Collinearity Reduction via Outliers Selection Regression (CROS-R): a Novel Approach to Handle Highly Correlated Predictors in Regression Modeling

Giorgio Melloni, Andrea Bellavia
TIMI Study Group, Brigham and Women's Hospital, Harvard Medical School

03/12/2024

Acknowledgments:

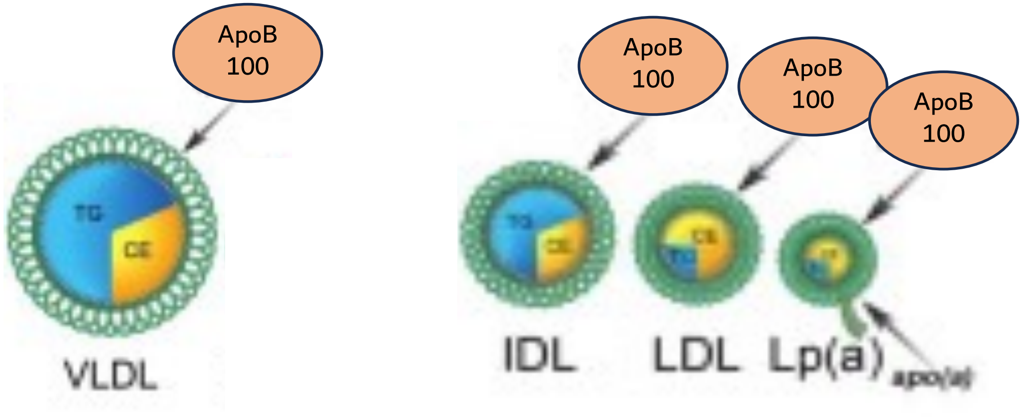
- **Dr. Nicholas A. Marston,** TIMI Study Group
- **Dr. Jakub Morze,** SGMK University
- **Sabina Murphy**, TIMI Study Group
- **- Dr. Marc S. Sabatine,** TIMI Study Group

- Dr. Patrick T. Ellinor, Broad Institute

Disclaimer:

The authors declare no conflict of interest. GEM, AB, NAM SM and MSS are members of the TIMI Study Group which has received institutional research grant support through Brigham and Women's Hospital from: Abbott, Abiomed, Inc., Amgen, Anthos Therapeutics, ARCA Biopharma, Inc., AstraZeneca, Boehringer Ingelheim, Daiichi-Sankyo, Ionis Pharmaceuticals, Inc., Janssen Research and Development, LLC, MedImmune, Merck, Novartis, Pfizer, Regeneron Pharmaceuticals, Inc., Roche, Saghmos Therapeutics, Inc., Siemens Healthcare Diagnostics, Inc., Softcell Medical Limited, The Medicines Company, Verve Therapeutics, Inc., Zora Biosciences

ApoB lipoproteins

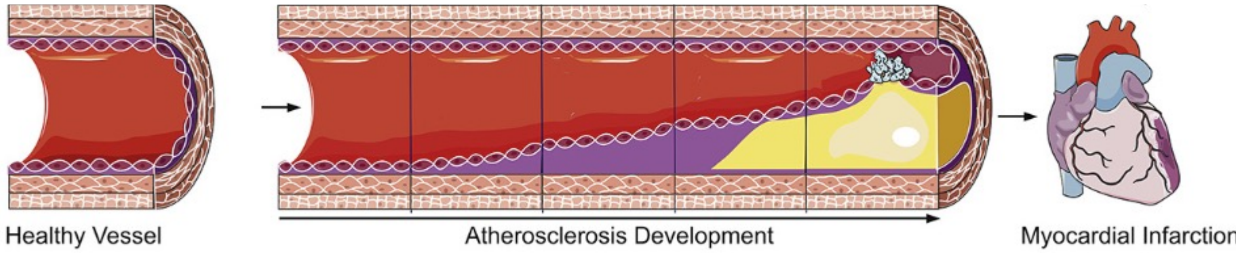


In the context of cardiovascular (CV) prevention and therapy,
ApoB lipoproteins are common biomarkers of **atherosclerosis**

Irrespective of class, all these lipoproteins share an ApoB100 molecule on their surface
Measuring ApoB can inform on the total abundance of the entire class of lipoprotein

Adapted from Glavinovic et al. , *JAHA*, 2022

ApoB lipoproteins



LDL, IDL and VLDL particles are all involved in plaque formation and associated with CV risk

Can ApoB recapitulate all of these single associations?

ApoB = LDL + IDL + VLDL

↓

ApoB risk $\overset{?}{=}$ LDL risk + IDL risk + VLDL risk

⬆ + ⬆ + ⬆

Adapted from Soppert et al. , *JADDR*, 2020

Why is it important?

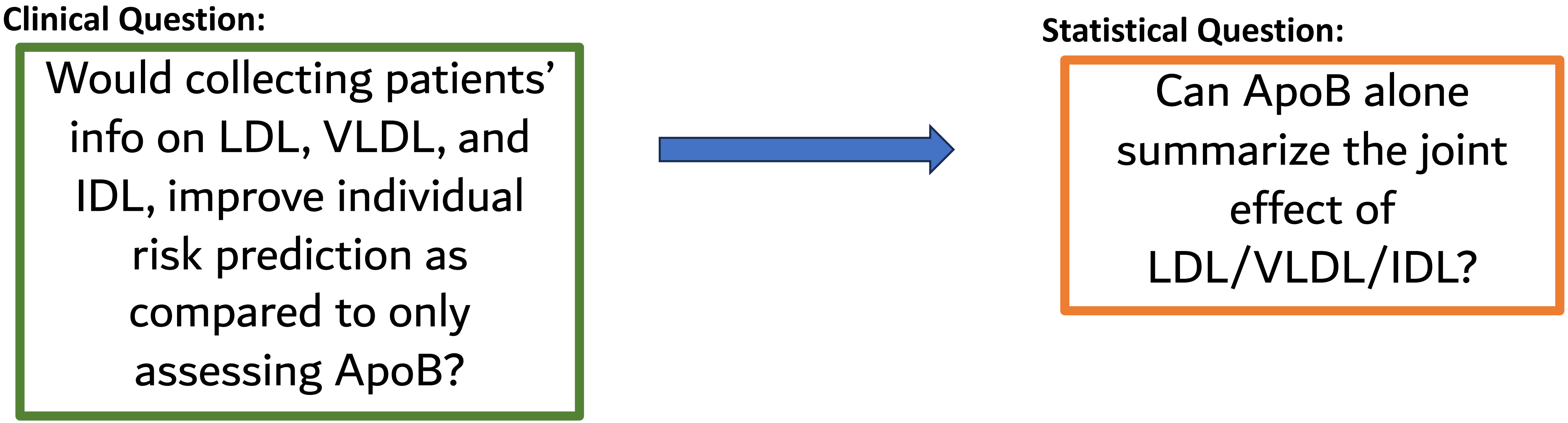
- The role of **LDL** (Low Density Lipoprotein, aka “bad” cholesterol) is well studied and understood
  - Agents that reduce LDL-Cholesterol (and ApoB) are part of the standard cardiovascular therapies
  - **ApoB** has been shown to be a superior risk biomarker of ASCVD

- **IDL** (Intermediate Density) and **VLDL** (Very Low Density) are less abundant
  - **Triglycerides** (**TG**) have been shown to carry CV risk beyond LDL cholesterol
  - Drugs that effectively reduce TG exists but their efficacy in clinical trial contests is unclear

Marston *et al.*, JAMA Cardiology, 2022

Objective 1

- 1) VARIABLE SELECTION PROBLEM:



Clinical Question:

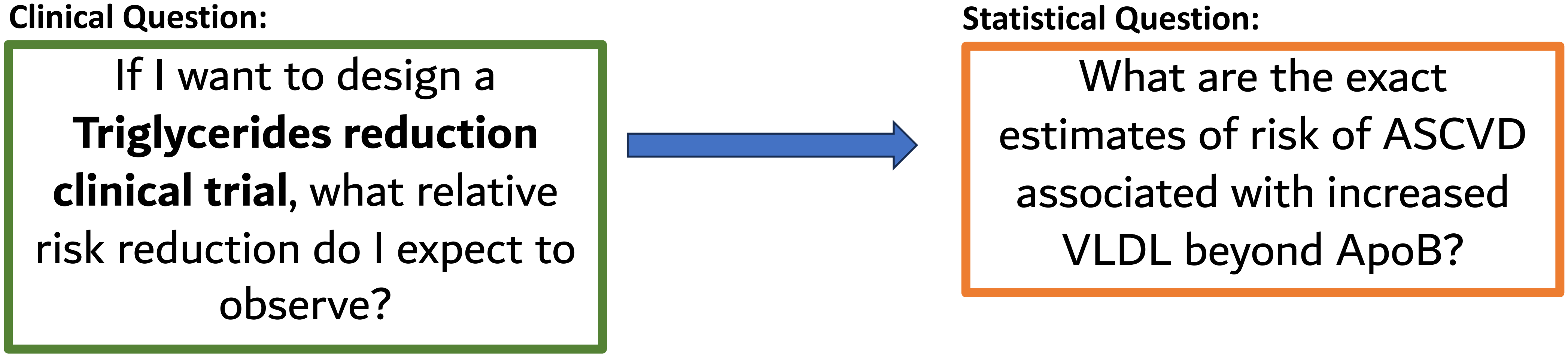
Would collecting patients' info on LDL, VLDL, and IDL, improve individual risk prediction as compared to only assessing ApoB?

Statistical Question:

Can ApoB alone summarize the joint effect of LDL/VLDL/IDL?

Objective 2

- 2) ESTIMATION PROBLEM:



Clinical Question:

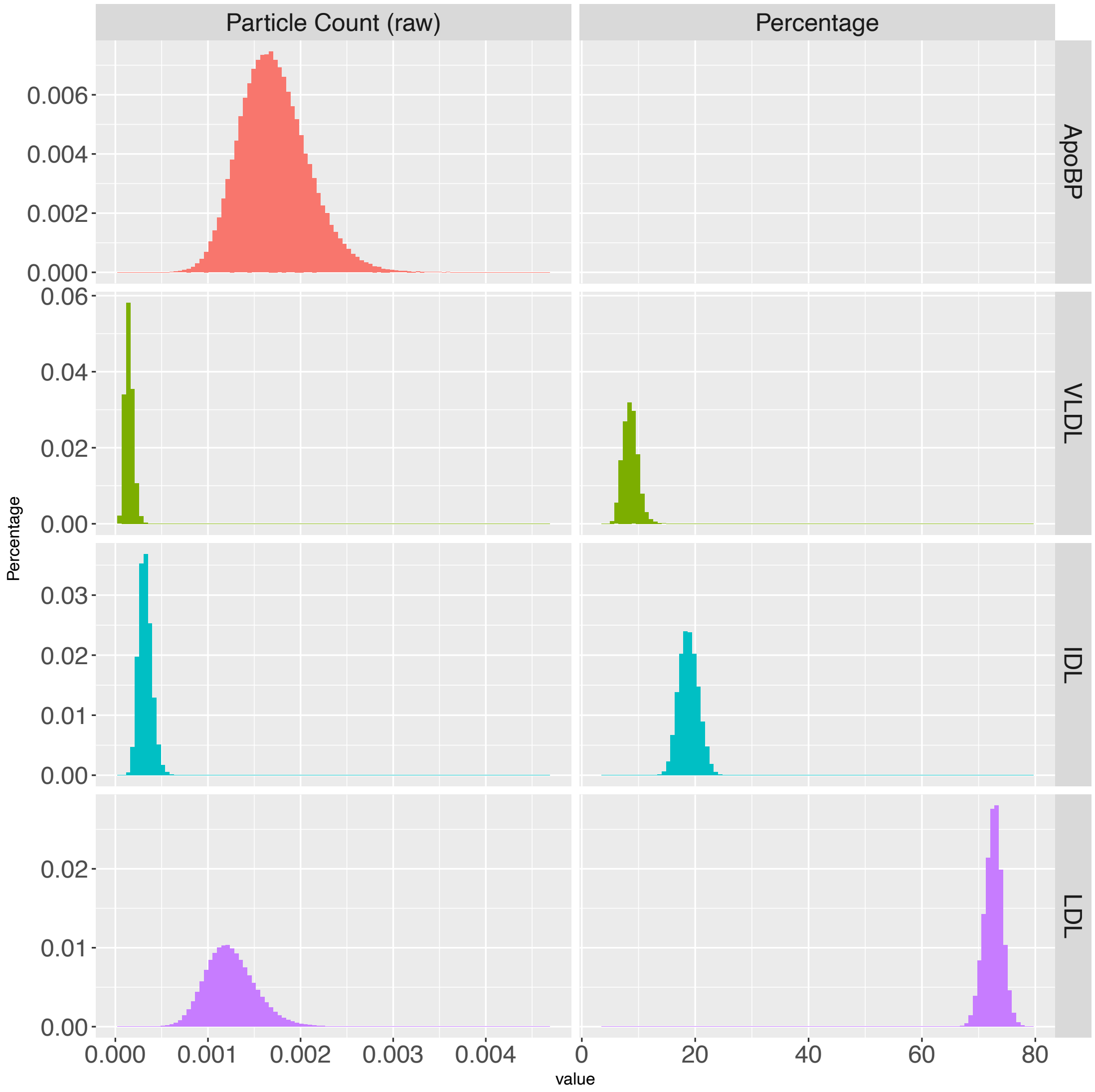
If I want to design a **Triglycerides reduction clinical trial**, what relative risk reduction do I expect to observe?

Statistical Question:

What are the exact estimates of risk of ASCVD associated with increased VLDL beyond ApoB?

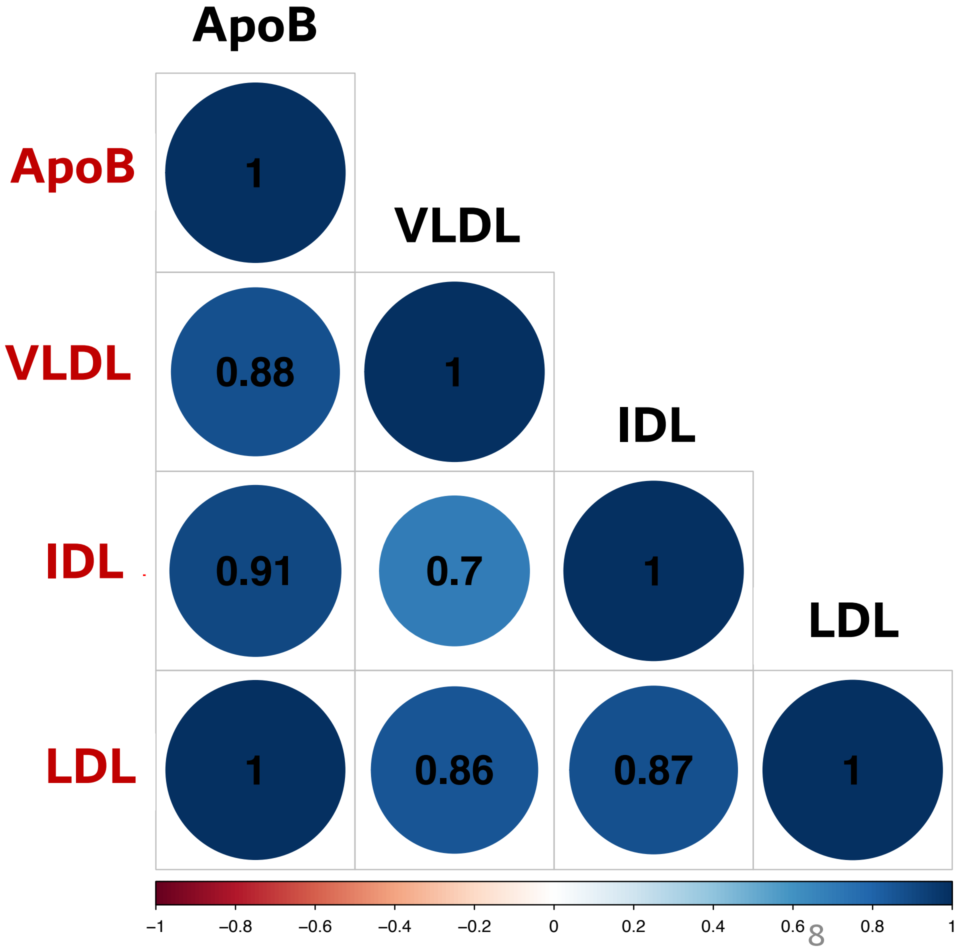
Lipoprotein distribution

The **extreme correlation** between lipoproteins is a major challenge to risk estimation because of **multicollinearity**



VLDL particles account for less than 10% of ApoB containing particles

LDL particles account for more than 70% of all ApoB containing particles



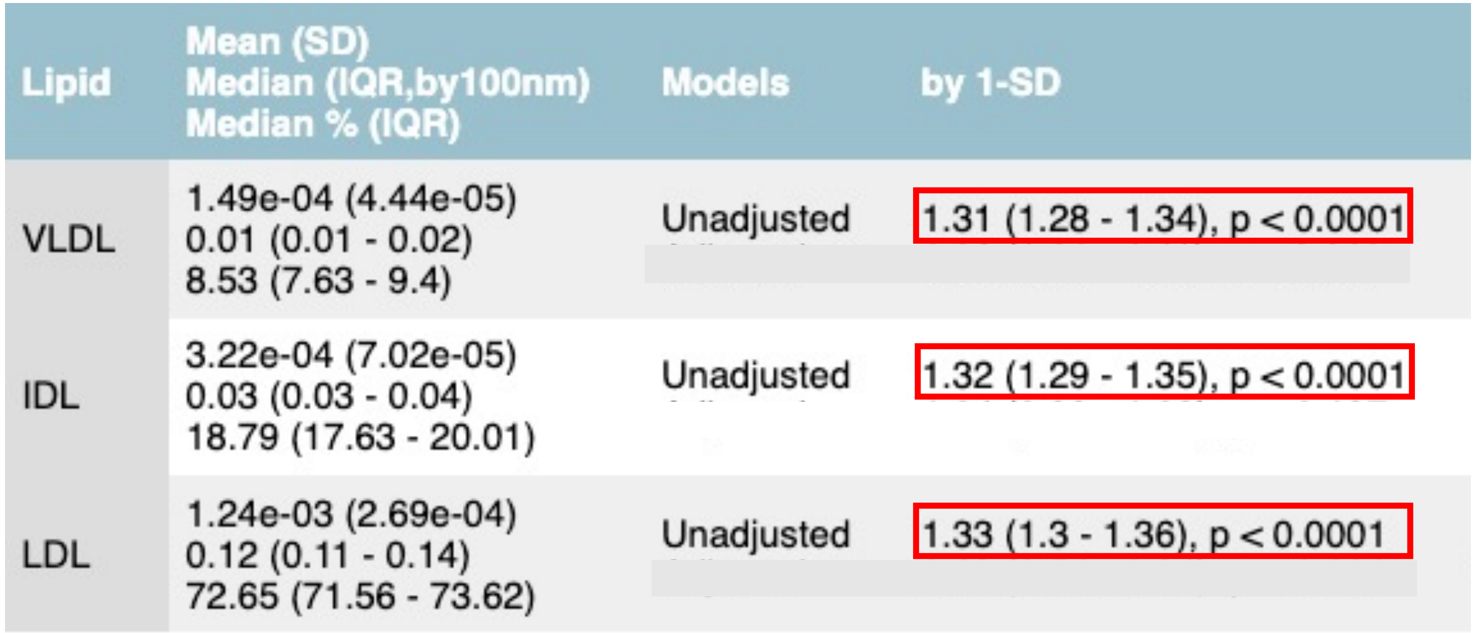
First Approach – direct Cox modeling

Using a dataset of lipid **NMR Spectroscopy** data from **210,732** primary prevention individuals from the **UK Biobank,** we wish to evaluate the joint effect of **VLDL** and **ApoB** in estimating Atherosclerotic Cardiovascular Events (ASCVD = MI, ischemic stroke or CV death):

Surv(ASCVD) ~ VLDL + ApoB + clinical adjustments

HR by 1-SD increase in each lipoprotein is reported to compare effect estimates across particle types

First Approach – direct Cox modeling

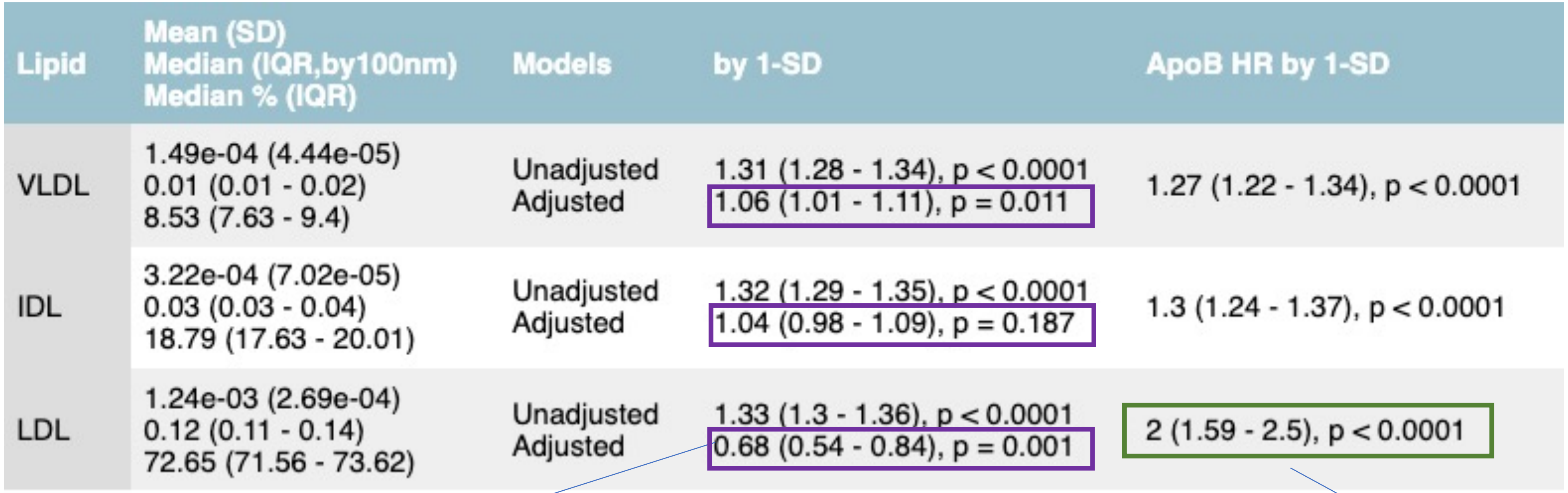


Lipid	Mean (SD) Median (IQR,by100nm) Median % (IQR)	Models	by 1-SD
VLDL	1.49e-04 (4.44e-05) 0.01 (0.01 - 0.02) 8.53 (7.63 - 9.4)	Unadjusted	1.31 (1.28 - 1.34), p < 0.0001
IDL	3.22e-04 (7.02e-05) 0.03 (0.03 - 0.04) 18.79 (17.63 - 20.01)	Unadjusted	1.32 (1.29 - 1.35), p < 0.0001
LDL	1.24e-03 (2.69e-04) 0.12 (0.11 - 0.14) 72.65 (71.56 - 73.62)	Unadjusted	1.33 (1.3 - 1.36), p < 0.0001

Adjusted only by clinical risk factors, all 3 particle types are associated with an increased risk of ~30% of developing ASCVD.

ApoB alone is associated with a 35% increased risk (HR 95% CI 1.35 [1.32 - 1.38])

First Approach – direct Cox modeling



Lipid	Mean (SD) Median (IQR,by100nm) Median % (IQR)	Models	by 1-SD	ApoB HR by 1-SD
VLDL	1.49e-04 (4.44e-05) 0.01 (0.01 - 0.02) 8.53 (7.63 - 9.4)	Unadjusted Adjusted	1.31 (1.28 - 1.34), p < 0.0001 1.06 (1.01 - 1.11), p = 0.011	1.27 (1.22 - 1.34), p < 0.0001
IDL	3.22e-04 (7.02e-05) 0.03 (0.03 - 0.04) 18.79 (17.63 - 20.01)	Unadjusted Adjusted	1.32 (1.29 - 1.35), p < 0.0001 1.04 (0.98 - 1.09), p = 0.187	1.3 (1.24 - 1.37), p < 0.0001
LDL	1.24e-03 (2.69e-04) 0.12 (0.11 - 0.14) 72.65 (71.56 - 73.62)	Unadjusted Adjusted	1.33 (1.3 - 1.36), p < 0.0001 0.68 (0.54 - 0.84), p = 0.001	2 (1.59 - 2.5), p < 0.0001

After adjusting for ApoB, 3 each model behaves differently:

- **VLDL** HR shrinks to 6%, $p < 0.05$
- **IDL** HR shrinks to 4%, $p > 0.05$
- **LDL** HR flips from >1 to 0.68, $p < 0.05$

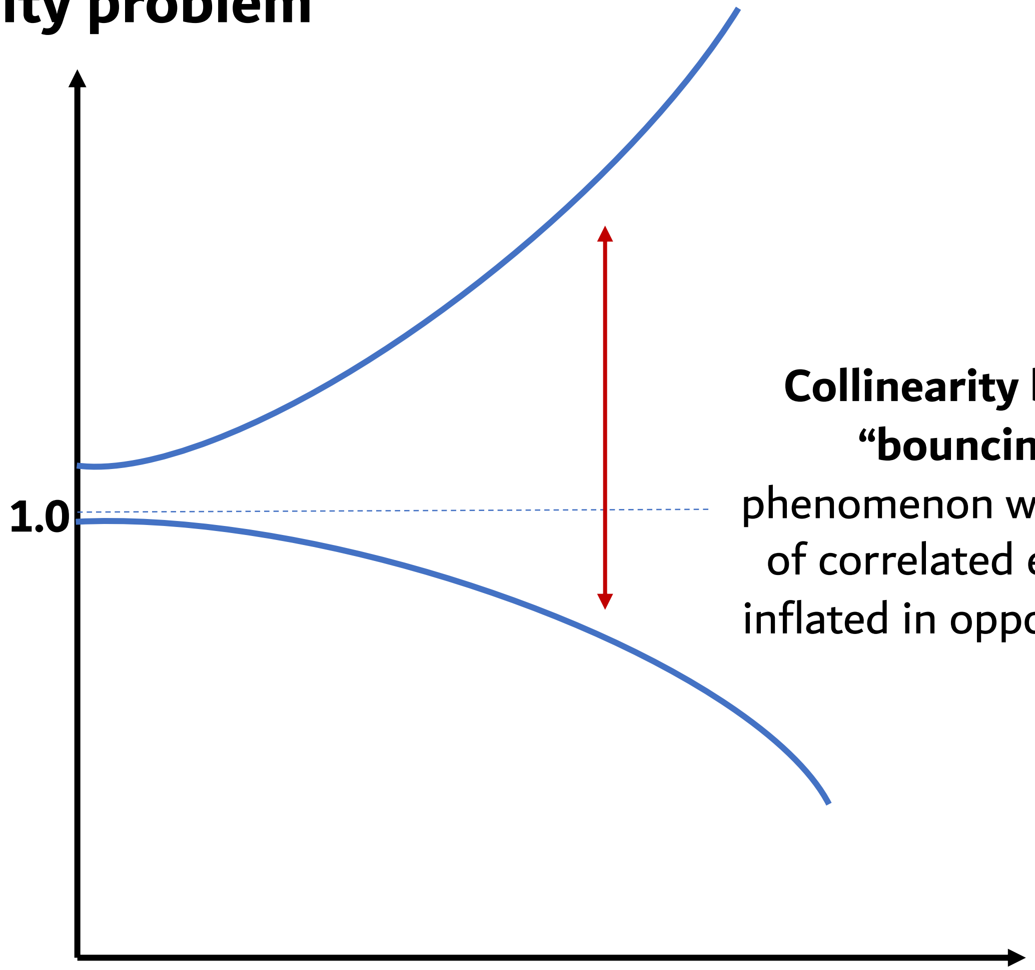
ApoB HR gets inflated by the collinearity

Collinearity problem

Models that include ApoB + another lipid have all **VIF > 4** for the variables of interest, a sign of collinearity

Lipid Model	VIF lipoprotein	VIF ApoB
VLDL	4.7	4.75
IDL	5.7	5.47
LDL	106.66	107.52

$$VIF_i = \frac{1}{1 - R_i^2}$$



Collinearity brings to the **“bouncing betas”** phenomenon where estimates of correlated exposures get inflated in opposite directions

<u>The use of non-parametric models (like bootstrap regression) or penalized models (LASSO/Elastic Net) nor alternative models (WQS) did not significantly reduce inflation</u>

Subsetting uncorrelated individuals (CROS-R)

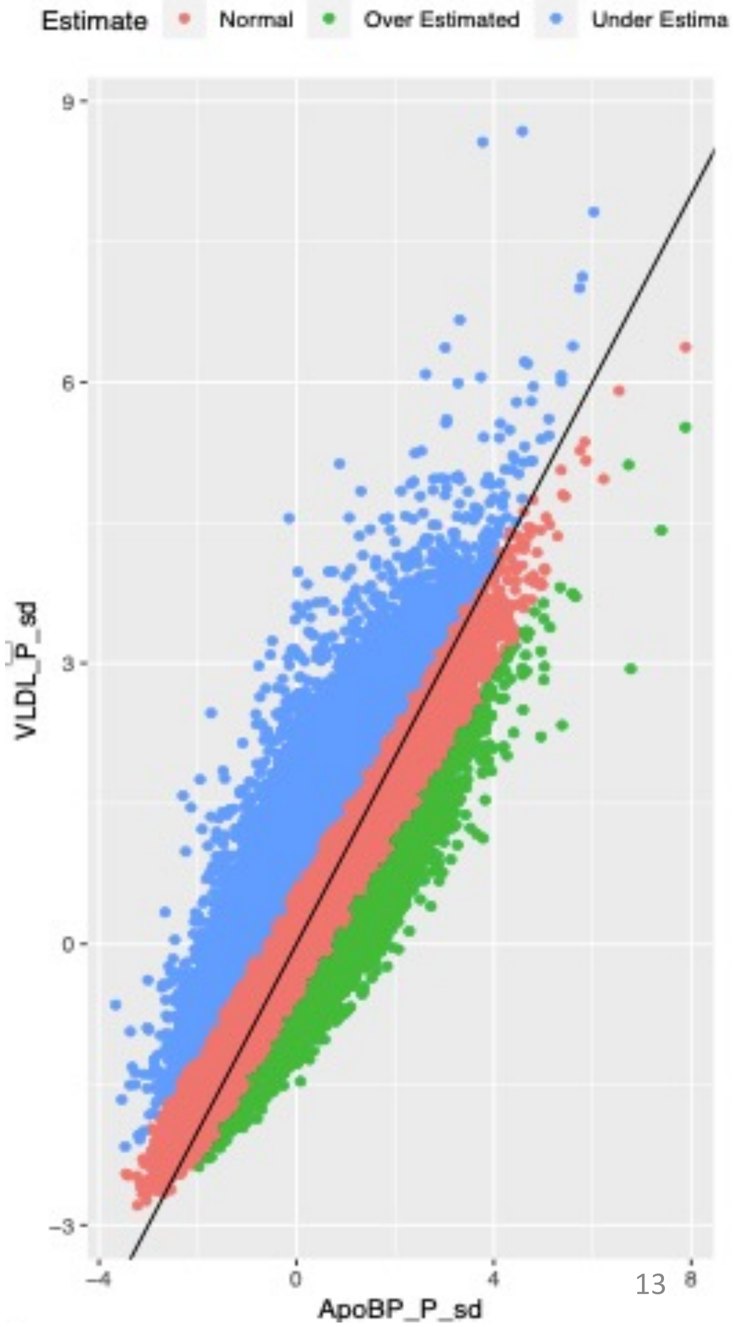
We run a linear regression between VLDL and Apob:

$$VLDL = \alpha + \beta * ApoB + \boldsymbol{\varepsilon}$$

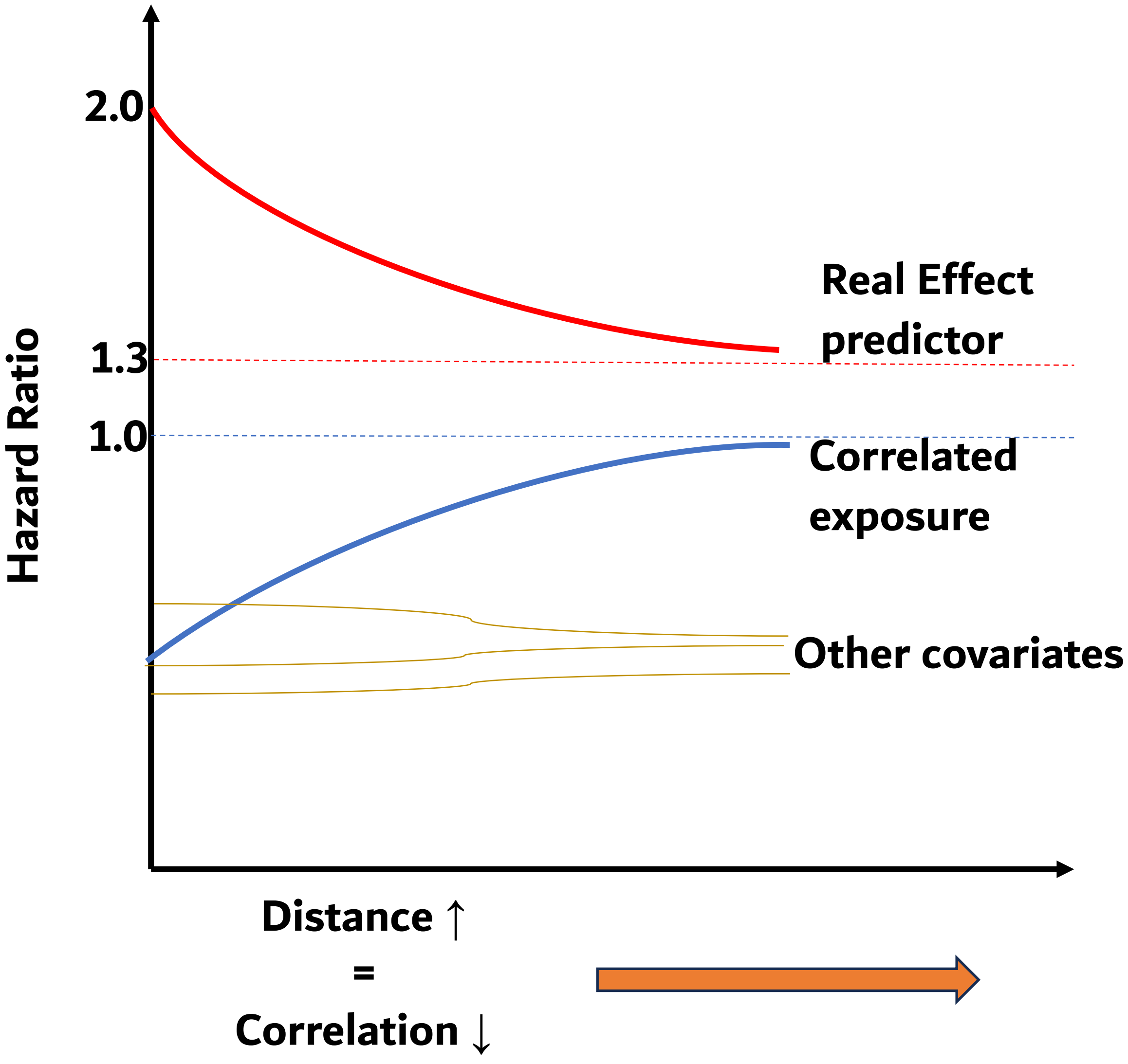
At varying **ε** we define 3 regions:

Individuals in **red** do not carry any information regarding the differential effect of ApoB and VLDL as one can easily be derived from the other.

Information in **blue** **(underestimated values)** and **green** **(overestimated values)** can be used to differentiate the effect of ApoB and VLDL



CROS-R: Collinearity Reduction via Outliers Selection Regression



By applying the same survival model on the regression outliers, we expect to see:

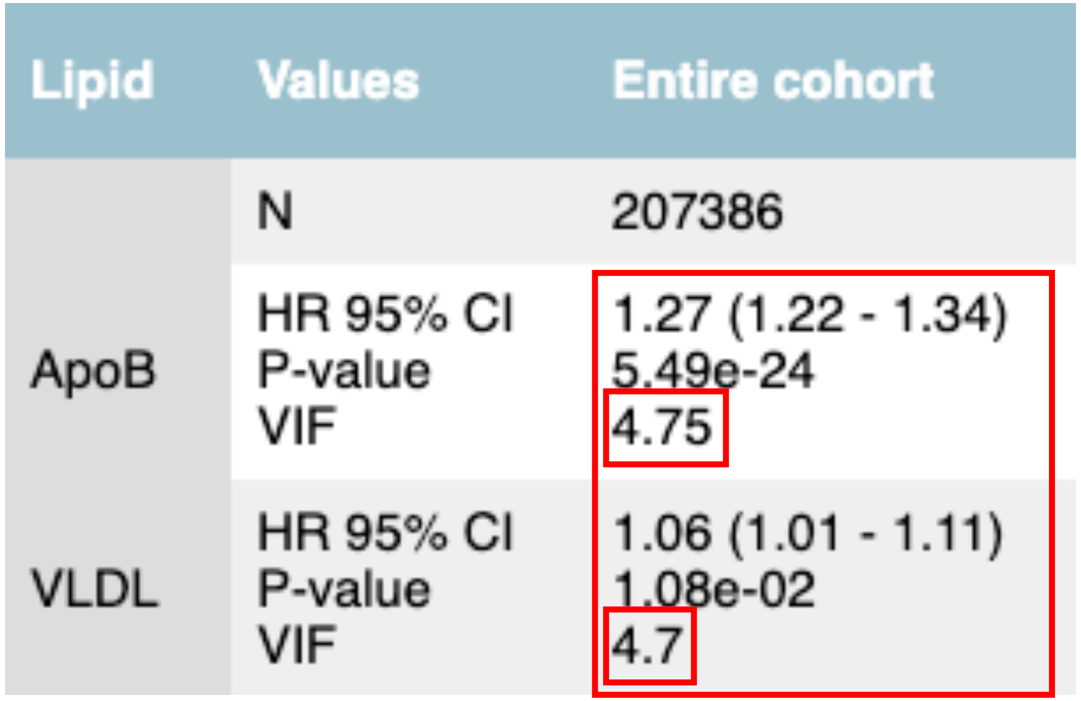
1) The estimate for the covariate with the real effect deflates and settles on the real effect size

2) The estimate of the correlated covariate should approach 1 (Null Effect in the presence of the real effect covariate)

3) All other covariates should stay fixed

CROS-R: Results

VLDL

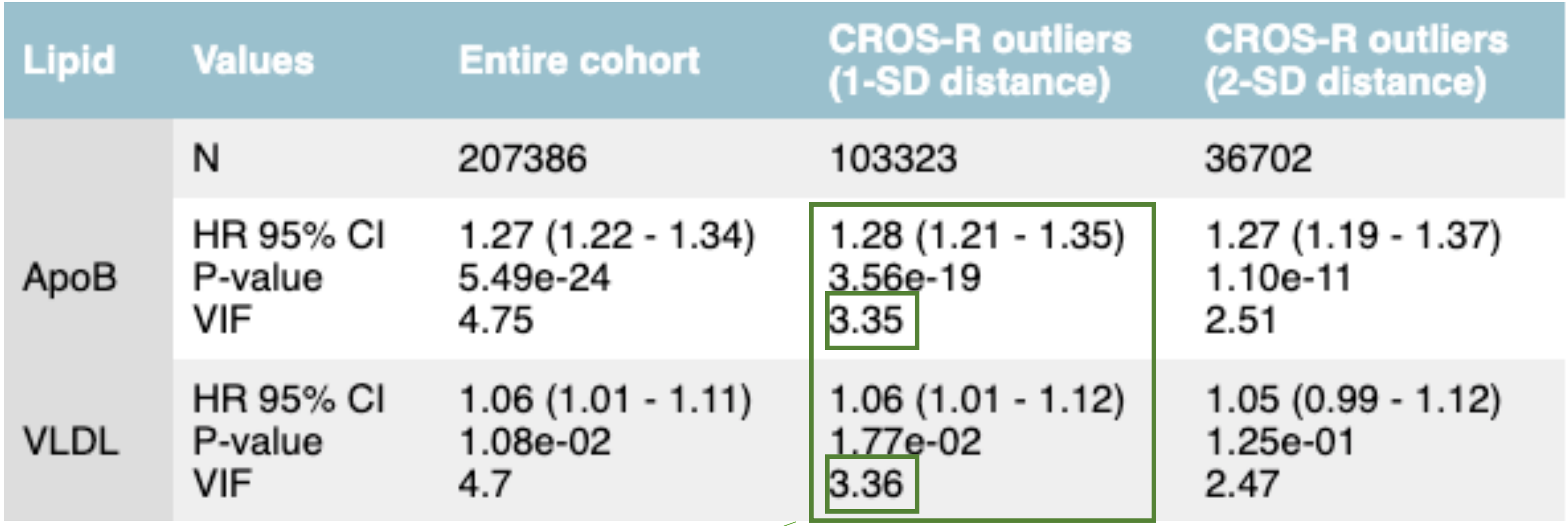


Lipid	Values	Entire cohort
	N	207386
ApoB	HR 95% CI P-value VIF	1.27 (1.22 - 1.34) 5.49e-24 4.75
VLDL	HR 95% CI P-value VIF	1.06 (1.01 - 1.11) 1.08e-02 4.7

VLDL and ApoB are affected by collinearity (VIF > 4)

VLDL

CROS-R: Results

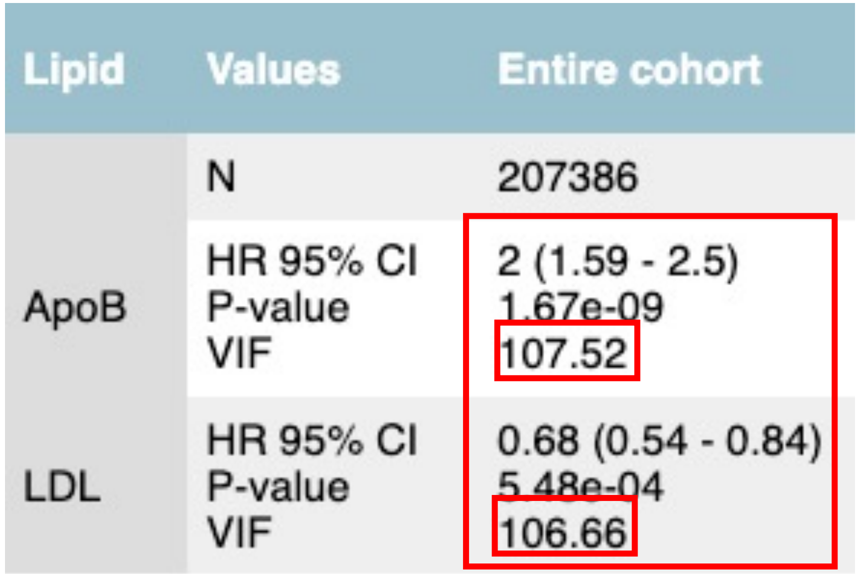


Lipid	Values	Entire cohort	CROS-R outliers (1-SD distance)	CROS-R outliers (2-SD distance)
	N	207386	103323	36702
ApoB	HR 95% CI P-value VIF	1.27 (1.22 - 1.34) 5.49e-24 4.75	1.28 (1.21 - 1.35) 3.56e-19 3.35	1.27 (1.19 - 1.37) 1.10e-11 2.51
VLDL	HR 95% CI P-value VIF	1.06 (1.01 - 1.11) 1.08e-02 4.7	1.06 (1.01 - 1.12) 1.77e-02 3.36	1.05 (0.99 - 1.12) 1.25e-01 2.47

At 1-SD distance, VIF is below 4 and estimates are unchanged, showing a small but significant contribution of VLDL beyond the risk conferred by ApoB

LDL

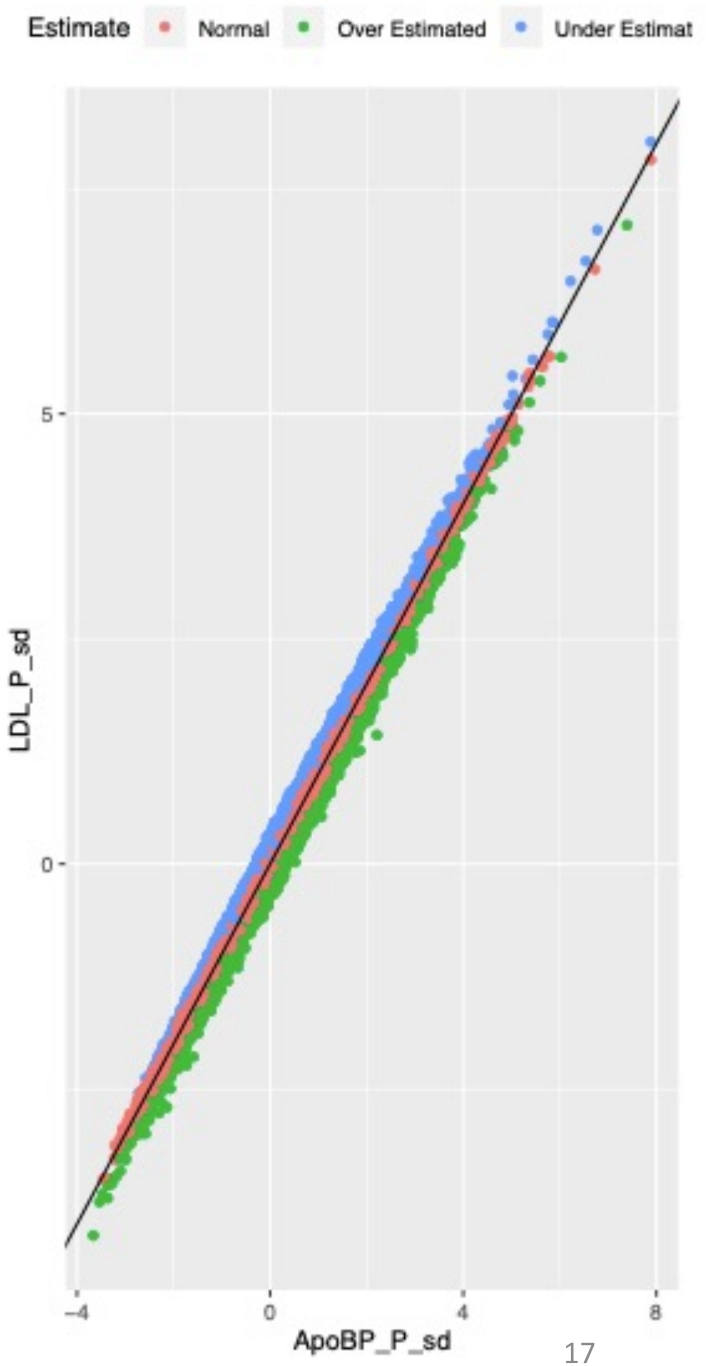
CROS-R: Results



Lipid	Values	Entire cohort
	N	207386
ApoB	HR 95% CI P-value VIF	2 (1.59 - 2.5) 1.67e-09 107.52
LDL	HR 95% CI P-value VIF	0.68 (0.54 - 0.84) 5.48e-04 106.66

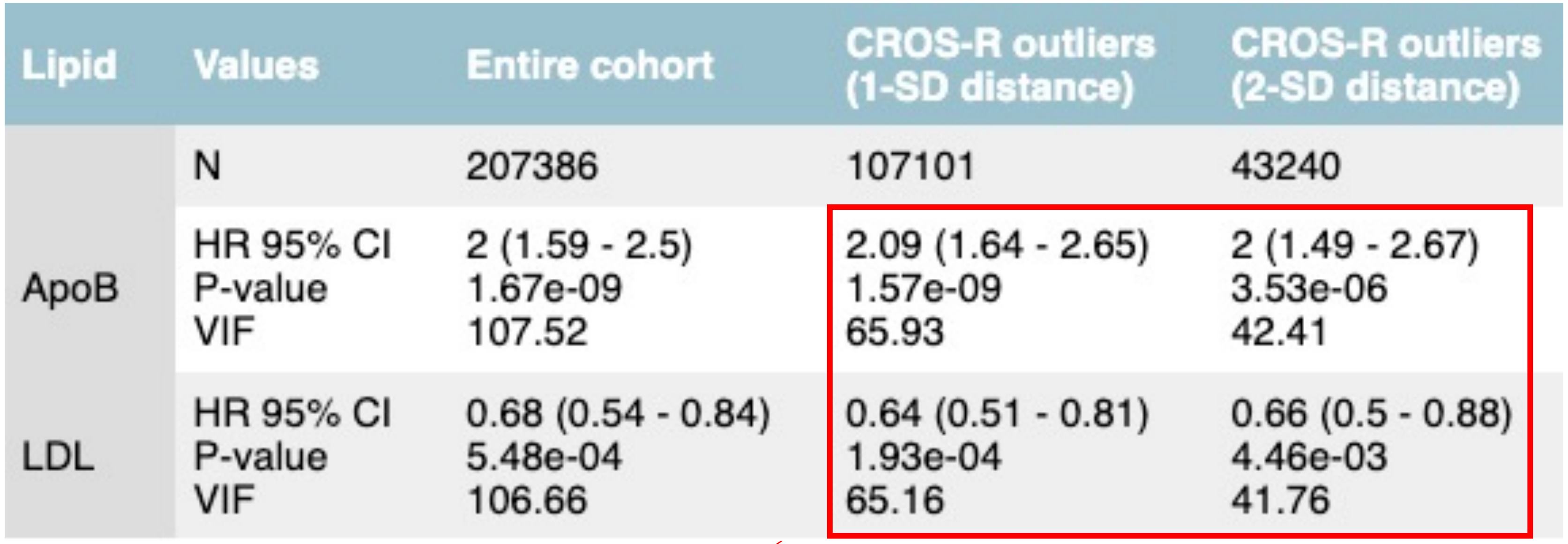
The original model was affected heavily by collinearity with VIF values above 100.

The correlation between the two variables is almost 1



LDL

CROS-R: Results

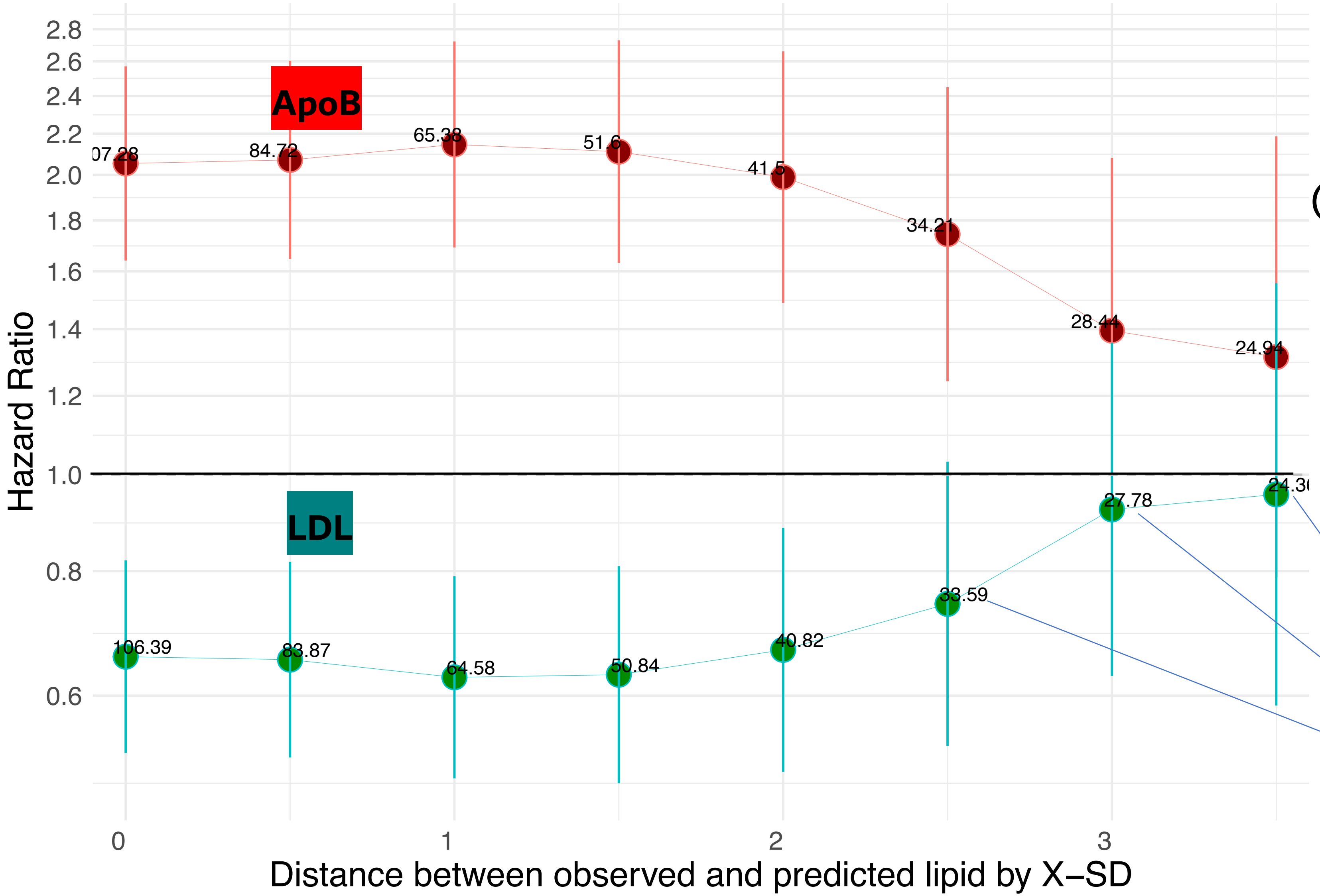


Lipid	Values	Entire cohort	CROS-R outliers (1-SD distance)	CROS-R outliers (2-SD distance)
	N	207386	107101	43240
ApoB	HR 95% CI P-value VIF	2 (1.59 - 2.5) 1.67e-09 107.52	2.09 (1.64 - 2.65) 1.57e-09 65.93	2 (1.49 - 2.67) 3.53e-06 42.41
LDL	HR 95% CI P-value VIF	0.68 (0.54 - 0.84) 5.48e-04 106.66	0.64 (0.51 - 0.81) 1.93e-04 65.16	0.66 (0.5 - 0.88) 4.46e-03 41.76

At 1 and 2-SD distance outliers, the VIF drops to ~40 but the estimates remain the same

CROS-R: Results

LDL



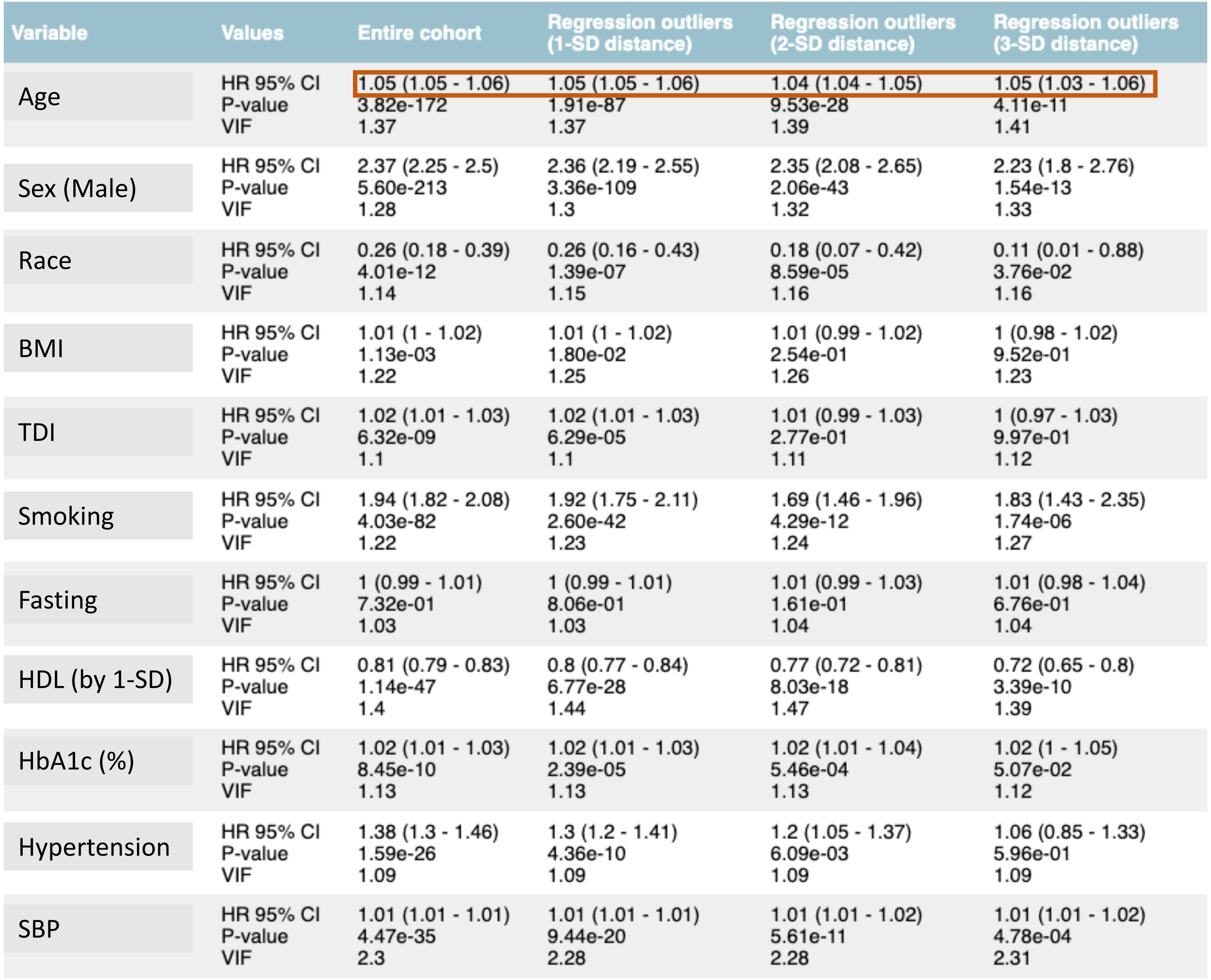
Final model estimates
(3.5-SD distance, VIF ~ 24)

ApoB HR
1.31(0.79,2.19) ~ 1.3

LDL HR
0.96(0.59,1.56) ~ 1

VIF values

CROS-R: other covariates



Variable	Values	Entire cohort	Regression outliers (1-SD distance)	Regression outliers (2-SD distance)	Regression outliers (3-SD distance)
Age	HR 95% CI	1.05 (1.05 - 1.06)	1.05 (1.05 - 1.06)	1.04 (1.04 - 1.05)	1.05 (1.03 - 1.06)
	P-value	3.82e-172	1.91e-87	9.53e-28	4.11e-11
	VIF	1.37	1.37	1.39	1.41
Sex (Male)	HR 95% CI	2.37 (2.25 - 2.5)	2.36 (2.19 - 2.55)	2.35 (2.08 - 2.65)	2.23 (1.8 - 2.76)
	P-value	5.60e-213	3.36e-109	2.06e-43	1.54e-13
	VIF	1.28	1.3	1.32	1.33
Race	HR 95% CI	0.26 (0.18 - 0.39)	0.26 (0.16 - 0.43)	0.18 (0.07 - 0.42)	0.11 (0.01 - 0.88)
	P-value	4.01e-12	1.39e-07	8.59e-05	3.76e-02
	VIF	1.14	1.15	1.16	1.16
BMI	HR 95% CI	1.01 (1 - 1.02)	1.01 (1 - 1.02)	1.01 (0.99 - 1.02)	1 (0.98 - 1.02)
	P-value	1.13e-03	1.80e-02	2.54e-01	9.52e-01
	VIF	1.22	1.25	1.26	1.23
TDI	HR 95% CI	1.02 (1.01 - 1.03)	1.02 (1.01 - 1.03)	1.01 (0.99 - 1.03)	1 (0.97 - 1.03)
	P-value	6.32e-09	6.29e-05	2.77e-01	9.97e-01
	VIF	1.1	1.1	1.11	1.12
Smoking	HR 95% CI	1.94 (1.82 - 2.08)	1.92 (1.75 - 2.11)	1.69 (1.46 - 1.96)	1.83 (1.43 - 2.35)
	P-value	4.03e-82	2.60e-42	4.29e-12	1.74e-06
	VIF	1.22	1.23	1.24	1.27
Fasting	HR 95% CI	1 (0.99 - 1.01)	1 (0.99 - 1.01)	1.01 (0.99 - 1.03)	1.01 (0.98 - 1.04)
	P-value	7.32e-01	8.06e-01	1.61e-01	6.76e-01
	VIF	1.03	1.03	1.04	1.04
HDL (by 1-SD)	HR 95% CI	0.81 (0.79 - 0.83)	0.8 (0.77 - 0.84)	0.77 (0.72 - 0.81)	0.72 (0.65 - 0.8)
	P-value	1.14e-47	6.77e-28	8.03e-18	3.39e-10
	VIF	1.4	1.44	1.47	1.39
HbA1c (%)	HR 95% CI	1.02 (1.01 - 1.03)	1.02 (1.01 - 1.03)	1.02 (1.01 - 1.04)	1.02 (1 - 1.05)
	P-value	8.45e-10	2.39e-05	5.46e-04	5.07e-02
	VIF	1.13	1.13	1.13	1.12
Hypertension	HR 95% CI	1.38 (1.3 - 1.46)	1.3 (1.2 - 1.41)	1.2 (1.05 - 1.37)	1.06 (0.85 - 1.33)
	P-value	1.59e-26	4.36e-10	6.09e-03	5.96e-01
	VIF	1.09	1.09	1.09	1.09
SBP	HR 95% CI	1.01 (1.01 - 1.01)	1.01 (1.01 - 1.01)	1.01 (1.01 - 1.02)	1.01 (1.01 - 1.02)
	P-value	4.47e-35	9.44e-20	5.61e-11	4.78e-04
	VIF	2.3	2.28	2.28	2.31

Subsetting by any criteria can introduce unwanted selection bias in the dataset.

Adjustment variables effects remain constant for any subset chosen

Limitations and future directions

- We used a large dataset, allowing to push this technique to smaller and smaller subsets
  - Smaller datasets might benefit from a **weighted approach** where each datapoint is weighted by a function of the residual **ε** rather than using a cutoff

- Predictors in the example showed a strong **linear correlation** but other interactions may be at play in different datasets
  - The first regression step can be modified to accommodate **non-linear effects**

- What if there is a **3-way interaction**?
  - Multivariable regression at step 1 can be adopted

- Regression outliers might coincide with real distribution outliers giving more weights than necessary to extreme measures
  - Outliers removal can be run before the first step

Conclusions

- We developed **CROS-R**, a two-step procedure grounded within classical regression to resolve multicollinearity issues

- **CROS-R doesn’t require any modification in the statistical modeling** of choice. After a first linear regression between the predictors under study, any regression strategy can be utilized

- **CROS-R can manage extremely high correlation** given sufficient data (0.88 to 1 in the example, similar results in simulation)

- By relying on simple statistical models, **CROS-R can return actionable and comparable effect measures** and not simply ranking predictors by their importance

Thank You

gmelloni@bwh.harvard.edu

https://timi.org/biostatistics/