

# 1 **Age-dependent topic modelling of comorbidities in UK Biobank identifies** 2 **disease subtypes with differential genetic risk**

3

4 Xilin Jiang<sup>1,2,3,4,5,6</sup>, Martin Jinye Zhang<sup>4,7§</sup>, Yidong Zhang<sup>1,8,9§</sup>, Arun Durvasula<sup>4,7,10,11§</sup>, Michael  
5 Inouye<sup>5,6,12,13,14,15</sup>, Chris Holmes<sup>1,2,16</sup>, Alkes L. Price<sup>4,7,17\*</sup>, Gil McVean<sup>1\*</sup>

6

## 7 **Affiliations**

8 <sup>1</sup> Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of  
9 Oxford, Oxford OX3 7LF, UK

10 <sup>2</sup> Department of Statistics, University of Oxford, Oxford OX1 3LB, UK

11 <sup>3</sup> Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK

12 <sup>4</sup> Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

13 <sup>5</sup> British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and  
14 Primary Care, University of Cambridge, Cambridge UK

15 <sup>6</sup> Heart and Lung Research Institute, University of Cambridge, Cambridge UK

16 <sup>7</sup> Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge,  
17 MA, USA

18 <sup>8</sup> CAMS China Oxford Institute, Nuffield Department of Medicine, University of Oxford,  
19 Oxford OX3 7BN, UK

20 <sup>9</sup> Department of Radiation Oncology, Peking Union Medical College Hospital, Chinese  
21 Academy of Medical Sciences and Peking Union Medical College, Beijing, China

22 <sup>10</sup> [Department of Genetics, Harvard Medical School, Cambridge, MA, USA](#)

23 <sup>11</sup> [Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA](#)

24 <sup>12</sup> Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary  
25 Care, University of Cambridge, Cambridge, UK.

26 <sup>13</sup> Health Data Research UK Cambridge, Wellcome Genome Campus and University of  
27 Cambridge, Cambridge, UK.

28 <sup>14</sup> British Heart Foundation Cambridge Centre of Research Excellence, Department of Clinical  
29 Medicine, University of Cambridge, Cambridge, UK.

30 <sup>15</sup> Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute,  
31 Melbourne, VIC, Australia.

32 <sup>16</sup> The Alan Turing Institute, London NW1 2DB, UK

33 <sup>17</sup> Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

34

35 § These authors contributed equally to this work

36 \*These authors jointly supervised the work

37 Corresponding authors:

38 [xilinjiang@hsph.harvard.edu](mailto:xilinjiang@hsph.harvard.edu)

39 [aprice@hsph.harvard.edu](mailto:aprice@hsph.harvard.edu)

40 [gil.mcvean@bdi.ox.ac.uk](mailto:gil.mcvean@bdi.ox.ac.uk)

41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60  
61  
62  
63  
64  
65  
66  
67  
68  
69  
70  
71  
72  
73  
74  
75  
76  
77

## Abstract

The analysis of longitudinal data from electronic health records (EHR) has potential to improve clinical diagnoses and enable personalised medicine, motivating efforts to identify disease subtypes from age-dependent patient comorbidity information. Here, we introduce an age-dependent topic modelling (ATM) method that provides a low-rank representation of longitudinal records of hundreds of distinct diseases in large EHR data sets. The model learns, and assigns to each individual, topic weights for several disease topics, each of which reflects a set of diseases that tend to co-occur within individuals as a function of age. Simulations show that ATM attains high accuracy in distinguishing distinct age-dependent comorbidity profiles. We applied ATM to 282,957 UK Biobank samples, analysing 1,726,144 disease diagnoses spanning all 348 diseases with  $\geq 1,000$  independent occurrences in the Hospital Episode Statistics (HES) data, identifying 10 disease topics under the optimal model fit. Analysis of an independent cohort, All of Us, with 211,908 samples and 3,098,771 disease diagnoses spanning 233 of the 348 UK Biobank diseases produced highly concordant findings. In UK Biobank we identified 52 diseases with heterogeneous comorbidity profiles ( $\geq 500$  occurrences assigned to each of  $\geq 2$  topics), including breast cancer, type 2 diabetes (T2D), hypertension, and hypercholesterolemia. For most of these diseases, topic assignments were highly age-dependent, suggesting differences in disease aetiology for early-onset vs. late-onset disease. We defined subtypes of the 52 heterogeneous diseases based on the topic assignments, and compared genetic risk across subtypes using polygenic risk scores (PRS). We identified 18 disease subtypes whose PRS differed significantly from other subtypes of the same disease, including a subtype of T2D characterised by cardiovascular comorbidities and a subtype of asthma characterised by dermatological comorbidities. We further identified specific variants underlying these differences such as a T2D-associated SNP in the *HMG2* locus that has a higher odds ratio in the top quartile of cardiovascular topic weight ( $1.18 \pm 0.02$ ) compared to the bottom quartile ( $1.00 \pm 0.02$ ) ( $P = 3 \times 10^{-7}$  for difference,  $FDR = 0.0002 < 0.1$ ). In conclusion, ATM identifies disease subtypes with differential genome-wide and locus-specific genetic risk profiles.

## 78 Introduction

79 Longitudinal electronic health record (EHR) data, encompassing diagnoses across hundreds of  
80 distinct diseases, offers immense potential to improve clinical diagnoses and enable personalised  
81 medicine<sup>1</sup>. Despite intense interest in both the genetic relationships between distinct diseases<sup>2–11</sup>  
82 and the genetic relationships between biological subtypes of disease<sup>12–15</sup>, there has been limited  
83 progress on classifying disease phenotypes into groups of diseases with frequent co-occurrences  
84 (comorbidities) and leveraging comorbidities to identify disease subtypes. Low-rank modelling  
85 has appealing theoretical properties<sup>16,17</sup> and has produced promising applications<sup>18–24</sup> to infer  
86 meaningful representations of high-dimensional data. In particular, low-rank representation is an  
87 appealing way to summarise data across hundreds of distinct diseases<sup>25–27</sup>, providing the  
88 potential to identify patient-level comorbidity patterns and distinguish disease subtypes. **The**  
89 **biological differentiation of** disease subtypes inferred from EHR data could be validated by  
90 comparing genetic profiles across subtypes, which is possible with emerging data sets that link  
91 genetic data with EHR data<sup>28–31</sup>.

92  
93 Previous studies have used low-rank representation to identify shared genetic components<sup>25–27</sup>  
94 across multiple distinct diseases, identifying relationships between diseases and generating  
95 valuable biological insights. However, age at diagnosis information in longitudinal EHR data has  
96 the potential to improve such efforts. For example, a recent study used longitudinal disease  
97 trajectories to identify disease pairs with statistically significant directionality<sup>32</sup>, suggesting that  
98 age information could be leveraged to infer comorbidity profiles that capture temporal  
99 information. In addition, patient-level comorbidity information could potentially be leveraged to  
100 identify biological subtypes of disease, complementing its application to increase power for  
101 identifying genetic associations<sup>12</sup> and to cluster disease-associated variants into biological  
102 pathways<sup>8</sup>; disease subtypes are fundamental to disease aetiology<sup>14,33–36</sup>.

103  
104 Here, we propose an age-dependent topic modelling (ATM) method to provide a low-rank  
105 representation of longitudinal disease records. ATM learns, and assigns to each individual, topic  
106 weights for several disease topics, each of which reflects a set of diseases that tend to co-occur  
107 **within individuals as a function of age**. We applied ATM to 1.7 million disease diagnoses  
108 spanning 348 diseases in UK Biobank, inferring 10 disease topics; **we validated ATM in All of**  
109 **Us**. We identified 52 diseases with heterogeneous comorbidity profiles that enabled us to define  
110 disease subtypes. We used genetic data to validate the disease subtypes, showing that they  
111 exhibit differential genome-wide and locus-specific genetic risk profiles.

112

113

## 114 **Results**

### 115 *Overview of methods*

116 We propose an age-dependent topic modelling (ATM) model, [which provides](#) a low-rank  
117 representation of longitudinal records of hundreds of distinct diseases in large EHR data sets  
118 (Figure 1, Methods). The model assigns to each individual *topic weights* for several *disease*  
119 *topics*; each disease topic reflects a set of diseases that tend to co-occur as a function of age,  
120 quantified by age-dependent *topic loadings* for each disease. The model assumes that for each  
121 disease diagnosis, a topic is sampled based on the individual's topic weights (which sum to 1  
122 across topics, for a given individual), and a disease is sampled based on the individual's age and  
123 the age-dependent topic loadings (which sum to 1 across diseases, for a given topic at a given  
124 age). The model generalises the latent dirichlet allocation (LDA) model<sup>37,38</sup> by allowing topic  
125 loadings for each topic to vary with age (Supplementary Note, Supplementary Figure 1).

126  
127 We developed a method to fit this model that addresses several challenges inherent to large EHR  
128 data sets; the method estimates topic weights for each individual, topic loadings for each disease,  
129 and posterior diagnosis-specific topic probabilities for each disease diagnosis. First, we derived a  
130 scalable deterministic method that uses numerical approximation approaches to fit the  
131 parameters of the model, addressing the challenge of computational cost. Second, we used the  
132 prediction odds ratio<sup>39</sup> to compare model structures (e.g. number of topics and parametric form  
133 of topic loadings as a function of age), addressing the challenge of appropriate model selection;  
134 roughly, the prediction odds ratio quantifies the accuracy of correctly predicting disease  
135 diagnoses in held-out [individuals](#) using comorbidity information, compared to a predictor based  
136 only on prevalence (see Methods and Supplementary Table 1). Third, we employed collapsed  
137 variational inference<sup>40</sup>, addressing the challenge of sparsity in the data (e.g. in UK Biobank data,  
138 the average patient has diagnoses for 6 of 348 diseases analysed); collapsed variational inference  
139 outperformed mean-field variational inference<sup>37</sup> in empirical data. Further details are provided in  
140 the Methods section and Supplementary Note; we have publicly released open-source software  
141 implementing the method (see Code Availability).

142  
143 We applied ATM to longitudinal records of [UK Biobank](#)<sup>29</sup> ([282,957 individuals with 1,726,144](#)  
144 [disease diagnoses spanning 348 diseases; the targeted individuals are those diagnosed with at](#)  
145 [least two of the 348 diseases studied](#)) and [All of Us](#)<sup>30</sup> ([211,908 individuals with 3,098,771](#)  
146 [disease diagnoses spanning 233 of the 348 diseases](#)). Each disease diagnosis has an associated  
147 age-at-diagnosis, defined as the earliest age of reported diagnosis of the disease in that  
148 individual; we caution that age at diagnosis may differ from age at disease onset (see  
149 Discussion). ATM does not use genetic data, but we used genetic data to validate the inferred  
150 topics (Methods).

151  
152 *Simulations*

153 We performed simulations to compare ATM with latent dirichlet allocation (LDA)<sup>37,38</sup>, a simpler  
154 topic modelling approach that does not model age. We simulated 61,000 disease diagnoses  
155 spanning 20 diseases in 10,000 individuals, using the ATM generative model; **we aimed to**  
156 **choose simulation parameters that resemble real data. In detail**, the average number of disease  
157 diagnoses per individual (6.1), ratio of #individuals/#diseases (500), topic loadings, and standard  
158 deviation in age at diagnosis (8.5 years for each disease) were chosen to match empirical UK  
159 Biobank data; **we varied the number of topics, number of individuals, and number of diseases in**  
160 **secondary analyses (see below)**. We assigned each disease diagnosis to one of two subtypes for  
161 the **target** disease based on age and other subtype differences, considering high, medium, or low  
162 age-dependent effects by specifying an average difference of 20, 10, or 5 years respectively in  
163 age at diagnosis for the two subtypes. For each level of age-dependent effects, we varied the  
164 proportion of diagnoses belonging to the first subtype (*i.e. the subtype that has an earlier average*  
165 *age-at-diagnosis*) from 10-50%. Further details of the simulation framework are provided in the  
166 Methods section. Our primary metric for evaluating the LDA and ATM methods **is the area**  
167 **under the precision-recall curve (AUPRC)<sup>41</sup> metric**, where precision is defined as the proportion  
168 of disease diagnoses that a given method assigned to the first subtype that were assigned  
169 correctly and recall is defined as the proportion of disease diagnoses truly belonging to the first  
170 subtype that were assigned correctly. We discretized the subtype assigned to each disease  
171 diagnosis by a given method by assigning the subtype with higher inferred probability. We note  
172 that AUPRC is larger when classifying the **minority** subtype; results using the second subtype as  
173 the classification target are also provided. We used AUPRC (instead of prediction odds ratio) in  
174 our simulations because the underlying truth is known. Further details and justifications of  
175 metrics used in this study are provided in the Methods Section and Supplementary Table 1.

176  
177 In simulations with high age-dependent effects, ATM attained much higher AUPRC than LDA  
178 across all values of subtype sample size proportion (AUPRC difference: 24%-42%), with both  
179 methods performing better at more balanced ratios (Figure 2, Supplementary Table 2).  
180 Accordingly, ATM attained both higher precision and higher recall than LDA (Supplementary  
181 Figure 2). Results were qualitatively similar when using the second subtype as the classification  
182 target (Supplementary Figure 3). In simulations with medium or low age-dependent effects,  
183 ATM continued to outperform LDA but with smaller differences between the methods. In  
184 simulations without age-dependent effects, ATM slightly underperformed LDA (Supplementary  
185 Figure 4A).

186  
187 We performed three secondary analyses. First, we varied the number of individuals, number of  
188 diseases, or number of disease diagnoses per individual. ATM continued to outperform LDA in  
189 each case, although increasing the number of individuals or the number of disease diagnoses per  
190 individual did not always increase AUPRC (Supplementary Figure 4B). Second, we performed  
191 simulations in which we increased the number of subtypes from two to five and changed the  
192 number of diseases to 50, and compared ATM models trained using different numbers of topics

193 (in 80% training data) by computing the prediction odds ratio; we used the prediction odds ratio  
194 (instead of AUPRC) in this analysis both because it is a better metric to evaluate the overall  
195 model fit to the data, and because it is unclear how to compare AUPRC across scenarios of  
196 varying topic numbers (see Supplementary Table 1). We confirmed that the prediction odds  
197 ratio was maximised using five topics, validating the use of the prediction odds ratio for model  
198 selection (Supplementary Figure 5A). Third, we computed the accuracy of inferred topic  
199 loadings, topic weights, and grouping accuracy (defined as proportion of pairs of diseases truly  
200 belonging to the same topic that ATM correctly assigned to the same topic), varying the number  
201 of individuals and number of diseases diagnoses per individual. We determined that ATM also  
202 performed well under these metrics (Supplementary Figure 5B-E).

203  
204 We conclude that ATM (which models age) assigns disease diagnoses to subtypes with higher  
205 accuracy than LDA (which does not model age) in simulations with age-dependent effects. We  
206 caution that our simulations largely represent a best-case scenario for ATM given that the  
207 generative model and inference model are very similar (although there are some differences, e.g.  
208 topic loadings were generated using a model different from the inference model), thus it is important  
209 to analyse empirical data to validate the method.

210  
211 *Age-dependent disease topic loadings capture comorbidity profiles in the UK Biobank*  
212 We applied ATM to longitudinal hospital records of 282,957 individuals from the UK Biobank  
213 with an average record span of 28.6 years<sup>29</sup>. We used Phecode<sup>42</sup> to define 1,726,144 disease  
214 diagnoses spanning 348 diseases with at least 1,000 diagnoses each; the average individual had  
215 6.1 disease diagnoses, and the average disease had a standard deviation of 8.5 years in age at  
216 diagnosis. The optimal inferred ATM model structure has 10 topics and models age-dependent  
217 topic loadings for each disease as a spline function with one knot, based on optimizing prediction  
218 odds ratio (see below). We assigned names (and corresponding acronyms) to each of the 10  
219 inferred topics based on the Phecode systems<sup>42</sup> assigned to diseases with high topic loadings  
220 (aggregated across ages) for that topic (Table 1, Supplementary Table 3).

221  
222 Age-dependent topic loadings across all 10 topics and 348 diseases (stratified into Phecode  
223 systems), summarised as averages across age < 60 and age ≥ 60, are reported in Figure 3,  
224 Supplementary Figure 6, and Supplementary Table 4. Some topics such as NRI span diseases  
225 across the majority of Phecode systems, while other topics such as ARP are concentrated in a  
226 single Phecode system. Conversely, a single Phecode system may be split across multiple topics,  
227 e.g. diseases of the digestive system are split across UGI, LGI, and MDS. We note that topic  
228 loadings in diseases that span multiple topics are heavily age-dependent. For example, type 2  
229 diabetes patients assigned to the CVD topic are associated with early onset of type 2 diabetes  
230 whereas type 2 diabetes patients assigned to the MGND topic are associated with late onset of  
231 type 2 diabetes.

232

233 We performed seven secondary analyses to validate the [integrity and reproducibility of](#) inferred  
234 comorbidity topics. First, we fit ATM models with different model structures using 80% training  
235 data, and computed their prediction odds ratios using 20% testing data. The ATM model  
236 structure with 10 topics and age-dependent topic loadings modelled as a spline function  
237 performed optimally (Supplementary Figure 7; see Methods). Second, we confirmed that ATM  
238 [\(which models age\)](#) attained higher prediction odds ratios than LDA [\(which does not model age\)](#)  
239 across different values of the number of topics (Supplementary Figure 8); [for the optimal model](#)  
240 [with 10 topics, ATM attained an average prediction odds ratio of 1.71, compared to a prediction](#)  
241 [odds ratio of 1.58 for LDA.](#) Third, we reached similar conclusions using evidence lower  
242 bounds<sup>39</sup> (ELBO; see Supplementary Table 1) (Supplementary Figure 9). Fourth, we confirmed  
243 that collapsed variational inference<sup>40</sup> outperformed mean-field variational inference<sup>37</sup>  
244 (Supplementary Figure 10). Fifth, we computed a co-occurrence odds ratio evaluating whether  
245 diseases grouped into the same topic by ATM in the training data have higher than random  
246 probability of co-occurring in the testing data (Supplementary Table 1). The co-occurrence odds  
247 ratio is consistently above one and increases with the number of comorbid diseases, for each  
248 inferred topic (Supplementary Figure 11). Sixth, we compared the topic loadings by repeating  
249 the inference on female-only or male-only populations and observed no major discrepancies,  
250 except for genitourinary topics MGND and FGND (topic loading  $R^2$  (female vs. all) = 0.788,  
251 topic loading  $R^2$  (male vs. all) = 0.773, Supplementary Figure 12). Lastly, we verified that BMI,  
252 sex, Townsend deprivation index, and birth year explained very little of the information in the  
253 inferred topics (Supplementary Table 3).

254  
255 Disease topics capture known biology as well as the age-dependency of comorbidities for the  
256 same diseases. For example, early onset of essential hypertension is associated with the CVD  
257 topic<sup>43</sup>, which captures the established connection between lipid dysfunction  
258 (“hypercholesterolemia”) and cardiovascular diseases<sup>44</sup>, while later onset of essential  
259 hypertension is associated with the CER topic, which pertains to type 2 diabetes, obesity and  
260 COPD (Figure 4A). Continuously varying age-dependent topic loadings for all 10 topics,  
261 restricted to diseases with high topic loadings, are reported in Supplementary Figure 13 and  
262 Supplementary Table 5. We note that most diseases have their topic loadings concentrated into a  
263 single topic (Figure 4B, Supplementary Figure 14A and Supplementary Table 4), and that most  
264 individuals have their topic weights concentrated into 1-2 topics (Figure 4C and Supplementary  
265 Figure 14B). For diseases spanning multiple topics (Supplementary Figure 6 and Supplementary  
266 Table 4), the assignment of type 2 diabetes patients to the CVD topic is consistent with known  
267 pathophysiology and epidemiology<sup>45,46</sup> and has been shown in other comorbidity clustering  
268 studies, e.g. with the Beta Cell and Lipodystrophy subtypes described in ref.<sup>35</sup> and the severe  
269 insulin-deficient diabetes (SIDD) subtype described in ref.<sup>14</sup>, which are characterised by early  
270 onset of type 2 diabetes and have multiple morbidities including hypercholesterolemia,  
271 hyperlipidemia, and cardiovascular diseases<sup>47</sup>. In addition, early-onset breast cancer and late-  
272 onset breast cancer are associated with different topics, e.g. NRI and FGND, consistent with

273 known treatment effects for breast cancer patients which increase susceptibility to infections,  
274 especially bacterial pneumonias<sup>48</sup> and hypothyroidism<sup>49</sup>. We conclude that ATM identifies  
275 latent disease topics that robustly compress age-dependent comorbidity profiles and capture  
276 disease comorbidities both within and across Phecode systems.

277  
278 *Age-dependent disease topic loadings capture concordant comorbidity profiles in All of Us*  
279 To assess the transferability of inferred topics between cohorts, we applied ATM to longitudinal  
280 data from 211,908 All of Us samples<sup>30</sup>. We analysed 3,098,771 diagnoses spanning 233 of the  
281 348 diseases analysed in UK Biobank for which data were available; the average individual had  
282 14.6 disease diagnoses, and an average disease had a standard deviation of 14.0 years in age at  
283 diagnosis. The optimal model for All of Us included 13 topics (Supplementary Figure 15,  
284 Supplementary Figure 16A-B, and Supplementary Table 6); most diseases have their topic  
285 loadings concentrated into a single topic and most individuals have their topic weights  
286 concentrated into 1-4 topics (Supplementary Figure 17).

287  
288 We assessed the concordance between each UK Biobank topic and each All of Us topic by  
289 computing the correlation between the respective topic loadings across the 233 diseases analysed  
290 in both data sets. Results are reported in Figure 5A-B, Supplementary Figure 18, and  
291 Supplementary Table 7. The median correlation between the ten UK Biobank topics and the  
292 most similar All of Us topic was 0.54, confirming qualitative alignment of topic loadings  
293 between All of Us and UK Biobank. For example, the topic loadings of CVD and CER topics  
294 were qualitatively similar to the most similar All of Us topics (Figure 5A vs. Figure 4A), even  
295 though disease prevalences differ between the two cohorts (Supplementary Table 8). When using  
296 the optimal All of Us model (13 topics) to predict diagnoses in UK Biobank, we obtained a  
297 prediction odds ratio that was significantly larger than 1 (mean = 1.32; jackknife s.e. = 0.0027,  
298 Supplementary Figure 16C). We note 3 key differences between All of Us and UK Biobank data:  
299 (i) All of Us contains primary care and hospital data encoded using SNOMED clinical terms,  
300 whereas UK Biobank uses hospitalization episode statistics (HES; encoded using ICD-10 clinical  
301 terms); (ii) All of Us is based on the U.S. population and U.S. health care system whereas UK  
302 Biobank is based on the UK population and UK health care system, which impacts diagnostic  
303 criteria and age at diagnosis; and (iii) All of Us individuals have different ancestries and  
304 socioeconomic backgrounds (including 26% African and 17% Latino; 78% of All of Us  
305 represents groups historically underrepresented in biomedical research based on race, ethnicity,  
306 age, gender identity, disability status, medical care access, income, and educational attainment)  
307 than UK Biobank individuals (94% European ancestry with higher than average income and  
308 educational attainment). We consider the cross-cohort prediction odds ratio of 1.32 to be an  
309 encouraging result given these key differences.

310  
311 For each of the 233 diseases, we assessed the concordance between UK Biobank and All of Us  
312 topic assignments for that disease by computing the correlation between UK Biobank topic  
313 assignments and All of Us topic assignments that were mapped to UK Biobank topics (by



314 weighting by correlations between topics; Methods). The average correlation between UK  
315 Biobank and All of Us topic assignments for the same disease was 0.70 (vs. average correlation  
316 of 0.02 for different diseases) (Figure 5C, Supplementary Figure 19, and Supplementary Table  
317 9). We conclude that ATM identifies latent disease topics from the All of Us cohort that align  
318 with topics from UK Biobank.

319  
320 *Disease subtypes defined by distinct topics are genetically heterogeneous*  
321 We sought to define disease subtypes in UK Biobank data based on the topic weights of each  
322 patient and diagnosis-specific topic probabilities of each disease diagnosis. In some analyses, we  
323 used continuous-valued topic weights to model disease subtypes. In analyses that require discrete  
324 subtypes, we assigned a discrete topic assignment to each disease diagnosis based on its  
325 maximum diagnosis-specific topic probability, and inferred the *comorbidity-derived subtype* of  
326 each disease diagnosis based on the discrete topic assignment; we note that discretizing  
327 continuous data loses information (see Discussion). We restricted our disease subtype analyses to  
328 52 diseases with at least 500 diagnoses assigned to each of two discrete subtypes; the average  
329 correlation between UK Biobank and All of Us disease subtypes (see above; same metric as  
330 Figure 5C) was 0.64 for the 41 (of 52) diseases that were shared between the two cohorts (Table  
331 2, Methods, Supplementary Figure 6, Supplementary Figure 20, and Supplementary Table 10).

332  
333 Age-dependent distributions of *comorbidity-derived* subtypes for four diseases (type 2 diabetes,  
334 asthma, hypercholesterolemia, and essential hypertension) are reported in Figure 6A and  
335 Supplementary Table 11; results for all 52 diseases are reported in Supplementary Figure 21 and  
336 Supplementary Table 11, and age-dependent distributions for the same four diseases in All of Us  
337 are reported in Supplementary Figure 22. The number of subtypes can be large, e.g. six subtypes  
338 for essential hypertension. Subtypes are often age-dependent, e.g. for the CVD and MGND  
339 subtypes of type 2 diabetes<sup>14,35</sup> (discussed above).

340  
341 ATM and the resulting subtype assignments do not make use of genetic data. However, we used  
342 genetic data to assess genetic heterogeneity across inferred subtypes of each disease. We used  
343 continuous-valued topic weights in this analysis. We first assessed whether PRS for overall  
344 disease risk varied with continuous-valued topic weights for each disease; PRS were computed  
345 using BOLT-LMM with five-fold cross validation<sup>50,51</sup> (see Methods and Code Availability).  
346 Results for four diseases (from Figure 6A) are reported in Figure 6B and Supplementary Table  
347 12; results for all 10 well-powered diseases (10 of 52 diseases with highest z-scores for nonzero  
348 SNP-heritability) are reported in Supplementary Figure 23 and Supplementary Table 12. We  
349 identified 18 disease-topic pairs (of 10\*10=100 disease-topic pairs analysed) for which PRS  
350 values in disease cases vary with patient topic weight. For example, for essential hypertension,  
351 hypercholesterolemia, and type 2 diabetes, patients assigned to the CVD subtype had  
352 significantly higher PRS values than patients assigned to other subtypes. For essential  
353 hypertension, patients assigned to the CER subtype had significantly higher PRS values; for type  
354 2 diabetes, patients assigned to the CER subtype had lower PRS values than the CVD subtype,

355 even though the majority of type 2 diabetes diagnoses are assigned to the CER subtype. We  
356 further verified that most of the variation in PRS values with disease subtype could not be  
357 explained by age<sup>52</sup> or differences in subtype sample size (Supplementary Figure 24). These  
358 associations between subtypes (defined using comorbidity data) and PRS (defined using genetic  
359 data) imply that disease subtypes identified through comorbidity are genetically heterogeneous,  
360 consistent with [phenomenological](#) differences in disease aetiology.

361  
362 We further investigated whether subtype assignments (defined using comorbidity data) revealed  
363 subtype-specific excess genetic correlations. [We used discrete subtypes in this analysis](#). We  
364 estimated excess genetic correlations between [disease-subtype and subtype-subtype](#) pairs  
365 (relative to genetic correlations between the underlying diseases). Excess [pairwise](#) genetic  
366 correlations for 15 [diseases and disease subtypes](#) (spanning 11 diseases and 3 topics: CER,  
367 MGND and CVD) are reported in Figure 7A and Supplementary Table 13 (relative to genetic  
368 correlations between the underlying diseases; Figure 7B), and excess [pairwise](#) genetic  
369 correlations for all 89 well-powered [diseases and disease subtypes](#) (89 of 378 [diseases and](#)  
370 [disease subtypes](#) with z-score > 4 for nonzero SNP-heritability; [378 = 348 diseases + 30 disease](#)  
371 [subtypes](#)) are reported in Supplementary Figure 25 and Supplementary Table 13. Genetic  
372 correlations between pairs of subtypes involving the same disease were significantly less than 1  
373 (FDR<0.1) for hypertension (CER vs. CVD:  $\rho = 0.86 \pm 0.04$ , P=0.0004; MGND vs. CVD:  $\rho =$   
374  $0.74 \pm 0.05$ , P= $3 \times 10^{-8}$ ) and type 2 diabetes (CER vs. MGND:  $\rho = 0.64 \pm 0.09$ , P= $8 \times 10^{-5}$ )  
375 (Figure 7A; Supplementary Table 13). In addition, we observed significant excess genetic  
376 correlations (FDR<0.1) for [8 disease-subtype and subtype-subtype pairs](#) involving different  
377 diseases (Figure 7A; Supplementary Table 13). [We sought to verify that genetic differences](#)  
378 [between subtypes were not due to partitions of the cohort that are unrelated to disease \(e.g. we](#)  
379 [expect a nonzero genetic correlation between tall vs. short type 2 diabetes cases, even if height is](#)  
380 [not genetically correlated to type 2 diabetes\)](#). Thus, we assessed whether the excess genetic  
381 [correlation could](#) be explained by non-disease-specific differences in the underlying topics  
382 (which are weakly heritable; Supplementary Table 3) by repeating the analysis using disease  
383 cases and controls with matched topic weights ([i.e. case and controls have matched topic weights](#)  
384 [distributions within each disease or disease subtypes](#)) (Methods). [We determined that the excess](#)  
385 [genetic correlations could not be explained by non-disease-specific differences](#) (Supplementary  
386 Figure 26). We also estimated subtype-specific SNP-heritability and identified some instances of  
387 differences between subtypes, albeit with limited power (Supplementary Table 14).

388  
389 Finally, we used the population genetic parameter  $F_{ST}$ <sup>53,54</sup> to quantify genome-wide differences  
390 in allele frequency between two subtypes of the same disease. [We used discrete subtypes in this](#)  
391 [analysis. We wished to avoid inferring genetic differences between subtypes that were due to](#)  
392 [partitions of the cohort that are unrelated to disease \(e.g. we expect a nonzero  \$F\_{ST}\$  between tall](#)  
393 [vs. short type 2 diabetes cases; see above\)](#). Thus, we assessed the statistical significance of  
394 [nonzero  \$F\_{ST}\$  estimates by comparing the observed  \$F\_{ST}\$  estimates \(between two subtypes of the](#)

395 same disease) to the expected  $F_{ST}$  based on matched topic weights (i.e.  $F_{ST}$  estimates between  
396 two sets of healthy controls with topic weight distributions matched to the respective disease  
397 subtypes) (excess  $F_{ST}$ ; Methods). We determined that 63 of 104 pairs of disease subtypes  
398 involving the same disease (spanning 29 of 49 diseases, excluding 3 diseases that did not have  
399 enough controls with matched topic weights) had significant excess  $F_{ST}$  estimates (FDR < 0.1)  
400 (Supplementary Figure 27, Supplementary Table 15). For example, the CVD, CER, and MGND  
401 subtypes of type 2 diabetes had significant excess  $F_{ST}$  estimates ( $F$ -statistic=0.0003,  $P$ =0.001  
402 based on 1,000 matched control sets). This provides further evidence that disease subtypes as  
403 determined by comorbidity have different molecular and physiological aetiologies. We conclude  
404 that disease subtypes defined by distinct topics are genetically heterogeneous.

405

#### 406 *Disease-associated SNPs have subtype-dependent effects*

407 We hypothesised that disease genes and pathways might differentially impact the disease  
408 subtypes identified by ATM. We investigated the genetic heterogeneity between disease  
409 subtypes at the level of individual disease-associated variants. We used continuous-valued topic  
410 weights in this analysis. We employed a statistical test that tests for SNP x topic interaction  
411 effects on disease phenotype in the presence of separate SNP and topic effects (Methods). We  
412 verified via simulations that this statistical test is well-calibrated under a broad range of scenarios  
413 with no true interaction, including direct effect of topic on disease, direct effect of disease on  
414 topic, pleiotropic SNP effects on disease and topic, and nonlinear effects (Supplementary Figure  
415 28). We also assessed the power to detect true interactions (Supplementary Figure 29). To limit  
416 the number of hypotheses tested, we applied this test to independent SNPs with genome-wide  
417 significant main effects on disease (Methods). We thus performed 2,530 statistical tests spanning  
418 888 disease-associated SNPs, 14 diseases, and 35 disease subtypes (Supplementary Table 16).  
419 We assessed statistical significance using global FDR<0.1 across the 2,530 statistical tests. We  
420 also computed main SNP effects specific to each quartile of topic weights across individuals and  
421 tested for different odds ratios in top vs. bottom quartiles, as an alternative way to represent SNP  
422 x topic interactions; the top/bottom quartile test is more intuitive, but less powerful in most  
423 cases.

424

425 We identified 43 SNP x topic interactions at FDR<0.1 (Figure 8, Supplementary Figure 30,  
426 Supplementary Table 17 and Supplementary Table 18). Here, we highlight a series of examples.  
427 First, the type 2 diabetes-associated SNP rs1042725 in the *HMG2* locus has a higher odds ratio  
428 in the top quartile of CVD topic weight (1.18±0.02) than in the bottom quartile (1.00±0.02)  
429 ( $P=3 \times 10^{-4}$  for interaction test (FDR = 0.04 < 0.1);  $P=3 \times 10^{-7}$  for top/bottom quartile test  
430 (FDR = 0.0002 < 0.1)). *HMG2* is associated with type 2 diabetes<sup>55</sup> and is reported to have  
431 functions in cardiac remodelling<sup>56</sup>, suggesting that shared pathways underlie the observed SNP x  
432 topic interaction. Second, the asthma-associated SNP rs1837253 in the *TSLP* locus has a higher  
433 odds ratio in the top quartile of SRD topic weight (1.17±0.02) than in the bottom quartile  
434 (1.05±0.02) ( $P=6 \times 10^{-6}$  for interaction test (FDR = 0.004 < 0.1);  $P=1 \times 10^{-3}$  for top/bottom

435 [quartile test \(FDR = 0.08 < 0.1\)](#)). *TSLP* plays an important role in promoting Th2 cellular  
436 responses and is considered a potential therapeutic target, which is consistent with assignment of  
437 asthma and atopic/contact dermatitis<sup>57</sup> to the SRD topic (Supplementary Table 4). Third, the  
438 hypertension-associated SNP rs3735533 within the *HOTTIP* long non-coding RNA has a lower  
439 odds ratio in the top quartile of CVD topic weight ( $1.07 \pm 0.02$ ) than in the bottom quartile  
440 ( $1.13 \pm 0.02$ ) ( $P = 0.0015$  for [interaction test \(FDR = 0.09 < 0.1\)](#);  $P=0.1$  for [top/bottom quartile](#)  
441 [test \(FDR = 0.55\)](#)). *HOTTIP* is associated with blood pressure<sup>27,58</sup> and conotruncal heart  
442 malformations<sup>59</sup>. Fourth, the hypothyroidism-associated SNP rs9404989 in the *HCG26* long non-  
443 coding RNA has a higher odds ratio in the top quartile of FGND topic weight ( $1.90 \pm 0.24$ ) than in  
444 the bottom quartile ( $1.19 \pm 0.13$ ) ( $P = 1 \times 10^{-4}$  for [interaction test \(FDR = 0.02 < 0.1\)](#);  
445  $P=3 \times 10^{-3}$  for [top/bottom quartile test \(FDR = 0.15\)](#)). Hypothyroidism associations have been  
446 reported in the HLA region<sup>27</sup>, but not to our knowledge in relation to the *HCG26*. To verify  
447 correct calibration, we performed control SNP x topic interaction tests using the same 888  
448 disease-associated SNPs together with random topics that did not correspond to disease subtypes,  
449 and confirmed that these control tests were well-calibrated ([Supplementary Figure 31B](#)). We  
450 conclude that genetic heterogeneity between disease subtypes can be detected at the level of  
451 individual disease-associated variants.  
452  
453

## 454 Discussion

455 We have introduced an age-dependent topic modelling (ATM) method to provide a low-rank  
456 representation of longitudinal disease records, leveraging age-dependent comorbidity profiles to  
457 identify and validate biological subtypes of disease. Our study builds on previous studies on  
458 topic modelling<sup>37,38,40,60</sup>, genetic subtype identification<sup>13-15</sup>, and low-rank modelling of multiple  
459 diseases to identify shared genetic components<sup>25-27</sup>. We highlight three specific contributions of  
460 our study. First, we incorporated age at diagnosis information into our low-rank representation,  
461 complementing the use of age information in other contexts<sup>32,52,61</sup>; we showed that age  
462 information is highly informative for our inferred comorbidity profiles in both simulated and  
463 empirical data, emphasising the importance of accounting for age in efforts to classify disease  
464 diagnoses. Second, we identified 52 diseases with heterogeneous comorbidity profiles that we  
465 used to define disease subtypes, many of which had not previously been identified  
466 (Supplementary Table 19); [comorbidity-derived disease subtypes were consistent between UK  
467 Biobank and All of Us, despite key differences between these cohorts](#). Third, we used genetic  
468 data (including PRS, genetic correlation and  $F_{ST}$  analyses) to validate these disease subtypes,  
469 confirming that the inferred subtypes reflect true differences in disease aetiology.

470  
471 We emphasise three downstream implications of our findings. First, it is of interest to perform  
472 disease subtype-specific GWAS on the disease subtypes that we have identified here, analogous  
473 to GWAS of previously identified disease subtypes<sup>13-15</sup>. Second, our findings motivate efforts to  
474 understand the functional biology underlying the disease subtypes that we identified; the recent  
475 availability of functional data that is linked to EHR is likely to aid this endeavor<sup>29,62</sup>. Third, [the  
476 efficient inference of ATM permits identifying](#) age-dependent comorbidity profiles and disease  
477 subtypes in [much larger](#) EHR data sets<sup>63</sup>; [though we acknowledge that](#) establishing  
478 [comprehensive](#) representations of disease topics that are transferable and robust across different  
479 healthcare systems and data sources represents a major future challenge.

480  
481 Our findings reflect a growing understanding of the importance of context, such as age, sex,  
482 socioeconomic status and previous medical history, in genetic risk<sup>52,64,65</sup>. To maximise power  
483 and ensure accurate calibration, context information needs to be integrated into clinical risk  
484 prediction tools that combine genetic information (such as polygenic risk scores<sup>1,66</sup>) and non-  
485 genetic risk factors. Our work focuses on age, but motivates further investigation of other  
486 contexts. We note that aspects of context are themselves influenced by genetic risk factors, hence  
487 there is an open and important challenge in determining how best to combine medical history  
488 and/or causal biomarker measurements with genetic risk to predict future events<sup>67</sup>.

489  
490 We note several limitations of our work. First, age at diagnosis information in EHR data may be  
491 an imperfect proxy for true age at onset, particularly for less severe diseases that may be detected  
492 as secondary diagnoses; although perfectly accurate age at onset information would be ideal, our  
493 study shows that that imperfect age at diagnosis information is sufficient to draw meaningful

494 conclusions. Second, raw EHR data may be inaccurate and/or difficult to parse<sup>1</sup>; again, although  
495 perfectly accurate EHR data would be ideal, our study shows that imperfect EHR data is  
496 sufficient to draw meaningful conclusions. Third, our ATM approach incurs substantial  
497 computational cost (Supplementary Table 20); however, analyses of biobank-scale data sets are  
498 computationally tractable, with our main analysis requiring only 4.7 hours of running time.  
499 Fourth, the genetic correlation and  $F_{ST}$  analyses were based on discrete subtypes, but discretizing  
500 continuous data loses information and may compromise power. However, definitions of disease  
501 often discretize continuous variables<sup>68</sup>. In addition, our PRS analysis (Figure 6) and SNP x topic  
502 interaction analysis (Figure 8) leveraged continuous-valued topic weights. Fifth, interpretability  
503 can be a potential downside of data reduction approaches. The interpretation of a particular  
504 disease topic is that it consists of diseases that tend to co-occur with a specified set of diseases as  
505 a function of age. Identifying the functional biology underlying these co-occurrences remains a  
506 direction for future research, but there is immediate utility in performing disease subtype-specific  
507 GWAS and downstream analyses using the subtypes that we have identified. Despite these  
508 limitations, ATM is a powerful approach for identifying age-dependent comorbidity profiles and  
509 disease subtypes.

510

## 511 Acknowledgements

512

513 This research has been conducted using the UK Biobank Resource; application number 12788.

514

515 Funded by Wellcome (BST00080- H503.01 to XJ, 100956/Z/13/Z to GM, [https://](https://wellcome.org)  
516 [wellcome.org](https://wellcome.org)); the Li Ka Shing Foundation (to GM, <https://www.lksf.org>); NIH grants R01  
517 HG006399, R01 MH101244, and R37 MH107649 (to ALP); The Alan Turing Institute  
518 (<https://www.turing.ac.uk>), Health Data Research UK (<https://www.hdruk.ac.uk>), the Medical  
519 Research Council UK (<https://mrc.ukri.org>), the Engineering and Physical Sciences Research  
520 Council (EPSRC <https://epsrc.ukri.org>) through the Bayes4Health programme Grant  
521 EP/R018561/1, and AI for Science and Government UK Research and Innovation (UKRI,  
522 <https://www.turing.ac.uk/research/asp>) (to CH); BHF Chair award CH/12/2/29428 (to XJ). This  
523 work was supported by core funding from the: British Heart Foundation (RG/13/13/30194;  
524 RG/18/13/33946), BHF Cambridge Centre of Research Excellence (RE/13/6/30180) and NIHR  
525 Cambridge Biomedical Research Centre (BRC-1215-20014). The funders had no role in study  
526 design, data collection and analysis, decision to publish, or preparation of the manuscript.

527

528 This work uses data provided by patients and collected by the NHS as part of their care and  
529 Support. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a  
530 joint development between the Wellcome Centre for Human Genetics and the Big Data Institute  
531 supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. The  
532 views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the  
533 Department of Health. We thank Kushal Dey, Luke Kelly and Yunlong Jiao for the discussion.

534

## 535 Data availability

536 UK Biobank data is publicly available at <https://www.ukbiobank.ac.uk/>.

537

## 538 Code availability

539 Open-source software implementing the ATM method is available at [https://github.com/Xilin-](https://github.com/Xilin-Jiang/ATM)

540 [Jiang/ATM](https://github.com/Xilin-Jiang/ATM). BOLT-LMM is available at <https://alkesgroup.broadinstitute.org/BOLT-LMM/>.

541 Heritability and genetic correlation analysis were performed using LDSC, which is available at

542 <https://github.com/bulik/ldsc>. PLINK v1.9, which was used for  $F_{ST}$  and association tests, is

543 available at <https://www.cog-genomics.org/plink/>.

544

545

## 546 Methods

### 547 Age-dependent topic model (ATM)

548 Our Age-dependent topic model (ATM) is a Bayesian hierarchical model to infer latent risk  
549 profiles for common diseases. The model assumes that each individual possesses several age-  
550 evolving disease profiles (topic loadings), which summarise the risk over age for multiple  
551 diseases that tend to co-occur within an individual's lifetime, namely the age specific multi-  
552 morbidity profiles. At each disease diagnosis, one of the disease profiles is first chosen based on  
553 individual weights of profile composition (topic weights), the disease is then sampled from this  
554 profile conditional on the age of the incidence.

555

556 We constructed a Bayesian hierarchical model to infer  $K$  latent risk profiles for  $D$  distinct  
557 common diseases. Each latent risk profile (comorbidity topics) is age-evolving and contains risk  
558 trajectories for all  $D$  diseases considered. Each individual might have a different number of  
559 diseases, while the disease risk is determined by the weighted combination of latent risk topics.

560 The indices in this note are as follows:

- 561 •  $s = 1, \dots, M$ ;
- 562 •  $n = 1, \dots, N_s$ ;
- 563 •  $i = 1, \dots, K$ ;
- 564 •  $j = 1, \dots, D$ ;

565 where  $M$  is the number of subjects,  $N_s$  is the number of records within  $s^{th}$  subject,  $K$  is the  
566 number of topics, and  $D$  is the total number of diseases we are interested in. The plate notation of  
567 the generative model is summarised in Supplementary Figure 1:

- 568 •  $\theta \in R^{M \times K}$  is the topic weight for all individuals (referred to as patient topic weights),  
569 each row of which ( $\in R^K$ ) is assumed to be sampled from a Dirichlet distribution with  
570 parameter  $\alpha$ .  $\alpha$  is set as a hyper parameter:  $\theta_s \sim Dir(\alpha)$ . **We used topic weights to**  
571 **assign continuous values for disease subtypes in PRS and SNP x Topic analyses.**
- 572 •  $z \in \{1, 2, \dots, K\}^{\sum_s N_s}$  (referred to as diagnosis-specific topic probability) is the topic  
573 assignment for each diagnosis  $w \in \{1, 2, \dots, D\}^{\sum_s N_s}$ . Note the total number of  
574 diagnoses across all patients are  $\sum_s N_s$ . The topic assignment for each diagnosis is  
575 generated from a categorical distribution with parameters equal to  $s^{th}$  individual topic  
576 weight:  $z_{sn} \sim Multi(\theta_s)$ . **We used diagnosis-specific topic probability to define**  
577 **discrete disease subtypes in excess genetic correlation and excess  $F_{ST}$  analyses.**
- 578 •  $\beta(t) \in F(t)^{K \times D}$  is the topic loading which is  $K \times D$  functions of age  $t$ .  $F(t)$  is the  
579 class of functions of  $t$ . At each plausible  $t$ , the following is satisfied:  $\sum_j \beta_{ij}(t) = 1$ . **In**  
580 **practice we ensure above is true and add smoothness by constrain  $F(t)$  to be a softmax of**  
581 **spline or polynomial functions:  $\beta_{ij}(t) = \frac{\exp(p_{ij}^T \phi(t))}{\sum_{j=1}^K \exp(p_{ij}^T \phi(t))}$ , where  $p_{ij}^T \phi(t)$  is**  
582 **polynomial and spline functions of  $t$ ;  $p_{ij} = \{p_{ijd}\}$ ;  $d = 1, 2, \dots, P$ ;  $P$  is the degree of**  
583 **freedom that controls the smoothness;  $\phi(t)$  is polynomial and spline basis for age  $t$ .**



- 584 •  $w \in \{1, 2, \dots, D\}^{\sum_s N_s}$  are observed diagnoses. The  $n^{th}$  diagnosis of  $s^{th}$  individual  $w_{sn}$  is  
585 sampled from the topic  $\beta_{z_{sn}}(t)$  chosen by  $z_{sn}$ :  $w_{sn} \sim Multi(\beta_{z_{sn}}(t_{sn}))$ , here  $t_{sn}$  is the  
586 age of the observed age at diagnosis of the observed diagnosis  $w_{sn}$ .

587

588 The values of interest in this model are global topic parameter  $\beta$ , individual (patient) level topic  
589 weight  $\theta$ , and diagnosis-specific topic probability  $z$ . Based on the generative process above, we  
590 notice that each patient is independent conditional on  $\alpha$ . Therefore, the inference of  $\theta$  and  
591  $z$  (discussed below) is performed by looping each individual in turn.

592

593 The key element in our model is age-evolving risk profiles, which is achieved by model the  
594 comorbidity trajectories  $\beta(t) \in F(t)^{K \times D}$  as functions of age. The functionals  $F(t)$  are  
595 parameterized as linear, quadratic, cubic polynomials, and cubic splines with one, two and three  
596 knots. We use prediction odds ratio to decide the optimal model structure including the function  
597 forms and the number of topics; we use ELBO to choose the optimal inference results (with  
598 random parameter initialization) for the same model structure (Supplementary Table 1).

599

## 600 Inference of ATM

601 The variables of interest are global topic parameter  $\beta(t)$ , individual (patient) level topic weight  
602  $\theta$ , and diagnosis-specific topic probability  $z$  of each diagnosis. We could adopt an EM strategy,  
603 where in the E-step we first estimate posterior distribution of  $\theta$  and  $z$ , then in the M-step we  
604 estimate  $\beta$  which maximises the evidence lower bound (ELBO).

605

606 The details of the inference is explained in Supplementary Note. In summary, in a Bayesian  
607 setting, We used the evidence function  $p(w|\alpha, \beta)$  to evaluate how well the model fits the data.  
608 The best  $\beta(t)$  is found by maximise the evidence function, while for  $\theta$  and  $z$  we aim to find or  
609 approximate their posterior distribution  $p(z, \theta | w, \alpha, \beta)$ . Given that the posterior distribution is  
610 intractable, we use variational distribution  $q(z, \theta)$  to approximate them. Now we could write the  
611 evidence function as:

$$612 \quad p(w | \alpha, \beta) = L(z, \theta, \beta, \alpha) + KL(q||p),$$

613 here  $KL(q||p) = - \int_{z, \theta} q(z, \theta) \ln \frac{p(z, \theta | w, \alpha, \beta)}{q(z, \theta)}$  is the KL divergence. Since KL divergence is  
614 always positive,  $L(z, \theta, \beta, \alpha)$  is a lower bound of the evidence function:

$$615 \quad L(z, \theta, \beta, \alpha) = E_q \{ \ln p(w, z, \theta | \alpha, \beta) - \ln q(z, \theta) \}.$$

616

617 When finding the posterior of  $\theta$  and  $z$ , we want  $\ln q(z, \theta)$  to be as close to the  
618 posterior  $p(z, \theta | w, \alpha, \beta)$  as possible. Since  $KL(q||p) = 0$  when  $q(z, \theta) = p(z, \theta | w, \alpha, \beta)$ ,  
619 this is achieved by minimising  $KL(q||p)$  or maximise  $L(z, \theta, \beta, \alpha)$ . The most commonly used  
620 form of  $q(z, \theta)$  assumes the distribution is factorised, which might cause instability when  
621 signal-to-noise ratio is low<sup>69</sup>. Therefore, more accurate inference methods such as collapsed  
622 variational inference is considered<sup>40</sup>. Comparison of the evidence lower bound  $L(z, \theta, \beta, \alpha)$

623 shows collapsed variational inference (CVB) is consistently more accurate than mean-field  
624 variational inference (VB) (Supplementary Figure 10). Therefore we chose the collapsed  
625 variational inference<sup>40</sup>. The collapsed variational inference is achieved by integrate out  $\theta$  from  
626 the likelihood function  $p(w, z, \theta | \alpha, \beta)$  and find the approximated posterior distribution  $q(z)$ .  
627 For detailed derivation, the comparison between collapsed variational inference and mean-field  
628 variational inference, and update algorithms, see Supplementary Note.

629  
630 When finding the  $\beta(t)$  that maximises the evidence function, we again maximise  $L(z, \theta, \beta, \alpha)$ .  
631 Maximising  $L(z, \theta, \beta, \alpha)$  with respect to  $\beta(t)$  does not have an analytical solution due to its  
632 softmax structure. We use local variational methods and numeric optimisation to find the  
633 distribution of  $\beta(t)$ . In summary,  $L(z, \theta, \beta, \alpha)$  is not tractable with respect to  $\beta(t)$  as it contains  
634 a log of softmax function (Section 3.2 of Supplementary Note). We introduced a local variational  
635 variable to obtain a tractable lower bound of  $L(z, \theta, \beta, \alpha)$  (equation 11 in Supplementary Note)  
636 and use gradient descent to approximate the lower bound. Details are provided in Supplementary  
637 Note.

638  
639 We extract topic weights at patient-level and diagnosis-level from the posterior distribution  
640 inferred from the data. Our model has the desired property that each patient and patient-diagnosis  
641 are assigned to comorbidity topics. The model estimates the posterior distribution  $q(z)$ , which is  
642 a categorical distribution (equation 8 of Supplementary Note). We listed following definitions in  
643 this paper that are derived from the  $q(z)$ :

- 644
- 645 • Each patient-diagnosis (incident disease) has a *diagnosis-specific topic probability*, which  
646 is computed as  $E_q\{z_n\}$ .
  - 647 • Each patient has a posterior *topic weights*  $\theta_s$ , which is a dirichlet distribution  $\theta_s \sim$   
648  $Dir(\alpha + \sum_{n=1}^{N_s} E_q\{z_n\})$ . The *topic weights* of each patient is defined as the mode of  
649 this Dirichlet distribution  $\frac{\sum_{n=1}^{N_s} E_q\{z_n\}}{\sum_{i=1}^K \sum_{n=1}^{N_s} E_q\{z_{ni}\}}$  (we used  $\alpha = 1$ , which puts an noninformative  
650 prior on the topic weights). *Topic weight is the low-rank representation of disease*  
651 *history*, for analyses including PRS association with comorbidity topics and SNP x Topic  
652 interaction analysis.
  - 653 • The *average topic assignments* of disease  $j$  is the mean over all  
654 incidences  $\overline{E_q\{z_{sn} \in \{w_{sn}=j\}\}}$ . This metric is used to measure which comorbidity topic a  
655 disease is associated with (Figure 4B), and it is equivalent to a weighted average of topic  
656 loadings (Supplementary Note equation 5 shows the link between diagnosis-specific topic  
657 probability and topic loading). A disease assigned to multiple topics is considered to have  
658 comorbidity subtypes.
  - 659 • A hard assignment of a patient-diagnosis to a *comorbidity-derived subtype* is based on the  
660 max value of the vector  $E_q\{z_n\}$ . The incident disease is assigned to topic  
661  $\text{argmax}_i (E_q\{z_{ni}\})$ .

662

### 663 **Metrics for evaluating ATM**

664 ATM is evaluated for different purposes, which requires different metrics (Supplementary Table  
665 1). Here we list the details of the four metrics considered: *Prediction odds ratio*, *Evidence Lower*  
666 *Bound (ELBO)*, *the Area under the Precision-Recall curve (AUPRC)*<sup>70</sup>, and *Co-occurrence odds*  
667 *ratio*.

668

669 *Prediction odds ratio*: To compare models of different topic numbers and configuration of age  
670 profiles, we compare the prediction odds ratio of each model. Briefly, prediction odds ratio is  
671 defined on 20% held-out test data as the odds that the true diseases are within the top 1%  
672 diseases predicted by ATM (trained on 80% of the training set and uses earlier diagnoses as  
673 input), divided by the odds that the true diseases are within the top 1% of diseases ranked by  
674 prevalence.

675

676 Specifically, we separate UK Biobank patients into a training set (80%) and a testing set (20%).  
677 On the training set, we estimate the comorbidity topic loadings. On the testing set, we fix the  
678 topic loadings and infer the patient topic weights to predict the next disease in chronological  
679 order. The topic loadings are estimated using the  $n$  diseases and compute the risk rank of  
680 diseases at the age of the  $n+1$  disease. The odds ratio is computed by the odds of the  $n+1$  disease  
681 being in the top 1% of diseases versus being in the top 1% most prevalent diseases. We use the  
682 top 1% most prevalent diseases instead of randomly chosen diseases as it represents a naive  
683 prediction model that predicts disease based on prevalence. The patient topic weights  
684 computation is in section Inference of ATM and the risk is computed as the linear combination  
685 of topics using topic weights as coefficients. We also compute the prediction odds ratio using the  
686 LDA model. We repeat the procedure for 10 times for each model configuration.

687

688 We compared the prediction odds ratio for topic number between 5 to 20, with linear, quadratic  
689 polynomial, cubic polynomial, and splines with one, two and three knots. We also compare the  
690 ATM model with the LDA model of topic number between 5 to 20.

691

692 *Evidence Lower Bound (ELBO)*: ELBO evaluated the accuracy of the variational inference  
693 method on a specific data set<sup>39</sup>. The mathematical expression of ELBO for ATM is presented in  
694 equation 9 in the Supplementary Note. To find the best model that fits the entire dataset, we  
695 evaluate the [ELBO for models with 19 choices of the number of topics: 5-20, 25, 30, and 50; 6](#)  
696 [choices of age profiles configuration: linear, quadratic polynomial, cubic polynomial, and splines](#)  
697 [with one, two and three knots](#). Each model is run for 10 times with random initialisations. We  
698 choose the model that has the highest ELBO after converging.

699

700 *AURPC*: To evaluate whether a model could capture the comorbidity subtypes in simulation  
701 analysis, we compute the precision, recall, and area under precision-recall curve (AUPRC) to

702 correctly classify disease diagnosis to be from the topic that it is generated from. The topic of  
703 each diagnosis is determined by diagnosis-specific topic probability. Note we could only  
704 evaluate AUPRC in simulations where the truth is known.

705

706 *Co-occurrence odds ratio:* To verify that the comorbidity profiles that the model captured are  
707 capturing diseases that are more likely to present within the same individual, we estimate the  
708 odds ratio of the disease duo, trio, quartet, and quintet that are captured by the topic versus that  
709 of random combinations. We divide the population into an 80% training set and a 20% testing  
710 set. We trained the ATM model with five random initialisations and kept the inference with the  
711 highest ELBO. Each disease is assigned to a topic by the highest average topic assignments.  
712 (section Inference of ATM) We focus on the top 100 diseases ranked by prevalence to avoid the  
713 combination being too rare to appear in the population. In the testing set, we computed the odds  
714 of individuals who have all diseases in the comorbidities versus the odds implied if all diseases  
715 are independent (computed as the product of disease prevalence). The odds ratio is computed for  
716 all combinations of duo, trio, quartet, and quintet that are assigned to the same topics. We  
717 perform the same analysis using PCA for comparison.

718

719

## 720 **Simulations of ATM method.**

721 To test whether the algorithm could assign disease diagnosis to correct comorbidity profiles, we  
722 simulated disease from two disease topics within a population of 10,000, using following  
723 parameters:

724 •  $M = 10,000$ ;

725 •  $\overline{N_S} = 6.1$ ;

726 •  $N_S \sim \exp\{\overline{N_S}\}$ ;

727 •  $D = 20$ ;

728 •  $K = 2$ ;

729 Here  $M$  is the number of individuals in the population,  $\overline{N_S}$  is the average number of diseases for  
730 each individual,  $D$  is the total number of diseases,  $K$  is the number of comorbidity topics. The  
731 distribution of disease number per-individual  $N_S$  is sampled from an exponential distribution,  
732 which matches those from UK Biobank data (Supplementary Figure 32). According to equation  
733 3.1 in Ghorbani et al.<sup>69</sup>, whether the topic model could capture the true latent structure is  
734 determined by the information signal-to-noise ratio and could be evaluated with limits  $M \rightarrow$   
735  $\infty$ ;  $D \rightarrow \infty$ ;  $\frac{D}{M} \rightarrow \delta$ , where  $\delta$  is a constant. Therefore we choose  $D$  and  $M$  at scales that make  $\frac{D}{M}$   
736 approximately similar to those of the UK Biobank dataset (Samples size = 282,957; distinct  
737 disease number = 349).

738

739 The simulated topics loadings are constructed as follows:

740 • All but  $K$  diseases are simulated to be associated with comorbidity profiles. Each of them  
741 has a risk period of 30 years and overlaps for 10 years with the next disease. For

742 example, if disease 1 has a risk period from 30 to 59 years of age, disease 2 will have a  
743 risk period between 50 to 79 years of age. When the risk period reaches the maximal age,  
744 the truncated part will be carried to the next disease to create diseases with shorter risk  
745 period. All risk periods are assigned a value 1. [The overlapping structure of topic](#)  
746 [loadings is chosen so that average standard deviation in age-at-diagnosis \(8.5 years\) and](#)  
747 [the age window under consideration \(30-80 years of age\) matches UK Biobank data.](#)

- 748 ●  $K$  diseases that are not associated with comorbidity are simulated to span all topics. The  
749 values of these diseases are sampled from  $Unif(0, \frac{0,1}{K})$  for each topic. Here  $K$  is the  
750 number of topics.
- 751 ● The age profiles are then normalised at each age point to ensure  $\sum_{j=1}^D \beta_j(t) = 1$  for all  
752  $t$ . With this constraint we could sample a disease at each age  $t$  using a multinomial  
753 probability with the topic loading as the parameter. The age range of the simulated topics  
754 is 30 to 81 years of age, which is the minimal and maximal age at diagnosis of incident  
755 disease in the UK Biobank population. An example of a simulated topic is shown in  
756 Supplementary Figure 33.

757  
758 For each individual, we sampled the Dirichlet parameter  $\alpha$  from a gamma distribution (shape =  
759 50, rate = 50). Topic loadings are sampled from the Dirichlet distribution for each patient as the  
760 generative process. For each patient, we first sample the number of diseases  $N_S$ . For each  
761 incident disease, we sample the disease age from uniform distribution between age 30 to 81 and  
762 a topic from the topic loading. We then choose the incident disease based on the age at diagnosis  
763 from the chosen topic. The procedure follows the generative process described in Supplementary  
764 Note.

765  
766 Since in real data we only use the first age at diagnosis for diseases that are recorded repeatedly  
767 within the same patient, we filter the simulated diseases accordingly. The filtered data are fed  
768 into the inference functions to infer the latent topics and disease assignments. The inferred topics  
769 resemble the true topics used to simulate diseases as shown in Supplementary Figure 33. For the  
770 initialisation of each inference, we first sample  $\beta$  and  $\theta$  from the Dirichlet distribution of non-  
771 informative hyperparameters, then initialise other variables parameters following the generative  
772 process. The variational inference converged where the relative increase of ELBO is below  $10^{-6}$ .

773  
774 To simulate disease having distinct comorbidity subtypes, we first simulate diseases using the  
775 procedure described above. We consider two scenarios: (1) the subtype of diseases have the  
776 same age at diagnosis distribution. (2) the subtypes of disease have distinct age at diagnosis  
777 distribution.

778  
779 We create diseases with distinct comorbidity profiles by combining diseases that are sampled  
780 from distinct topics and labelling them as a single disease. We first chose one disease (**disease**  
781 **A**) then sampled a proportion of a second disease (**disease B**) to label as **disease A**. The

782 proportion is varied to create a different sample size ratio of the two subtypes. In scenario one,  
783 **disease B** is a disease that has the exact same age distribution as **disease A** but from the other  
784 topic. In scenario two, **disease B** is from the other topic and has a different age distribution (age  
785 at diagnosis moves up for 20 years, 10 years, or 5 years, respectively) than **disease A**. After  
786 changing the labels of **disease B** to be the same as **disease A**, we used the inference procedure  
787 described as above to get the posterior distribution.

788

789 To evaluate whether a model could capture the comorbidity subtypes, we compute the precision,  
790 recall, and area under precision-recall curve (AUPRC) to correctly classify incident **disease B** to  
791 be from the topic that it is generated from. The topic of each diagnosis is determined by  
792 diagnosis-specific topic probability. We use other diseases from the topic of **disease B** to  
793 benchmark the topic label. Topic modelling on the simulated data is performed with both ATM  
794 and LDA (both implemented using collapsed variational inference for fair comparison) to  
795 compare the performances.

796

797 We evaluate the subtype classification with varying values for three simulation parameters:

- 798 ● ratio of sample sizes between the two subtypes. We change the ratio of the two subtypes  
799 by a grid between 0 to 0.9 with a step size 0.1. The default value of sample size ratio is  
800 set as 0.1 in other simulations to test for other parameters that have impacts on the  
801 precision and recall.
- 802 ● Simulated population size. We simulated population sizes equal to 200, 500, 1000, 2000,  
803 5000, and 10,000. The default population size is 10,000 in other simulations.
- 804 ● Number of distinct diseases. We simulated datasets with 20, 30, 40, and 50 distinct  
805 diseases, with 2, 3, 4 and 5 underlying disease topics respectively. The default number of  
806 distinct diseases is 20 in other simulations.
- 807 ● Difference of age distribution. We considered three scenarios of subtype age distribution,  
808 with 0, 10, and 20 years of difference in the average age at diagnosis.

809

810

### 811 **UK Biobank comorbidity data.**

812 We analysed comorbidity data from 282,957 UK Biobank samples with diagnoses for at least  
813 two of the 348 focal diseases that we studied (see below). We use the hospital episode statistics  
814 (HES) data within the UK Biobank dataset, which records diseases using the ICD-10/ICD-10CM  
815 coding system; [the average record span of HES data is 28.6 years](#). Codes started with letters from  
816 A to N are kept as they correspond to disease code (opposed to procedure codes). The disease  
817 records were mapped from ICD-10/ICD-10CM codes to PheCodes using a three-step procedure:  
818 [First, we mapped the first four letters of each ICD-10 records to the phecodes, using the map file](#)  
819 [downloaded from phewascatalog.org; second, we mapped the remaining records using ICD-](#)  
820 [10CM map file downloaded from phewascatalog.org; last, we mapped remaining records using](#)  
821 [the same ICD-10CM map system but only use the first four character of each ICD-10CM codes.](#)

822 We also noticed (ICD-10/ICD-10CM)-Phecode pairs are not always one-to-one; when a single  
823 ICD-10/ICD-10CM code is mapped to more than one PheCodes, we chose the Phecode with the  
824 largest number of links to ICD-10/ICD-10CM codes to reduce redundancy of the mapping result.  
825 Using the procedure above, we mapped 99.7% ICD-10/ICD-10CM code to PheCodes, with  
826 4,637,127 records in total.

827  
828 The mapped Phecodes are filtered to keep only the first age at diagnosis for the same diseases  
829 within a patient. The age at diagnosis for each record is computed as the difference between  
830 month of birth to the episode starting date. We then computed the occurrence of each disease in  
831 the UK Biobank and kept 348 that have more than 1,000 occurrences (Supplementary Table 4).  
832 Starting with all 488,377 UK Biobank patients (including both European and non-European  
833 ancestries), we filtered the patients to keep only those who have at least two distinct diseases  
834 from the 348 focal diseases, as we are most interested in the comorbidity information. We treated  
835 the death as an additional disease (8,666 records) to evaluate if certain comorbidities are more  
836 likely to lead to fatal events. After these procedures, there are in total 1,726,144 distinct records  
837 across 282,957 patients.

838  
839 To name the topics inferred from the UK Biobank, we take the sum of *average topic assignments*  
840 (section Inference of ATM) over diseases for each Phecode system and extract the three most  
841 common Phecode disease systems. Six topics are named using the three most common Phecode  
842 disease systems: NRI “neoplasms, respiratory, infectious diseases”, CER “cardiovascular,  
843 endocrine/metabolic, respiratory”, SRD “sense organs, respiratory, dermatologic”, FGND  
844 “female genitourinary, neoplasms, digestive”, MGND “male genitourinary, digestive,  
845 neoplasms”, MDS “musculoskeletal, digestive, symptoms”, and ARP. For four topics that are  
846 predominantly associated with one system, we name them based on their top associated Phecode  
847 system: LGI “lower gastrointestinal”, UGI “upper gastrointestinal”, CVD “cardiovascular”, and  
848 ARP “arthropathy”.

849  
850 We present focal diseases for each topic in two ways. Firstly, we filter each topic using the  
851 profile mean value between age 30 to 81 to keep the top seven diseases. We chose seven for  
852 visualisation, as we found more diseases would be harder to read on a plot. Secondly, we also  
853 show seven diseases that have the highest average assignment to each topic. This will give a  
854 picture of diseases that are not the most prevalent in the population but are predominantly  
855 associated with the target topic.

856  
857 To compare the comorbidity heterogeneity between age groups, we group the incidences for each  
858 disease to two age groups: young group (<60 years of age) and old group (≥60 years of age). We  
859 compute the average topic assignment of each group as described in section Inference of ATM.  
860 Additionally, we inferred topics for male (984,554 records in 156,366 individuals) and female  
861 (741,590 records in 126,591 individuals) populations respectively using a model with 10 topics

862 and spline function with one knot. We extract the average topic assignment for each disease, and  
863 use Pearson's correlation to match the topics for both sexes to the topics inferred on the entire  
864 population.

865  
866 Each diagnosis is assigned to a specific topic using max diagnosis-specific topic probability. We  
867 focus our disease heterogeneity analysis on 52 diseases that have at least 500 incidences assigned  
868 to a secondary topic.

869

### 870 **All of Us comorbidity data.**

871 We analysed EHR data collected in the EHR domain of All of Us samples, which includes both  
872 primary care and secondary care data. The average distance between first and last diagnoses is  
873 7.9 years (vs. 7.0 years in UK Biobank); the average record span period is unknown, but we  
874 hypothesized that it is likely to be considerably larger than 7.9 years (vs. 28.6 years in UK  
875 Biobank). Disease codes in the All of Us EHR domain are coded in SNOMED CT. We first  
876 mapped All of Us disease codes from SNOMED CT to ICD-10CM code using map version  
877 20220901 downloaded from  
878 [https://www.nlm.nih.gov/research/umls/mapping\\_projects/snomedct\\_to\\_icd10cm.html](https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html). When a  
879 single SNOMED CT code was mapped to multiple ICD-10CM codes, we choose the code with  
880 the highest UK Biobank prevalence from these ICD-10CM codes. We then mapped ICD-10CM  
881 codes to Phecodes, using the same procedure described in the section above. We kept 233  
882 Phecodes that overlap with the 348 diseases analysed in the UK Biobank. We kept the first  
883 diagnosis for recurrent diseases in each patient. After mapping, we are left with 3,098,771  
884 diagnoses spanning 211,908 All of Us samples. We run ATM with topic number from 5 to 20  
885 and spline with two knots (degree of freedom = 5) on the All of Us comorbidity data and  
886 computed prediction odds ratio (using five-fold cross validation) and ELBO (on all 211,908  
887 samples).

888

### 889 **Comparing disease topics between UK Biobank and All of Us.**

890 We compared the optimal models from UK Biobank (10 topics, degree of freedom = 5) and All  
891 of Us (13 topics, degree of freedom = 5). We constrained our analyses on 233 of the 348 diseases  
892 that are shared between the two data sets. We performed three analyses to compare the  
893 comorbidity patterns from the two data sets.

894

895 First, we computed the correlation of topic loadings from two data sets. Since the topic loadings  
896 are functions of age, we computed their correlations using four different ways to summarise age  
897 information: topic loadings averaged across age; topic loadings at age 50, 60, and 70. For each  
898 UK Biobank topic, we found its most similar All of Us topic that has max correlation of topic  
899 loadings (averaged across age).

900



901 Second, we computed the cross-population prediction odds ratio, using the All of Us topics to  
 902 predict on UK Biobank comorbidity data. We divided the UK Biobank samples into 10 jackknife  
 903 blocks and computed prediction odds ratios on each leave-one-out sample.

904

905 Third, we compared the correlation of comorbidity profiles (measured by average topic  
 906 assignments; see Methods for definition) for 233 diseases that are shared between the two  
 907 populations. We define *correlations between topic assignments* as the correlation between UK  
 908 Biobank average topic assignments and All of Us average topic assignments after mapped to UK  
 909 Biobank topic space (see below).

910

911 Comparing disease topics inferred from different data sets is challenging due to the  
 912 exchangeability of topics (i.e. distinct topic configurations have the same likelihood for a given  
 913 data set). To compute *correlations between topic assignments* from ATM inference on different  
 914 populations, we first mapped the topics to the same topic space. Suppose there are two topic  
 915 spaces  $\{T^1\}$  and  $\{T^2\}$ . We create a map  $s(\cdot)$  from  $\{T^1\}$  and  $\{T^2\}$  by computing the normalised  $R^2$   
 916 between topic loadings:

917 
$$s(\{T^1\}, \{T^2\})_{i1,i2} = \frac{\text{cor}(T_{i1}^1, T_{i2}^2)_+^2}{\sum_{k2}^{K_2} \text{cor}(T_{i1}^1, T_{k2}^2)_+^2},$$

918 here  $T_{i1}^1$  is the  $i1^{th}$  topic loading from  $\{T^1\}$ ,  $T_{i2}^2$  is the  $i2^{th}$  topic loading from  $\{T^2\}$ ;  $K_1, K_2$  is the  
 919 number of topics in  $\{T^1\}, \{T^2\}$ ;  $\text{cor}(\cdot)_+$  is the positive part of the correlation, where we set the  
 920 negative correlations to zero as negative correlations between topic loadings are uninformative  
 921 consequences of the multinomial distribution in the model. Intuitively,  $s(\{T^1\}, \{T^2\})_{i1,i2}$  maps  
 922 each  $T_{i1}^1$  to  $\{T^2\}$  based on the proportion of  $T_{i1}^1$  variance explained by the  $K_2$  topics in  $\{T^2\}$ .  
 923  $s(\{T^1\}, \{T^2\}) \in R^{K_1 \times K_2}$ .

924

925 Suppose we have the comorbidity profile of disease 1 and disease 2, which lies in  $\{T^1\}$  and  $\{T^2\}$   
 926 respectively. We could map disease 1 diagnosis-specific topic probability  $z_{sn,1} \in R^{K_1}$  (or  
 927 average topic assignments of disease 1; see these definitions in Methods) to topic space 2:  
 928  $z'_{sn,1} = s(\{T^1\}, \{T^2\})^T z_{sn,1}$ ,  $z'_{sn,1} \in R^{K_2}$ . The correlations between topic assignments in topic  
 929 space 2 between disease 1 and disease 2 is the correlation  $\text{cor}(\overline{z'_{sn,1}}, \overline{z_{sn,2}})$ ; here  $\overline{z'_{sn,1}}$  is the  
 930 average topic assignments for disease 1 after mapped to topic space 2, which is the average of  
 931  $z'_{sn,1}$  across all diagnoses;  $\overline{z_{sn,2}}$  is the average topic assignments for disease 2. For correlations  
 932 between topic assignments within the same topic,  $s(\{T^1\}, \{T^1\})$  is an identical matrix.

933

934 **UK Biobank genotype data.**

935 For genetic correlation analysis,  $F_{ST}$ , and SNP x Topic interaction analyses, we used genetic  
 936 data from 488,377 UK Biobank participants (prior to restricting to 282,957 samples with at least  
 937 two of the 348 diseases studied). For PRS and heritability estimation of the 10 topics, we used  
 938 mixed-effect association model implemented in BOLT-LMM software<sup>50,51</sup>, where we

939 **constrained** our analysis to 409,694 British Isle ancestry individuals to **adjust for** population  
940 structure. For  $F_{ST}$  analysis with PLINK we used 805,426 genotyped SNPs; for BOLT-LMM  
941 PRS analysis we used 727,882 genotyped SNP with  $MAF > 0.1\%$ ; for genetic correlation analysis  
942 using LDSC, we used 157,756 Genotyped SNPs mapped to HapMap3 SNPs; for **computing**  
943 **heritability where mixed-effect association are performed using BOLT-LMM**<sup>50,51</sup> **subsequent**  
944 **heritability estimation was performed using LDSC**<sup>2</sup>, we used 1,201,838 imputed SNPs mapped  
945 to HapMap3 SNPs SNPs.

946

#### 947 **Polygenic risk scores (PRS) analysis.**

948 **Despite population stratification cannot be excluded**<sup>71</sup>, **to adjusted for and minimize the impact**  
949 **of population stratification, we applied mixed-effect association model to samples of British Isle**  
950 **ancestry group (N = 409,694) to compute PRS, for 10 heritable diseases that have the highest**  
951 **heritability z-scores. We used a mixed model to estimate effect size implemented by BOLT-**  
952 **LMM and constructed genome-wide PRS**<sup>50</sup>. **For four diseases with more than 20,485 case**  
953 **(essential hypertension, arthropathy, asthma, and hypercholesterolemia), we downsampled**  
954 **controls to make the total sample size half of that of British isle ancestry population (N =**  
955 **204,847) for computation efficiency; for other diseases, we sampled 9 controls for each case to**  
956 **ensure case proportion at or above 10% as recommended by BOLT-LMM (type 2 diabetes,**  
957 **varicose veins of lower extremity, hypothyroidism, other peripheral nerve disorders, major**  
958 **depressive disorder, and GRED). We used PLINK to select genotyped SNPs with  $MAF > 0.1\%$**   
959 **as recommended in BOLT-LMM. For each disease, we used 5-fold cross validation to estimate**  
960 **effect sizes using BOLT-LMM and computed the PRS on the held-out testing set. We used**  
961 **continuous-valued topic weights to analyse association between disease subtypes and PRS.** The  
962 predictive PRS are used to compute the excess PRS over different topic loadings, by a linear  
963 regression where PRS is the response variable and topic weights is the predictor.

964

965 We compute the relative risk for each percentile of PRS using the following formula:

966

$$RR_{pt,s} = \frac{n_{pt,s} \times 100}{n_s},$$

967 where  $RR_{pt,s}$  is the relative risk of  $s$  subtype for the  $pt^{th}$  PRS percentile (computed for the entire  
968 population);  $n_{pt,s}$  is the number of cases in  $s$  subtype that has PRS within the  $pt^{th}$  percentile;  $n_s$   
969 is the number of cases in the  $s$  subtype.

970

#### 971 **Genetic correlation analysis.**

972 **We used discrete subtypes in genetic correlation analysis (different from the PRS analysis**  
973 **above).** For each disease and disease subtype, we use a case-control matching strategy to  
974 construct data to estimate coefficients for genetic correlation analysis. For each case in the  
975 disease group, we pick four nearest neighbors (without replacement) from the control group,  
976 matching sex, BMI, year of birth and 40 genetic principal components. The covariates are  
977 available within the UK Biobank data set, over which we computed the principal components.

978 We then compute the Euclidean distance of the principal components to find the nearest  
979 neighbours in the population. All cases are matched with four controls except for 401.1 essential  
980 hypertension which has a sample size larger than 20% of the population. We match only one  
981 control for each hypertension case.

982  
983 We perform logistic regression with sex and top 10 principal components as covariates to  
984 estimate the main variant effect of the 805,426 variants that are genotyped. We used PLINK 1.9  
985 for association analysis<sup>72</sup>. With the summary statistics from the association analysis, we use  
986 LDSC to map the summary statistics to HapMap3 SNPs and match the effect and non-effect  
987 alleles<sup>2,73</sup>. Since UK Biobank is mostly of British Isle ancestry, we use the pre-computed LD  
988 score from the LDSC website. We estimated the heritability for each disease or disease subtype  
989 which has more than 1000 incidences ([378 = 30 diseases subtypes + 348 diseases](#)). We use 1000  
990 incidence threshold as LDSC are more accurate with larger sample size. We focus on 71 disease  
991 and 18 disease subtypes [of the 378 diseases subtypes and diseases](#) that have heritability z-score  
992 above 4 for genetic correlation analysis.

993  
994 The genetic correlation is computed for each pair of [disease-disease, disease-subtype, and](#)  
995 [subtype-subtype](#) using the [logistic regression](#) summary statistics and LD score regression. We  
996 report the estimate of genetic correlation and z-scores. Additionally, for pairs that involve  
997 subtypes (disease-subtype or subtype-subtype), we compute the excess genetic correlation,  
998 defined as the difference between the genetic correlation involving subtypes ([disease-subtype](#)  
999 [and subtype-subtype](#)) and the genetic correlation involving all disease diagnoses ([disease-](#)  
1000 [disease](#)). For example, the genetic correlation between T2D-CER and hypertension-CVD is  
1001 compared to the genetic correlation between all T2D and all hypertension. The z-score and p-  
1002 value of the genetic correlation differences are reported. We note that genetic correlations  
1003 between subtypes of the same disease are compared to 1. We only reported p-values of excess  
1004 genetic correlation when both genetic correlation estimation has standard error <0.1 and at least  
1005 one of the genetic correlation has  $|z\text{-score}|>4$ .

1006  
1007 To avoid potential collider effects where subtypes are defined by topic components that are  
1008 independent of the diseases, we [performed the same genetic correlation analyses but](#) match cases  
1009 in each subtype with controls [with similar topic loadings](#). We computed PCs from 23 variables  
1010 (10 topic loadings, 10 PCs, year of birth, sex, and BMI) and used the nearest neighbour  
1011 procedure (by Euclidean Distance) to find controls for each case. Here controls are chosen from  
1012 individuals without the targeting disease, i.e. an individual with one subtype of the target disease  
1013 could not be a control for the other subtypes. We performed the same analysis using this case-  
1014 control matching procedure and compared the genetic correlation with the case-control  
1015 procedure described above. We perform the analysis for four diseases that have evidence for  
1016 genetic subtypes: asthma, type 2 diabetes, hypercholesterolemia, and hypertension. For one

1017 subtype (hypertension-CVD), the heritability (0.0313, s.e. = 0.0289) is below threshold after  
1018 matching the topic, which was excluded in genetic correlation analysis.

1019

### 1020 **$F_{ST}$ analysis.**

1021 We used discrete subtypes in genetic correlation analysis (same as genetic correlation analysis  
1022 above; different from the PRS analysis). To evaluate the genetic heterogeneity between disease  
1023 subtypes, we estimated the  $F_{ST}$  for 52 diseases that have at least 500 incidences assigned to a  
1024 secondary topic. To test the statistical significance of  $F_{ST}$ , we adopted a permutation strategy and  
1025 sampled the same number of controls of similar topic weights distribution for each subtype. The  
1026 topic weights are matched by sampling (without replacement) the same number of controls for  
1027 each dominant topic weight quartile of the cases (i.e. matching the topic that defines the  
1028 subtype), which ensures the controls have the same topic weight stratification as the disease  
1029 subtypes. We then compute the  $F_{ST}$  across the control groups matched for subtypes. We excluded  
1030 three diseases, “hypertension”, “hypercholesterolemia”, and “arthropathy”, from  $F_{ST}$  analysis as  
1031 we do not have enough controls that match topic weight distribution. The  $F_{ST}$ s are computed  
1032 using PLINK 1.9's weighted mean across all genotyped SNPs, which report  $F$  statistics across all  
1033 subtypes.

1034

1035 We obtained 1,000 permutation samples and reported the permutation p-value. Under the  
1036 assumption that causal and non-causal variants have similar allele frequency differences across  
1037 the subtypes,  $F_{ST}$  is a measure of causal genetic effect heterogeneity across subtypes.

1038

### 1039 **SNP x topic interaction test.**

1040 We used continuous-valued topic weights in the SNP x topic interaction analysis (same as the  
1041 PRS analysis; different from the genetic correlation and  $F_{ST}$  analyses). For the diseases that have  
1042 heritability z-score above 4 in the UK Biobank, we further investigated whether there are  
1043 interactions between genetic risk factors with the topic loadings. We used a fit a logistic  
1044 regression model using following model:

$$1045 \quad \text{logit}(p) = \beta_0 + \beta_1 * T + \beta_2 * T^2 + \beta_3 * G + \beta_4 * G * T,$$

1046 where  $T$  is individual topic weights for a specified topic,  $G$  is the genotype, and  $p$  is the  
1047 probability of getting the disease. We computed the test statistics under the null that  $\beta_4 = 0$ .

1048

1049 Since the simulation shows the interaction test is underpowered when the variant effects are  
1050 small, we focus on the set of SNP that reaches genome-wide significance level to increase power  
1051 to detect interaction effects. We performed LD-clumping using  $r^2 > 0.6$  to remove variants that  
1052 are in strong LD with the lead variants. We computed the test statistics using the model above  
1053 (for testing  $\beta_4 = 0$ ) and computed study-wise FDR across 2530 disease-topic pairs. We used QQ  
1054 plots to check that interaction test statistics computed using all non-subtype topics for each  
1055 disease (which are expected to be null) were well-calibrated. (Supplementary Figure 31B).

1056

1057  
1058 To verify the significant interactions, we divided cases into quartiles based on topic loading for  
1059 each disease-topic pair, and randomly sampled two controls that match the topic loading for each  
1060 case. We estimated the main effect sizes for all GWAS-SNP within each quartile of topic  
1061 loadings to capture effects that are modulated by topic weights. We focus on the SNPs that have  
1062 significant interaction test statistics computed in the previous step and compare it with  
1063 background SNPs that have genome-wide significant main effects but no interaction effect  
1064 ( $P > 0.05$ ).

### 1065 1066 **Simulations of SNP x topic interaction**

1067 We simulate comorbidity with genetics to test interaction between genetic and comorbidity  
1068 topics. We simulated 100 independent variants with MAF randomly sampled from [the MAF of](#)  
1069 [888 independent disease associated SNPs](#). We assumed an additive model and simulated  
1070 genotypes for the population using Hardy-Weinberg equilibrium. We simulated three types of  
1071 genetic effects on topic and diseases on topic of the simulation framework described in  
1072 Simulations of ATM method section:

- 1073 ● Genetics-topic effect: each variant is simulated to have an linear effect of 0.04 on the  
1074 topic loading. We choose this value as after normalising the topic, a regression of causal  
1075 variant to topic would have an effect size approximately 0.01 which is similar to our  
1076 observation in the UK Biobank. The number of variants that are causal to the topic varies  
1077 between 2 to 20. We simulated the effect on one topic by adding additive SNP effects and  
1078 normalise the topic loadings of each patient. The topic-disease causality is a natural  
1079 consequence following the generative process of sampling data.
- 1080 ● Genetic-disease-topic effect: we simulated a heritable disease that is causal to the topic.  
1081 The disease is simulated with 20 causal variants each of effect size 0.15. We vary the  
1082 disease-to-topic causal effect from 0.05 to 0.5, with a default value of 0.1 in other  
1083 analyses (similar to the correlation we found in UK Biobank analysis). We simulated the  
1084 effect on one topic by adding additive causal disease effects and normalise the topic  
1085 loadings of each patient.
- 1086 ● The genetic effect could interact with the topic when contributing to disease risk. We  
1087 simulated four additional diseases to represent different structures (Supplementary Figure  
1088 28).
  - 1089 ○ Genetic effects interact with topic loading on altering disease risk. The interaction  
1090 term is added to the mean of disease liability, which is sampled from a Gaussian  
1091 distribution. The disease is then sampled by a threshold on the liability, where the  
1092 incidence rate is by default 0.5. The interaction effect is varied from 0.4 to 4, with  
1093 default value equal to 2.
  - 1094 ○ Pleiotropy effects are simulated with a variant that have both genetic-disease and  
1095 genetic-topic-disease effects. Both genetic and topic effects are added to the mean  
1096 of disease liability. A disease is sampled by a threshold with default incidence rate

1097 equal to 0.5. The topic-disease effect is varied from 0.4 to 4, with default value  
1098 equal to 2.

1099 ○ Pleiotropy effect with nonlinear topic-disease effect. A quadratic term of topic-  
1100 disease effect added to the second model.

1101 ○ Pleiotropy effect with nonlinear genetic-disease effect. A quadratic term of  
1102 genetic-disease effect added to the second model.

1103

1104 For disease-topic or topic-disease causal effects, we simulated 50 repetition at each causal effect  
1105 size. For interaction analysis, we repeated 10 times at each parameter value, as there are more  
1106 SNPs for uncertainty estimation. The simulated disease sets are fed into the inference procedure  
1107 to infer the patient topic weights.

1108

1109

1110

1111

1112 **Tables**

1113

Acronym	Disease systems	Representative diseases	Number of associated diseases
NRI	Neoplasms, respiratory, infectious diseases	Secondary malignancy of lymph nodes; Pneumococcal pneumonia; Bacterial infection NOS	53
CER	Circulatory system, endocrine/metabolic, respiratory	Type 2 diabetes; Obesity; Chronic airway obstruction	41
SRD	Sense organs, respiratory, dermatologic	Cataract; Septal Deviations/Turbinates Hypertrophy; Benign neoplasm of skin	38
CVD	Cardiovascular disease	Hypercholesterolemia; Coronary atherosclerosis; Myocardial infarction	27
UGI	Upper gastrointestinal disease	Diaphragmatic hernia; Benign neoplasm of other parts of digestive system; Gastritis and duodenitis;	22
LGI	Lower gastrointestinal disease	Irritable Bowel Syndrome; Benign neoplasm of colon; Anal and rectal polyp;	13
FGND	Female genitourinary, neoplasms, digestive	Uterine leiomyoma; Malignant neoplasm of female breast; Hypothyroidism NOS	34
MGND	Male genitourinary, neoplasms, digestive	Urinary tract infection; Cancer of prostate; Other disorders of bladder	33
MDS	Musculoskeletal, digestive, symptoms	Back pain; Cholelithiasis; Other disorders of soft tissues	29
ARP	Arthropathy-related disease	Arthropathy NOS; Rheumatoid arthritis; Enthesopathy	26

1114

1115  
1116  
1117  
1118  
1119  
1120  
1121  
1122  
1123  
1124

**Table 1. Summary of 10 inferred disease topics in the UK Biobank.** For each topic, we list its 3-letter acronym, disease systems, representative diseases, and number of associated diseases (defined as diseases with average diagnosis-specific topic probability >50% for that topic). Topics are ordered by the Phecode system (see Figure 3). 316 of 348 diseases analysed are associated with a topic; the remaining 32 diseases do not have a topic with average diagnosis-specific topic probability >50%.

Phecode	Name of disease	Subtypes in UKB (≥500 diagnoses)	UKB-AOU correlation
41	Bacterial infection NOS	NRI, UGI	41.9%
41.1	Staphylococcus infections	NRI, CER	NA
153.2	Colon cancer	NRI, LGI	32.7%
153.3	Malignant neoplasm of rectum, rectosigmoid junction, and anus	NRI, LGI	44.7%
174.1	Breast cancer [female]	NRI, FGND	98.6%
174.11	Malignant neoplasm of female breast	NRI, FGND	86.9%
180.3	Cervical intraepithelial neoplasia [CIN] [Cervical dysplasia]	SRD, FGND	46.2%
214.1	Lipoma of skin and subcutaneous tissue	SRD, MGND	NA
244.4	Hypothyroidism NOS	FGND, MDS	64.7%
250.2	Type 2 diabetes	CER, CVD, MGND	81.5%
272.11	Hypercholesterolemia	CER, CVD, FGND, MGND	69.8%
278.1	Obesity	CER, CVD, ARP	61.1%
285	Other anemias	NRI, UGI	70.5%
296.22	Major depressive disorder	CER, FGND, MDS	94.5%
300.1	Anxiety disorder	CER, FGND, MDS	88.2%
340	Migraine	CER, MDS	81.7%
351	Other peripheral nerve disorders	MDS, ARP	96.7%
401.1	Essential hypertension	CER, SRD, CVD, LGI, FGND, MGND, MDS, ARP	85.6%

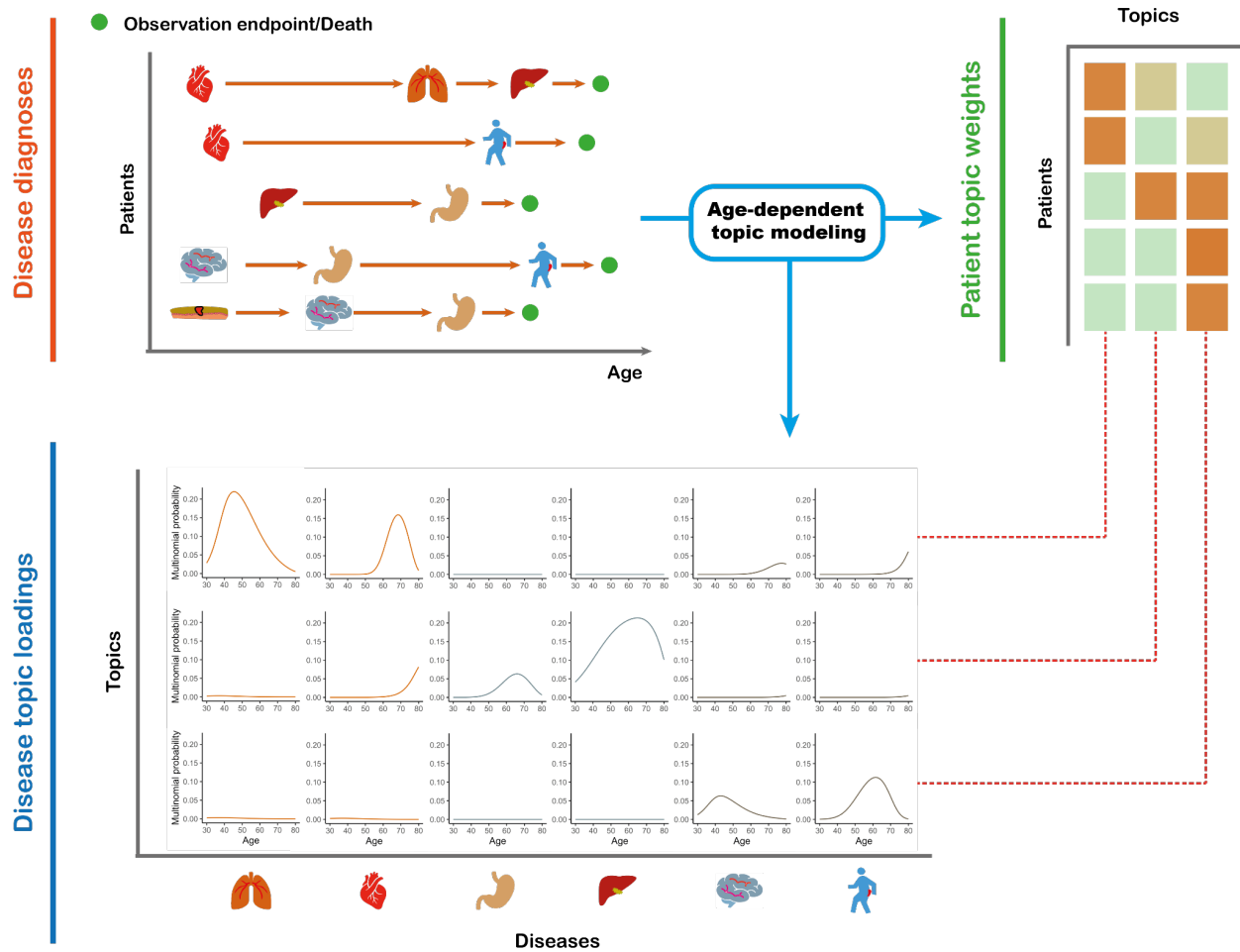


443.9	Peripheral vascular disease, unspecified	CER, CVD	83.0%
451.2	Phlebitis and thrombophlebitis of lower extremities	NRI, CER	NA
454.1	Varicose veins of lower extremity	SRD, FGND, MGND, ARP	71.1%
480	Pneumonia	NRI, CER	95.7%
480.11	Pneumococcal pneumonia	NRI, CER	NA
495	Asthma	CER, SRD, FGND, MGND, MDS, ARP	75.9%
496.3	Bronchiectasis	NRI, CER	94.3%
502	Postinflammatory pulmonary fibrosis	NRI, CER	95.6%
509.1	Respiratory failure	NRI, CER	93.2%
519.8	Other diseases of respiratory system, NEC	NRI, CER	90.4%
521.1	Dental caries	FGND, MGND	5.1%
525	Other diseases of the teeth and supporting structures	FGND, MGND	-11.8%
530.11	GERD	UGI, MDS	-4.8%
550.5	Ventral hernia	NRI, LGI, MGND	60.3%
558	Noninfectious gastroenteritis	NRI, LGI	66.4%
560.4	Other intestinal obstruction	NRI, LGI	NA
563	Constipation	NRI, LGI, MDS	-18.7%
564.1	Irritable Bowel Syndrome	LGI, FGND, MDS	55.2%
568.1	Peritoneal adhesions (postoperative) (postinfection)	NRI, MDS	NA
571.5	Other chronic nonalcoholic liver disease	CER, MDS	-9.0%
585.1	Acute renal failure	NRI, CER	95.7%
585.3	Chronic renal failure [CKD]	NRI, CER	67.0%
591	Urinary tract infection	NRI, MGND	53.9%
595	Hydronephrosis	NRI, MGND	92.1%
599.4	Urinary incontinence	FGND, MGND, MDS	NA
610.4	Benign neoplasm of breast	SRD, FGND	NA

611.3	Lump or mass in breast	NRI, FGND	NA
618.1	Prolapse of vaginal walls	FGND, MDS	38.4%
626.12	Excessive or frequent menstruation	FGND, MDS	NA
706.2	Sebaceous cyst	SRD, FGND, MGND	78.4%
716.9	Arthropathy NOS	FGND, MDS, ARP	90.9%
729	Other disorders of soft tissues	CER, MDS	25.5%
743.11	Osteoporosis NOS	NRI, FGND, MDS	73.7%
745	Pain in joint	MDS, ARP	NA

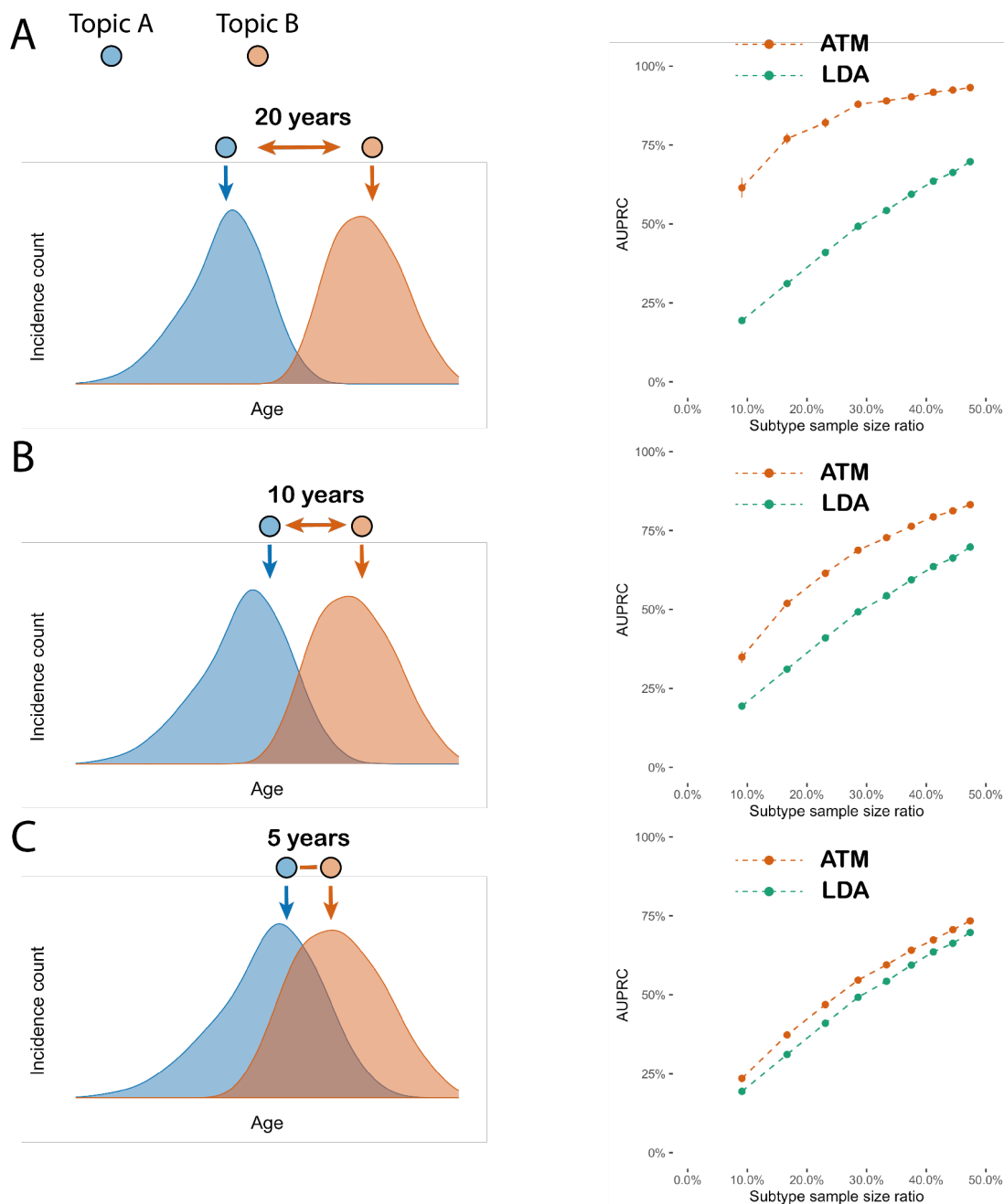
1125 **Table 2. List of 52 diseases with comorbidity subtypes in UK Biobank.** For each disease with  
1126 at least 500 diagnoses assigned to each of two discrete subtypes, we list its Phecode, disease  
1127 description, subtypes with at least 500 diagnoses, and correlations between UK Biobank topic  
1128 assignments and All of Us topic assignments that were mapped to UK Biobank topics (Methods).  
1129  
1130

1131 **Figures**

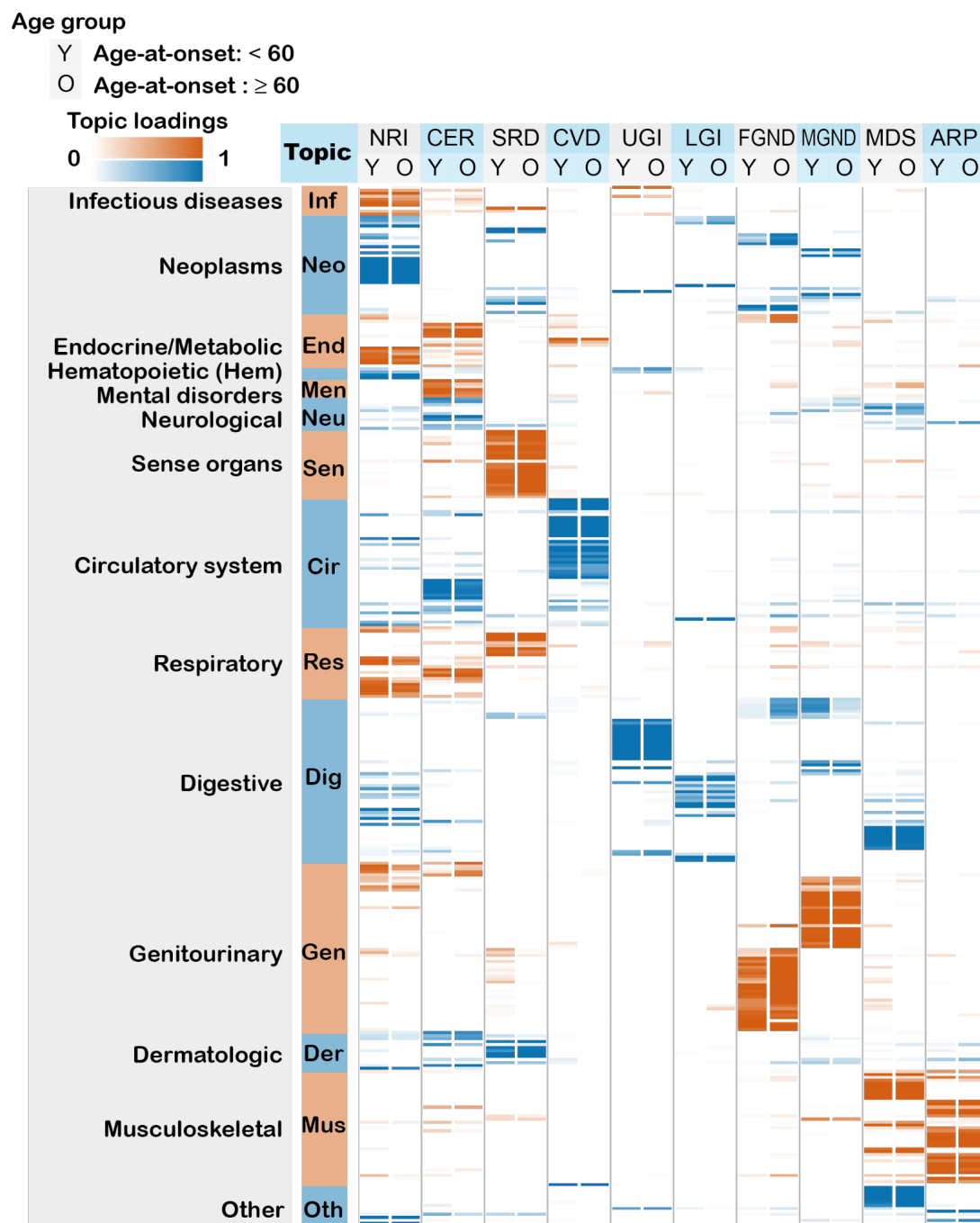


1132  
1133 **Figure 1: ATM provides an efficient way to represent longitudinal comorbidity data.** Top  
1134 left: input consists of disease diagnoses as a function of age. Top right: ATM assigns a topic  
1135 weight to each patient. Bottom: ATM infers age-dependent topic loadings.

1136  
1137  
1138  
1139

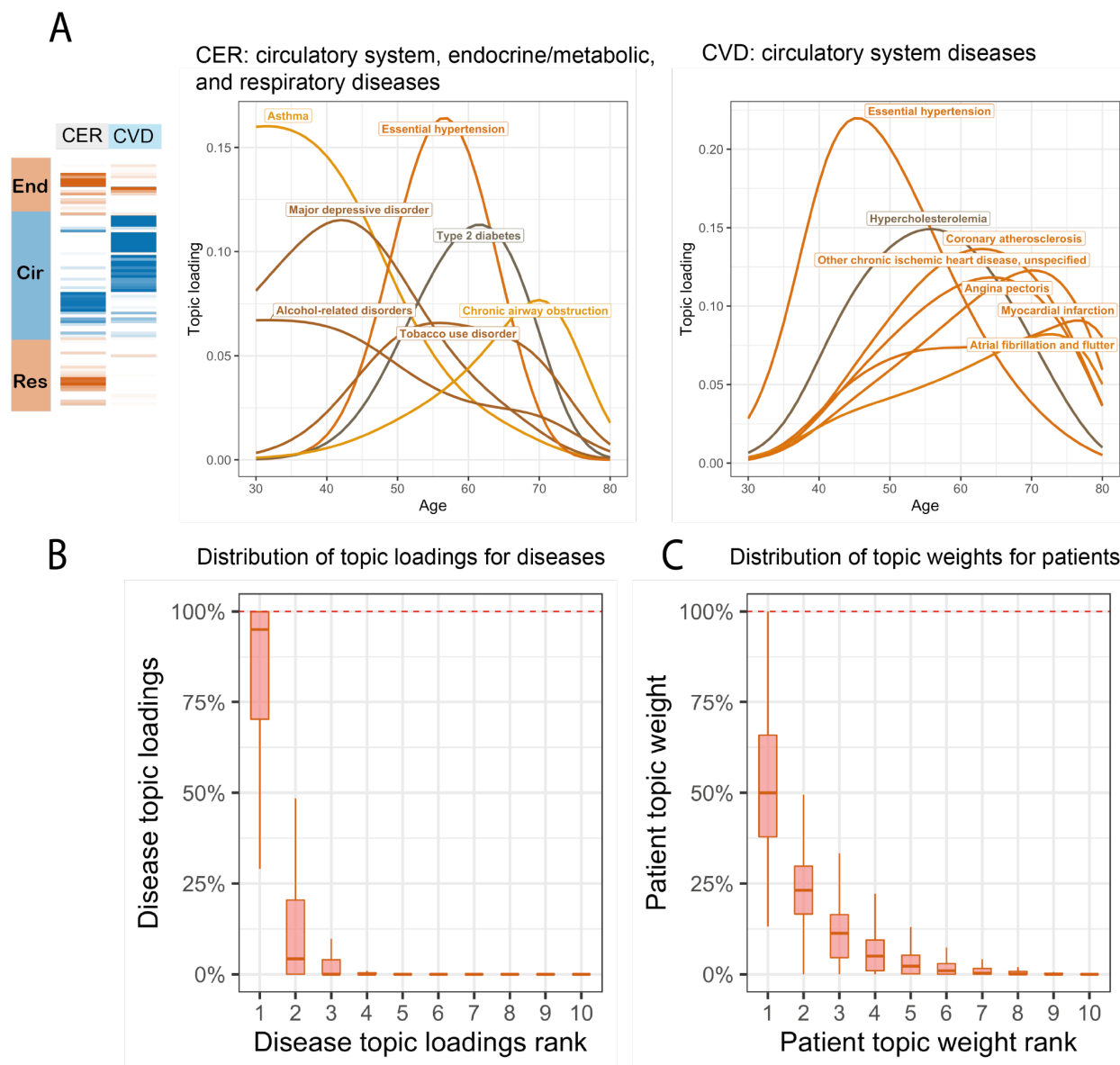


1140  
 1141 **Figure 2: ATM outperforms LDA in simulations with age-dependent effects.** In simulations  
 1142 at different levels of age-dependent effects (left panels), we report the area under the precision  
 1143 and recall curve (AUPRC) for ATM vs. LDA as a function of subtype sample size proportion  
 1144 (the proportion of diagnoses belonging to the smaller subtype) (right panels). Each dot represents  
 1145 the mean of 100 simulations of 10,000 individuals. Error bars denote 95% confidence intervals.  
 1146 (A) 20-year difference in age at diagnosis for the two subtypes. (B) 10-year difference in age at  
 1147 diagnosis for the two subtypes. (C) 5-year difference in age at diagnosis for the two subtypes.  
 1148 Numerical results are reported in Supplementary Table 2.  
 1149



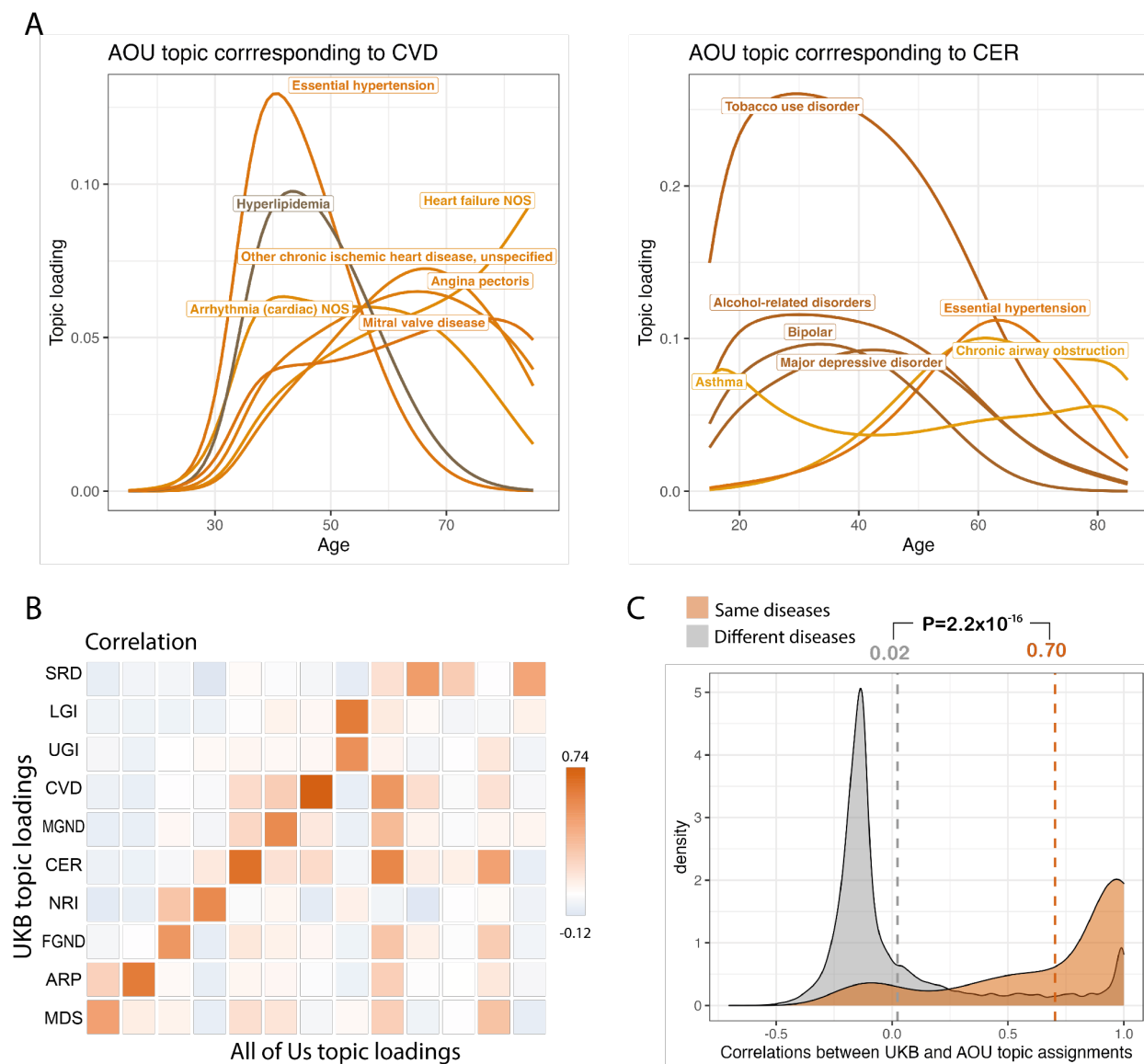
1150  
 1151 **Figure 3. Age-dependent topic loadings of 10 inferred disease topics across 348 diseases in**  
 1152 **the UK Biobank.** We report topic loadings averaged across younger ages (age at diagnosis < 60)  
 1153 and older ages (age at diagnosis > 60). Row labels denote disease categories ordered by Phecode  
 1154 systems, with alternating blue and red color for visualisation purposes; “Other” is a merge of five  
 1155 Phecode systems: “congenital anomalies”, “symptoms”, “injuries & poisoning”, “other tests”,  
 1156 and “death” (which is treated as an additional disease, see Methods). Topics are ordered by the  
 1157 corresponding Phecode system. Further details on the 10 topics are provided in Table 1. Further

1158 details on the diseases discussed in the text (type 2 diabetes and breast cancer) are provided in  
 1159 Supplementary Figure 6. Numerical results are reported in Supplementary Table 4.  
 1160  
 1161



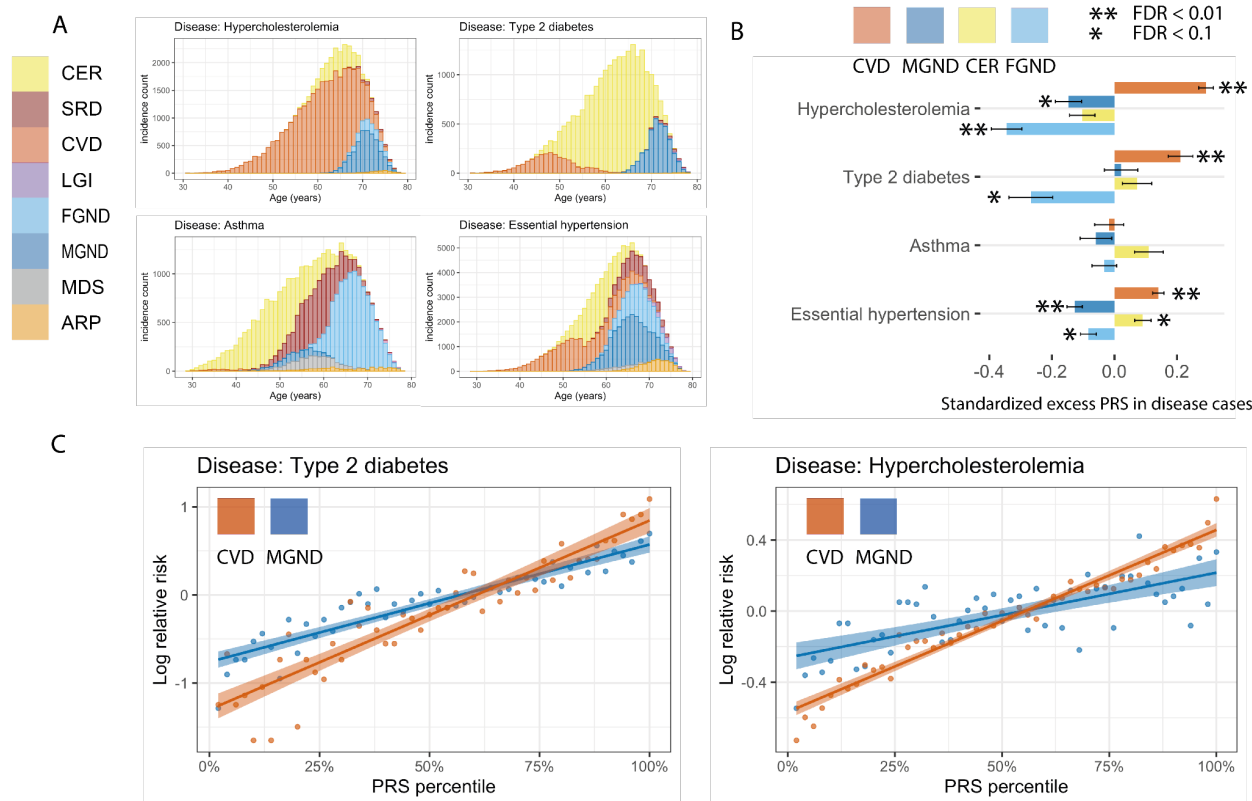
1162  
 1163 **Figure 4. Topic loadings in UK Biobank capture age-dependent comorbidities.** (A) Age-  
 1164 dependent topic loadings for two representative topics, CER and CVD; for each topic, we  
 1165 include the top seven diseases with highest topic loadings. Results for all 10 topics are reported  
 1166 in Supplementary Figure 13. (B) Box plot of disease topic loading as a function of rank; disease  
 1167 topic loadings are computed as a weighted average across all values of age at diagnosis. (C) Box  
 1168 plot of patient topic weight as a function of rank. Numerical results are reported in  
 1169 Supplementary Table 5.  
 1170

1171  
1172



1173  
1174 **Figure 5. Topic loadings in All of Us capture age-dependent comorbidities that are**  
1175 **concordant with UK Biobank (A)** Age-dependent topic loadings for two All of Us topics  
1176 corresponding to CVD and CER (Figure 4A); for each topic, we include the top seven diseases  
1177 with highest topic loadings. Correlations of topic loadings between UK Biobank and All of Us  
1178 topics were 0.74 for CVD and 0.65 for CER. Numerical results for all 13 topics are reported in  
1179 Supplementary Table 6. (B) Topic loading correlations between UK Biobank (UKB) and All of  
1180 Us (AOU). The y-axis reflects the 10 topics from the optimal UK Biobank model; the x-axis  
1181 reflects the 13 topics from the optimal All of Us model. Numerical results are reported in  
1182 Supplementary Table 7. (C) Correlations between UKB and AOU topic assignments were higher  
1183 for same diseases (red shading, average = 0.70) than for different diseases (grey shading, average  
1184 = 0.02)<sup>[PA1]</sup>. Numerical results are reported in Supplementary Table 9.

1185



1186

1187 **Figure 6. Polygenic risk scores vary across disease subtypes defined by distinct topics. (A)**

1188 Stacked barplots of age-dependent subtypes (defined by topics) for 4 representative diseases

1189 (type 2 diabetes, asthma, hypercholesterolemia, and essential hypertension); for each disease, we

1190 include all subtypes with at least one diagnosis. Results for all 52 diseases are reported in

1191 Supplementary Figure 21. (B) Standardised excess PRS values in disease cases (s.d. increase in

1192 PRS per unit increase in patient topic weight) for 4 representative diseases and 4 corresponding

1193 topics. (C) Relative risk for cases of type 2 diabetes and hypercholesterolemia of CVD and

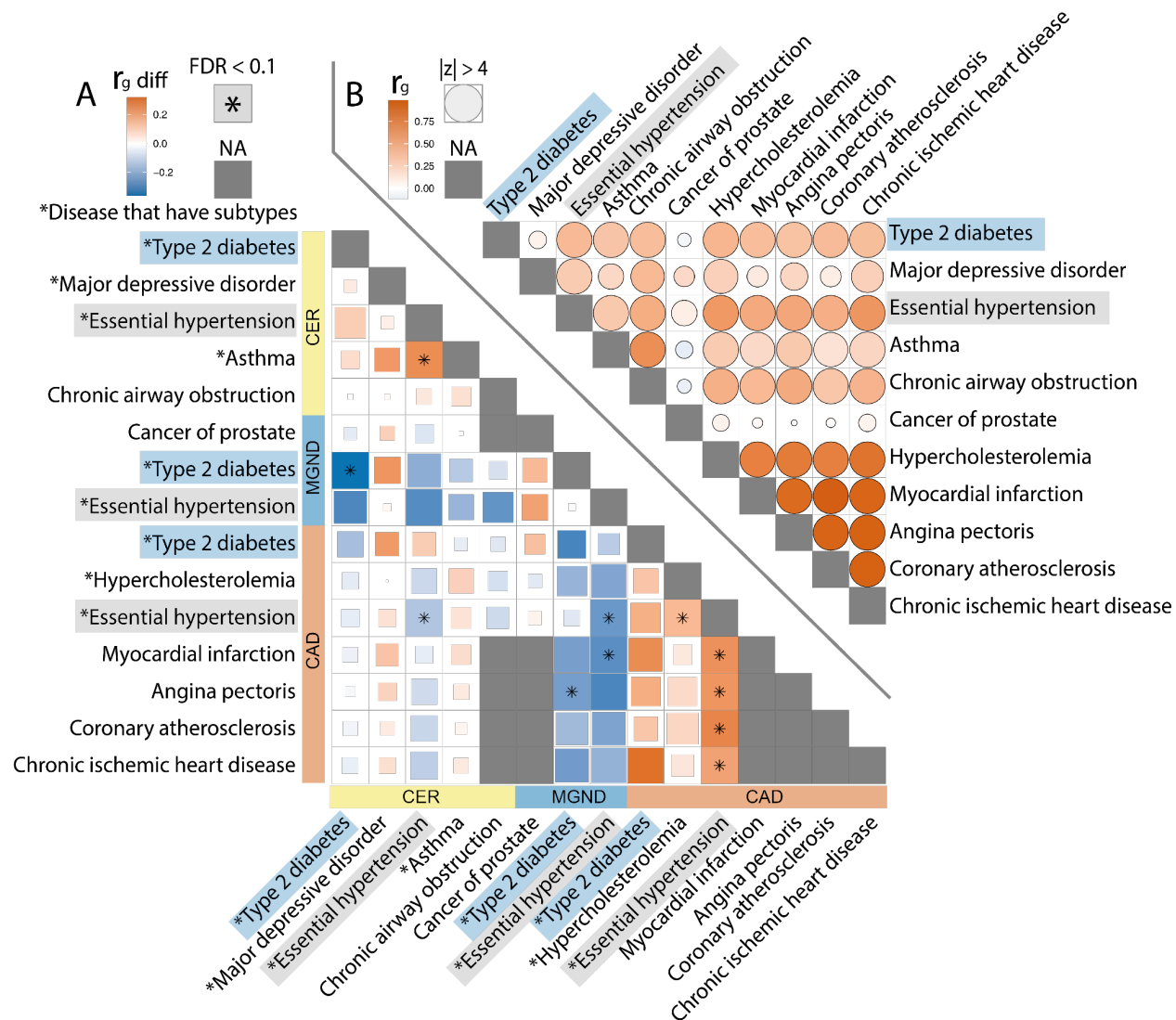
1194 MGND subtypes (vs. controls) across PRS percentiles. Each point spans 2 PRS percentiles.

1195 Lines denote regression on log scale. Error bars denote 95% confidence intervals. Numerical

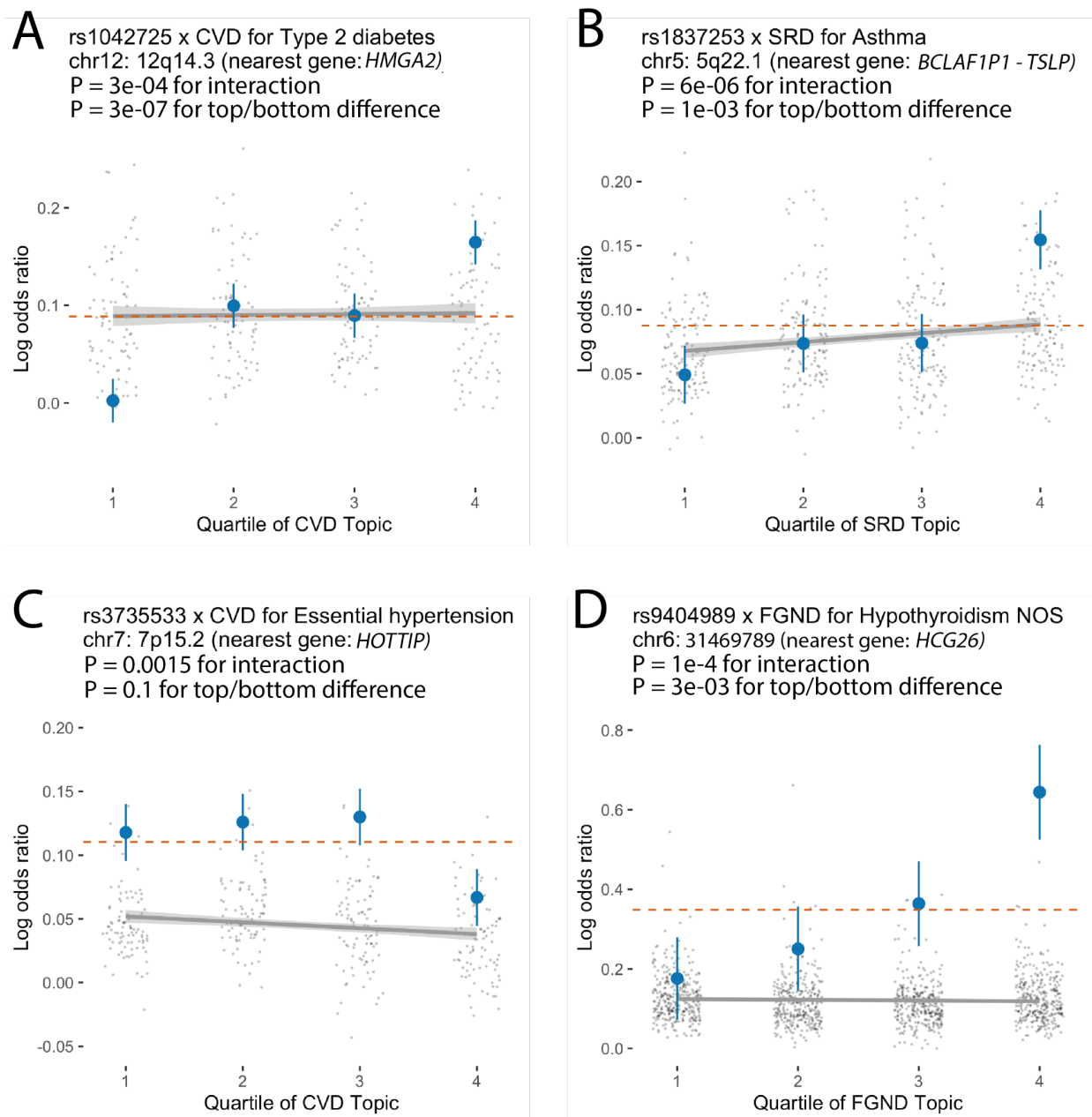
1196 results are reported in Supplementary Table 10-12.

1197





1198  
 1199 **Figure 7. Genetic correlations vary across disease subtypes defined by distinct topics.** (a)  
 1200 Excess genetic correlations for pairs of 15 disease subtypes or diseases (9 disease subtypes  
 1201 (denoted with asterisks) + 6 diseases without subtypes), relative to genetic correlations between  
 1202 the underlying diseases. Full square with asterisk denotes FDR < 0.1; less than full squares have  
 1203 area proportional to z-scores for difference. Grey squares denote NA (pair of diseases without  
 1204 subtypes or pair of same disease subtype or disease). (b) Genetic correlations between the  
 1205 underlying diseases. Full circle denotes  $|z\text{-score}| > 4$  for nonzero genetic correlation; less than  
 1206 full circles have area proportional to  $|z\text{-score}|$ . Numerical results are reported in Supplementary  
 1207 Table 13.  
 1208



1209  
1210  
1211  
1212  
1213  
1214  
1215  
1216  
1217  
1218

**Figure 8. Examples of SNP x topic interaction effects on disease phenotypes.** For each example, we report main SNP effects (log odds ratios) specific to each quartile of topic weights across individuals, for both the focal SNP (blue dots) and background SNPs for that disease and topic (genome-wide significant main effect ( $P < 5 \times 10^{-8}$ ) but non-significant SNP x topic interaction effect ( $P > 0.05$ ); grey dots). Dashed red lines denote aggregate main SNP effects for each focal SNP. Error bars denote 95% confidence intervals. Grey lines denote linear regression of grey dots, with grey shading denoting corresponding 95% confidence intervals. Numerical results are reported in Supplementary Table 18.

## 1219 References

- 1220 1. Abul-Husn, N. S. & Kenny, E. E. Personalized Medicine and the Power of Electronic Health  
1221 Records. *Cell* **177**, 58–69 (2019).
- 1222 2. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits.  
1223 *Nat. Genet.* **47**, 1236–1241 (2015).
- 1224 3. Wang, K., Gaitsch, H., Poon, H., Cox, N. J. & Rzhetsky, A. Classification of common  
1225 human diseases derived from shared genetic and environmental determinants. *Nat. Genet.*  
1226 **49**, 1319–1325 (2017).
- 1227 4. Zhao, W. *et al.* Identification of new susceptibility loci for type 2 diabetes and shared  
1228 etiological pathways with coronary heart disease. *Nat. Genet.* **49**, 1450–1457 (2017).
- 1229 5. Zhu, Z. *et al.* A genome-wide cross-trait analysis from UK Biobank highlights the shared  
1230 genetic architecture of asthma and allergic diseases. *Nat. Genet.* **50**, 857–864 (2018).
- 1231 6. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using  
1232 MTAG. *Nat. Genet.* **50**, 229–237 (2018).
- 1233 7. O'Connor, L. J. & Price, A. L. Distinguishing genetic correlation from causation across 52  
1234 diseases and complex traits. *Nat. Genet.* **50**, 1728–1734 (2018).
- 1235 8. Cortes, A., Albers, P. K., Dendrou, C. A., Fugger, L. & McVean, G. Identifying cross-  
1236 disease components of genetic risk across hospital data in the UK Biobank. *Nat. Genet.* **52**,  
1237 126–134 (2019).
- 1238 9. Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M. & He, X. Mendelian randomization  
1239 accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary  
1240 statistics. *Nat. Genet.* **52**, 740–747 (2020).
- 1241 10. Peyrot, W. J. & Price, A. L. Identifying loci with different allele frequencies among cases of  
1242 eight psychiatric disorders using CC-GWAS. *Nat. Genet.* **53**, 445–454 (2021).
- 1243 11. Mattheisen, M. *et al.* Identification of shared and differentiating genetic architecture for

- 1244 autism spectrum disorder, attention-deficit hyperactivity disorder and case subgroups. *Nat.*  
1245 *Genet.* **54**, 1470–1478 (2022).
- 1246 12. Cortes, A. *et al.* Bayesian analysis of genetic association across tree-structured routine  
1247 healthcare data in the UK Biobank. *Nat. Genet.* **49**, 1311–1318 (2017).
- 1248 13. Zhang, H. *et al.* Genome-wide association study identifies 32 novel breast cancer  
1249 susceptibility loci from overall and subtype-specific analyses. *Nat. Genet.* **52**, 572–581  
1250 (2020).
- 1251 14. Mansour Aly, D. *et al.* Genome-wide association analyses highlight etiological differences  
1252 underlying newly defined subtypes of diabetes. *Nat. Genet.* **53**, 1534–1542 (2021).
- 1253 15. Hautakangas, H. *et al.* Genome-wide analysis of 102,084 migraine cases identifies 123 risk  
1254 loci and subtype-specific risk alleles. *Nat. Genet.* **54**, 152–160 (2022).
- 1255 16. Srebro, N. & Shraibman, A. Rank, Trace-Norm and Max-Norm. in *Learning Theory* 545–  
1256 560 (Springer Berlin Heidelberg, 2005).
- 1257 17. Candès, E. & Recht, B. Exact matrix completion via convex optimization. *Commun. ACM*  
1258 **55**, 111–119 (2012).
- 1259 18. Yan, J. & Pollefeys, M. A General Framework for Motion Segmentation: Independent,  
1260 Articulated, Rigid, Non-rigid, Degenerate and Non-degenerate. in *Computer Vision – ECCV*  
1261 *2006* 94–106 (Springer Berlin Heidelberg, 2006).
- 1262 19. Ma, Y., Derksen, H. & Hong, W. Segmentation of multivariate mixed data via Lossy data  
1263 coding and compression. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**, 1546–1562 (2007).
- 1264 20. Rao, S., Tron, R., Vidal, R. & Ma, Y. Motion segmentation in the presence of outlying,  
1265 incomplete, or corrupted trajectories. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**, 1832–  
1266 1845 (2010).
- 1267 21. Liu, G. & Yan, S. Latent Low-Rank Representation for subspace segmentation and feature  
1268 extraction. in *2011 International Conference on Computer Vision* 1615–1622  
1269 ([ieeexplore.ieee.org](http://ieeexplore.ieee.org), 2011).

- 1270 22. Liu, Z. *et al.* Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. *arXiv*  
1271 *[cs.AI]* (2018).
- 1272 23. Chen, Y. & Chi, Y. Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix  
1273 Estimation: Recent Theory and Fast Algorithms via Convex and Nonconvex Optimization.  
1274 *IEEE Signal Process. Mag.* **35**, 14–31 (2018).
- 1275 24. Chen, Y. & Candès, E. J. The projected power method: An efficient algorithm for joint  
1276 alignment from pairwise differences. *Commun. Pure Appl. Math.* **71**, 1648–1714 (2018).
- 1277 25. Jia, G. *et al.* Estimating heritability and genetic correlations from large health datasets in the  
1278 absence of genetic data. *Nat. Commun.* **10**, 1–11 (2019).
- 1279 26. Tanigawa, Y. *et al.* Components of genetic associations across 2,138 phenotypes in the UK  
1280 Biobank highlight adipocyte biology. *Nat. Commun.* **10**, 4064 (2019).
- 1281 27. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human  
1282 phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
- 1283 28. Jia, G. *et al.* Discerning asthma endotypes through comorbidity mapping. *Nat. Commun.*  
1284 **13**, 1–19 (2022).
- 1285 29. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data.  
1286 *Nature* **562**, 203–209 (2018).
- 1287 30. of Us Research Program, A. The ‘All of Us’ Research Program. *N. Engl. J. Med.* **381**, 668–  
1288 676 (2019).
- 1289 31. Ishigaki, K. *et al.* Large-scale genome-wide association study in a Japanese population  
1290 identifies novel susceptibility loci across different diseases. *Nat. Genet.* **52**, 669–679  
1291 (2020).
- 1292 32. Siggaard, T. *et al.* Disease trajectory browser for exploring temporal, population-wide  
1293 disease progression patterns in 7.2 million Danish patients. *Nat. Commun.* **11**, 1–10 (2020).
- 1294 33. Posey, J. E. *et al.* Resolution of Disease Phenotypes Resulting from Multilocus Genomic  
1295 Variation. *N. Engl. J. Med.* **376**, 21–31 (2017).

- 1296 34. Cook, E. K. *et al.* Comorbid and inflammatory characteristics of genetic subtypes of clonal  
1297 hematopoiesis. *Blood Adv* **3**, 2482–2486 (2019).
- 1298 35. Udler, M. S. *et al.* Type 2 diabetes genetic loci informed by multi-trait associations point to  
1299 disease mechanisms and subtypes: A soft clustering analysis. *PLoS Med.* **15**, e1002654  
1300 (2018).
- 1301 36. Wani, B., Aziz, S. A., Ganaie, M. A. & Mir, M. H. Metabolic Syndrome and Breast Cancer  
1302 Risk. *Indian J. Med. Paediatr. Oncol.* **38**, 434–439 (2017).
- 1303 37. Blei, Ng & Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.* (2003).
- 1304 38. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using  
1305 multilocus genotype data. *Genetics* **155**, 945–959 (2000).
- 1306 39. Bishop, C. M. *Pattern Recognition and Machine Learning*. (Springer New York, 2006).
- 1307 40. Teh, Y., Newman, D. & Welling, M. A collapsed variational Bayesian inference algorithm for  
1308 latent Dirichlet allocation. *Adv. Neural Inf. Process. Syst.* **19**, (2006).
- 1309 41. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and  
1310 receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).
- 1311 42. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development  
1312 and Initial Evaluation. *JMIR Med Inform* **7**, e14325 (2019).
- 1313 43. Suvila, K. *et al.* Early Onset Hypertension Is Associated With Hypertensive End-Organ  
1314 Damage Already by MidLife. *Hypertension* HYPERTENSIONAHA11913069 (2019).
- 1315 44. Wong, B. *et al.* Cardiovascular Disease Risk Associated With Familial  
1316 Hypercholesterolemia: A Systematic Review of the Literature. *Clin. Ther.* **38**, 1696–1709  
1317 (2016).
- 1318 45. Shah, M. S. & Brownlee, M. Molecular and Cellular Mechanisms of Cardiovascular  
1319 Disorders in Diabetes. *Circ. Res.* **118**, 1808–1829 (2016).
- 1320 46. Shah, A. D. *et al.* Type 2 diabetes and incidence of cardiovascular diseases: a cohort study  
1321 in 1.9 million people. *The Lancet Diabetes & Endocrinology* **3**, 105–113 (2015).

- 1322 47. Dabelea, D. & Hamman, R. F. Elevated Cardiometabolic Risk Profile Among Young Adults  
1323 With Diabetes: Need for Action. *Diabetes care* vol. 42 1845–1846 (2019).
- 1324 48. Wong, J. L. & Evans, S. E. Bacterial Pneumonia in Patients with Cancer: Novel Risk  
1325 Factors and Management. *Clin. Chest Med.* **38**, 263–277 (2017).
- 1326 49. Falstie-Jensen, A. M. *et al.* Incidence of hypothyroidism after treatment for breast cancer—  
1327 a Danish matched cohort study. *Breast Cancer Res.* **22**, 1–10 (2020).
- 1328 50. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association  
1329 for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).
- 1330 51. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in  
1331 large cohorts. *Nat. Genet.* **47**, 284–290 (2015).
- 1332 52. Jiang, X., Holmes, C. & McVean, G. The impact of age on genetic risk for common  
1333 diseases. *PLoS Genet.* **17**, e1009723 (2021).
- 1334 53. Weir, B. S. & Cockerham, C. C. ESTIMATING F-STATISTICS FOR THE ANALYSIS OF  
1335 POPULATION STRUCTURE. *Evolution* **38**, 1358–1370 (1984).
- 1336 54. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST:  
1337 the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).
- 1338 55. Mahajan, A. *et al.* Fine-mapping type 2 diabetes loci to single-variant resolution using high-  
1339 density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
- 1340 56. Wu, Q.-Q. *et al.* The protective effect of high mobility group protein HMGA2 in pressure  
1341 overload-induced cardiac remodeling. *J. Mol. Cell. Cardiol.* **128**, 160–178 (2019).
- 1342 57. Indra, A. K. Epidermal TSLP: a trigger factor for pathogenesis of atopic dermatitis. *Expert*  
1343 *Rev. Proteomics* **10**, 309–311 (2013).
- 1344 58. Kanai, M. *et al.* Genetic analysis of quantitative traits in the Japanese population links cell  
1345 types to complex human diseases. *Nat. Genet.* **50**, 390–400 (2018).
- 1346 59. Oluwafemi, O. O. *et al.* Genome-Wide Association Studies of Conotruncal Heart Defects  
1347 with Normally Related Great Vessels in the United States. *Genes* **12**, (2021).

- 1348 60. Blei, D. M. & Lafferty, J. D. A correlated topic model of Science. *Ann. Appl. Stat.* **1**, 17–35  
1349 (2007).
- 1350 61. Zaitlen, N. *et al.* Informed conditioning on clinical covariates increases power in case-  
1351 control association studies. *PLoS Genet.* **8**, e1003032 (2012).
- 1352 62. Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank  
1353 participants. *bioRxiv* 2022.06.17.496443 (2022) doi:10.1101/2022.06.17.496443.
- 1354 63. Ross, J. S. Covid-19, open science, and the CVD-COVID-UK initiative. *BMJ* vol. 373 n898  
1355 (2021).
- 1356 64. Mostafavi, H. *et al.* Variable prediction accuracy of polygenic scores within an ancestry  
1357 group. *Elife* **9**, e48376 (2020).
- 1358 65. Dumitrescu, L. *et al.* Evidence for age as a modifier of genetic associations for lipid levels.  
1359 *Ann. Hum. Genet.* **75**, 589–597 (2011).
- 1360 66. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk  
1361 prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).
- 1362 67. Lin, J. *et al.* Integration of biomarker polygenic risk score improves prediction of coronary  
1363 heart disease in UK Biobank and FinnGen. *bioRxiv* (2022)  
1364 doi:10.1101/2022.08.22.22279057.
- 1365 68. Falconer, D. S. The inheritance of liability to diseases with variable age of onset, with  
1366 particular reference to diabetes mellitus. *Ann. Hum. Genet.* **31**, 1–20 (1967).
- 1367 69. Ghorbani, B., Javadi, H. & Montanari, A. An Instability in Variational Inference for Topic  
1368 Models. in *Proceedings of the 36th International Conference on Machine Learning* (eds.  
1369 Chaudhuri, K. & Salakhutdinov, R.) vol. 97 2221–2231 (PMLR, 09--15 Jun 2019).
- 1370 70. Davis, J. & Goadrich, M. The relationship between Precision-Recall and ROC curves. in  
1371 *Proceedings of the 23rd international conference on Machine learning* 233–240  
1372 (Association for Computing Machinery, 2006).
- 1373 71. Haworth, S. *et al.* Apparent latent structure within the UK Biobank sample has implications



- 1374 for epidemiological analysis. *Nat. Commun.* **10**, 333 (2019).
- 1375 72. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer  
1376 datasets. *Gigascience* **4**, 7 (2015).
- 1377 73. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in  
1378 diverse human populations. *Nature* **467**, 52–58 (2010).
- 1379

1380 **Supplementary Tables (see Excel file)**

1381 **Supplementary Table 1. List of metrics for evaluating ATM performance.** The name,  
1382 purpose, and implementation details of each metric are listed for comparison. For more details of  
1383 each metric, see Methods.

1384  
1385 **Supplementary Table 2. Simulation results for ATM on identifying disease subtypes.** We  
1386 show the area under the precision-recall curve (AUPRC) for ATM in simulated data with two  
1387 subtypes that have 20/10/5 years of age at diagnosis differences. Results for LDA (fifth column)  
1388 are also shown for comparison. Rows show results for varying proportions of samples that  
1389 belong to the smaller subtype. The results correspond to Figure 2.

1390  
1391 **Supplementary Table 3. Characteristics of disease topics inferred from the UK Biobank.**  
1392 For each topic, we listed the top 10 representative diseases (by topic loadings), heritability  
1393 estimates, average topic weights (across all individuals), average age (weighted across all disease  
1394 diagnosis assigned to the topic), proportional of variance explained by BMI, sex, Townsend  
1395 deprivation index, and birth year.

1396  
1397 **Supplementary Table 4. Topic loadings of 10 inferred disease topics across 348 diseases in**  
1398 **the UK Biobank.** For each disease we reported the topic loading across diagnoses before 60  
1399 years old and after 60 years old. The Phecode, number of incidences, ICD-10 code, disease  
1400 name, and Phecode systems are also listed for each disease. The values in this table correspond  
1401 to Figure 3.

1402  
1403 **Supplementary Table 5. Topic loadings as functions of age for 10 inferred disease topics.**  
1404 For each topic, we listed the topic loading of each disease from age 30 to 80 years old. At each  
1405 age point, the topic loadings add to one across diseases for each topic. The values correspond to  
1406 Figure 4A and Supplementary Figure 13.

1407  
1408 **Supplementary Table 6. Topic loadings as functions of age for 13 inferred disease topics**  
1409 **from the All of Us data.** For each topic, we listed the topic loading of each disease between age  
1410 20 to 85. At each age point, the topic loadings add to one across diseases for each topic.

1411  
1412 **Supplementary Table 7. Correlation of topic loadings between each pair of All of Us and**  
1413 **UK Biobank topics.** Numeric values for Figure 5B.

1414  
1415 **Supplementary Table 8. Prevalences in All of Us and UK Biobank for the 233 diseases that**  
1416 **are shared between the two data sets.**

1417  
1418 **Supplementary Table 9. Correlations between topic assignments for pairs of All of Us**  
1419 **disease and UK Biobank disease across 233 diseases that are shared between the two data**

1420 sets. Disease associations to topics are measured using average topic assignments (see Methods  
1421 for definition) for both UK Biobank and All of Us. Average topic assignments in All of Us are  
1422 mapped to UK Biobank using topic loading correlation; the correlation for each disease pair in  
1423 the UK Biobank topic space (Methods).  
1424

1425 **Supplementary Table 10. Number of diagnoses assigned to each subtypes for 52 diseases.**

1426 We listed the number of diagnoses assigned to each disease subtypes by the diagnosis-specific  
1427 topic probability, for the 52 diseases that have at least two subtypes with >500 diagnosis.  
1428

1429 **Supplementary Table 11. Average age at diagnosis for each subtypes of the 52 diseases.** We

1430 listed the age at diagnosis across all diagnoses within each disease subtype, for the 52 diseases  
1431 that have at least two subtypes with >500 diagnosis.  
1432

1433  
1434 **Supplementary Table 12. Excess PRS in cases for all topics across 10 diseases (selected by**  
1435 **heritability z-score).** We report the estimated changes in s.d. of PRS per unit changes in the  
1436 patient topic weight, which is estimated through regression across disease diagnoses. The PRS  
1437 was estimated using BOLT-LMM and all the cases of British Isle Ancestry. Numbers correspond  
1438 to Figure 6B and Supplementary Figure 23.  
1439

1440  
1441 **Supplementary Table 13. Excess genetic correlations.** Columns are: Phecode of the first  
1442 disease (trait1), Phecode of the second trait (trait2), subtype of the first trait (topic1), subtype of  
1443 the second trait (topic2), the z-score of genetic correlation between two disease subtypes  
1444 (subtype.rho.zscore), estimate of genetic correlation between two disease subtypes  
1445 (subtype.rho.est), standard error of genetic correlation between two disease subtypes  
1446 (subtyp.rho.err), the z-score of genetic correlation between two diseases (all.rho.zscore), estimate  
1447 of genetic correlation between two diseases (all.rho.est), standard error of genetic correlation  
1448 between two diseases (all.rho.err), z-score of excess genetic correlation (diff.zscore), estimate of  
1449 excess genetic correlation (diff.rg), absolute value of excess genetic correlation  
1450 (diff.rg.zscore.abs), p-values for excess genetic correlation (P), FDR for excess genetic  
1451 correlation (FDR), name of the first disease (phenotype.1), and name of the second disease  
1452 (phenotype.2). We only reported p-values of excess genetic correlation when both genetic  
1453 correlation estimation has standard error <0.1 and at least one of the genetic correlation has |z-  
1454 score|>4. Numbers correspond to Figure 7 and Supplementary Figure 25.  
1455

1456 **Supplementary Table 14. Heritability estimation for disease subtypes.** For each disease we  
1457 list heritability estimates for two subtypes using LDSC. Topic.x refer to the subtype with the  
1458 highest heritability, topic.y refer to subtype with the lowest heritability. We report the point  
1459 estimate, standard error, z-scores of both disease subtypes. The z-score of heritability differences

1460 between the two subtypes are also reported. Note we used a different sample threshold 1000 (due  
1461 to the power of LDSC), which includes 26 of the 52 diseases that have subtypes.

1462

1463 **Supplementary Table 15. Excess  $F_{ST}$  estimation across disease subtypes.** We report the  
1464 estimate of excess  $F_{ST}$  (computed as the  $F_{ST}$  across subtypes subtracted by the  $F_{ST}$  from controls  
1465 with matched topic weights). The p-values are for excess  $F_{ST}>0$ , which is computed from 1000  
1466 randomly sampled control sets.

1467

1468 **Supplementary Table 16. GxTopic interaction tests across independent GWAS SNPs.** Each  
1469 row represents one SNP x topic weight pair (disease subtype). OR, SE, STAT, and P represent  
1470 the odds ratio, standard error, test statistics, and p-value of the main effects. SNPxTopic OR,  
1471 SNPxTopic STAT, SNPxTopic P, SNPxTopic FDR represent the odds ratio, test statistics, p-  
1472 value, and genome-wide FDR of testing the interaction effect in model 2 of Supplementary  
1473 Figure 28. We use Studywise FDR which adjusts for multiple testing across GWAS SNPs of all  
1474 disease subtypes.

1475

1476 **Supplementary Table 17. Significant SNP x topic interactions.** Same table as supplementary  
1477 Table 16, but filtered to SNP-topic pairs with interaction effect passing  $FDR<0.1$ . Reported  
1478 SNP-topic pairs were selected for topic weights specific effect estimation in Figure 8 and  
1479 Supplementary Figure 30.

1480

1481 **Supplementary Table 18. Effect size estimation across topic weight quartiles for significant  
1482 SNP x topic interactions.** Quartile, mean\_effect, se\_effect, refer to the quartile of topic weight,  
1483 estimate of effect sizes of the SNP using case-controls from this quartile, and standard error of  
1484 the effect size estimation. We also reported the nearest genes reported by GWAS Catalog. The  
1485 last two columns report the P-value of effect size being different between the top and bottom  
1486 quartiles and the [FDR \(across 2530 tests\)](#).

1487

1488 **Supplementary Table 19. Literature search of disease subtypes identified by ATM.** We  
1489 searched on Pubmed using the description (ignoring conjunctions) AND “subtype” in  
1490 title/abstract and manually screened the top 10 relevant results between 2012 and 2022. The  
1491 studies that mentioned subtypes of the searched diseases are included in the “Published  
1492 references” column. If there is no reference of target disease subtypes among the top 10 search  
1493 results, we use “NA”. We note our search is not exhaustive but nevertheless provides  
1494 information on whether subtypes of the target disease are described in studies involving target  
1495 disease and subtypes.

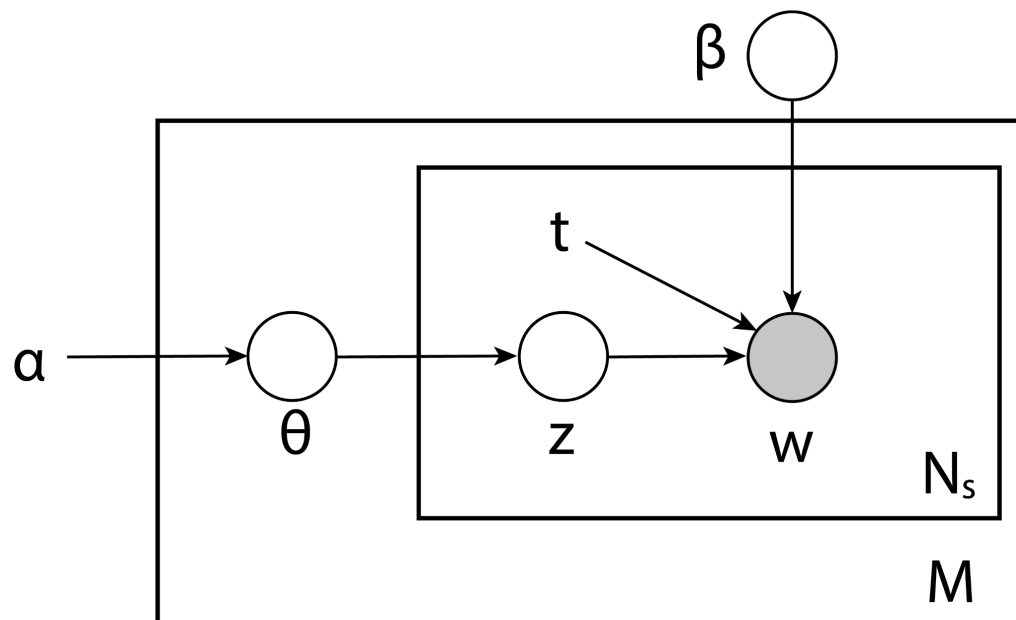
1496

1497 **Supplementary Table 20. ATM running time.** We tested the running time on the UK Biobank  
1498 data using ATM of varying topic number and parametric form of topic loadings. Degrees of  
1499 freedom from 2 to 7 represent linear, quadratic polynomial, cubic polynomial, spline with one

1500 knot, spline with two knots, and spline with three knots. Note a few models with 50 topics did  
1501 not converge.  
1502

1503 Supplementary Figures

1504



1505

1506

1507

1508

1509

1510

1511

1512

**Supplementary Figure 1. Plate notation of ATM generative model.**  $M$  is the number of subjects;  $N_s$  is the number of records within  $s^{th}$  subject. All plates (circles) are variables in the generative process, where the plates with shade  $w$  is the observed variable and plates without shade are unobserved variables to be inferred;  $\theta$  is the topic weight for all individuals;  $z$  is diagnosis-specific topic probability;  $t$  is the age at onset for each diagnosis;  $\beta$  is the topic loadings which are functions of age  $t$ ;  $\alpha$  is the (non-informative) hyperparameter of the prior distribution of  $\theta$ . The generative process is described in the Methods and Supplementary Note.

1513

1514

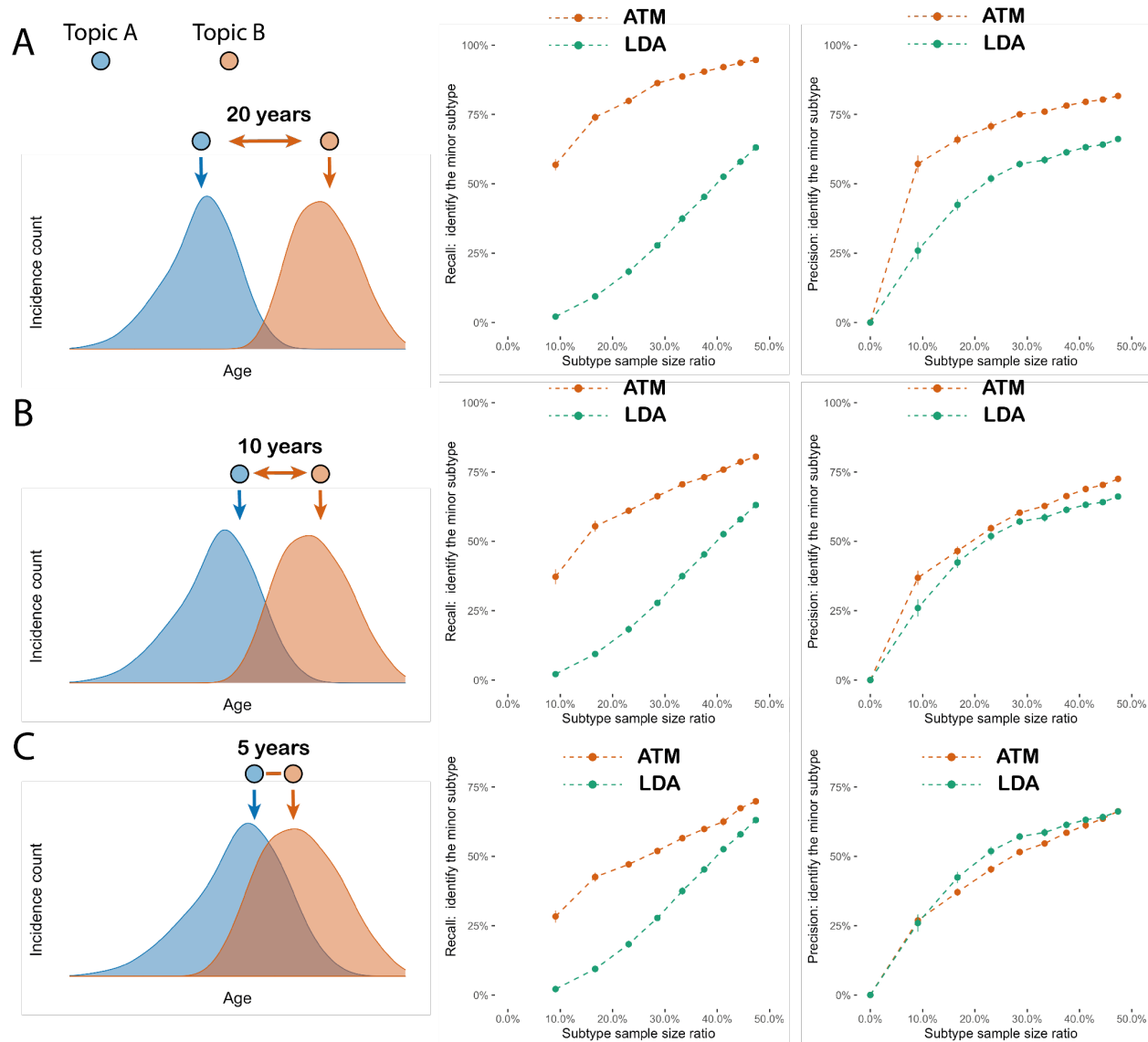
1515

1516

1517

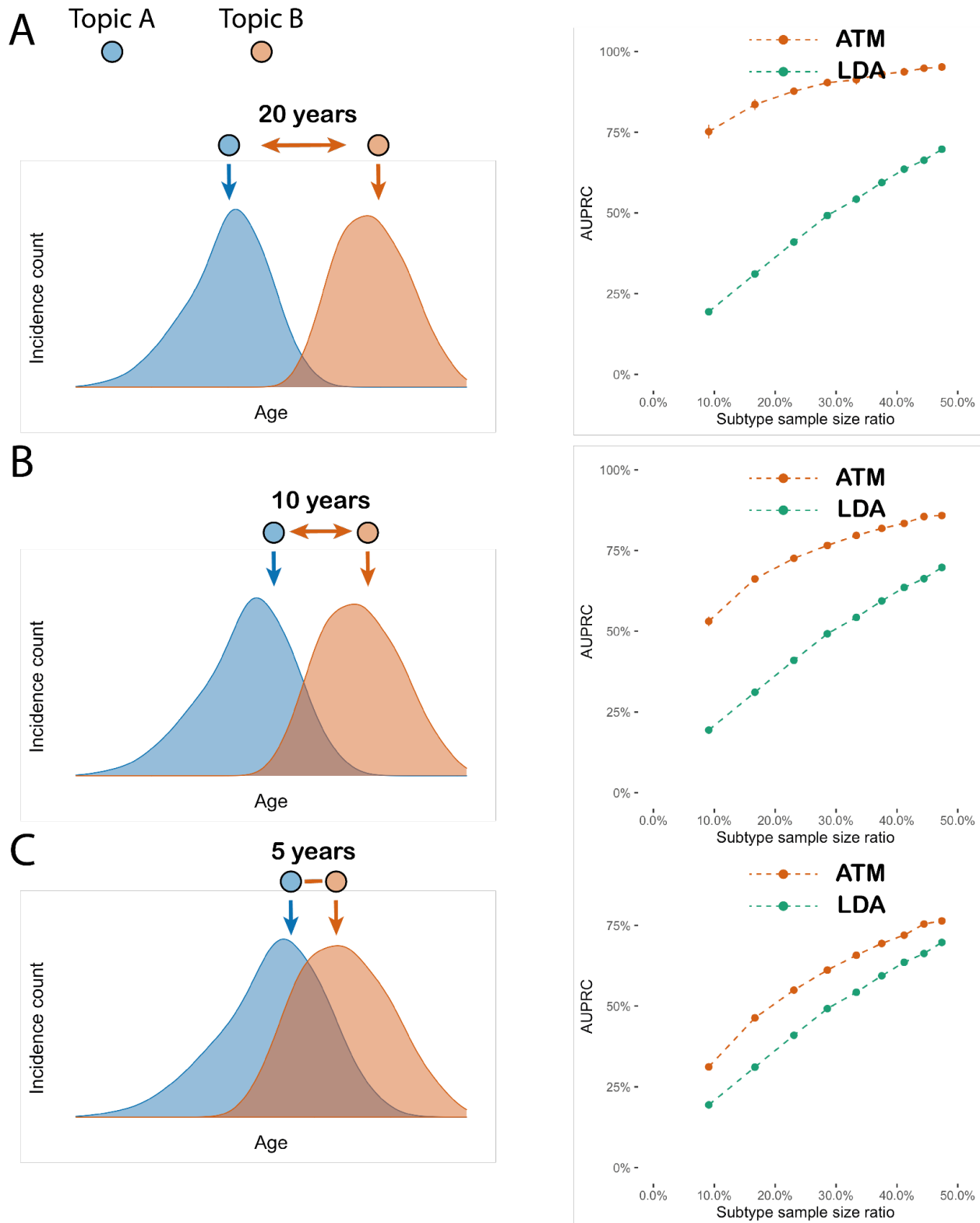
1518

1519



1520  
 1521 **Supplementary Figure 2. Additional simulation studies established the power of the**  
 1522 **method to identify comorbidity.** The precision and recall rate to correctly assign incident  
 1523 disease to correct comorbidity profiles using Latent Dirichlet Allocation (LDA) and our method  
 1524 (ATM). X-axis refers to the proportion of cases that belong to the small subgroup; precision and  
 1525 recall are computed for the label incidences in the small subgroup. Each dot represents the mean  
 1526 of 100 simulations of 10,000 people, the bar shows the 95% confidence intervals. Red refers to  
 1527 the ATM and green refers to the LDA model.(a) Scenario where two subtypes are simulated with  
 1528 20 years of difference in age at diagnosis. (b) Scenario where two subtypes are simulated with 10  
 1529 years of difference in age at diagnosis. (c) Scenario where two subtypes are simulated with 5

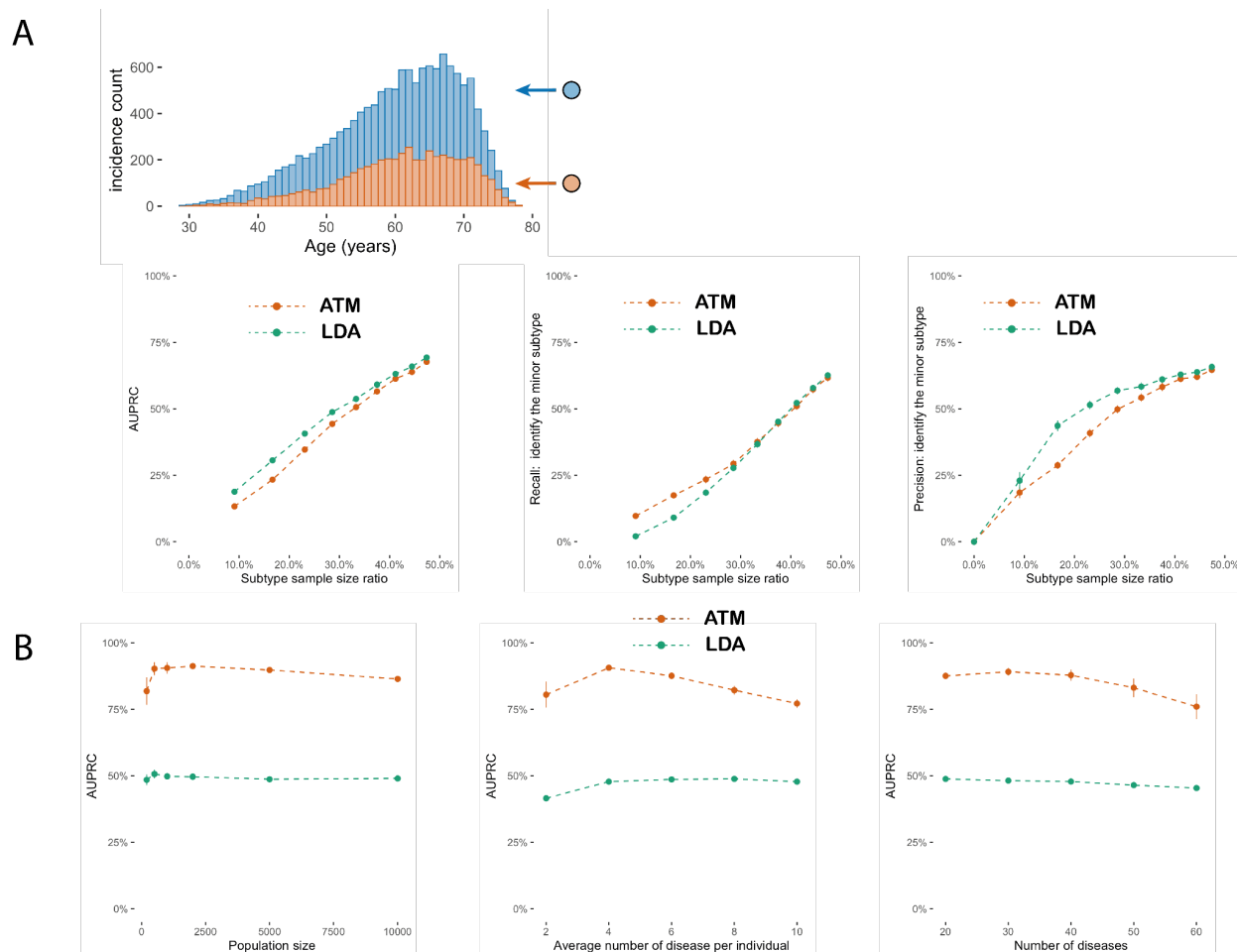
1530 years of age difference.



1531  
1532 **Supplementary Figure 3. Same analysis as in Figure 2 but simulating the smaller subtype**  
1533 **to have older age at diagnosis.** The area under precision and recall curve (AUPRC) to correctly

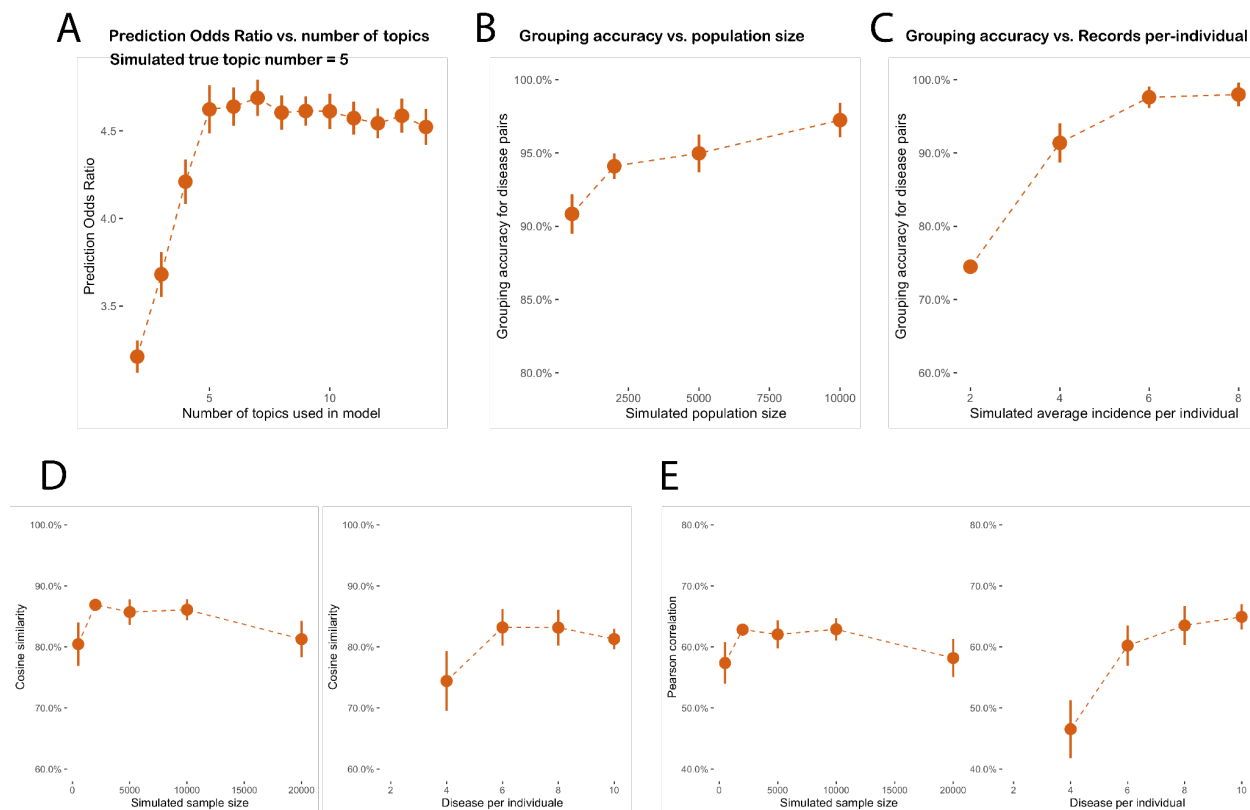


1534 assign incident disease to correct comorbidity profiles using Latent Dirichlet Allocation (LDA)  
 1535 and ATM. X-axis refers to the proportion of cases that belong to the older subtype (the orange  
 1536 subtype); precision and recall are computed for classifying the incidences in the older subgroup.  
 1537 Each dot represents the mean of 100 simulations of 10,000 people, the bar shows the 95%  
 1538 confidence intervals. In the right column red refers to the ATM and green refers to the LDA  
 1539 model. Note AUPRC is only meaningful when precision and recall pertains to classifying the  
 1540 smaller subtype, therefore we simulate with the smaller subtype taking up to 50% of cases. (a)  
 1541 Scenario where two subtypes are simulated with 20 years of difference in age at diagnosis. (b)  
 1542 Scenario where two subtypes are simulated with 10 years of difference in age at diagnosis. (c)  
 1543 Scenario where two subtypes are simulated with 5 years of age difference.  
 1544  
 1545  
 1546  
 1547



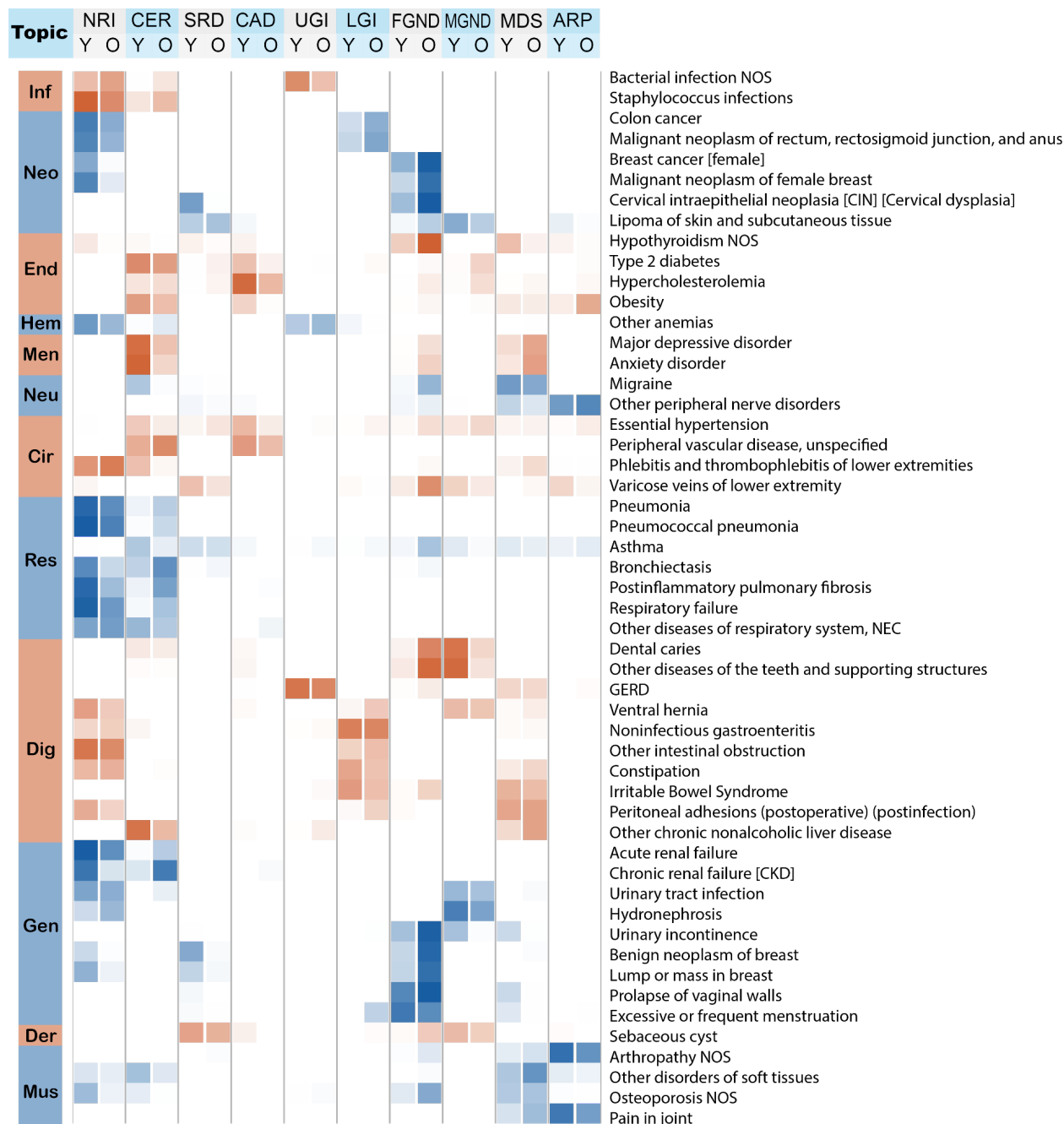
1548 **Supplementary Figure 4. Additional simulation studies established the power of the**  
 1549 **method to identify comorbidity.** (a) Same analysis as Figure 2 but simulated subtypes with  
 1550 same age at diagnosis distribution. LDA outperforms ATM slightly as we have additional  
 1551 regularisation when modelling topic loading as functions of age, while for LDA age is not  
 1552

1553 modelled. (b) AUPRC computed as in Figure 2A with varying population size, average number  
1554 of diseases per individual, and number of distinct diseases. Each dot shows the mean of 20  
1555 simulations and the bar shows 95% confidence interval.  
1556  
1557  
1558



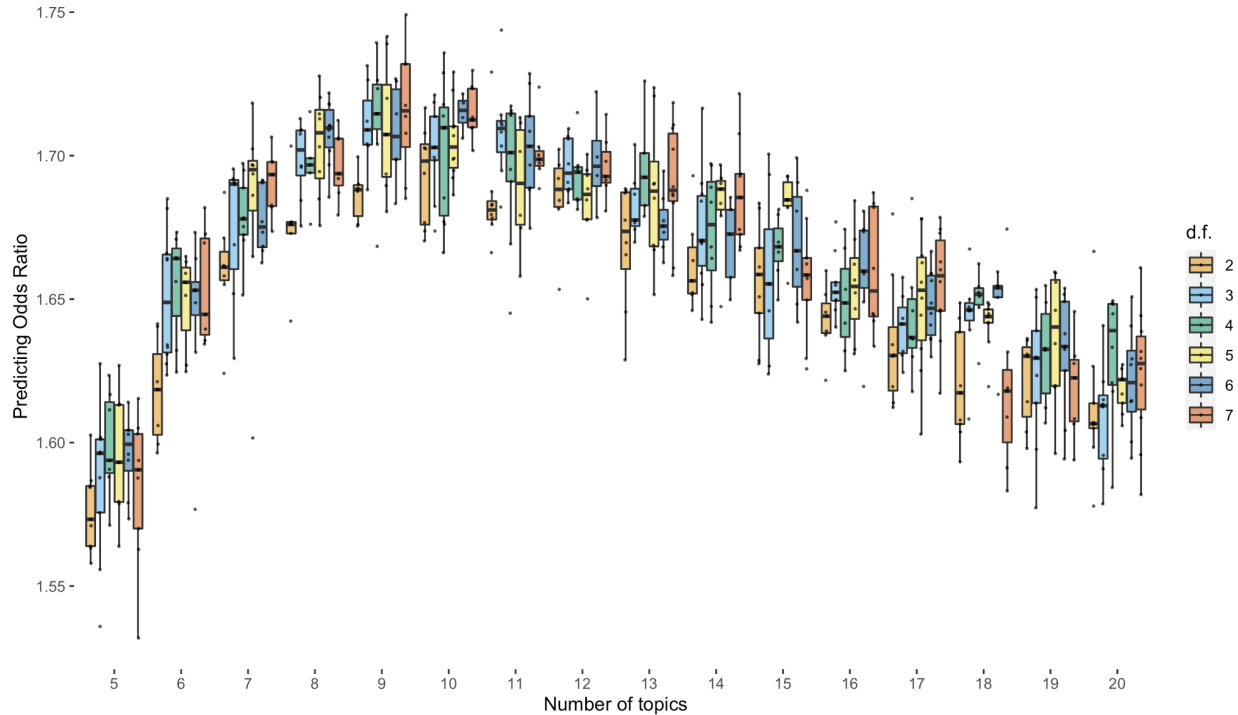
1559 **Supplementary Figure 5. Simulations confirming that ATM could accurately recover topic**  
1560 **loadings and topic weights.** (a) We simulated data using 5 topics while fitting models of  
1561 varying topic numbers. To compute prediction odds ratios (see Methods), we used 80% of data  
1562 as training data to fit ATM and computed prediction odds ratio in the held out data, where we  
1563 use the topic loading computed from the training data and prior diseases to infer the topic  
1564 weights to predict the target diseases. The simulation was performed for 20 replications for each  
1565 topic number in the inference. (b-c) We assign each disease to a single topic based on topic  
1566 loading and compute the grouping accuracy as the proportion of disease pairs that are correctly  
1567 grouped to the same topic. The grouping accuracy remains high for varying simulated population  
1568 size and average disease per individual. (d) Recovery of topic loadings. We evaluate the  
1569 accuracy of topic loading inference by computing the cosine similarity between inferred topic  
1570 loading with the underlying truth. We match the inferred topics with the true topics using  
1571 correlation of topic weights, using a greedy procedure (matching the first inferred topic from all  
1572 true topics and then matching the next topic from the remaining not-matched true topics) to  
1573 ensure the matching is bijectively. (e) Recovery of topic weights. We evaluate the accuracy of

1575 topic weight inference by computing the correlation of inferred topic weights and with the  
 1576 underlying truth. The ordering of topics uses the same strategy as in panel d.  
 1577  
 1578



1579  
 1580 **Supplementary Figure 6. Posterior topic distributions are different between age groups for**  
 1581 **diseases that have subtypes.** The figure has the same legends as Figure 3A but focusing on 52  
 1582 diseases that have a subtype with at least 500 incidences.  
 1583

1584



1585

1586 **Supplementary Figure 7. Prediction odds ratio across different model configurations.** Each

1587 dot represents one inference on a random training and testing split of the UK Biobank

1588 individuals. The models are run with different topic numbers and parametric configurations of

1589 topic loadings. Degrees of freedom (d.f.) from 2 to 7 represent linear, quadratic polynomial,

1590 cubic polynomial, spline with one knot, spline with two knots, and spline with three knots. The

1591 prediction odds ratios are computed on the testing data using topic loadings inferred from the

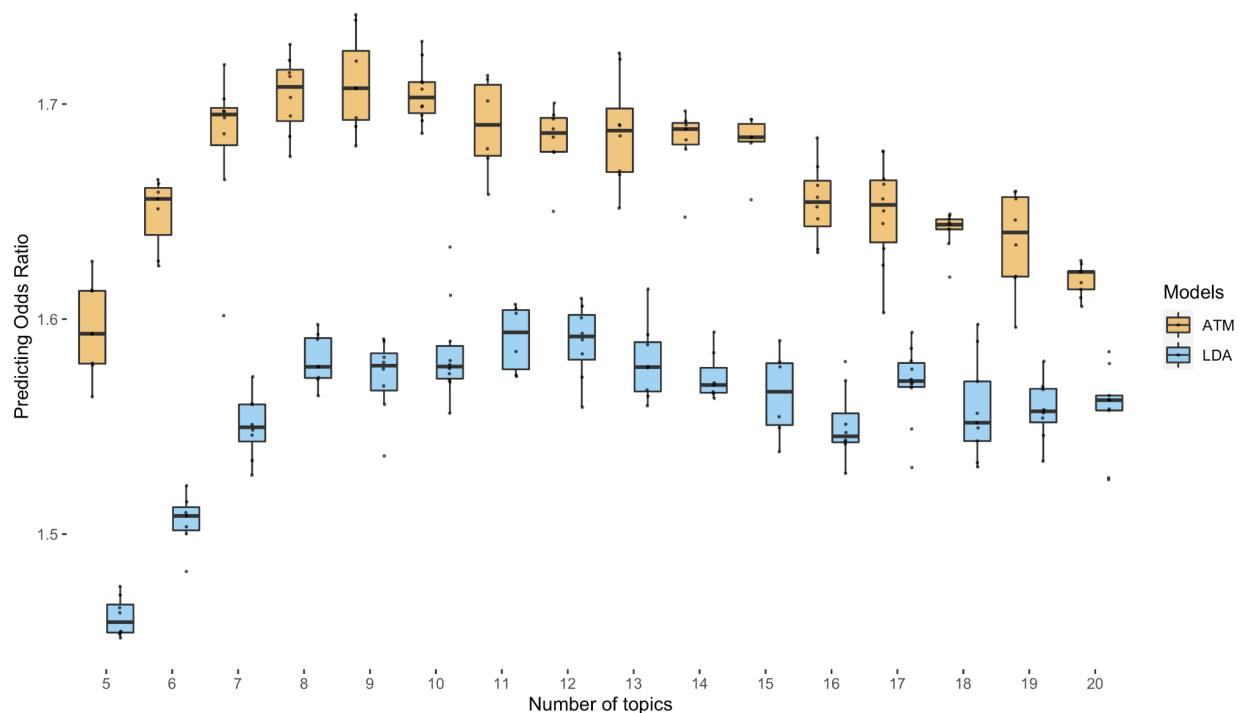
1592 training data and topic weights inferred using previous diseases of testing individuals. The odds

1593 ratios are between the odds that target diseases are within model-predicted top percentile disease

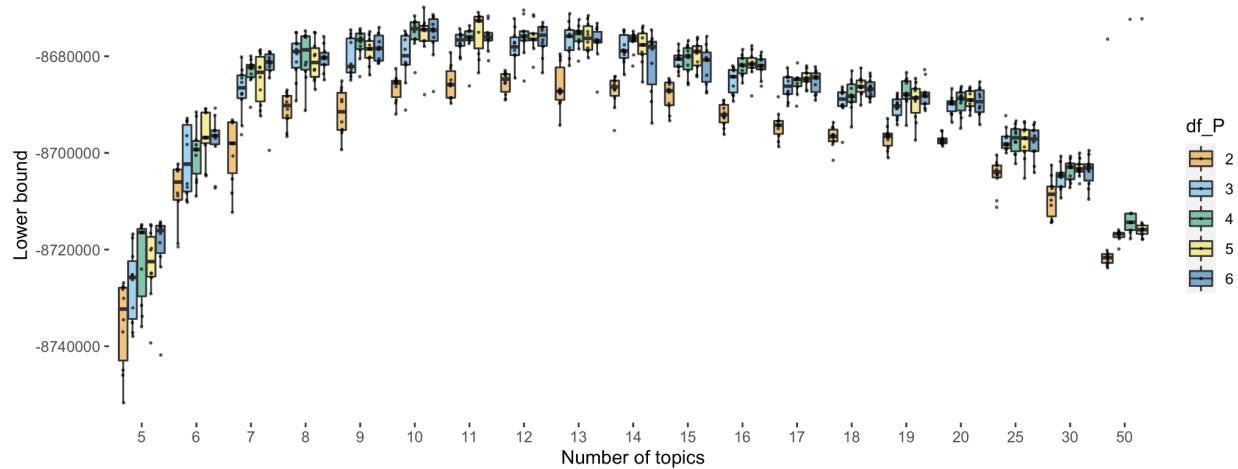
1594 set versus the odds that target diseases are within the prevalence-ordered top percentile disease

1595 set.

1596

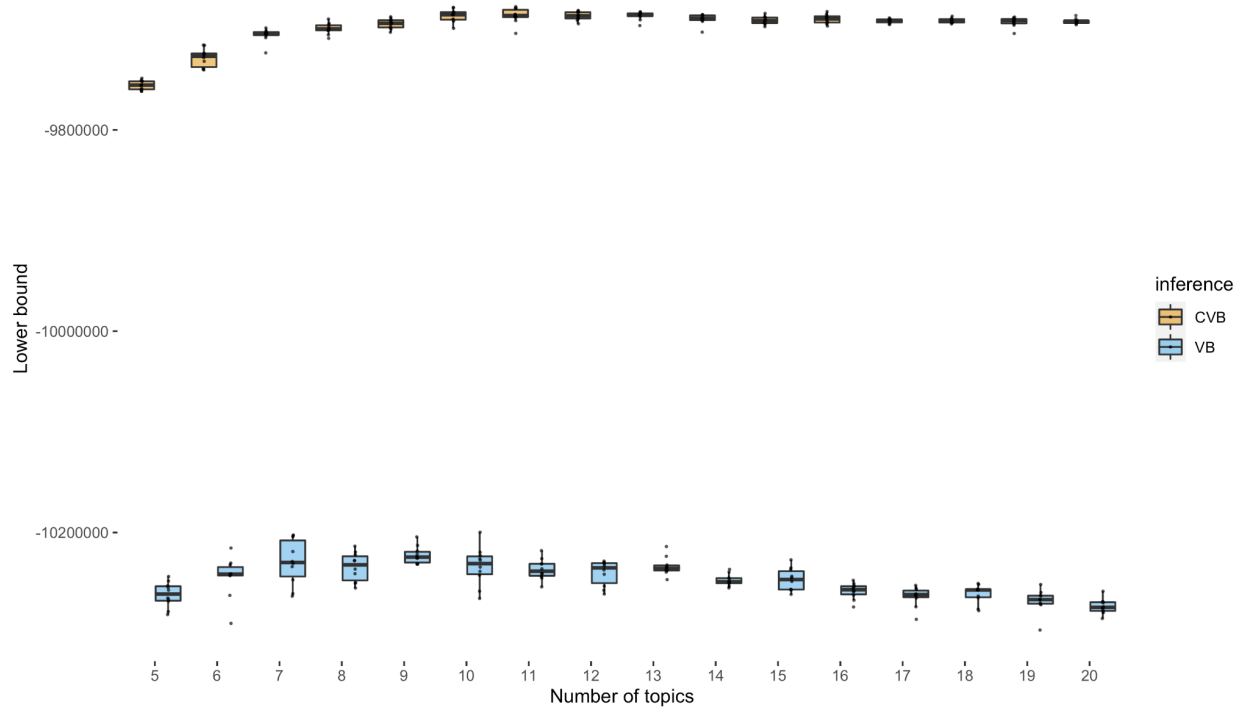


1597  
1598 **Supplementary Figure 8. Comparison of prediction odds ratio between LDA and ATM.**  
1599 Each dot represents results from running either ATM or LDA on the same random training and  
1600 testing split. The models were run with different topic numbers and we chose a cubic spline with  
1601 one knot for configuring ATM topic loadings. The prediction odds ratios are computed on the  
1602 testing data using topic loadings inferred from the training data and topic weights inferred using  
1603 previous diseases of testing individuals. The odds ratios are between the odds that target diseases  
1604 are within model-predicted top percentile disease set versus the odds that target diseases are  
1605 within the prevalence-ordered top percentile disease set. [For the optimal model with 10 topics,](#)  
1606 [ATM has an average prediction odds ratio 1.71 \(across 10 random training-testing splits\); LDA](#)  
1607 [has an average prediction odds ratio 1.58 \(across 10 random training-testing splits\).](#)  
1608  
1609  
1610



1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618

**Supplementary Figure 9. Evidence lower bound (ELBO) of different model configurations on the entire dataset.** Each dot represents one inference on a random training and testing split of the UK Biobank individuals. The models are run with different topic numbers and parametric configurations of topic loadings. Degrees of freedom (d.f.) from 2 to 7 represent linear, quadratic polynomial, cubic polynomial, spline with one knot, spline with two knots, and spline with three knots.

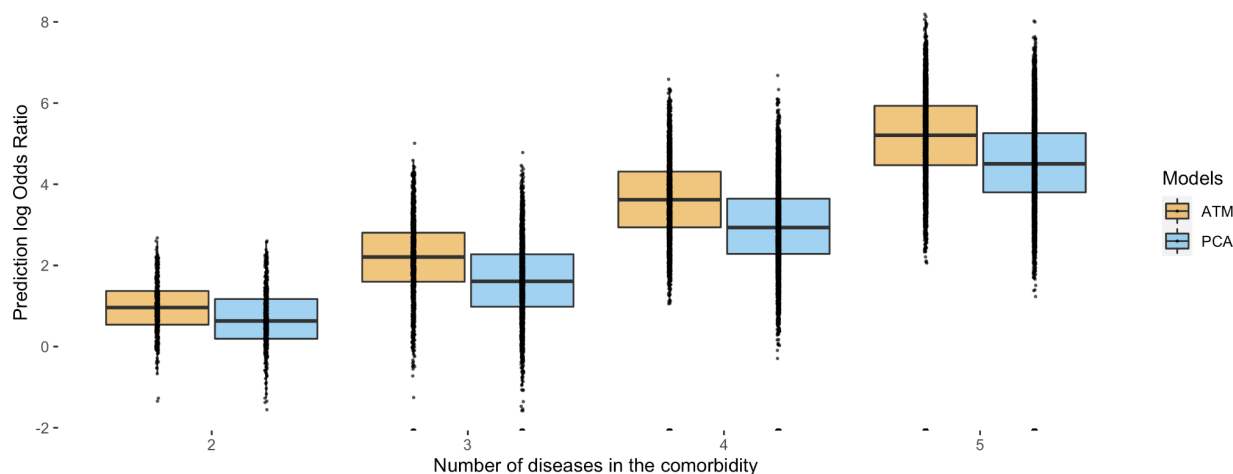


1619  
1620  
1621  
1622  
1623

**Supplementary Figure 10. Comparison of ELBO for collapsed variational inference and mean-field variational inference.** ELBO is computed by fitting the ATM using two inference methods on the entire UK Biobank dataset, where the topic loadings are configured as cubic polynomials. Models of different numbers of topics are fitted with 10 random initialisations for

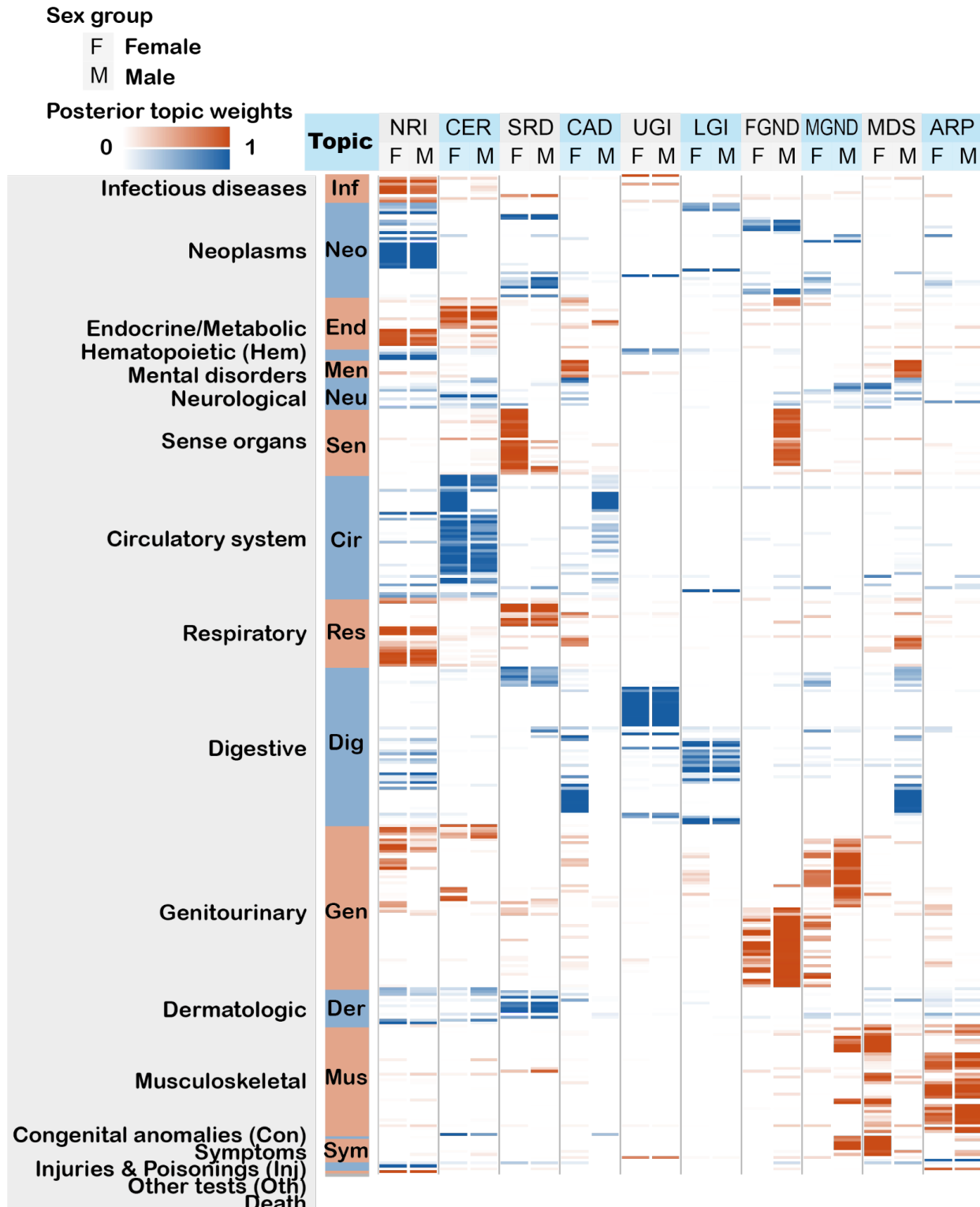
1624 both CVB and the VB (mean-field variational inference, which is a more commonly used  
1625 inference method for Bayesian models). The ELBO of an inference methods is a lower bound  
1626 that approximates the evidence function, which depends on the number of topics and parametric  
1627 form of topic loading, but not the inference methods; higher ELBO means better inference  
1628 accuracy.

1629  
1630



1631  
1632 **Supplementary Figure 11. Prediction log odds ratio of comorbidities.** Diseases combinations  
1633 (comorbidities) are extracted from topics loadings that are trained on a training set, using max  
1634 value of average topic assignments (Methods). Roughly, the odds ratio of each disease  
1635 combination is computed by selecting all disease sets containing combinations of 2, 3, 4, and 5  
1636 diseases assigned to the same topic by max topic loading, and dividing incidences where the  
1637 disease sets appeared in one patient by the expected number in an independent testing set. We  
1638 show the comparison of ATM and PCA for all combinations of 2, 3, 4, and 5 diseases; here we  
1639 use PCA as we wish to show the superiority of topic modelling in identifying clusters of disease  
1640 compared to other low-rank methods that are not based on multinomial distribution.

1641  
1642

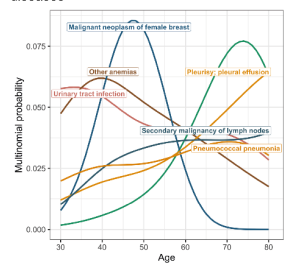


1643  
 1644  
 1645  
 1646  
 1647

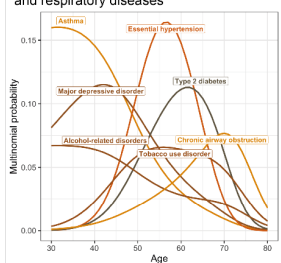
**Supplementary Figure 12: Posterior topic distributions of female and male populations.**  
 The figure is the same as Figure 3A but comparing the topics that are inferred from female and male populations separately.



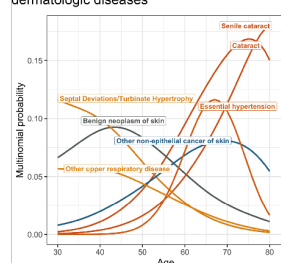
NRI: Neoplasm, respiratory, and infectious diseases



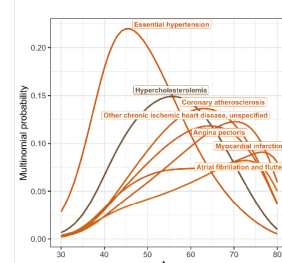
CER: circulatory system, endocrine/metabolic, and respiratory diseases



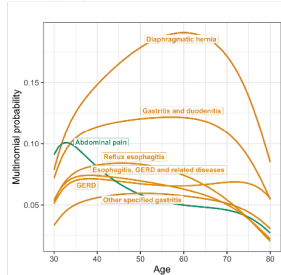
SRD: Sense organ, respiratory, and dermatologic diseases



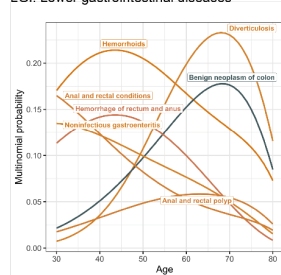
CAD: circulatory system diseases



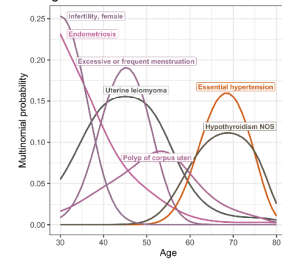
UGI: Upper gastrointestinal diseases



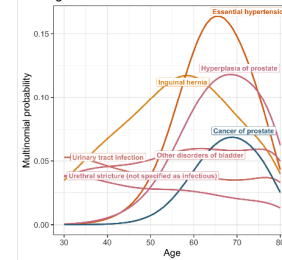
LGI: Lower gastrointestinal diseases



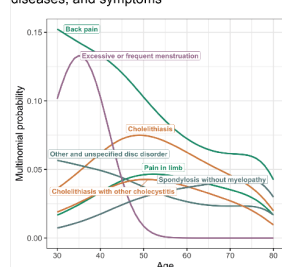
FGND: Female genitourinary, neoplasms, and digestive diseases



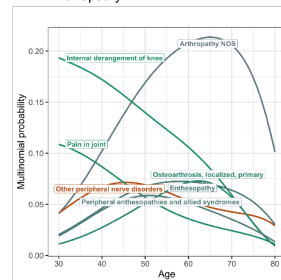
MGND: Male genitourinary, neoplasms, and digestive diseases



MDS: Musculoskeletal diseases, digestive diseases, and symptoms



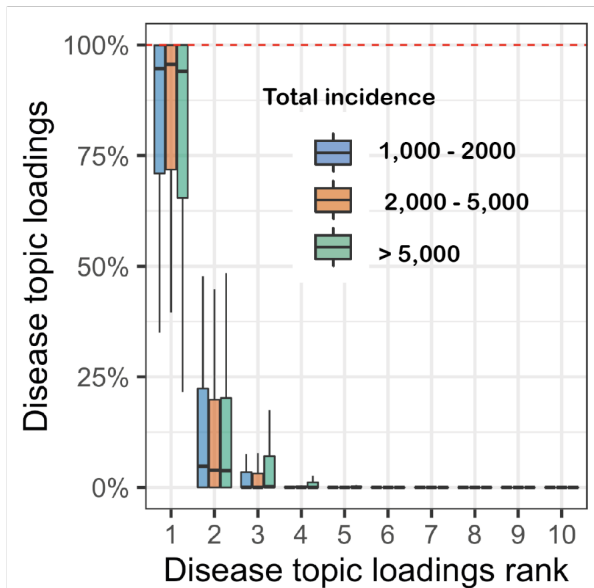
APR: Arthropathy



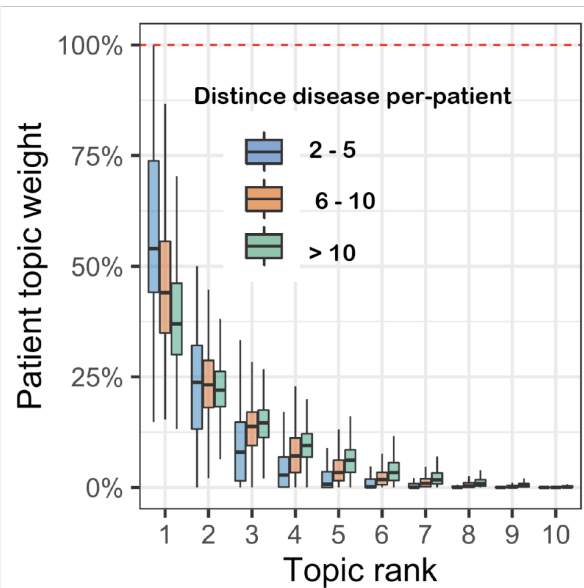
1648  
1649  
1650  
1651  
1652  
1653

**Supplementary Figure 13. Top seven diseases in each comorbidity topic.** Seven diseases that have highest loading within the topic are shown for each comorbidity topic. Colour of the curves reflect the ordering of Phecodes. We chose seven for best visual presentation. Numerical results are reported in Supplementary Table 5.

**A** Distribution of topic weights for diseases



**B** Distribution of topic weights for patients



1654

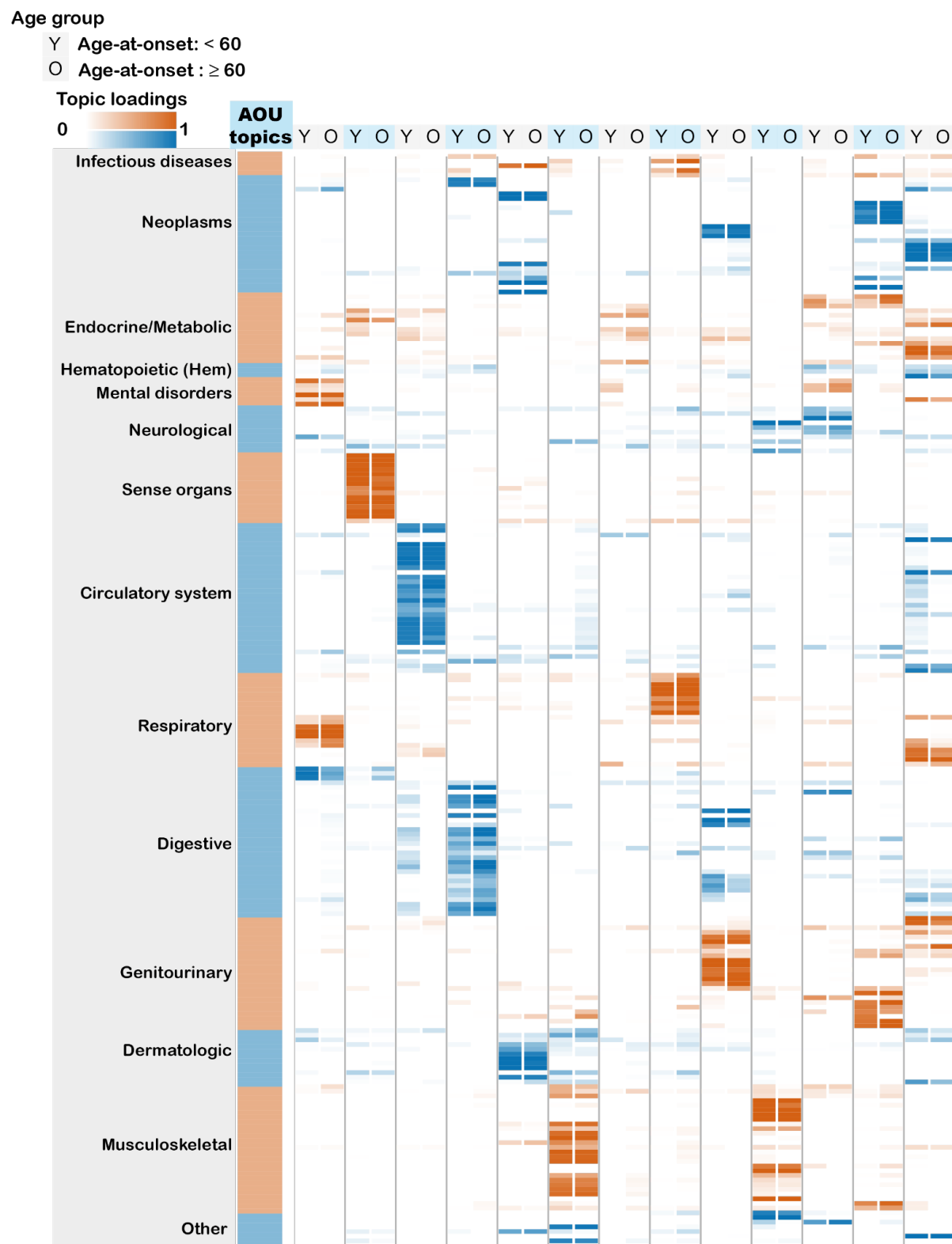
1655

1656

1657

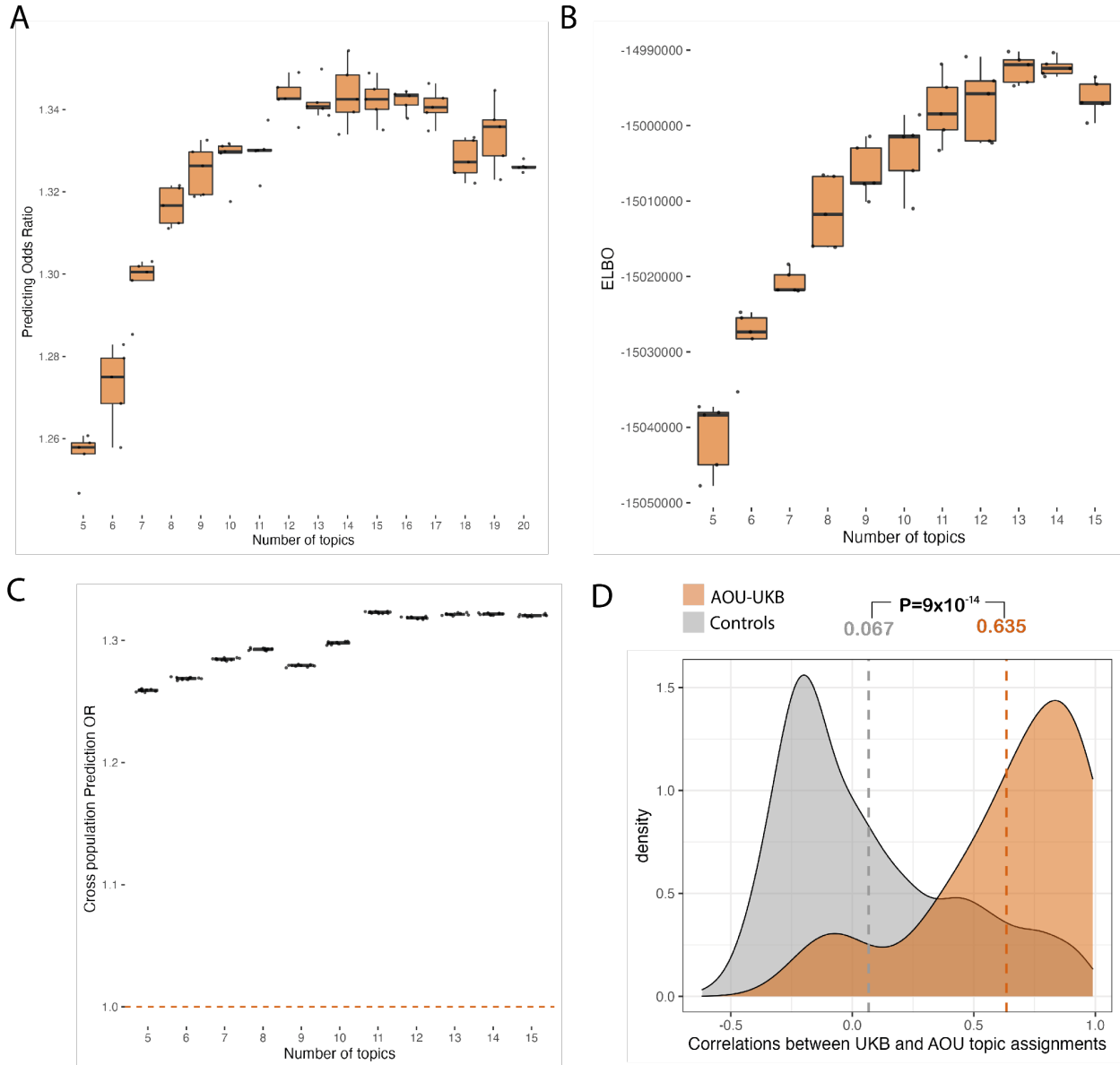
1658

**Supplementary Figure 14. Additional topic sparsity analysis.** (a) Sparsity of disease topic loadings. Box plot shows the distribution of topic loading for disease of different incidence numbers. (b) Sparsity of patient topic weights. Box plot shows the topic weight distribution in decreasing order for individuals with different numbers of diagnosis.



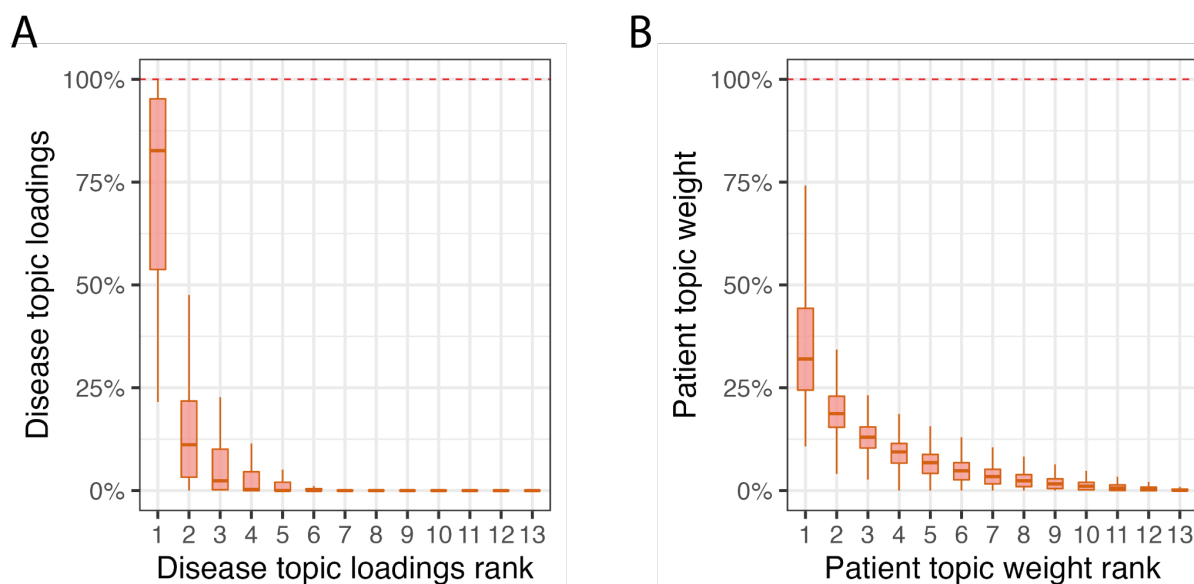
1659  
 1660 **Supplementary Figure 15. Age-dependent topic loadings of 13 inferred disease topics across**  
 1661 **233 diseases in the All of Us.** We report topic loadings averaged across younger ages (age at  
 1662 diagnosis < 60) and older ages (age at diagnosis > 60). Row labels denote disease categories

1663 ordered by Phecode systems, with alternating blue and red color for visualisation purposes;  
 1664 “Other” is a merge of five Phecode systems: “congenital anomalies”, “symptoms”, “injuries &  
 1665 poisoning”, “other tests”, and “death” (which is treated as an additional disease, see Methods).  
 1666 Topics are ordered by the corresponding Phecode system. This figure is an All of Us equivalent  
 1667 of Figure 3.  
 1668

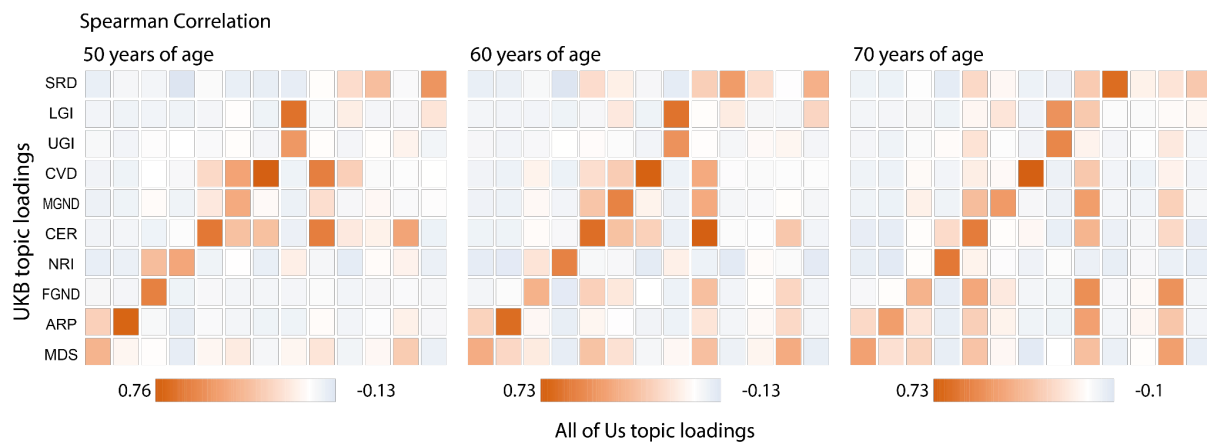


1669 **Supplementary Figure 16. ATM infers disease topics from All of Us cohort which align**  
 1670 **with topics from UK Biobank.** (A) Prediction odds ratio using ATM model with different topic  
 1671 numbers in All of Us. Each dot represents one of the five-fold cross validation within the All of  
 1672 Us individuals. (B) Evidence lower bound (ELBO) of different ATM model configurations on  
 1673 the entire All of Us dataset. Each dot represents one inference with random initialization. The  
 1674 models are run with different topic numbers and same configurations of topic loadings (spline  
 1675

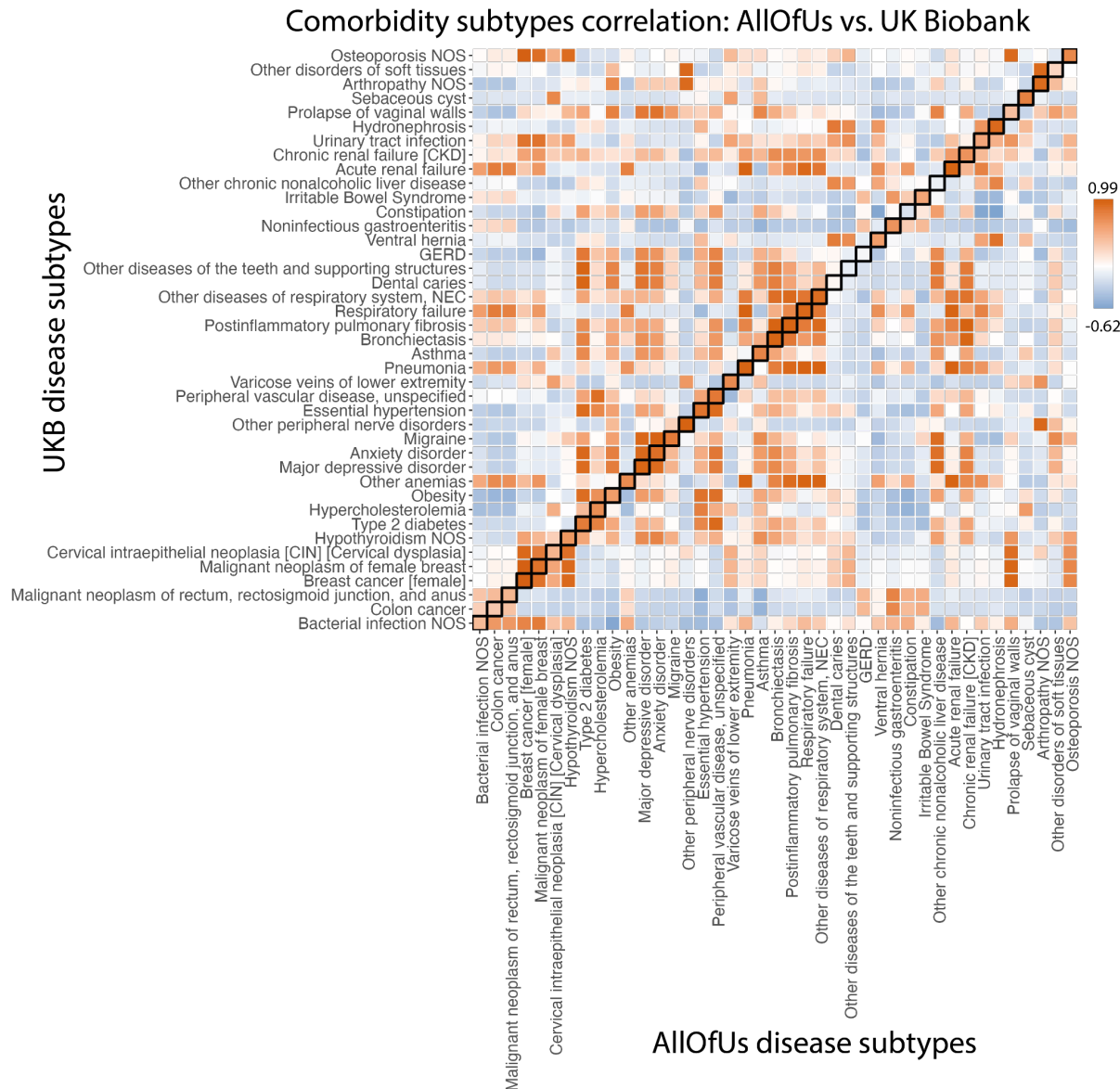
1676 model with one knot). (C) Prediction odds ratio on UK Biobank individuals using All of Us topic  
1677 loadings. We divide the UKBiobank population into 10 jackknife blocks and each dot represents  
1678 the prediction odds ratio on one leave-one-out jackknife sample. Topic weights are inferred using  
1679 prior diseases of UKBB individuals, using loadings trained from All of Us. The odds ratios are  
1680 between the odds that target diseases are within model-predicted top two-percentile disease set  
1681 versus the odds that target diseases are within the prevalence-ordered top two-percentile disease  
1682 set. Prediction odds ratio is 1.32 (s.e. = 0.0027) when using the optimal 13 All of Us topic to  
1683 predict UK Biobank diagnoses. We chose the top-two percentile to match the UK Biobank  
1684 analysis as All of Us has 233 of the 348 diseases analysed in the UK Biobank. (D) Correlations  
1685 between UKB and AOU topic assignments for 41 diseases with subtypes between AOU and  
1686 UKB (red shade) are significantly higher than expected (grey shade). The correlation between 41  
1687 AOU-UKB disease pairs are reported in Table 2 and Supplementary figure 19. Grey shade is the  
1688 distribution of non-diagonal correlations in Supplementary Figure 19. Grey and red vertical  
1689 dashed line reports the mean of the grey and red shades; P-value is for the difference of the  
1690 mean.  
1691



1692 **Supplementary Figure 17. Distribution of topic loading across diseases and topic weights**  
1693 **across patients for All of Us.** (A) Box plot of disease topic loading as a function of rank;  
1694 disease topic loadings are computed as a weighted average across all values of age at diagnosis.  
1695 (B) Box plot of patient topic weight as a function of rank. This figure is an All of Us equivalent  
1696 of Figure 4B-C.  
1697  
1698

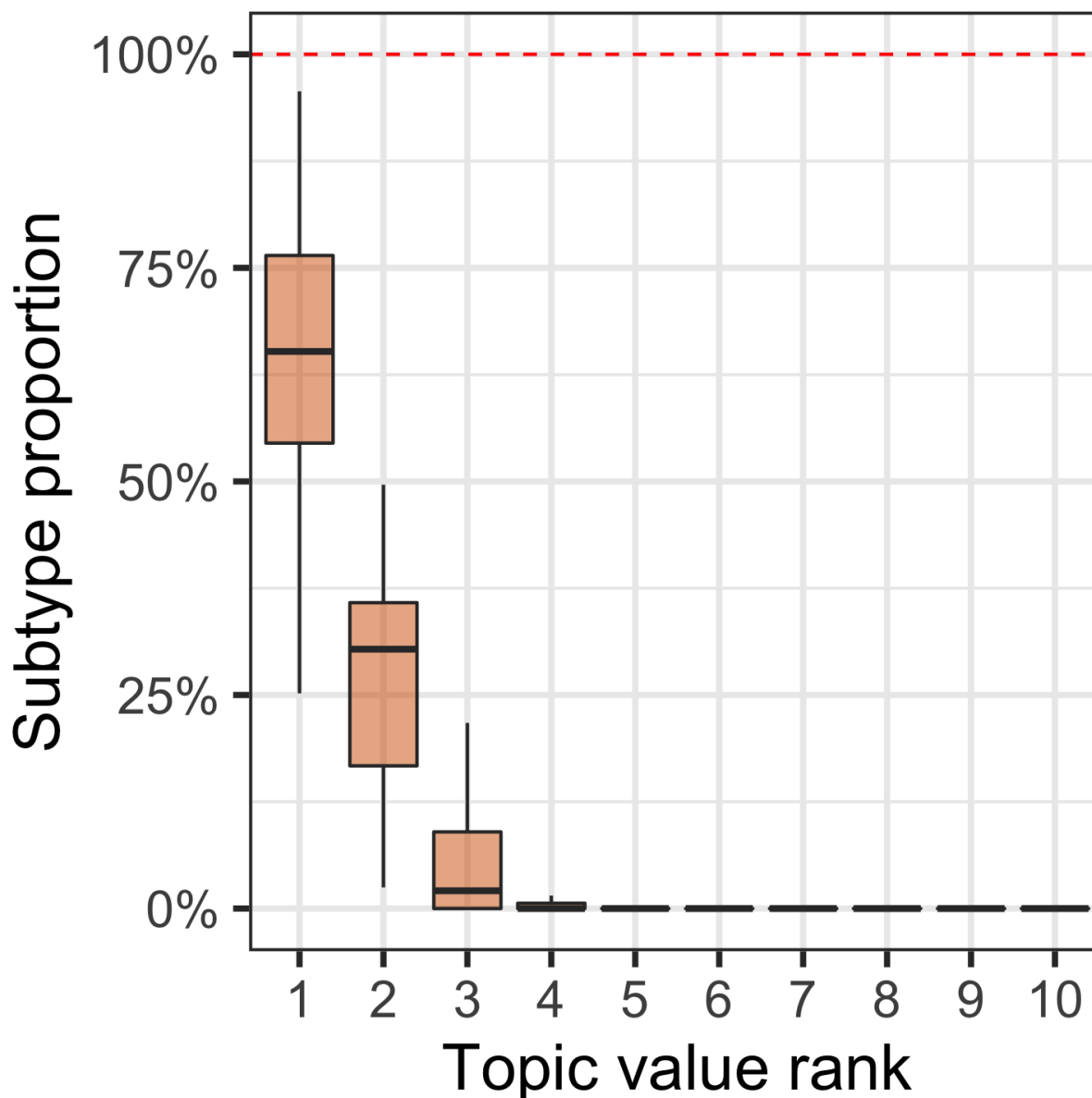


1699  
1700 Supplementary Figure 18. correlation between topic loadings from UK Bibank (y-axis) and All  
1701 of Us (x-axis) for three age slices. The figures are the age-specific versions for Figure 5C.  
1702



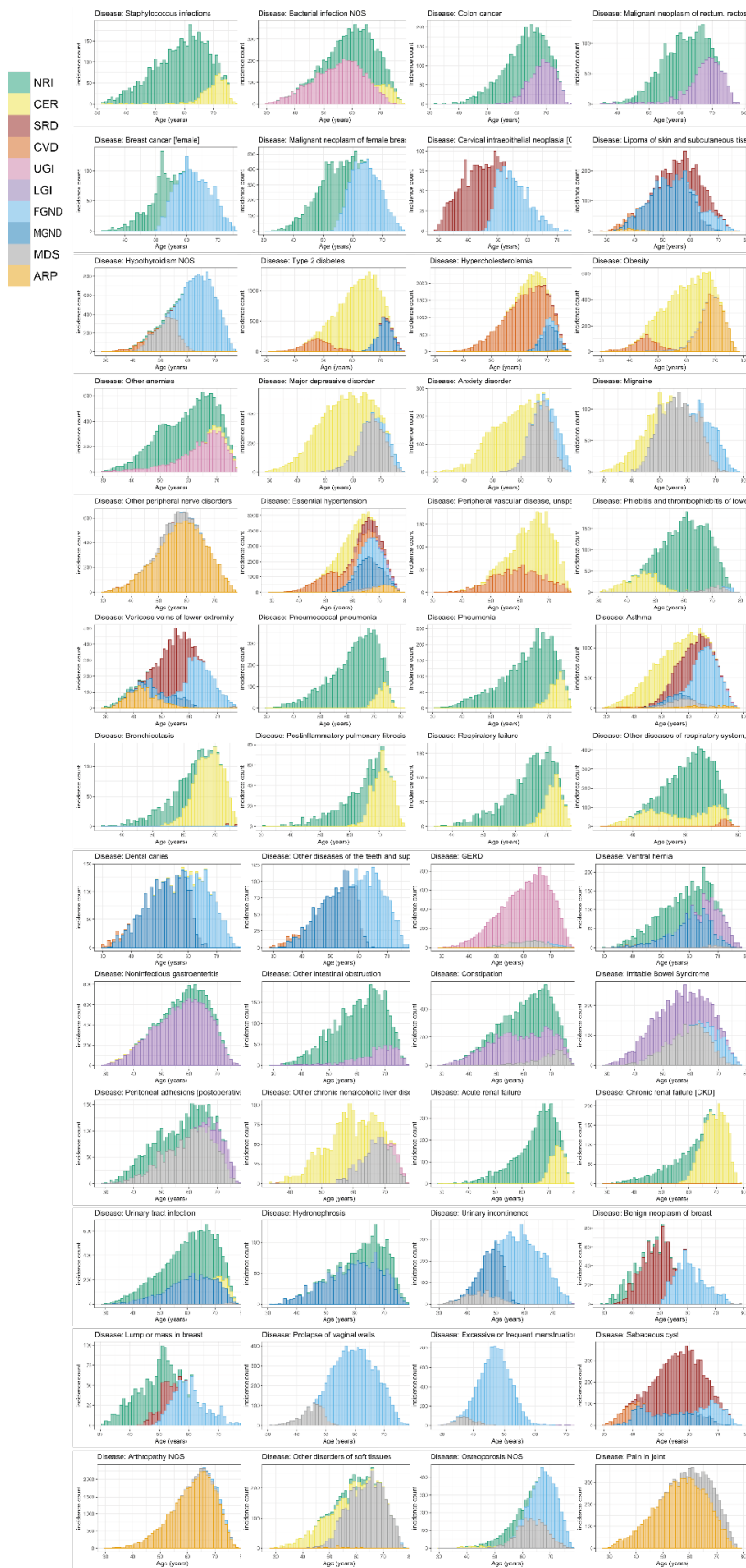
1703  
 1704 **Supplementary Figure 19. Subtype correlations between UK Biobank and All of Us for 41**  
 1705 **diseases that are presented in both datasets and have subtypes in UK Biobank.** Each box of  
 1706 the heatmap shows the correlation of average diagnosis-specific topic probability between a  
 1707 disease from All of Us and the other disease from UK Biobank. The diagnosis-specific topic  
 1708 probabilities from All of Us were mapped to UK Biobank based on proportional variance  
 1709 between the two topic spaces (Methods).

1710  
 1711  
 1712  
 1713

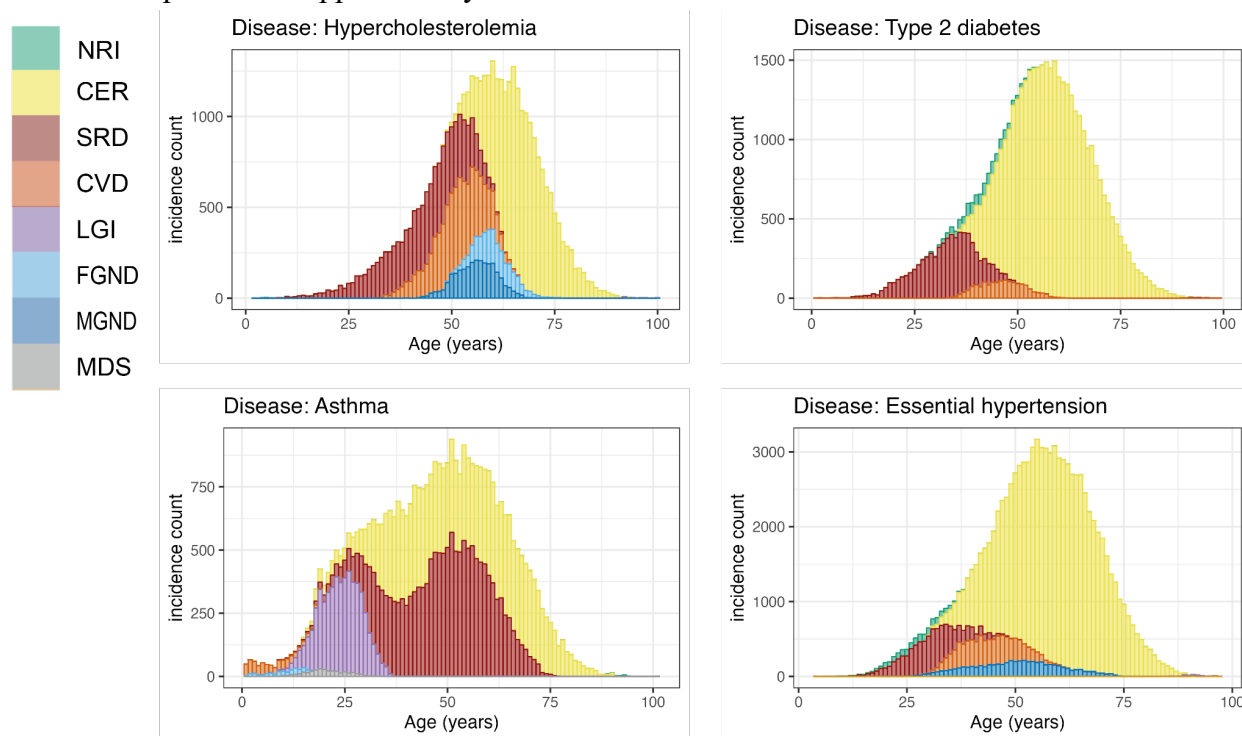


1714  
1715 **Supplementary Figure 20. Topic distribution for the 52 diseases that have at least 500 cases**  
1716 **assigned to distinct topics.** For each disease of the 52 diseases, we computed the proportion of  
1717 diagnoses assigned to each subtype. The box plot shows the distribution of the subtype  
1718 proportion from the largest (leftmost boxes) to the smallest. For nearly all diseases, the cases are  
1719 concentrated into three subtypes, with very few cases assigned to other topics.



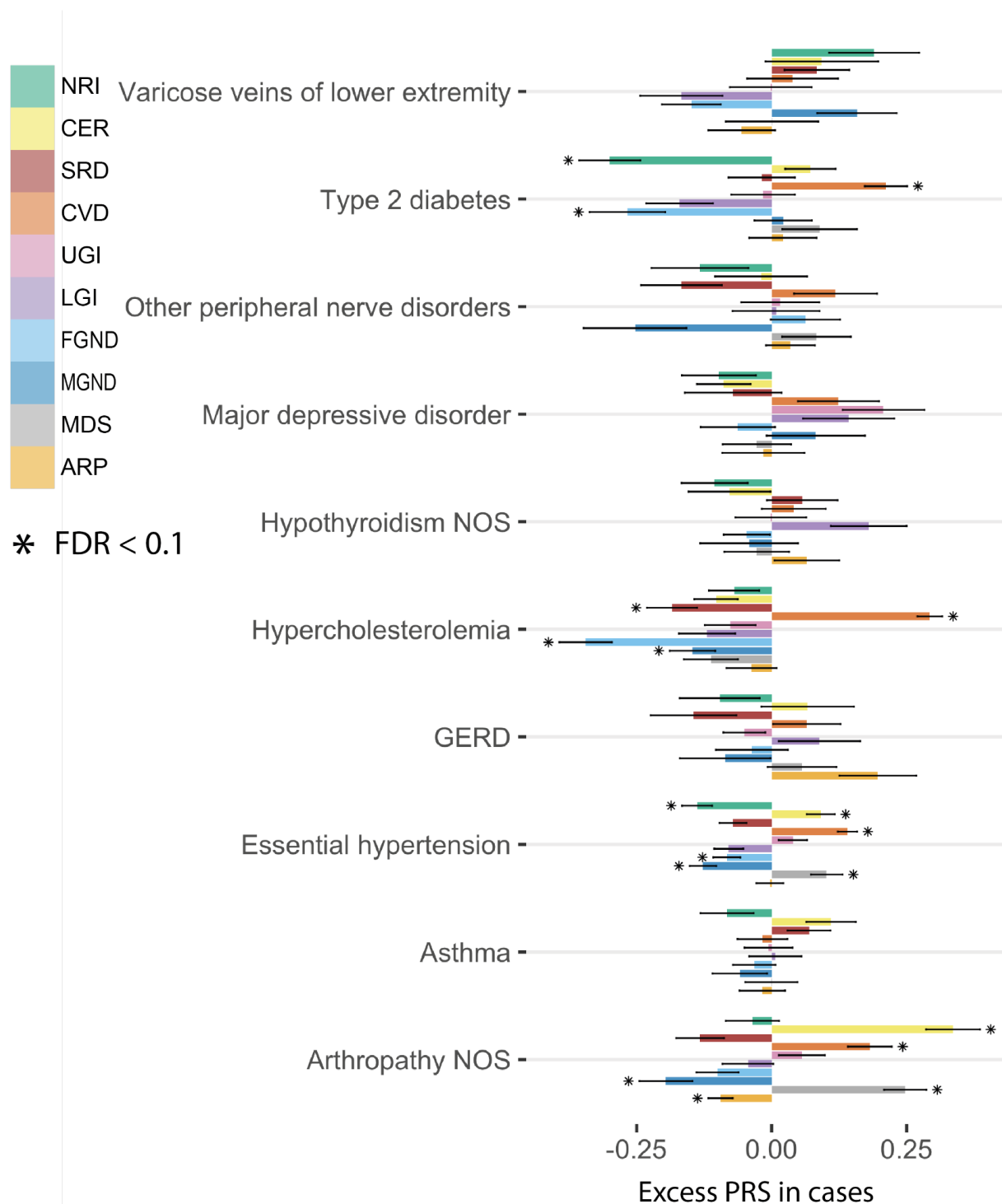


1721 **Supplementary Figure 21. Comorbidity subtype distribution over age for 52 diseases.**  
1722 Diseases shown are ordered by 13 phecode systems: infectious diseases, neoplasms,  
1723 endocrine/metabolic, hematopoietic, mental disorders, neurological, circulatory system,  
1724 respiratory, digestive, neoplasms, genitourinary, dermatologic, and musculoskeletal. Numerical  
1725 results are reported in Supplementary Table 10-11.

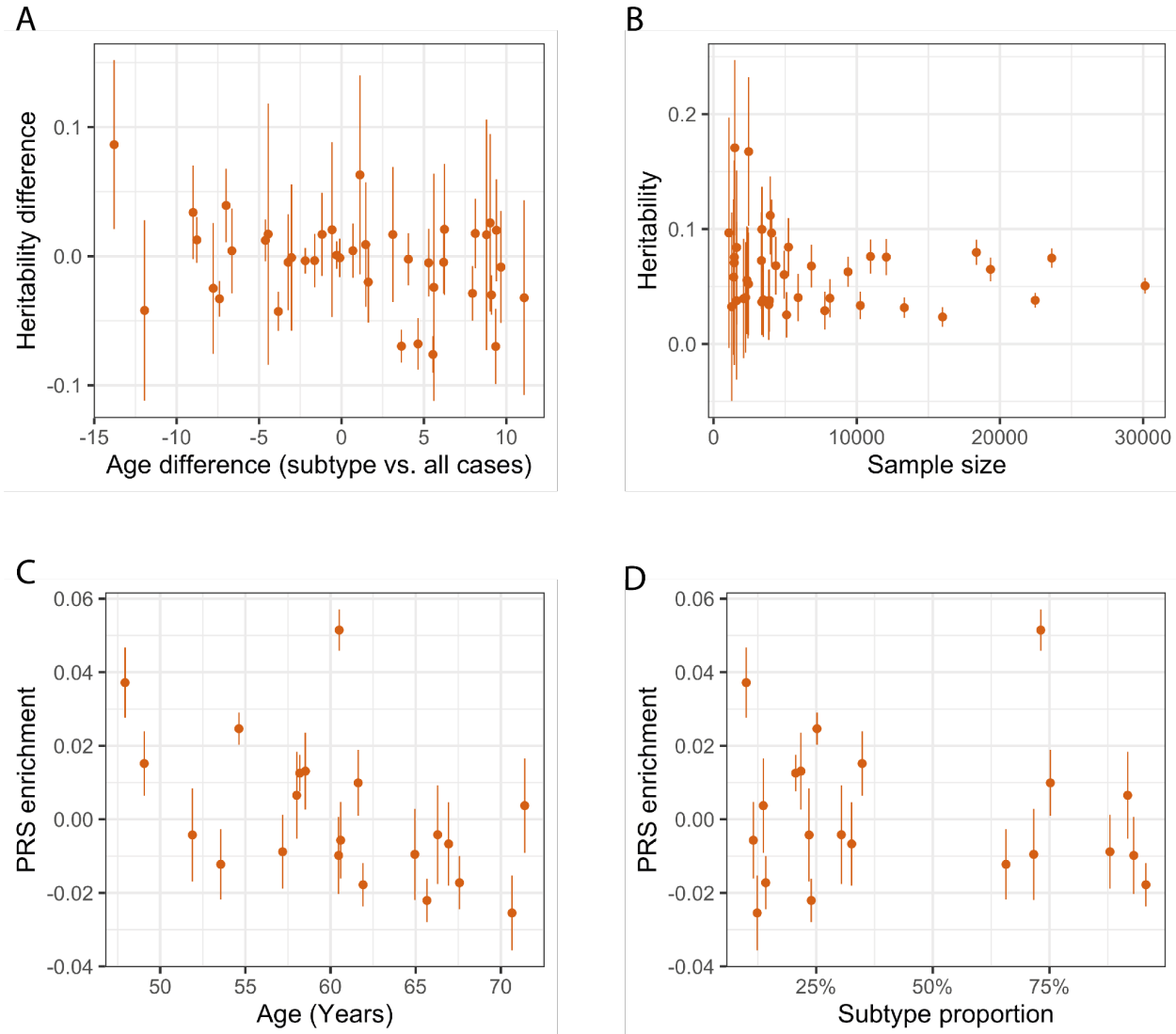


1726 **Supplementary Figure 22. Stacked barplots of age-dependent subtypes in All of Us.** Disease  
1727 topics in All of Us are mapped to their most similar UK Biobank topics; colours are the same as  
1728 Supplementary Figure 21. The figures are for 4 representative diseases in Figure 6A (type 2  
1729 diabetes, asthma, hypercholesterolemia, and essential hypertension); for each disease, we include  
1730 all subtypes with at least one diagnosis.  
1731

1732  
1733  
1734  
1735  
1736  
1737  
1738



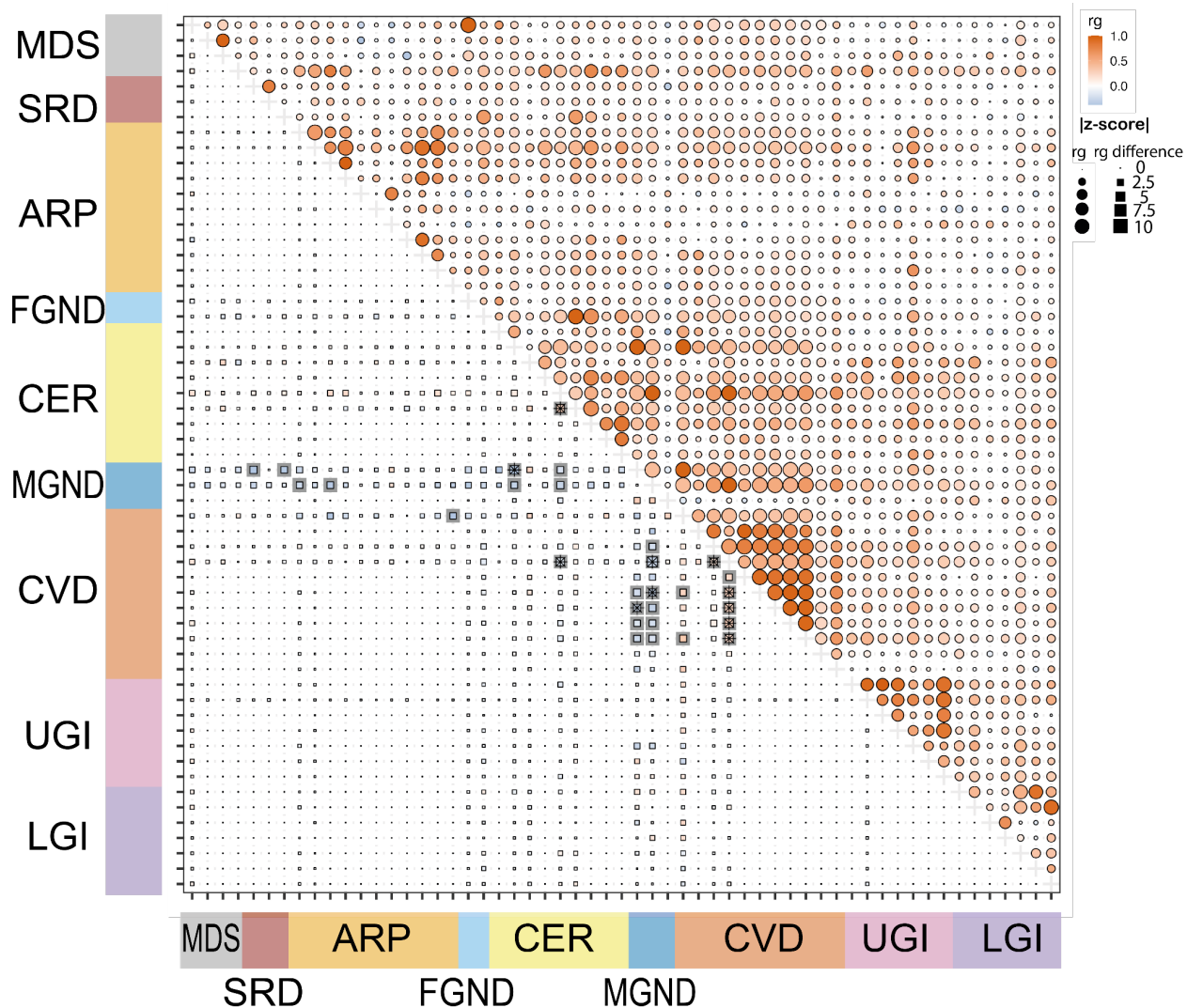
1739  
 1740 **Supplementary Figure 23. Excess PRS analysis for all topics across 10 diseases (selected by**  
 1741 **heritability z-score).** The bar plot shows the estimated changes in s.d. of PRS per unit changes  
 1742 in the patient topic weight in disease cases. The PRS is estimated using all the cases in British  
 1743 Isle Ancestry. The stars show disease-topic pairs that are significant at FDR = 0.05. Numerical  
 1744 results are reported in Supplementary Table 8.  
 1745



1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754

**Supplementary Figure 24. Heritability and PRS are not associated with age and subtype size.** (a) We plot heritability deviation from all cases versus age deviation from all cases of 41 subtypes (from 14 diseases that have heritability z-score above 5). The heritability is estimated by first performing mixed-effect association analysis using BOLT-LMM on imputed SNPs from the British Isle Ancestry then using LDSC. (b) Heritability for the subtypes plotted with the sample size of the subtype. (c-d) Excess PRS from Figure 6B plotted against the age and the sample size (denoted by the ratio of samples between subtype and all cases) for the subtypes. The dots and the bars show the mean and 95% confidence interval across all subfigures.

1755



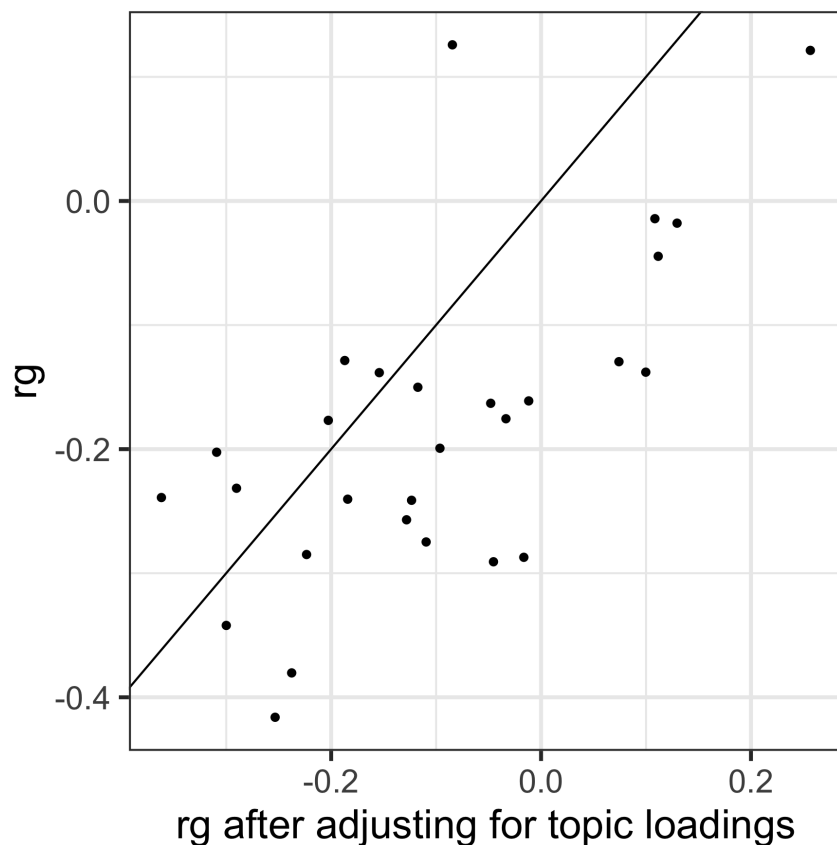
1756

1757

1758 **Supplementary Figure 25. Excess genetic correlation ( $r_g$ ) between disease subtypes across**  
 1759 **disease subtypes.** Lower left panel shows the  $r_g$  of disease-subtype or subtype-subtype pairs  
 1760 subtracted by the  $r_g$  of corresponding disease-disease pairs. Each row or column represents a  
 1761 disease or subtype. For disease-disease pairs, the excess  $r_g$  is not defined in the lower left panel,  
 1762 since the difference is 0. The upper right panel shows  $r_g$  of corresponding disease-disease pairs,  
 1763 where values could be duplicated as the same disease could have multiple subtypes. The 89  
 1764 diseases or subtypes are chosen here by heritability z-score > 4 and  $r_g$  z-score > 4 with at least  
 1765 one other disease or subtypes. We kept 57 rows and columns for better visualisation by  
 1766 removing 32 diseases that have subtypes included. A star means FDR < 0.1, while a shade means  
 1767 a nominal statistical significance at P = 0.05.

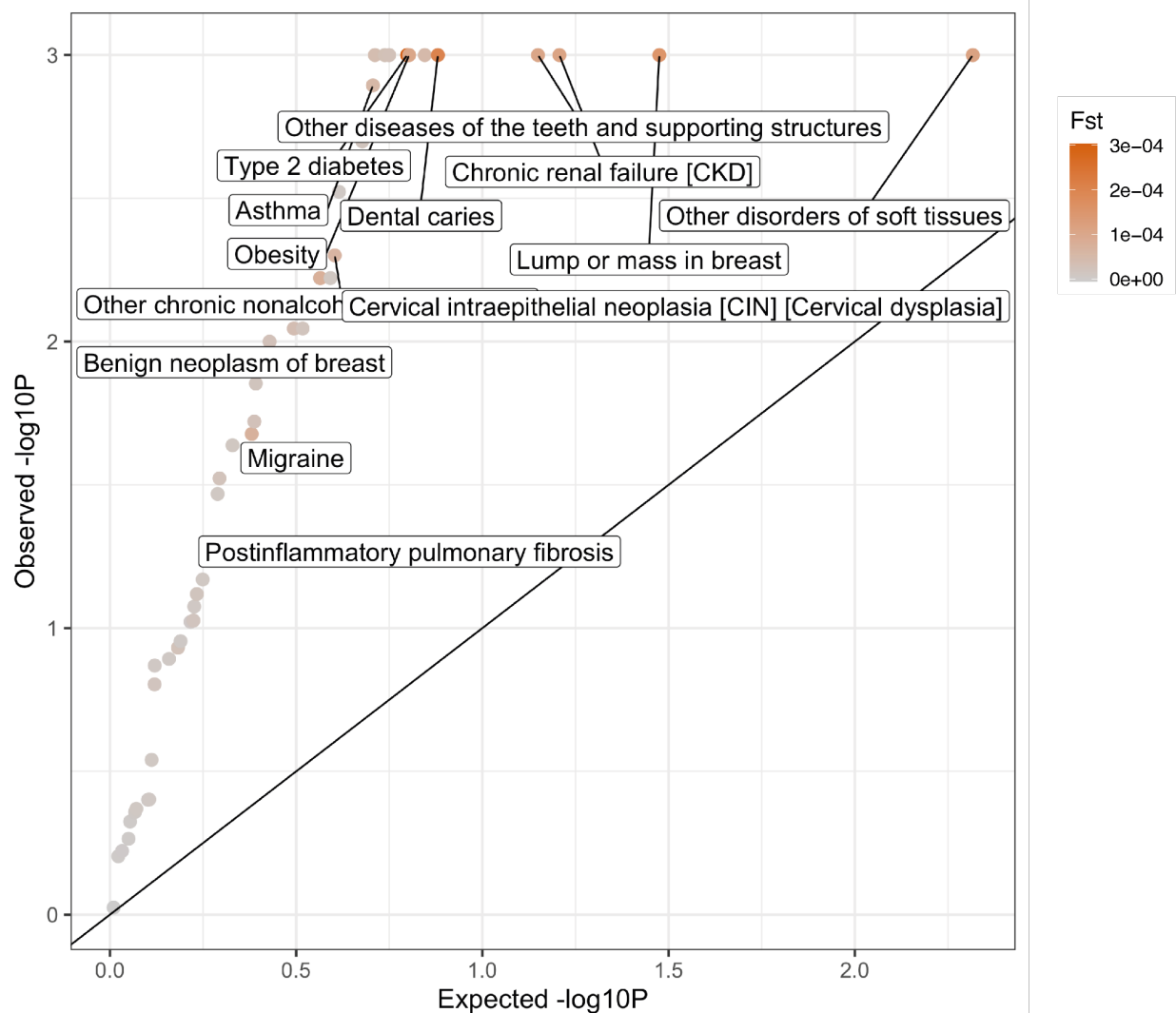
1768

1769

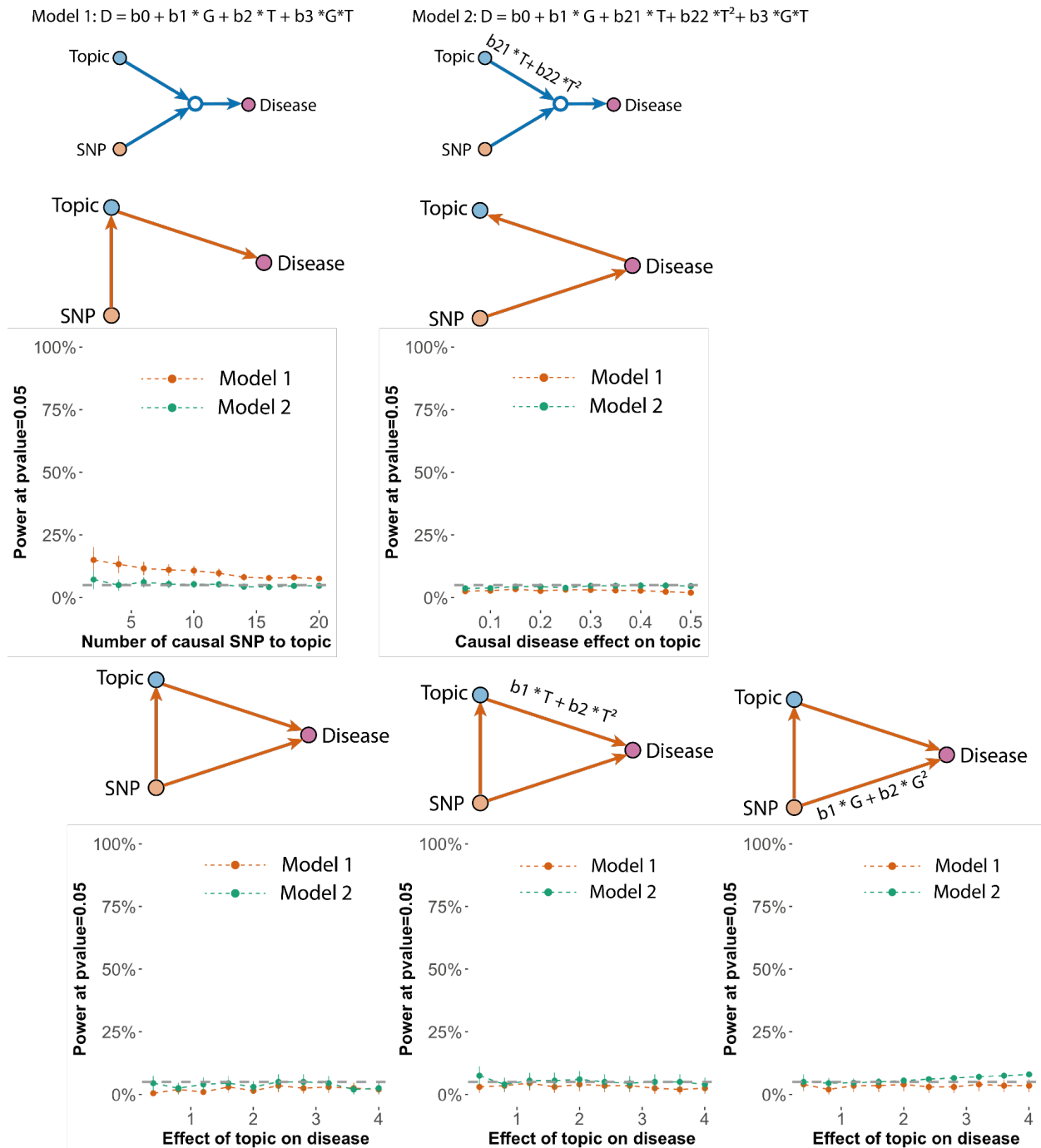


1770  
1771

1772 **Supplementary Figure 26. Comparison of excess genetic correlation ( $r_g$ ) between subtypes**  
1773 **using summary statistics from case-control matched by topic weights (x-axis) and not**  
1774 **matched by topic weights.** The analysis is performed on subtypes of four diseases: type 2  
1775 diabetes, hypercholesterolemia, hypertension, and asthma. The excess  $r_g$  (shown in panel (A) of  
1776 Figure 7) of two case-control matching strategies across 28 subtype pairs are compared and the  
1777 effect lies along the diagonal line. The excess genetic correlation attenuated slightly when topic  
1778 weights are matched, while it can not explain all the excess  $r_g$ .



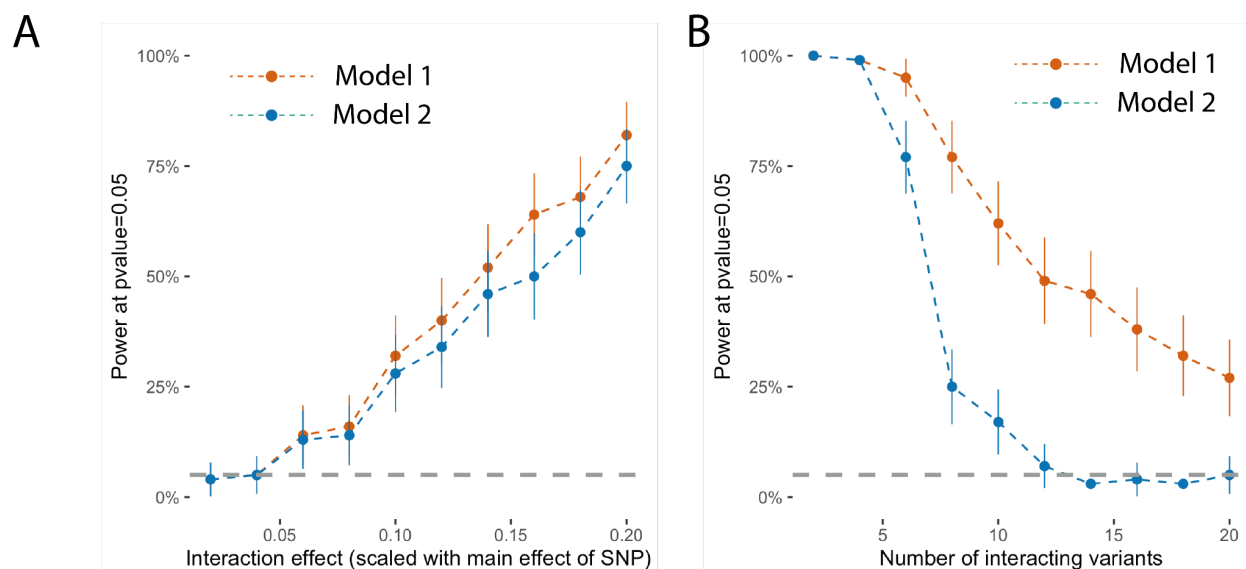
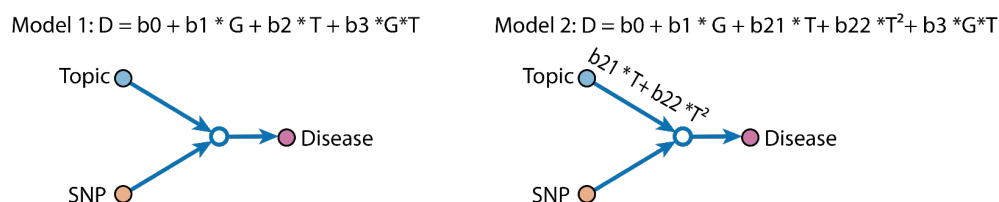
1779  
1780 **Supplementary Figure 27. Excess  $F_{ST}$  of disease subtypes compared with controls with**  
1781 **matched topic weights.** P-values are for testing case- $F_{ST}$  significantly higher than controls of  
1782 similar topic weight distribution. The permutation controls are sampled for 1,000 times with the  
1783 same topic weights distribution and sample size to the disease subtypes. We focus on 49 of the  
1784 52 diseases which have more than one subgroup of at least 500 cases. Subtypes are defined based  
1785 on the max value of the diagnosis-specific topic probability. Three diseases (“hypertension”,  
1786 “hypercholesterolemia”, and “arthropathy”) are excluded as there are not enough controls that  
1787 match the topic weights of cases. The colour shows the value of  $F_{ST}$  across subtypes.  
1788  
1789



1790  
 1791 **Supplementary Figure 28. Simulation analysis verified the SNP x topic interaction tests are**  
 1792 **calibrated when no actual interaction exists.** We show the false positive rate of two models  
 1793 under five simulated model structures where no actual interaction exists (Methods). The false  
 1794 positive rate is computed as the power to detect interaction effects using model 1 (red; linear  
 1795 model) and model 2 (green; with non-linear main effect term) under P-value=0.05. The five  
 1796 structures evaluated are (1) SNP causal to topic and topic causal to disease; (2) SNP causal to  
 1797 disease and disease causal to topic; (3) SNP is causal to both topic and disease; (4) and (5) SNP  
 1798 is causal to both topic and disease with nonlinear effects. Genotypes are simulated using the

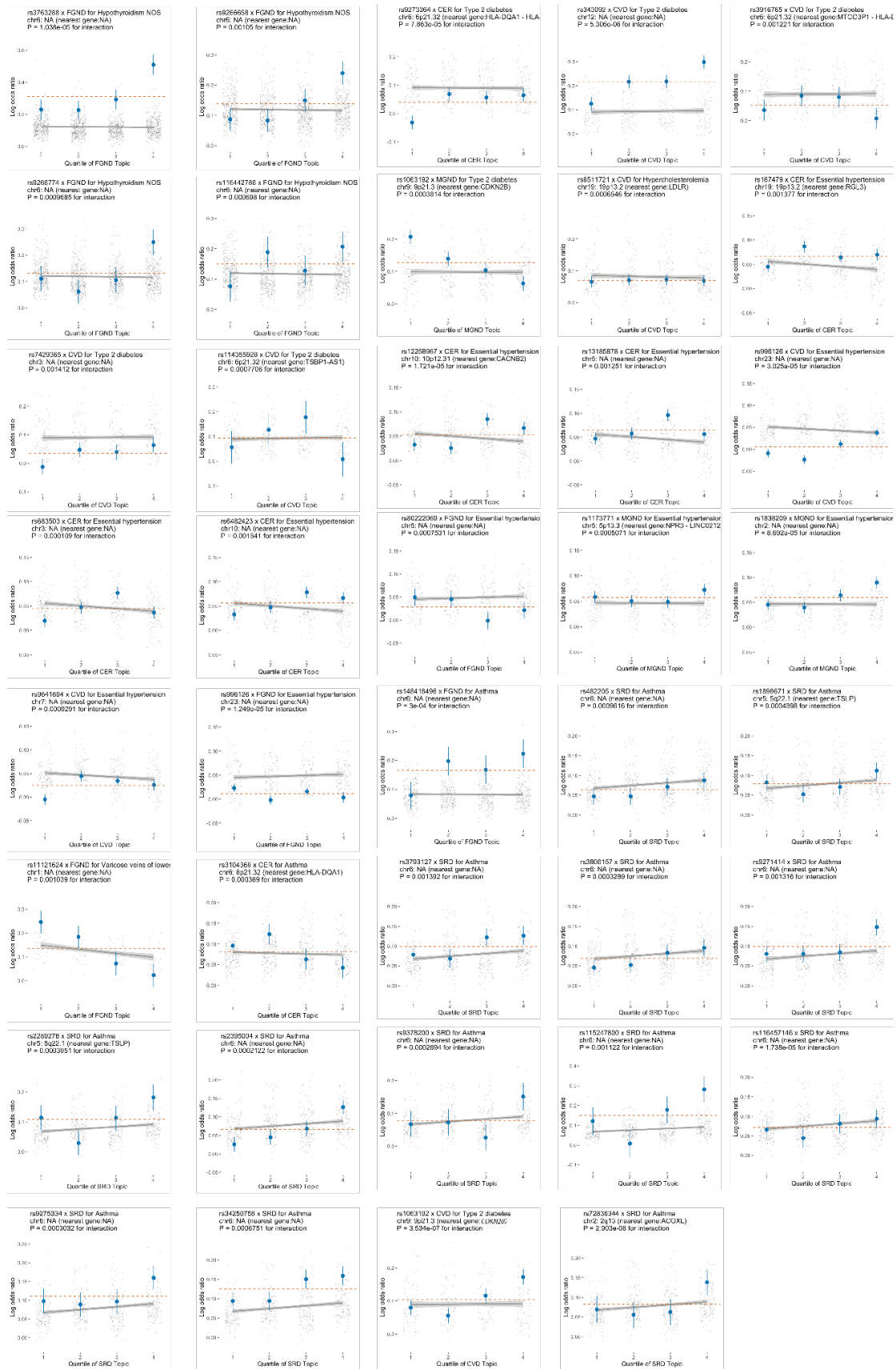


1799 MAF from the 888 disease associated SNPs that were analysed in the SNPxTopic interaction  
 1800 tests.  
 1801

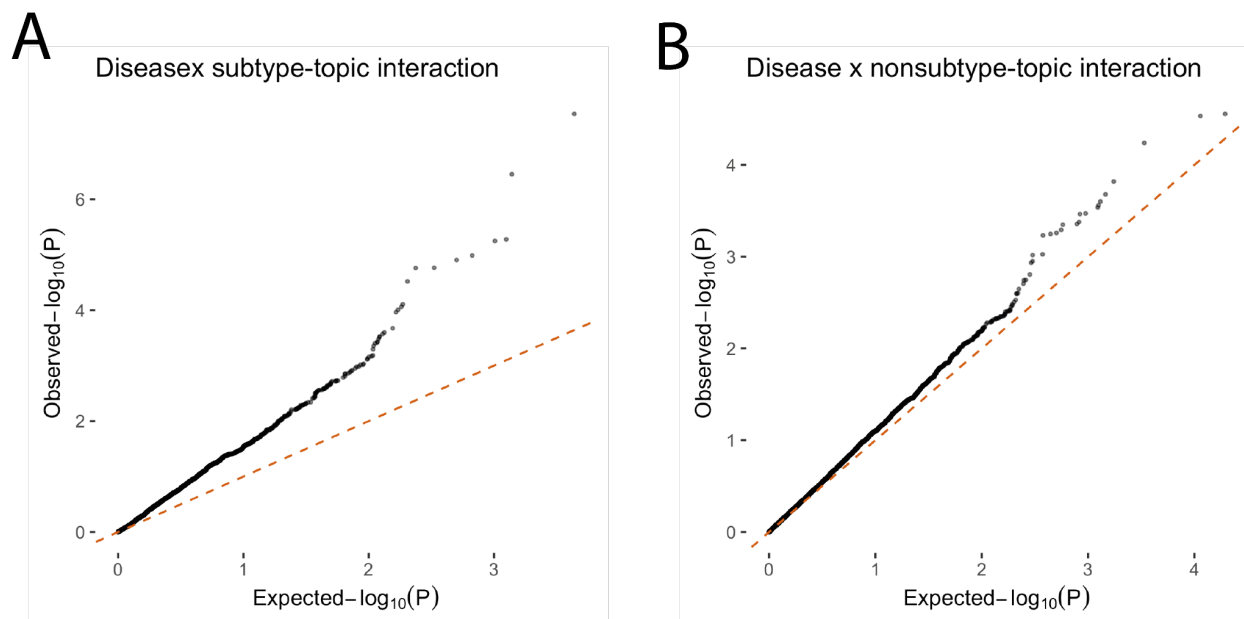


1802 **Supplementary Figure 29: Power to detect true SNP x topic interaction effect under**  
 1803 **simulation.** (a) We simulated data with 10,000 individuals, topic-to-disease main effect size  
 1804 equal to 2, interaction effect from 0.04 to 0.4, and SNP-to-disease main effect size proportional  
 1805 to the interaction effect (0.02 to 0.2); disease diagnoses are generated using gaussian liability  
 1806 with top 20 percentile as cases. We tested for the SNP x topic interaction using model 1 and  
 1807 model 2 and computed the power of discovering the true interaction. (b) We simulated data with  
 1808 10,000 individuals and an interaction effect equal to 0.4. Instead of simulating a single SNP  
 1809 effect, we simulated 2 to 20 variants that all interact with topic weight. We then test the SNP x  
 1810 Topic interaction in model 1 and model 2 with one variant at a time, which is the same strategy  
 1811 as most GWAS interaction tests. We note the power of model 2 is lower than model 1, while we  
 1812 still choose model 2 as it is better calibrated (Supplementary Figure 28).  
 1813

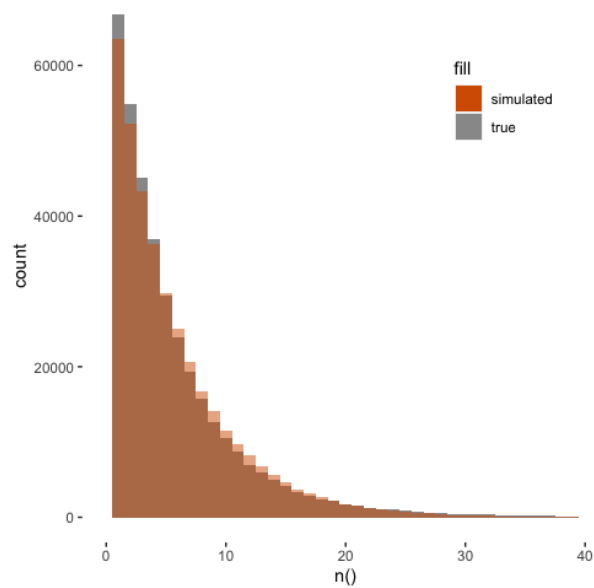
1814  
 1815  
 1816  
 1817



1819 **Supplementary Figure 30. Additional 39 SNPs (mapped genes in the parentheses) that have**  
1820 **different effect sizes in different quantiles of topic weights.** Blue dots are the effect sizes of  
1821 the target SNP within each topic weights quartile; grey dots are the effect sizes of background  
1822 SNPs which are genome-wide significant for the traits but do not have evidence supporting  
1823 interaction with topic weights ( $P > 0.05$ ). The grey line shows the regression line of the grey dots  
1824 and its 95% confidence interval. The topic weights (of the topic being tested) are matched for  
1825 both case and controls within each quartile. Numerical results are reported in Supplementary  
1826 Table 18.

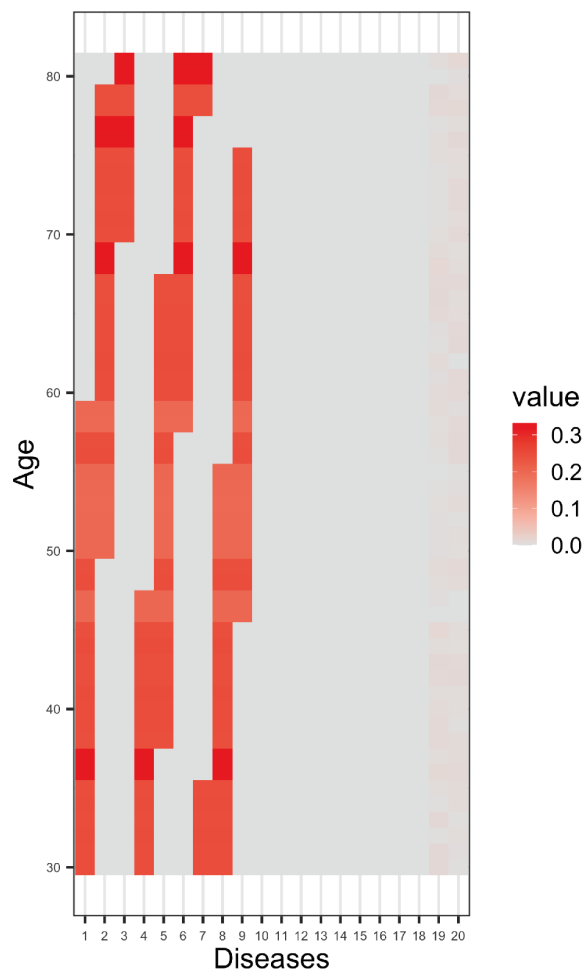


1827 **Supplementary Figure 31: QQ plot of SNP x topic interaction for all GWAS SNPs ( $P <$**   
1828  **$5 \times 10^{-8}$ ).** (a) We show the interaction between SNP-topic where the topics define disease  
1829 subtypes. We focus on the subset of subtypes whose disease have  $h^2$  z-score larger than 4 to  
1830 ensure there is enough GWAS signal for testing. The P-values are for testing the interaction  
1831 effects with nonlinear topic-to-disease main effects (Model 2 in Supplementary Figure 28). **The**  
1832 **median for observed p-value is 0.35.** (b) As a control to show the calibration of the tests, we plot  
1833 the QQ-plot over the same set of GWAS SNPs, but over the topic that are not identified as  
1834 subtypes of the disease by ATM. **The median for observed p-value in the Null test is 0.47. The**  
1835 **observed small inflation of test statistics ( $0.47 < 0.5$ ) is caused by the correlation between topics**  
1836 **(i.e. a SNP that interacts with a subtype-topic is expected to have weak interaction with other**  
1837 **non-subtype topics as the topic weights sum to one).**  
1838  
1839

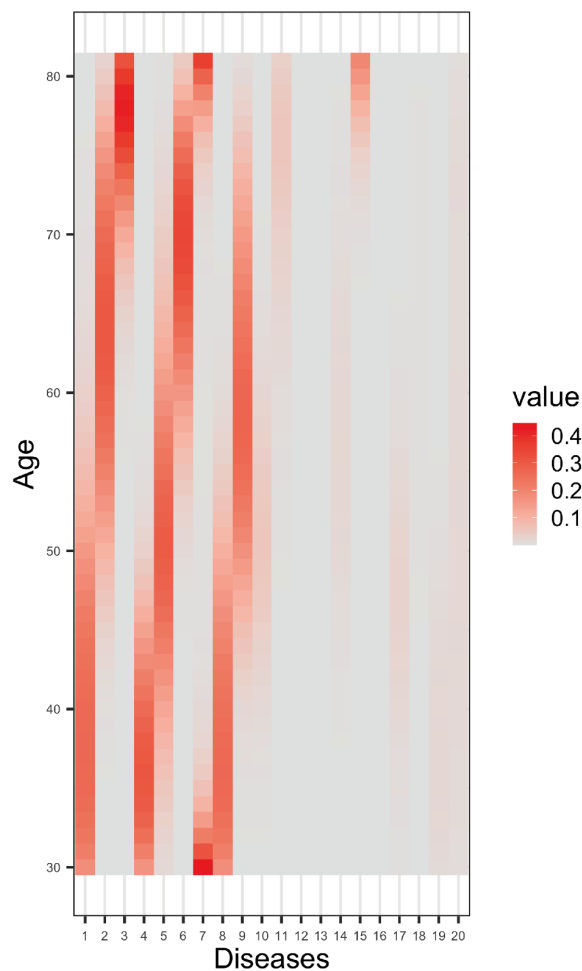


1840  
1841 **Supplementary Figure 32. Comparison of simulated diagnosis per individual versus true**  
1842 **UK Biobank data.** Histogram of number of distinct diseases per patient from the UK Biobank  
1843 HES dataset and from the simulated exponential distribution with mean = 6.1.  
1844

### True disease topic



### Inferred disease topic



1845  
1846  
1847  
1848  
1849

**Supplementary Figure 33. Examples of simulated vs. inferred topic loadings from ATM.** Left panel shows the topic loadings used to simulated 10,000 individuals; right panel shows the inferred topic loadings using topic loadings parametrized as cubic polynomials.