LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Reduce, score*, regress, repeat: using factor analysis to tackle multicollinear HDL metabolomics data in seven CHD studies

Amy Mulick[†]
Research Fellow, Medical Statistics Department
London School of Hygiene and Tropical Medicine
London, UK

[†]With JP Casas and David Prieto at UCL Institute of Health Informatics and George Ploubidis at UCL Institute of Education

Cholesterol: bad and good?

Observational and causal evidence shows that low levels of 'bad' cholesterol (carried in low-density lipoproteins, LDL-C) is associated with a lower risk of coronary heart disease (CHD)

Frustratingly, the same has not been shown for 'good' (high-density, HDL-C) cholesterol. Causal studies failed to prove that high levels reduce this risk

Maybe HDL-C is the wrong biomarker.

<u>Hypothesis: more information from the HDL pathway is needed to adequately describe the risk of CHD</u>

Metabolomic components of HDL

HDL-C

XL-HDL = CE + FC + TG + PL

L-HDL = CE + FC + TG + PL

M-HDL = CE + FC + TG + PL

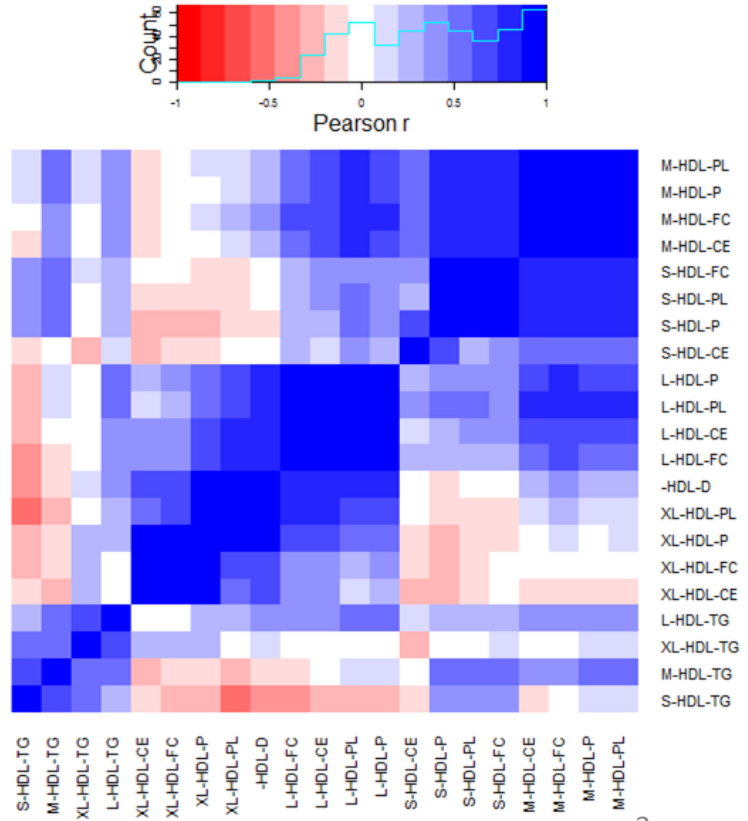
S-HDL = CE + FC + TG + PL

CE: cholesterol esters

FC: free cholesterol

TG: triglycerides

PL: phospholipids



Collinearity

New (independent) biomarkers often identified through joint modelling, but this does not work in the presence of strong collinearity.

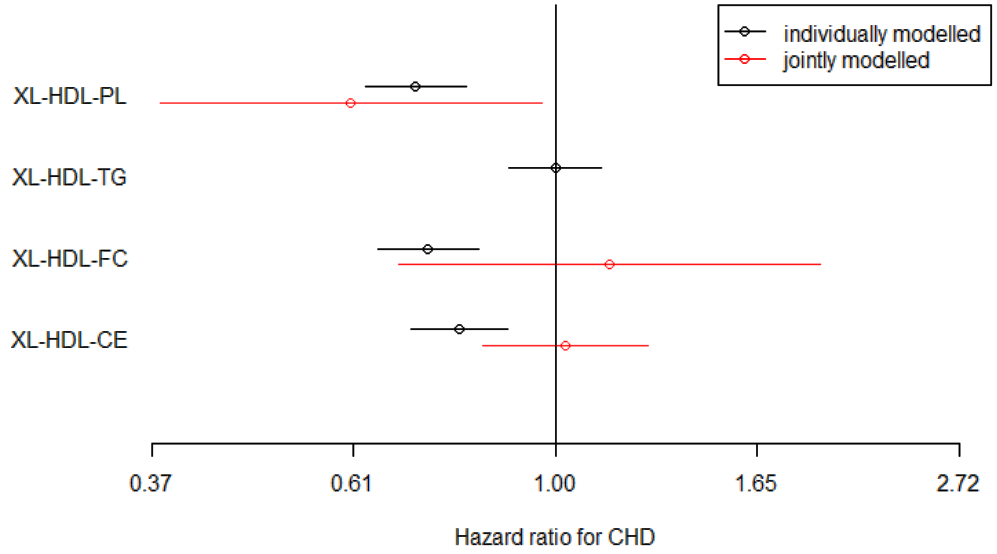
With high-dimensional data, it may be more appropriate to think of biomarkers as patterns of expressions.

Methodological problem:

It is unclear how to detect patterns when data come from multiple studies.

Pearson correlations

	XL-HDL-CE	XL-HDL-FC	XL-HDL-TG
XL-HDL-FC	**0.84**		
XL-HDL-TG	0.37	0.34	
XL-HDL-PL	**0.81**	**0.97**	0.30



N=3780; CHD events=313

Analysis plan

REDUCE: perform Factor Analysis of HDL metabolites, reducing them to a smaller number of latent factors (metabolite patterns)

SCORE*: use factor analysis solution to predict values (scores) for the latent factors

REGRESS: model latent factor scores in covariate-adjusted Cox regression

REPEAT: do this for all seven studies

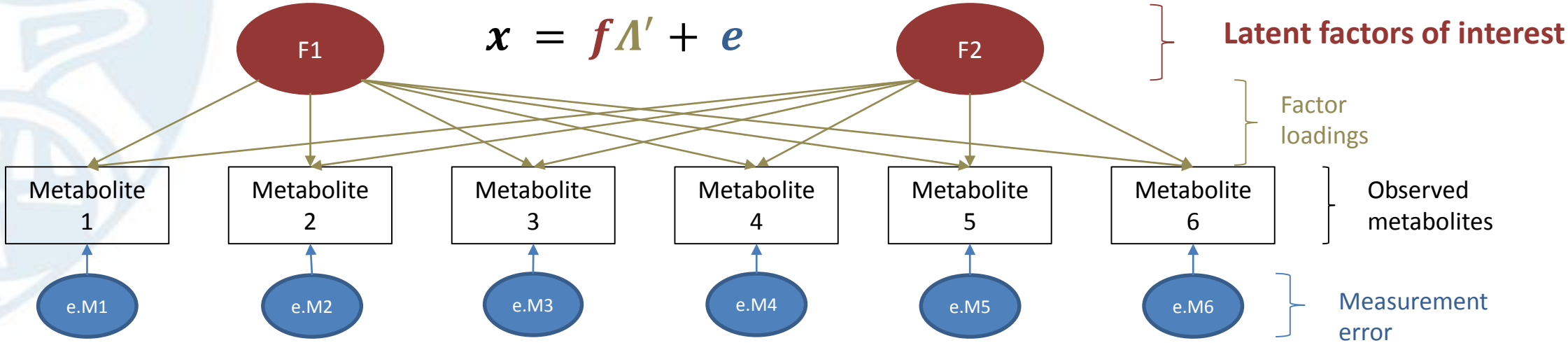
Then:

Pool log hazard ratios in a meta-analysis

(Exploratory) Factor Analysis

Suppose *p* metabolites are expressed by *k*<*p* underlying metabolic processes

Factor analysis can estimate qualitative and quantitative information on them



LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

(Exploratory) Factor Analysis

$$x = f\Lambda' + e$$

STEP 1: Λ ESTIMATION

We estimate Λ indirectly by estimating $\mathbf{e}$ (details omitted) using maximum likelihood

STEP 2: Λ ROTATION

For interpretability we obliquely rotate $\widehat{\Lambda}$ using the quartimin criterion, producing $\widehat{\Lambda}^*$

STEP 3: $\mathbf{f}$ PREDICTION

$\mathbf{f}$ is predicted from $\widehat{\Lambda}^*$ and $\widehat{\Phi}$ (the k x k factor correlation matrix) using the 'regression' method: $\hat{f} = \widehat{\Phi}\widehat{\Lambda}^{*\prime}\Sigma^{-1}x$

The 'loadings' of $\widehat{\Lambda}^*$ are measures of association between observed variables and latent factors

SCORE*

***It is unclear how to handle this across multiple studies.**

To predict the 'same' thing, we need one estimate of $\widehat{\boldsymbol{\Phi}}$ and $\widehat{\boldsymbol{\Lambda}}^*$.

But seven studies = seven solutions. Which to use?

$$\hat{f} = \widehat{\boldsymbol{\Phi}}\widehat{\boldsymbol{\Lambda}}^{*\prime}\boldsymbol{\Sigma}^{-1}\boldsymbol{x}$$

We considered some options:

- Perform a Confirmatory Factor Analysis and validate the first study's factors.
  - **Yes, but no**. Collinearity too strong. Failed to converge on a solution
- Pool all IPD, get one correlation matrix (Σ), and perform one factor analysis?
  - **No**. Difficult to define the 'population', mean metabolite concentrations bound to differ.
- Don't pool IPD but get seven correlation matrices and pool them, then perform one factor analysis?
  - **No**. Correlation matrices are positive semi-definite and no guarantee that pooling them retains this necessary property.
- Perform seven factor analyses, pool the loading matrices (Λ) and factor correlation matrix (Φ), and use these to predict scores?
  - **No**. Unknown how to do this in a principled way. Difficult if more studies added.
- Perform seven factor analyses and predict scores separately in each study?
  - **Yes**. We compare $\widehat{\boldsymbol{\Lambda}}^*$ for 3-, 4-, 5-factor solutions between the studies

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Survival analysis

Factor scores (for 3-, 4- and 5-factor solutions)

- scaled to unit SD
- modelled jointly in an age-adjusted Cox regression model
- restricted to individuals free from CHD at baseline and with complete data
- progressive adjustments by known/probable confounders: sex, ethnicity, smoking, systolic blood pressure (SBP), BMI, diabetes, LDL-C

CHD

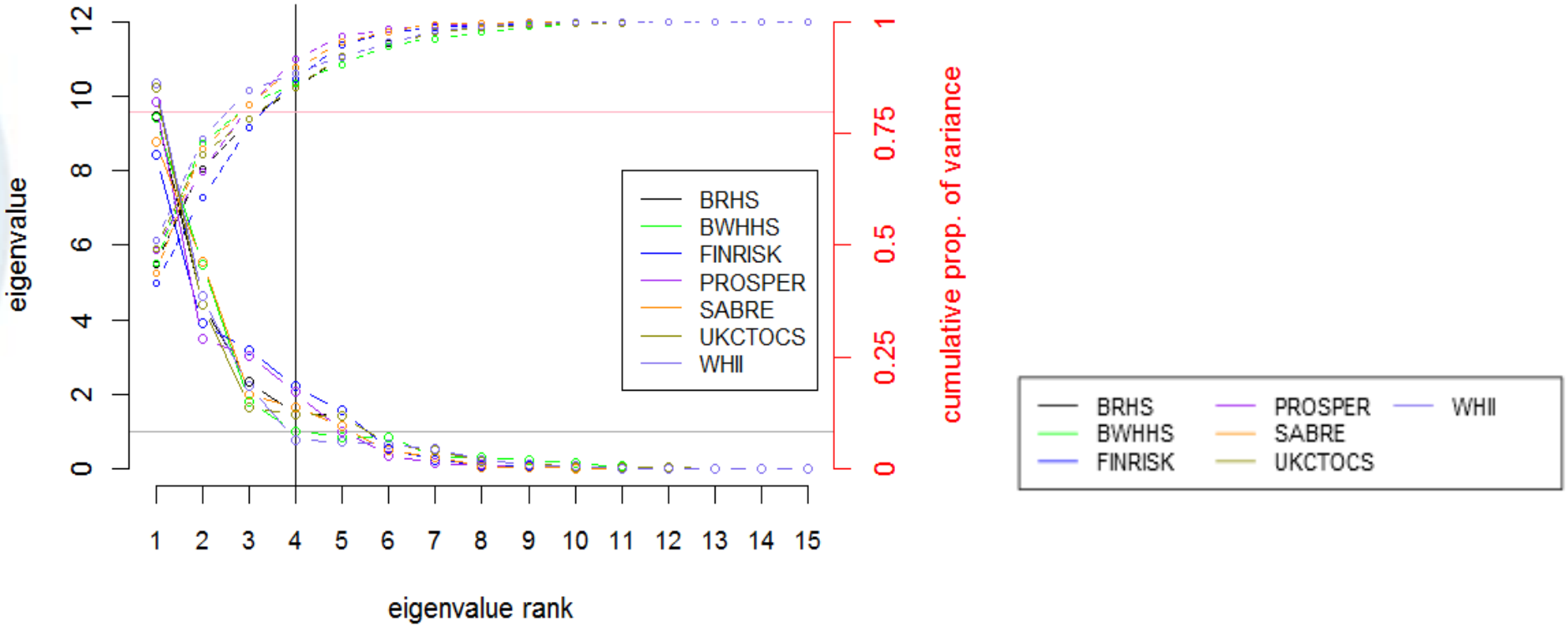
- fatal
- non-fatal, generally
  - myocardial infarction (MI)
  - coronary artery bypass graft (CABG)
  - percutaneous transluminal coronary angioplasty (PTCA)

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

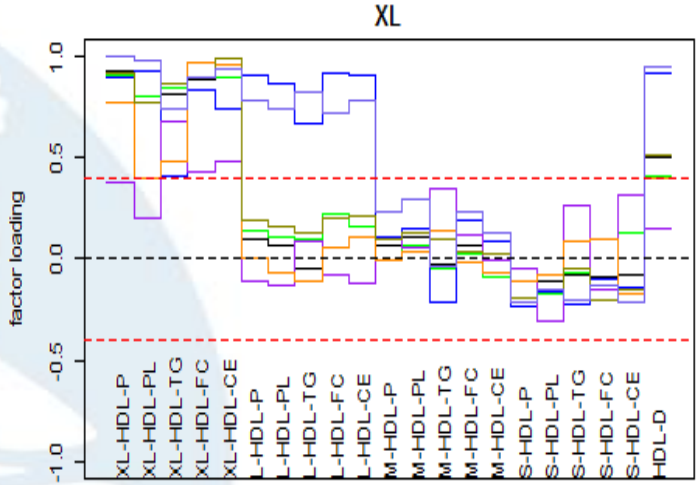
Seven studies

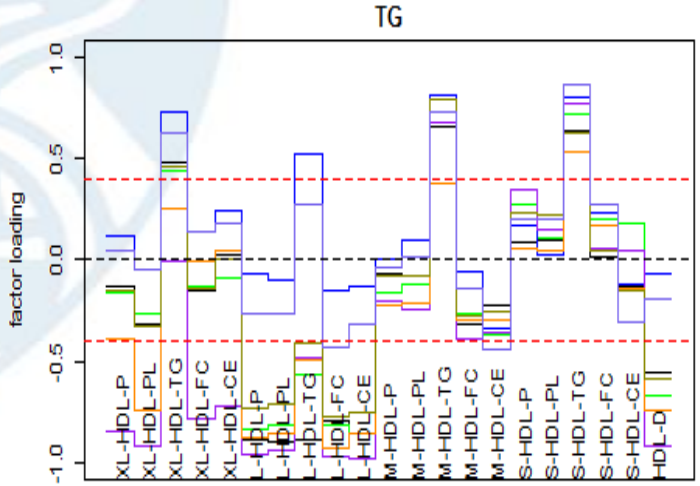
STUDY	STUDY TYPE	N	GENDER	MEAN AGE (years)	MEAN FOLLOW-UP (years)
BRHS	cohort	3965	Men	69	9
BWHHS	cohort	3777	Women	69	10
FINRISK (1997)	population cohort	7602	Both	48	13
PROSPER	RCT (statin)	5359	Both	76	3
SABRE	cohort	3297	Both	52	17
UKCTOCS	RCT (cancer screening): nested case-control	3194	Women	65	5
WHII (Wave 5)	cohort	6170	Both	56	6
TOTAL	**mixed**	**33 364**	**47% female**	**61**	**8.9**

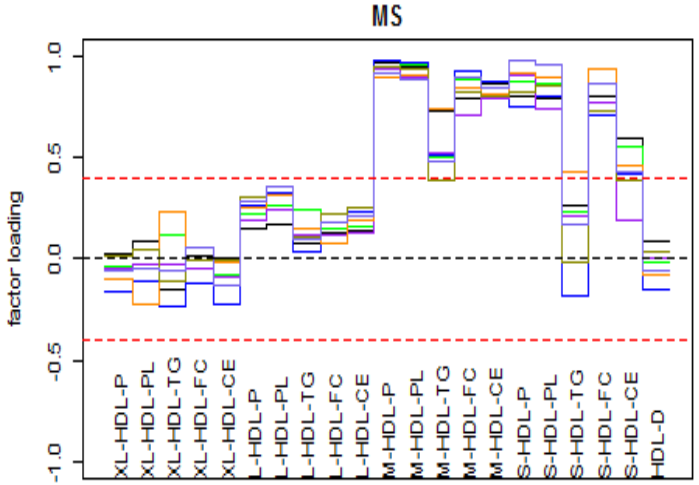
Factor Analysis: variability explained



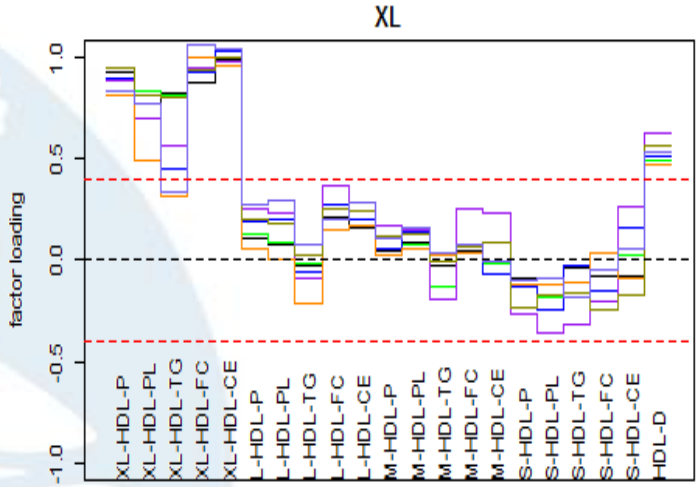
RESULTS: FACTOR ANALYSIS 3

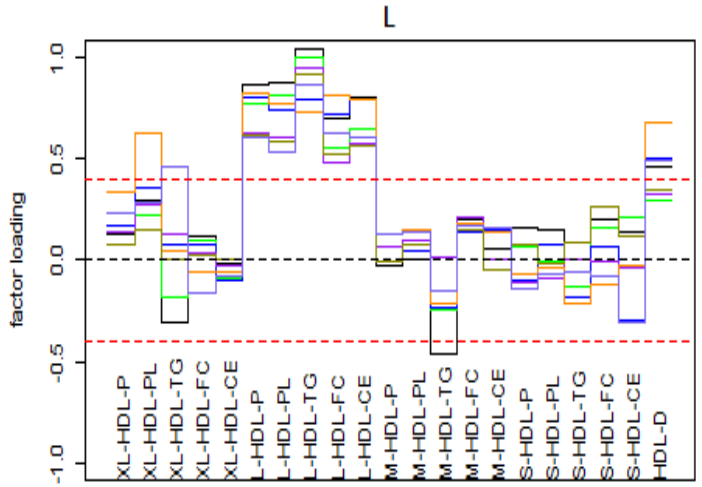


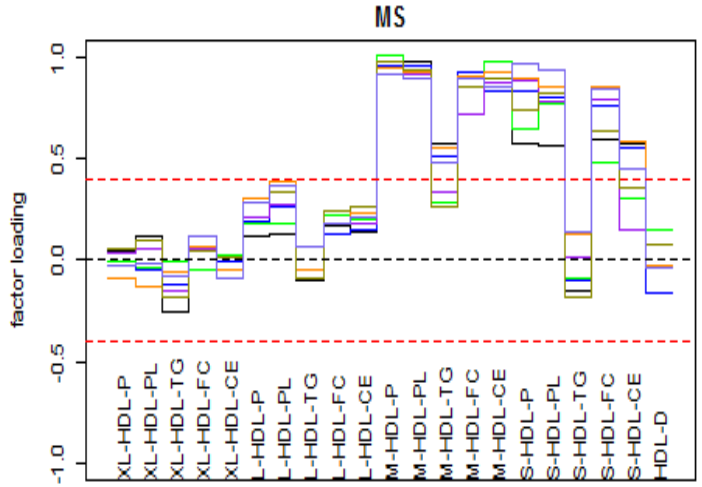


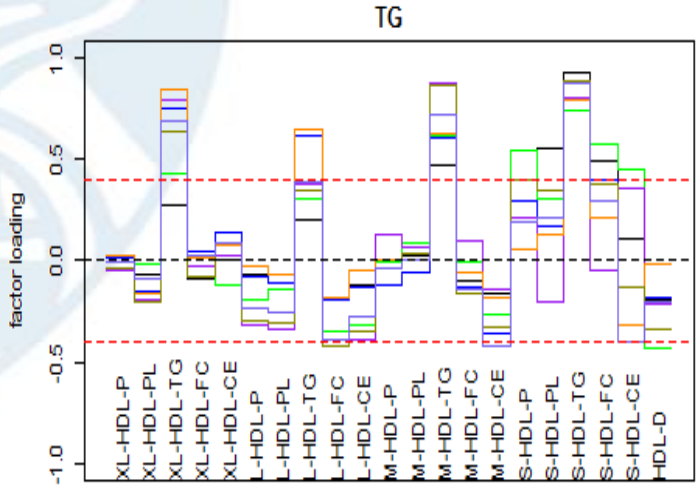


RESULTS: FACTOR ANALYSIS 4

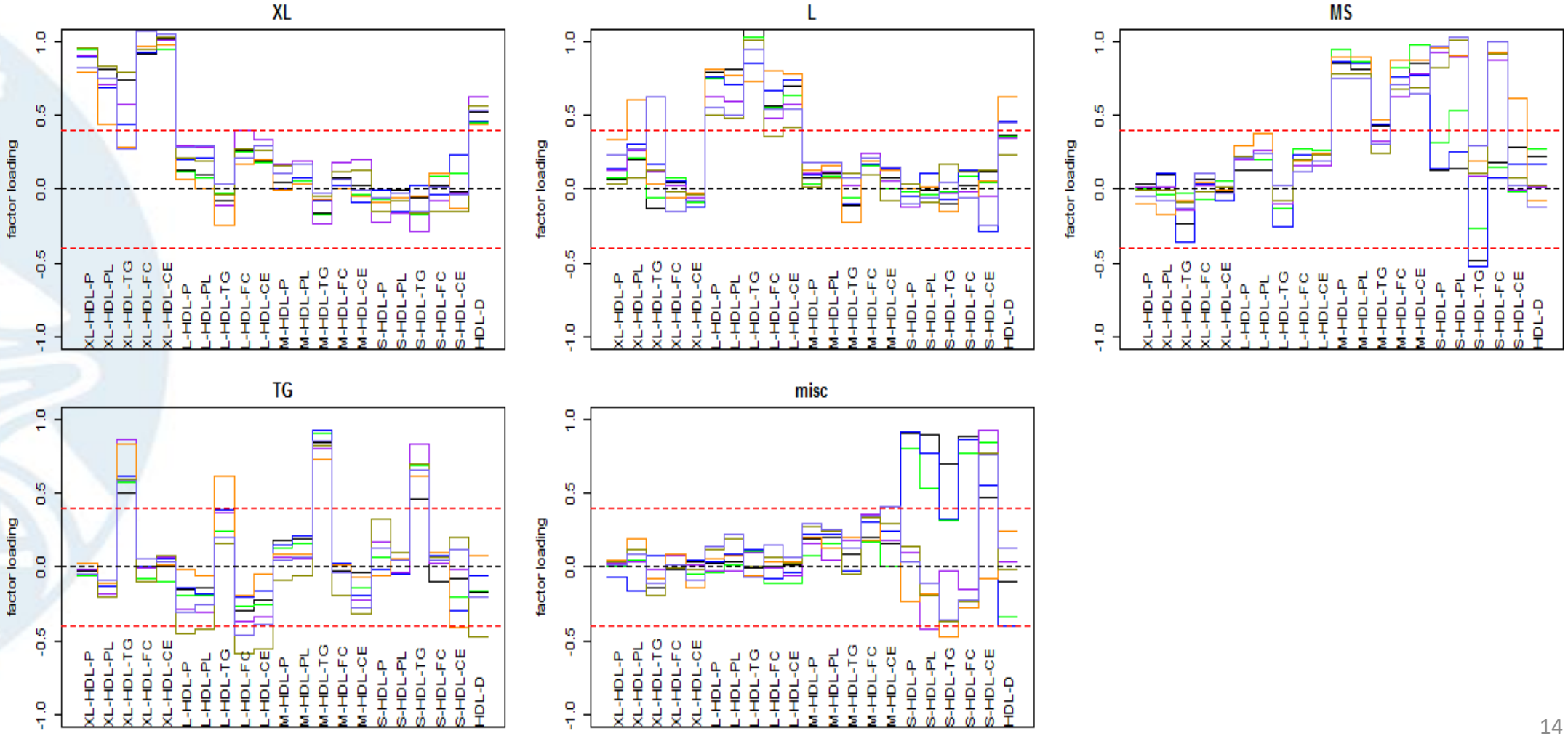








RESULTS: FACTOR ANALYSIS 5



RESULTS: FACTOR ANALYSIS

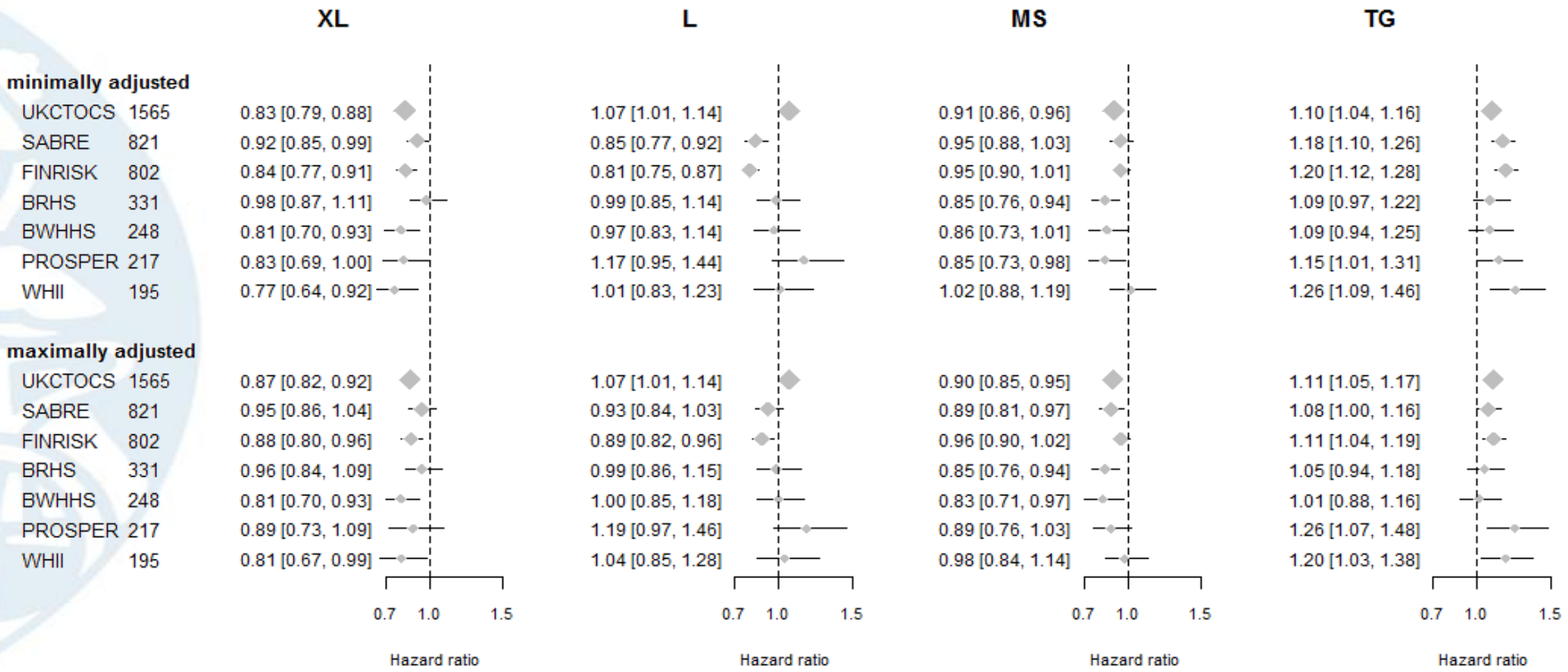
1. 4 factors explain a sufficient amount of variance and composition is nearly identical in all seven studies, therefore

2. This is evidence for the presence of 4 consistent patterns of HDL expression, so

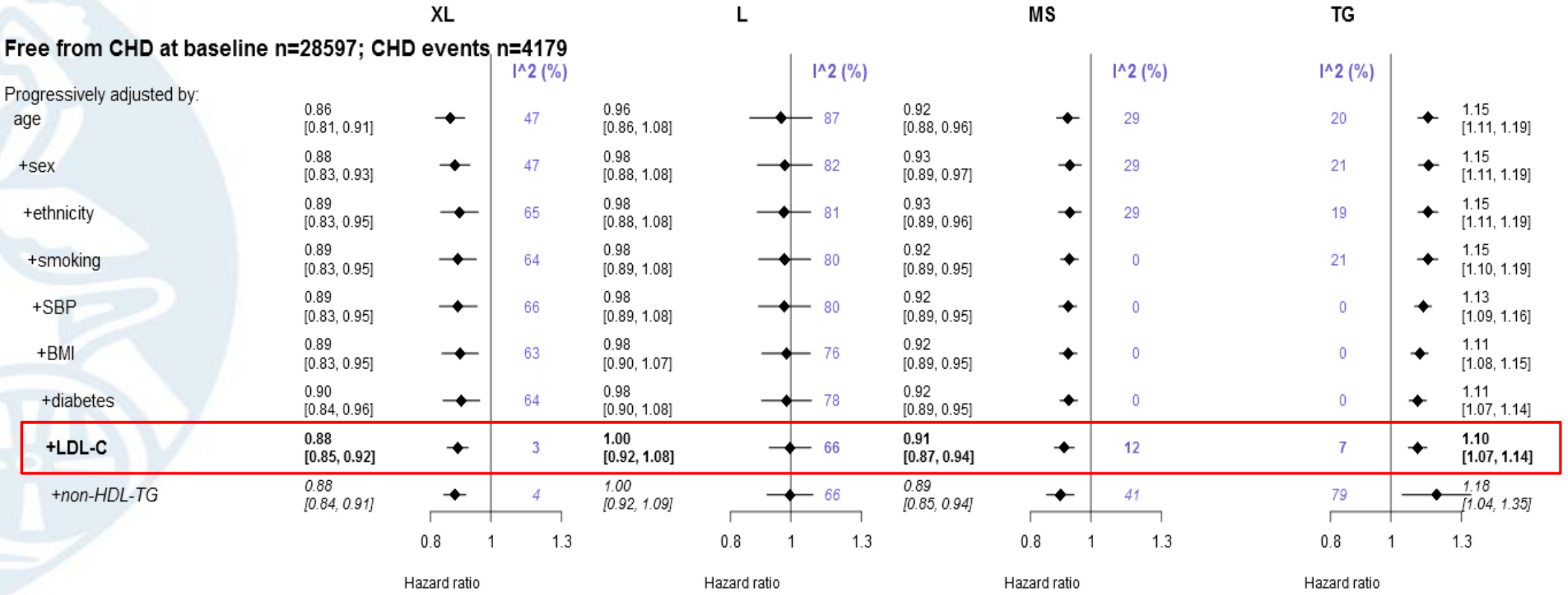
3. We can accept the small degree of variability this might have added to our regression results by estimating and predicting them separately in each study, and

4. We can combine results in a random-effects meta-analysis
   - log HRs pooled using inverse variance method
   - Between-study heterogeneity estimated using method of DerSimonian and Laird and reported with I^2 statistic

RESULTS: META-ANALYSIS



RESULTS: META-ANALYSIS



SUMMARY

- We used (exploratory) factor analysis to estimate patterns in the HDL pathway from highly-dimensional metabolomics data in seven datasets
- We identified four patterns that were remarkably consistent across very diverse studies
- Three were associated with the incidence of CHD, one of which was in the opposite direction to the other two
- Our study shows that our present understanding of the relationship between HDL-C and CHD may be oversimplified.

NEXT STEPS (ONGOING)

Compare these HDL metabolomic patterns with HDL genomics.

ACKNOWLEDGMENTS

Research directed by **Professor JP Casas** and statistical analysis directed by **Dr David Prieto** at **UCL Institute of Health Informatics** with funding from the **British Heart Foundation**

Special thanks to:

Aroon Hingorani	**UCLEB**
Jorgen Engmann	**UCLEB**
Tina Shah	**UCLEB**
Therese Tillin	**SABRE**
Mika Kivimaki	**WHII**
Aki Havulinna	**FINRISK**
Veikko Salomaa	**FINRISK**
Charles Boachie	**PROSPER**
Naveed Sattar	**PROSPER**
Goya Wannamethee	**BRHS**
Barbara Jefferis	**BRHS**
Usha Menon	**UKCTOCS**
Andy Ryan	**UKCTOCS**
Aleksandra Gentry-Maharaj	**UKCTOCS**

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Factor Analysis

Goal to estimate **f** for every observation

The 'common factor' model is (for one observation):

$$\boldsymbol{x} = \boldsymbol{f}\boldsymbol{\Lambda}' + \boldsymbol{e}$$

Where

x is a (1 x p) vector of observed metabolites

f is a (1 x k) vector of latent factors

$\boldsymbol{\Lambda}$ is a (p x k) 'loading' matrix

e is a (1 x p) vector of 'uniquenesses'

Factor Analysis

$$x = f\Lambda' + e$$

STEP 1: ESTIMATION

The factor analysis algorithm estimates $\mathbf{e}$, and thus $\mathbf{\Lambda}$, using the fact that the correlation matrix $\mathbf{\Sigma}$ of the p observed variables can be decomposed into:

$$\Sigma = \Lambda\Lambda' + \Psi$$

Where $\mathbf{\Psi}$ is a (p x p) diagonal matrix of $\mathbf{e}$.

$\mathbf{\Lambda}$ is constructed from the k leading eigenvectors of $\mathbf{\Sigma - \Psi}$ after choosing the 'best' $\mathbf{\Psi}$ using, e.g., maximum likelihood estimation

Factor Analysis

$$\Lambda^* = \Lambda(T')^{-1}$$

STEP 2: Λ ROTATION

Λ is rotated for interpretability: the matrix is transformed by re-projecting its coordinates in Euclidean space. 'Looking at the data from a different angle'

The (rotated) loadings occur generally between -1 and 1 and are measures of association between observed variables and latent factors

Solution obliquely rotated (allowing the final factors to be correlated) using the quartimin criterion

Variable	Factor1	Factor2	Factor3	Factor4
metabolite1	0.3806	-0.0559	-0.0742	**0.7706**
metabolite2	**0.7207**	-0.0533	-0.2440	0.3744
metabolite3	-0.1958	0.2444	-0.0236	**0.4145**
metabolite4	-0.0099	0.1222	-0.0433	**0.9643**
metabolite5	-0.0396	-0.1256	0.1407	**0.9907**
metabolite6	**0.9388**	0.0683	0.1124	-0.0124
metabolite7	**0.9173**	0.0960	0.1583	-0.0792
metabolite8	**0.5730**	0.0669	0	-0.1672
metabolite9	**0.9596**	-0.0666	0.0702	0.0648
metabolite10	**0.9070**	0.0473	0.0622	0.0939

SCORE*

…

$$\hat{f} = \hat{\Phi}\hat{\Lambda}^{*\prime}\Sigma^{-1}x$$

We decided to perform the factor analysis and predict factor scores separately in ALL studies

Pilot data suggested there would be between 3 and 5 factors: we compare Λ results for those solutions between the studies

If we find ‘same’ factors, we use within-study Λ to predict scores within studies and accept the small degree of variability this might add to our regression results between studies