A method for comparing multiple imputation techniques: a case study on the U.S. National COVID Cohort Collaborative

Elena Casiraghi[1,2,3*], Rachel Wong[4,*], Margaret Hall[4], Ben Coleman[5,6], Marco Notaro[1,2], Michael D. Evans[7], Jena S. Tronieri[8], Hannah Blau[5], Bryan Laraway[9], Tiffany J. Callahan[9], Lauren E. Chan[10], Carolyn T. Bramante[11], John B. Buse[12,13], Richard A. Moffitt[4], Til Stürmer[14], Steven G. Johnson[15], Yu Raymond Shao[16,17], Justin Reese[3], Peter N. Robinson[5,6], Alberto Paccanaro[18,19], Giorgio Valentini[1,2], Jared D. Huling[20,**] and Kenneth J. Wilkins[21,**] on behalf of the N3C Consortium

[1] AnacletoLab, Department of Computer Science "Giovanni degli Antoni", Università degli Studi di Milano, Milan, ITALY
[2] CINI, Infolife National Laboratory, Roma, ITALY
[3] Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[4] Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA
[5] The Jackson Laboratory for Genomic Medicine, Farmington, USA
[6] Institute for Systems Genomics, University of Connecticut, Farmington, CT, USA
[7] Biostatistical Design and Analysis Center, Clinical and Translational Science Institute, University of Minnesota, Minneapolis, MN, USA
[8] Department of Psychiatry, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA
[9] University of Colorado, Anschutz Medical Campus, Aurora, CO, USA
[10] College of Public Health and Human Sciences, Oregon State University, Corvallis, USA
[11] Division of General Internal Medicine, University of Minnesota, Minneapolis, MN, USA
[12] NC Translational and Clinical Sciences Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[13] Division of Endocrinology, Department of Medicine, University of North Carolina School of Medicine, USA
[14] Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[15] Institute for Health Informatics, University of Minnesota, Minneapolis, MN, USA
[16] Harvard-MIT Division of Health Sciences and Technology (HST), 260 Longwood Ave, Boston, USA
[17] Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, USA
[18] School of Applied Mathematics (EMAp), Fundação Getúlio Vargas, Rio de Janeiro, BRAZIL
[19] Department of Computer Science, Royal Holloway, University of London, Egham, UK
[20] Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, USA
[21] Biostatistics Program, Office of the Director, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda, MD, USA

N3C Consortium: Tell Bennet, Christopher Chute, Peter DeWitt, Kenneth Gersing, Andrew Girvin, Melissa Haendel, Jeremy Harper, Janos Hajagos, Stephanie Hong, Emily Pfaff, Jane Reusch, Corneliu Antoniescu, Kimberly Robaski
Corresponding Authors (*): Elena Casiraghi, Rachel Wong
Co-senior Authorship (**): Jared D Huling and Kenneth J. Wilkins equally contributed to the work

Supplementary files

S1.xlsx:

This Excel file contains all the results we obtained when using a number of imputation m = 42.
The file is composed by:

- sheet "mean_measures_m42" containing the colored table (Figure 5) showing and detailing the average measures obtained by the tested imputation algorithms across the three outcomes.
- sheets "RB_mean" (see also Supplementary Figure S1), "MSE_mean" (see also Supplementary Figure S2), "ER_mean" (see also Supplementary Figure S3), and "CR_mean" contain the four win-tie-loss tables (for the RB measure, the MSE measure, the ER measure, the CR measure) obtained by summing the wins, ties, losses obtained by each model over the three outcome variables.
  On the right, each of the four sheets contains the mean of the win-tie-loss tables over the three outcomes, where the wins, ties, and losses are computed by comparing the models on the rows to the models on the column by a paired-sided paired rank sign test .
  The grid shows numbers in the range [-3, +3]; they are computed by representing each win by a +1 value, each tie as a 0 value, each loss as a -1 value.

S2.xlsx:

This Excel file has the same structure of S1.xlsx; it details all the results we obtained when using a number of imputation m = 5.

S3_MCAR.xlsx:

This Excel file has the same structure of S1.xlsx and S2.xslx; it details all the results we obtained when simulating MCAR missingness in the amputated datasets.

S4_MNAR.xlsx:

This Excel file has the same structure of S1.xlsx and S2.xslx and S3_MCAR.xlsx; it details all the results we obtained when simulating MNAR missingness in the amputated datasets.

Supplementary Figures:

MI algorithm	univariate imputation method	use outcomes	one-hot encode binned numeric predictors	one-hot encode categorical predictors	univariate imputation order	pmm donors	average of absolute values of RB across outcomes	wins	ties	losses
amelia		F	T				0.025	1	4	-94
			F	T			0.017	7	16	-50
		T	T				0.021	8	26	-28
			F				0.014	44	21	-10
mice	default	F	F	F	monotone		0.012	6	16	-63
					revmonotone		0.012	5	18	-60
		T			monotone		0.012	7	14	-63
					revmonotone		0.012	7	15	-64
	logreg	F	T	T	monotone		0.013	5	6	-85
					revmonotone		0.013	5	3	-89
		T			monotone		0.011	18	21	-34
					revmonotone		0.011	22	18	-31
	norm	F	T	T	monotone		0.012	7	8	-76
					revmonotone		0.012	6	5	-80
			F		monotone		0.012	10	11	-63
					revmonotone		0.013	9	10	-66
		T			monotone		0.011	34	16	-23
					revmonotone		0.011	35	17	-23
			F		monotone		0.011	40	18	-16
					revmonotone		0.011	41	17	-15
missRanger	extratrees	F	T	T	monotone		0.007	109	4	0
					monotone	3	0.008	40	8	-38
					monotone	5	0.009	19	16	-49
					revmonotone		0.007	107	4	0
					revmonotone	3	0.008	51	8	-32
					revmonotone	5	0.009	22	13	-49
			F	F	monotone		0.007	113	4	-2
					monotone	3	0.010	26	17	-30
					monotone	5	0.010	12	20	-38
					revmonotone		0.007	114	4	-2
					revmonotone	3	0.010	27	15	-29
					revmonotone	5	0.010	10	18	-45
		T	T	T	monotone		0.006	83	16	0
					monotone	3	0.006	81	10	-9
					monotone	5	0.007	66	14	-14
					revmonotone		0.006	83	16	0
					revmonotone	3	0.006	84	8	-8
					revmonotone	5	0.007	59	14	-15
					monotone		0.006	51	13	-9
					monotone	3	0.008	47	18	-12
			F	F	monotone	5	0.009	36	18	-21
					revmonotone		0.006	50	14	-8
					revmonotone	3	0.008	46	18	-13
					revmonotone	5	0.008	34	18	-25
IPW	logreg	F	T				0.001	31	24	-6
			F				0.001	32	24	-6
		T	T				0.001	28	27	-6
			F				0.001	28	27	-6
	RF	F	T				0.002	5	11	-48
			F				0.004	3	4	-60
		T	T				0.013	0	3	-79
			F				0.017	0	2	-82

Figure S1: Column "average absolute value of RB across predictors and outcomes" reports the average RB measure across the hospitalization, invasive ventilation, and patients' survival outcomes (the table is also made available in Supplementary file S1 – sheet "RB_mean"). Columns "wins", "ties", "losses" report the sum of, respectively, wins, ties, and losses computed by comparing the (absolute value of the) RB measures over the three outcomes (the corresponding win-tie-loss grid is shown in the Supplementary material). The comparison between two models over an outcome variable is performed with a sided Wilcoxon signed-rank test comparing the distribution of the (absolute) RB values for all the predictor variables. The winner is the model achieving the lowest RB distribution. All the models but missRanger with no pmm and using the outcome variables in the imputation model are obtaining RB ≤ 0, meaning that the computed estimates are systematically lower than those computed on the complete dataset. missRanger with outcome variable in the imputation model and no pmm is instead bringing to the computation of inflated estimates**.**

MI algorithm	univariate imputation method	use outcomes	one-hot encode binned numeric predictors	one-hot encode categorical predictors	univariate imputation order	pmm donors	average MSE across outcomes	wins	ties	losses
amelia		F	T				0.007	18	5	-76
			F	T			0.004	25	28	-22
		T	T				0.005	26	23	-22
			F				0.003	36	19	-1
mice	default	F	F	F	monotone		0.002	26	21	-32
					revmonotone		0.002	25	20	-34
		T			monotone		0.002	30	18	-32
					revmonotone		0.002	26	19	-34
	logreg	F	T	T	monotone		0.003	25	9	-58
					revmonotone		0.003	25	7	-60
		T			monotone		0.002	37	26	-15
					revmonotone		0.002	39	25	-13
	norm	F	T	T	monotone		0.002	28	13	-43
					revmonotone		0.002	27	9	-48
		T	F		monotone		0.002	30	16	-33
					revmonotone		0.002	28	17	-34
		F	T		monotone		0.002	54	15	-11
					revmonotone		0.002	53	16	-9
		T	F		monotone		0.002	67	12	-5
					revmonotone		0.002	72	14	-2
missRanger	extratrees	F	T	T	monotone		0.001	93	2	0
					monotone	3	0.001	38	22	-22
					monotone	5	0.001	28	21	-34
					revmonotone		0.001	92	3	0
					revmonotone	3	0.001	48	20	-18
					revmonotone	5	0.001	31	21	-29
					monotone		0.001	88	6	-2
			F	F	monotone	3	0.002	34	27	-13
					monotone	5	0.002	30	22	-20
					revmonotone		0.001	89	6	-2
					revmonotone	3	0.002	43	25	-8
					revmonotone	5	0.002	30	23	-24
		T	T	T	monotone		0.001	47	20	-2
					monotone	3	0.001	70	12	-11
					monotone	5	0.001	64	13	-12
					revmonotone		0.001	46	21	-2
					revmonotone	3	0.001	74	8	-8
					revmonotone	5	0.001	65	10	-17
			F	F	monotone		0.002	36	7	-28
					monotone	3	0.002	47	15	-14
					monotone	5	0.002	42	16	-18
					revmonotone		0.002	36	5	-30
					revmonotone	3	0.002	44	17	-15
					revmonotone	5	0.002	37	18	-20
IPW	logreg	F	T				0.007	13	3	-130
			F				0.007	14	2	-130
		T	T				0.007	12	2	-132
			F				0.007	12	3	-131
	RF	F	T				0.008	6	1	-144
			F				0.007	7	1	-143
		T	T				0.012	1	1	-150
			F				0.012	0	1	-151

Figure S2: Column "average MSE across predictors and outcomes" reports the average MSE measure across the hospitalization, invasive ventilation, and patients' survival outcomes (the table is also made available in Supplementary file S1 – sheet "MSE_mean"). Columns "wins", "ties", "losses" report the sum of, respectively, wins, ties, and losses computed by comparing the MSE measures over the three outcomes (the corresponding win-tie-loss grid is shown in the Supplementary material). The comparison between two models over an outcome variable is performed with a sided Wilcoxon signed-rank test comparing the distribution of the MSE values for all the predictor variables. The winner is the model achieving the lowest MSE distribution. The detailed win-tie-loss grids are reported in the Supplementary material.

MI algorithm	univariate imputation method	use outcomes	one-hot encode binned numeric predictors	one-hot encode categorical predictors	univariate imputation order	pmm donors	average ER across outcomes	wins	ties	losses
amelia		F	T				0.900	0	5	-84
			F	T			0.949	4	21	-43
		T	T				0.914	6	31	-25
			F				0.971	34	23	-13
mice	default	F	F	F	monotone		0.962	6	14	-54
					revmonotone		0.964	6	14	-52
		T			monotone		0.963	8	17	-47
					revmonotone		0.962	9	12	-53
	logreg	F	T	T	monotone		0.950	5	10	-76
					revmonotone		0.950	5	7	-77
		T			monotone		0.952	16	21	-31
					revmonotone		0.952	17	25	-30
	norm	F	T	T	monotone		0.952	9	13	-67
					revmonotone		0.950	7	12	-68
			F		monotone		0.963	7	16	-53
					revmonotone		0.961	6	14	-55
		T	T		monotone		0.952	30	20	-20
					revmonotone		0.951	28	22	-20
			F		monotone		0.964	37	22	-15
					revmonotone		0.963	39	21	-15
missRanger	extratrees	F	T	T	monotone		0.952	110	6	0
					monotone	3	0.958	44	15	-23
					monotone	5	0.956	29	19	-33
					revmonotone		0.951	111	6	0
					revmonotone	3	0.958	54	18	-17
					revmonotone	5	0.955	30	20	-29
			F	F	monotone		0.955	103	6	0
					monotone	3	0.960	24	21	-19
					monotone	5	0.961	13	24	-24
					revmonotone		0.956	103	6	0
					revmonotone	3	0.959	25	21	-19
					revmonotone	5	0.960	11	26	-29
		T	T	T	monotone		1.037	45	20	0
					monotone	3	0.971	79	9	-7
					monotone	5	0.967	65	12	-14
					revmonotone		1.038	42	22	0
					revmonotone	3	0.972	82	9	-6
					revmonotone	5	0.968	60	13	-16
			F	F	monotone		1.039	12	27	-15
					monotone	3	0.979	29	24	-10
					monotone	5	0.980	24	21	-15
					revmonotone		1.038	14	25	-15
					revmonotone	3	0.980	31	21	-12
					revmonotone	5	0.979	24	21	-19
IPW	logreg	F	T				0.946	18	35	-6
			F				0.946	19	34	-6
		T	T				0.950	17	36	-6
			F				0.950	19	34	-6
	RF	F	T				0.937	4	8	-52
			F				0.946	3	5	-54
		T	T				0.835	0	3	-84
			F				0.857	0	3	-89

Figure S3: In column "average ER across outcomes" we report the average ER measure across the hospitalization, invasive ventilation, and patients' survival outcomes (the table is also made available in Supplementary file S1 – sheet "ER_mean"). Columns "wins", "ties", "losses" report the sum of, respectively, wins, ties, and losses computed by comparing the distributions of the ER measures over the three outcomes. Since we would like each $[ER_i]$ $(i \in \{1, \dots, d\})$ estimate to be as nearest as possible to 1, for the comparison between two models over an outcome variable we used a sided Wilcoxon signed-rank test to compare the following distribution for each model $f(ER_i) = \|1 - ER_i\|$. The detailed win tie loss grids are reported in the supplementary material.

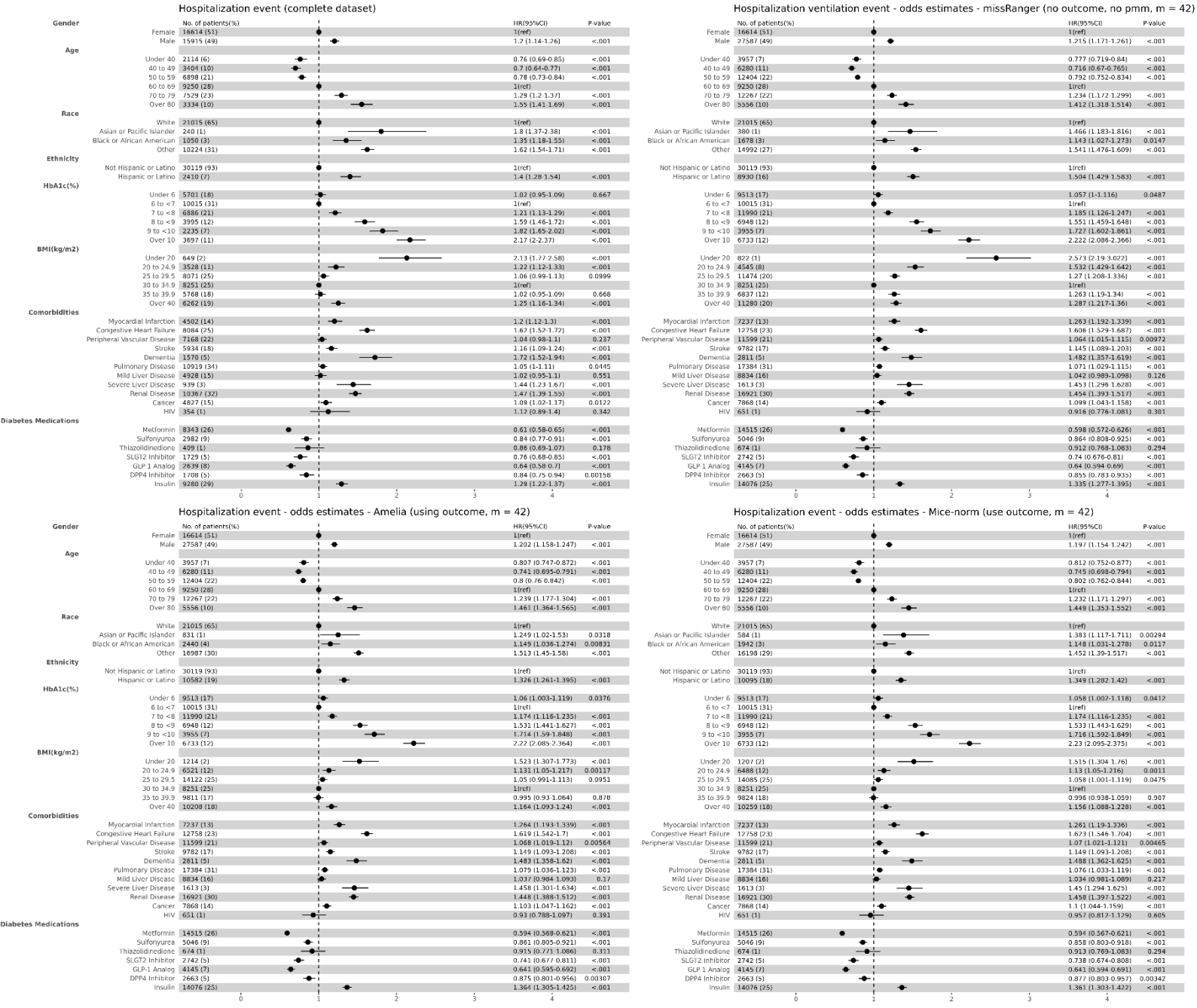

Figure S4: Hospitalization event: estimates obtained on the complete dataset obtained by listwise deletion (top-left) and on the full dataset by the MI estimation pipelines that use the best missRanger (top-right), Amelia (bottom-left), and Mice (bottom-right) models. For missRanger we used no pmm, we did not use the outcome variables in the imputation model, we one-hot encoded categorical predictors and binned numeric predictors (age, BMI, and Hba1c), and we used an univariate imputation order given by the decreasing number of missing values; for Mice-norm we included the outcome variables in the imputation model, we used an univariate imputation order given by the increasing number of missing values; for Amelia we included the outcome variables in the imputation model.

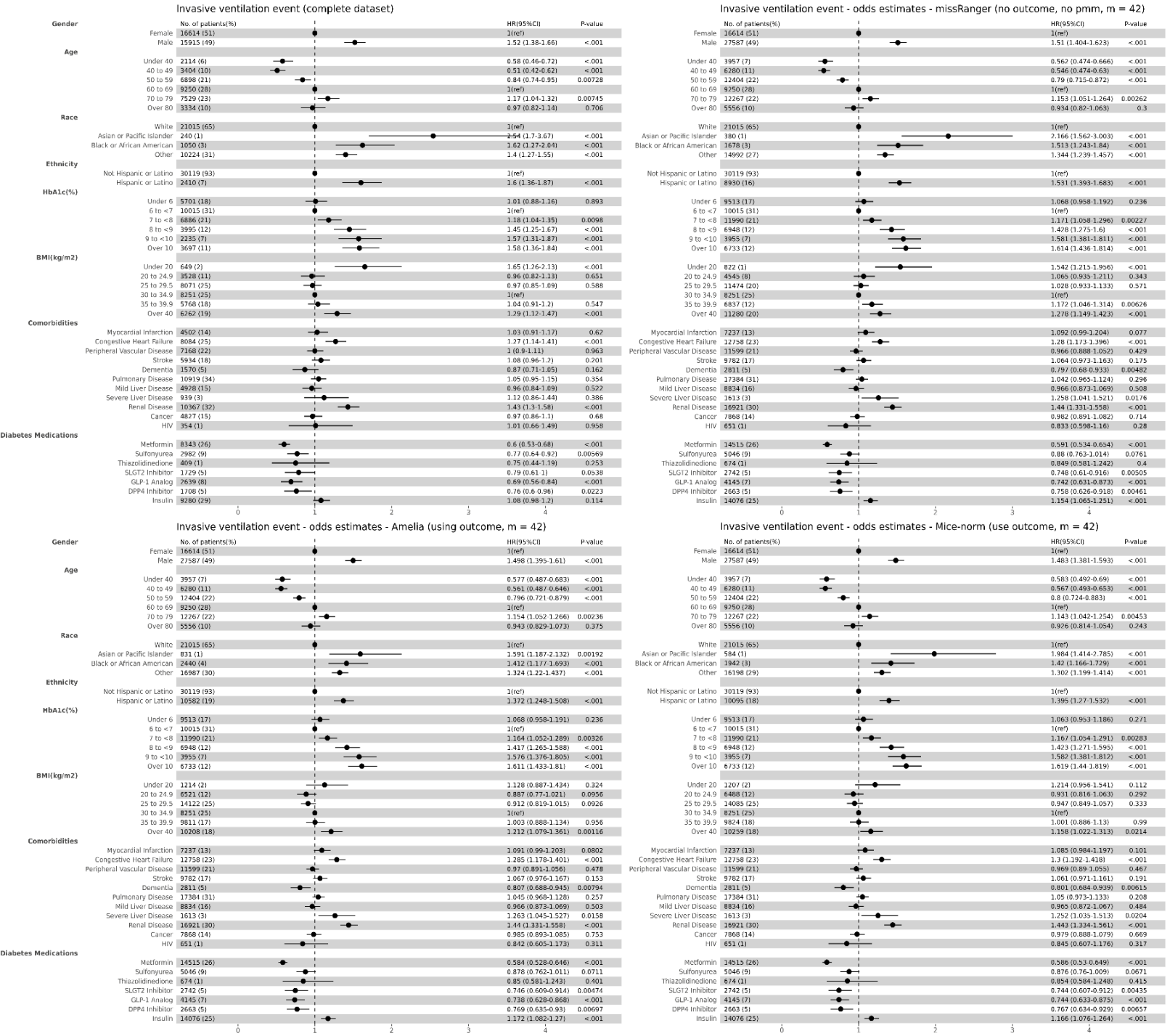

Figure S5: Invasive ventilation event: estimates obtained on the complete dataset obtained by listwise deletion (top-left) and on the full dataset by the MI estimation pipelines that using the best missRanger (top-right), Amelia (bottom-left), and Mice (bottom-right) models. For missRanger we used no pmm, we did not use the outcome variables in the imputation model, we one-hot encoded categorical predictors and binned numeric predictors (age, BMI, and Hba1c), and we used an univariate imputation order given by the decreasing number of missing values; for Mice-norm we included the outcome variables in the imputation model, we used an univariate imputation order given by the increasing number of missing values; for Amelia we included the outcome variables in the imputation model.

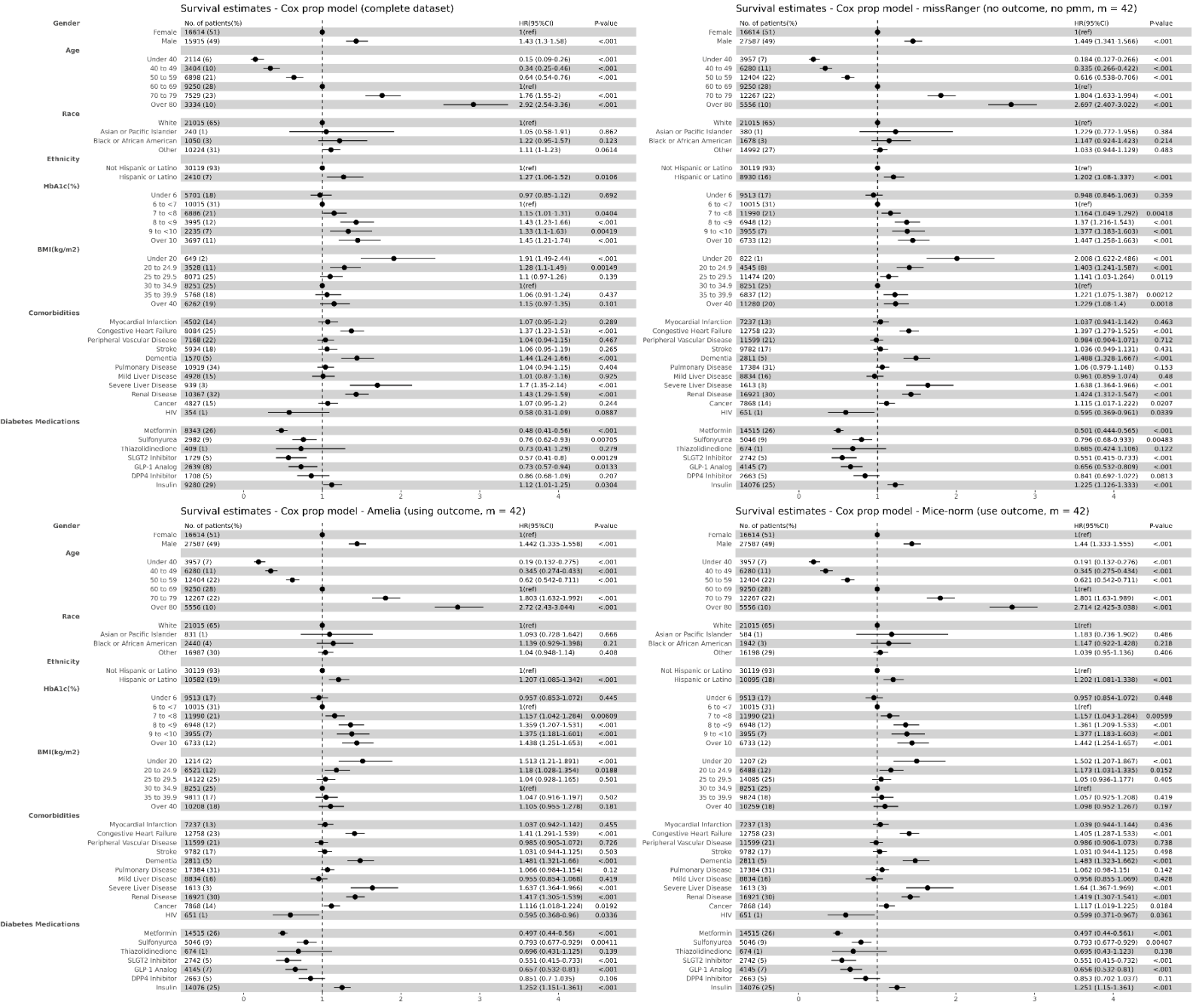

Figure S6: Death event: estimates obtained on the complete dataset obtained by listwise deletion (top-left) and on the full dataset by the MI estimation pipelines that using the best missRanger (top-right), Amelia (bottom-left), and Mice (bottom-right) models. For missRanger we used no pmm, we did not use the outcome variables in the imputation model, we one-hot encoded categorical predictors and binned numeric predictors (age, BMI, and Hba1c), and we used an univariate imputation order given by the decreasing number of missing values; for Mice-norm we included the outcome variables in the imputation model, we used an univariate imputation order given by the increasing number of missing values; for Amelia we included the outcome variables in the imputation model.