DOI:
10.1186/s12860-019-0210-7

Document Version
Peer reviewed version

Metabolic profiles to predict long-term cancer and mortality: the use of latent class analysis

Aida Santaolalla[1], Hans Garmo[1, 2], Anita Grigoriadis[3], Sundeep Ghuman[1,4], Niklas Hammar[5,], Ingmar Jungner[6], Göran Walldius[7], Mats Lambe[8], Lars Holmberg[1], Mieke Van Hemelrijck[1, 7]

Affiliations:

1. King's College London, School of Cancer & Pharmaceutical Sciences, Translational Oncology and Urology Research, London, UK
2. Regional Oncologic Centre, Uppsala University, Uppsala, Sweden
3. King's College London, School of Cancer & Pharmaceutical Sciences, Cancer Bioinformatics, Breast Cancer Now, London, UK
4. Guy's and St Thomas' NHS Foundation Trust
5. Department of Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
6. Department of Medicine, Clinical Epidemiological Unit, Karolinska Institutet and CALAB Research, Stockholm, Sweden
7. Unit of Cardiovascular Epidemiology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden
8. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Corresponding author: sundeep.ghuman@kcl.ac.uk (SG)

Email addresses co-authors: hans.garmo@kcl.ac.uk; anita.grigoriadis@kcl.ac.uk; sundeep.ghuman@kcl.ac.uk ;niklas.hammar@ki.se; ingmar.jungner@ki.se; goran.walldius@ki.se; mats.lambe@ki.se; lars.holmberg@kcl.ac.uk; mieke.vanhemelrijck@kcl.ac.uk

Word count: 250 (abstract) and 3742 (manuscript).

Abstract

Background: Metabolites are genetically and environmentally determined. Consequently, they can be used to characterize environmental exposures and reveal biochemical mechanisms that link exposure to disease. To explore disease susceptibility and improve population risk stratification, we aimed to identify metabolic profiles linked to carcinogenesis and mortality and their intrinsic associations by characterizing subgroups of individuals based on serum biomarker measurements. We included 13,615 participants from the Swedish Apolipoprotein MOrtality RISk Study who had measurements for 19 biomarkers representative of central metabolic pathways. Latent Class Analysis (LCA) was applied to characterise individuals based on their biomarker values (according to medical cut-offs), which were then examined as predictors of cancer and death using multivariable Cox proportional hazards models.

Results: LCA identified four metabolic profiles within the population: (1) normal values for all markers (63% of population); (2) abnormal values for lipids (22%); (3) abnormal values for liver functioning (9%); (4) abnormal values for iron and inflammation metabolism (6%). All metabolic profiles (classes 2-4) increased risk of cancer and mortality, compared to class 1 (e.g. HR for overall death was 1.26 (95%CI: 1.16 - 1.37), 1.67 (95%CI: 1.47 - 1.90), and 1.21 (95%CI: 1.05 - 1.41) for class 2, 3, and 4, respectively).

Conclusion: We present an innovative approach to risk stratify a well-defined population based on LCA metabolic-defined subgroups for cancer and mortality. Our results indicate that standard of care baseline serum markers, when assembled into meaningful metabolic profiles, could help assess long term risk of disease and provide insight in disease susceptibility and etiology

Keywords:

Risk stratification, biomarkers, metabolic profiles, latent class analysis, disease susceptibility, cancer epidemiology.

Background

Cancer is a multi-pathway disease, assembled as a heterogeneous and hierarchically organized system, and still one of the major causes of death worldwide – with an increasing burden given the aging population (1-3). Cancer data has grown exponentially in the last decade with new advanced technologies resulting in highly diverse, mixed data types and huge volumes of information (e.g.: 542045 is the number of publications retrieved in PubMed when searching the terms 'cancer' AND 'data' on August 2017). Due to the nature of this emerged "Big Cancer Data" and the demand for high-sensitive and high-specific biomarkers, there is a need for significant sample sizes and advanced mathematical and statistical models (4, 5) capable of extracting relevant clinical and biological information (6, 7). These more systematic-based approaches, replacing single biomarker analyses by multiple profiling testing, may provide new avenues for biomarker development in cancer diagnosis and management (8, 9). Recent studies have adopted these integrative approaches assessing multiple serum markers simultaneously for cancer diagnosis (10-13). Furthermore, the concept of the exposome has been introduced into the field of cancer epidemiology (14). It refers to every non-genetic exposure to which an individual is subjected from conception to death (14, 15) . Specifically, metabolites, part of the internal exposome, are both genetically and environmentally determined and can consequently be used to characterize environmental exposures and reveal biochemical mechanisms that link exposure to disease (15-18). Hence, the internal distribution of metabolites and their interactions might help unravelling cancer susceptibility in a population.

With the overall goal of identifying statistical methods to stratify individuals based on their underlying risk of developing cancer and risk of increasing mortality, we conducted a data driven approach utilizing standard serum markers available from routine health check-ups to study susceptibility to cancer and death in a well-defined cohort of 13,615 participants from the AMORIS study (Apolipoprotein MOrtality RISk) (19, 20). More specifically, the study was set out to explore population heterogeneity and cancer susceptibility by investigating serum metabolic profiles using latent class analysis (LCA). This data reduction method clusters covariates based on models of data

distribution probabilities. It allows for evaluation of clusters of biomarkers linked to carcinogenesis and their intrinsic associations, which ultimately helps us assess their possible role in predicting long-term cancer and mortality.

Results

Characteristics of the study population

A total of 1,956 individuals (14.37%) developed cancer after at least 3 years of follow-up, including 655 breast and genito-urinary cancers, 330 cases of digestive cancer, 133 cases of respiratory cancers and 129 lymphatic and hematopoietic cancers during a mean follow-up time for cancer of 16.6 years, median follow-up time in the cohort of 17.22 years with a minimum of 3.01 years and a maximum of 24.77. 3,158 participants (23.20%) died during a mean follow-up of 17.3 years, comprising 706 cancer-specific deaths. Study population characteristics by cancer status is illustrated in Table 1.

Table 1| Characteristics of the study population by cancer status defined at the end of the follow up period. All the serum markers are dichotomized using standard clinical cut-offs.

	Total N=13,615 (100%)	**No Cancer N=11,659 (85.63%)**	**Cancer N=1,956 (14.37%)**
Age (years)			
Mean (SD)	51.91 (14.80)	50.86 (15.00)	58.14 (11.75)
Under 40	2951 (21.67)	2841 (24.37)	110 (5.62)
40-50	3550 (26.07)	3148 (27.00)	402 (20.55)
50-60	3065 (22.51)	2491 (21.37)	574 (29.35)
Above 60	4049 (29.74)	3179 (27.27)	870 (44.48)
Sex			
Female	7588 (55.73)	6636 (56.92)	952 (48.67)
Male	6027 (44.27)	5023 (43.08)	1004 (51.33)
Socio-economics Status			
High	6493 (47.69)	5416 (46.45)	1077 (55.06)
Low	5007 (36.78)	4368 (37.46)	639 (32.67)
Not employed or missing	2115 (15.53)	1875 (16.08)	240 (12.27)
Educational Status			
High	4313 (33.42)	3688 (33.40)	625 (33.57)

Middle	5495 (42.58)	4725 (42.79)	770 (41.35)
Low	3097 (24.00)	2630 (23.82)	467 (25.08)
Missing [b]	710 (5.21)	616 (5.28)	94 (4.80)
CCI			
0	12258 (90.03)	10520 (90.23)	1738 (88.85)
1	963 (7.07)	807 (6.92)	156 (7.98)
2	221 (1.62)	188 (1.61)	33 (1.69)
3+	173 (1.27)	144 (1.24)	29 (1.48)
Total Cholesterol (mmol/L)			
Mean(SD)	5.82 (1.17)	5.79 (1.18)	6.00 (1.13)
< 6.50	9774 (71.79)	8453 (72.50)	1321 (67.54)
≥ 6.50	3841 (28.21)	3206 (27.50)	635 (32.46)
Triglycerides (mmol/L)			
Mean(SD)	1.44 (1.00)	1.43 (1.00)	1.48 (0.93)
< 1.71	10128 (74.39)	8716 (74.76)	1412 (72.19)
≥ 1.71	3487 (25.61)	2943 (25.24)	544 (27.81)
Apolipoprotein A-1 (g/L)			
Mean(SD)	1.44 (0.23)	1.44 (0.23)	1.43 (0.23)
< 1.05	328 (2.41)	278 (2.38)	50 (2.56)
≥ 1.05	13287 (97.59)	11381 (97.62)	1906 (97.44)
Apolipoprotein B (g/L)			
Mean(SD)	1.22 (0.35)	1.22 (0.35)	1.29 (0.34)
< 1.50	10902 (80.07)	9431 (80.89)	1471 (75.20)
≥ **1.50**	2713 (19.93)	2228 (19.11)	485 (24.80)
HDL Cholesterol (mmol/L)			
Mean(SD)	1.54 (0.43)	1.54 (0.43)	1.52 (0.43)
< 1.03	1457 (10.70)	1231 (10.56)	226 (11.55)
≥ 1.03	12158 (89.30)	10428 (89.44)	1730 (88.45)
LDL Cholesterol (mmol/L)			
Mean(SD)	3.64 (1.06)	3.61 (1.06)	3.82 (1.04)
< 4.10	9345 (68.64)	8128 (69.71)	1217 (62.22)
≥ **4.10**	4270 (31.36)	3531 (30.29)	739 (37.78)
Glucose (mmol/L)			
Mean(SD)	5.22 (1.53)	5.21 (1.53)	5.30 (1.53)
< 6.11	12223 (89.78)	10488 (89.96)	1735 (88.70)
≥ **6.11**	1392 (10.22)	1171 (10.04)	221 (11.30)
Fructosamine (mmol/L)			
Mean(SD)	2.09 (0.27)	2.08 (0.27)	2.10 (0.25)
< 2.6	13184 (96.83)	11291 (96.84)	1893 (96.78)
≥ **2.6**	431 (3.17)	368 (3.16)	63 (3.22)
GGT (IU/L) *			
Mean(SD)	33.21 (48.12)	32.74 (48.09)	36.03 (48.21)
Normal (<18)	5511 (40.48)	4827 (41.40)	684 (34.97)
Normal high (18-36)	4983 (36.60)	4236 (36.33)	747 (38.19)
Elevated (36-72)	2098 (15.41)	1750 (15.01)	348 (17.79)

Highly elevated (>72)	1023 (7.51)	846 (7.26)	177 (9.05)
AST (IU/L)			
Mean(SD)	22.84 (19.23)	22.70 (19.60)	23.64 (16.88)
< 45	13155 (96.62)	11271 (96.67)	1884 (96.32)
≥ 45	460 (3.38)	388 (3.33)	72 (3.68)
ALT (IU/L)			
Mean(SD)	29.02 (34.35)	28.95 (35.73)	29.41 (24.54)
< 50	12296 (90.31)	10546 (90.45)	1750 (89.47)
≥ **50**	1319 (9.69)	1113 (9.55)	206 (10.53)
Albumin (g/L)			
Mean(SD)	43.05 (2.82)	43.13 (2.83)	42.58 (2.72)
<35	28 (0.21)	23 (0.20)	5 (0.26)
>35	13587 (99.79)	11636 (99.80)	1951 (99.74)
Leukocytes (10^9 cells/L)			
Mean(SD)	6.52 (1.97)	6.49 (1.96)	6.65 (2.01)
<10	12956 (95.16)	11106 (95.26)	1850 (94.58)
≥ **10**	659 (4.84)	553 (4.74)	106 (5.42)
C-Reactive Protein (mg/L)			
Mean(SD)	5.86 (15.14)	5.82 (14.25)	6.16 (19.58)
<10	11858 (87.1)	10193 (87.43)	1665 (85.12)
10-15	1196 (8.78)	993 (8.52)	203 (10.38)
15-25	265 (1.95)	223 (1.91)	42 (2.15)
25-50	200 (1.47)	167 (1.43)	33 (1.69)
>50	96 (0.71)	223 (0.71)	13 (0.66)
Iron (µmol/L) *			
Mean(SD)	18.13 (5.80)	18.13 (5.83)	18.11 (5.59)
Low	636 (4.67)	540 (4.63)	96 (4.91)
Normal	12512 (91.90)	10715 (91.90)	1797 (91.87)
High	467 (3.43)	404 (3.47)	63 (3.22)
TIBC (mg/dL) *			
Mean(SD)	0.39 (0.11)	0.31 (0.11)	0.31 (0.10)
Low	4067 (29.87)	3494 (29.97)	573 (29.29)
Normal	6650 (48.84)	5683 (48.74)	967 (49.44)
High	2898 (21.29)	2482 (21.29)	416 (21.27)
Creatinine (µmol/L) *			
Mean(SD)	79.65 (16.16)	79.38 (16.37)	81.26 (14.74)
Low	40 (0.29)	31 (0.27)	9 (0.46)
Normal	12088 (88.78)	10392 (89.13)	1696 (86.71)
High	1487 (10.92)	1236 (10.60)	251 (12.83)
Phosphate (mmol/L) *			
Mean(SD)	1.07 (0.17)	1.07 (0.17)	1.05 (0.17)
Low	95 (0.70)	76 (0.65)	19 (0.97)
Normal	12796 (93.98)	10948 (93.90)	1848 (94.48)
High	724 (5.32)	635 (5.45)	89 (4.55)
Calcium (mmol/L) *			

Mean(SD)	2.38 (0.09)	2.38 (0.09)	2.38 (0.10)
Low	191 (1.40)	167 (1.43)	24 (1.23)
Normal	13195 (96.92)	11300 (96.92)	1895 (96.88)
High	229 (1.68)	192 (1.65)	37 (1.89)
Log (triglycerides/HDL) c			
mean(SD)	(-)0.19 (0.81)	(-)0.20 (0.82)	(-)0.14 (0.80)
< 0.5	11197 (82.24)	9618 (82.49)	1579 (80.73)
≥ 0.5	2418 (17.76)	2041 (17.51)	377 (19.27)
ApoB/ApoA-I c			
mean(SD)	0.87 (0.29)	0.87 (0.29)	0.92 (0.30)
< 1.00	9584 (70.39)	8347 (71.59)	1237 (63.24)
≥ 1.00	4031 (29.61)	3312 (28.41)	719 (36.76)
Life Status			
Alive	10457 (76.80)	9385 (80.50)	1072 (54.81)
Death	3158 (23.20)	2274 (19.50)	884 (45.19)
Cancer	1956 (14.90)	11659 (0.00)	1956 (100.00)

The following abbreviations have been used in Table 1: High Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Gamma-Glutamyl transferase (GGT), Alanine aminotransferase (ALT), Aspartate aminotransferase (AST) and Total iron binding capacity (TIBC).

a Clinically abnormal cut-off values are highlighted for each biomarker.

b The missing values are not included in the percentage of the Educational Status categories

c Ratios are dimensionless

*Clinical cut-offs

The following cut-offs criteria was applied:

GGT reference interval:

Low [GGT < 18 IU/L]

Normal high [18 IU/L ≥ GGT <36 IU/L]

Elevated [36 IU/L ≥ GGT <72 IU/L]

High elevated [GGT ≥ 72 IU/L]

Iron reference interval:

Men [Low ≤ 11, Normal = 11-31, High ≥ 31]

Women [Low ≤ 9, Normal = 9-30, High≥ 30]

TIBC reference interval:

Men [Low ≤ 0.257, Normal = 0.257-0.379, High ≥ 0.379]

Women [Low ≤ 0.246, Normal = 0.246- 0.391, High ≥ 0.391]

Creatinine reference interval:

Men [Low ≤ 60, Normal = 60-100, High ≥ 100]

Women [Low ≤ 45, Norma l= 45-90, High ≥ 90]

Phosphate reference interval:

Men [Low ≤ 0.7, Normal = 0.7-1.4, High ≥ 1.4]

Women [Low ≤ 0.8, Normal = 0.8-1.4, High ≥1.4]

Calcium reference interval per gender by age:

Men

[Age < 40, Low ≤ 2.22, Normal = 2.22-2.60, High ≥2.60]

[Age 40-60, Low ≤ 2.20, Normal = 2.20 -2.59, High ≥2.59]

[Age > 60, Low ≤ 2.19, Normal= 2.19 -2.58, High ≥ 2.58]

Women

[Age < 40, Low ≤ 2.17, Normal = 2.17-2.56, High ≥2.56]

[Age 40-60, Low ≤2.19, Normal = 2.19-2.60, High ≥2.60]

[Age > 60, Low ≤ 2.21, Normal = 2.21-2.60, High ≥2.60]

Latent Class Analysis characterizes the study population into four metabolic profiles

LCA was executed using the dichotomized values of the biomarkers to facilitate the biological interpretation of the results. The Chi-squared distribution criterion for model selection indicated a best fit model comprehend of four LCA classes, while AIC and BIC stabilized at 4 classes (Figure 1A, Figure 1B) (43). All the criterions did not converge to a local maximum from class 12 onwards. The class allocation of the observations (individuals), the class conditional probability of each biomarker and the latent mixing proportions were obtained when running poLCA package in R statistical language.

Table 2 and Figure 2 outline the LCA-derived classes with the estimated class population proportions, the class conditional probabilities of belonging to each latent class for each of the biomarkers and the biological interpretation of the LCA-derived classes. The four mutually exclusive classes characterize the population in metabolic profiles based on class conditional probabilities: (1) those with probabilities for all abnormal values of the markers under 0.3; therefore, considered the normal class (63% of population); (2) those with abnormal values for lipid markers (22%); (3) those with abnormal values for liver function markers (9%); (4) those with abnormal values for iron and inflammation metabolism (6%).

A validation of the characterization of the population performed with the Latent class methodology is outlined in Appendix Table S3. The baseline clinical characteristics of the individuals by LCA-derived metabolic classes (supplementary Table S3) replicate the results displayed in Table 2 for the class conditional probabilities.

Table 2| Predicted class memberships of the clinically abnormal biomarkers cut-off values for the 4 latent classes model. Estimated class population shares for the four different LCA classes.

LCA-derived Classes	**Class 1**	**Class 2**	**Class 3**	**Class 4**
% on the population	63%	22%	9%	6%
Biological interpretation	Normal	Lipids	Liver	Iron/ Inflammation
ApoB/ApoA-I ≥ 1.00 b	0.1320	**0.6840**	0.4519	0.2480
Log (Triglycerides/HDL) ≥ 0.50 b	0.0126	**0.5436**	0.3852	0.1421
Glucose ≥6.11 mmol/L	0.0342	0.2401	0.2174	0.0919
Fructosamine ≥ 2.60 mmol/L	0.0039	0.0967	0.0555	0.0280
ALT ≥ 50 IU/L	0.0051	0.0107	**1.0000**	0.0291
GGT Elevated36-72 IU/L	0.0848	0.2532	**0.3521**	0.1732

GGT Highly elevated ≥72 IU/L	0.0240	0.0843	**0.4098**	0.0619
AST ≥ 45 IU/L	0.0052	0.0045	**0.3168**	0.0180
CRP >10 mg/L	0.0282	0.0715	0.0771	**0.2740**
Albumin <35 g/L	0.0007	0.0022	0.0024	0.0114
Leukocytes ≥ 10^9 cells/L	0.0265	0.0786	0.0438	0.1344
Iron low µmol/L	0.0001	0.0040	0.0281	**0.5527**
Iron high µmol /L	0.0404	0.0155	0.0712	0.0000
TIBC low mg/dL	0.2201	0.2807	0.2622	**1.0000**
TIBC high mg/dL	0.2438	0.1707	0.2984	0.0000
Creatinine low µmol /L	0.0022	0.0037	0.0041	0.0051
Creatinine high µmol /L	0.0822	0.1765	0.1166	0.1116
Phosphate low mmol/L	0.0078	0.0041	0.0063	0.0098
Phosphate high mmol/L	0.0425	0.0611	0.0544	0.1110
Calcium low mmol/L	0.0124	0.0092	0.0099	0.0458
Calcium high mmol/L	0.0121	0.0253	0.0299	0.0135

a High probabilities of the biomarkers to belong to a class are highlighted.
b Ratios are dimensionless

LCA derived metabolic profiles as cancer and mortality predictors

We then investigated the prediction capabilities of the four LCA-derived metabolic profiles to estimate overall cancer risk, specific cancer types risk, cancer mortality and overall mortality, assigning the reference level to the healthy metabolic profile Class 1 (Tables 3A - 3B).

Table 3A| Hazard ratios and 95 % confidence interval for the association of LCA-derived metabolic classes and overall cancer risk and cancer specific risk.

	Hazard Ratios (95% CI) [a]	**Hazard Ratios (95% CI) [b]**
Cancer Risk: All cancer types		
Number of events	1956	1956
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.09 (0.98 - 1.22)	1.05 (0.94 - 1.17)
3 - Liver	1.28 (1.10 - 1.50)	1.28 (1.09 - 1.49)
4 – Inflammation & Iron	1.17 (0.97 - 1.41)	1.17 (0.97 - 1.41)
Cancer Risk: Buccal cavity and pharynx		
Number of events	34	34
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.79 (0.77 - 4.14)	1.70 (0.73 - 1.17)
3 - Liver	2.66 (0.96 - 7.35)	2.60 (0.94 - 7.16)
4 - Inflammation & Iron	3.94 (1.38 - 11.30)	3.77 (1.31 - 10.82)

Cancer Risk: Digestive organs and peritoneum		
Number of events	133	133
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	0.83 (0.62 - 1.11)	0.83 (0.62 - 1.11)
3 - Liver	2.12 (1.54 - 2.91)	2.12 (1.54 - 2.91)
4 - Inflammation & Iron	0.86 (0.51 - 1.46)	0.86 (0.51 - 1.46)
Cancer Risk: Respiratory system		
Number of events	133	133
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.40 (0.94 - 2.08)	1.32 (0.88 -1.96)
3 - Liver	0.90 (0.44 - 1.82)	0.87 (0.43 - 1.77)
4 - Inflammation & Iron	1.48 (0.76 - 2.88)	1.46 (0.75 - 2.84)
Cancer Risk: Skin melanoma		
Number of events	205	205
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	0.78 (0.56 - 1.10)	0.78 (0.56 - 1.11)
3 - Liver	0.71 (0.40 - 1.26)	0.73 (0.41 - 1.31)
4 - Inflammation & Iron	0.70 (0.35 - 1.37)	0.70 (0.35 - 1.37)
Cancer Risk: Breast and genito-urinary organs		
Number of events	655	655
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.19 (0.99 - 1.42)	1.12 (0.94 - 1.33)
3 - Liver	1.04 (0.80 - 1.37)	1.04 (0.80 - 1.37)
4 - Inflammation & Iron	1.25 (0.91 - 1.71)	1.25 (0.91 - 1.71)
Cancer Risk: Brain & nervous system, Thyroids		
Number of events	34	34
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.01 (0.51 - 1.99)	0.96 (0.48 - 1.00)
3 - Liver	1.01 (0.38 - 2.67)	0.99 (0.38 - 2.59)
4 - Inflammation & Iron	0.92 (0.28 - 2.99)	0.91 (0.28 - 2.96)
Cancer Risk: Connective and endocrine tissue		
Number of events	56	56
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	0.65 (0.21 - 1.95)	0.64 (0.21 - 1.94)
3 - Liver	2.65 (1.00 - 7.02)	2.67 (1.01 - 7.07)
4 - Inflammation & Iron	3.00 (1.11 - 8.11)	2.96 (1.10 - 8.00)
Cancer Risk: Lymphatic and hematopoietic tissues: Hodgkin lymphoma, Non-H lymphoma, Leukemia and Myeloma		
Number of events	129	129
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.72 (1.15 - 2.56)	1.68 (1.12 - 2.51)
3 - Liver	1.65 (0.91 - 3.00)	1.68 (0.93 - 3.05)
4 - Inflammation & Iron	1.23 (0.56 - 2.68)	1.25 (0.57 - 2.73)

175
176 a Time scale adjusted for age, sex and CCI

b Age scale adjusted for age, sex and CCI

Table 3B| Hazard ratios and 95 % confidence interval for the association of LCA- derived metabolic classes and all causes death, Cancer death and CVD death.

	Hazard Ratios (95% CI) [a]	Hazard Ratios (95% CI) [b]
All causes death		
Number of events	3158	3158
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.26 (1.16 - 1.37)	1.29 (1.19 - 1.40)
3 - Liver	1.67 (1.47 - 1.90)	1.70 (1.49 - 1.93)
4 - Inflammation & Iron	1.21 (1.05 - 1.41)	1.20 (1.04 - 1.40)
Cancer death		
Number of events	706	706
1 - Normal class	1.00 (ref)	1.00 (ref)
2 - Lipids	1.22 (1.02 - 1.45)	1.20 (1.01 - 1.42)
3 - Liver	1.44 (1.11 - 1.86)	1.46 (1.13 - 1.90)
4 - Inflammation & Iron	0.93 (0.66 - 1.32)	0.93 (0.66 - 1.32)

a Time scale adjusted for age, sex and CCI
b Age scale adjusted for age, sex and CCI

All metabolic profiles increased risk of cancer and mortality compared to Class 1. For instance, individuals in Class 3 (abnormal liver function profile) had a higher risk of overall cancer (HR: 1.28 (95%CI: 1.10- 1.50)), but also a worse cancer-specific survival and overall survival as compared to those in Class 1 (Tables 3A – 3B). Class 2 (abnormal lipid profile) and Class 4 (abnormal iron markers and inflammatory) were positively associated with overall death, while Class 2 was also associated with cancer–specific death. The results were consistent for both time-scales (Tables 3A – 3B).

When assessing the risk of specific cancer types, several patterns occurred (Tables 3A –3B). Individuals in Class 2 (abnormal lipid markers) presented a higher risk of lymphatic and hematopoietic tissue cancer (HR: 1.72 (95%CI: 1.15 - 2.56)). There was a greater risk of digestive cancers in individuals in Class 3 (abnormal values of liver enzymes) (HR: 2.12 (95%CI: 1.54 - 2.91)), while individuals in Class 4 (abnormal iron markers and inflammation) were exposed to a higher risk of buccal and oral system cancers in comparison with the individuals in Class 1 (HR: 3.94 (95%CI 1.38 - 11.30)) (Table 3A).

Moreover, the connective tissue and endocrine glands cancer risk was higher in individuals grouped in liver metabolic profile (HR: 2.65 (95%CI: 1.00 - 7.02) and in participants belonging to the iron markers and inflammation (HR: 3.00 (95%CI: 1.11 - 8.11)). Similar associations were observed when using the age scale for the multivariable cox proportional hazard regression model (Table 3A – 3B).

Discussion

We demonstrated that standard of care baseline serum markers when assembled into meaningful metabolic profiles can help stratify the population for cancer risk, cancer mortality and overall mortality. More specifically, we observed that abnormal values for markers of the lipid metabolism, liver function and inflammatory and iron metabolism distinguish participants into metabolic profiles, which are predictive of long term cancer risk and/or mortality.

Metabolic profiles

Among the biological pathways addressed in our LCA, abnormalities in the lipid metabolism were the most common. Hyperlipidemia was present in about a quarter of the study population explaining the largest abnormal metabolic profile. The weight of the lipid profile in the analysis was consistent with the reported global prevalence of hypercholesterolemia among adults (37% for males and 40% for females) as reported in the Global Health Observatory in 2008 estimates by the World Health Organization (WHO) and the results from the Swedish population in the WHO MONICA project (46). Dyslipidemias are associated with higher risk of CVD and other chronic diseases such as cancer, as also observed in our study (47). Liver dysfunction, iron deficiency and altered inflammatory markers profiles also distinguished important subgroups in our study population. About 9% of our population had abnormal values for markers of liver functioning (GGT, AST and ALT), which is similar to the results obtained in a population-based survey in the United States that estimated abnormal alanine aminotransferase (ALT) was present in 9% of respondents in absence of viral hepatitis C or excessive alcohol consumption (48). Moreover, these enzymes are known to be linked to cancer because of their

role in preserving the intracellular homeostasis of the oxidative stress (49-51), which is concordant with the results of these analyses. The iron profile and inflammatory markers clustered 6% of individuals in the study, which was predominantly driven by low levels of serum iron and TIBC, as well as high levels of CRP and leukocytes. This could potentially point towards anemia of inflammation, a chronic inflammation presenting low iron values, that occurs because the iron deficiency provides the body with infection resistance, which demonstrates the tightly connection between the inflammatory response and the iron and its homeostasis (52). This condition has been reported in more than 30% of cancer patients at time of diagnosis.

Metabolic profiles as a risk factor for long term cancer and mortality

The above-described three classes of abnormal metabolic profiles were all associated with an increased risk of cancer and worse survival, as compared to the healthy class. The findings therefore confirm the key importance of these metabolisms in the maintenance of the intracellular homeostasis and how their unbalance can be related with the etiology of cancer disease and mortality (2). The LCA adapted in this study thus illustrates how a biomarker-wide approach can help assess markers of the blood exposome in the context of carcinogenesis and mortality (53) (Figure 3).

More specifically, individuals presenting **abnormal liver function** markers carried worse outcomes in terms of overall cancer risk and cancer death, and a positive association with digestive, connective and endocrine cancers diagnosis. Moreover, the participants with this profile had a higher probability of overall death. These results are consistent with previous published data. A positive association between elevated GGT and overall cancer risk, with no interaction of ALT, was found in the AMORIS cohort previously (24), and it was also reported in other large cohort studies (54, 55). These studies also found strong associations with elevated levels of GGT and digestive and respiratory cancer incidence. Elevated GGT has been associated with mortality from all causes, liver disease, cancer and diabetes, while ALT only showed associations with liver disease death in a large US cohort (56). However, in a study based on an elderly population it was found that GGT was associated with increased cardiovascular disease mortality, and ALP and AST with increased cancer-related mortality (57).

Moreover, a meta-analysis evaluating the associations between liver enzymes and all-cause mortality found positive independent associations of baseline levels of GGT and ALP with all-cause mortality (58). In the present study, the liver biomarker profile was positive associated with all the outcomes studied, suggesting a key role of this pathway in the development of cancer, probably related with its active role maintaining the intracellular redox regulation. Further investigations are necessary to establish the potential of the altered liver enzyme profile as a tool for cancer risk stratification.

Individuals allocated to the **lipid profile** presented positive associations with cancer mortality, and overall mortality and higher risk of lymphatic and hematopoietic cancers. The link between hyperlipidemia and mortality has been studied broadly, with associations with established links for cancer and all-cause mortality (59-61). The association between lipids and lymphatic and hematopoietic cancers is more controversial, as other studies found an inverse association for these cancers and high levels of serum cholesterol (62, 63). However, a systematic literature review from 2016 found no association (64).

Participants clustered in the **unbalanced iron profile and inflammation** had an increased risk of endocrine, buccal and oral cancers and were observed to have a higher risk of all-causes death. Altered inflammation and iron metabolisms are key metabolic 'hallmarks of cancer' (2, 34, 65). Our observation of an association with an increased risk of buccal and oral cancer corroborates previous findings in AMORIS (34).

Population heterogeneity and risk stratification: the need for data reduction techniques

The modulation effect of population heterogeneity on the association between potential risks factors and disease is a new avenue to understand the variability of risk in the population (66). For instance, in a targeted metabolomics exercise Shan et al. performed a principal component analysis and time to event analysis identifying metabolic profiles to predict risk of CVD (13). Another study used Monte Carlo Cross Validation and Lasso logistic regression to evaluate serum biomarkers as an alternative to fecal immunochemical testing to improve detection of colorectal cancer (11). In 2010, the European

Prospective Investigation on Cancer and Nutrition (EPIC) cohort reported that a specific prediagnostic plasma phospholipid fatty acid profile could predict the risk of gastric cancer (67). As rationalized in the HELIX project, these multiple profiling approaches aim to identify groups of individuals in the population that share a similar exposome that might account for differences on the specific risk of study (68).Together with these studies, our systematic data integration approach based on LCA demonstrates the potential of investigating population heterogeneity using metabolic profiling as risk factors for long term cancer risk and mortality prediction. However, in order to establish the prediction capability of these LCA metabolic profiles and implement their use in a clinical setting, further studies to validate the results whilst allowing to measure sensitivity and specificity, will need to be conducted such as a nested case-control in AMORIS that could determine the predictive capabilities of the metabolic profiles to estimate cancer risk and mortality.

Strengths and limitations

The present study has been conducted in a large and well-defined population, applying a multi-faced approach covering main biological pathways to assess biomarker profiles that could indicate cancer risk, cancer survival and mortality. The major strength of these analyses lies in the innovative avenue to study population heterogeneity and susceptibility to disease and mortality in a large cohort of participants with multiple measurements, all measured on fresh blood samples on the same day at the same clinical laboratory. We included all the markers available in the cohort for a large population (n>13000), however not every marker of the central metabolic pathways was available in the database (i.e. Complete Blood Count). Life-style factors established as cancer risk factors such as tobacco smoking, low physical activity, poor diet, alcohol intake, obesity and hypertension were partially available in AMORIS which limited their used in the study. To mitigate the lack of some of these external factors such as BMI, the analyses have been adjusted for Charlson Comorbidity Index which includes comorbidities such as obesity and hypertension. The lack of others life-style factors such as alcohol consumption was mitigated by using information on serum biomarkers such as gamma glutamyl transferase and other liver enzymes. All participants were selected by analyzing blood samples from

health check-ups in non-hospitalized individuals from the greater Stockholm area ensuring good internal validity in the study. Future studies will benefit from a longitudinal approach with repeated serum markers measurements that will capture the population phenotypic variations in relation to disease over long periods of time and will help to improve our understanding of the biomarkers' impact on carcinogenesis and mortality.

Conclusion

Our findings support the recently expressed need for a shift from the classical epidemiological approach of assessing one exposure to a systemic approach with multiple exposures. The LCA adapted in this study illustrates how a biomarker-wide approach can help assess population susceptibility to disease and provide insight into disease etiology in the context of carcinogenesis and mortality (Figure 3). Given the environmental and genetic modulation of metabolic molecules, metabolic profiling based on standard of care serum markers could become a useful non-invasive predictive signature for risk stratification and an important area of research for mechanisms and clinical relevance.

Methods

Study design and study population

The AMORIS study, a large prospective cohort study, has been described in detail elsewhere (19, 21, 22). Briefly, the AMORIS database is based on linkages with the Central Automation Laboratory (CALAB) database, which analyzed fresh blood samples from subjects from the greater Stockholm area. All individuals were either healthy individuals referred for clinical laboratory testing as part of a general health check-up or outpatients between 1985 and 1996. The AMORIS cohort has been linked to several Swedish national registries such as the National Cancer Register, the Patient Register, the Cause of Death Register, the consecutive Swedish Censuses during 1970-1990, and the National Register of Emigration, using the Swedish 10-digit personal identity number. These linkages provide detail information on demographics, lifestyle, socio-economic status, vital status, cancer diagnosis,

comorbidities and emigration. The AMORIS study conformed to the declaration of Helsinki and was approved by the ethics board of the Karolinska Institute.

From the AMORIS cohort, we included all individuals aged 20 years or older with measurements for the following serum biomarkers (n=13,615), which were all measured on the same day, using fully automated methods with automatic calibration performed on fresh blood samples, at the same laboratory (CALAB) of high quality according to international blinded testing (23) (Appendix Table S1 and S2): total cholesterol (TC) (mmol/L), triglycerides (TG) (mmol/L), apolipoprotein A-1 (ApoA-I) (g/L), apolipoprotein B (ApoB) (g/L), high density lipoprotein (HDL) (mmol/L), low density lipoprotein (LDL) (mmol/L), glucose (mmol/L), fructosamine (FAMN) (mmol/L), gamma-glutamyl transferase (GGT) (IU/L), alanine aminotransferase (ALT) (IU/L), aspartate aminotransferase (AST) (IU/L), albumin (g/L), leukocytes (WBC) (10^9 cells/L), C-reactive protein (CRP) (mg/L), iron (FE) (µmol/L), total iron binding capacity (TIBC) (mg/dL), creatinine (µmol/L), phosphate (mmol/L) and calcium (mmol/L). All methods have previously been described (22).

These biomarkers were selected to reflect common metabolic pathways: lipid (TC, TG, ApoA-I, ApoB, HDL and LDL) and glucose metabolism (Glucose, FAMN), liver function (GGT, ALT and AST), inflammation (Albumin, WBC and CRP), iron metabolism (FE and TIBC), kidney function (Creatinine) and phosphate (Phosphate and Calcium). The blood metabolites included in the analysis were all the standard serum markers available from routine health check-ups. Most of the markers included have been previously studied individually in AMORIS, however no systemic integrative approach to examine the metabolic markers interactions and susceptibility to cancer has been conducted to date (24-35). All participants were free from cancer at time of study entry and none were diagnosed with cancer within the first three years of follow-up to avoid reverse causation.

The main exposure variables for the analyses were the above-mentioned metabolic biomarkers, for which the values were categorized using standardized clinical cut-offs based on recognized medical criteria to facilitate interpretation of the results (Appendix Table S2). The main outcomes were first cancer diagnosis, as registered in the National Cancer Register using ICD-9 for the years 1987-1992,

ICD-O/2 for years 1993-2004 and for year 2005 onwards has been coded in ICD-O/3), and mortality. As secondary outcomes, we explored those cancer types for which there were more than 30 events during follow-up. Likewise, cancer mortality was explored. Follow-up time was assessed specifically for each of the outcomes studied. For cancer diagnosis, follow-up time was defined as time from blood drawn until date of first cancer diagnosis, death, emigration or study closing date (31st of December 2012), whichever occurred first. The follow-up time for death was described as time from blood drawn until date of death, emigration or study closing date (31st of December 2012), whichever occurred first.

Information on the following potential confounders was also incorporated: age, sex and comorbidities. The latter was quantified using the Charlson Comorbidity Index (CCI) calculated based on data from the National Patient Register. The CCI comprises 17 disease categories, all assigned a weight. The sum of an individual's weights was used to create the CCI ranging from no comorbidity to severe comorbidity (0, 1, 2, and ≥3) (36).

Data Analysis

First, we calculated **Pearson correlation coefficients** to measure the strength of association between the biomarkers included in the analysis. Pearson's correlation analyses showed strong correlation between the different biomarkers in the lipid metabolism (TC, LDL and ApoB ($r > 0.7$); HDL and ApoA-I ($r > 0.8$)). We replaced the individual lipid biomarkers by the established ApoB/ApoA-I ratio and log (TG/HDL) ratio (20, 23, 37, 38) to avoid collinearity and to comply with the principle of local independence as required by latent class analysis (39). Most of the markers were normally distributed except from the liver biomarkers.

Latent Class Analysis (LCA) (39, 40) is a model-based clustering method that reduces the dimension of the data by clustering covariates into latent classes, using a probabilistic model that describes the data distribution, and it assesses the probability that individuals belong to certain latent classes. LCA avoids the use of a linear combination or a random distance definition to reduce the number of covariates

(41) and has recently been employed in health sciences (42, 43). More specifically, we applied LCA to characterize different classes of individuals based on their metabolic profiles (44) and to evaluate intrinsic associations between the biomarkers, using the poLCA package (45) in R statistical programming language. We first determined the optimal number of LCA-derived classes by executing step-wise models with different numbers of classes, starting with the null model and adding one extra class in each model until reaching the total number of biomarkers in the data, while the model kept converging into a local maximum likelihood. The criterions used for model selection (Akaike information criterion (AIC), Bayesian information criterion (BIC) and Chi-squared distribution) were evaluated to estimate the best goodness of fit model and to define the optimal number of LCA-derived metabolic classes that characterized our dataset. To identify which sets of biomarkers predominantly explained each latent class, how the classes were distributed across the study population and which individuals were allocated to each class, we assessed the conditional probabilities, mixed proportions and class memberships of the best fitted latent class model.

Once each subject was assigned to its LCA-derived metabolic class, we conducted **multivariable Cox proportional hazard regression** to examine whether the LCA-derived metabolic classes were associated with long term risk of overall cancer as well as specific cancer types. In addition, we evaluated how the classes were associated with all cause-death and cancer-specific death. All models were adjusted for age, sex, and CCI. We performed a sensitivity analysis using age as a time-scale, as this is potentially a strong confounder. Moreover, Schoenfeld residuals were tested to ensure the proportional hazard assumption of the multivariable cox proportional hazard regression analysis.

Data management and statistical analyses were performed using Statistical Analysis Systems (SAS) release 4.3 (SAS Institute, Cary, NC) and R version 3.0.2 (R Foundation for Statistical Computing, Vienna, Austria).

Abbreviations

AIC - Akaike information criterion

ALT - Alanine Aminotransferase

AMORIS - Apolipoprotein MOrtality RISk Study

ApoA-I - Apolipoprotein A-1

ApoB - Apolipoprotein B

AST - Aspartate Aminotransferase

BIC - Bayesian information criterion

CALAB - Central Automation Laboratory

CCI - Charlson Comorbidity Index

CRP - C-reactive protein

FAMN – Fructosamine

FE - Iron

GGT - Gamma-glutamyl Transferase

HDL - High Density Lipoprotein

ICD-9 - International Classification of Diseases 9th Revision

ICD-O/2 International Classification of Diseases for Oncology 2nd Revision

ICD-O/3 International Classification of Diseases for Oncology 3rd Revision

LCA - Latent Class analysis

LDL - Low Density Lipoprotein

SAS - Statistical Analysis Systems

TC - Total Cholesterol

TIBC - Total Iron Binding Capacity

TG - Triglycerides

WBC - Leukocytes

WHO - World Health Organization

Declarations

Ethical approval and consent to participate

The study was approved by the ethics board of Karolinska Institutet who waived the need for consent and conformed to the declaration of Helsinki.

Availability of data and materials

The authors can confirm that for ethical and legal reasons imposed there are restrictions to the allowance of general public access to the data underlying the findings of this study. The database is formed of not only the AMORIS cohort but is a merged database. This includes AMORIS plus information from the Swedish National Patient Registry, the National Cause of Death Registry, SWEDEHEART, the Work Lipids, Fibrinogen study, the Cohort of Swedish Men Study, the Swedish Mammography Cohort, the cohort of 60-year-old subjects in Stockholm, the Sollentuna Primary Prevention study and the National Prescribed Drug Register.

The merged database from these sources contain sensitive information and is therefore anonymized and located in a security server with restricted access at the institute of Environmental medicine, Karolinska Institutet in Stockholm.

Professor Maria Feychting (maria.feychting@ki.se) and Sofia Carlsson (sofia.carlsson@ki.se) are both members of the Steering Committee of the AMORIS cohort and are based on the Unit of Epidemiology, Institute of Environmental Medicine hosting the database. They would both be able to respond to external requests for data access given that the interested party can obtain approval from the data owners including the National Board of Health and Welfare in Sweden (http://www.socialstyrelsen.se/english) and Statistics Sweden (http://www.scb.se/en_/) as well as from the owners of the research registers at Karolinska Institutet, Stockholm. Sweden.

To ensure persistent and long-term database storage and availability, AMORIS cohort database is stored at the Institute of Environmental Medicine and the storage follows the principles kept at Karolinska Institutet. The database can be accessed after permission and considering the restrictions by remote access through a secure LAN solution.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by King's College London (Salaries for AS, HG, AG, MVH), Karolinska Institutet (IJ, GW, ML LH), Cancer Research UK grant (C45074/A26553) (PI:MVH) and the Gunner and Ingmar Jungner Foundation for Laboratory Medicine (https://ki.se/en/about/the-gunnar-ingmar-jungner-foundation-for-laboratory-medicine) who provide donations to fund the AMORIS database.

Author contribution

AS designed the study, analysed the data and wrote the primary manuscript. AG was responsible in designing and conceptualising study, reviewing manuscript and supervising the project. MVH conceptualised and designed the study, over saw the study and was a major contributor in writing the manuscript. SG contributed to revising and reviewing manuscript and was responsible for submission. HG provided analysis, statistical input and reviewed the manuscript. ML, LH, IJ, GW, NH provided data acquisition, quality control of data, study design, data interpretation as well as reviewing the manuscript. All authors read and approved the final manuscript.

Acknowledgement The authors are grateful to all sample and data donors who participated in the AMORIS study.

References

1. Global Burden of Disease Cancer C, Fitzmaurice C, Dicker D, Pain A, Hamavid H, Moradi-Lakeh M, et al. The Global Burden of Cancer 2013. JAMA Oncol. 2015 Jul;1(4):505-27. PubMed PMID: 26181261. Pubmed Central PMCID: 4500822.

2. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000 Jan 7;100(1):57-70. PubMed PMID: 10647931.

3. Global Burden of Disease Cancer C, Fitzmaurice CA, C., Barber RM, Barregard L, Bhutta ZA, Brenner H, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015: A Systematic Analysis for the Global Burden of Disease Study. JAMA Oncol. 2017 Apr 01;3(4):524-48. PubMed PMID: 27918777.

4. Blair RH, Trichler DL, Gaille DP. Mathematical and statistical modeling in cancer systems biology. Front Physiol. 2012 06/28

03/30/received

06/05/accepted;3:227. PubMed PMID: 22754537. Pubmed Central PMCID: 3385354.

5. Dupont WD, Blume JD, Smith JR. BUilding and validating complex models of breast cancer risk. JAMA Oncology. 2016.

6. Poste G. Bring on the biomarkers. Nature. 2011 Jan 13;469(7329):156-7. PubMed PMID: 21228852. Epub 2011/01/14. eng.

7. Zhang Y. News & Views: Bring on the Biomarkers—It's Time for the “Big Science” Approach. Clin Chem. 2011 June 1, 2011;57(6):928-9.

8. Brooks JD. Translational genomics: the challenge of developing cancer biomarkers. Genome Res. 2012 Feb;22(2):183-7. PubMed PMID: 22301132. Pubmed Central PMCID: 3266026.

9. Beltran H, Rubin MA. New strategies in prostate cancer: translating genomics into the clinic. Clin Cancer Res. 2013 Feb 1;19(3):517-23. PubMed PMID: 23248095. Epub 2012/12/19. eng.

10. Bünger S, Haug U, Kelly M, Posorski N, Klempt-Giessing K, Cartwright A, et al. A novel multiplex-protein array for serum diagnostics of colon cancer: a case–control study. BMC Cancer. 2012 09/07

05/11/received

518 08/31/accepted;12:393-. PubMed PMID: PMC3502594.
519 11. Wild N, Andres H, Rollinger W, Krause F, Dilba P, Tacke M, et al. A combination of serum
520 markers for the early detection of colorectal cancer. Clin Cancer Res. 2010 Dec 15;16(24):6111-21.
521 PubMed PMID: 20798228. Epub 2010/08/28. eng.
522 12. Noto D, Cefalu AB, Barbagallo CM, Ganci A, Cavera G, Fayer F, et al. Baseline metabolic
523 disturbances and the twenty-five years risk of incident cancer in a Mediterranean population. Nutrition,
524 metabolism, and cardiovascular diseases : NMCD. 2016 Jul 12. PubMed PMID: 27511705. Epub
525 2016/08/12. Eng.
526 13. Shah SH, Sun JL, Stevens RD, Bain JR, Muehlbauer MJ, Pieper KS, et al. Baseline
527 metabolomic profiles predict cardiovascular events in patients at risk for coronary artery disease. Am
528 Heart J. 2012 May;163(5):844-50 e1. PubMed PMID: 22607863. Epub 2012/05/23. Eng.
529 14. Wild CP. The exposome: from concept to utility. Int J Epidemiol. 2012 Feb;41(1):24-32.
530 PubMed PMID: 22296988. Epub 2012/02/03. Eng.
531 15. Wild CP, Scalbert A, Herceg Z. Measuring the exposome: a powerful basis for evaluating
532 environmental exposures and cancer risk. Environ Mol Mutagen. 2013 Aug;54(7):480-99. PubMed
533 PMID: 23681765. Epub 2013/05/18. eng.
534 16. Nicholson G, Rantalainen M, Maher AD, Li JV, Malmodin D, Ahmadi KR, et al. Human
535 metabolic profiles are stably controlled by genetic and environmental variation. Mol Syst Biol. 2011
536 Aug 30;7:525. PubMed PMID: 21878913. Pubmed Central PMCID: PMC3202796. Epub 2011/09/01.
537 Eng.
538 17. Cui Y, Balshaw DM, Kwok RK, Thompson CL, Collman GW, Birnbaum LS. The Exposome:
539 Embracing the Complexity for Discovery in Environmental Health. Environ Health Perspect. 2016 Aug
540 01;124(8):A137-40. PubMed PMID: 27479988. Pubmed Central PMCID: PMC4977033. Epub
541 2016/08/02. eng.
542 18. Czene K, Lichtenstein P, Hemminki K. Environmental and heritable causes of cancer among
543 9.6 million individuals in the Swedish Family-Cancer Database. Int J Cancer. 2002 May 10;99(2):260-
544 6. PubMed PMID: 11979442. Epub 2002/04/30. Eng.
545 19. Walldius G, Jungner I, Kolar W, Holme I, Steiner E. High cholesterol and triglyceride values
546 in Swedish males and females: increased risk of fatal myocardial infarction. First report from the
547 AMORIS (Apolipoprotein related MOrtality RISk) study. Blood Press Suppl. 1992;4:35-42. PubMed
548 PMID: 1345333. Epub 1992/01/01. eng.
549 20. Walldius G, Jungner I, Holme I, Aastveit AH, Kolar W, Steiner E. High apolipoprotein B, low
550 apolipoprotein A-I, and improvement in the prediction of fatal myocardial infarction (AMORIS study):
551 a prospective study. Lancet. 2001 Dec 15;358(9298):2026-33. PubMed PMID: 11755609. Epub
552 2002/01/05. eng.
553 21. Van Hemelrijck M, Harari D, Garmo H, Hammar N, Walldius G, Lambe M, et al. Biomarker-
554 based score to predict mortality in persons aged 50 years and older: a new approach in the Swedish

555 AMORIS study. Int J Mol Epidemiol Genet. 2012;3(1):66-76. PubMed PMID: 22493753. Pubmed
556 Central PMCID: 3316450. Epub 2012/04/12. eng.
557 22. Walldius G, Malmstrom H, Jungner I, de Faire U, Lambe M, Van Hemelrijck M, et al. The
558 AMORIS cohort. Int J Epidemiol. 2017 Feb 02. PubMed PMID: 28158674. Epub 2017/02/06. eng.
559 23. Jungner I, Marcovina SM, Walldius G, Holme I, Kolar W, Steiner E. Apolipoprotein B and A-
560 I values in 147576 Swedish males and females, standardized according to the World Health
561 Organization-International Federation of Clinical Chemistry First International Reference Materials.
562 Clin Chem. 1998 Aug;44(8 Pt 1):1641-9. PubMed PMID: 9702950. Epub 1998/08/14. eng.
563 24. Van Hemelrijck M, Jassem W, Walldius G, Fentiman IS, Hammar N, Lambe M, et al. Gamma-
564 glutamyltransferase and risk of cancer in a cohort of 545,460 persons - the Swedish AMORIS study.
565 Eur J Cancer. 2011 Sep;47(13):2033-41. PubMed PMID: 21486691. Epub 2011/04/14. eng.
566 25. Van Hemelrijck M, Walldius G, Jungner I, Hammar N, Garmo H, Binda E, et al. Low levels of
567 apolipoprotein A-I and HDL are associated with risk of prostate cancer in the Swedish AMORIS study.
568 Cancer Causes Control. 2011 Jul;22(7):1011-9. PubMed PMID: 21562751. Epub 2011/05/13. eng.
569 26. Van Hemelrijck M, Garmo H, Holmberg L, Walldius G, Jungner I, Hammar N, et al. Prostate
570 cancer risk in the Swedish AMORIS study: the interplay among triglycerides, total cholesterol, and
571 glucose. Cancer. 2011 May 15;117(10):2086-95. PubMed PMID: 21523720. Epub 2011/04/28. eng.
572 27. Van Hemelrijck M, Holmberg L, Garmo H, Hammar N, Walldius G, Binda E, et al. Association
573 between levels of C-reactive protein and leukocytes and cancer: three repeated measurements in the
574 Swedish AMORIS study. Cancer Epidemiol Biomarkers Prev. 2011 Mar;20(3):428-37. PubMed PMID:
575 21297038. Pubmed Central PMCID: PMC3078551. Epub 2011/02/08. eng.
576 28. Melvin JC, Seth D, Holmberg L, Garmo H, Hammar N, Jungner I, et al. Lipid profiles and risk
577 of breast and ovarian cancer in the Swedish AMORIS study. Cancer Epidemiol Biomarkers Prev. 2012
578 Aug;21(8):1381-4. PubMed PMID: 22593241. Epub 2012/05/18. eng.
579 29. Van Hemelrijck M, Garmo H, Hammar N, Jungner I, Walldius G, Lambe M, et al. The interplay
580 between lipid profiles, glucose, BMI and risk of kidney cancer in the Swedish AMORIS study. Int J
581 Cancer. 2012 May 1;130(9):2118-28. PubMed PMID: 21630265. Epub 2011/06/02. eng.
582 30. Wulaningsih W, Garmo H, Holmberg L, Hammar N, Jungner I, Walldius G, et al. Serum Lipids
583 and the Risk of Gastrointestinal Malignancies in the Swedish AMORIS Study. Journal of cancer
584 epidemiology. 2012;2012:792034. PubMed PMID: 22969802. Pubmed Central PMCID: 3437288.
585 31. Van Hemelrijck M, Hermans R, Michaelsson K, Melvin J, Garmo H, Hammar N, et al. Serum
586 calcium and incident and fatal prostate cancer in the Swedish AMORIS study. Cancer Causes Control.
587 2012 Aug;23(8):1349-58. PubMed PMID: 22710746. Epub 2012/06/20. eng.
588 32. Wulaningsih W, Michaelsson K, Garmo H, Hammar N, Jungner I, Walldius G, et al. Inorganic
589 phosphate and the risk of cancer in the Swedish AMORIS study. BMC Cancer. 2013;13:257. PubMed
590 PMID: 23706176. Pubmed Central PMCID: PMC3664604. Epub 2013/05/28. eng.

591 33. Wulaningsih W, Michaelsson K, Garmo H, Hammar N, Jungner I, Walldius G, et al. Serum
592 calcium and risk of gastrointestinal cancer in the Swedish AMORIS study. BMC Public Health. 2013
593 Jul 17;13(1):663. PubMed PMID: 23866097. Pubmed Central PMCID: 3729677. Epub 2013/07/20.
594 Eng.
595 34. Gaur A, Collins H, Wulaningsih W, Holmberg L, Garmo H, Hammar N, et al. Iron metabolism
596 and risk of cancer in the Swedish AMORIS study. Cancer Causes Control. 2013 Jul;24(7):1393-402.
597 PubMed PMID: 23649231. Pubmed Central PMCID: PMC3675271. Epub 2013/05/08. eng.
598 35. Wulaningsih W, Holmberg L, Garmo H, Zethelius B, Wigertz A, Carroll P, et al. Serum glucose
599 and fructosamine in relation to risk of cancer. PLoS ONE. 2013;8(1):e54944. PubMed PMID:
600 23372798. Pubmed Central PMCID: PMC3556075. Epub 2013/02/02. eng.
601 36. Quan H, Li B, Couris CM, Fushimi K, Graham P, Hider P, et al. Updating and validating the
602 Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data
603 from 6 countries. Am J Epidemiol. 2011 Mar 15;173(6):676-82. PubMed PMID: 21330339.
604 37. Kuyl JM, Mendelsohn D. Observed relationship between ratios HDL-cholesterol/total
605 cholesterol and apolipoprotein A1/apolipoprotein B. Clin Biochem. 1992 Oct;25(5):313-6. PubMed
606 PMID: 1490290. Epub 1992/10/01. eng.
607 38. Dobiášová M, Frohlich J. The plasma parameter log (TG/HDL-C) as an atherogenic index:
608 correlation with lipoprotein particle size and esterification rate inapob-lipoprotein-depleted plasma
609 (FERHDL). Clin Biochem. 2001 10//;34(7):583-8.
610 39. Vermunt JK, Magidson J. Latent class cluster analysis. Applied latent class analysis. 2002:89-
611 106.
612 40. Wood PK. J. A. Hagenaars and A. L. McCutcheon, Applied Latent Class Analysis, Kluwer,
613 Dordrecht, 2002, pp. 476. Journal of Classification. 2008 2008/06/01;25(1):143-5. English.
614 41. Chadeau-Hyam M, Campanella G, Jombart T, Bottolo L, Portengen L, Vineis P, et al.
615 Deciphering the complex: Methodological overview of statistical models to derive OMICS-based
616 biomarkers. Environ Mol Mutag. 2013;54(7):542-57.
617 42. Kongsted A, Nielsen AM. Latent Class Analysis in health research. Journal of physiotherapy.
618 2016 Aug 12. PubMed PMID: 27914733. Epub 2016/12/05. eng.
619 43. Lacey RJ, Strauss VY, Rathod T, Belcher J, Croft PR, Natvig B, et al. Clustering of pain and
620 its associations with health in people aged 50 years and older: cross-sectional results from the North
621 Staffordshire Osteoarthritis Project. BMJ open. 2015 November 1, 2015;5(11).
622 44. Haughton D, Legrand P, Woolford S. Review of three latent class cluster analysis packages:
623 Latent Gold, poLCA, and MCLUST. The American Statistician. 2009;63(1).
624 45. Linzer DA, Lewis JB. poLCA: An R package for polytomous variable latent class analysis.
625 46. Tolonen H, Keil U, Ferrario M, Evans A, Project WM. Prevalence, awareness and treatment of
626 hypercholesterolaemia in 32 populations: results from the WHO MONICA Project. Int J Epidemiol.
627 2005 Feb;34(1):181-92. PubMed PMID: 15333620.

47. Reiner Ž, Catapano AL, De Backer G, Graham I, Taskinen M-R, Wiklund O, et al. ESC/EAS Guidelines for the management of dyslipidaemias. The Task Force for the management of dyslipidaemias of the European Society of Cardiology (ESC) and the European Atherosclerosis Society (EAS). 2011 2011-07-01 00:00:00;32(14):1769-818.

48. Ioannou GN, Boyko EJ, Lee SP. The prevalence and predictors of elevated serum aminotransferase activity in the United States in 1999-2002. Am J Gastroenterol. 2006 Jan;101(1):76-82. PubMed PMID: 16405537. Epub 2006/01/13. eng.

49. Mason JE, Starke RD, Van Kirk JE. Gamma-glutamyl transferase: a novel cardiovascular risk biomarker. Preventive cardiology. 2010 Winter;13(1):36-41. PubMed PMID: 20021625. Epub 2009/12/22. eng.

50. Teppala S, Shankar A, Li J, Wong TY, Ducatman A. Association between serum gamma-glutamyltransferase and chronic kidney disease among US adults. Kidney Blood Press Res. 2010;33(1):1-6. PubMed PMID: 20090360. Epub 2010/01/22. eng.

51. Lim JS, Yang JH, Chun BY, Kam S, Jacobs DR, Jr., Lee DH. Is serum gamma-glutamyltransferase inversely associated with serum antioxidants as a marker of oxidative stress? Free Radic Biol Med. 2004 Oct 01;37(7):1018-23. PubMed PMID: 15336318. Epub 2004/09/01. eng.

52. Wessling-Resnick M. Iron Homeostasis and the Inflammatory Response. Annu Rev Nutr. 2010;30:105-22. PubMed PMID: PMC3108097.

53. Rappaport SM, Barupal DK, Wishart D, Vineis P, Scalbert A. The blood exposome and its role in discovering causes of disease. Environ Health Perspect. 2014 Aug;122(8):769-74. PubMed PMID: 24659601. Pubmed Central PMCID: 4123034.

54. Strasak AM, Rapp K, Brant LJ, Hilbe W, Gregory M, Oberaigner W, et al. Association of gamma-glutamyltransferase and risk of cancer incidence in men: a prospective study. Cancer Res. 2008 May 15;68(10):3970-7. PubMed PMID: 18483283. Epub 2008/05/17. eng.

55. Strasak AM, Pfeiffer RM, Klenk J, Hilbe W, Oberaigner W, Gregory M, et al. Prospective study of the association of gamma-glutamyltransferase with cancer incidence in women. Int J Cancer. 2008 Oct 15;123(8):1902-6. PubMed PMID: 18688855. Epub 2008/08/09. eng.

56. Ruhl CE, Everhart JE. Elevated serum alanine aminotransferase and gamma-glutamyltransferase and mortality in the United States population. Gastroenterology. 2009 Feb;136(2):477-85 e11. PubMed PMID: 19100265. Epub 2008/12/23. eng.

57. Koehler EM, Sanna D, Hansen BE, van Rooij FJ, Heeringa J, Hofman A, et al. Serum liver enzymes are associated with all-cause mortality in an elderly population. Liver international : official journal of the International Association for the Study of the Liver. 2014 Feb;34(2):296-304. PubMed PMID: 24219360. Epub 2013/11/14. eng.

58. Kunutsor SK, Apekey TA, Seddoh D, Walley J. Liver enzymes and risk of all-cause mortality in general populations: a systematic review and meta-analysis. Int J Epidemiol. 2014 February 1, 2014;43(1):187-201.

59. Rose G, Shipley MJ. Plasma lipids and mortality: a source of error. Lancet. 1980 Mar 08;1(8167):523-6. PubMed PMID: 6102243. Epub 1980/03/08. eng.
60. Schupf N, Costa R, Luchsinger J, Tang MX, Lee JH, Mayeux R. Relationship between plasma lipids and all-cause mortality in nondemented elderly. J Am Geriatr Soc. 2005 Feb;53(2):219-26. PubMed PMID: 15673344. Epub 2005/01/28. eng.
61. Akerblom JL, Costa R, Luchsinger JA, Manly JJ, Tang M-X, Lee JH, et al. Relation of plasma lipids to all-cause mortality in Caucasian, African-American and Hispanic elders. Age Ageing. 2008;37(2):207-13. PubMed PMID: PMC2715146.
62. Neaton JD, Blackburn H, Jacobs D, et al. Serum cholesterol level and mortality findings for men screened in the multiple risk factor intervention trial. Arch Intern Med. 1992;152(7):1490-500.
63. Kagan A, McGee DL, Yano K, Rhoads GG, Nomura A. Serum cholesterol and mortality in a Japanese-American population: the Honolulu Heart program. Am J Epidemiol. 1981 Jul;114(1):11-20. PubMed PMID: 7246518. Epub 1981/07/01. eng.
64. Radišauskas R, Kuzmickienė I, Milinavičienė E, Everatt R. Hypertension, serum lipids and cancer risk: A review of epidemiological evidence. Medicina (Mex). 2016 //;52(2):89-98.
65. Beguin Y, Aapro M, Ludwig H, Mizzen L, Osterborg A. Epidemiological and nonclinical studies investigating effects of iron in carcinogenesis--a critical review. Crit Rev Oncol Hematol. 2014 Jan;89(1):1-15. PubMed PMID: 24275533.
66. Wang M, Spiegelman D, Kuchiba A, Lochhead P, Kim S, Chan AT, et al. Statistical methods for studying disease subtype heterogeneity. Stat Med. 2016 Feb 28;35(5):782-800. PubMed PMID: 26619806. Pubmed Central PMCID: 4728021. Epub 2015/12/02. eng.
67. Chajès V, Jenab M, Romieu I, Ferrari P, Dahm CC, Overvad K, et al. Plasma phospholipid fatty acid concentrations and risk of gastric adenocarcinomas in the European Prospective Investigation into Cancer and Nutrition (EPIC-EURGAST). The American Journal of Clinical Nutrition. 2011 November 1, 2011;94(5):1304-13.
68. Vrijheid M, Slama R, Robinson O, Chatzi L, Coen M, van den Hazel P, et al. The human early-life exposome (HELIX): project rationale and design. Environ Health Perspect. 2014 Jun;122(6):535-44. PubMed PMID: 24610234. Pubmed Central PMCID: PMC4048258. Epub 2014/03/13. eng.

Figure 1A| Line-graph depicting the goodness of fit indicators AIC and BIC. The model that best fits the dataset comprehends of four latent classes as determined by the minimum value reached by AIC and BIC criterions before stabilization of the values. The criterion did not converge to a local maximum from class 12 onwards.

Figure 1B| Line-graph depicting the goodness of fit indicators (X^2(1) (Chi-square). The model that best fits the dataset comprehends of four latent classes as determined by the minimum value reached by Chi-square. The criterions did not converge to a local maximum from class 12 onwards.

696

Figure 2| Class Membership Probabilities for abnormal clinical values of the serum markers for the four LCA – derived metabolic classes. The four different biomarker profiles are represented in the graph.

Figure 3| Study statistical pipeline describing the methodology followed in the project. We explored the blood exposome using metabolic markers of the population to assess how population heterogeneity is associated with cancer risk and mortality.

Additional Files

Table S1.docx| Laboratory fully automated methods with automatic calibration were performed at one accredited laboratory (CALAB to measure the serum biomarkers examine in the study.

Table S2.docx| Panel of serum markers describing standard medical cut-offs information.

Table S3.docx| Characteristics of the study population by LCA-derived metabolic classes.

725
726

727

728
729
730
731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746