Subgrouping multimorbid patients with ischemic heart disease by means of unsupervised clustering: A cohort study of 72,249 patients defined by 3,046 diagnoses

Short title: Unsupervised clustering of patients with ischemic heart disease

Amalie D. Haue1,2,¶, Peter C. Holm1,5, Karina Banasik1, Agnete T. Lundgaard1, Victorine P. Muse3, Timo Röder1, David Westergaard1, Piotr J. Chmura1, Alex H. Christensen2,3, Peter EK. Weeke2, Erik Sørensen4, Ole BV. Pedersen4,5, Sisse R. Ostrowski3,6, Kasper K. Iversen3, Lars V. Køber2,6, Henrik Ullum7, Henning Bundgaard2,5,*, Søren Brunak, PhD1,8*

1Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
2Department of Cardiology, The Heart Center, Rigshospitalet, Copenhagen, Denmark
3Department of Cardiology, Copenhagen University Hospital, Herlev, Denmark
4Department of Clinical Immunology, Copenhagen University Hospital, Copenhagen, Denmark
5Department of Clinical Immunology, Zealand University Hospital, Køge, Denmark
6Department of Clinical Medicine, University of Copenhagen, Rigshospitalet, Copenhagen, Denmark
7Statens Serum Institut, Copenhagen, Denmark
8Copenhagen University Hospital, Rigshospitalet, Copenhagen, Denmark

¶These authors contributed equally to this work.

E-mail: soren.brunak@cpr.ku.dk (SB)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Background: There are no methods for classifying multimorbid patients with ischemic heart disease (IHD), although such methods might be clinically useful due to the marked differences in presentation and disease-course.

Methods: A population-based cohort study from a Danish secondary care setting of patients with IHD (2004-2016) and subjected to a coronary angiography (CAG) or coronary computed tomography angiography (CCTA). Data sources were The Danish National Patient Registry, in-hospital laboratory data, and genetic data from Copenhagen Hospital Biobank. Comorbidities included diagnoses assigned prior to presentation of IHD. Patients were clustered my means of the Markov Clustering Algorithm based on the entire spectrum of registered multimorbidity. The two prespecified outcomes were: New ischemic events (including death from IHD causes) and death from non-IHD causes. Patients were followed from date of CAG/CCTA until one of the two outcomes occurred or end of follow-up, whichever came first. Biological and clinical appropriateness of clusters was assessed by comparing risks (estimated from Cox proportional hazard models) in clusters and by phenotypic and genotypic enrichment analyses, respectively.

Findings: In a cohort of 72,249 patients with IHD (mean age 63.9 years, 63.1% males), 31 distinct clusters (C1-31, 67,136 patients) were identified. Comparing each cluster to the others, eight clusters (9,590 patients) had statistically significantly higher (five clusters) or lower (three clusters) risk of new ischemic events; 18 clusters (35,982 patients) had a higher (11 clusters) or lower (seven clusters) risk of death from non-IHD causes. All clusters at increased risk of new ischemic events, associated with risk of death from non-IHD causes as well. Cardiovascular or inflammatory diseases were the commonly enriched in clusters (13), and distributions for 24 laboratory test results differed significantly across clusters. Polygenic risk scores for atrial fibrillation and diabetes were increased in x and y clusters respectively.
Conclusions: Clustering of patients with IHD based on comorbidities identified subgroups of patients with significantly different clinical outcomes. This novel approach may support differentiation of treatment intensity dependent on expected outcomes.
Non-standard abbreviations

CAG: Coronary arteriography

CCTA: Coronary computed tomography angiography

ICD-10: International Statistical Classification of Diseases and Related Health Problems 10th Revision

IHD: Ischemic heart disease

MCL: Markov clustering

NPR: Danish National Patient Registry (NPR)

O/E-ratio: Observed-expected-ratio

PRS: Polygenic risk score
Introduction

Ischemic heart disease (IHD) is a common, chronic, complex disease and mode of onset, disease burden and disease progression vary considerably between patients (1–3). This heterogeneity relates to several factors, but a major contribution is multimorbidity as more than 85% of IHD patients are diagnosed with other chronic diseases; a phenomenon coined cardiometabolic multimorbidity (4,5). The increased mortality in patients with cardiometabolic multimorbidity is generally only related to single disease states, such as obstructive lung disease, diabetes, or stroke, although it is known that the risk of cardiovascular diseases is increased in many chronic, inflammatory disorders (6,7). As more patients at older age and with more and more co-morbidities are seen, new methods for characterizing and studying cardiometabolic multimorbidity are needed (8–11).

Unsupervised clustering algorithms can systematically reveal structures in large, feature-rich datasets and may be used to identify distinct patient subgroups within a heterogeneous population (12). Proof-of-concept analyses of cardiovascular phenotypes, including IHD, heart failure, diabetes, and atrial fibrillation have already been performed (13–19). While these studies successfully identify subgroups resembling those from traditional analyses, they often fail to demonstrate that clustering analysis leads to novel understanding of a given dataset. Rather, they are typically restricted to characterize high-, medium-, and low-risk subgroups which by and large resemble more conservative approaches from an earlier, less data-rich, epoch (20).

For decades, Danish healthcare registries have had a strong position within epidemiological research (20–22). Given the opportunities for using clinical data more extensively, we carried out an unsupervised clustering analysis of 72,249 patients with IHD based on their entire
disease history until IHD onset. Explicitly, we wanted to classify IHD based on the entire spectrum of multimorbidity. We identified distinct patient subgroups derived from a total of 3,046 diagnoses assigned prior to IHD onset. The biological and clinical appropriateness of the patient subgroups identified by unsupervised clustering analysis was asserted by assessments of their associations with clinical outcomes and clinical characteristics, laboratory data, and genetics (Figure 1).
Fig 1: **Graphical overview of study.** Conceptual figure displaying the cohort of patients with IHD. A.: Assemblage of patient-specific vectors and an $f \times p$ matrix, where $f$ corresponds to the number of diagnoses and $p$ corresponds to the number of included patients. B: Unsupervised clustering of IHD patients using the MCL algorithm. That were the basis for

C. Association between clusters and clinical outcomes

D. Phenotypic and genetic characterization of clusters
construction of a matrix. C: which was the basis for performing unsupervised clustering to identify distinct clusters. D: Cluster-based risk stratification of cohort. Red: Increased risk of both outcomes. Orange: Increased risk of death from non-IHD causes, only. Yellow: No statistically significant associations. Light green: Reduced risk of death from non-IHD causes. Dark green: Reduced risk of new ischemic events or non-IHD causes. E: Phenotypic and genetic characterization of clusters. IHD: Ischemic heart disease. MCL: Markov Clustering.
Methods

Data sources, study population, and outcomes

Data from the Danish National Patient Registry (NPR) and the Danish Registry for Causes of Death were linked to in-hospital electronic health data covering the two Danish healthcare regions in Eastern Denmark (~2.9 mil inhabitants), and the Copenhagen Hospital Biobank Cardiovascular Disease Cohort(21,23,24). Linkage of different healthcare data sources was obtained via the personal identification number and only patients admitted to a hospital in Eastern Denmark in years 2004 to 2016 were considered(25). We identified all patients in NPR who were assigned an ICD-10 code for IHD(26). To increase the positive predictive value of IHD diagnoses and align included patients in time, we further required that patients had been subjected to coronary arteriography (CAG) or coronary computed tomography angiography (CCTA). To qualify that CAG/CCTAs were conclusive for IHD, patients were only included if the CAG/CCTA was performed during a contact where patients were assigned an ICD-10 code for IHD. We set the earliest CAG/CCTA fulfilling this criterium as the index date and excluded patients with an index date before year 2004 or after 2016 (Fig 2).
There were two predefined outcomes: 1) New ischemic events and 2) Death from other causes than IHD (non-IHD causes). The outcome “new ischemic event” was a composite outcome of a) hospitalization minimum 30 days after index for myocardial infarction or unstable angina pectoris (i.e., hospitalization with myocardial infarction or unstable angina pectoris as the primary diagnosis), b) revascularization not related to the index date and c) any death where IHD was listed as the primary or secondary cause. Outcomes were obtained from NPR and Danish Registry for Causes of Death. Eligible codes for inclusion, outcomes and specific cutoffs are available in S1 Fig and S1 Table.

Data preprocessing and application of the Markov cluster algorithm

We performed a clustering analysis of included patients based on their multimorbidity prior to their IHD diagnosis (index) using the Markov cluster (MCL) algorithm (27). Multimorbidity was represented as patient-specific vectors using diagnoses assigned prior to or at index. ICD-10 codes assigned to less than five patients (n=1,673) were excluded from the analysis owing to legal regulations in Denmark. As we focused the studies on multimorbidity in IHD, ICD-10 codes for IHD (I20-I25) were excluded from patients-specific vectors. Thus, a total of 3,046 ICD-10 codes were the basis for constructing a patient similarity network that was used as MCL algorithm input. Patient-specific vectors consisted of length 3,046 with integers indicating the number of times a patient had been assigned a particular ICD-10 code. The length of the vectors corresponded to the number of input features (ICD-10 codes). By combining the patient-specific vectors from all included patients, a matrix of size p x f was constructed, where p indicates the number of included patients and f indicates the number of input features (ICD-10 codes). Following a series of preprocessing steps, a patient similarity network was created based on the p x f matrix and used as input for the MCL algorithm (28). Resulting clusters were denoted C followed by an
integer indicating the rank of the clusters with respect to cluster size (number of patients).

Thus, $C_1$ denotes the largest cluster and cluster-membership was used to denote a cluster as a covariate in subsequent analyses. Robustness of clustering was assessed by generating a series of diluted and shuffled versions of the resulting clusters (reference clustering), and their similarity was quantified using the variance of information measure as previously described(29). Explicitly, a series of diluted and shuffled versions of the input graph were generated(30). In total, 500 variations of the input graph were constructed by shuffling and deleting edges, respectively. The variation in the graphs was then quantified by means of variation of information parameter. Details regarding the MCL settings and a description of cluster robustness assessment are available in the S1 Appendix.

Preprocessing of laboratory and genetic data

Clusters were characterized by laboratory and genetic data based on the subset of patients where these data types were available. A panel of 25 different lab parameters was included in the analyses. Only tests taken up to 90 days before index or at the day of index were included. Included lab tests were plasma levels of potassium, sodium, hemoglobin, estimated glomerular filtration rate (eGFR), creatinine, carbamide, glucose, troponin (I/T), HDL cholesterol, LDL cholesterol, total cholesterol, leukocytes, C-reactive protein, lymphocytes, monocytes, neutrophils, basophils, platelets, INR, alanine transaminase, albumin, alkaline phosphatase, bilirubin, and triglyceride. For every cluster, a score was computed based on the number of patients with a lab test below, within, or above the standard reference value, indicated by -1, 0 and 1, respectively. Score was defined as the mean of the summarized values per cluster.
Autosomal genotype data were obtained by identifying included patients who were also among the study participants in the Copenhagen Hospital Biobank – Cardiovascular Disease Cohort(24). For included patients with genetic data available, we calculated polygenic risk scores (PRSs) for 14 traits, obtained from nine GWAS meta-analyses (atrial fibrillation, BMI-adjusted non-insulin diabetes, chronic kidney disease, HDL cholesterol levels, heart failure, LDL cholesterol levels, stroke, total cholesterol levels, triglyceride levels)(31–36) and five GWAS (acute myocardial infarction, coronary artery disease, diastolic blood pressure, non-alcoholic fatty liver disease, systolic blood pressure)(37–40). PRSs were calculated using the “LDpred2-auto” algorithm, implemented in the R package “bigsnp” (version 1.11.6) with R version 4.0.0 and the workflow management system Snakemake(41–43). Each trait’s PRS distribution was scaled to a mean of zero and a standard deviation of one.

Statistical analyses of clusters identified by the MCL algorithm

As the study was designed to identify patient subgroups and not individual variation, clusters of size < 500 were excluded from the remaining analyses. Mean age at IHD onset in each cluster was compared to the mean age at onset in all the other clusters using Tukey’s Honest Significant Difference (HSD) method. Significance level was set to 0.05 and P-values were adjusted assuming 465 tests (adj. P-val.).

To investigate the association between cluster-membership and the competing risks of new ischemic events and death from non-IHD causes, we used Cox proportional-hazards models (Cox models). Patients were followed from index until occurrence of either of the two outcomes, or end of follow-up (year 2018), whichever came first. The dependent variable was either risk of new ischemic events or death from non-IHD causes, and the independent variables were cluster, sex, and age at index. To age-adjust the models, analyses were
performed using restricted cubic spline with three knots for age at index. Follow-up time was truncated to a maximum of five years. For each cluster, hazard ratios (HRs) and 95% confidence intervals (CIs) were estimated by comparing HRs for the members of the cluster with the HRs with that of non-members.

Further characterization of clusters consisted of: (1) phenotypic enrichment analysis, (2) characterization of clusters with respect to their laboratory profiles and (3) a test for genetic enrichment. The phenotypic enrichment analysis was carried out based on ratios between observed (O) and expected (E) frequencies of diagnoses in the clusters (O/E-ratios). That is, ratios between the frequencies of ICD-10 codes in each cluster (observed frequencies) and the frequencies of ICD-10 codes in the entire population (expected frequencies) were calculated and expressed as O/E-ratios(44). In subsequent characterization of clusters, enrichment denoted O/E-ratios > 2, and clusters were characterized as having little enrichment if the sum of the ten largest O/E-ratios < 50. Inverse changes were used to denote O/E-ratios between 0 and 1.

Hierarchical clustering was applied to estimate the cluster similarity with respect to the laboratory tests using the Euclidean distance between the score of each cluster for each test.

For each of the fourteen traits we calculated PRSs for, we used Wilcoxon rank-sum tests to compare the PRS distribution of each cluster to the combined PRS distribution of PRSs in all other clusters. Resulting P-values were converted to the false discovery rate (FDR) to account for multiple testing, with a total of 434 tests. We report effect sizes as calculated by the “wilcox.test” function built into R version 4.0.0. Level of significance was set to FDR < 0.05, assuming 434 tests.
Further details regarding preprocessing and analyses of laboratory and genetic data are available in the S2 Appendix.

Results

Cohort demographics and co-morbidities

A total of 72,249 patients (63.1% males, mean age 63.9 years) were included (Table 1).

Angina pectoris (I20) was the most common IHD diagnosis (38,239 patients, 52.9%), followed by acute myocardial infarction (I21) (33,229 patients, 46.0%) and chronic IHD (I25) (22,750 patients, 31.5%). The most common co-morbidity prior to the IHD index was hypertension (I10.9) (24,818 patients, 34.4%) followed by dyslipidemia (E78.0) (12,780 patients, 17.7%) and non-insulin dependent diabetes (E11.9) (7,551 patients, 10.5%). Prior to index, the mean number of diagnoses per patient was 8.1. A total of 68,103 patients (94.3%) were co-morbid at index. The overall incidence (new ischemic events and death from non-IHD causes) was 94 events per 1000 person-years (Table 1).
Table 1: Patient demographics, comorbidities, and outcomes

<table>
<thead>
<tr>
<th>Cohort demographics</th>
<th>Total</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients (%)</td>
<td>72,249</td>
<td>45,576 (63.1)</td>
<td>26,673 (36.1)</td>
</tr>
<tr>
<td>Mean age at index (SD)</td>
<td>63.9 (11.9)</td>
<td>62.9 (11.6)</td>
<td>65.6 (12.1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>IHD manifestations (ICD-10)</th>
<th>Total</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angina pectoris (I20)</td>
<td>38,239</td>
<td>22,628</td>
<td>15,611</td>
</tr>
<tr>
<td>Acute myocardial infarction (I21)</td>
<td>33,299</td>
<td>27,720</td>
<td>10,579</td>
</tr>
<tr>
<td>Subsequent myocardial infarction (I22)</td>
<td>61</td>
<td>34</td>
<td>27</td>
</tr>
<tr>
<td>Certain current complications following acute myocardial infarction (I23)</td>
<td>138</td>
<td>92</td>
<td>46</td>
</tr>
<tr>
<td>Other acute ischemic heart diseases (I24)</td>
<td>1,341</td>
<td>814</td>
<td>527</td>
</tr>
<tr>
<td>Chronic ischemic heart disease (I25)</td>
<td>22,750</td>
<td>14,589</td>
<td>8,152</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Common comorbidities (ICD-10)</th>
<th>Total</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary (essential) hypertension (I10.9)</td>
<td>24,818</td>
<td>14,508</td>
<td>10,310</td>
</tr>
<tr>
<td>Hypercholesterolemia (E78.0)</td>
<td>12,780</td>
<td>7,842</td>
<td>4,938</td>
</tr>
<tr>
<td>Non-insulin dependent diabetes (E11.9)</td>
<td>7,551</td>
<td>4,891</td>
<td>2,660</td>
</tr>
<tr>
<td>Atrial fibrillation and atrial flutter, unspecified (I48.9)</td>
<td>7,075</td>
<td>4,509</td>
<td>2,566</td>
</tr>
<tr>
<td>Heart failure, unspecified (I50.9)</td>
<td>6,160</td>
<td>4,059</td>
<td>2,101</td>
</tr>
<tr>
<td>Chest pain, unspecified (R07.9)</td>
<td>5,863</td>
<td>3,441</td>
<td>2,422</td>
</tr>
<tr>
<td>Senile cataract, unspecified (H25.9)</td>
<td>5,764</td>
<td>2,795</td>
<td>2,969</td>
</tr>
<tr>
<td>Pneumonia, unspecified (J18.9)</td>
<td>5,469</td>
<td>3,236</td>
<td>2,260</td>
</tr>
<tr>
<td>Hyperlipidemia, unspecified (E78.5)</td>
<td>5,002</td>
<td>3,066</td>
<td>1,936</td>
</tr>
<tr>
<td>Chronic obstructive pulmonary disease (J44.9)</td>
<td>4,621</td>
<td>2,449</td>
<td>2,172</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcomes, number of cases</th>
<th>Total</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>New ischemic events (%)</td>
<td>14,679</td>
<td>10,152</td>
<td>4,527</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>5,833</td>
<td>3,709</td>
<td>2,124</td>
</tr>
<tr>
<td>Revascularization</td>
<td>6,282</td>
<td>4,718</td>
<td>2,124</td>
</tr>
<tr>
<td>Death caused by IHD</td>
<td>2,563</td>
<td>1,724</td>
<td>839</td>
</tr>
<tr>
<td>Death from non-IHD causes (%)</td>
<td>10,684</td>
<td>6,710</td>
<td>3,974</td>
</tr>
<tr>
<td>Censored (%)</td>
<td>46,886</td>
<td>28,713</td>
<td>18,172</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Outcomes, time to event</th>
<th>Mean time to event in years (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>Males</td>
</tr>
<tr>
<td>New ischemic events</td>
<td>1.48 (1.40)</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>2.40 (1.87)</td>
</tr>
<tr>
<td>Revascularization</td>
<td>2.25 (1.88)</td>
</tr>
<tr>
<td>Death caused by IHD</td>
<td>1.92 (1.13)</td>
</tr>
<tr>
<td>Death from non-IHD causes</td>
<td>2.16 (1.50)</td>
</tr>
<tr>
<td>Censored</td>
<td>4.37 (1.08)</td>
</tr>
<tr>
<td>Total</td>
<td>3.72 (1.64)</td>
</tr>
</tbody>
</table>
Unsupervised clustering of multimorbid patients with IHD

In the cohort, the MCL algorithm identified 36 distinct clusters based on the set of 3,046 ICD-10 codes assigned to the patients prior to or at index. The 36 clusters contained a total of 68,084 patients. The remaining 4,365 patients in the cohort (6.0%) that did not cluster were primarily patients with no diagnoses prior to index (>99%). This observation served as a negative control as the MCL algorithm correctly identified patients without registered co-morbidity prior to index. Further, clusters robustness was assessed as described in Methods with a reasonable variation of information. Next, the 31 of the 36 clusters with >500 patients (67,136 patients) were characterized (Table 2). Using Tukey’s HSD to compare the age at index between all 31 clusters (a total of 466 combinations), we found significant differences in 391 comparisons (S3 Table). For demographics of patients that did not cluster or were in clusters of size < 500, see S4 Table.
Table 2: Cluster demographics, characteristics, and associations with outcomes

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Size</th>
<th>Mean age at index in years (SD)</th>
<th>Males</th>
<th>Females</th>
<th>New ischemic events</th>
<th>Death from non-IHD causes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>HR</td>
<td>Adj. P-val.</td>
</tr>
<tr>
<td>C1</td>
<td>7,191</td>
<td>64.8 (11.3)</td>
<td>3,897</td>
<td>3,294</td>
<td>1.000</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C2</td>
<td>5,990</td>
<td>58.6 (11.5)</td>
<td>2,862</td>
<td>3,127</td>
<td>0.825</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>C3</td>
<td>4,641</td>
<td>56.8 (11.4)</td>
<td>2,727</td>
<td>1,914</td>
<td>0.757</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>C4</td>
<td>4,401</td>
<td>69.6 (10.2)</td>
<td>2,853</td>
<td>1,548</td>
<td>0.920</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C5</td>
<td>4,290</td>
<td>63.9 (10.7)</td>
<td>2,803</td>
<td>1,487</td>
<td>1.402</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>C6</td>
<td>3,589</td>
<td>59.7 (10.9)</td>
<td>2,388</td>
<td>1,201</td>
<td>0.969</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C7</td>
<td>3,309</td>
<td>63.8 (11.0)</td>
<td>2,025</td>
<td>1,284</td>
<td>0.889</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C8</td>
<td>2,802</td>
<td>71.1 (10.9)</td>
<td>1,867</td>
<td>935</td>
<td>0.943</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C9</td>
<td>2,581</td>
<td>63.7 (11.8)</td>
<td>1,803</td>
<td>778</td>
<td>1.314</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>C10</td>
<td>2,562</td>
<td>74.2 (9.6)</td>
<td>1,225</td>
<td>1,337</td>
<td>0.978</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C11</td>
<td>2,292</td>
<td>66.1 (11.0)</td>
<td>2,186</td>
<td>106</td>
<td>0.926</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C12</td>
<td>2,213</td>
<td>70.3 (10.2)</td>
<td>2,068</td>
<td>145</td>
<td>0.920</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C13</td>
<td>2,070</td>
<td>58.6 (10.2)</td>
<td>1,348</td>
<td>722</td>
<td>0.946</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C14</td>
<td>2,070</td>
<td>68.2 (9.6)</td>
<td>1,030</td>
<td>1,010</td>
<td>1.146</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C15</td>
<td>2,040</td>
<td>63.9 (10.1)</td>
<td>1,208</td>
<td>805</td>
<td>1.031</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C16</td>
<td>1,654</td>
<td>64.1 (12.1)</td>
<td>1,013</td>
<td>641</td>
<td>1.107</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C17</td>
<td>1,281</td>
<td>65.3 (9.9)</td>
<td>714</td>
<td>567</td>
<td>1.001</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C18</td>
<td>1,251</td>
<td>68.2 (9.8)</td>
<td>802</td>
<td>449</td>
<td>1.790</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>C19</td>
<td>1,168</td>
<td>58.5 (9.7)</td>
<td>995</td>
<td>173</td>
<td>0.752</td>
<td>&lt; 0.050</td>
</tr>
<tr>
<td>C20</td>
<td>1,119</td>
<td>71.5 (11.3)</td>
<td>713</td>
<td>406</td>
<td>1.213</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C21</td>
<td>1,000</td>
<td>61.0 (11.0)</td>
<td>769</td>
<td>231</td>
<td>1.116</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C22</td>
<td>988</td>
<td>69.2 (10.4)</td>
<td>516</td>
<td>472</td>
<td>1.023</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C23</td>
<td>935</td>
<td>58.7 (12.2)</td>
<td>588</td>
<td>347</td>
<td>1.609</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>C24</td>
<td>932</td>
<td>67.9 (10.1)</td>
<td>28</td>
<td>904</td>
<td>0.787</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C25</td>
<td>860</td>
<td>56.2 (9.9)</td>
<td>664</td>
<td>196</td>
<td>0.978</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C26</td>
<td>852</td>
<td>58.7 (12.1)</td>
<td>391</td>
<td>461</td>
<td>0.939</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C27</td>
<td>823</td>
<td>65.1 (10.9)</td>
<td>532</td>
<td>291</td>
<td>1.201</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C28</td>
<td>686</td>
<td>71.7 (8.0)</td>
<td>673</td>
<td>13</td>
<td>0.866</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C29</td>
<td>550</td>
<td>57.2 (11.1)</td>
<td>435</td>
<td>115</td>
<td>0.906</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>C30</td>
<td>533</td>
<td>61.2 (11.7)</td>
<td>391</td>
<td>172</td>
<td>1.874</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>C31</td>
<td>520</td>
<td>64.4 (11.2)</td>
<td>213</td>
<td>307</td>
<td>1.052</td>
<td>&gt; 0.050</td>
</tr>
<tr>
<td>NA*</td>
<td>5,113</td>
<td>60.1 (11.1)</td>
<td>3,878</td>
<td>1,235</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

*Patients that did not cluster or were in clusters of size < 500

HR: Hazard Ratio; Adj. P-val: Adjusted P-value.
Clusters, clinical outcomes, and phenotypic enrichment

To assess if the unsupervised clustering identified patient subgroups at different risks of disease progression, we used cluster-membership (C1-C31) as a covariate in a series of Cox models. A total of 14,679 patients experienced a new ischemic event during follow-up and 10,684 patients died from other causes than IHD. Mean follow-up times was 3.72 years (Table 1). Risks for new ischemic events and death from non-IHD causes in each cluster were compared to the pooled risk for patients in the remaining 30 clusters. The survival analysis demonstrated that the MCL algorithm stratified patients according to risk of new ischemic events and death from non-IHD causes (Fig 3). Comparing each cluster to all the others (n=30), seven clusters (20,221 patients) had a statistically significantly higher (five clusters) or lower (two clusters) risk of new ischemic events; and 18 clusters (43,173 patients) had a higher (14 clusters) or lower (four clusters) risk of death from non-IHD causes. All clusters at increased risk of new ischemic events, associated with risk of death from non-IHD causes as well. Thus, 11 clusters (23,963 patients) did not have altered risk of the two outcomes, when compared to the other clusters (Fig 1D).
**Fig 3: Risk of new ischemic events and non-IHD causes stratified by cluster.**

Forest plots where clusters are shown against HR for new ischemic events (left) and death from non-IHD causes (right). X-axis: HR for a single cluster relative to mean HR of the 30 other clusters. Y-axis: Clusters arranged by risk of new ischemic events, increasing risk from top to bottom. IHD: Ischemic heart disease. HR: Hazzard ratio.
The distribution of O/E-ratios was heavily left-skewed as less than 99% (n=101) of all O/E-ratios were >10 and roughly 7% (n=887) of all O/E-ratios were >2. About 60% of all O/E-ratios (n=8,056) were in the range of 0 and 1 corresponding to inverse changes. Generally, clusters that had high risk of new ischemic events or death from non-IHD causes were also characterized by large, summarized O/E-values corresponding to a high degree of multimorbidity (S5 Table). To obtain an overview of the results of the risk stratification in conjunction with the enrichment analysis, clusters were described according to three categories: (1) clusters enriched for cardiometabolic or -vascular diseases, (2) clusters enriched for degenerative or inflammatory diseases and (3) clusters enriched characterized by little enrichment and inverse changes.

Clusters enriched for cardiometabolic and -vascular diseases

Four of the five clusters at increased risk of new ischemic events (and death from non-IHD causes) were enriched for diabetes (C5, C18, C23, and C30). In these four clusters, HRs ranged from 1.40 (C5, 95%CI: 1.30;1.50, adj. P-val. < 0.001) to 1.88 (C30, 95%CI: 1.60;2.00, adj. P-val. < 0.001) with a significant difference in age at index (C5: 63.9 years, C30: 61.2 years, Adj. P-val. < 0.001, TukeyHSD). C18 and C23 were only enriched for insulin-dependent diabetes, but differed in that C18 was also enriched for insulin-dependent diabetes with vascular complications and periphery atherosclerosis. In contrast, C5 was only enriched for non-insulin dependent diabetes and included diabetes with as well as without complications. Lastly, C30 was only enriched for diabetes with complications (insulin and non-insulin dependent) and was the diabetes cluster enriched for chronic kidney disease and bacterial infections, as well (S5 Table).
Other cardiac diseases that displayed enrichment were supraventricular arrhythmias (C4), cardiomyopathies (C9), and valve diseases (C20). Of the three clusters, only C9 had increased risk of new ischemic events (HR: 1.31 (C9, 95%CI: 1.20;1.44, adj. P-val.: < 0.001). Risk of death from non-IHD causes was 1.79 (95%CI: 1.60;2.00, adj. P-val. < 0.001). In contrast, C4 and C20 only had increased risk of death from non-IHD causes with HRs of 1.49 (C4, 95%CI: 1.34;1.59, adj. P-val. < 0.001) and 1.78 (C20, 95%CI: 1.54;2.04, adj. P-val. < 0.001). Interestingly, the cluster enriched for cerebrovascular diseases (C27) did not have altered risk of any of the two outcomes. In sum, all clusters that had increased risk of new ischemic events were enriched for cardiometabolic diseases, albeit not all clusters enriched for cardiometabolic and -vascular diseases had increased risk of new ischemic events (Table 2 and S5 Table).

Clusters enriched for degenerative or inflammatory diseases

Six clusters (C7, C13, C14, C22, C26, and C31) were enriched for diagnoses describing degenerative or inflammatory diseases, i.e., osteoarthritis (C7), degenerative spine disease (C13 and C22), chronic obstructive pulmonary disease (C14), asthma (C26), and rheumatoid arthritis (C31). Remarkably, none of the four clusters had increased risk of new ischemic events and only one cluster (C14) had increased risk of death from non-IHD causes (HR: 3.39, 95%CI: 3.09;3.71, adj. P-val. < 0.001). Conversely, C7 and C13 had reduced risk of death from non-IHD causes (C7, HR: 0.61, 95%CI: 0.52;0.72, adj. P-val. < 0.001 and C13, HR: 0.58, 95%CI: 0.45;0.74, adj. P-val. < 0.001). Age at index for the clusters enriched for degenerative or inflammatory diseases range between 58.6 years (C13) and 69.2 years (C22) (Table 2). Taken together, these findings hint to the dual nature of inflammation as a potential disease modifier as well as a risk factor.
Clusters characterized by little enrichment and inverse changes

Six clusters (C1, C2, C3, C6, C15, and C17) were characterized by little enrichment, which included the two clusters with reduced risk of new ischemic events (C2, HR: 0.82, 95%CI: 0.76;0.89, adj. P-val. < 0.001 and C3, HR: 0.76, 95%CI: 0.52;0.69, adj. P-val. < 0.001). Not surprisingly, none of these six clusters had increased risk of either of the two outcomes, but three clusters (C2, C3, and C6) had reduced risk of death from non-IHD causes (C2, HR: 0.60, 95%CI: 0.52;0.69, adj. P-val. < 0.001, C3, HR: 0.59, 95%CI: 0.59;0.69, adj. P-val. < 0.001 and C6, HR: 0.68, 95%CI: 0.57;0.79, adj. P-val. < 0.001) (Table 2). It was a common attribute of the clusters without altered risk of any of the two outcomes that O/E-ratios for hypertension and dyslipidemia were among the largest. In contrast, diabetes, heart failure, and chronic obstructive pulmonary disease frequently displayed inverse changes (O/E-ratios < 1) in these clusters (S5 Table). Taken together, these observations indicate that risk of disease progression in this populations necessitates a more sophisticated analysis of multimorbidity.

For a list with results of the enrichment analysis for all clusters, including the 13 clusters not described above, S5 Table.

Clusters and their association with laboratory measurements and genetic data

Clusters were also characterized by means of datatypes not included among the MCL algorithm input features. For patients in the 31 clusters, we had laboratory measurements on 30,755 (49.5%) and genetic data on 19,422 (31.3%). To assess if the phenotypic differences captured by the MCL algorithm were also reflected in laboratory measurements, we tested if the distributions of test results within and out of reference ranges differed significantly. There were significantly different distributions of tests within and out of reference ranges in clusters.
for the 24 most frequent tests. Overall, this indicates that the phenotypic patterns within the
entire spectrum of cardiovascular multimorbidity registered before index correlate with
results of clinical laboratory tests (S6 Table). Thus, these findings are a strong indicator that
the patterns captured by the MCL algorithm are biologically relevant. For a graphical
summary of the laboratory scores in each cluster, see S4 Figure.

Finally, we identified 41 cases (out of 434 tests) where the PRS distribution for a specific
trait in cluster was significantly different from that trait’s combined PRS distribution of the
other 30 clusters. Among these cases, we found the largest effects size to be a higher genetic
risk for atrial fibrillation in cluster C4 (0.57, FDR < 0.001) as well as a higher genetic risk for
non-insulin dependent diabetes in cluster C5 (0.55, FDR < 0.001). These findings are
congruent with the results of the enrichment analysis for C4 and C5, respectively. In contrast,
C1 (phenotypically characterized by inverse changes) had relatively large, positive effect
sizes for systolic as well as diastolic blood pressure (0.20 and 0.16, FDR < 0.001). Similarly,
there were positive effect sizes for total cholesterol and triglycerides in C6, which was also
characterized by little phenotypic enrichment as well as a high degree of inverse changes. A
list of significant effect sizes for the 41 significant cases, see S7 Table.

**Discussion**

In this study, we developed a novel, data-driven method for structuring the entire spectrum of
multimorbidity by means of an unsupervised clustering analysis. In a cohort of 72,249
patients with IHD patients, we identified 31 distinct clusters (67,136 patients) based on 3,046
diagnoses assigned prior to or at index. By comparing risk of new ischemic events and death
from non-IHD causes across clusters and then performing an enrichment analysis, we found
that clusters at increased risk of new ischemic events were enriched for diabetes (four
clusters) or cardiomyopathies (one cluster). Neither the cluster enriched for supraventricular
arrythmias, nor valve diseases had increased risk of new ischemic events. Degenerative and inflammatory diseases were enriched in a total of six clusters and displayed no clear trend in their relation to the outcomes. The results of the enrichment analysis were supported by trends in laboratory test results and clusters enriched for supraventricular arrhythmias and non-insulin diabetes also had congruently, higher genetic risks.

The results of the study agree with common knowledge on risk of IHD, while also adding insights to the disease-disease associations, which are currently underappreciated in the literature. The fact that clusters enriched for diabetes were generally the most high-risk clusters serves as a methodological reality check (6). Added value of the study lies in the fact that the method allows for a more sophisticated description of such associations, as the method allows to study the entire spectrum of multimorbidity. For example, four clusters were enriched for diabetes, which is in line with the current paradigm that a single term is insufficient to describe a multifactorial disease, such as diabetes (16, 29). By integrating different data types, the findings indicate how phenotypic and genetic data complement each other, by exemplifying (1) that clustering analysis facilitates stronger genetic signals in patient subgroups and (2) that genetic data may unveil patterns not captured by phenotypic data alone.

In addition, the method developed in this study and subsequent findings add perspective to the relatively limited body of literature regarding associations between chronic inflammatory and cardiovascular diseases (7). While previous studies have concluded that the risk of cardiovascular diseases is increased in most chronic inflammatory disorders, the results of our study indicate that pre-existing degenerative or inflammatory disorders in patients with IHD do not increase the risk of new ischemic events.
The pre-selected outcomes in the present study are also a unique aspect of the study, as previous clustering analyses within the cardiovascular domain studies have mainly analyzed all-cause mortality (17,18). This aspect of the study allows to distinguish between risk of progression related to IHD and risk of progression that is related to comorbidity drawing attention to important aspects of multimorbidity in this domain. For example, clusters enriched for supraventricular arrhythmias and chronic obstructive pulmonary disease, respectively, only had increased risk of death from non-IHD causes. The study design, including the enrichment analysis, also revealed that classical risk factors for IHD (e.g., hypertension and dyslipidemia) did not drive the clustering. This finding agrees with previously published comorbidity phenotypes in patients with IHD (18). We argue that the present study displays that continuous exploration and characterization of multimorbidity in IHD are key elements in optimizing the exploit the full potential of continuously developing treatment strategies.

Previous clustering analyses within the cardiovascular domain have typically included either thousands of patients or hundreds of input features, but not both (14,15). For example, Hall et al. defined multimorbidity using only eight different chronic conditions, whereas Crowe et al. defined multimorbidity with reference to 20 predefined conditions (17,18). Thus, the scale of our study exceeds that of previous work, as it includes more than 70,000 patients and more than 3,000 input features. And further, we limited the risk of introducing bias by not exerting feature selection prior to clustering.

The two main limitations with respect to the data foundation are that (1) owing to the novelty of the method, there were no standardized way of assessing the representation of multimorbidity and (2) it was only a subset for which laboratory and genetic data were
available. These challenges are naturally overcome in clustering analyses based on data from randomized controlled trials, such as the studies by Inohara et al, and Karwath et al. (15,19)

However, in the present, data-rich era, we argue that it is highly important to develop methods for structuring and studying other data than what is being collected for trials. Ideally, the two approaches, based on nationwide data and randomized controlled trials, respectively, will complement each other; and will facilitate more precise identification of patients who are likely to benefit from different treatment options as well as guide optimized selection of patients for randomized controlled trials.

In sum, the study further showcases the strengths of a more fine-grained analysis of patient subgroups, which, in turn, may pave the way for successful implementation of precision medicine. Owing to its flexibility, the comprehensive, data-driven analysis of cardiovascular multimorbidity represents a novel method for characterizing multimorbidity in IHD with great potential of applying it to other diseases of interest or other clinical data. Such trends may guide clinical decision making in cases, where for example it is not obvious how to manage the angiographic findings or the combination of drugs that a specific patient will benefit most from.

In conclusion, the present study cements the complexity of multimorbid patients with IHD and exemplifies the clinical relevance of a more fine-grained patient subgrouping by carrying out a cluster-based risk-stratifying the cohort. Further, owing to its flexibility, the comprehensive, data-driven method of cardiovascular multimorbidity presented here represents a novel method for characterizing multimorbidity in IHD with great potential. Improved patient subgrouping may be critical guide future clinical decision making in cases,
where it is non-trivial how to manage the angiographic findings or to find the optimal combination of drugs for a given patient.
Acknowledgement

The authors would like to thank (1) research programmer, Troels Siggaard, Novo Nordisk Foundation Center for Research, University of Copenhagen, Denmark for continuous and reliable infrastructure support, and (2) Head of Cardiovascular Research, Hilma Hólm, deCODE genetics, Iceland for insightful comments.
References


28. MCL - a cluster algorithm for graphs [Internet]. [cited 2023 Jan 30]. Available from: http://micans.org/mcl/


42. R Core Team. R: A Language and Environment for Statistical Computing [Internet].
   https://www.R-project.org/

   Sustainable data analysis with Snakemake [Internet]. F1000Research; 2021 [cited 2023
   Mar 16]. Available from: https://f1000research.com/articles/10-33

44. Violán C, Roso-Llorach A, Foguet-Boreu Q, Guisado-Clavero M, Pons-Viguès M, Pujol-
Supporting information captions

S1 Fig: Classification of new ischemic events.

S1 Table: Eligible codes for inclusion and outcomes

S1 Appendix: MCL algorithm settings and assessment of cluster robustness
- S2 Fig: Selection of number of components.
- S3 Fig: Limiting edge-density and average node degree in sex-specific similarity networks.

S2 Appendix: Preprocessing of laboratory data
- S2 Table: Laboratory codes included in assessment of data quality and completeness

S3 Appendix: Calculation of polygenetic risk scores for 14 traits

S3 Table: Comparison of mean age at index in 31 cluster using Tukey’s HSD

S4 Table: Demographics for patients not cluster or were in clusters of size < 500

S5 Table: Cluster-wise summarized O/E-ratios, 10 largest O/E-ratios and 10 lowest O/E-ratios.

S6 Table: Chi-squared test for distribution laboratory values in clusters

S7 Table: Traits with significantly different PGS distributions in clusters