Supplementary Information

Supplemental file 1: Histograms of the number of individuals with each Dx-encoded variable each month from 2000 through 2022. Also indicated are the top 10 most prevalent codes contributing to that variable's signal, in the order of decreasing prevalence. ICD-9 and ICD-10 codes are interspersed for the purposes of ranking, and the ICD description is provided. Figures are provided as a separate pdf file with 47 pages, named CVD.pdf. These figures allow one to assess the utilization of the different codes over time and detect any discrepancies.

Supplementary Discussion

Biological Drivers of MACE

Our extensive curation and harmonization of the laboratory results has enabled us to characterize the relative likelihood of major adverse cardiovascular events (MACE) for different levels of a range of laboratory results and to compare the importance of abnormal labs and vital signs across a breadth of risk factors. Our nested retrospective case control (Rcc) study design matches year of birth and level of healthcare utilization. We also model across numerous other healthcare utilization patterns, some behaviors, and varying levels of treatment. Many risk factors are independent predictors of comparable amplitude, and we grouped them in Figure 3 for the metabolic variables and in Figure S1 for variables relating to disease progression and mental health conditions.

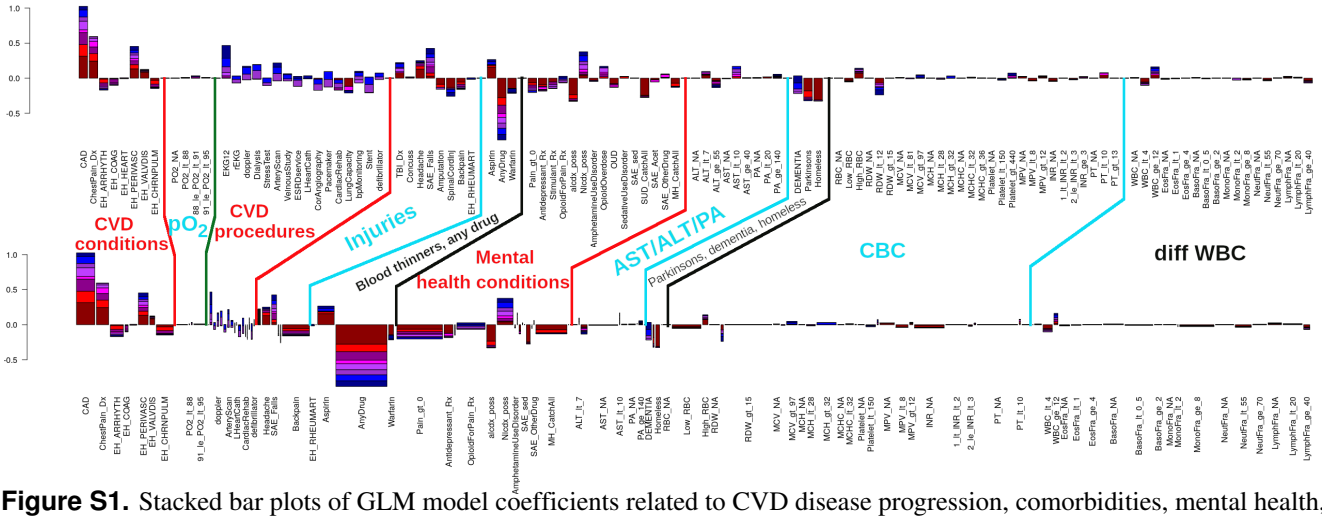

Figure S1. Stacked bar plots of GLM model coefficients related to CVD disease progression, comorbidities, mental health, complete blood count, and selected items from differential white blood counts from the **Rcc cohort**. Panels and color-scheme for bars are as defined in Figure 3, except that the y-axis scale spans a broader range to accommodate CAD and individuals who have taken any drug. Analogous plots for LDL-DS, Visit-DS, and C-17 can be found in Supplemental Figures S3, S4 and S5

Abnormal **Cholesterol levels** such as low values of high-density lipoprotein (HDL) (<40 mg/dL), high values of low-density lipoprotein (LDL) (>130 mg/dL), high values of triglycerides (>200 mg/dL), and taking high-intensity statins are all predictive of MACE events, with model coefficients of 0.1-0.2 and 5% to 25% of individuls in each group. High-intensity statin usage is 30% to 50% higher in individuls with high LDL and triglycerides, compared to average, similar to the level of enrichment to that of a coronary artery disease (CAD) diagnosis. The predictive value of the cholesterol labs is similar across study designs.

Lipid metabolism includes measures of HDL cholesterol, LDL cholesterol, total cholesterol, and triglycerides and is reviewed in the context of cardiovascular disease (CVD) by Soppert *et al.*[1]. Elevated LDL cholesterol is a major cause of heart disease, causing both atherosclerosis and inflammation. We observe a steady increase in MACE risk with greater LDL concentrations and even a slight decrease in risk from below-normal LDL levels. HDL particles promote cholesterol efflux and reduce inflammation. HDL metabolism is more complex, with lower MACE risk associated with high HDL levels, as long as both triglycerides and LDL cholesterol are not elevated. As seen in Figure 3, very high levels of HDL cholesterol are predictive of MACE events, as reported by others[2]. Two studies carefully examined the probability of MACE outcomes on lipid levels[2,3] and a comparison of their Figure 1 to our coefficients characterizing lipid dependence shows striking agreement on the non-linear (U-shaped) effect of HDL and LDL on MACE outcomes. As we found, elevated triglycerides were also linked to elevated atherosclerotic cardiovascular disease (ASCVD) risk[1,4].

Stage 2 **hypertension**, defined as having a median systolic blood pressure measurement during a given time window greater than 140 mmHg, is one of the strongest MACE risk factors (Figure 3), with a coefficient of 0.43 in Rcc and 0.53 in calendar 2017 (C17) (Table S1), and a prevalence of 59% (Table S2). It has long been known to be among the important CVD risk

Conditions	% tot	C.C17	C.V	C.Rcc	OR.Rcc	Cases	Controls	%Cas/%Cnt	Enr.AMI	Enr.S	Enr.dth
Total						1,078,498	1,648,330		31.5%	26.7%	7.2%
Female	3.4	-0.31	-0.34	-0.47	0.62	33102	59619	0.7	0.62	0.82	0.41
Black	11.5	0.13	0	0.01	1.01	151916	160427	0.97	0.93	1.22	0.64
Hispanic	3.6	0	-0.07	-0.07	0.93	45140	51951	0.93	0.95	1.05	0.65
Alcohol depen.	11.4	0	0.02	-0.33	0.72	140322	171104	0.89	0.9	1.01	0.73
Nicotine Dx	20.6	0.47	0.41	0.38	1.46	285800	275297	1.03	1.11	1.12	0.97
OBESITY Dx	23.3	-0.08	-0.13	-0.37	0.69	291662	343753	0.88	0.98	0.95	0.8
CAD Dx	23.8	0.87	0.58	1.03	2.79	393600	254971	1.35	1.52	1.18	1.98
ChestPain Dx	15.9	0.47	0.16	0.59	1.81	226909	205638	1.06	1.22	1.11	0.73
Hypertension 2	43.3	0.53	0.33	0.43	1.53	598373	581602	1.04	1.04	1.14	1.32
PulsePr $\geq$ 90 mmHg	6.5	0.06	0.22	0.17	1.18	100079	78377	1.14	1.17	1.31	1.9
HYPERTENS Dx	54.8	0.15	0.1	-0.02	0.98	749502	745751	1.01	1.03	1.1	1.32
Aspirin	21.4	0.16	0.26	0.27	1.31	322468	262152	1.15	1.21	1.29	1.17
ACEInhibitors	32.8	0.07	0.07	0.11	1.12	469526	423910	1.1	1.13	1.15	1.38
AntiAnginals	11.7	0.16	0.22	0.34	1.4	197898	119921	1.31	1.6	1.2	1.93
BetaB 1st gen	31.2	0.19	0.17	0.27	1.31	470978	380469	1.2	1.26	1.18	1.49
Na $<$ 134	6.2	0.1	0.22	0.14	1.15	87795	80740	1.05	1.08	1.18	1.41
K $<$ 3.5	4.6	0.15	0.13	0.23	1.25	61277	63932	0.98	0.96	1.15	1.04
Cl $\leq$ 92 mEq/L	0.9	0.06	0.29	0.16	1.17	13449	11441	1.08	1.11	1.18	2.07
TotChol $\geq$ 220 mg/dl	6.5	0.17	0.12	0.21	1.24	89184	86983	1.01	1.09	1.11	1.05
LDL 130-160 mg/dl	9.9	0.06	0.06	0.06	1.06	128165	140553	0.95	0.98	1.05	0.98
LDL 160-190 mg/dl	3.7	0.16	0.15	0.16	1.17	51600	50506	1.01	1.08	1.12	1
LDL $\geq$ 190 mg/dl	1.4	0.19	0.22	0.24	1.27	20351	17725	1.07	1.18	1.2	1.06
Trig $\geq$ 200 mg/dl	24.1	0.15	0.12	0.15	1.16	329803	327522	1.01	1.09	1.07	1.04
Statin High	20.8	0.02	0.06	0.14	1.15	307052	260463	1.12	1.2	1.17	1.4
Statin Medium	23.4	-0.08	-0.03	-0.04	0.96	317539	319767	0.99	1.04	1.08	1.3
UNCDIAB Dx	26.9	0.02	0.01	-0.15	0.86	375857	357250	1.04	1.08	1.13	1.29
COMDIAB Dx	13.7	0.12	0.07	-0.01	0.99	199280	172997	1.09	1.17	1.21	1.18
Metformin	16.4	-0.02	-0.02	-0.06	0.94	227720	218833	1.03	1.09	1.14	0.99
Sensitizers	13.6	0	0	-0.06	0.94	196330	175201	1.07	1.14	1.17	1.35
Insulin	9.7	0.02	0.05	0.02	1.02	148362	116546	1.14	1.25	1.28	1.36
HbA1c 8-9%	8.4	0.16	0.16	0.16	1.17	125231	105006	1.1	1.2	1.24	1.19
HbA1c $\geq$ 9%	6.5	0.63	0.47	0.52	1.67	102192	73862	1.18	1.31	1.38	1.23
Glucose $>$ 150 mg/dl	18.6	0.28	0.26	0.21	1.23	269424	238577	1.08	1.15	1.19	1.25
RR NA	5.7	0	0	0.45	1.57	112236	42472	1.51	1.76	1.87	3.94
AnyDrug	66.5	0.04	-0.07	-0.88	0.41	869836	942797	0.69	0.98	1.03	1.2
EKG12	2.7	0.27	0.22	0.47	1.6	38576	34394	1.06	1.24	1.13	0.62
Troponin NA	65.7	-0.25	-0.07	-0.21	0.81	879744	912713	0.85	0.99	1.07	1.41
Troponin $>$ 0.4	0.3	-0.08	-0.01	0.07	1.07	5829	2926	1.33	1.83	1.23	1.68
WBC $\geq 12 \times 10^9$/L	5.1	0.13	0.21	0.16	1.18	72137	67169	1.04	1.14	1.1	1.23
EEstradiol	0.1	-0.2	-0.3	-0.22	0.8	849	2720	0.48	0.38	0.56	0.05

Table S1. Attributes of selected predictor variables. For each of 40 variables shown, we provide 11 columns of information, including the percentage of total individuals in each category (% tot), logistic regression model coefficients for C17 (C.C17), Visit-DS (C.V), and Rcc (C.Rcc) cohorts, the calculated odds ratio for Rcc (exp($C.Rcc$) in OR.Rcc), number of cases and controls in the Rcc cohort, the percentage of cases divided by the percentage of controls for each variable, and an enrichment factor for each element of our combined outcome (Enr.AMI, Enr.Stroke, and Enr.dth), computed as the number with the respective outcome divided by the number of controls for each variable, in turn, divided by that quantity for all individuals.

factors, with the stresses on the arteries causing a similar type of damage to atherosclerosis, but in a complementary manner[5]. It is a stronger risk factor in younger adults, and the risk rises steeply for those with a systolic blood pressure of >140 mmHg (stage 2 hypertension)[6], as we saw in Figures 3 and S8. While systolic blood pressure continues to increase with age, diastolic tends to decrease, and our observation of high pulse pressure being an independent MACE predictor is consistent with the existing literature[7]. High pulse pressure is only a strong predictor at higher levels (≥ 90 mmHg), where it is present in only 6.5% of individuals. We found that hypotension, although present only in a small minority of individuals, was also a strong predictor of MACE events. This observation has been controversial for some time, especially in terms of the aggressiveness of therapy to reduce high blood pressure[8,9]. A meta-analysis by Bohm *et al.*[10] showed that a target blood pressure of 120–130 mm Hg systolic and 70–80 mm Hg diastolic is associated with the lowest rate of CVD events. We also observed a high pulse/heart rate to be independently predictive of MACE events, in agreement with Sun *et al.*[11], although we did not observe elevated risk for low heart rates.

Rcc	overall p	Overweig	Obese	EH_Obes	Angioten	AntiAngi	BetaBloc	BetaBlo	EH_Uncor	EH_Comp	HbA1c 6	HbA1c 7	HbA1c 8	HbA1c >9	Glucose 1	Glucose >	eGFR 60-(	eGFR 45-(	eGFR 30-	Metform	Sensitize	Insulin	CAD	ChestPair	HDL < 40	LDL 160-1	LDL > 19	Trig >=20	Statin Hi	Statin Mi	Statin Lo
overall	100	49.6	46.2	31.8	12.3	15.9	42.6	7.1	36.7	18.4	30.5	18.3	11.5	8.8	30.1	25.4	65.5	24.4	11.3	22.4	18.6	13.2	32.5	21.6	33.3	5.1	1.9	32.9	28.4	31.9	4.4
Female	4.7	44	53	44	12	10	33	4	27	12	25	12	8	7	21	17	58	22	9	19	12	9	15	28	18	9	4	30	20	37	4
Black	15.6	49	47	35	14	13	52	10	42	22	36	20	13	13	31	28	53	20	10	25	20	17	24	27	29	6	2	24	26	39	3
Hispanic	4.9	53	46	38	13	15	39	6	46	25	34	23	16	14	35	32	59	22	11	30	25	18	27	25	40	3	1	38	28	36	5
Homeless	6.6	56	46	36	9	14	42	8	32	17	29	14	11	12	28	25	61	17	8	22	16	14	22	36	37	8	3	37	20	28	4
Alcohol	15.6	56	45	35	10	14	47	7	41	16	29	14	9	8	30	22	62	19	8	20	25	11	26	32	33	7	3	39	26	31	4
Nicotine	28.1	55	45	34	10	18	47	8	46	18	33	17	12	10	32	26	61	21	9	24	18	14	32	30	38	7	3	41	31	35	4
Overweigh	49.6		35	24	13	17	46	7	46	18	32	17	10	8	31	25	60	27	13	21	18	12	26	23	36	6	2	35	31	36	5
Obese	46.2	37		59	17	19	51	9	51	27	40	27	18	14	40	36	61	27	13	34	27	20	27	26	42	6	2	45	35	37	5
EH_Obese	31.8	38	85		18	20	53	9	55	31	44	30	20	15	43	39	64	28	13	38	30	23	27	30	44	6	2	47	37	39	5
Hypertens	59	53	52	36	16	18	51	8	43	23	35	22	14	11	36	31	59	29	14	26	23	16	25	24	37	6	2	38	32	36	5
EH_Hypert	74.8	53	53	37	16	20	53	9	45	24	37	23	15	11	36	31	59	29	14	28	24	17	29	25	38	5	2	38	34	38	5
EH_Uncon	36.7	49	64	48	20	23	56	11		49	58	49	31	24	61	63	59	32	17	60	50	36	44	27	44	4	2	47	42	42	6
EH_Comp.	18.4	48	68	53	25	26	62	14	97		56	58	42	32	65	76	57	37	22	66	60	52	47	31	47	4	2	50	44	43	6
HbA1c 8	11.5	45	71	55	24	26	61	13	99	68	46	75		49	67	95	59	36	21	76	73	65	46	31	50	4	2	57	48	45	6
HbA1c >9	8.8	45	71	56	23	24	60	14	99	68	36	61	64		58	97	59	34	19	78	73	72	43	32	50	6	3	60	48	44	6
Glucose 12	30.1	51	62	46	19	22	57	10	74	40	58	45	26	17		55	61	33	17	50	42	28	42	28	44	5	2	47	40	42	6
Glucose >1	25.4	48	66	49	22	23	58	12	91	56	54	60	43	34	65		59	34	18	65	60	47	44	28	46	5	2	53	43	42	6
CAD	32.5	54	53	36	18	41	72	15	49	27	39	26	16	12	39	34	57	33	17	29	26	20		34	41	4	2	39	46	42	6
ChestPain	21.6	43	56	44	17	37	63	12	46	27	39	24	16	13	39	33	64	28	13	31	25	20	51		41	7	3	43	38	41	6
HDS < 40	33.3	54	58	42	16	21	53	9	48	26	40	27	17	13	40	35	61	30	15	31	27	19	40	27		7	2	49	35	40	6
LDL 160-19	5.1	57	52	37	11	16	42	7	32	15	30	15	10	10	28	23	67	24	10	21	16	12	28	28	43		16	48	41	44	6
LDL > 190	1.9	58	53	37	12	18	45	7	35	17	31	16	12	12	29	26	66	25	11	23	18	14	31	29	41	45		53	48	45	6
Trig >=200	32.9	52	63	46	16	21	54	9	52	28	41	29	20	16	43	41	62	30	14	36	30	22	38	29	50	7	3		39	40	5
StressTest	3.8	53	53	40	15	24	56	8	47	27	37	25	16	12	39	34	60	27	13	29	27	20	40	35	40	6	2	41	39	45	7

C17	overall p	Overweig	Obese	EH_Obes	Angioten	AntiAngi	BetaBloc	BetaBlo	EH_Uncor	EH_Comp	HbA1c 6	HbA1c 7	HbA1c 8	HbA1c >9	Glucose 1	Glucose >	eGFR 60-(	eGFR 45-(	eGFR 30-	Metform	Sensitize	Insulin	CAD	ChestPair	HDL < 40	LDL 160-1	LDL > 19	Trig >=20	Statin Hi	Statin Mi	Statin Lo
overall	100	55	56	40	11	9	33	5	33	16	31	17	11	8	28	23	66	22	9	22	15	11	21	21	34	6	2	36	23	25	3
Female	11	42	49	39	6	4	19	2	15	6	16	6	4	4	13	9	54	13	4	11	6	4	5	19	14	7	3	20	10	14	2
Black	20	49	54	41	12	7	30	6	34	17	35	16	11	10	25	22	56	16	7	22	15	12	14	24	26	6	2	23	20	22	2
Hispanic	7	52	53	44	9	7	24	3	33	16	27	16	11	10	26	23	62	16	7	23	16	12	14	20	37	4	1	38	20	25	3
Homeless	9	55	49	39	7	8	31	5	25	12	26	11	8	9	24	19	61	14	6	18	11	10	13	30	34	7	3	36	15	21	3
Alcohol	19	57	49	38	9	8	35	5	25	12	26	11	7	7	25	18	64	15	6	17	11	8	16	27	31	7	3	39	20	24	3
Nicotine	31	55	49	37	8	10	34	5	29	14	29	14	9	8	27	21	65	17	7	20	13	10	19	26	36	7	3	40	23	25	3
Overweigh	55		36	27	10	9	31	5	28	13	27	13	8	6	24	18	65	22	9	17	12	9	21	20	32	6	2	33	22	25	3
Obese	56	35		61	13	10	36	6	41	21	36	22	14	11	33	29	65	22	9	28	19	15	22	22	38	6	2	43	25	26	3
EH_Obese	40	37	85		14	11	38	6	45	24	39	24	16	12	36	32	68	23	9	32	22	17	23	26	40	6	2	45	28	29	3
Hypertens	59	52	57	41	15	10	39	6	38	20	34	20	13	10	32	27	65	25	10	25	17	14	23	22	34	6	2	38	26	28	3
EH_Hypert	71	51	59	43	15	12	43	7	42	22	37	22	14	10	34	29	66	27	11	28	19	15	27	23	36	6	2	39	29	30	3
EH_Uncon	33	46	69	54	19	15	46	9		48	60	49	32	24	61	62	65	30	14	63	43	34	32	25	42	4	2	48	37	35	4
EH_Comp.	16	45	71	57	23	18	53	12	97		57	58	43	33	65	75	63	35	18	69	55	50	37	29	46	4	2	51	43	38	4
HbA1c 8	11	42	75	60	23	17	51	11	99	66	46	76		50	68	95	65	34	17	79	68	63	34	28	48	4	2	57	45	39	4
HbA1c >9	8	42	75	61	21	16	49	11	99	66	38	62	66		59	98	63	31	16	82	69	70	31	29	48	6	2	60	43	38	4
Glucose 12	28	48	66	51	18	14	46	8	72	38	59	44	26	17		54	67	30	13	51	36	26	30	26	42	5	2	48	35	34	4
Glucose >1	23	45	71	55	20	15	48	10	91	54	55	61	45	35	66		65	31	15	69	54	45	32	26	45	5	2	54	38	36	4
CAD	21	53	57	42	18	34	66	15	49	28	42	26	17	12	39	34	66	33	15	31	23	20		36	40	4	2	39	43	33	3
ChestPain	21	52	59	48	15	25	50	9	40	23	36	21	14	11	34	28	69	24	9	28	19	16	36		38	7	2	43	31	32	3
HDL < 40	34	51	63	47	13	12	38	7	41	22	37	23	15	11	35	30	66	25	11	29	20	15	25	24		7	2	50	24	29	3
LDL 160-19	6	56	56	41	9	7	29	4	24	11	26	11	7	8	23	17	71	18	6	17	10	8	15	24	39		17	48	28	32	3
LDL > 190	2	57	57	42	10	8	31	5	27	13	28	13	9	9	24	19	71	20	8	19	12	10	17	25	39	50		54	36	34	3
Trig >=200	36	50	67	49	13	11	39	6	44	23	37	24	17	13	37	34	68	25	10	32	22	17	23	25	47	8	3		29	30	3
StressTest	1	48	63	53	18	21	54	11	58	42	43	32	23	17	44	42	67	30	14	38	30	27	37	46	38	5	2	41	36	36	4

Table S2. Co-occurrence of selected variables in the Rcc (top) and the C17 (bottom) cohorts. Each row defines a subgroup of individuals, where the value in each column is the percentage of that subgroup with the indicated attribute. For example, 12% of homeless individuals have at least one time window with a median HbA1c level above 9%, compared to 8.8% of all individuals in the Rcc cohort. Note that Rcc is enriched to 50% cases and is matched on year of birth. This information helps us to understand which interaction terms among variables may be important to MACE risk prediction, as indicated in the Results section. The first row and column provide the percent of indicated variables with each attribute, so that relative enrichment can be determined. The overlap percentage of each variable with itself is 100%, but this is left blank to guide the eye. Shading and highlighting of rows and columns identifies related variables, and is done to guide the eye.

Each of the major classes of blood pressure medications are themselves strong and independent predictors of MACE events. Examination of the joint probability distributions in Table S2 shows, however, that the likelihood of having a recorded pharmacy fill for blood pressure medications is only 30% elevated in subjects with stage 2 hypertension, perhaps indicating that the medications are effective in reducing blood pressure below 140 mmHg. While less prevalent than beta blockers, antianginal medications are both stronger MACE predictors and more acute (higher contribution of the dark red shaded area at the bottom of the bar in Figure 3).

Decreased renal function, measured by reduced eGFR and increased urinary albumin-to-creatinine ratio were shown to be independent predictors of MACE events in type-2 diabetics in the ADVANCE study with a 2.2 hazard ratio for cardiovascular events for every halving of eGFR[12]. As seen in Figure 3, reduced eGFR is an acute MACE predictor, and MACE risk increases at eGFR levels of 60 and below. Although our coefficients in Figure 3 are small (0.1-0.2), 44% of the individuals in Rcc with an eGFR between 45 and 60 also had a diagnosis of CAD, with a model coefficient of 1.0. A follow-up retrospective study by Currie, *et al.*[13] reported that the MACE risk for a given level of eGFR was 2.4-4.6 times higher in the presence of type-2 diabetes. The importance of understanding the interdependencies of risk factors is evident in meta-analyses of cardiovascular outcome trials with SGLT-2 inhibitors to reduce CVD events and mortality[14,15]. Understanding these is even more important in type-2 diabetes individuals with impaired renal function, where reductions ranging from 7% to 20% beyond the standard of care were achieved, but the relevant differences among trials were still unclear[14,15].

Electrolyte imbalances, especially low sodium, potassium, and chloride, are present in only a few percent of individuals but have coefficients in the 0.14-0.23 range in the Rcc study. High bicarbonate levels, which are associated with acid-base imbalance, are much more prevalent than the other imbalances and slightly protective. Declining **renal function** in each of three categories of eGFR between 60 and 15 ml/min/1.73m^2 of body surface area, is acutely predictive of MACE events. The 30% of individuals with eGFR $<$ 60 ml/min/1.73m^2 are 15% more likely to have stage 2 hypertension. From Table S2, we see that they are also 50% more likely than the average patient to be taking first-generation beta blockers.

Low sodium[16], potassium and magnesium[17] levels are associated with poor outcomes in individuals with CVD. Community prevalence and the distinct risk factors of the various electrolyte disorders are discussed by Liamis *et al.*[18]. Numerous classes of medications are available to treat hypertension, and electrolyte imbalance is one of several important side effects considered in the choice of medication[19]. The protective effects of low serum calcium levels with respect to MACE outcomes that we observe may be related to the direct appearance of Ca^{2+} ions in the arterial plaques, promoting both stroke[20] and acute myocardial infarction (AMI)[21].

The contribution of **diabetes** risk to MACE events is multi-factorial and indicated by multiple diagnostic codes, laboratory results, and medications. Similar to the other categories of MACE risk factors, recently diagnosed diabetes and recently started diabetes treatments are protective, while abnormal laboratory results are predictive of poor outcomes, with the highest levels (e.g., HbA1c) showing the strongest predictive coefficient. Diabetes has been reported to increase women's chronic heart disease risk by a factor of approximately 2.8, and men's by 2.2[22]. Although the presence of diabetes is highly correlated with many other predictor variables, mechanistically, it augments microvascular and macrovascular risk and is considered to be an independent predictor of many vascular disease outcomes[23]. The risk coefficients in Figure 3 show that both high HbA1c and high glucose are predictive of MACE events and that the effect is positively associated with blood glucose levels. Only rarely is hypoglycemia seen, but when it occurs, it is strongly predictive of MACE events. Both diabetes diagnostic codes and medications protect against MACE events, and our interpretation is that this is because diagnosis and treatment are highly correlated. Our results are consistent with the observation of Galbete *et al.*, who noted in a review of 19 studies of CVD risk in populations with diabetes and 46 studies of the general population that many studies have a high risk of bias due to methodological shortcomings and scarcity of independent validations[24].

Although **body mass index (BMI)** is a well-established risk factor for CVD[25–30], we see here that it is protective as a model coefficient (bars pointing down), whether characterized as a diagnosis code (EH_OBESITY; 278.0 or E66.x) or a vital sign. The observation that being overweight or obese can improve chances of survival is known as the obesity paradox[31]. While the association of obesity with the metabolic syndrome[32] and its co-linearity with other MACE risk factors certainly reduces our observed model coefficients, this effect is still observed in an univariate analysis and across multiple study designs. Several effects appear relevant to this point. First, in agreement with Park *et al.*[33], we observe that being underweight (BMI $< 18.5\ kg/m^2$) is a risk factor for CVD, perhaps due to poor nutritional status and the relatively high age of our cohort (median age of C17 is 65). Second, BMI, as a predictive variable, ignores important issues of muscle mass and the distribution of adipose tissue that strongly impacts the prognosis associated with obesity[30]. Finally, selection biases are likely occurring because we restricted our analysis to the first recorded MACE event. The observation that obesity has a much larger model coefficient in Rcc than C17 supports this supposition. Although the coefficient for the Dx-code-based variable is more than double that of the vital signs, it is important to recall from the Methods section that vital signs were imputed across time bins while the Dx-code-based variables were not. This has the effect of increasing the impact of vital-sign-based coefficients in comparison to Dx- or medication-based. Also, 85% of the individuals with a Dx code for obesity have at least one record of

BMI > 30, while 60% of those with a BMI > 30 have a diagnosis code of obesity.

Figure S1 is constructed in a manner similar to Figure 3, but characterizes **predictor variables not directly related to metabolic syndrome**, including specific CVD conditions and procedures, medications, mental health conditions, injuries and other laboratory assays. The strongest predictors in this figure are diagnoses of CAD or chest pains, with coefficients of 1.03 and 0.59, respectively. The most protective variable is a measure of healthcare utilization: whether the patient had taken any drug during the 7.5 years prior to the time of prediction.

Analysis of **complete blood counts (CBC)** identified a number of statistical associations with MACE outcome, including liver enzymes, platelets, elevated white blood cell (WBC) counts, red blood cell (RBC) counts, and red cell distribution width (RDW). These conditions are statistically significant but comprise a much smaller population-attributable risk than the other variables discussed, and we consider them no further in this work.

Several **mental health conditions** correlate with MACE outcomes and are generally protective. This may be because CVD and mental health care are two of the largest categories of care sought in the US Department of Veterans Healthcare Administration (VHA), and since we are not comparing them to healthy controls, mental health is protective against MACE events. A notable contrast is nicotine dependence, which is strongly predictive of MACE events with a coefficient of $\sim$0.4 in all study designs, and with a prevalence $> 20\%$ in Rcc (see Table S1). A diagnosis of alcohol use disorder is protective (-0.33 in Rcc and neutral in C-17). Aspirin is strongly predictive, presumably because it is prescribed to at-risk individuals.

Additional Supplementary Tables and Figures

	Rcc		LDL-DS		Visit-DS		C-17	
	Control	**Cases**	**Control**	**Cases**	**Control**	**Cases**	**Control**	**Cases**
Total	1,648,330	1,078,498 (39.6%)	2,340,480	391,803 (14.3%)	3,585,628	548,554 (13.2%)	3,491,024	210,084 (5.7%)
AMI		519,225		186,274		255,752		109,284
Stroke		440,035		166,278		241,672		92,514
ASCVD death		119,238		38,058		51,120		8,286
Female	98,263	35,713 (26.7%)	183,353	12,900 (6.6%)	222,540	15,817 (6.6%)	384,516	9,045 (2.1%)
Black	264,330	163,829 (38.3%)	368,813	66,450 (15.3%)	511,674	87,694 (14.6%)	629,516	40,916 (6.1%)
Hispanic	133,052	48,670 (36.2%)	133,052	22,673 (14.6%)	175,052	25,684 (12.8%)	56,892	11,818 (4.9%)
Age (median)	69.3	69.8	60.1	63.1	62.7	64.6	64.2	69.2
Female	58.9	61.2	47.8	55.3	49.4	56.8	49.4	60.1
Black	63.7	64.7	53.9	59.0	55.5	60.5	58.1	65.2
Hispanic	66.8	69.3	57.6	64.0	58.5	64.3	56.7	69.2
Youngest quartile	56.1	56.9	42.2	47.7	44.8	50.4	40.2	47.9
Oldest quartile	84.9	85.2	79.6	79.0	81.0	80.5	79.9	80.1
Computed dataset	1,000,000	1,000,000 (50%)	1,568,000	386,658 (19.8%)	1,001,838	331,691 (24.9%)	872,756	210,084 (19.4%)

Table S3. Selected characteristics of each of the four study populations, broken down by cases and controls. The first row provides the total number of cases and controls in each study, together with the percentage of cases in each group, which drops from 39.6 % in the enriched Rcc study, to the 5.7% in C-17, which has only 5 years of followup. The next 3 rows provides the breakdown of the combined MACE outcome into AMI, stroke and ASCVD death. The breakdown of component outcomes is roughly 48:42:10 for all studies except C-17, which only has 4% ASCVD deaths, because our NDI data reporting ends after only 2 of the 5 years of followup, at the end of 2018. Below this are case and control breakdowns for three important and overlapping subgroups, females, Blacks, and Hispanics. From the reported percentage with MACE events for each group, we see that women have a much lower rate of events than overall and than each of the other subgroups. Some of this difference is determined by the age distribution, which is characterized in the bottom portion of the table.

Method \ Data set	Rcc	LDL-ds 10yr	Visit-ds 10yr	C-17 5 yr
AHA/PCE-like*	0.580	0.662	0.610	0.705
Rcc - glm	0.686	0.607	0.610	0.620
fine tuned	0.685	0.699	0.687	0.731
Rcc - Neural Net	0.723	0.645	0.635	0.658
fine tuned	0.723	0.701	0.688	0.739
Rcc - RF	0.712	0.637	0.642	0.657
fine tuned	0.714	0.701	0.692	0.740
Rcc - TabNet	0.713	0.640	0.637	0.654
fine tuned	0.713	0.701	0.692	0.744
LDL-ds	0.547	0.760	0.742	0.744
fine tuned	0.652	0.776	0.755	0.745
Visit-ds	0.552	0.732	0.719	0.636
fine tuned	0.654	0.747	0.732	0.709
C-17-glm	0.546	0.722	0.722	0.751
fine tuned	0.654	0.744	0.739	0.751
C-17-NN	0.579	0.678	0.672	0.759
fine tuned	0.662	0.693	0.687	0.760
C-17-RF	0.545	0.680	0.677	0.763
fine tuned	0.648	0.694	0.689	0.768
C-17-TabNet	0.556	0.683	0.662	0.756
fine tuned	0.648	0.698	0.680	0.756

Table S4. C-statistics for MACE prediction methodologies using our four study designs, including assessing the transferability of models from one study design to another. The first row provides C-statistics for an AHA-PCE-like model for the indicated study designs, where models were trained on half the data and evaluated on the other half for each study design. For Rcc and C-17, models were trained with four prediction methodologies (GLM, NN, RF, TabNet), and evaluated in each of the study designs. Within each table cell, the top number indicates C-statistic with the trained model used directly on each study design, while the lower number is the C-statistic with each model score included together with Gender, Race, Ethnicity, Age, Age^2, Pain gt 0, AnyDrug, HbA1c gt 9%, DEMENTIA, EH HYPERTENSION in a glm fine tuning. Typically, this re-optimization improves the score by less than two percentage points, but for the case of Rcc models used to predict on the non-age-matched models, model performance increases by 8-10 percentage points. About half the improvement comes from the age variable. For our time-to-event models (LDL-DS and Visit-DS), scores were evaluated as interval predictions at 10 years past the time of prediction.

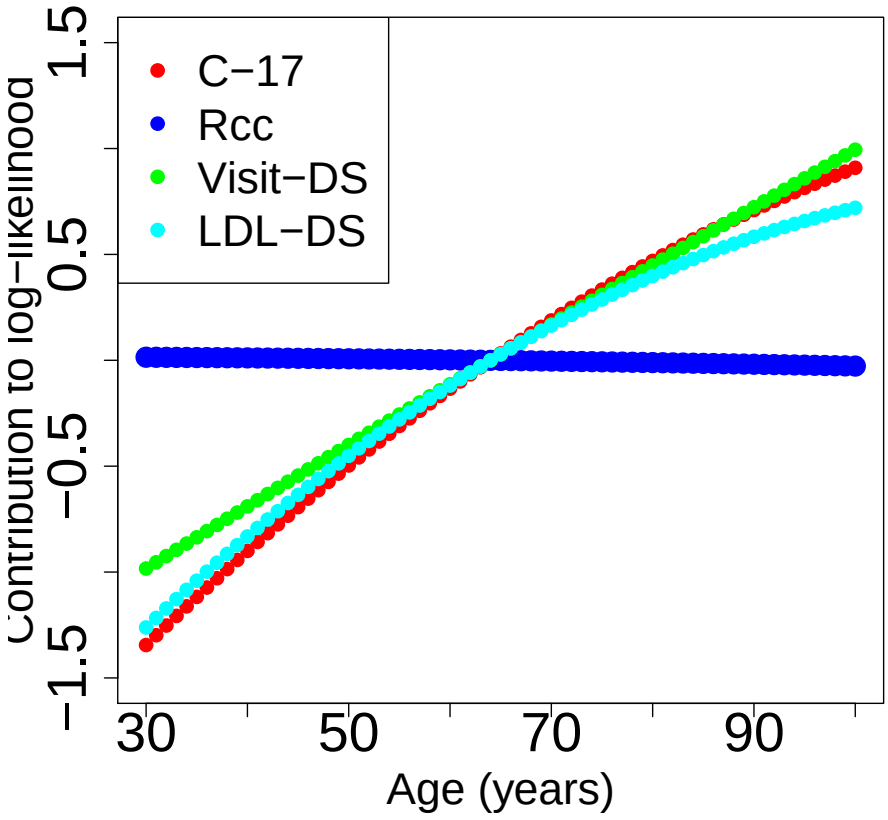

Figure S2. Age dependent contribution to log likelihood of MACE events from age and age^2 terms of the generalized linear models for each of the four study designs. Need to plot Rcc with age-dependence from fine tuning.

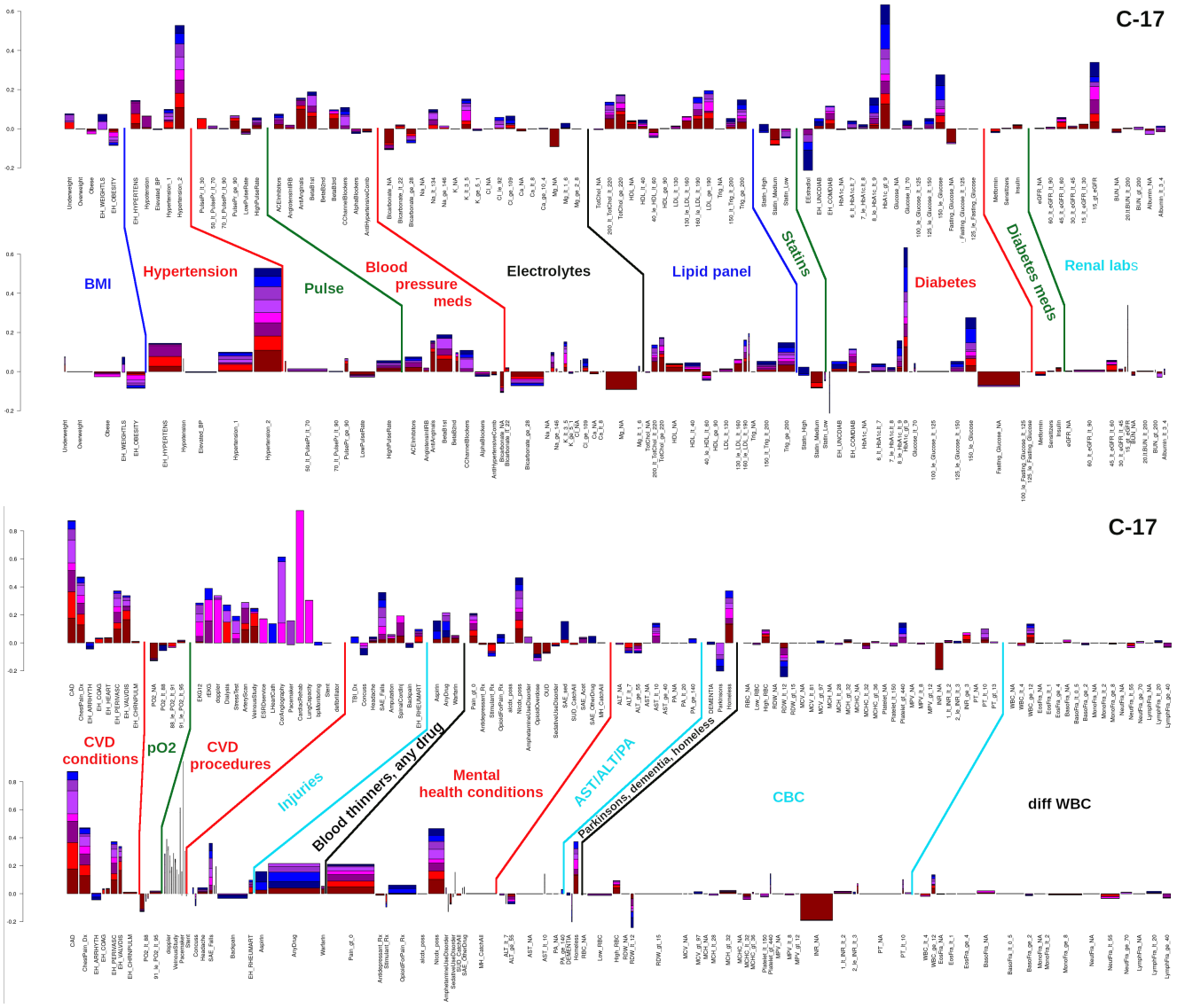

Figure S3. Stacked bar plots of model coefficients related to (top) metabolic syndrome, and (bottom) CVD disease progression, comorbidities, mental health, CBC, and dWBC, from the GLM model and C-17 cohort. Descriptions are the same as for Figs 3 and S1

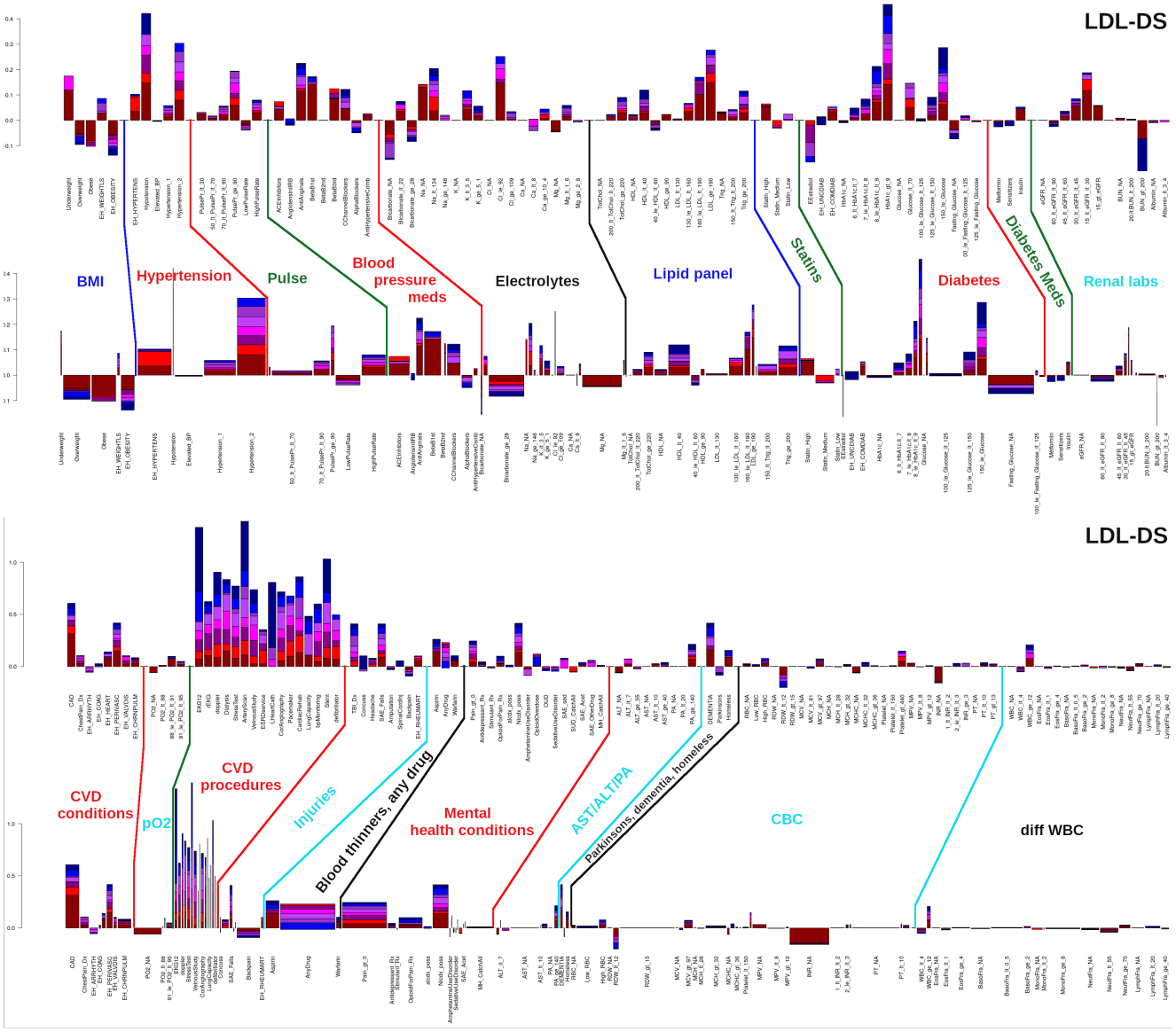

Figure S4. Stacked bar plots of model coefficients related to (top) metabolic syndrome, and (bottom) CVD disease progression, comorbidities, mental health, CBC, and dWBC, from the GLM model and LDL-DS cohort. Descriptions are the same as for Figs 3 and S1

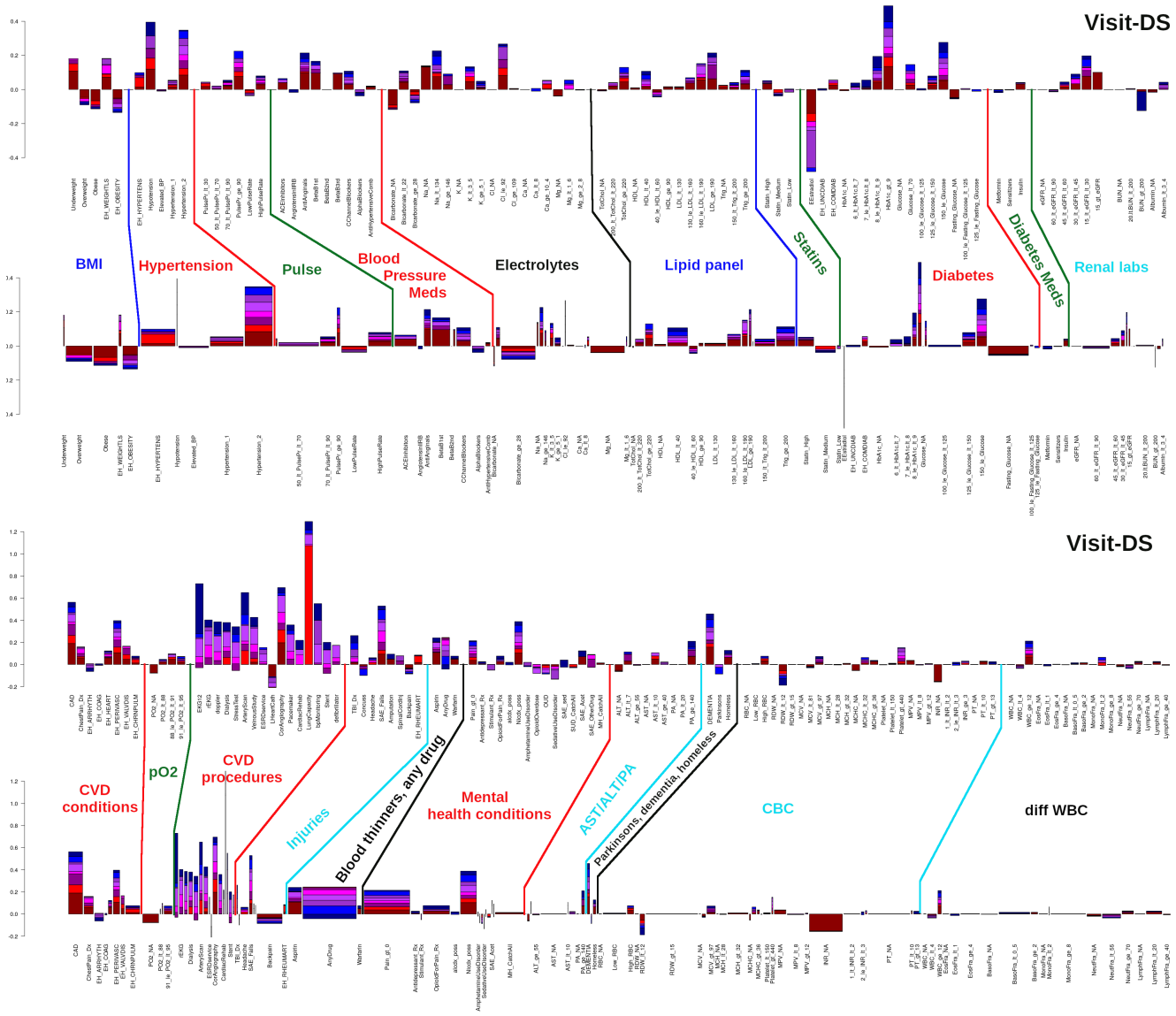

Figure S5. Stacked bar plots of model coefficients related to (top) metabolic syndrome, and (bottom) CVD disease progression, comorbidities, mental health, CBC, and dWBC, from the GLM model for **Visit-DS** cohort. Descriptions are the same as for Figs 3 and S1

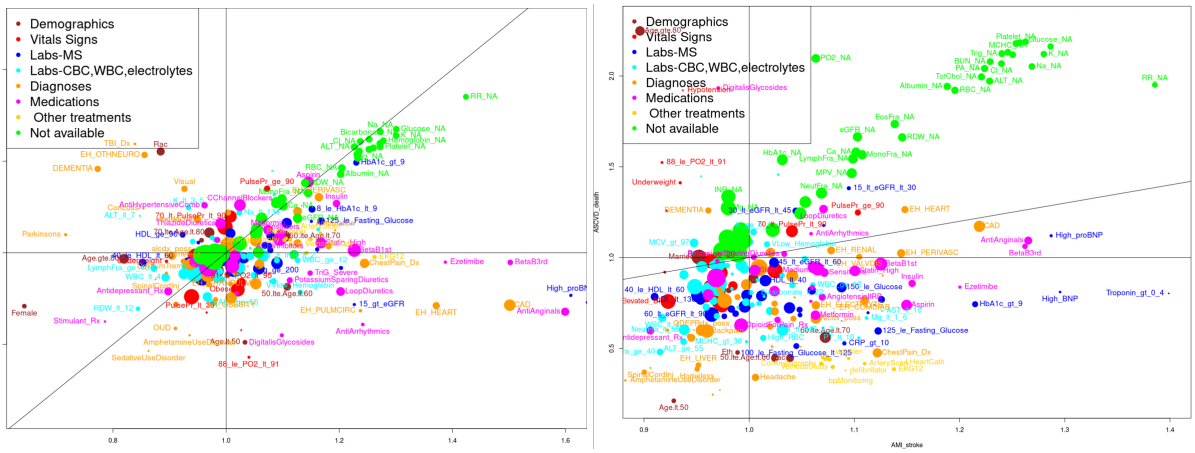

Figure S6. Association of each predictor variable with (left) stroke on the y-axis and AMI on the x-axis, and (right) ASCVD death on the y-axis and AMI or stroke on the x-axis, for the Rcc study design. Note that these are not model coefficients but the relative risk, or fraction of individuals with the indicated suboutcome divided by total, for each variable, compared to that quantity for all individuals. Vertical and horizontal axes and a line of slope one through the origin are provided for reference. On the left, variables above the diagonal are more predictive of stroke and include TBI, neuropathy, dementia, Parkinsons, blindness, and being >80 years old, female or Black. Below the diagonal are relative predictors of AMI, including diagnoses of CAD, heart failure, chest pain, and COPD, low SPO2 and high proBNP, and prescription of antianginals, 3rd generation beta blockers, Ezetimbe, anti arrhythmics, diuretics, and digitalis glycosides. Indicators of diabetes, behavioral health problems, and hypertension are equally predictive of stroke and AMI. Indicators of little healthcare or missingness of data (green) are also equally predictive of stroke and AMI, and are among the strongest predictors of outcome, which appears puzzling until one remembers that Rcc was matched on the level of healtchare utilization, and it is the rarer missing variables that are predictive of outcome, with the more common protective. More insight into this behavior is provided in the plot to the right, where missing data (green) are much more predictive of ASCVD death than stroke and AMI, presumably indicating a confounding effect of healthcare utilization with Dx-coded variables. Indeed, this figure can be interpreted largely by understanding how different modalities of care are predictive of whether a Dx-coded outcome will be recorded, working from bottom to top through CVD-related procedures, labs associated with metabolic syndromes, CVD medications, and finally missing laboratory measures. Similar plots are shown in Supplementary Figure S7 for the C17 cohort. This shows the strong impact of matching on healthcare utilization, in that the numerous NA categories are associated with decreased risk for AMI and stroke, and neutral risk for ASCVD death. **risk for different MACE suboutcomes** in subgroups defined by each predictor variable, compared to the overall average risk and in the Rcc cohort.

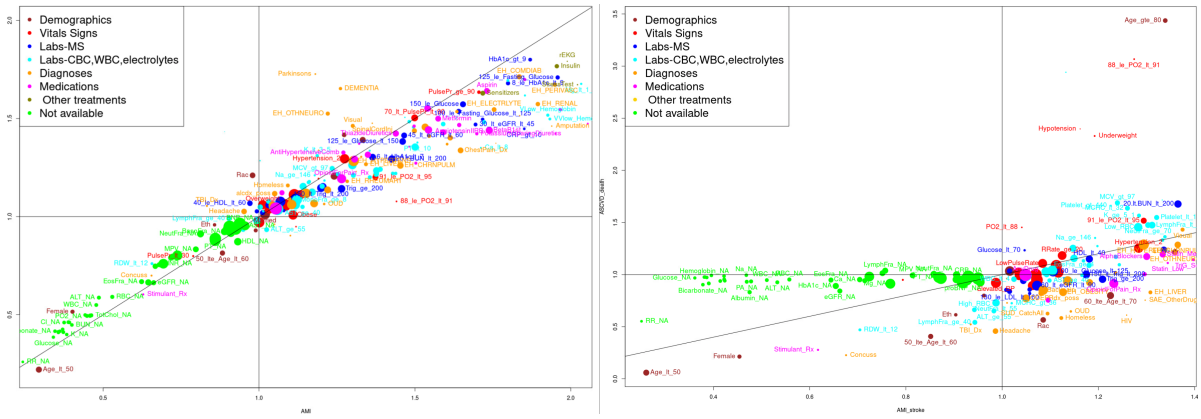

Figure S7. Association of each predictor variable with (left) AMI on the y-axis and stroke on the x-axis, and (right) ASCVD death on the y-axis and AMI + stroke on the x-axis, for the C17 study design. The color code and symbol sizes are the same as in Figure S6. Note that these are not model coefficients but the fraction of individuals with the indicated suboutcome divided by the number of controls, for each variable, compared to that quantity for all individuals. Vertical and horizontal axes and a line of slope one through the origin are provided for reference.

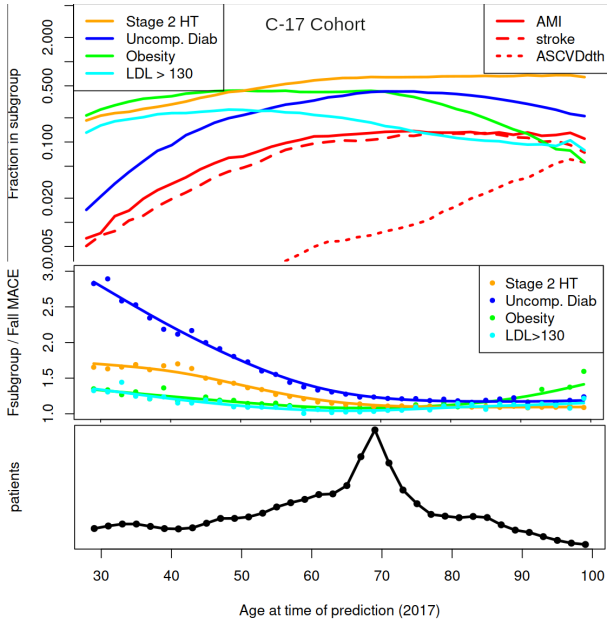

Figure S8. Dependence of selected variables on age for the C-17 cohort. The top panel shows on a logarithmic scale the age dependence of the fraction of each component of our combined outcome (in red), and variables representative of each of four risk factor groups, hypertension, diabetes, obesity, and high LDL cholesterol. The middle panel compares the age-dependent relative risk for MACE for each of the above subgroups vs. the overall C-17 cohort. The bottom panel is a histogram of total number of individuals in each age group. We placed the x-axis on a log10 scale, for which the histogram peaks at 1.3, or 20 codes present. All curves are strongly age dependent, indicating both confounding effects and the need for interaction terms in our model. Indeed, this is why Rcc is matched on date of birth and likely is one reason for the improved performance of the non-linear models. The age-dependence of the relative risk for these four risk factors can be seen in the middle panel, where we present the risk ratio for MACE (defined as the MACE prevalence within each subgroup to the MACE prevalence across all individuals) in the C-17 cohort. The strongest relative risk is in the younger diabetic and hypertensive individuals, with those under 45 years of age experiencing approximately 2-fold elevated MACE risk. The bottom panel shows a histogram of how the million individuals are distributed across age, with the lowest numbers around age 40 and 100, consistent with the larger scatter seen in the relative risk curves.

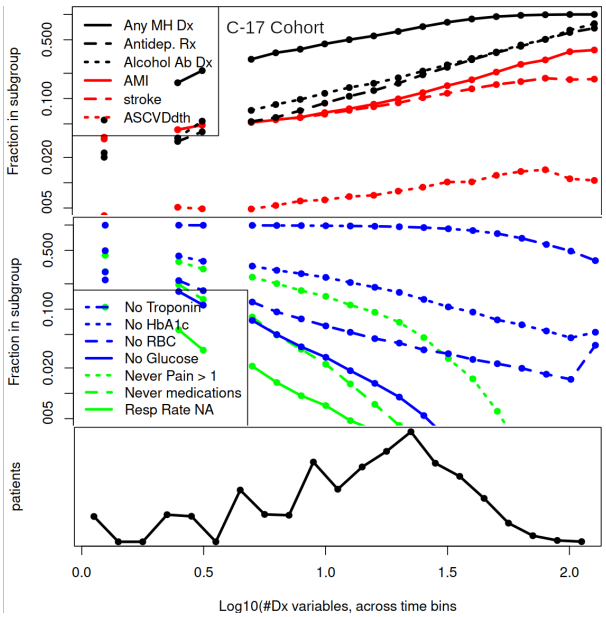

Figure S9. Dependence of selected variables on level of health care utilization for the C17 cohort. The top panel shows the fraction of each indicated subgroup on a logarithmic scale as a function of an aggregated measure of healthcare utilization for each component of our combined outcome (in red) and for three behavioral healthcare usage metrics (in black). The middle panel is similar but shows the number of individuals **not** receiving the indicated laboratory, vital signs, and pharmacy usage. The bottom panel shows a histogram of the number of individuals at each level of healthcare utilization. Our healthcare utilization metric (x-axis, right panel) was computed by summing the number of diagnosis codes present over the 47 distinct diagnosis variables across each of the eight time windows of data collection for a maximum possible score of 376 (each bin is encoded as presence/absence, or 0/1). We placed the x-axis on a log10 scale, for which the histogram peaks at 1.3, or 20 codes present. Supplementary Figure S9 provides similar information for our other strongly interacting and confounded effect - missingness of data and level of healthcare utilization.

References

1. Soppert, J., Lehrke, M., Marx, N., Jankowski, J. & Noels, H. Lipoproteins and lipids in cardiovascular disease: From mechanistic insights to therapeutic targeting. *Adv. Drug Deliv. Rev.* **159**, 4–33 (2020).
2. Liu, C., Dhindsa, D., Almuwaqqat, Z., Sun, Y. V. & Quyyumi, A. A. Very high high-density lipoprotein cholesterol levels and cardiovascular mortality. *Am. J. Cardiol.* **167**, 43–53 (2022).
3. Madsen, C. M., Varbo, A. & Nordestgaard, B. G. Extreme high high-density lipoprotein cholesterol is paradoxically associated with high mortality in men and women: Two prospective cohort studies. *Eur. Hear. J.* **38**, 2478–2486 (2017).
4. Grundy, S. N. & *et al.* 2018 AHA/ACC/AACVPR/AAPA/ABC/ACPM/ADA/AGS/APhA/ASPC/NLA/PCNA guideline on the management of blood cholesterol: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Circulation* **139**, e1082–e1143 (2019).
5. Hollander, W. Role of hypertension in atherosclerosis and cardiovascular disease. *Am. J. Cardiol.* **38**, 786–800 (1978).
6. Kim, H. *et al.* Age and sex specific target of blood pressure for the prevention of cardiovascular event among the treatment naive hypertensive patients. *Nat. Sci. Rep.* **12**, 21538 (2020).
7. Domanski, M. *et al.* Pulse pressure and cardiovascular disease-related mortality. follow-up study of the Multiple Risk Factor Intervention Trial (MRFIT). *J. Am. Med. Assoc.* **287**, 2677–2683 (2002).
8. Mancia, G. Should blood pressure reduction be aggressive in patients with hypertension and coronary artery disease? *Lancet* **388**, 2061–2062 (2016).
9. Vidal-Petiot, E. *et al.* Cardiovascular event rates and mortality according to achieved systolic and diastolic blood pressure in patients with stable coronary artery disease: An international cohort study. *Lancet* **388**, 2142–2152 (2016).
10. Böhm, M. *et al.* Achieved blood pressure and cardiovascular outcomes in high-risk patients: results from ONTARGET and TRANSCEND trials. *Lancet* **389**, 2226–2237 (2017).
11. Sun, N., Yuanuyan, C., Xi, Y., Wang, H. & Wang, L. Association between heart rate and major adverse cardiovascular events among 9,991 hypertentive patients: A multicenter retrospective follow-up study. *Front. Cardiovasc. Med.* **8**, 1–10 (2021).
12. Ninomiya, T. *et al.* Albuminuria and kidney function independently predict cardiovascular and renal outcomes in diabetes. *J. Am. Soc. Nephrol.* **20**, 1813–1821 (2009).
13. Currie, C. J. *et al.* Major adverse cardiovascular events in people with chronic kidney disease in relation to disease severity and diabetes status. *PLoS ONE* **14**, e0221044 (2019).
14. Giugliano, D. *et al.* Preventing major adverse cardiovascular events by sglt-2 inhibition in patients with type 2 diabetes: the role of the kidney. *Cardiovasc. Diabetol* **19**, 1–6 (2020).
15. Giugliano, D. *et al.* Sglt-2 inhibitors and cardiorenal outcomes in patients with or without type 2 diabetes: a meta-analysis of 11 cvots. *Cardiovasc. Diabetol.* **20**, 1–11 (2021).
16. Rodriguez, M. *et al.* Hyponatremia in heart failure: Pathogenesis and management. *Cur. Cardiol. Rev.* **15**, 252–261 (2019).
17. Laslett, D. B. *et al.* Electrolyte abnormalities in patients presenting with ventricular arrhythmia (from the lyte-vt study). *Am. J. Cardiol.* **129**, 36–41 (2020).
18. Liamis, G. *et al.* Electrolyte disorders in community subjects: prevalence and risk factors. *Am. J. Med.* **126**, 256–263 (2013).
19. H, K. & R., Z. Antihypertensive medications. (2022).
20. Wang, X., Chen, X., Chen, Z. & Zhang, M. Arterial calcification and its association with stroke: Implication of risk, prognosis, treatment response, and prevention. *Front. Cell. Neurosci.* **16** (2022).
21. Javaid, A., Mitchell, J. D. & Villines, T. Predictors of coronary artery calcium and long-term risks of death, myocardial infarction, and stroke in young adults. *J. Am. Hear. Assoc.* **10**, e022513 (2021).
22. Peters, S. A., Huxley, R. R., Sattar, N. & Woodward, M. Sex differences in the excess risk of cardiovascular diseases associated with type 2 diabetes: Potential explanations and clinical implications. *Curr. Cardiovasc. Risk Rep.* **9**, 1–39 (2015).
23. Leon, B. M. & Maddox, T. M. Diabetes and cardiovascular disease: Epidemiology, biological mechanisms, treatment recommendations and future research. *World J. Diab.* **6**, 1246–1258 (2015).

24. Galbete, A. *et al.* Cardiovascular risk in patients with type 2 diabetes: A systematic review of prediction models. *Diabetes Res. Clin. Pract.* **184**, 109089, DOI: https://doi.org/10.1016/j.diabres.2021.109089 (2022).

25. Wilson, P. W. F. *et al.* Prediction of coronary heart disease using risk factor categories. *Circulation* **97**, 1837–1847 (1998).

26. D'Agostino, R. B., Lee, M.-L. & Belanger, A. J. Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham Heart Study. *Stat. Medicine* **9**, 1501–1515 (1990).

27. Goff, D. C. *et al.* 2013 ACC/AHA guideline on the assessment of cardiovascular risk. *Circulation* **129**, S49–S73, DOI: 10.1161/01.cir.0000437741.48606.98 (2014). https://www.ahajournals.org/doi/pdf/10.1161/01.cir.0000437741.48606.98.

28. Flora, G. D. & Nayak, M. K. A brief review of cardiovascular diseases, associated risk factors and current treatment regimes. *Curr. Pharm. Des.* **25**, 4063–4084 (2019).

29. Roth, G. A., Mensah, G. A. & Others. Global burden of cardiovascular diseases and risk factors, 1990-2019: Update from the GBD 2019 study. *J. Am. Coll. Cardiol.* **76**, 2982–3021 (2020).

30. Flier, J. S. & Maratos-Flier, E. Pathobiology of obesity. In Jameson, J. L. *et al.* (eds.) *Harrison's Principles of Internal Medicine*, Chapter 394 (McGraw-Hill, New York, 2018).

31. Carbone, S. *et al.* Obesity paradox in cardiovascular disease: Where do we stand? *Vasc. Heal. Risk Manag.* **15**, 89–100 (2019).

32. Eckel, N. *et al.* Transition from metabolic healthy to unhealthy phenotypes and association with cardiovascular disease risk across BMI categories in 90 257 women (the Nurses' Health Study): 30 year follow-up from a prospective cohort study. *Lancet Diabetes Endocrinol* **6**, 714–724 (2018).

33. Park, D., Lee, J. H. & Han, S. Underweight: another risk factor for cardiovascular disease?: A cross-sectional 2013 Behavioral Risk Factor Surveillance System (BRFSS) study of 491,773 individuals in the USA. *Medicine (Baltimore)* **96**, e8769 (2017).

Glossary

AMI acute myocardial infarction
ASCVD atherosclerotic cardiovascular disease
BMI body mass index
C17 calendar 2017
CAD coronary artery disease
CBC complete blood counts
CVD cardiovascular disease
HDL high-density lipoprotein
LDL low-density lipoprotein
MACE major adverse cardiovascular events
RBC red blood cell
Rcc nested retrospective case control
VHA US Department of Veterans Healthcare Administration
Visit-DS Visit - decision support
WBC white blood cell