point system, using Equation (2) in the text. Estimates of beta parameters from the Weibull model are naturally defined such that higher values increase survival likelihood. In order to make higher points reflect higher risk of mortality, a few adjustments were made in order to directly apply Equation (2). First, point scales were shifted for predictors to be non-positive (if necessary), and the adjusted intercept was then modified to compensate for this (indicated with the prime in Equation (3) below). The signs on these points were then removed (those shown in Tables 2 and S5) so that the corresponding point total was also nonnegative ("– *shifted point total*$_i$" below). To compensate for this flip, the approximation becomes

$$\boldsymbol{x}_i'\boldsymbol{\beta} \approx adjusted\ intercept' - \left(-\ shifted\ point\ total_i\right) * S$$

Equation (3)

Equation (3) reexpresses Equation (2), but where " – *shifted point total*$_i$" is obtained from Table 2 or Table S5. Once Equation (3) is applied, risk estimates can be computed based on Equation (1). Both the direct calculation of the linear predictor as well as the point approximation method using Equation (3) are illustrated below. For further detail on the calculations, please contact the corresponding author.

Example 1: 72.6 year-old female with CVD, an FEV1 of 0.969, BMI of 30.07, sAWT z-score of -0.818, distance walked of 1675 feet, former smoker with 49.5 pack-years, dyspnea score of 0, and 'low' visual emphysema score. Calculations would be as follows:

Predictor	Subject value	Point approach	$\boldsymbol{x}_i'\boldsymbol{\beta}$ approach
Intercept			5.468337
FEV_1, % pred.	0.969	0	+ 2.680588(0.969) – 1.111009(0.969)^2
Age, years	72.6	12	+ 0.039362(72.6) – 0.000529(72.6)^2
BMI, kg/m2	30.07	2	+ 0.094463(30.07) – 0.001205(30.07)^2
sAWT z-score	⁻0.818	2	+ 0.085056(⁻0.818)
6MWD, feet	1675	7	+ 0.000681(1675)
Dyspnea score	0	0	– 0.082935(0)
CT visual emphysema	Low[a]	3	+ 0.193291
Current smoking	No	0	+ 0
Pack-years smoking	49.5	2	– 0.002588(49.5)
Diabetes	No	0	+ 0
CVD	Yes	2	– 0.103437
Cancer	No	0	+ 0
		—	———
Total		30	10.0168

[a]Low=trace, mild or moderate.

10-year risk estimates for two approaches:

Exact ($\boldsymbol{x}_i'\boldsymbol{\beta}$) approach: P(10 yr mortality)=1–exp(–exp[(ln(3652.5)–10.0168)/0.667965]) = **6.4%**

Point approach: Linear predictor = Adj. intercept – (– *shifted point total* * S)

= 11.52395203 – (30*0.05)

= 10.023952

P(10 yr mortality) = 1–exp(–exp[(ln(3652.5)–10.02395)/0.667965])

= **6.3%**

Example 2: 47.7 year-old male with CVD, an FEV1 of 0.85, BMI of 24.15, sAWT z-score of 0.528, distance walked of 1620 feet, current smoker with 33.7 pack-years, dyspnea score of 0, and 'low' visual emphysema score. Calculations would be as follows:

Predictor	Subject value	Point approach	$\boldsymbol{x}_i'\boldsymbol{\beta}$ approach
Intercept			2.734014
FEV_1. % pred.	0.85	0	+ 2.456290(0.85)–1.329384(0.85)^2
Age, years	47.7	0	+ 0.114866(47.7) – 0.001127(47.7)^2
BMI, kg/m2	24.15	6	+ 0.119280(24.15) – 0.001471(24.15)^2
sAWT z-score	0.528	7	– 0.120124(0.528)
6MWD, feet	1620	6	+ 0.000588(1620)
Dyspnea score	0	0	– 0.088768(0)
CT visual emphysema	Low[a]	2	+ 0.169568
Current smoking	Yes	5	– 0.234298
Pack-years smoking	33.7	1	– 0.001066(33.7)
Diabetes	No	0	+ 0
CVD	Yes	2	– 0.108811
Cancer	No	0	+ 0
		—	_____
Total		29	9.4796

[a]Low=trace, mild or moderate.

10-year risk estimates for two approaches:

Exact ($\boldsymbol{x}_i'\boldsymbol{\beta}$) approach: P(mortality)=1–exp(–exp[(ln(3652.5)–9.4796)/0.684146]) = **14.3%**

Point approach: Linear predictor = Adj. intercept – (– *shifted point total* * S)

= 10.94352157 – (29*0.05)

= 9.493522

P(10 yr mortality) = 1–exp(–exp[(ln(3652.5)–9.493522)/0.684146])

= **14.1%**

Note: For both examples, the more decimal places that are kept in the calculation, the greater the accuracy. (Pre-calculation numbers shown above are rounded, but more decimal places were retained in actual calculations.)

Appendix C. Supplemental tables and figures

Table S1. Demographics of COPDGene and SPIROMICS subjects, using sample sizes restricted by full and reduced models (due to missing data for some variables). Entries are Mean (SD) for continuous variables, and % (as indicated) for counts.

		Full model		Reduced model	
		COPDGene	**SPIROMICS**	**COPDGene**	**SPIROMICS**
Category	**Variable**	n=9074	n=846	n=9867	n=2630
Outcome	6-yr mortality rate[a]	11.9%	10.0%	12.7%	10.7%
	Max years followed	10.6	7.3	10.6	7.3
Demographic	Age, years	59.7 (9.0)	64.7 (8.7)	59.7 (9.0)	63.6 (8.8)
	BMI, kg/m^2	28.8 (6.2)	27.9 (5.2)	28.9 (6.3)	27.9 (5.3)
	Males	53.1%	55.2%	53.0%	54.0%
Smoking	Current smoker	52.1%	36.2%	52.2%	39.6%
	Pack-years	44.2 (24.8)	49.0 (22.7)	44.4 (25.0)	49.4 (26.0)
Symptoms	Dyspnea score				
	0 (Low)	45.1%	34.3%	44.3%	31.4%
	1	14.2%	42.9%	13.8%	43.0%
	2	13.0%	13.6%	13.0%	15.0%
	3	18.1%	7.1%	18.7%	8.3%
	4 (High)	9.6%	2.1%	10.3%	2.4%
Lung function	FEV_1, % predicted	76.9 (25.2)	72.9 (25.7)	76.2 (25.5)	73.0 (26.3)
	FEV_1/FVC, %	66.9 (15.9)	59.8 (16.4)	66.6 (16.2)	60.0 (16.6)
	GOLD group				
	PRISm	12.2%	2.8%	12.4%	2.6%
	0	43.7%	28.4%	43.1%	30.6%
	1	8.1%	16.6%	7.8%	14.7%
	2	19.3%	30.1%	19.1%	29.8%
	3	11.4%	17.0%	11.5%	15.7%
	4	5.4%	5.1%	6.1%	6.7%
CT	Visual presence of emphysema	66.5%	91.4%		
	sAWT[b], mm	1.1 (0.23)	1.5 (0.13)		
Exercise	6MWD, feet	1361.8 (395.0)	1347.0 (380.5)		
Comorbidities	Diabetes	12.8%	12.7%	13.2%	13.4%
	CVD	49.1%	59.0%	49.4%	58.2%
	Cancer	5.0%	12.3%	5.0%	11.3%

[a]Based on Kaplan-Meier estimates, to account for those lost to follow-up; 6-year estimates chosen so that comparisons could be made between cohorts.

[b]Thirona used for COPDGene, VIDA for SPIROMICS. Note that a subset of subjects in COPDGene also had VIDA measurements, for which the mean and SD were similar to that of SPIROMICS; Thirona measurements were used for COPDGene due to greater availability of data, whereas VIDA was the only approach used in SPIROMICS.

Abbreviations: BMI – body mass index; CVD – cardiovascular disease; FEV_1 – forced expiratory volume in 1 second; GOLD – Global initiative for chronic Obstructive Lung Disease; PRISm – Preserved ratio and impaired spirometry; sAWT – mean segmental airway wall thickening; SD – standard deviation; 6MWD – 6-minute walk distance.

Table S2. Demographics of SPIROMICS subjects used for validation in full and reduced models, by gender. Differences in sample sizes between full and reduced models was due to missing data for some variables. Entries are Mean (SD) for continuous variables, and % (as indicated) for counts.

		Full model		Reduced model	
		Women	**Men**	**Women**	**Men**
Category	**Variable**	379	467	1211	1419
Outcome	6-yr mortality rate[a]	7.3%	12.8%	9.1%	12.2%
	Max years followed	7.3	7.3	7.3	7.3
Demographic	Age, years	64.2 (9.1)	65.0 (8.4)	63.0 (9.0)	64.1 (8.6)
	BMI, kg/m^2	27.5 (5.6)	28.2 (4.9)	27.8 (5.7)	28.0 (4.9)
Smoking	Current smoker	38.5%	34.3%	41.5%	38.1%
	Pack-years	45.2 (20.2)	52.0 (24.2)	45.4 (20.3)	52.8 (29.7)
Symptoms	Dyspnea score				
	0 (Low)	29.6%	38.1%	26.6%	35.5%
	1	44.1%	42.0%	44.3%	41.9%
	2	17.4%	10.5%	19.0%	11.6%
	3	6.9%	7.3%	7.7%	8.7%
	4 (High)	2.1%	2.1%	2.5%	2.3%
Lung function	FEV_1, % predicted	72.9 (25.6)	72.9 (25.8)	74.0 (25.8)	72.2 (26.8)
	FEV_1/FVC, %	61.7 (16.7)	58.3 (16.1)	62.0 (16.2)	58.2 (16.7)
	GOLD group				
	PRISm	4.2%	1.7%	3.5%	1.9%
	0	32.5%	25.1%	35.0%	26.8%
	1	13.7%	18.8%	11.7%	17.2%
	2	25.9%	33.6%	28.3%	31.0%
	3	19.0%	15.4%	16.4%	15.1%
	4	4.8%	5.4%	5.0%	8.0%
CT	Visual presence of emph.	90.0%	92.5%		
	sAWT[b], mm	1.4 (0.09)	1.6 (0.10)		
Exercise	6MWD, feet	1290.4 (374.8)	1392.9 (379.2)		
Comorbidities	Diabetes	11.1%	13.9%	11.7%	14.8%
	CVD	58.1%	59.7%	56.9%	59.3%
	Cancer	11.6%	12.9%	11.6%	11.1%

[a]Based on Kaplan-Meier estimates, to account for those lost to follow-up.
[b]VIDA used for SPIROMICS.
Abbreviations: BMI – body mass index; CVD – cardiovascular disease; FEV_1 – forced expiratory volume in 1 second; GOLD – Global initiative for chronic Obstructive Lung Disease; mMRC – modified Medical Research Council; PRISm – Preserved ratio and impaired spirometry; sAWT – mean segmental airway wall thickening; SD – standard deviation; 6MWD – 6-minute walk distance.

Table S3. Parameter estimates for Weibull models, using full set of predictors. Increasing parameter estimate values increase survival likelihood. These estimates provide a more accurate way of determining subject risk of mortality based on the fitted model. For examples, see Appendix B.

	Women				Men			
Parameter	**Estimate**	**SE**	**Chi-Sq**	**P-value**	**Estimate**	**SE**	**Chi-Sq**	**P-value**
Intercept	5.468337	1.3634	16.09	<.0001	2.734014	1.1832	5.34	0.0208
6MDW, feet	0.000681	0.0001	57.55	<.0001	0.000588	0.0001	72.94	<.0001
Current smoker[a]	-0.330043	0.0722	20.87	<.0001	-0.234298	0.0584	16.12	<.0001
FEV_1, % of pred.	2.680588	0.5392	24.71	<.0001	2.456290	0.4281	32.93	<.0001
FEV_1^2, % of pred.	-1.111010	0.4030	7.60	0.0058	-1.329380	0.3149	17.82	<.0001
Age, years	0.039362	0.0425	0.86	0.3542	0.114866	0.0351	10.69	0.0011
Age^2, years	-0.000529	0.0003	2.51	0.1134	-0.001127	0.0003	16.60	<.0001
BMI, kg/m^2	0.094463	0.0234	16.29	<.0001	0.119280	0.0269	19.70	<.0001
BMI^2, kg/m^2	-0.001205	0.0004	11.05	0.0009	-0.001471	0.0004	11.30	0.0008
Dyspnea score	-0.082935	0.0255	10.59	0.0011	-0.088830	0.0202	19.41	<.0001
sAWT, z-score	-0.085056	0.0365	5.42	0.0199	-0.120124	0.0295	16.63	<.0001
Visual emphysema[b]								
None	0.342787	0.1095	9.80	0.0017	0.269973	0.0901	8.97	0.0027
Low	0.193291	0.0753	6.58	0.0103	0.169568	0.0612	7.67	0.0056
High	0				0			
Diabetes[a]	-0.168710	0.0865	3.80	0.0511	-0.190287	0.0617	9.51	0.0020
CVD[a]	-0.103437	0.0601	2.96	0.0854	-0.108811	0.0501	4.72	0.0297
Cancer[a]	-0.241954	0.1070	5.11	0.0238	-0.161668	0.0804	4.04	0.0445
Pack-years	-0.002588	0.0011	5.25	0.0219	-0.001066	0.0008	1.89	0.1692
Scale[c]	0.667965	0.0252	-	-	0.684146	0.0205	-	-

[a]Indicator variable for stated condition.
[b]Low=trace, mild or moderate; High=confluent or advanced destructive.
[c]The Weibull shape parameter for this model is the inverse of the scale parameter.
Abbreviations: BMI – body mass index; CVD – cardiovascular disease; FEV_1 – forced expiratory volume in 1 second; sAWT – mean segmental airway wall thickening; 6MWD – 6-minute walk distance.

Table S4. Parameter estimates for Weibull models, using reduced set of predictors. Increasing parameter estimate values increase survival likelihood.

	Women				Men			
Parameter	**Estimate**	**SE**	**Chi-Sq**	**P-value**	**Estimate**	**SE**	**Chi-Sq**	**P-value**
Intercept	6.306427	1.3518	21.76	<.0001	2.916401	1.1322	6.63	0.0100
Current smoker[a]	-0.406114	0.0669	36.87	<.0001	-0.326372	0.0532	37.67	<.0001
FEV_1, % of pred.	4.418283	0.5087	75.44	<.0001	3.471928	0.3990	75.70	<.0001
FEV_1^2, % of pred.	-1.955450	0.3816	26.27	<.0001	-1.649120	0.2977	30.68	<.0001
BMI, kg/m^2	0.082303	0.0218	14.26	0.0002	0.141649	0.0245	33.30	<.0001
BMI^2, kg/m^2	-0.001119	0.0003	10.85	0.0010	-0.001995	0.0004	25.46	<.0001
Age, years	0.050555	0.0419	1.46	0.2276	0.124480	0.0336	13.70	0.0002
Age^2, years	-0.000696	0.0003	4.48	0.0343	-0.001236	0.0003	21.71	<.0001
Dyspnea score	-0.167368	0.0246	46.11	<.0001	-0.144363	0.0189	58.27	<.0001
Diabetes[a]	-0.164885	0.0834	3.91	0.0480	-0.208005	0.0597	12.15	0.0005
CVD[a]	-0.176855	0.0591	8.96	0.0028	-0.105932	0.0486	4.75	0.0294
Cancer[a]	-0.134860	0.1097	1.51	0.2190	-0.186861	0.0787	5.64	0.0175
Pack-years	-0.003300	0.0011	8.81	0.0030	-0.002204	0.0007	8.69	0.0032
Scale[b]	0.707303	0.0249	-	-	0.707654	0.0200	-	-

[a]Indicator variable for stated condition.
[b]The Weibull shape parameter for this model is the inverse of the scale parameter.
Abbreviations: BMI – body mass index; CVD – cardiovascular disease; FEV_1 – forced expiratory volume in 1 second.

Table S5. Converting regression coefficients and predictors into points, for full and reduced models (full model points are the same as in Table 2; they are presented here for comparison). For a subject with given characteristics, determine points associated with each predictor and then add for a total score. The risk of death within 10 years can then be computed based on the total score. For intervals of continuous variables, the closed bracket includes the value, while the open bracket does not. For example, [50, 55) means at least 50 but less than 55.

			Full		Reduced	
Category	Risk factor	Level	W	M	W	M
Demographic	Age, years	<50	0	0	0	0
		[50, 55)	1	0	2	0
		[55, 60)	3	1	5	1
		[60, 65)	6	3	8	3
		[65, 70)	9	6	12	7
		[70, 75)	12	10	16	12
		[75, 80)	16	16	22	18
		≥80	21	22	28	25
	BMI, kg/m^2	<20	11	15	8	12
		[20, 24)	7	10	5	7
		[24, 28)	4	6	3	3
		[28, 32)	2	3	1	1
		[32, 36)	1	1	0	0
		[36, 40)	0	0	0	0
		[40, 44)	0	0	1	1
		[44, 48)	1	1	2	4
		≥48	3	2	4	8
Smoking	Current smoker	Yes	7	5	8	7
	Pack-years	[10, 25)	1	0	1	1
		[25, 50)	2	1	2	2
		[50, 75)	3	1	4	3
		[75, 100)	5	2	6	4
		≥100	6	2	7	5
Symptoms	Dyspnea score	0 (Low)	0	0	0	0
		1	2	2	3	3
		2	3	4	7	6
		3	5	5	10	9
		4 (High)	7	7	13	12

			Full		Reduced	
Category	Risk factor	Level	W	M	W	M
Lung function	FEV1, % pred	<20%	24	16	37	27
		[20, 30)	19	12	30	21
		[30, 40)	15	9	23	16
		[40, 50)	12	6	17	12
		[50, 60)	9	4	12	8
		[60, 70)	6	2	8	5
		[70, 80)	4	1	5	3
		[80, 90)	2	0	2	1
		≥90	0	0	0	0
CT	Visual emphysema[a]	None	0	0		
		Low	3	2		
		High	7	5		
	sAWT, z-score	< −1.5	0	0		
		[−1.5, −0.5)	2	2		
		[−0.5, 0.5)	3	5		
		[0.5, 1.5)	5	7		
		≥1.5	7	10		
Exercise	6MWD, feet	<500	24	21		
		[500, 750)	20	18		
		[750, 1000)	17	15		
		[1000, 1250)	14	12		
		[1250, 1500)	10	9		
		[1500, 1750)	7	6		
		[1750, 2000)	3	3		
		≥2000	0	0		
Comorbidities	Diabetes		3	4	3	4
	CVD		2	2	4	2
	Cancer		5	3	3	4

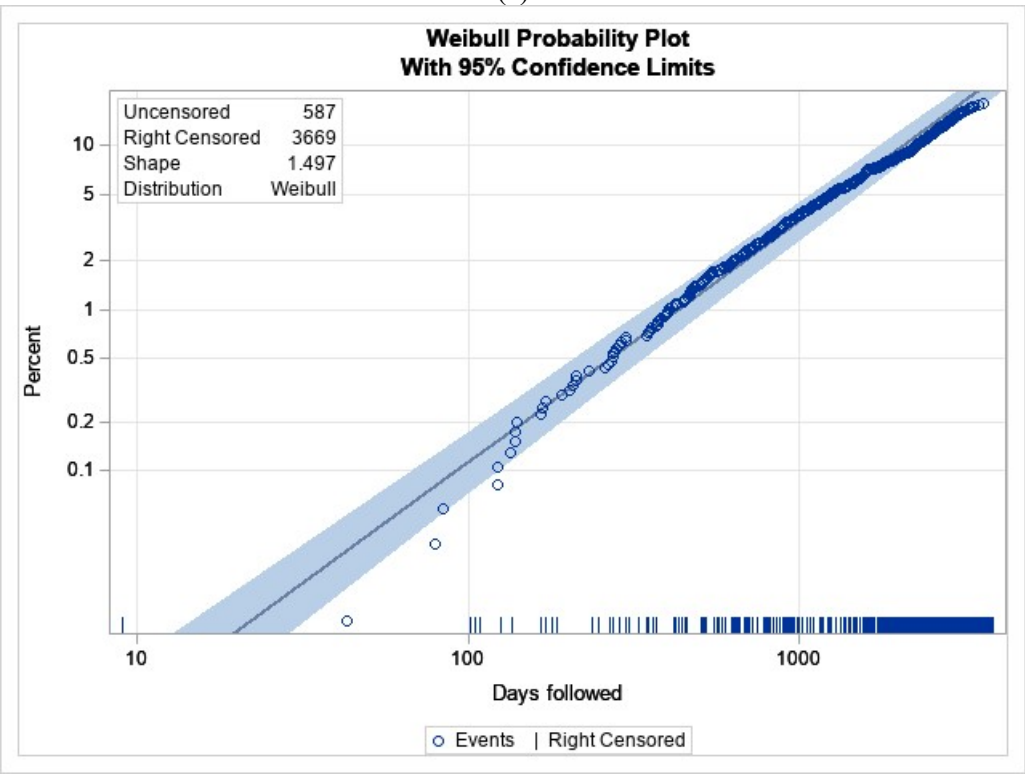
[a]Low=trace, mild or moderate; High=confluent or advanced destructive.
Abbreviations: BMI – body mass index; CVD – cardiovascular disease; FEV_1 – forced expiratory volume in 1 second; sAWT – mean segmental airway wall thickening; 6MWD – 6-minute walk distance.

Table S6. Risk estimates associated with Point totals for reduced model (see Table S5).

Point	Women	Men	Point	Women	Men
0	1.6%	4.6%	38	21.4%	49.8%
1	1.7%	4.9%	39	22.7%	52.3%
2	1.9%	5.3%	40	24.2%	54.8%
3	2.0%	5.7%	41	25.7%	57.4%
4	2.1%	6.1%	42	27.3%	59.9%
5	2.3%	6.5%	43	29.0%	62.5%
6	2.5%	6.9%	44	30.7%	65.1%
7	2.7%	7.4%	45	32.6%	67.7%
8	2.8%	7.9%	46	34.5%	70.3%
9	3.0%	8.5%	47	36.5%	72.8%
10	3.3%	9.1%	48	38.6%	75.3%
11	3.5%	9.7%	49	40.7%	77.7%
12	3.8%	10.4%	50	43.0%	80.0%
13	4.0%	11.1%	51	45.3%	82.2%
14	4.3%	11.9%	52	47.6%	84.4%
15	4.6%	12.7%	53	50.1%	86.3%
16	4.9%	13.6%	54	52.5%	88.2%
17	5.3%	14.5%	55	55.0%	89.9%
18	5.7%	15.5%	56	57.6%	91.5%
19	6.1%	16.5%	57	60.2%	92.9%
20	6.5%	17.6%	58	62.8%	94.1%
21	7.0%	18.7%	59	65.4%	95.2%
22	7.5%	20.0%	60	68.0%	96.2%
23	8.0%	21.3%	61	70.5%	97.0%
24	8.5%	22.6%	62	73.1%	97.7%
25	9.1%	24.1%	63	75.5%	98.2%
26	9.8%	25.6%	64	77.9%	98.7%
27	10.5%	27.2%	65	80.2%	99.0%
28	11.2%	28.8%	66	82.5%	99.3%
29	11.9%	30.6%	67	84.6%	99.5%
30	12.8%	32.4%	68	86.5%	99.7%
31	13.6%	34.3%	69	88.4%	99.8%
32	14.6%	36.3%	70	90.1%	99.9%
33	15.5%	38.4%	71	91.6%	99.9%
34	16.6%	40.5%	72	93.0%	
35	17.7%	42.8%	73	94.2%	
36	18.8%	45.1%	74	95.3%	
37	20.1%	47.4%	75	96.3%	

Figure S1. Probability plots for the Weibull full predictive model, for (a) women, and (b) men. The plots show good fits for the empirical approach (K-M estimates, circles), relative to the Weibull model (line in middle, with 95% confidence bands shaded). Only a small handful of points are not within the confidence bands at early days followed, out of over 9,000 points modeled.

(a)



(b)

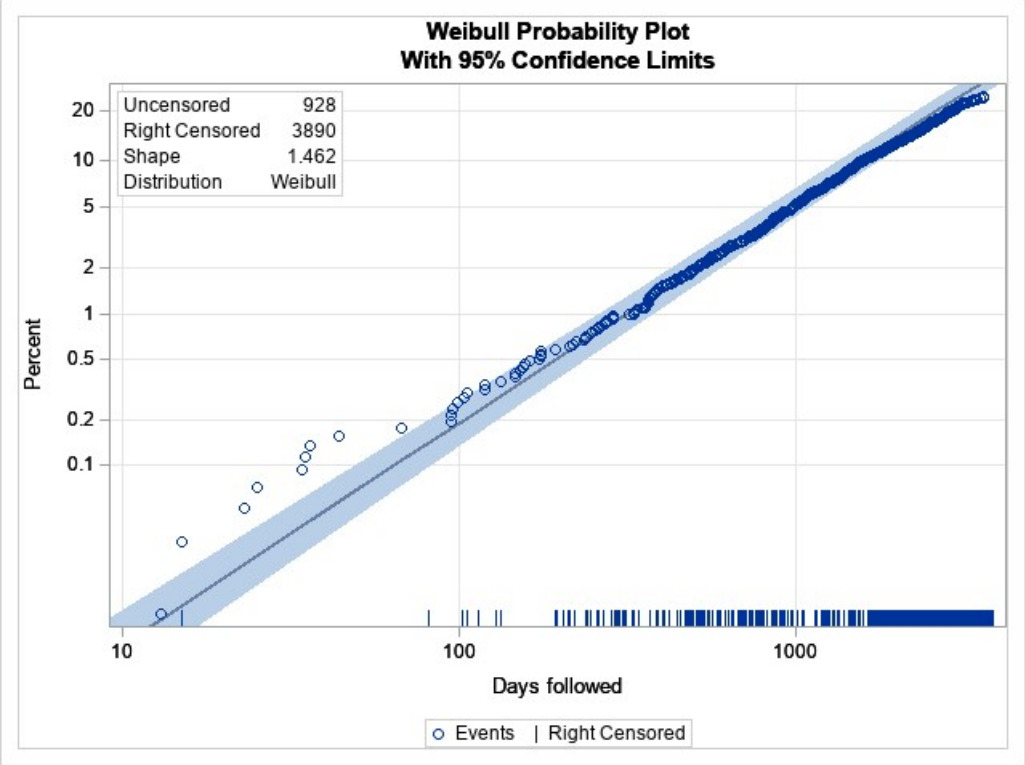
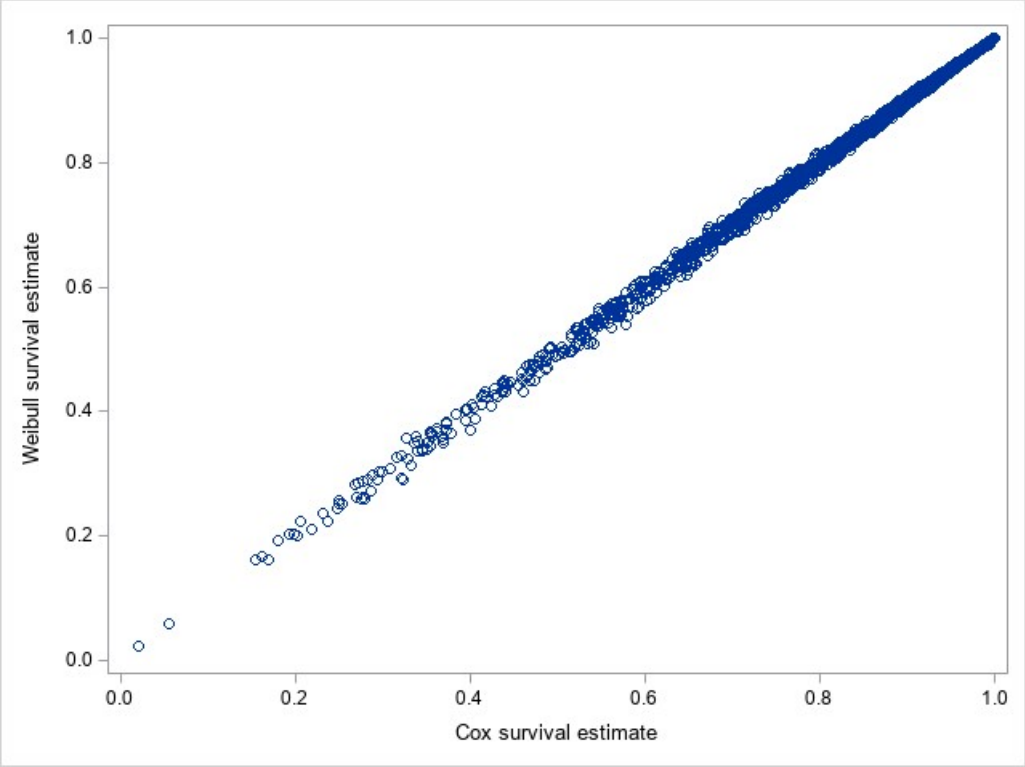


Figure S2. Survival (1–risk) estimates for the parametric survival model with Weibull distribution (y-axis) versus the proportional hazards (Cox) survival model (x-axis) for (a) women and (b) men. Survival estimates were computed at the last day of follow-up for subjects. Graphs show high consistency between approaches.

(a)



(b)