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

Xilin Jiang[1,2,3,4,5,6], Martin Jinye Zhang[4,7§], Yidong Zhang[1,8,9§], Micheal Inouye[5,6,10,11,12,13], Chris Holmes[1,2,14], Alkes L. Price[4,7,15]*, Gil McVean[1]*

Affiliations

[1] Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford OX3 7LF, UK
[2] Department of Statistics, University of Oxford, Oxford OX1 3LB, UK
[3] Wellcome Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK
[4] Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[5] British Heart Foundation Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge UK
[6] Heart and Lung Research Institute, University of Cambridge, Cambridge UK
[7] Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
[8] CAMS China Oxford Institute, Nuffield Department of Medicine, University of Oxford, Oxford OX3 7BN, UK
[9] Department of Radiation Oncology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China
[10] Cambridge Baker Systems Genomics Initiative, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.
[11] Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK.
[12] British Heart Foundation Cambridge Centre of Research Excellence, Department of Clinical Medicine, University of Cambridge, Cambridge, UK.
[13] Cambridge Baker Systems Genomics Initiative, Baker Heart and Diabetes Institute, Melbourne, VIC, Australia.
[14] The Alan Turing Institute, London NW1 2DB, UK
[15] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[§] These authors contributed equally to this work
*These authors jointly supervised the work
Corresponding authors:
xilinjiang@hsph.harvar.edu
aprice@hsph.harvard.edu
gil.mcvean@bdi.ox.ac.uk

Abstract
Longitudinal data from electronic health records (EHR) has immense potential to improve clinical diagnoses and personalised medicine, motivating efforts to identify disease subtypes from age-dependent patient comorbidity information. 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 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 348 diseases with ≥1,000 incidences. We inferred 10 disease topics optimising model fit. We identified 52 diseases with heterogeneous comorbidity profiles (≥500 incidences assigned to each of ≥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 SNPs underlying these differences. For example, the T2D-associated SNP rs1063192 in the *CDKN2B* locus has a higher odds ratio in the top quartile of cardiovascular topic weight (1.19±0.02) than in the bottom quartile (1.08±0.02) (P=4×10^{-5} for difference). In conclusion, ATM identifies disease subtypes with differential genome-wide and locus-specific genetic risk profiles.

Introduction

Longitudinal electronic health record (EHR) data, encompassing diagnoses across hundreds of distinct diseases, offers immense potential to improve clinical diagnoses and personalized medicine[1]. Despite intense interest in both the genetic relationships between distinct diseases[2–11] and the genetic relationships between biological subtypes of disease[12–15], there has been limited progress on classifying disease phenotypes into groups of diseases with frequent co-occurrences (comorbidities) and leveraging comorbidities to identify disease subtypes. Low-rank modelling has appealing theoretical properties[16,17] and has produced promising applications[18–24] to infer meaningful representations of high-dimensional data. In particular, low-rank representation is an appealing way to summarise data across hundreds of distinct diseases[25–27], providing the potential to identify patient-level comorbidity patterns and distinguish disease subtypes. Disease subtypes inferred from EHR data could be validated by comparing genetic profiles across subtypes, which is possible with emerging data sets that link genetic data with EHR data[28–30].

Previous studies have used low-rank representation to identify shared genetic components[25–27] across multiple distinct diseases, identifying relationships between diseases and generating valuable biological insights. However, age at diagnosis information in longitudinal EHR data has the potential to improve such efforts. For example, a recent study used longitudinal disease trajectories to identify disease pairs with statistically significant directionality[31], suggesting that age information could be leveraged to infer comorbidity profiles that capture temporal information. In addition, patient-level comorbidity information could potentially be leveraged to identify biological subtypes of disease, complementing its application to increase power for identifying genetic associations[12] and to cluster disease-associated variants into biological pathways[8]; disease subtypes are fundamental to disease aetiology[14,32–35].

Here, we propose an age-dependent topic modelling (ATM) method to provide a low-rank representation of longitudinal disease records. ATM 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 as a function of age. We applied ATM to 1.7 million disease diagnoses spanning 348 diseases in the UK Biobank, inferring 10 disease topics. We identified 52 diseases with heterogeneous comorbidity profiles that enabled us to define disease subtypes. We used genetic data to validate the disease subtypes, showing that they exhibit differential genome-wide and locus-specific genetic risk profiles.

Results

Overview of methods

We propose an age-dependent topic modelling (ATM) model, providing a low-rank representation of longitudinal records of hundreds of distinct diseases in large EHR data sets (Figure 1, Methods). The model assigns to each individual *topic weights* for several *disease topics*; each disease topic reflects a set of diseases that tend to co-occur as a function of age, quantified by age-dependent *topic loadings* for each disease. The model assumes that for each disease diagnosis, a topic is sampled based on the individual's topic weights (which sum to 1 across topics, for a given individual), and a disease is sampled based on the individual's age and the age-dependent topic loadings (which sum to 1 across diseases, for a given topic at a given age). The model generalises the latent dirichlet allocation (LDA) model[36,37] by allowing topic loadings for each topic to vary with age (Supplementary Note, Supplementary Figure 25).

We developed a method to fit this model that addresses several challenges inherent to large EHR data sets; the method estimates topic weights for each individual, topic loadings for each disease, and posterior diagnosis-specific topic probabilities for each disease diagnosis. First, we derived a scalable deterministic method that uses numerical approximation approaches to fit the parameters of the model, addressing the challenge of computational cost. Second, we used the prediction odds ratio[38] to compare model structures (e.g. number of topics and parametric form of topic loadings as a function of age), addressing the challenge of appropriate model selection; roughly, the prediction odds ratio quantifies the accuracy of correctly predicting disease diagnoses in held-out patients using comorbidity information, compared to a predictor based only on prevalence (see Methods). Third, we employed collapsed variational inference[39], addressing the challenge of sparsity in the data (e.g. in UK Biobank data that we analyzed, the average patient has diagnoses for 6 of 348 diseases analysed); collapsed variational inference outperformed mean-field variational inference[36] in empirical data. Further details are provided in the Methods section and Supplementary Note; we have publicly released open-source software implementing the method (see Code Availability).

We applied ATM to longitudinal records of 282,957 individuals from the UK Biobank[28], containing a total of 1,726,144 disease diagnoses spanning 348 diseases (see Data Availability). Each disease diagnosis had an associated age at diagnosis, defined as the earliest age of reported diagnosis of the disease in that individual; we caution that age at diagnosis may differ from age at disease onset (see Discussion). ATM does not use genetic data, but we used genetic data to analyse the inferred topics (Methods).

Simulations

We performed simulations to compare ATM with latent dirichlet allocation (LDA)[36,37], a simpler topic modelling approach that does not model age. We simulated 61,000 disease diagnoses spanning 20 diseases in 10,000 individuals, using the ATM generative model; the average number of disease diagnoses per individual (6.1), ratio of #individuals/#diseases (500), topic loadings, and standard deviation in age at diagnosis (8.5 years for each disease) were chosen to match empirical UK Biobank data. We assigned each disease diagnosis to one of two subtypes for the underlying disease based on age and other subtype differences, considering high, medium, or low age-dependent effects by specifying an average difference of 20, 10, or 5 years respectively in age at diagnosis for the two subtypes. For each level of age-dependent effects, we varied the proportion of diagnoses belonging to the first subtype (*subtype sample size proportion*) from 10-50%. Further details of the simulation framework are provided in the Methods section. Our primary metric for evaluating the LDA and ATM methods was area under the precision-recall curve (AUPRC)[40], where precision is defined as the proportion of disease diagnoses that a given method assigned to the first subtype that were assigned correctly and recall is defined as the proportion of disease diagnoses truly belonging to the first subtype that were assigned correctly. We discretized the subtype assigned to each disease diagnosis by a given method by assigning the subtype with higher inferred probability. We note that AUPRC is larger when classifying the smaller subtype; results using the second subtype as the classification target are also provided. We used AUPRC (instead of prediction odds ratio) in our simulations because the underlying truth is known. Further details and justifications of metrics used in this study are provided in the Methods section.

Results are reported in Figure 2 and Supplementary Table 1. In simulations with high age-dependent effects, ATM attained much higher AUPRC than LDA across all values of subtype sample size proportion (AUPRC difference: 24%-42%), with both methods performing better at more balanced ratios. Accordingly, ATM attained both higher precision and higher recall than LDA (Supplementary Figure 3). Results were qualitatively similar when using the second subtype as the classification target (Supplementary Figure 2). In simulations with medium or low age-dependent effects, ATM continued to outperform LDA but with smaller differences between the methods. In simulations without age-dependent effects, ATM slightly underperformed LDA (Supplementary Figure 4A).

We performed three secondary analyses. First, we varied the number of individuals, number of diseases, or number of disease diagnoses per individual. ATM continued to outperform LDA in each case, although increasing the number of individuals or the number of disease diagnoses per individual did not always increase AUPRC (Supplementary Figure 4B). Second, we performed simulations in which we increased the number of subtypes from two to five and changed the number of diseases to 50, and compared ATM models trained using different numbers of topics (in 80% training data) by computing the prediction odds ratio; we used the prediction odds ratio

(instead of AUPRC) in this analysis both because it is unclear how to map an underlying truth to an incorrect number of subtypes, and because we wished to evaluate the broader utility of prediction odds ratio. We confirmed that the prediction odds ratio was maximized using five topics, validating the use of the prediction odds ratio for model selection (Supplementary Figure 5A). Third, we computed the accuracy of inferred topic loadings, topic weights, and grouping accuracy (defined as proportion of pairs of diseases truly belonging to the same topic that ATM correctly assigned to the same topic), varying the number of individuals and number of diseases diagnoses per individual. We determined that ATM also performed well under these metrics (Supplementary Figure 5B-E).

We conclude that ATM (which models age) assigns disease diagnoses to subtypes with higher accuracy than LDA (which does not model age) in simulations with age-dependent effects. We caution that our simulations largely represent a best-case scenario for ATM given that the generative model and inference model are very similar (although there are some differences), thus it is important to analyse empirical data to validate the method.

Age-dependent disease topic loadings capture comorbidity profiles in the UK Biobank
We applied ATM to longitudinal records of 282,957 individuals from the UK Biobank[28]. We used Phecode[41] to define 1,726,144 disease diagnoses spanning 348 diseases with at least 1,000 diagnoses each; the average individual had 6.1 disease diagnoses, and the average disease had a standard deviation of 8.5 years in age at diagnosis. The optimal ATM model structure included 10 topics and modelled age-dependent topic loadings for each disease as a spline function with one knot (see below). We assigned names (and corresponding acronyms) to each of the 10 inferred topics based on the Phecode systems[41] assigned to diseases with high topic loadings (aggregated across ages) for that topic (Table 1).

Age-dependent topic loadings across all 10 topics and 348 diseases (stratified into 18 Phecode systems), summarised as averages across age<60 and age≥60, are reported in Figure 3, Supplementary Figure 9, and Supplementary Table 3. Some topics such as NRI span diseases across the majority of Phecode systems, while other topics such as ARP are concentrated in a single Phecode system. Conversely, a single Phecode system may be split across multiple topics, e.g. the digestive system is split across UGI, LGI, and MDS. We note that topic loadings in diseases that span multiple topics are heavily age-dependent. For example, the CVD subtype of type 2 diabetes is associated with early onset whereas the MGND subtype of type 2 diabetes is associated with late onset. The CVD subtype of type 2 diabetes resembles the Beta Cell and Lipodystrophy subtypes described in ref. [34] and the severe insulin-deficient diabetes (SIDD) subtype described in ref.[14], which are characterised by early onset and have multiple morbidities including hypercholesterolemia, hyperlipidemia, and cardiovascular diseases. In addition, early-onset breast cancer and late-onset breast cancer are associated to different topics, consistent with their distinct aetiologies.

We performed five secondary analyses to validate the inferred comorbidity topics. First, we fit ATM of different configurations on the training data, and computed their prediction odds ratios on the testing data. ATM with 10 topics and topic loadings as spline function is among the optimal models. (Supplementary Figure 6; see Methods). Second, we confirmed that ATM attained higher prediction odds ratios than LDA across different topic numbers (Supplementary Figure 7). Third, we reached similar conclusions using evidence lower bounds (ELBO) [38] when fitting the model without training and testing data split (Supplementary Figure 8). Fourth, to verify that the sets of diseases appearing in the same topics actually co-occur in individuals in independent data, we tested if the diseases grouped in the same topic by ATM on the training set have higher than expected probability to appear as comorbidity in testing data. We select all disease sets containing combinations of 2, 3, 4, and 5 diseases assigned to the same topic by max topic loading, and divide incidences where the disease sets appeared in one patient by the expected number in an independent testing set. The test odds ratio is consistently above one and increases with the number of diseases in the comorbidities, for each inferred topic (Supplementary Figure 10). We also compared the topic loadings by repeating the inference on

female- or male-only populations and found no major discrepancy except for genitourinary comorbidities MGND and FGND. (topic loading R^2 (female vs. all) = 0.788, topic loading R^2 (male vs. all) = 0.773, Supplementary Figure 11).

Age-dependent topic loadings for two representative topics that both include circulatory-related diseases (CER and CVD), restricted to diseases with high topic loadings, are reported in Figure 4A and Supplementary Table 4; results for all 10 topics are reported in Supplementary Figure 12 and Supplementary Table 4. Disease topics capture known biology as well as the age-dependency of comorbidities for the same diseases. For example, early onset of essential hypertension is associated with the CVD topic, which captures the established connection between lipid dysfunction ("hypercholesterolemia"; "Hyperlipidemia") and cardiovascular diseases[42], while later onset of essential hypertension is associated with the CER topic, which pertains to type 2 diabetes, obesity and COPD. We note that most diseases have their topic loadings concentrated into a single topic (Figure 4B, Supplementary Figure 13A and Supplementary Table 3), and that most individuals have their topic weights concentrated into 1-2 topics (Figure 4C and Supplementary Figure 13B).

We conclude that ATM identifies latent disease topics that robustly compress age-dependent comorbidity profiles and capture disease comorbidities both within and across Phecode systems.

Disease subtypes defined by distinct topics are genetically heterogeneous
We sought to define disease subtypes based on the diagnosis-specific topic probabilities of each disease diagnosis. We assigned a discrete topic assignment to each disease diagnosis based on its maximum diagnosis-specific topic probability, and defined the disease subtype of each disease diagnosis based on the topic assignment. We restricted our disease subtype analyses to 52 diseases with at least 500 diagnoses assigned to each of two distinct subtypes (Methods, Supplementary Figure 9, Supplementary Figure 14, and Supplementary Table 5).

Age-dependent distributions of subtypes (topics) for four diseases (type 2 diabetes, asthma, hypercholesterolemia, and essential hypertension) are reported in Figure 5A and Supplementary Table 6; results for all 52 diseases are reported in Supplementary Figure 15 and Supplementary Table 6. The number of subtypes can be large, e.g. six subtypes for essential hypertension. Subtypes are often age-dependent, e.g. for the CVD and MGND subtypes of type 2 diabetes (discussed above).

ATM and the resulting subtype assignments do not make use of genetic data. However, we used genetic data to assess genetic heterogeneity across inferred subtypes of each disease. We first assessed whether polygenic risk scores (PRS) for overall disease risk varied across subtypes of each disease; PRS were computed using BOLT-LMM with five-fold cross validation [43,44] (see Methods and Code Availability). Results for four diseases (from Figure 5A) are reported in Figure 5B and Supplementary Table 7; results for all 10 well-powered diseases (10 of 52 diseases with highest z-scores for nonzero SNP-heritability) are reported in Supplementary Figure 16 and Supplementary Table 7. We identified 18 disease subtypes (of 100 disease subtypes analysed) for which PRS values in disease cases vary with patient topic weight. For example, for essential hypertension, hypercholesterolemia, and type 2 diabetes, patients assigned to the CVD subtype had significantly higher PRS values than patients assigned to other subtypes. For essential hypertension, patients assigned to the CER subtype had significantly higher PRS values; for type 2 diabete, patients assigned to the CER subtype had lower PRS values than CVD subtype, even if majority of type 2 diabete are CER subtype. We further verified that most of the variation in PRS values with disease subtype could not be explained by age[45] or differences in subtype sample size (Supplementary Figure 17). These associations between subtypes (defined using comorbidity data) and PRS (defined using genetic data) imply that disease subtypes identified through comorbidity are genetically heterogeneous, consistent with differences in disease aetiology.

We further investigated whether subtype assignments (defined using comorbidity data) revealed subtype-specific excess genetic correlations. We estimated excess genetic correlations between pairs of disease subtypes (relative to genetic correlations between the underlying diseases). Excess genetic correlations for 15 disease subtypes (spanning 11 diseases and 3 topics: CER, MGND and CVD) are reported in Figure 6A and Supplementary Table 8 (relative to genetic correlations between the underlying diseases; Figure 6B), and excess genetic correlations for all

89 well-powered disease subtypes (89 of 378 disease subtypes with z-score > 4 for nonzero SNP-heritability) are reported in Supplementary Figure 18 and Supplementary Table 8. Genetic correlations between pairs of subtypes involving the same disease were significantly less than 1 (FDR<0.1) for hypertension (CER vs. CVD: $\rho = 0.86 \pm 0.04$, P=0.0004; MGND vs. CVD: $\rho = 0.74 \pm 0.05$, P=3×10^{-8}) and type 2 diabetes (CER vs. MGND: $\rho = 0.64 \pm 0.09$, P=8×10^{-5}) (Figure 6A; Supplementary Figure 18). In addition, we observed significant excess genetic correlations (FDR<0.1) for 8 pairs of disease subtypes involving different diseases (Figure 6A; Supplementary Figure 18). We verified that the excess genetic correlations could not be explained by non-disease-specific differences in the underlying topics (which are weakly heritable; Supplementary Table 2) by repeating the analysis using disease cases and controls with matched topic weights (Methods, Supplementary Figure 19). We also estimated subtype-specific SNP-heritability and identified some instances of differences between subtypes, albeit with limited power.

Finally, we used the population genetic parameter F_{ST}[46,47] to quantify genome-wide differences in allele frequency between two subtypes of the same disease; we used F_{ST} on control sets with matched topic weights to assess statistical significance while accounting for non-disease-specific differences in the underlying topics (excess F_{ST}; Methods). We determined that 63 of 104 pairs of disease subtypes involving the same disease (spanning 29 of 49 diseases, excluding 3 diseases that did not have enough controls with matched topic weights) had significant excess F_{ST} estimates (FDR < 0.1) (Supplementary Figure 20, Supplementary Table 9). For example, each pair of the CVD, CER, and MGND subtypes of type 2 diabetes had significant excess F_{ST} estimates (0.0003, P=0.001 based on 1,000 matched control sets). This provides further evidence that disease subtypes as determined by comorbidity have different molecular and physiological aetiologies.

We conclude that disease subtypes defined by distinct topics are genetically heterogeneous.

Disease-associated SNPs have subtype-dependent effects

We hypothesised that disease genes and pathways might differentially impact the disease subtypes identified by ATM. We investigated the genetic heterogeneity between disease subtypes at the level of individual disease-associated variants. We employed a statistical test that tests for SNP x topic interaction effects on disease phenotype in the presence of separate SNP and topic effects (Methods). We verified via simulations that this statistical test is well-calibrated under a broad range of scenarios with no true interaction, including direct effect of topic on disease, direct effect of disease on topic, pleiotropic SNP effects on disease and topic, and nonlinear effects (Supplementary Figure 21). We also assessed the power to detect true interactions (Supplementary Figure 22). To limit the number of hypotheses tested, we applied this test to independent SNPs with genome-wide significant main effects on disease (Methods) We thus performed 2,530 statistical tests spanning 888 disease-associated SNPs, 14 diseases, and 35 disease subtypes (Supplementary Table 11). We assessed statistical significance using global FDR<0.1 across the 2,530 statistical tests. We also computed main SNP effects specific to each quartile of topic weights across individuals, as an alternative way to represent SNP x topic interactions.

We identified 43 SNP x topic interactions at FDR<0.1 (Figure 7, Supplementary Figure 23, Supplementary Table 10 and Supplementary Table 12). We highlight 4 examples. First the type 2 diabetes-associated SNP rs1063192 in the *CDKN2B* locus has a higher odds ratio in the top quartile of CVD topic weight (1.19±0.02) than in the bottom quartile (1.08±0.02) (P=$4 \times 10^{-4}$for difference). Second, the asthma-associated SNP rs1837253 in the *TSLP* locus has a higher odds ratio in the top quartile of SRD topic weight (1.17±0.02) than in the bottom quartile (1.05±0.02) (P=1×10^{-4} for difference). Third, the hypertension-associated SNP rs3735533 in the *HOTTIP* locus has a lower odds ratio in the top quartile of CVD topic weight (1.07±0.02) than in the bottom quartile (1.13±0.02). Fourth, the hypothyroidism-associated SNP rs9404989 in the *HCG26* locus has a higher odds ratio in the top quartile of FGND topic weight (1.90±0.24) than in the bottom quartile (1.19±0.13)(P=$3 \times 10^{-3}$for difference). To verify correct calibration, we performed control SNP x topic interaction tests using the same 888 disease-associated SNPs together with random topics that did not correspond to disease subtypes, and confirmed that these control tests were well-calibrated (Supplementary Figure 24B).

We conclude that genetic heterogeneity between disease subtypes can be detected at the level of individual disease-associated variants.

Discussion

We have introduced an age-dependent topic modelling (ATM) method to provide a low-rank representation of longitudinal disease records, leveraging age-dependent comorbidity profiles to identify and validate biological subtypes of disease. Our study builds on previous studies on topic modelling[36,37,39,48], genetic subtype identification[13–15], and low-rank modelling of multiple diseases[25–27]. We highlight three specific contributions of our study. First, we incorporated age at diagnosis information into our low-rank representation, complementing the use of age information in other contexts[31,45,49]; we showed that age information is highly informative for our inferred comorbidity profiles in both simulated and empirical data, emphasising the importance of accounting for age in efforts to classify disease diagnoses. Second, we identified 52 diseases with heterogeneous comorbidity profiles that we used to define disease subtypes, many of which had not previously been identified. Third, we used genetic data (including PRS, genetic correlation and F_{ST} analyses) to validate these disease subtypes, confirming that the inferred subtypes reflect true differences in disease aetiology.

We emphasise three downstream implications of our findings. First, it is of interest to perform disease subtype-specific GWAS on the disease subtypes that we have identified here, analogous to GWAS of previously identified disease subtypes[13–15]. Second, our findings motivate efforts to understand the functional biology underlying the disease subtypes that we identified; the recent availability of functional data that is linked to EHR is likely to aid this endeavor[28,50]. Third, it is of interest to apply ATM to identify age-dependent comorbidity profiles and disease subtypes in other EHR data sets[29,30]; establishing representations of disease topics that are transferable and robust across different healthcare systems and data sources represents a major future challenge.

Our findings reflect a growing understanding of the importance of context, such as age, sex, socioecnomic status and previous medical history, in genetic risk [45,51,52]. To maximize power and avoid systematic bias, such context needs to be integrated into clinical risk prediction tools that combine genetic information, such as polygenic risk scores [1,53], and non-genetic conventional risk factors. We note that aspects of context are themselves influenced by genetic risk factors, hence there is an open and important challenge in determining how best to combine medical history and/or causal biomarker measurements with genetic risk to predict future events[54].

We note several limitations of our work. First, age at diagnosis information in EHR data may be an imperfect proxy for true age at onset, particularly for less severe diseases that may be detected as secondary diagnoses; although perfectly accurate age at onset information would be ideal, our study shows that that imperfect age at diagnosis information is sufficient to draw meaningful conclusions. Second, raw EHR data may be inaccurate and/or difficult to parse[1]; again, although perfectly accurate EHR data would be ideal, our study shows that imperfect EHR data is sufficient to draw meaningful conclusions. Third, our ATM approach incurs substantial computational cost; however, we show that analyses of biobank-scale data sets are

computationally tractable. Finally, we have applied ATM to a UK population of predominantly European ancestry; it is of interest to apply ATM to diverse populations[29,30]. Despite these limitations, ATM is a powerful approach for identifying age-dependent comorbidity profiles and disease subtypes.

Acknowledgements

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

Funded by Wellcome (BST00080- H503.01 to XJ, 100956/Z/13/Z to GM, https://wellcome.org); the Li Ka Shing Foundation (to GM, https://www.lksf.org); NIH grants R01 HG006399, R01 MH101244, and R37 MH107649 (To ALP); The Alan Turing Institute (https://www.turing.ac.uk), Health Data Research UK (https://www.hdruk.ac.uk), the Medical Research Council UK (https://mrc.ukri.org), the Engineering and Physical Sciences Research Council (EPSRChttps://epsrc.ukri.org) through the Bayes4Health programme Grant EP/R018561/1, and AI for Science and Government UK Research and Innovation (UKRI, https://www.turing.ac.uk/ research/asg) (to CH).The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Code availability

The code generated in this study is available at: https://github.com/Xilin-Jiang/ATM

Methods

Age-dependent topic model (ATM)

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

We constructed a Bayesian hierarchical model to infer *K* latent risk profiles for *D* distinct common diseases. Each latent risk profile (comorbidity topics) is age-evolving and contains risk trajectories for all *D* diseases considered. Each individual might have a different number of diseases, while the disease risk is determined by the weighted combination of latent risk topics. The indices in this note are as follows:

- $s = 1, \ldots, M$;
- $n = 1, \ldots, N_s$;
- $i = 1, \ldots, K$;
- $j = 1, \ldots, D$;

where M is the number of subjects, N_S is the number of records within s^{th} subject, *K* is the number of topics, and *D* is the total number of diseases we are interested in.
The plate representation of the generative model is summarised in Supplementary Figure 25.

- $\theta \in R^{M \times K}$ is the topic weight for all individuals (referred to as patient topic weights), each row of which ($\in R^K$) is assumed to be sampled from a Dirichlet distribution with parameter α. α is set as a hyper parameter: $\theta_s \sim Dir(\alpha)$.
- $z \in \{1,2,\ldots,K\}^{\sum_S N_S}$ (referred to as diagnosis topic weights) is the topic assignment for each diagnosis $w \in \{1,2,\ldots,D\}^{\sum_S N_S}$. Note the total number of diagnoses across all patients are $\sum_S N_S$. The topic assignment for each diagnosis is generated from a multinoulli distribution with parameters equal to s^{th} individual topic weight: $z_{sn} \sim Multi(\theta_s)$.
- $\beta(t) \in F(t)^{K \times D}$ is the topic which is $K \times D$ functions of age t. $F(t)$ is the class of functions of t. At each plausible t, the following is satisfied: $\sum_j \beta_{ij}(t) = 1$. In practice we use softmax function to ensure above is true and add smoothness by constrain $F(t)$ to be spline or polynomial functions: $\beta_{ij}(t) = \frac{exp(p_{ij}{}^T \phi(t))}{\{\sum_{j=1}^{D} exp(p_{ij}{}^T \phi(t))}$, where $p_{ij} = \{p_{ijd}\}$; $d = 1,2,\ldots,P$; P is the degree of freedom that controls the smoothness; $\phi(t)$ is polynomial and spline basis for age t.
- $w \in \{1,2,\ldots,D\}^{\sum_S N_S}$ are observed diagnoses. The nth diagnosis of sth individual w_{sn} is 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 age of the observed age at diagnosis of the observed diagnosis w_{sn}.

The values of interest in this model are global topic parameter β, individual (patient) level topic weight θ, and diagnosis-specific topic probability z. Based on the generative process above, we notice that each patient is independent conditional on α. Therefore, we could adopt an EM strategy, where we first estimate θ and z, then estimate β which maximises the evidence lower bound (ELBO).

The key element in our model is age-evolving risk profiles, which is achieved by model the comorbidity trajectories $\beta(t) \in F(t)^{K \times D}$ as functions of age. The functionals $F(t)$ considered are linear, quadratic, cubic polynomials, and cubic splines with one, two and three knots.

Inference of ATM

The variables of interest are global topic parameter $\beta(t)$, individual (patient) level topic weight θ, and topic weight z of each diagnosis. In a Bayesian setting, the model could be evaluated by the evidence function $p(w|\alpha, \beta)$. The best $\beta(t)$ is found by maximise the evidence function, while for θ and z we aim to find or approximate their posterior distribution $p(z, \theta \mid w, \alpha, \beta)$. Given that the posterior distribution is intractable, we use variational distribution $q(z, \theta)$ to approximate them. Now we could write the evidence function as:$p(w| \alpha, \beta) = L(z, \theta, \beta, \alpha) + KL(q||p)$.

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

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

When finding the posterior of θ and z, we want the$ln\ q(z, \theta)$ to be as close to the posterior$p(z, \theta \mid w, \alpha, \beta)$ as possible. Since $KL(q||p) = 0$ when $q(z, \theta) = p(z, \theta \mid w, \alpha, \beta)$, this could be achieved by minimising $KL(q||p)$ or maximise $L(z, \theta, \beta, \alpha)$. The most commonly used form of $q(z, \theta)$ assume the distribution is factorised, which might cause instability when signal-to-noise ratio is low[55]. Comparison of the evidence lower bound $L(z, \theta, \beta, \alpha)$ shows it is consistently higher for collapsed variational inference.(Supplementary Figure 1) Therefore we choose collapsed variational inference which is more accurate[39]. The collapsed variational inference is achieved by integrate out θ from the likelihood function $p(w, z, \theta \mid \alpha, \beta)$ and find the approximated posterior distribution $q(z)$. For detailed derivation, the comparison between the collapsed variational methods and mean-field method, and update algorithms, see Supplementary Note.

When finding the $\beta(t)$ that maximises the evidence function, we again maximise $L(z, \theta, \beta, \alpha)$. Maximising $L(z, \theta, \beta, \alpha)$ with respect to $\beta(t)$ does not have an analytical solution due to its softmax structure. We use local variational methods and numeric optimisation to find the distribution of $\beta(t)$. Details are provided in Supplementary Note.

We extract topic weights at patient-level and diagnosis-level from the posterior distribution inferred from the data. Our model has the desired property that each patient and patient-diagnosis are assigned to comorbidity topics. The model estimates the posterior distribution $q(z)$, which is a categorical distribution (Supplementary Note.) Several metrics related to topic assignments could be derived from the $q(z)$:

- Each patient-diagnosis (incident disease) has a diagnosis-specific topic probability, which is computed as $E_q\{z_n\}$.
- Each patient has a posterior topic assignment θ_s , which is a dirichlet distribution $\theta_s \sim Dir(\alpha + \sum_{n=1}^{N_S} E_q\{z_n\})$. The topic weights of each patient is the mode of 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 $\alpha = 1$). The value is used as the patient low-rank representation of disease history, for analysis including PRS association with comorbidity within cases and G x Topic interaction analysis.
- The average topic assignments of disease j is the mean over all incidences $\underline{E_q\{ z_{sn \in \{w_{sn}=j\}}\}}$. This metric is used to measure which comorbidity topic a disease is associated with (Figure 4B), and it is equivalent to a weighted average of topic loadings (for the specific weighted average expression, see equation 5 of Supplementary Note). A disease assigned to multiple topics is considered to have comorbidity subtypes.
- A hard assignment of a patient-diagnosis to a subtype is based on the max value of the vector $E_q\{z_n\}$. The incident disease is assigned to topic $argmax_i\ (E_q\{z_{ni}\})$.

Metrics for evaluating ATM

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

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

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

Evidence Lower Bound (ELBO): Mathematical expression of ELBO is presented in equation 9 in Supplementary Note. To find the best model that fit to the entire dataset, we evaluate the ELBO for models with topic numbers between 5 to 20, 25, 30, and 50 topics and age profiles configured by linear, quadratic polynomial, cubic polynomial, and splines with one, two and three knots. Each model is run for 10 times with random initialisations. We choose the model that has the highest ELBO after converging.

AURPC: To evaluate whether a model could capture the comorbidity subtypes in simulation analysis, we compute the precision, recall, and area under precision-recall curve (AUPRC) to correctly classify disease diagnosis to be from the topic that it is generated from. The topic of each diagnosis is determined by diagnosis-specific topic probability.

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

Simulations of ATM method.

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

- $M = 10{,}000$;
- $\underline{N_S} = 6.1$;
- $N_S \sim exp\{\underline{N_S}\}$;
- $D = 20$;
- $K = 2$;

Here M is the number of individuals in the population, $\underline{N_S}$ is the average number of diseases for each individual, D is the total number of diseases, K is the number of comorbidity topics. The distribution of disease number per-individual N_S is sampled from an exponential distribution, which matches those from UK Biobank data (Supplementary Figure 26). According to equation

3.1 in Ghorbani et al.[55], whether the topic model could capture the true latent structure is determined by the information signal-to-noise ratio and could be evaluated with limits $M \rightarrow \infty$; $D \rightarrow \infty$; $\frac{D}{M} \rightarrow \delta$, where δ is a constant. Therefore we choose D and M at scales that make $\frac{D}{M}$ approximately similar to those of the UK Biobank dataset (Samples size = 282,957; distinct disease number = 349).

The simulated topics loadings are constructed as follows:

- All but K diseases are simulated to be associated with comorbidity profiles. Each of them has a risk period of 30 years and overlaps for 10 years with the next disease. For example, if disease 1 has a risk period from 30 to 59 years of age, disease 2 will have a risk period between 50 to 79 years of age. When the risk period reaches the maximal age, the truncated part will be carried to the next disease to create diseases with shorter risk period. All risk periods are assigned a value 1.
- K diseases that are not associated with comorbidity are simulated to span all topics. The values of these diseases are sampled from $Unif(0, \frac{0.1}{K})$ for each topic. Here K is the number of topics.
- The age profiles are then normalised at each age point to ensure $\sum_{j=1}^{D} \beta_j(t) = 1$ for all t. With this constraint we could sample a disease at each age t using a multinomial probability with the topic loading as the parameter. The age range of the simulated topics is 30 to 81 years of age, which is the minimal and maximal age at diagnosis of incident disease in the UK Biobank population. An example of simulated topic is shown in Supplementary Figure 27

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

Since in real data we only use the first age at diagnosis for diseases that are recorded repeatedly within the same patient, we filter the simulated diseases accordingly. The filtered data are fed into the inference functions to infer the latent topics and disease assignments. The inferred topics resemble the true topics used to simulate diseases as shown in Supplementary Figure 27. For the initialisation of each inference, we first sample β and θ from the Dirichlet distribution of non-informative hyperparameters, then initialise other variables parameters following the generative process. The variational inference converged where the relative increase of ELBO is below 10^{-6}.

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

We create diseases with distinct comorbidity profiles by combining diseases that are sampled from distinct topics and labelling them as a single disease. We first chose one disease (**disease A**) then sampled a proportion of a second disease (**disease B**) to label as **disease A**. The proportion is varied to create a different sample size ratio of the two subtypes. In scenario one, **disease B** is a disease that has the exact same age distribution as **disease A** but from the other topic. In scenario two, **disease B** is from the other topic and has a different age distribution (age at diagnosis moves up for 20 years, 10 years, or 5 years, respectively) than **disease A**. After changing the labels of **disease B** to be the same as **disease A**, we used the inference procedure described as above to get the posterior distribution.

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

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

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

UK Biobank comorbidity data.

We use the hospital episode statistics (HES) data within the UK Biobank dataset, which records diseases using the ICD-10/ICD-10CM coding system. Codes started with letters from A to N are

kept as they correspond to disease code (opposed to procedure codes). The disease records were mapped from ICD-10/ICD-10CM codes to PheCodes using a three-step procedure: Firstly, we map the first four substring of each ICD-10 records to the phecodes, using the map file downloaded from phewascatalog.org; Secondly, we map the remaining records using ICD-10CM map file downloaded from phewascatalog.org; Lastly, we map remaining records to a collapsed ICD-10CM mapping system which only use the first four character of ICD-10CM codes. We also noticed an ICD-10/ICD-10CM code could map to multiple PheCodes. When a single ICD-10/ICD-10CM code s mapped to more than one PheCodes, we only kept the Phecode that are mapped to the most ICD-10 codes (i.e. PheCode is constructed by combining ICD-10 that represent similar diseases. The Phecode that represent a larger number of ICD-10 codes are more likely to be a well defined disease, which we chose to keep.), which ensure that one ICD-10(CM) code only maps to one PheCode. Using the procedure above, we mapped 99.7% ICD-10/ICD-10CM code to PheCodes, with 4,637,127 records in total.

The mapped Phecodes are filtered to keep only the first age at diagnosis for the same diseases within a patient. The age at diagnosis for each record is computed as the difference between month of birth to the episode starting date. We then computed the occurrence of each disease in the UK Biobank and kept 348 that have more than 1000 occurrences. We then filter the patients to keep only those who have at least two distinct diseases from the 348 focal diseases, as we are most interested in the comorbidity information. We treated the death as an additional disease (8,666 records) to evaluate if certain comorbidities are more likely to lead to fatal events. After these procedures, there are in total 1,726,144 distinct records across 282,957 patients.

To name the topics inferred from the UK Biobank, we take the sum of average topic assignments (section Inference of ATM) over diseases that are within each phecode system and extract the top 3 systems. Most of comorbidity topics are named using the first three topics (e.g. CER: cardiovascular, endocrine/metabolic, respiratory), except for topics that are predominantly associated with one system (LGI: lower gastrointestinal; UGI: upper gastrointestinal; CVD: cardiovascular).

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

To compare the comorbidity heterogeneity between age groups, we group the incidences for each disease to two age groups: young group (<60 years of age) and old group (≥60 years of age). We compute the average topic assignment of each group as described in section Inference of ATM.

Additionally, we inferred topics for male (984,554 records in 156,366 individuals) and female (741,590 records in 126,591 individuals) populations respectively using a model with 10 topics and spline function with one knot. We extract the average topic assignment for each disease, and use Pearson's correlation to match the topics for both sexes to the topics inferred on the entire population.

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

Polygenic risk scores (PRS) analysis.
To exclude the possibility of overfitting the data and population stratification, we compute PRS using mixed-effect association on the British Isle ancestry group for the 10 heritable diseases that have the highest heritability z-scores. We used a mixed model to estimate effect size implemented by BOLT-LMM and constructed genome-wide PRS [43]. For the computation of PRS, we randomly sampled half of the British isle ancestry population (N = 204,847) for computation efficiency or sample 9 controls for each case to ensure the minimal prevalence rate above 10% as recommended by BOLT-LMM. We used PLINK to select genotyped SNPs with MAF > 0.1% as recommended in BOLT-LMM. For each disease, we performed 5 fold cross validation to perform BOLT-LMM analysis and compute the PRS on the held-out testing set. The predictive PRS are then used to compute the excess PRS over different topic loadings, by a linear regression where PRS is the response variable and topic weights is the predictor.

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

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

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

Genetic correlations analysis.
For each disease and disease subtypes, we use a case-control matching strategy to construct data to estimate coefficients for genetic correlation analysis. For each case in the disease group, we pick four nearest neighbors (without replacement) from the control group, matching sex, BMI, year of birth and 40 genetic principal components. The covariates are available within the UK Biobank data set, over which we computed the principal components across the British Isle ancestry population. We then compute the Euclidean distance of the principal components to find the nearest neighbours in the population. All cases are matched with four controls except for 401.1 essential hypertension which has a sample size larger than 20% of the population. We match only one control for each hypertension case.

We perform logistic regression with sex and top 10 principal components as covariates to estimate the main variant effect of the 805,426 variants that are genotyped. We used PLINK 1.9 for association analysis[56]. With the summary statistics from the association analysis, we use LDSC to map the summary statistics to HapMap3 SNPs and match the effect and non-effect alleles[2,57]. Since UK Biobank is mostly of British Isle ancestry, we use the pre-computed LD score from the LDSC website. We estimated the heritability for each disease or disease subtype which has more than 1000 incidences (378 diseases subtypes and diseases). We use 1000 incidence threshold as LDSC are more accurate with larger sample size. We focus on 71 disease and 18 disease subtypes that have heritability z-score above 4 for genetic correlation analysis.

The genetic correlation is computed for each duo of disease/subtypes using the same summary statistics and LD score regression. We report the estimate of genetic correlation and z-scores. Additionally, for duos that involve subtypes (disease-subtype or subtype-subtype), we compute the differences between genetic correlation with subtypes and that of all disease diagnoss. For example, genetic correlation of T2D-CER versus hypertension-CVD are compared to genetic correlation between all T2D and hypertension incidences. The z-score and p-value of the genetic correlation differences are reported. Note that genetic correlation of subtypes of the same disease are compared to 1.

To avoid potential collider effects where subtypes are defined by topic components that are independent of the diseases, we further match cases in each subtype with controls that match the topic loadings. We computed PCs from 23 variables (10 topic loadings, 10 PCs, year of birth, sex, and BMI) and use the nearest neighbour procedure (by Euclidean Distance) to find controls for each case. Here controls are chosen from individuals without the targeting disease, i.e. an individual with one subtype of the target disease could not be a control for the other subtypes. We performed the same analysis using this case-control matching procedure and compared the genetic correlation with the case-control procedure described above. We perform the analysis for four diseases that have evidence for genetic subtypes: asthma, type 2 diabetes, hypercholesterolemia, and hypertension. For one subtype (hypertension-CVD), the heritability (0.0313, s.e. = 0.0289) is below threshold after matching the topic, which was excluded in genetic correlation analysis.

F_{ST} analysis.

To evaluate the genetic heterogeneity between disease subtypes, we estimated the Fst for 52 diseases that have at least 500 incidences assigned to a secondary topic. To test the statistical significance of Fst, we adopted a permutation strategy and sampled the same number of controls of similar topic weights distribution for each subtype. The topic weights are matched by sampling (without replacement) the same number of controls for each dominant topic weight quartile of the cases (i.e. matching the topic that defines the subtype), which ensures the controls

have the same topic weight stratification as the disease subtypes. We then compute the F_{ST} across the control groups matched for subtypes. We excluded three diseases, "hypertension", "hypercholesterolemia", and "arthropathy", from F_{ST} analysis as we do not have enough controls that match topic weight distribution. The F_{ST}s are computed using PLINK 1.9's weighted mean across all genotyped SNPs.

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

SNP x topic interaction test.

For the diseases that have heritability z-score above 4 in the UK Biobank, we further investigated whether there are interactions between genetic risk factors with the topic loadings. We used a fit a logistic regression model using following model:

$$logit(p) = \beta_0 + \beta_1 * T + \beta_2 * T^2 + \beta_3 * G + \beta_4 * G * T,$$

where T is individual topic weights for a specified topic, G is the genotype, and p is the probability of getting the disease. We computed the test statistics under the null that $\beta_4 = 0$. We used QQ plots to check that the test statistics are well calibrated for each disease-topic pair.

Since the simulation shows the interaction test is underpowered when the variant effects are small, we focus on the set of SNP that reaches genome-wide significance level to increase power to detect interaction effects. We performed LD-clumping using $r^2 > 0.6$ to remove variants that are in strong LD with the lead variants. We computed the test statistics using the model above (for testing $\beta_4 = 0$) and computed study-wise FDR across disease-topic pairs.

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

Simulations of SNP x topic interaction

We simulate comorbidity with genetics to test interaction between genetic and comorbidity topics. We simulated 100 independent variants with MAF randomly sampled from $Unif(0, 0.5)$. We assumed an additive model and simulated genotypes for the population using Hardy-Weinberg equilibrium. We simulated three types of genetic effects on topic and diseases on topic of the simulation framework described in Simulations of ATM method section:

- Genetics-topic effect: each variant is simulated to have an linear effect of 0.04 on the topic loading. We choose this value as after normalising the topic, a regression of causal variant to topic would have an effect size approximately 0.01 which is similar to our observation in the UK Biobank. The number of variants that are causal to the topic varies between 2 to 20. We simulated the effect on one topic by adding additive SNP effects and normalise the topic loadings of each patient. The topic-disease causality is a natural consequence following the generative process of sampling data.
- Genetic-disease-topic effect: we simulated a heritable disease that is causal to the topic. The disease is simulated with 20 causal variants each of effect size 0.15. We vary the disease-to-topic causal effect from 0.05 to 0.5, with a default value of 0.1 in other analyses (similar to the correlation we found in UK Biobank analysis). We simulated the effect on one topic by adding additive causal disease effect and normalise the topic loadings of each patient.
- The genetic effect could interact with the topic when contributing to disease risk. We simulated four additional diseases to represent different structures (Supplementary Figure 21).
    - Genetic effects interact with topic loading on altering disease risk. The interaction term is added to the mean of disease liability, which is sampled from a Gaussian distribution. The disease is then sampled by a threshold on the liability, where the incidence rate is by default 0.5. The interaction effect is varied from 0.4 to 4, with default value equal to 2.
    - Pleiotropy effects are simulated with a variant that have both genetic-disease and genetic-topic-disease effects. Both genetic and topic effects are added to the mean of disease liability. A disease is sampled by a threshold with default incidence rate equal to 0.5. The topic-disease effect is varied from 0.4 to 4, with default value equal to 2.
    - Pleiotropy effect with nonlinear topic-disease effect. A quadratic term of topic-disease effect added to the second model.
    - Pleiotropy effect with nonlinear genetic-disease effect. A quadratic term of genetic-disease effect added to the second model.

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

Tables

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/Turbinate 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

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 Phecode system (see Figure 3).316 of 348 diseases analysed are associated to a topic; the remaining 32 diseases do not have a topic with average diagnosis-specific topic probability >50%.

Figures

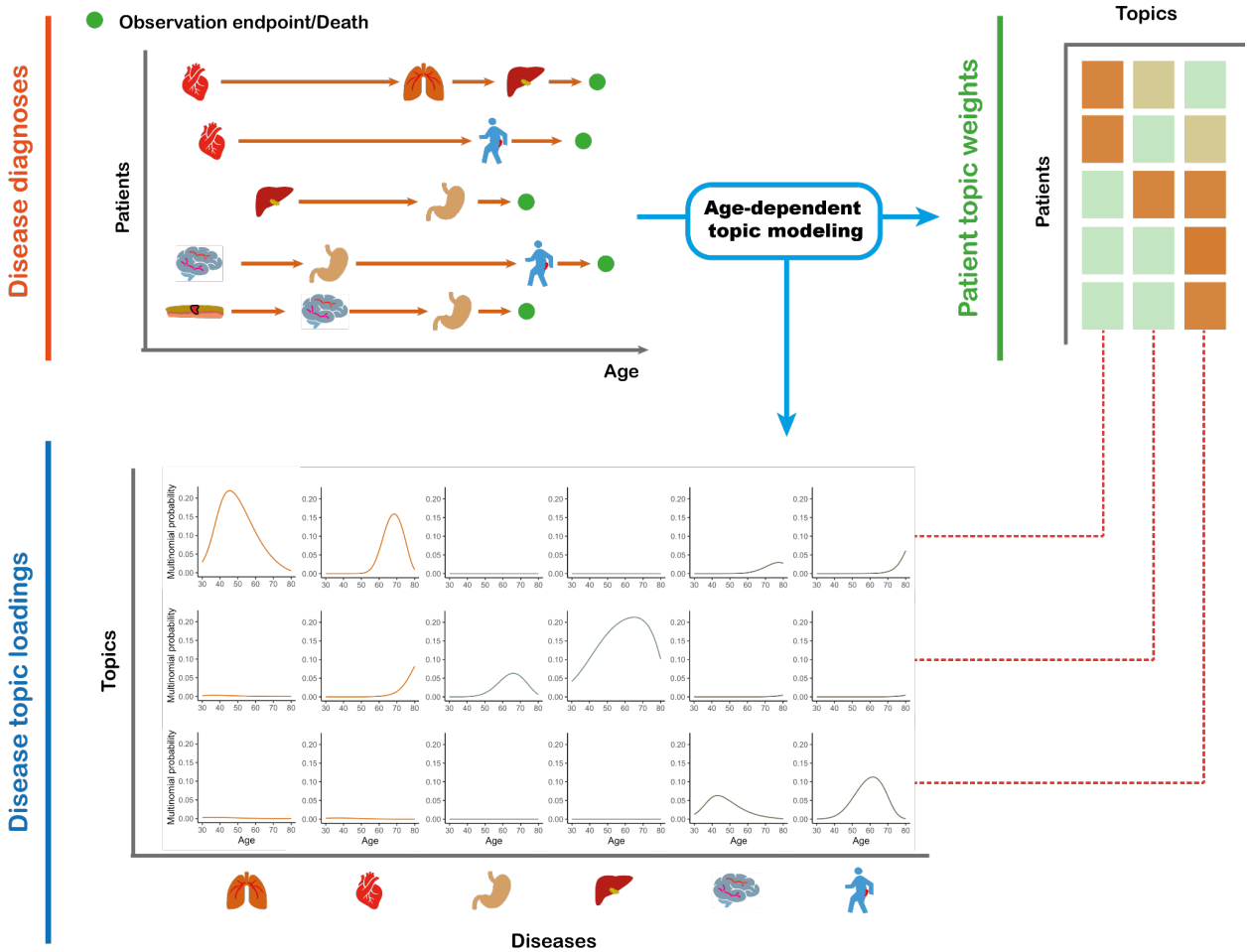


Figure 1: ATM provides an efficient way to represent longitudinal comorbidity data. Top left: input consists of disease diagnoses as a function of age. Top right: ATM assigns a topic weight to each patient. Bottom: ATM infers age-dependent topic loadings.

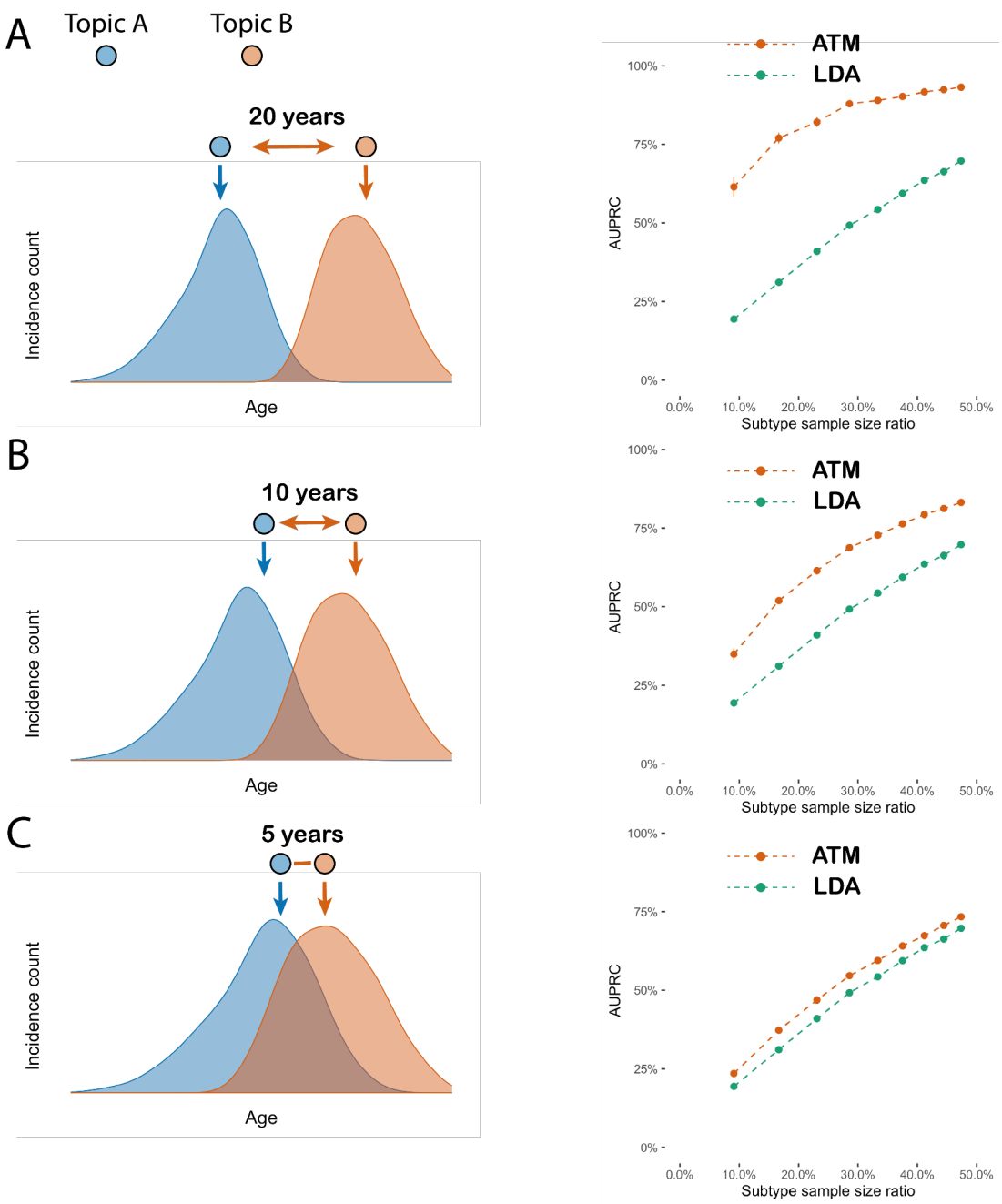


Figure 2: ATM outperforms LDA in simulations with age-dependent effects. In simulations at different levels of age-dependent effects (left panels), we report the area under the precision and recall curve (AUPRC) for ATM vs. LDA as a function of subtype sample size proportion (the proportion of diagnoses belonging to the smaller subtype) (right panels). Each dot represents the mean of 100 simulations of 10,000 individuals. Error bars denote 95% confidence intervals. (a) 20-year difference in age at diagnosis for the two subtypes. (b) 10-year difference in age at diagnosis for the two subtypes. (c) 5-year difference in age at diagnosis for the two subtypes. Numerical results are reported in Supplementary Table 1.

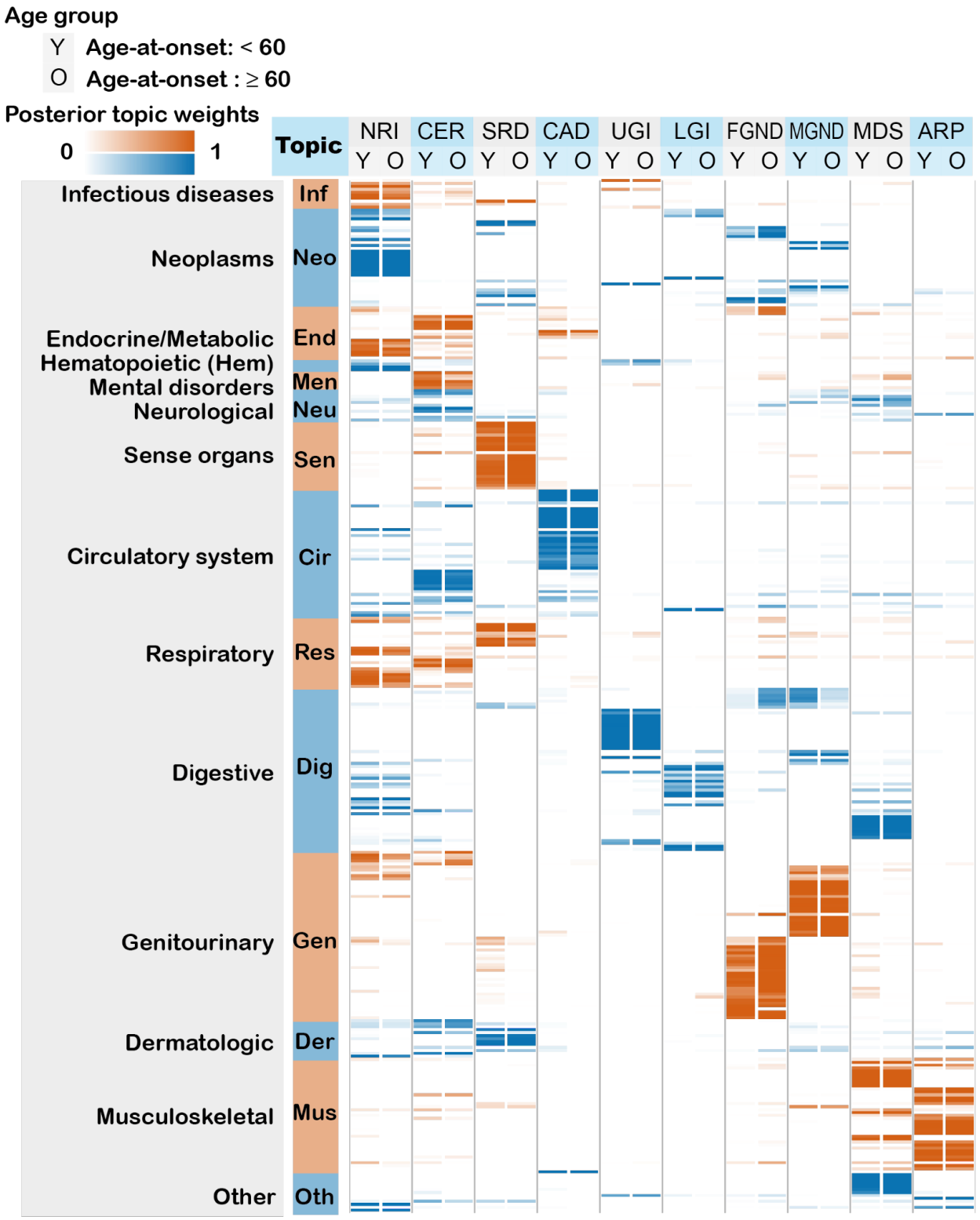


Figure 3. Age-dependent topic loadings of 10 inferred disease topics across 348 diseases in the UK Biobank. We report topic loadings averaged across younger ages (age at diagnosis < 60)

and older ages (age at diagnosis > 60). Row labels denote 13 Phecode systems, with alternating blue and red color for visualisation purposes; “Other” includes “congenital anomalies”, “symptoms”, “injuries & poisoning”, “other tests”, and “death”. Topics are ordered by Phecode system. Further details on the 10 topics are provided in Table 1. Further details on the diseases discussed in the text (e.g. breast cancer) are provided in Supplementary Figure 9. Numerical results are reported in Supplementary Table 3.

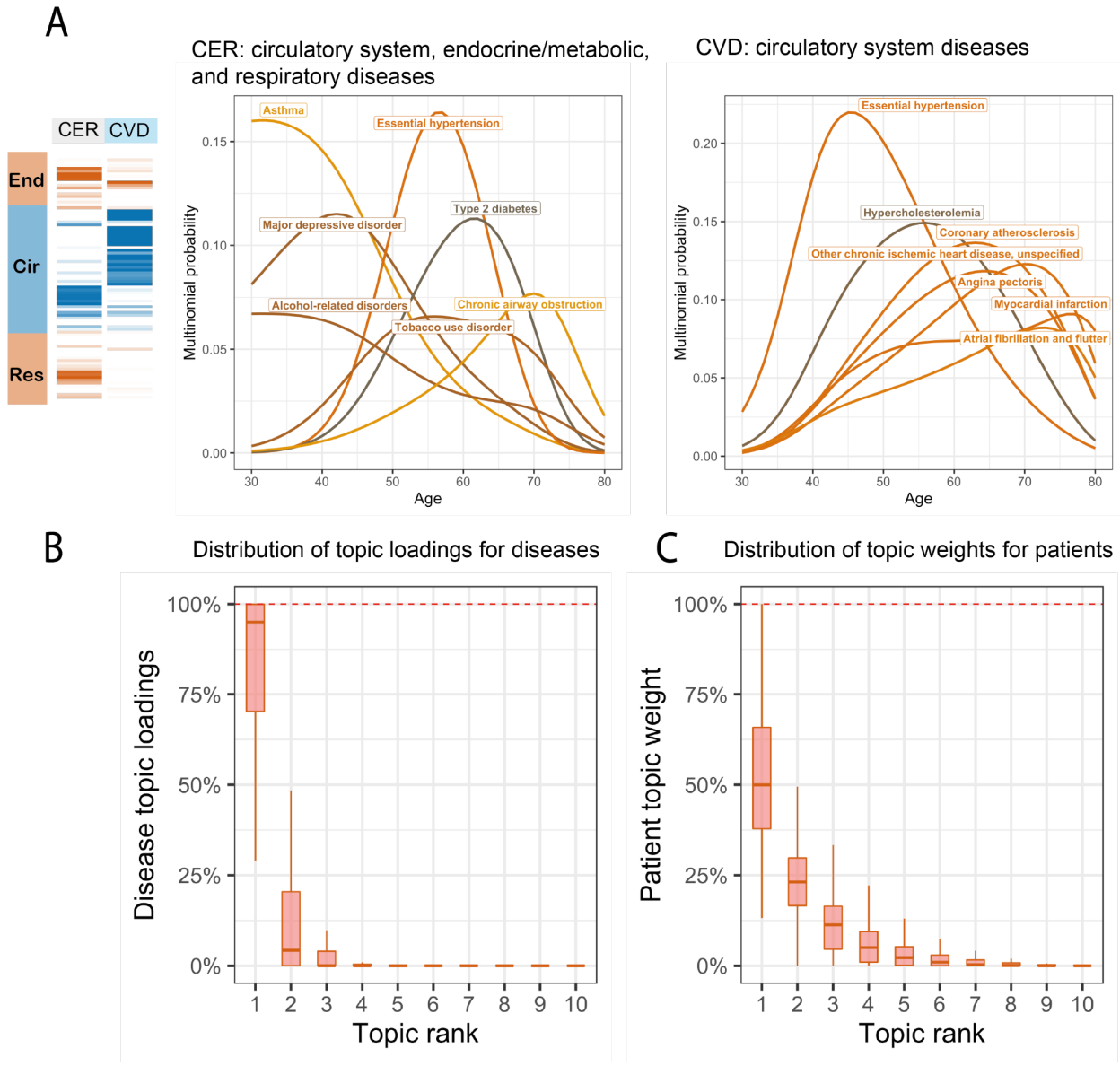


Figure 4. Topic loadings capture age-dependent comorbidities. (a) Age-dependent topic loadings for two representative topics, CER and CVD; for each topic, we include all diseases with high topic loadings. Results for all 10 topics are reported in Supplementary Figure 12. (b) Box plot of disease topic loading as a function of rank; disease topic loadings are computed as a weighted average across all values of age at diagnosis. (c) Box plot of patient topic weight as a function of rank. Numerical results are reported in Supplementary Table 4.

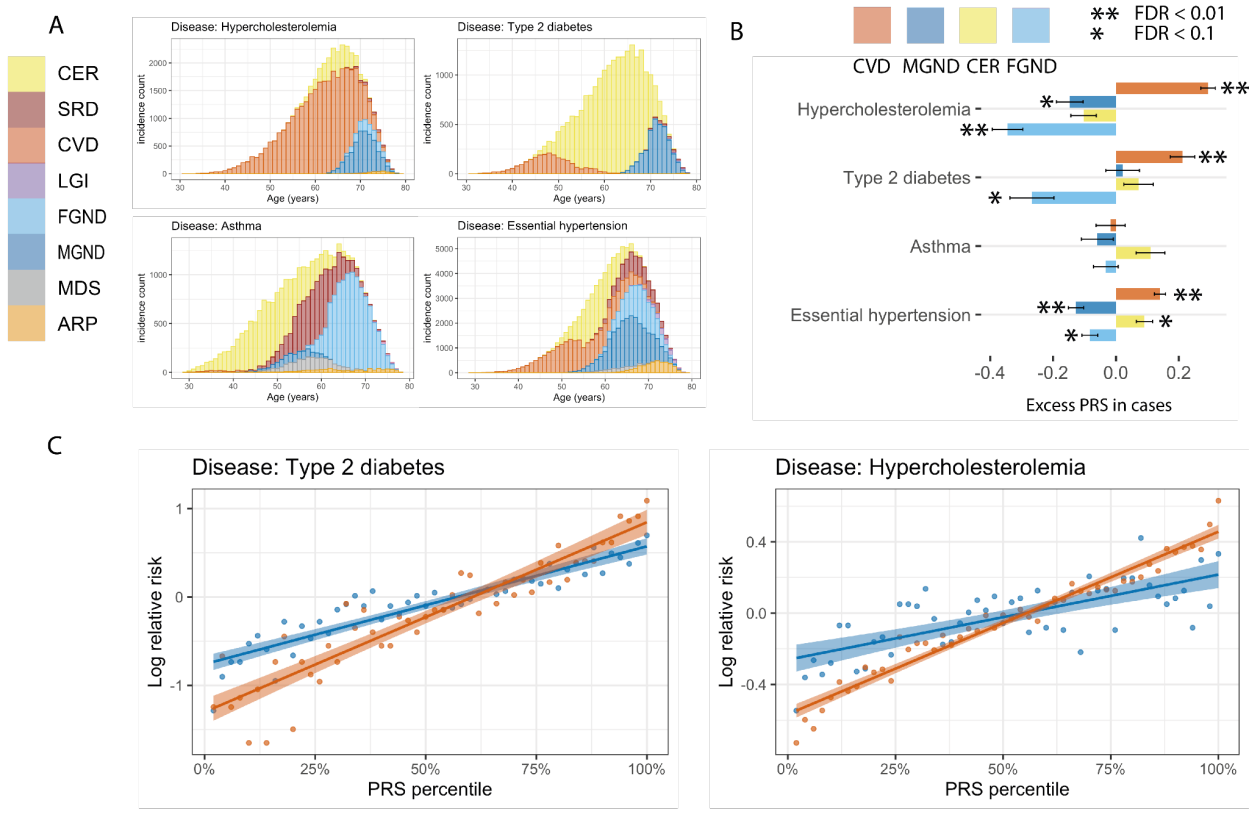


Figure 5. Polygenic risk scores vary across disease subtypes defined by distinct topics. (a) Stacked barplots of age-dependent topic loadings for 4 representative diseases (type 2 diabetes, asthma, hypercholesterolemia, and essential hypertension); for each disease, we include all topics with high topic loadings. Results for all 52 diseases are reported in Supplementary Figure 15. (b) Standardised excess PRS values in disease cases (s.d. increase in PRS per s.d. increase in patient topic weight) for 4 representative diseases and 4 corresponding topics. (c) Relative risk for cases of type 2 diabetes and hypercholesterolemia of CVD and MGND subtypes (vs. controls) across PRS percentiles. Each spans 2 PRS percentiles. Lines denote regression on log scale. Error bars denote 95% confidence intervals. Color codings are as in (a) and (b) (red for CVD and blue for MGND). Numerical results are reported in Supplementary Table 6.

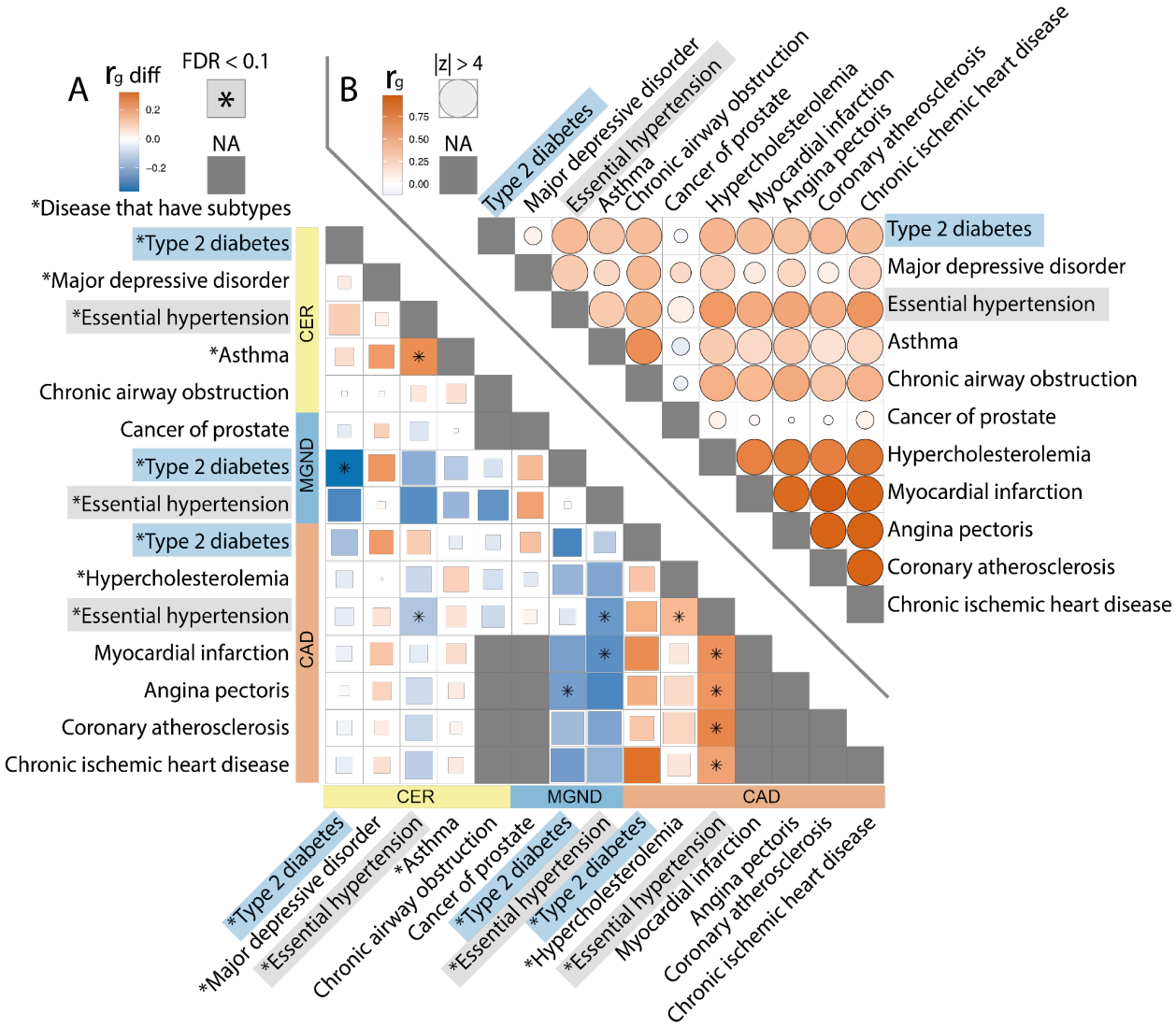


Figure 6. Genetic correlations vary across disease subtypes defined by distinct topics. (a) Excess genetic correlations for 15 disease subtypes (relative to genetic correlations between the underlying diseases). Full square with star denotes FDR < 0.1; less than full squares have area proportional to z-scores for difference. Grey squares denote NA (disease pairs that are both without subtypes). (b) Genetic correlations between the underlying diseases. Full circle denotes |z-score| > 4 for nonzero genetic correlation; less than full circles have area proportional to |z-score|. Numerical results are reported in Supplementary Table 8.

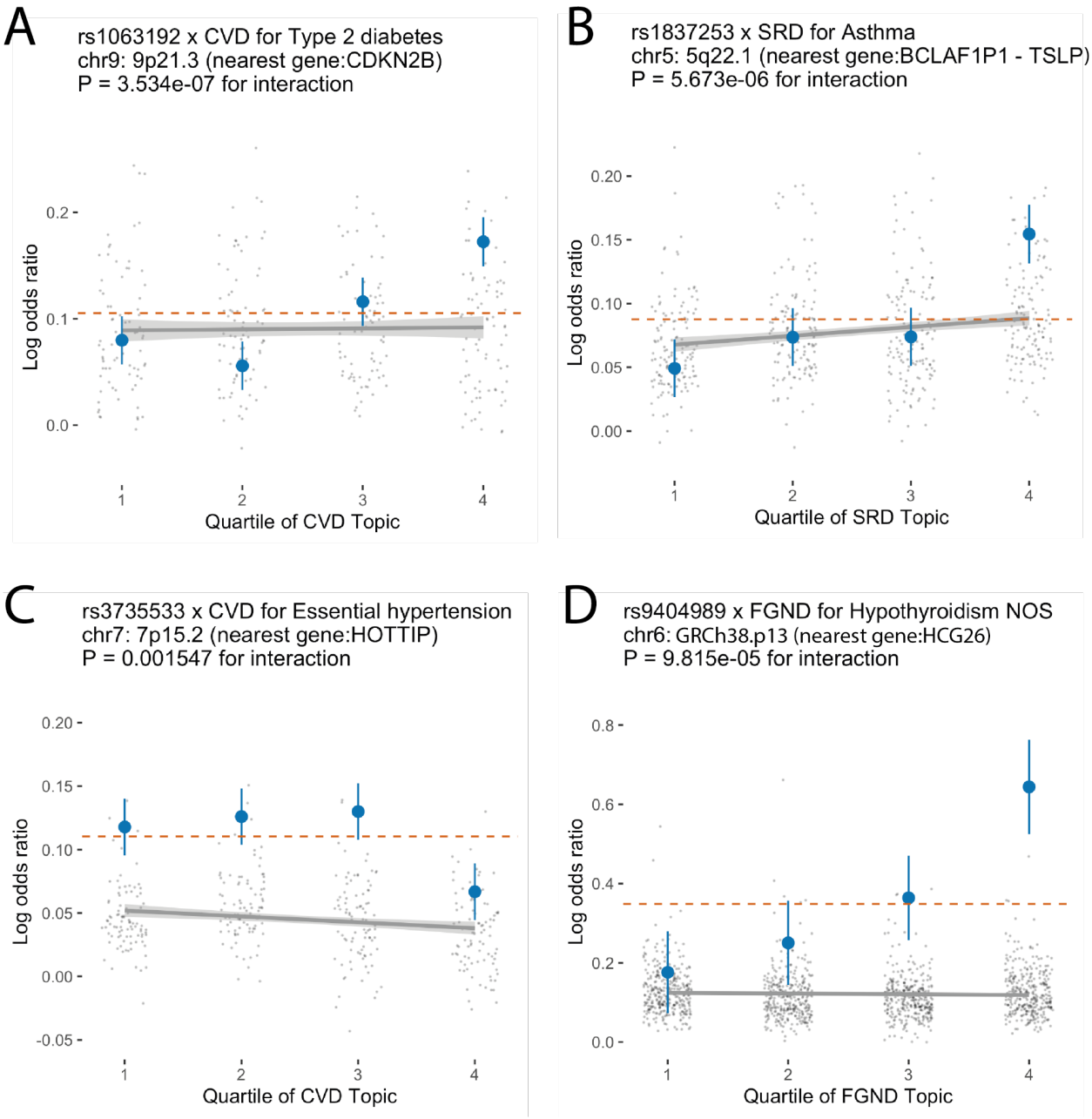


Figure 7. Examples of SNP x topic interaction effects on disease phenotypes. For each example, we report main SNP effects (log odds ratios) specific to each quartile of topic weights across individuals, for both the focal SNP (blue dots) and background SNPs for that disease and topic (genome-wide significant 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.

References

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

autism spectrum disorder, attention-deficit hyperactivity disorder and case subgroups. *Nat. Genet.* **54**, 1470–1478 (2022).

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

22. Liu, Z. *et al.* Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. *arXiv [cs.AI]* (2018).
23. Chen, Y. & Chi, Y. Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix Estimation: Recent Theory and Fast Algorithms via Convex and Nonconvex Optimization. *IEEE Signal Process. Mag.* **35**, 14–31 (2018).
24. Chen, Y. & Candès, E. J. The projected power method: An efficient algorithm for joint alignment from pairwise differences. *Commun. Pure Appl. Math.* **71**, 1648–1714 (2018).
25. Jia, G. *et al.* Estimating heritability and genetic correlations from large health datasets in the absence of genetic data. *Nat. Commun.* **10**, 1–11 (2019).
26. Tanigawa, Y. *et al.* Components of genetic associations across 2,138 phenotypes in the UK Biobank highlight adipocyte biology. *Nat. Commun.* **10**, 4064 (2019).
27. Sakaue, S. *et al.* A cross-population atlas of genetic associations for 220 human phenotypes. *Nat. Genet.* **53**, 1415–1424 (2021).
28. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
29. of Us Research Program, A. The 'All of Us' Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
30. Ishigaki, K. *et al.* Large-scale genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases. *Nat. Genet.* **52**, 669–679 (2020).
31. Siggaard, T. *et al.* Disease trajectory browser for exploring temporal, population-wide disease progression patterns in 7.2 million Danish patients. *Nat. Commun.* **11**, 1–10 (2020).
32. Posey, J. E. *et al.* Resolution of Disease Phenotypes Resulting from Multilocus Genomic Variation. *N. Engl. J. Med.* **376**, 21–31 (2017).
33. Cook, E. K. *et al.* Comorbid and inflammatory characteristics of genetic subtypes of clonal hematopoiesis. *Blood Adv* **3**, 2482–2486 (2019).

34. Udler, M. S. *et al.* Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Med.* **15**, e1002654 (2018).

35. Wani, B., Aziz, S. A., Ganaie, M. A. & Mir, M. H. Metabolic Syndrome and Breast Cancer Risk. *Indian J. Med. Paediatr. Oncol.* **38**, 434–439 (2017).

36. Blei, Ng & Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.* (2003).

37. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

38. Bishop, C. M. *Pattern Recognition and Machine Learning*. (Springer New York, 2006).

39. Teh, Y., Newman, D. & Welling, M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Adv. Neural Inf. Process. Syst.* **19**, (2006).

40. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics* **31**, 2595–2597 (2015).

41. Wu, P. *et al.* Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. *JMIR Med Inform* **7**, e14325 (2019).

42. Wong, B. *et al.* Cardiovascular Disease Risk Associated With Familial Hypercholesterolemia: A Systematic Review of the Literature. *Clin. Ther.* **38**, 1696–1709 (2016).

43. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).

44. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

45. Jiang, X., Holmes, C. & McVean, G. The impact of age on genetic risk for common diseases. *PLoS Genet.* **17**, e1009723 (2021).

46. Weir, B. S. & Cockerham, C. C. ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE. *Evolution* **38**, 1358–1370 (1984).

47. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A. L. Estimating and interpreting FST: the impact of rare variants. *Genome Res.* **23**, 1514–1521 (2013).

48. Blei, D. M. & Lafferty, J. D. A correlated topic model of Science. *Ann. Appl. Stat.* **1**, 17–35 (2007).

49. Zaitlen, N. *et al.* Informed conditioning on clinical covariates increases power in case-control association studies. *PLoS Genet.* **8**, e1003032 (2012).

50. Sun, B. B. *et al.* Genetic regulation of the human plasma proteome in 54,306 UK Biobank participants. *bioRxiv* 2022.06.17.496443 (2022) doi:10.1101/2022.06.17.496443.

51. Mostafavi, H. *et al.* Variable prediction accuracy of polygenic scores within an ancestry group. *Elife* **9**, e48376 (2020).

52. Dumitrescu, L. *et al.* Evidence for age as a modifier of genetic associations for lipid levels. *Ann. Hum. Genet.* **75**, 589–597 (2011).

53. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).

54. Lin, J. *et al.* Integration of biomarker polygenic risk score improves prediction of coronary heart disease in UK Biobank and FinnGen. *bioRxiv* (2022) doi:10.1101/2022.08.22.22279057.

55. Ghorbani, B., Javadi, H. & Montanari, A. An Instability in Variational Inference for Topic Models. in *Proceedings of the 36th International Conference on Machine Learning* (eds. Chaudhuri, K. & Salakhutdinov, R.) vol. 97 2221–2231 (PMLR, 09--15 Jun 2019).

56. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).

57. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).

58. Marisa, L. *et al.* Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med.* **10**, e1001453 (2013).

59. Crockett, S. D. & Nagtegaal, I. D. Terminology, Molecular Features, Epidemiology, and Management of Serrated Colorectal Neoplasia. *Gastroenterology* **157**, 949–966.e4 (2019).

60. Zhang, H.-C. *et al.* Identification and Validation in a Novel Quantification System of Ferroptosis Patterns for the Prediction of Prognosis and Immunotherapy Response in Left- and Right-Sided Colon Cancer. *Front. Immunol.* **13**, 855849 (2022).

61. Harbeck, N. & Gnant, M. Breast cancer. *Lancet* **389**, 1134–1150 (2017).

62. Gao, J. J. & Swain, S. M. Luminal A Breast Cancer and Molecular Assays: A Review. *Oncologist* **23**, 556–565 (2018).

63. Morra, A. *et al.* Breast Cancer Risk Factors and Survival by Tumor Subtype: Pooled Analyses from the Breast Cancer Association Consortium. *Cancer Epidemiol. Biomarkers Prev.* **30**, 623–642 (2021).

64. Lim, E. H. *et al.* Cervical dysplasia: assessing methylation status (Methylight) of CCNA1, DAPK1, HS3ST2, PAX1 and TFPI2 to improve diagnostic accuracy. *Gynecol. Oncol.* **119**, 225–231 (2010).

65. Ondič, O. *et al.* Bizarre cell dysplasia of the cervix. *J. Obstet. Gynaecol. Res.* **43**, 345–351 (2017).

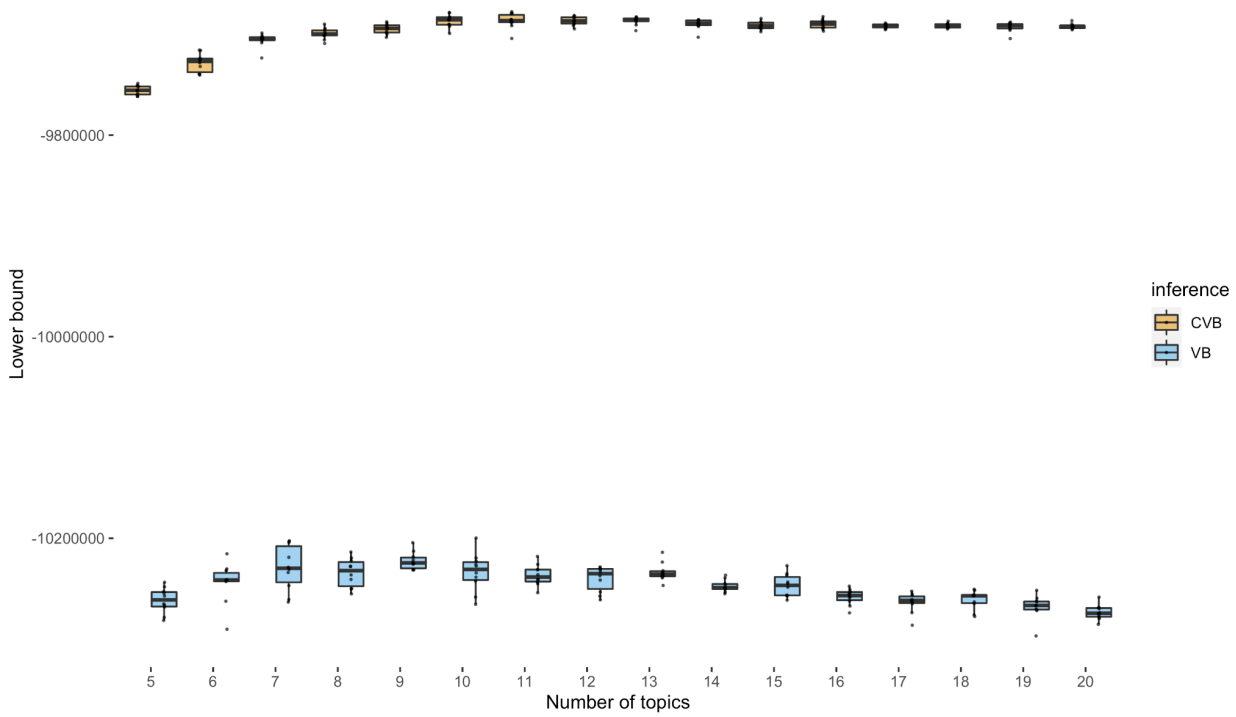
66. Kim, H. I. *et al.* Non-immune-related hypothyroidism and its relationship with excess iodine. *Eur. J. Nutr.* **58**, 2851–2858 (2019).

67. Mastrototaro, L. & Roden, M. Insulin resistance and insulin sensitizing agents. *Metabolism* **125**, 154892 (2021).

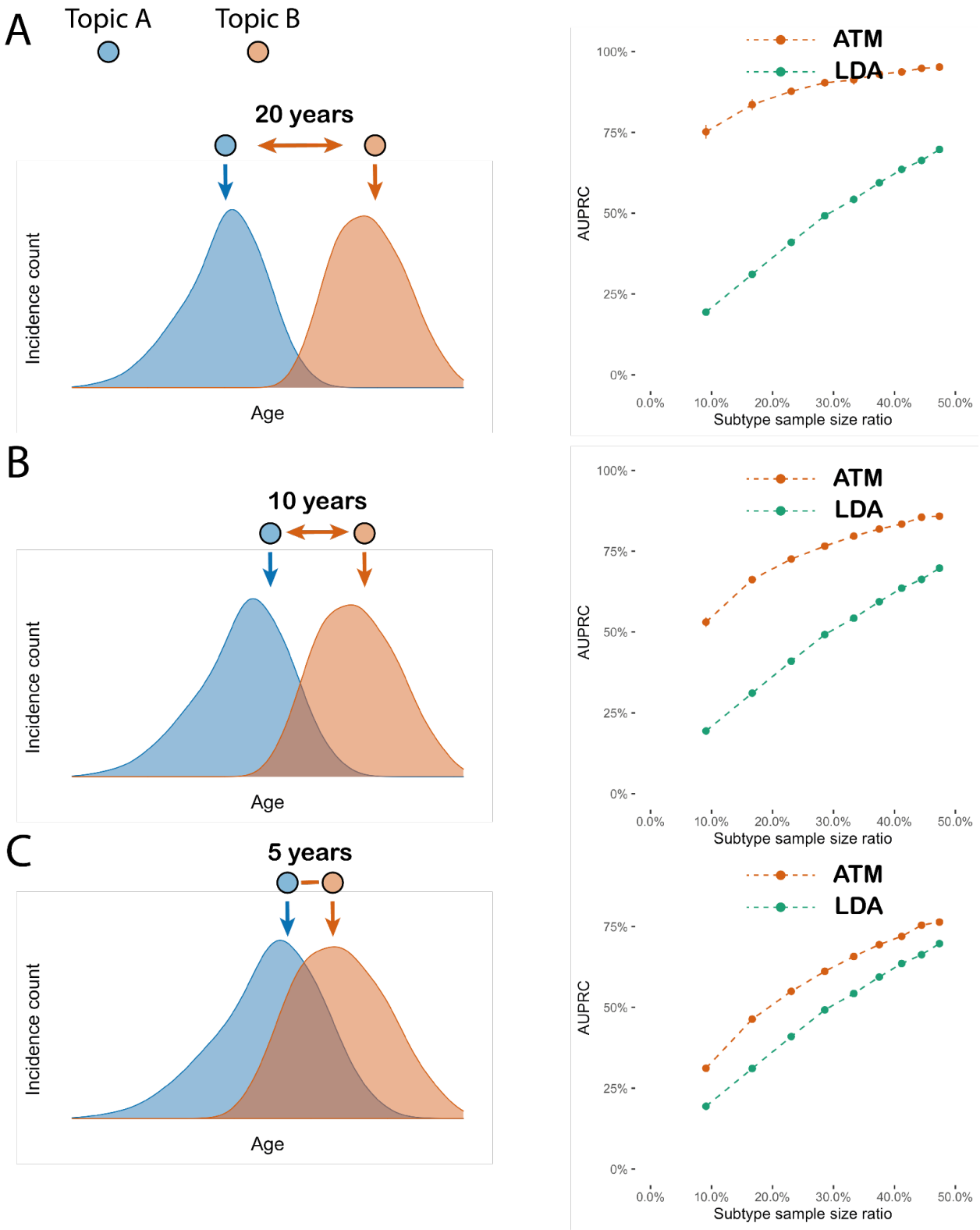
68. Mariam, A. *et al.* A Type 2 Diabetes Subtype Responsive to ACCORD Intensive Glycemia Treatment. *Diabetes Care* **44**, 1410–1418 (2021).

69. Geiss, H. C., Schwandt, P. & Parhofer, K. G. Influence of simvastatin on LDL-subtypes in patients with heterozygous familial hypercholesterolemia and in patients with diabetes mellitus and mixed hyperlipoproteinemia. *Exp. Clin. Endocrinol. Diabetes* **110**, 182–187 (2002).

Supplementary Figures

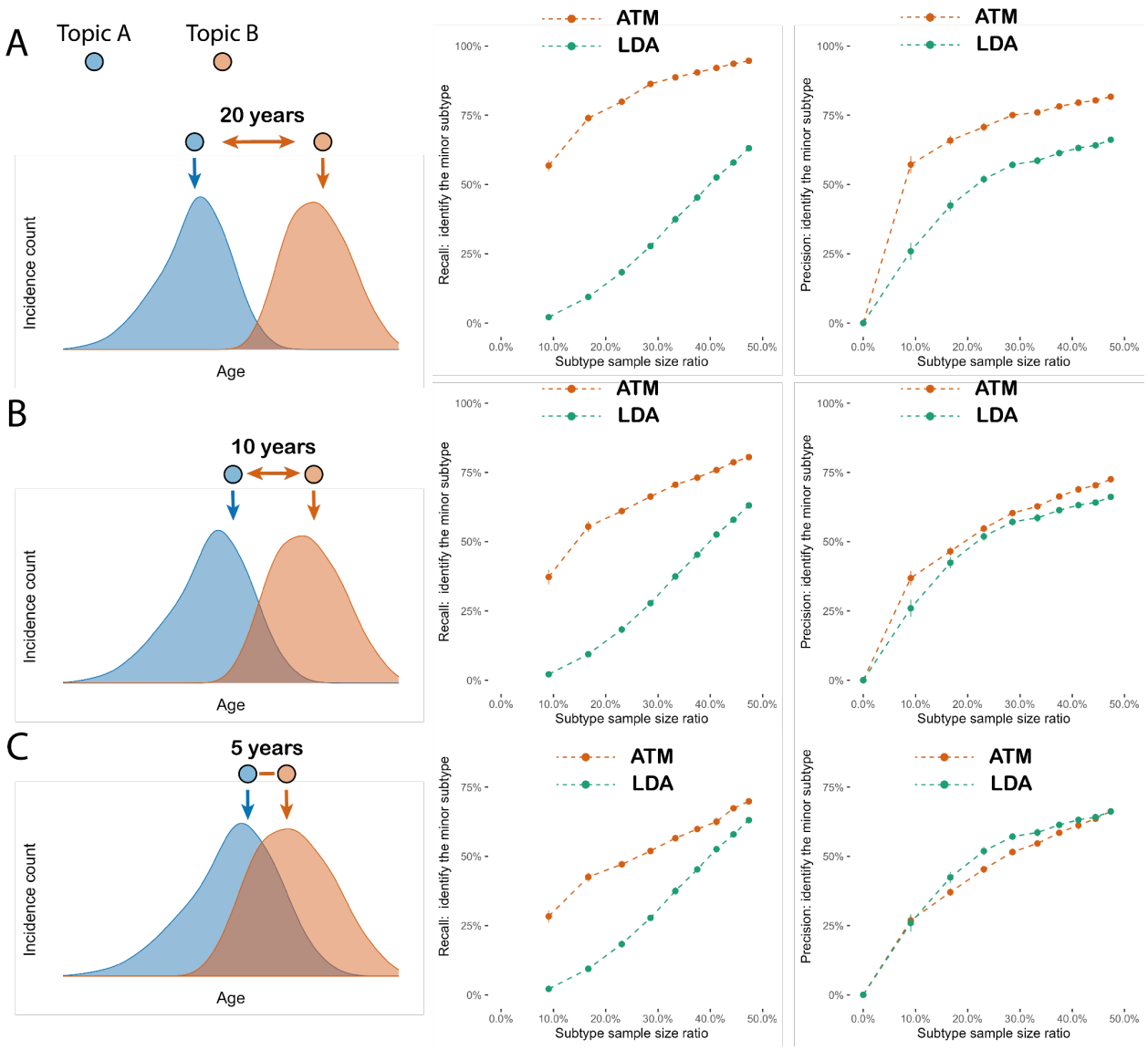


Supplementary Figure 1. Comparison of ELBO for collapsed variational inference and mean-field variational inference. ELBO is computed from the entire UK Biobank dataset and with cubic polynomials. Model with each number of topics is inferred with 10 random initialisations with both CVB and the VB (mean-field).



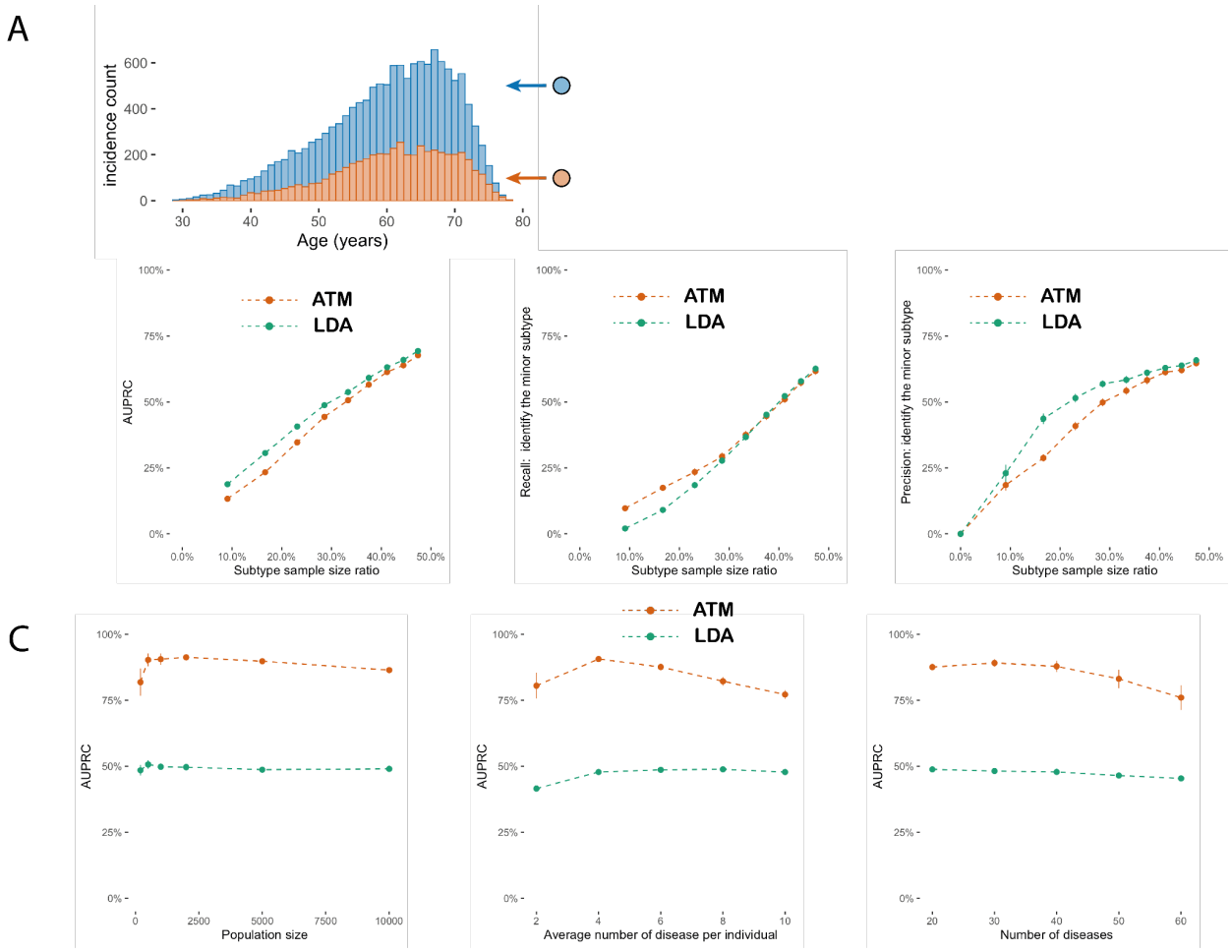
Supplementary Figure 2. Same analysis as in Figure 2 but classifying the second disease. The area under precision and recall curve (AUPRC) to correctly assign incident disease to correct comorbidity profiles using Latent Dirichlet Allocation (LDA) and our methods. X-axis refers to the proportion of cases that belong to the smaller subtype; precision and recall are computed for

classifying the incidences in the smaller subgroup. Each dot represents the mean of 100 simulations of 10,000 people, the bar shows the 95% confidence intervals. Red refers to the ATM and green refers to the LDA model. (a) Scenario where two subtypes are simulated with 20 years of difference in age-at onset. (b) Scenario where two subtypes are simulated with 10 years of difference in age-at onset. (c) Scenario where two subtypes are simulated with 5 years of age difference.

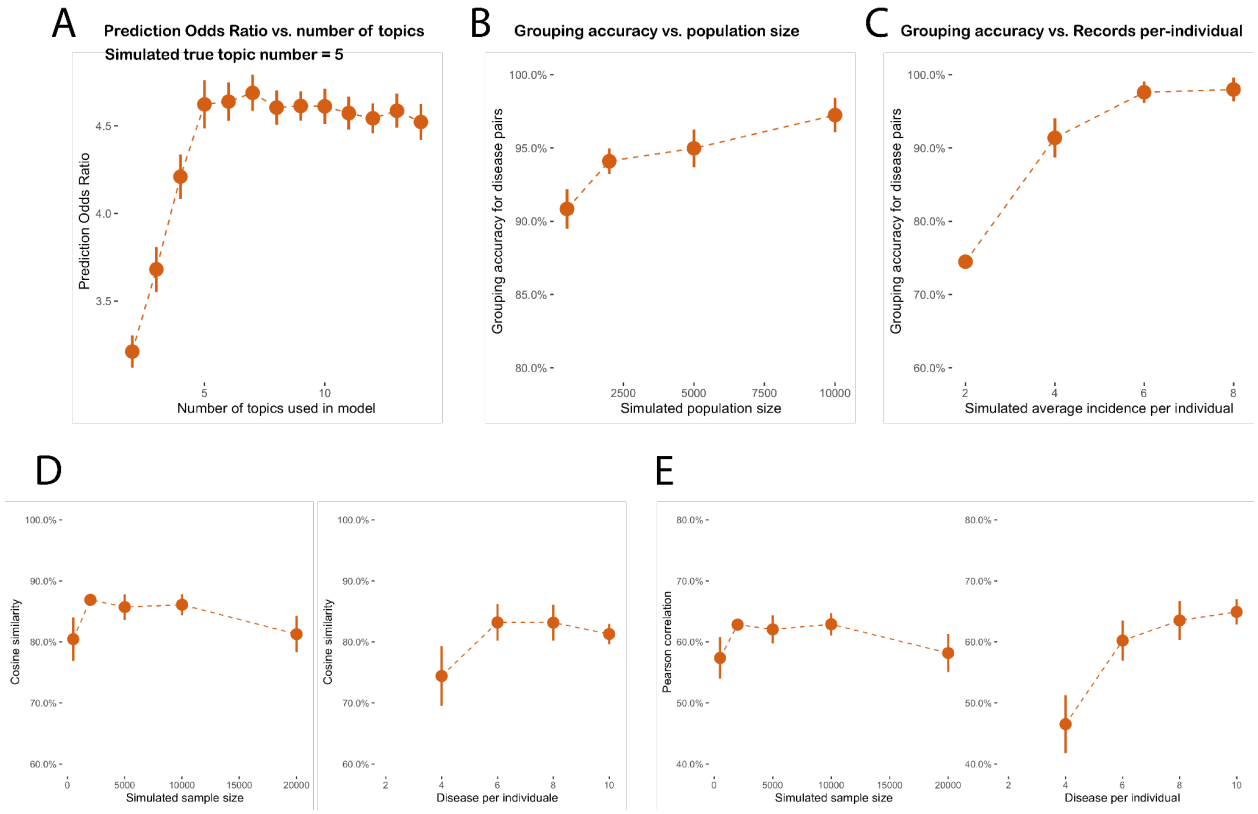


Supplementary Figure 3. Additional simulation studies established the power of the method to identify comorbidity. (a) The precision and recall rate to correctly assign incident disease to correct comorbidity profiles using Latent Dirichlet Allocation (LDA) and our methods. X-axis refers to the size ratio of the small subgroup and the large subgroup, precision and recall are computed for the label incidences in the small subgroup. Each dot represents the mean of 100 simulations of 10,000 people, the bar shows the 95% confidence intervals. Red refers to the

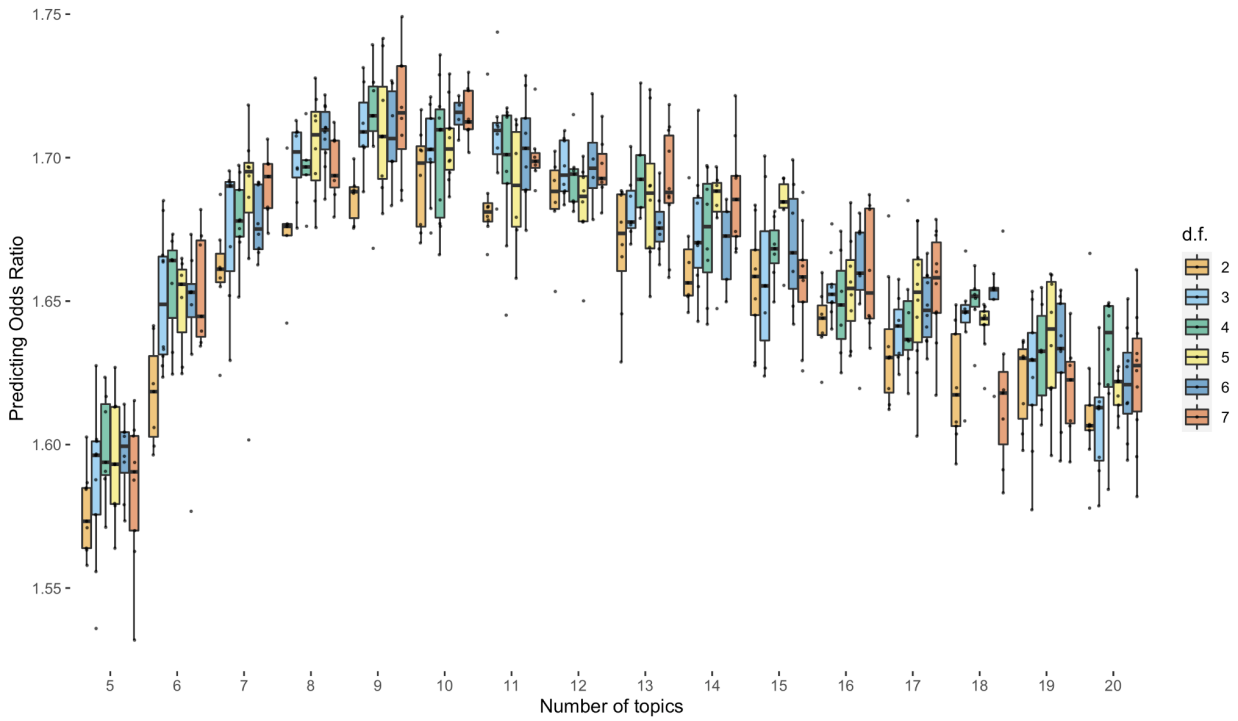
ATM and green refers to the LDA model. (b) Scenario where two subtypes are simulated with 10 years of difference in age-at onset. (c) Scenario where two subtypes are simulated with 5 years of age difference.



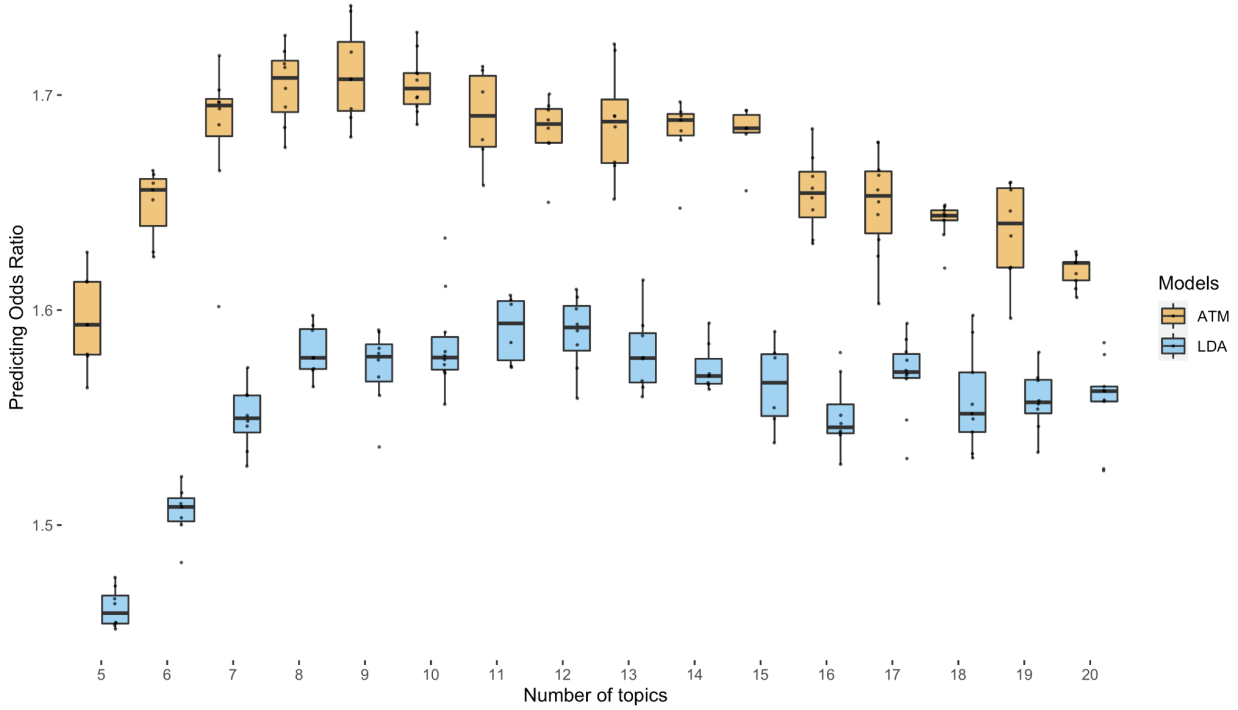
Supplementary Figure 4. Additional simulation studies established the power of the method to identify comorbidity.(a) Same analysis as panel a but simulated subtypes with same age at diagnosis distribution. LDA outperforms ATM slightly as we have additional regularisation when modelling topic loading as functions of age, while for LDA age is not modelled. (b) AUPRC computed as Figure 2A with varying population size, average number of diseases per individual, and number of distinct diseases.



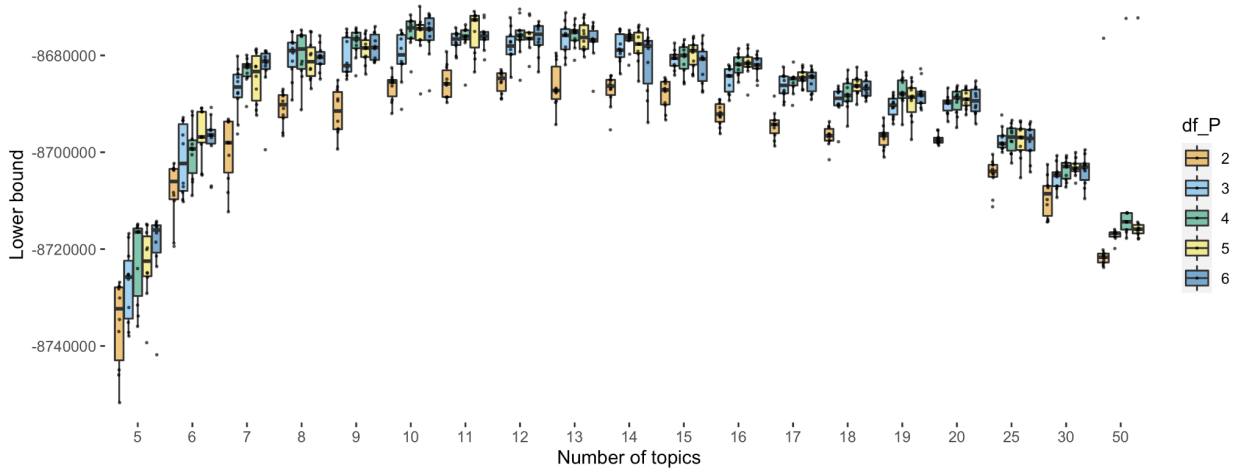
Supplementary Figure 5. Confirmatory simulations. (a) We simulated data using 5 topics while fitting models of varying topic numbers. We use 80% of data as training data and computed prediction odds ratio in the held out data, where we use the topic loading computed from the training data and prior diseases to infer the topic weights to predict the target diseases. The simulation was performed for 20 replications for each topic number in the inference. (b-c) We assign each disease to a single topic based on topic loading and compute the grouping accuracy as the proportion of disease pairs that are correctly grouped to the same topic. The grouping accuracy remains high for varying simulated population size and average disease per individual. (d) Recovery of topic loadings. We evaluate the accuracy of topic loading inference by computing the cosine similarity between inferred topic loading with the underlying truth. We match the inferred topics with the true topics using correlation of topic weights, using a greedy procedure (matching the first inferred topic from all true topics and then matching the next topic from the remaining not-matched true topics) to ensure the matching is bijectively. (e) Recovery of topic weights. We evaluate the accuracy of topic weight inference by computing the correlation of inferred topic weights and with the underlying truth.



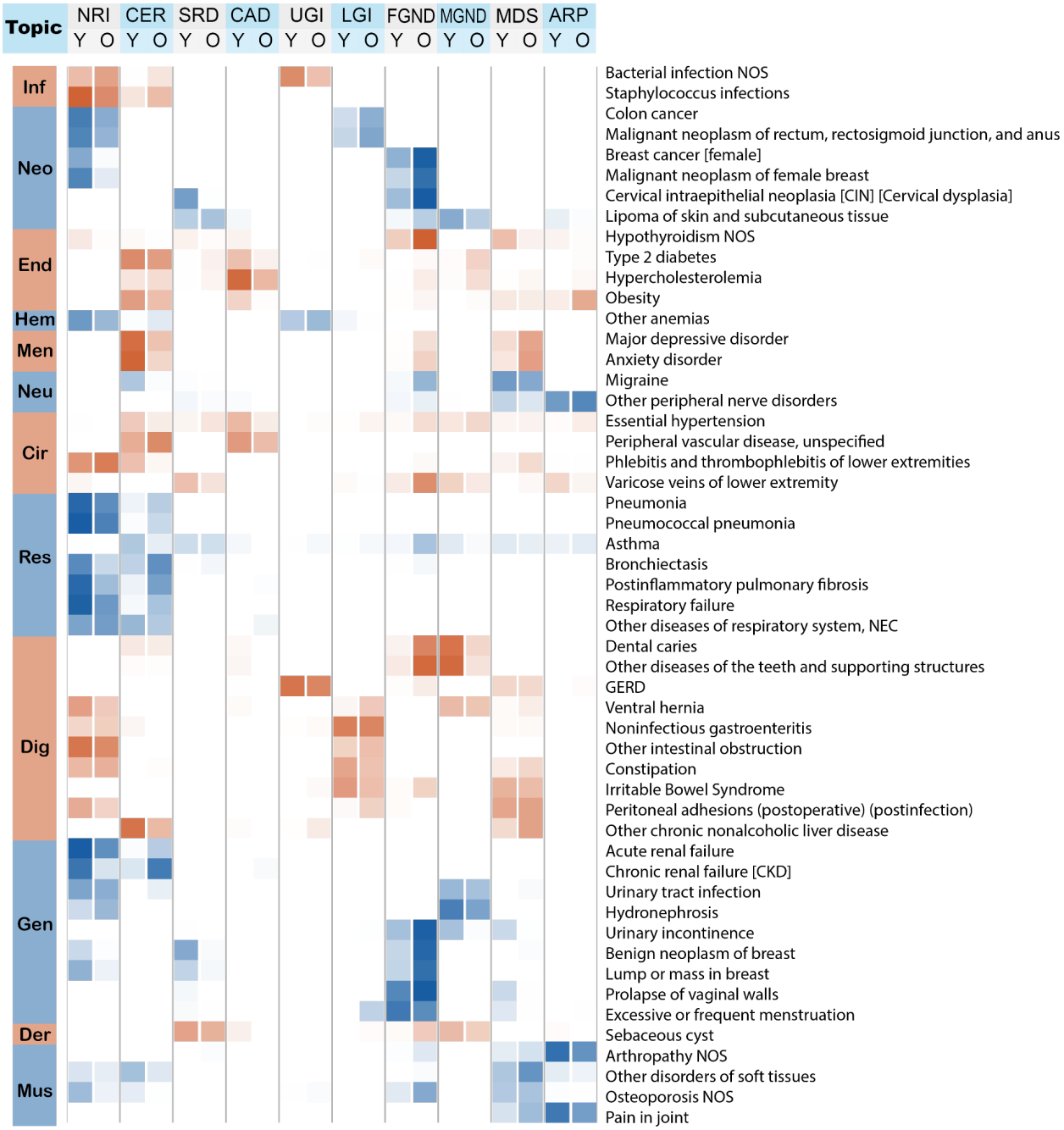
Supplementary Figure 6. Prediction odds ratio of different model configurations. Each dot represents one inference on a random training and testing split. The models are run with different topic numbers and configurations of age risk profiles. Degree of freedom (d.f.) from 2 to 7 represent linear, quadratic polynomial, cubic polynomial, spline with one knot, spline with two knots, and spline with three knots.The prediction odds ratios are between model-predicted top percentile disease set versus the prevalence-ordered top percentile disease set.



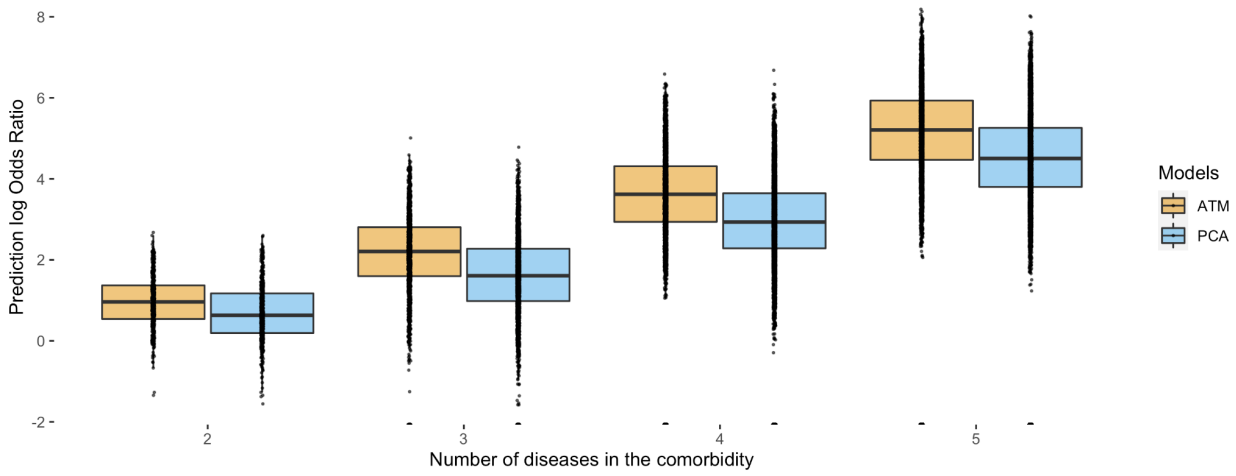
Supplementary Figure 7. Comparison of prediction odds ratio between LDA and ATM. Each dot represents one inference on a random training and testing split. The models are run with different topic numbers and we choose a cubic spline with one knot for configuring ATM age profiles. The prediction odds ratios are between the model-predicted top percentile disease set versus the prevalence-ordered top percentile disease set.



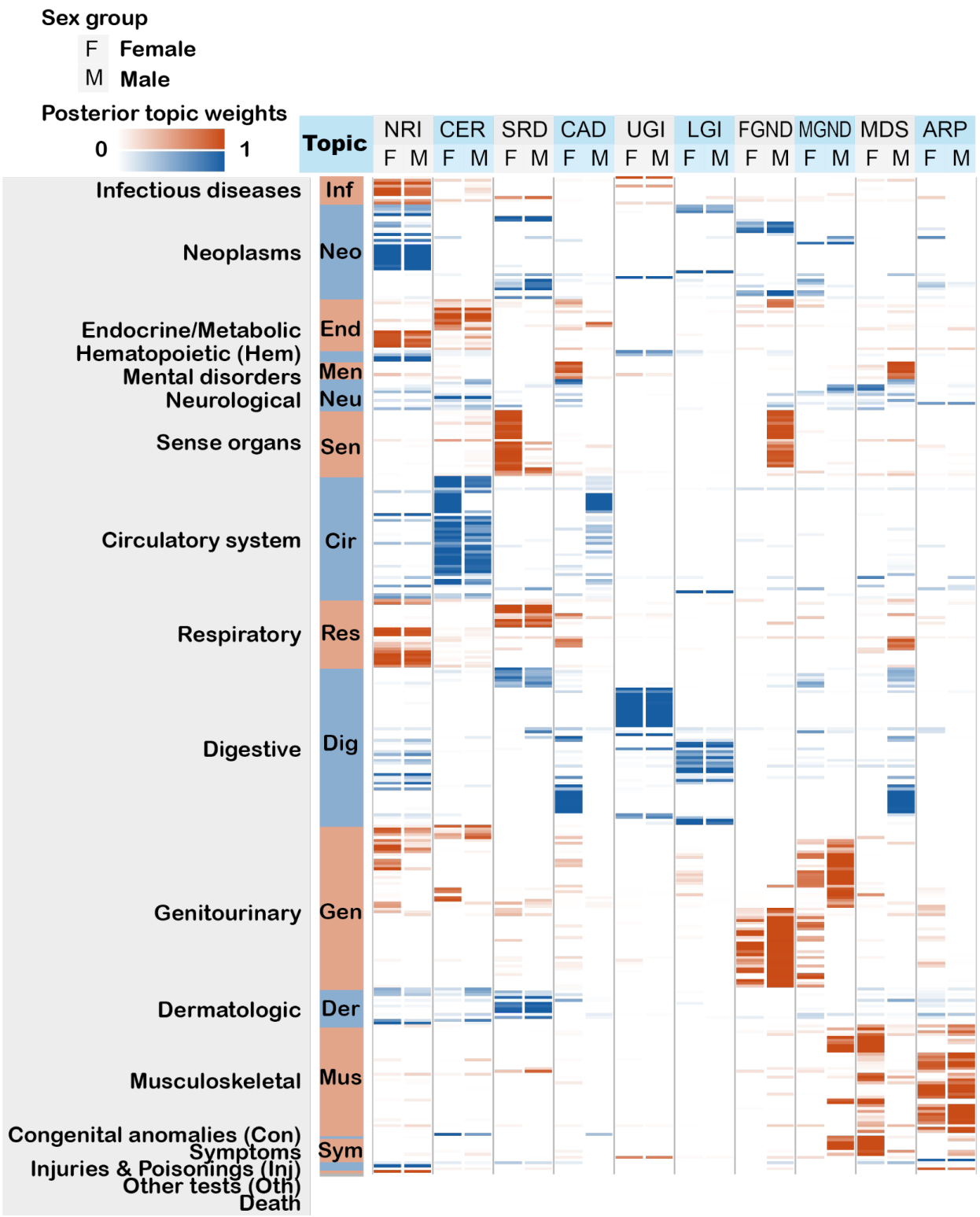
Supplementary Figure 8. Evidence lower bound of different model configurations on the entire dataset. Each dot represents an inference with random initialisation. The models are run with different topic numbers and configurations of age risk profiles. Degrees of freedom (d.f.) from 2 to 6 represent linear, quadratic polynomial, cubic polynomial, spline with one knot, and spline with two knots.



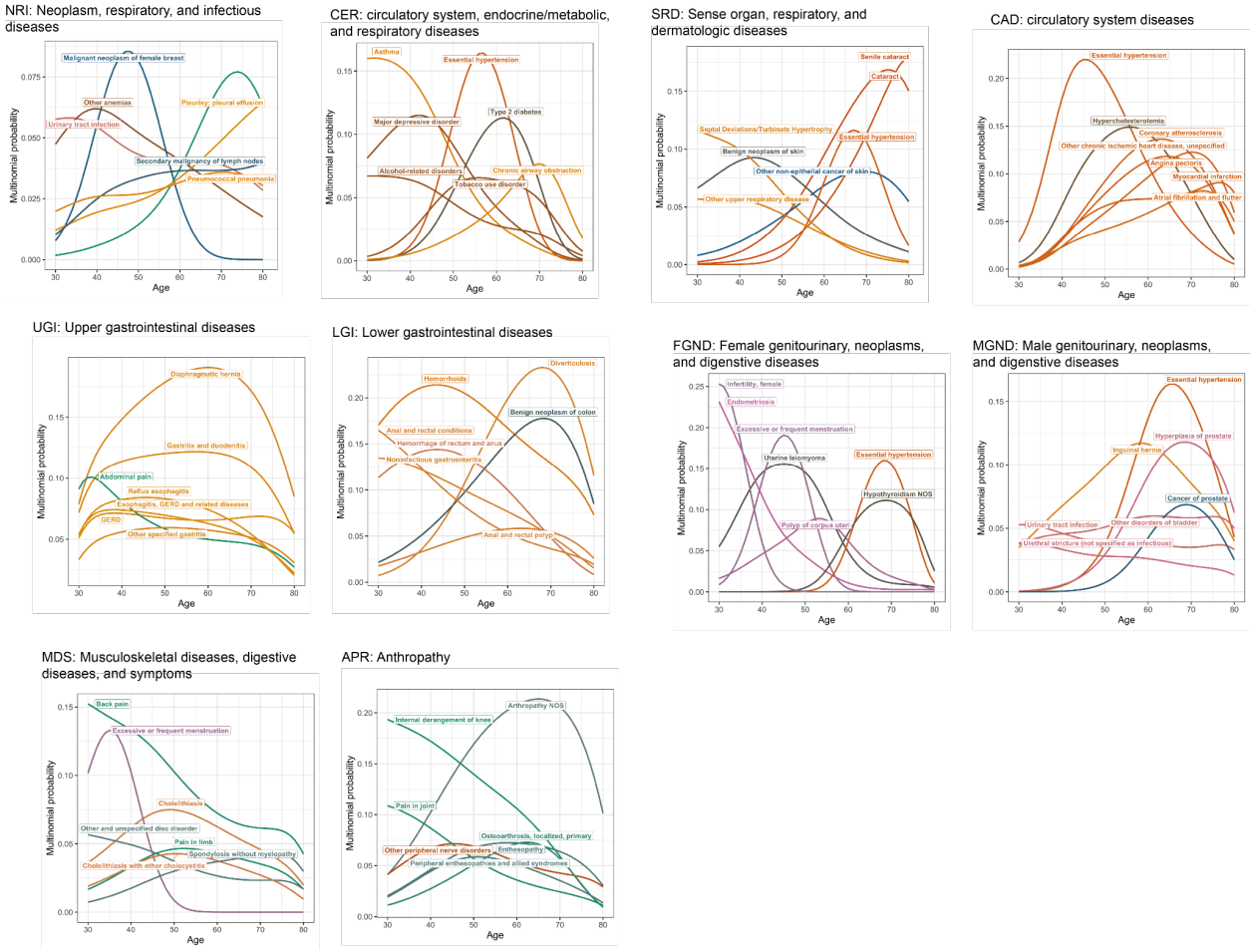
Supplementary Figure 9. Posterior topic distributions are different between age groups for diseases that have subtypes. The figure is the same as Figure 3A but focusing on 52 diseases that have a subtype with at least 500 incidences.



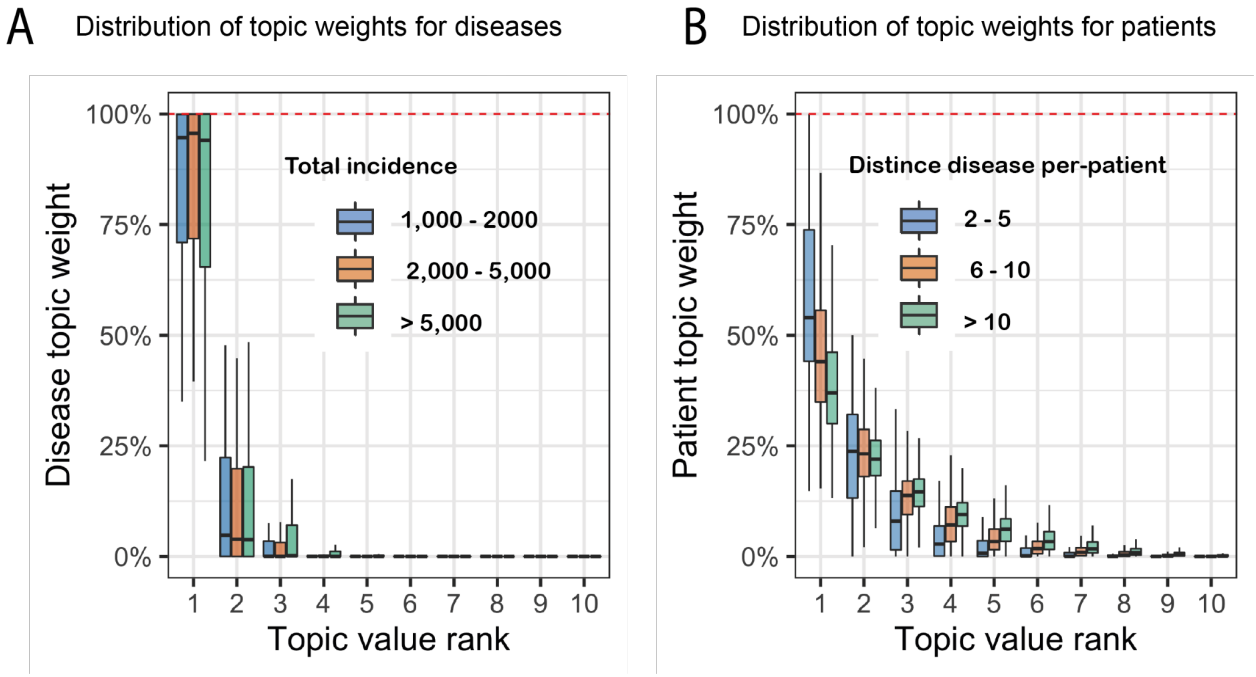
Supplementary Figure 10. Prediction log odds ratio of comorbidities. Diseases combinations (comorbidities) are extracted from topics that are trained on a training set. The odds ratio of each disease combination is computed by comparing the prevalence of cases that have all diseases in the combination versus that implied when all diseases occur independently.



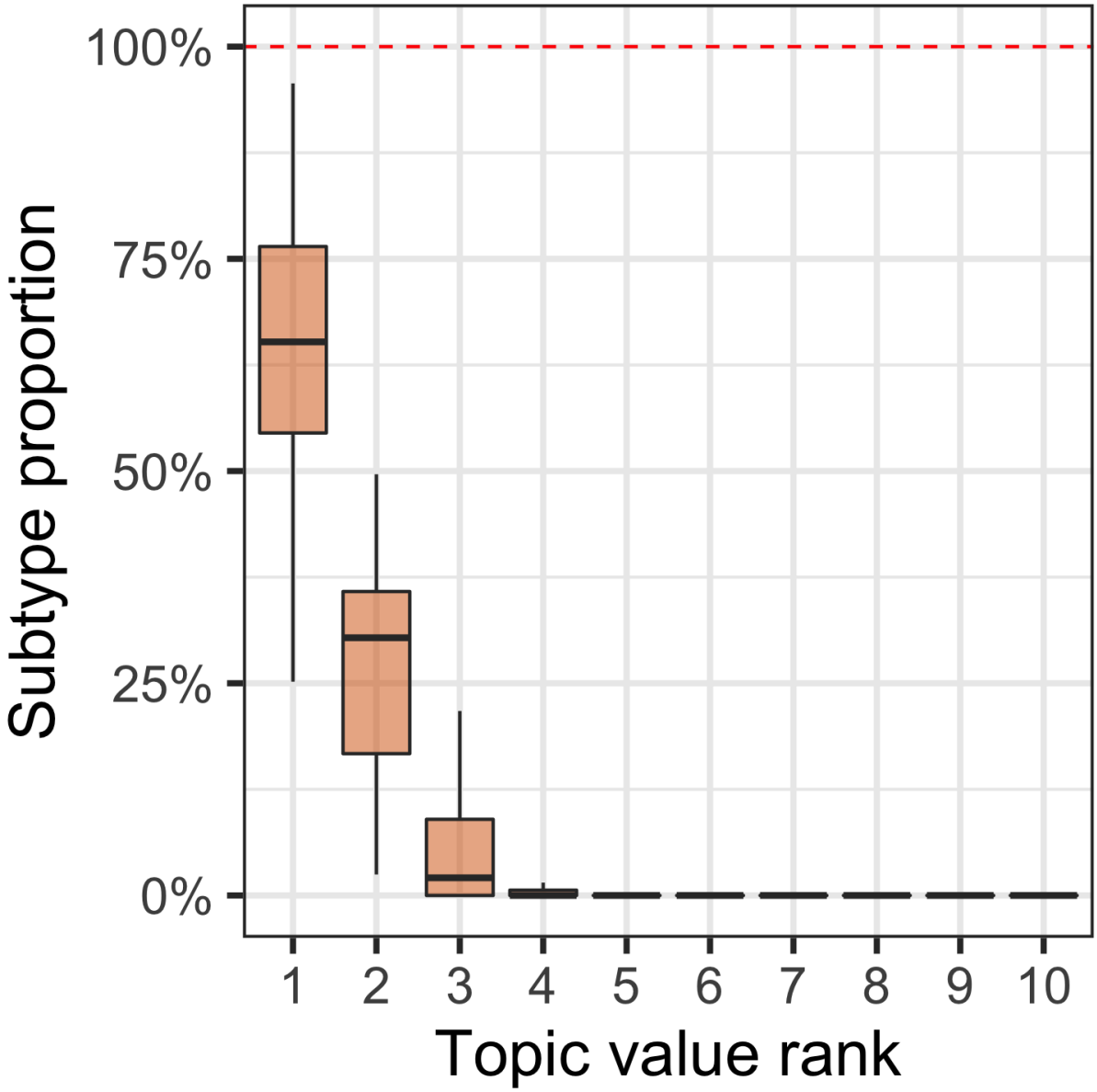
Supplementary Figure 11: Posterior topic distributions of female and male populations. The figure is the same as Figure 3A but comparing the topics that are inferred from female and male populations separately.



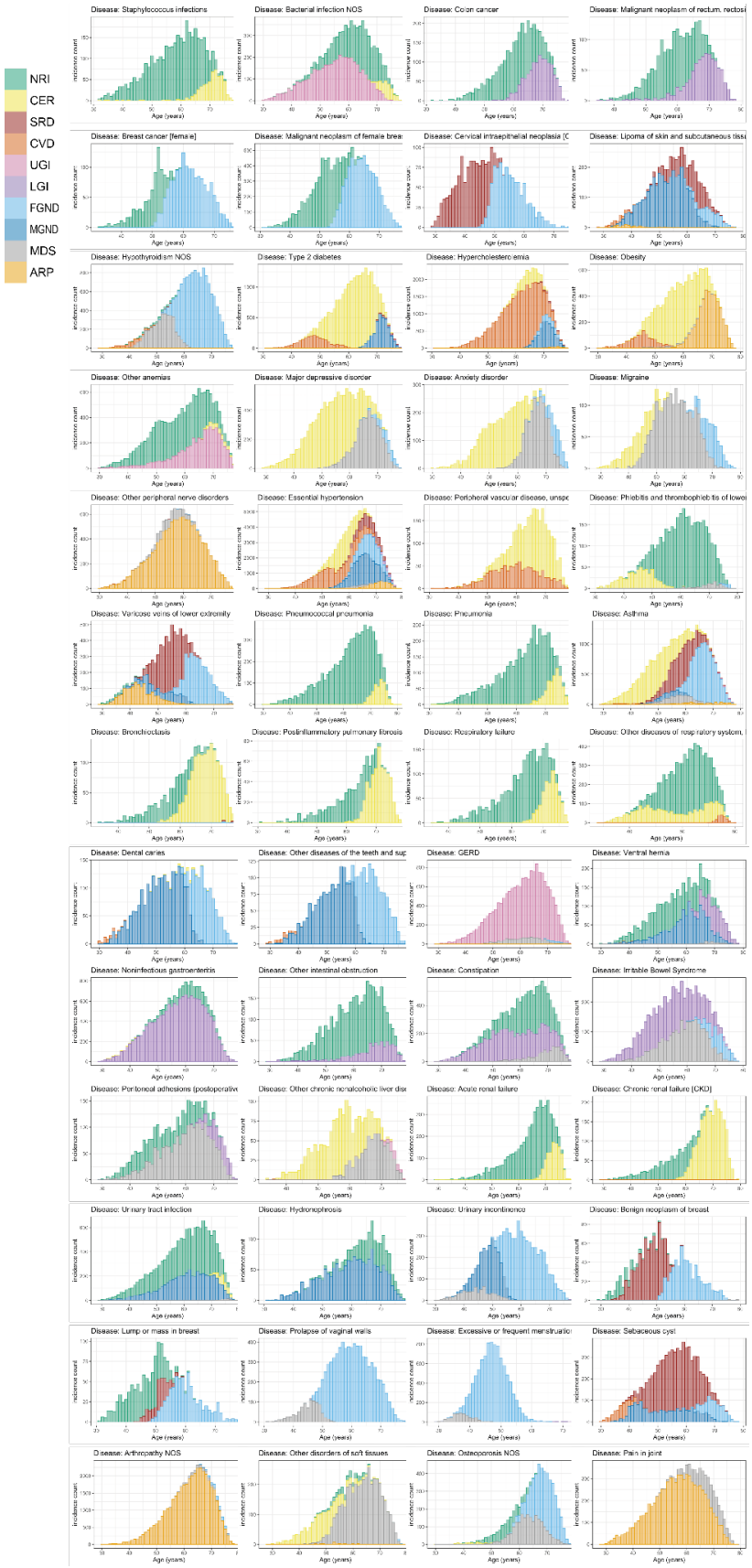
Supplementary Figure 12. Top seven diseases in each comorbidity topic. Seven diseases that have highest loading within the topic are shown for each comorbidity topic. Colour of the curves reflect the ordering of Phecodes.



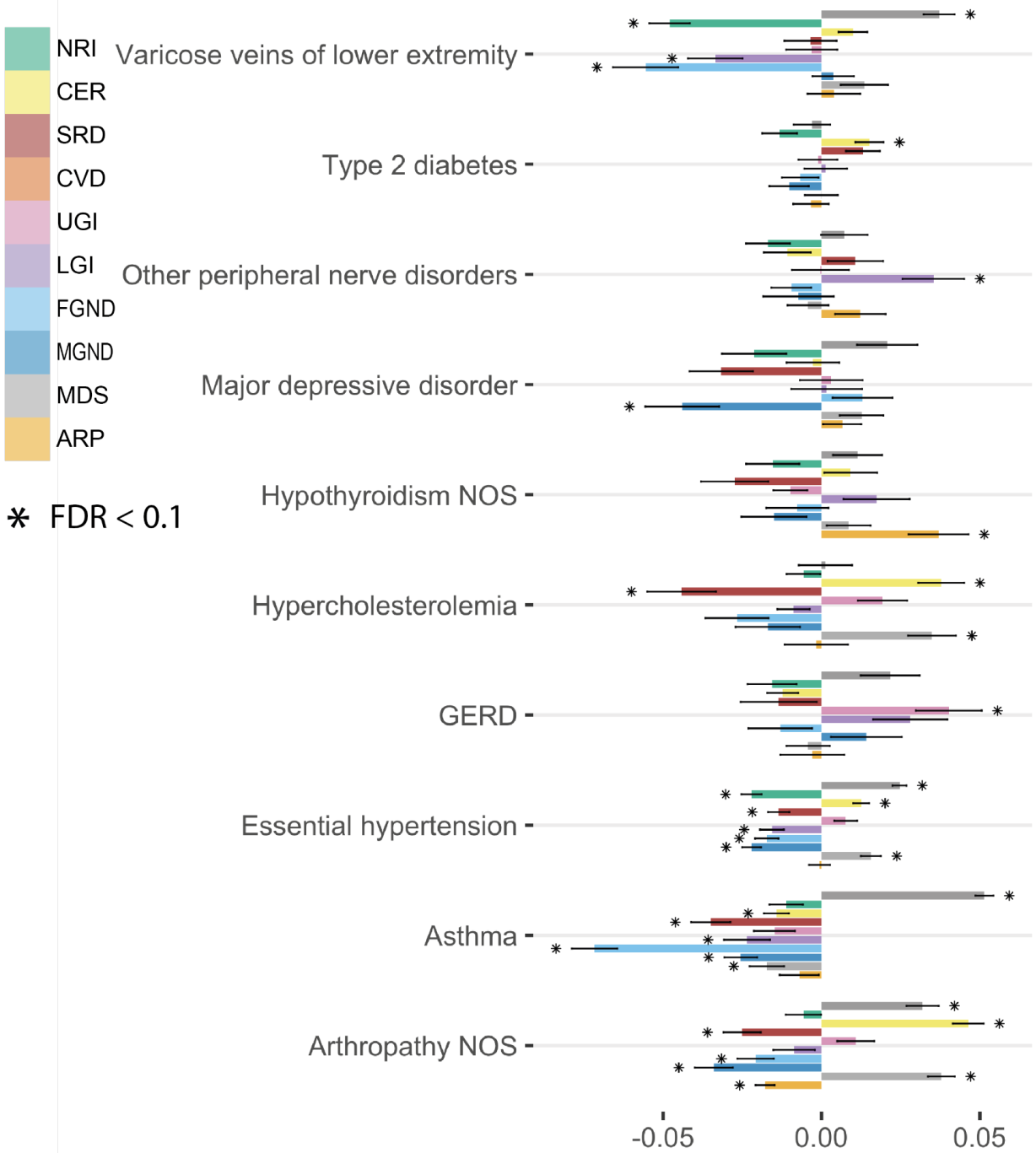
Supplementary Figure 13. Additional topic sparsity analysis. (a) Sparsity of disease topic loadings. Box plot shows the distribution of topic loading for disease of different incidence numbers. (b) Sparsity of patient topic loading. Box plot shows the topic loading distribution in decreasing order for individuals with different numbers of diagnosis.



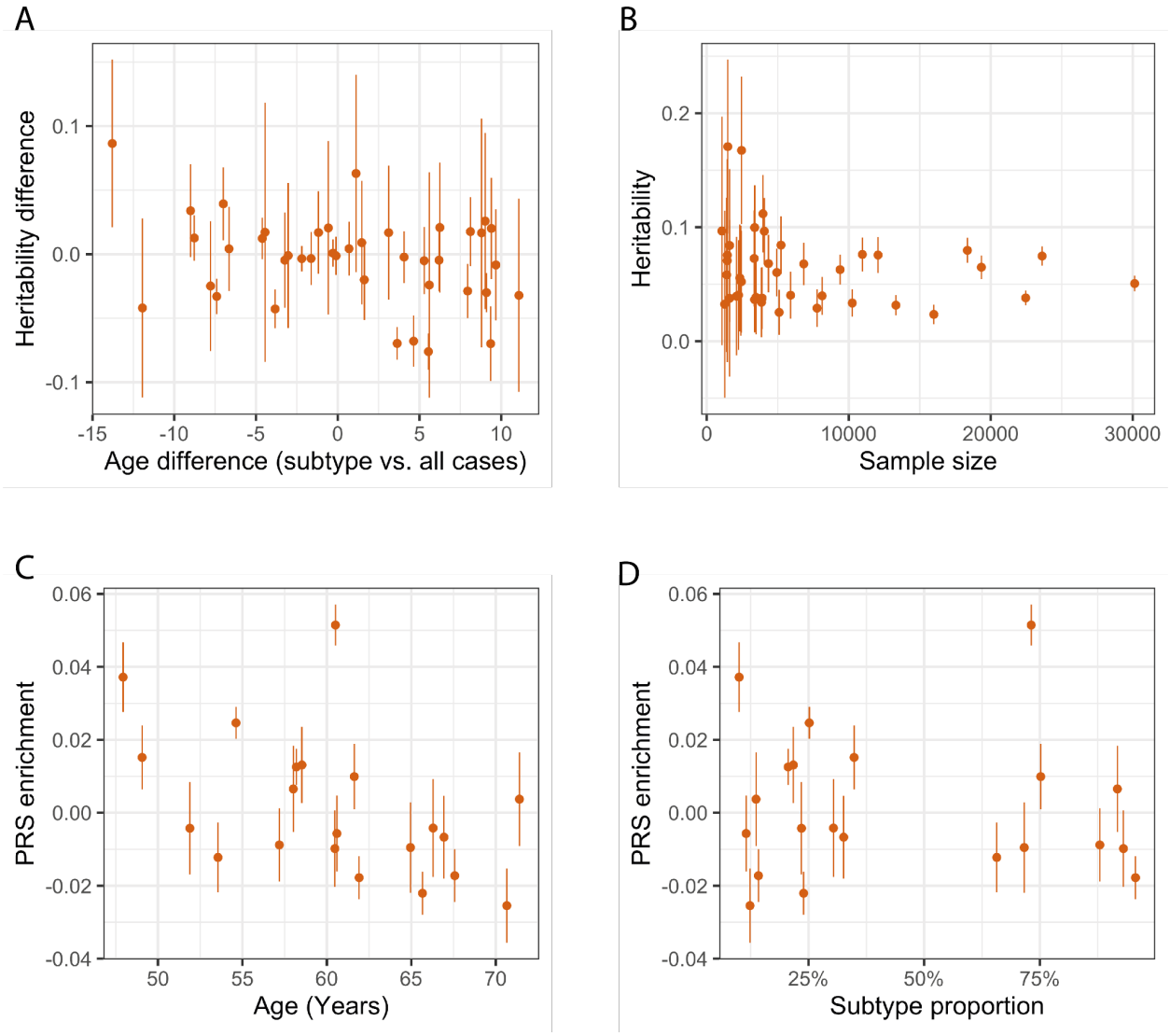
Supplementary Figure 14. Topic weight distribution for 52 diseases that have at least 500 cases assigned to distinct topics. The box plot shows the distribution of the subtype proportion from the largest (leftmost boxes) to the smallest. For nearly all diseases, the cases are concentrated into three subtypes, with very few cases assigned to other topics.


NRI
CER
SRD
CVD
UGI
LGI
FGND
MGND
MDS
ARP
Disease: Staphylococcus infections
Disease: Bacterial infection NOS
Disease: Colon cancer
Disease: Breast cancer [female]
Disease: Hypothyroidism NOS
Disease: Type 2 diabetes
Disease: Hypercholesterolemia
Disease: Obesity
Disease: Other anemias
Disease: Major depressive disorder
Disease: Anxiety disorder
Disease: Migraine
Disease: Other peripheral nerve disorders
Disease: Essential hypertension
Disease: Varicose veins of lower extremity
Disease: Pneumococcal pneumonia
Disease: Pneumonia
Disease: Asthma
Disease: Bronchiectasis
Disease: Respiratory failure
Disease: Dental caries
Disease: GERD
Disease: Ventral hernia
Disease: Noninfectious gastroenteritis
Disease: Other intestinal obstruction
Disease: Constipation
Disease: Irritable Bowel Syndrome
Disease: Acute renal failure
Disease: Chronic renal failure [CKD]
Disease: Urinary tract infection
Disease: Hydronephrosis
Disease: Urinary incontinence
Disease: Benign neoplasm of breast
Disease: Lump or mass in breast
Disease: Prolapse of vaginal walls
Disease: Sebaceous cyst
Disease: Arthropathy NOS
Disease: Osteoporosis NOS
Disease: Pain in joint
Age (years)
incidence count

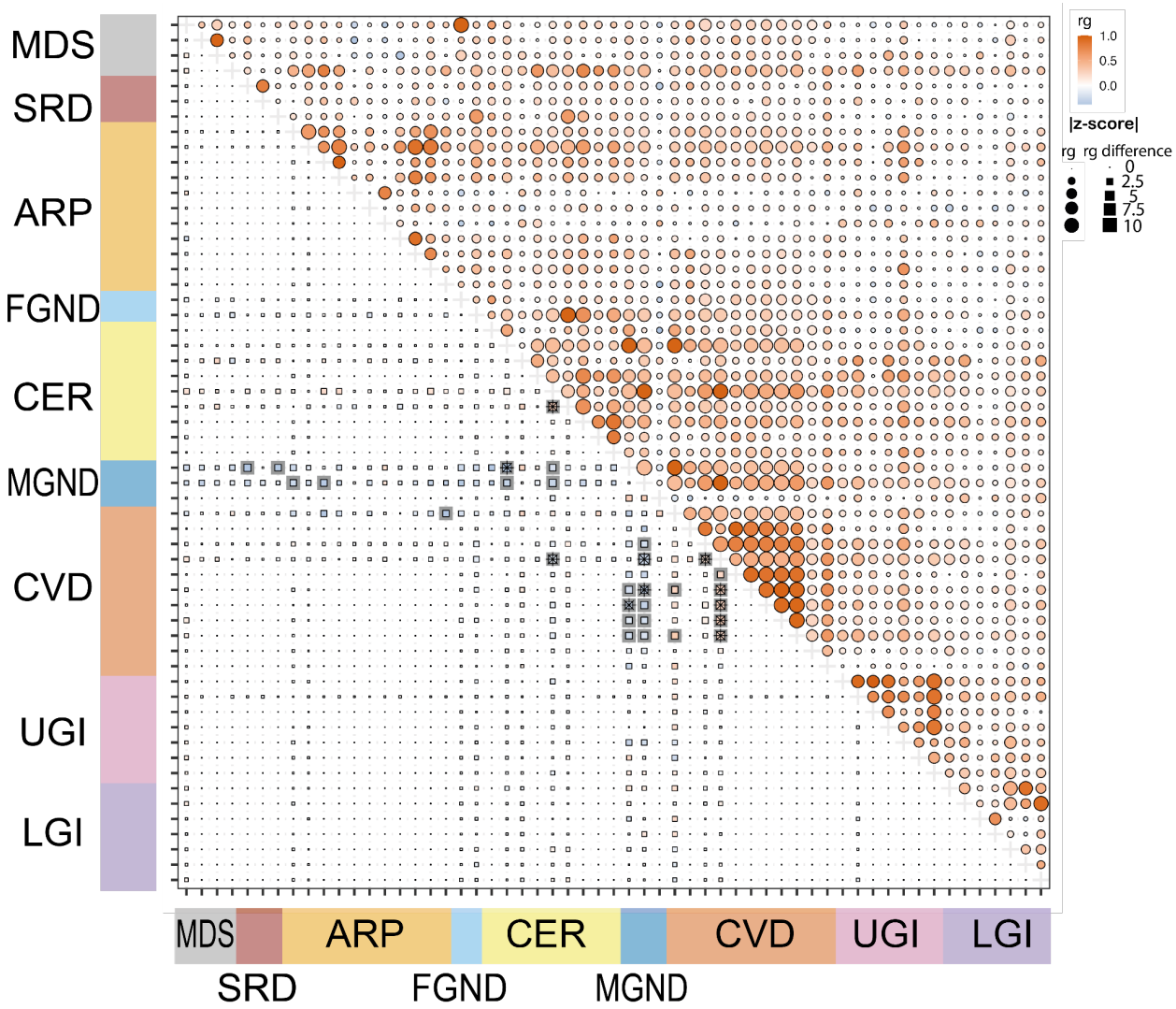
Supplementary Figure 15. Heterogeneity in topic distribution over age for disease subtypes (Part one). Diseases shown are from eight phecode systems: infectious diseases, neoplasms, endocrine/metabolic, hematopoietic, mental disorders, neurological, circulatory system, respiratory, digestive, neoplasms, genitourinary, dermatologic, and musculoskeletal.



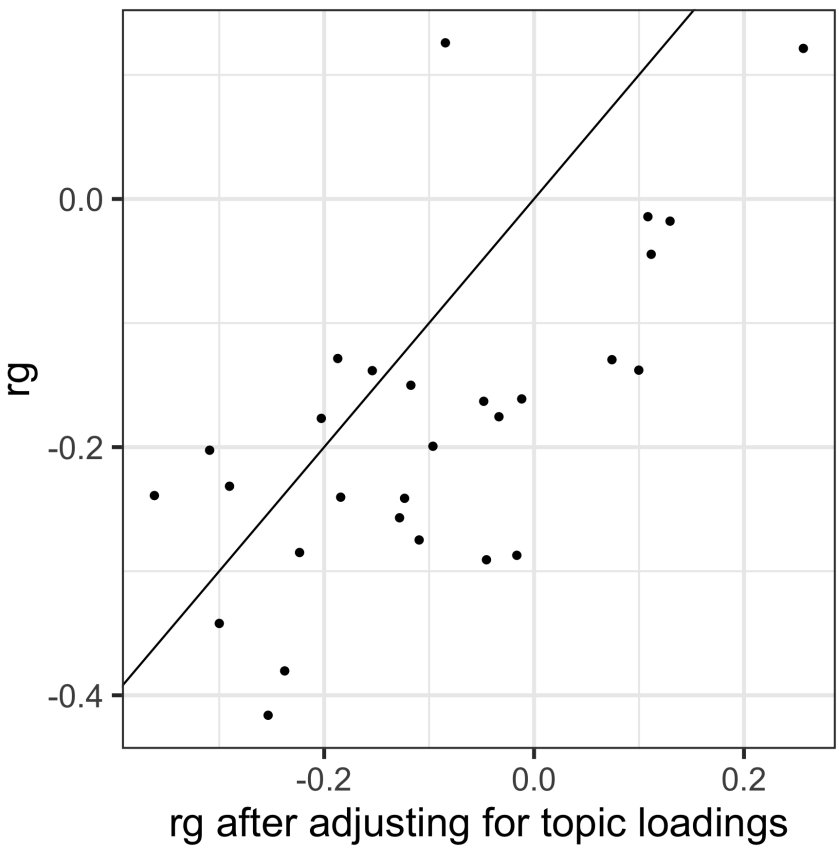
Supplementary Figure 16. PRS enrichment analysis for all topics across 10 diseases (selected by heritability z-score). The bar plot shows the estimated changes in s.d. of PRS per s.d. changes in the patient topic weight in all cases. The PRS is estimated using all the cases in British Isle Ancestry. The stars show disease-topic pairs that are significant at FDR = 0.1. The color of the bars encode the same topic as Figure 5.



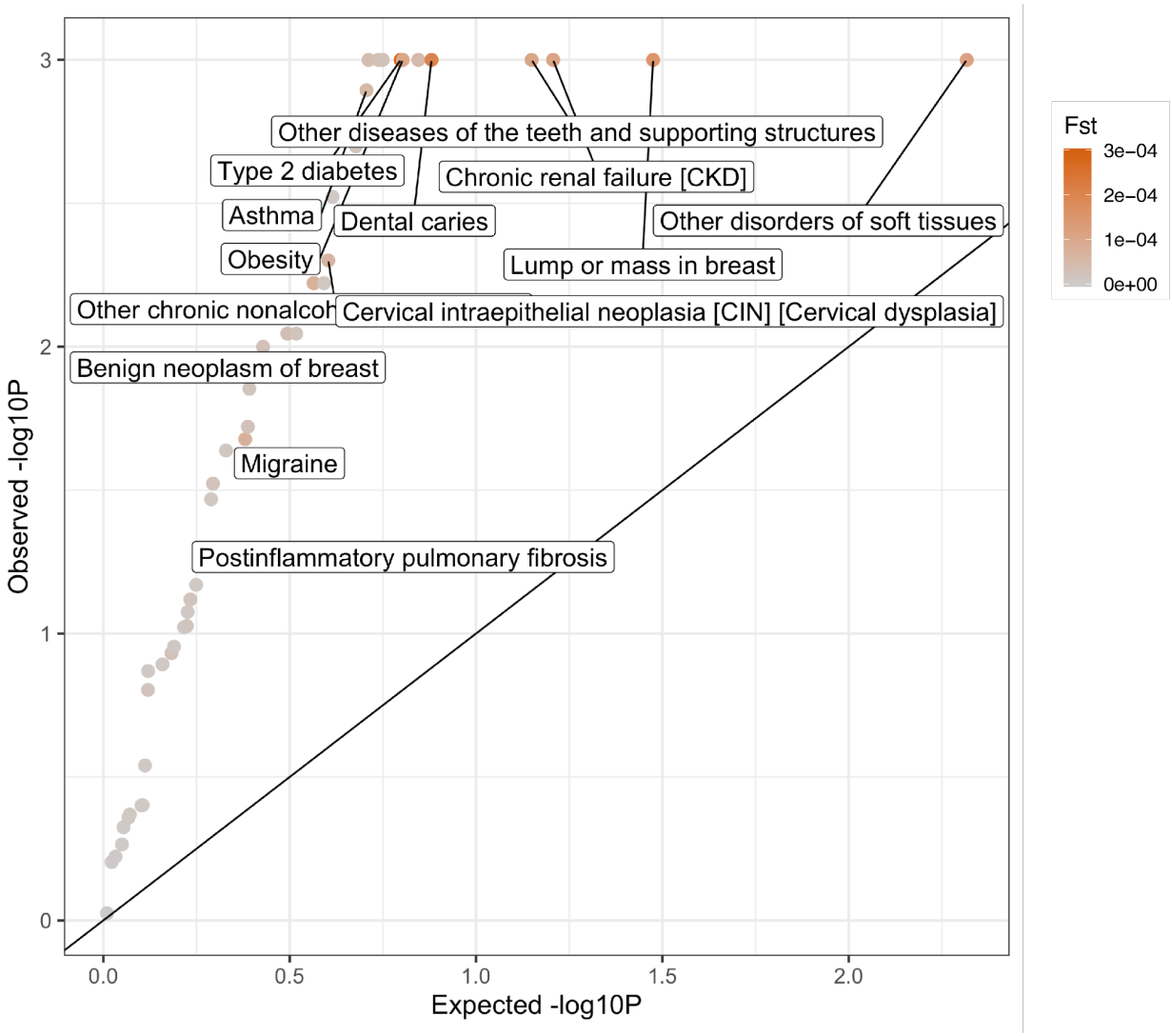
Supplementary Figure 17. Heritability and PRS are not associated with age and subtype size. (a) We plot heritability deviation from all cases versus age deviation from all cases of 41 subtypes (from 14 diseases that have heritability z-score above 5). The heritability is estimated by first performing 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) PRS enrichment from Figure 5B 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.



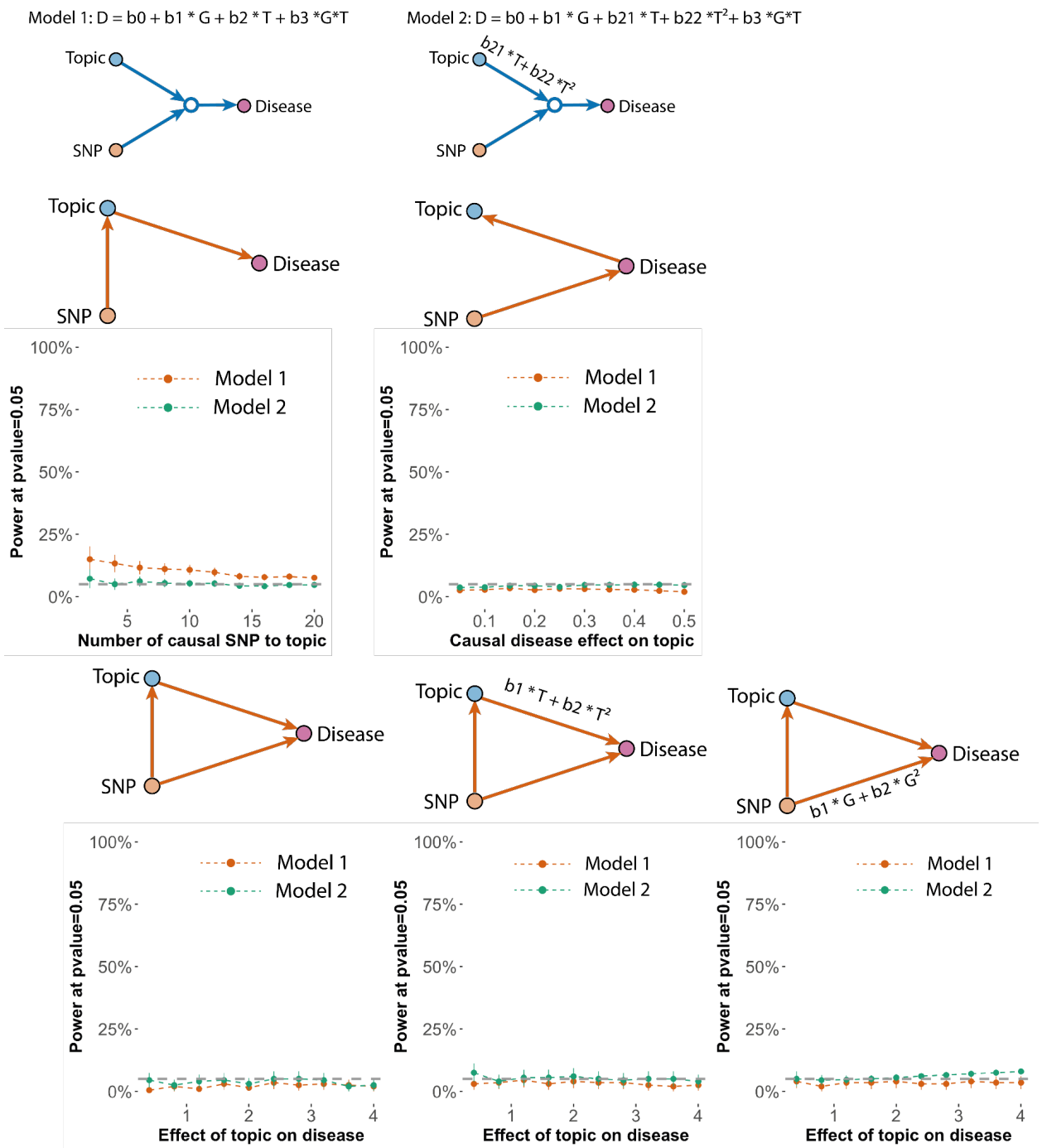
Supplementary Figure 18. Lower left panel shows the r_g of disease-subtype or subtype-subtype pairs subtracted by the r_g of corresponding disease-disease pairs. Each row or column represents a disease or subtype and the r_g difference is not defined for disease-disease pairs in the lower left panel. The upper right panel shows the corresponding r_g of corresponding disease-disease pairs, where values could be duplicated as the same disease could have multiple subtypes. The 57 diseases or subtypes are chosen here by heritability z-score > 4 and r_g z-score > 4 with at least one other disease or subtypes.



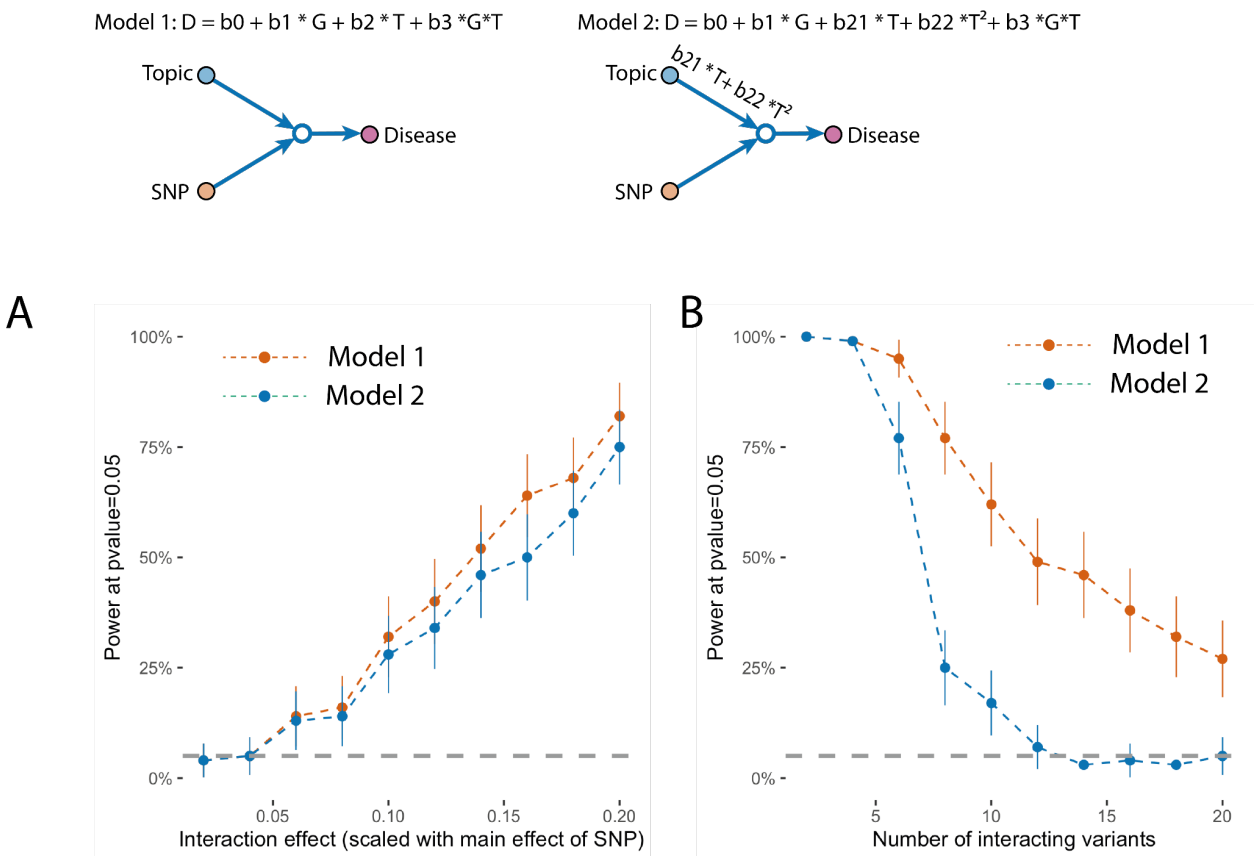
Supplementary Figure 19. Comparison of genetic correlation (ρ) between subtypes using summary statistics from case-control matched by topic weights (x-axis) and not matched by topic weights. The analysis is performed on subtypes of four diseases: type 2 diabetes, hypercholesterolemia, hypertension, and asthma. The ρ difference (shown in panel (a) of Figure 6) of 28 subtype pairs are compared and the effect lies along the diagonal line. The genetic correlation difference attenuated slightly when using topic weights are matched, while it can not explain all the ρ differences.



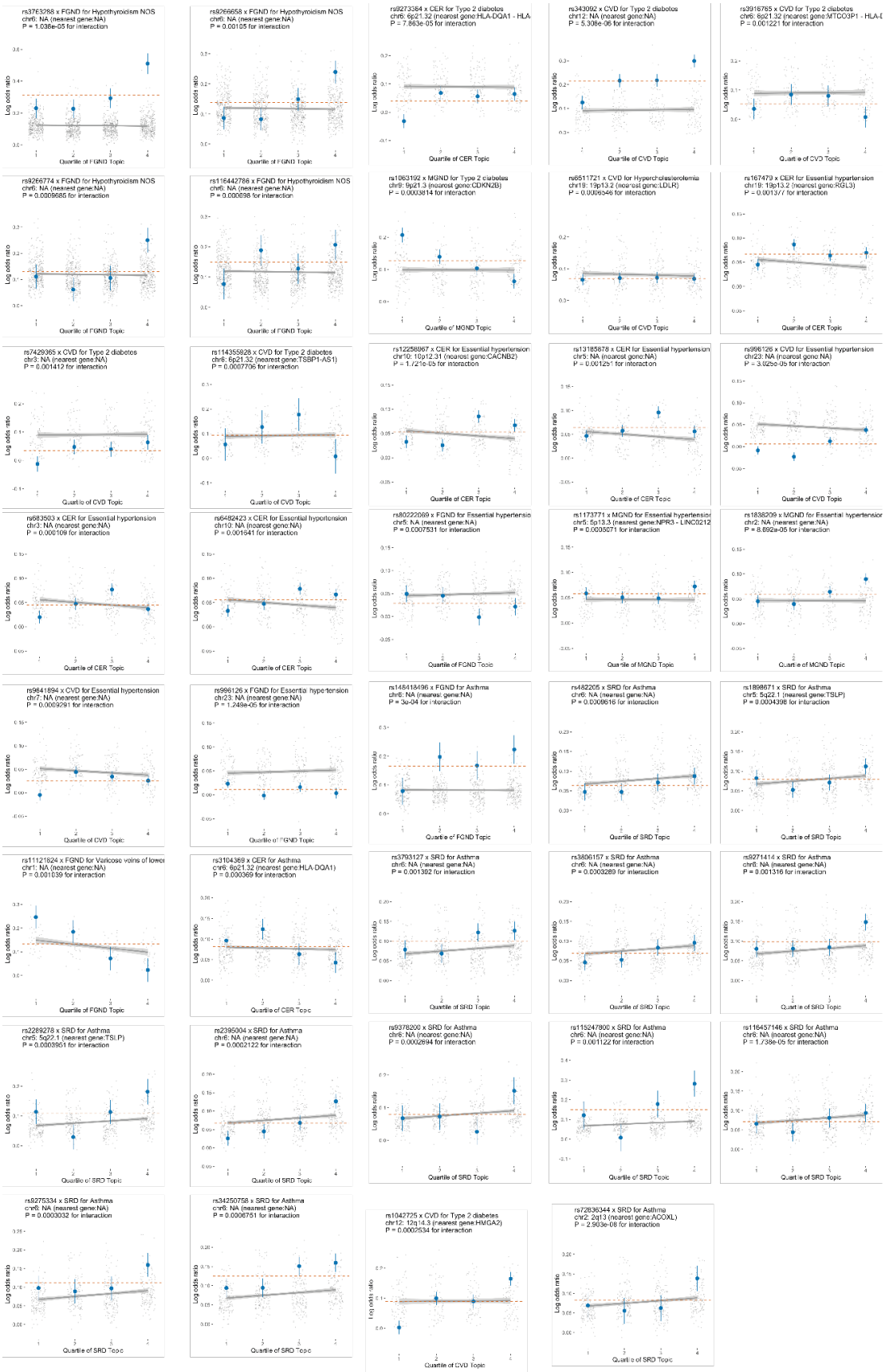
Supplementary Figure 20. F_{ST} analysis of diseases stratified by topic liabilities. P-values are for testing case-Fst significantly different from controls of similar topic weight distribution. The permutation controls are sampled for 1,000 times with the same size to each case-subtype. We focus on 49 of the 52 diseases which have a subgroup of at least 500 cases. We define the subgroups based on the max value of the disease posterior assignment to topics. Three diseases (“hypertension”, “hypercholesterolemia”, and “arthropathy”) are excluded as there are not enough controls that match the topic loadings of the cases. The color shows the value of F_{ST} across subtypes.



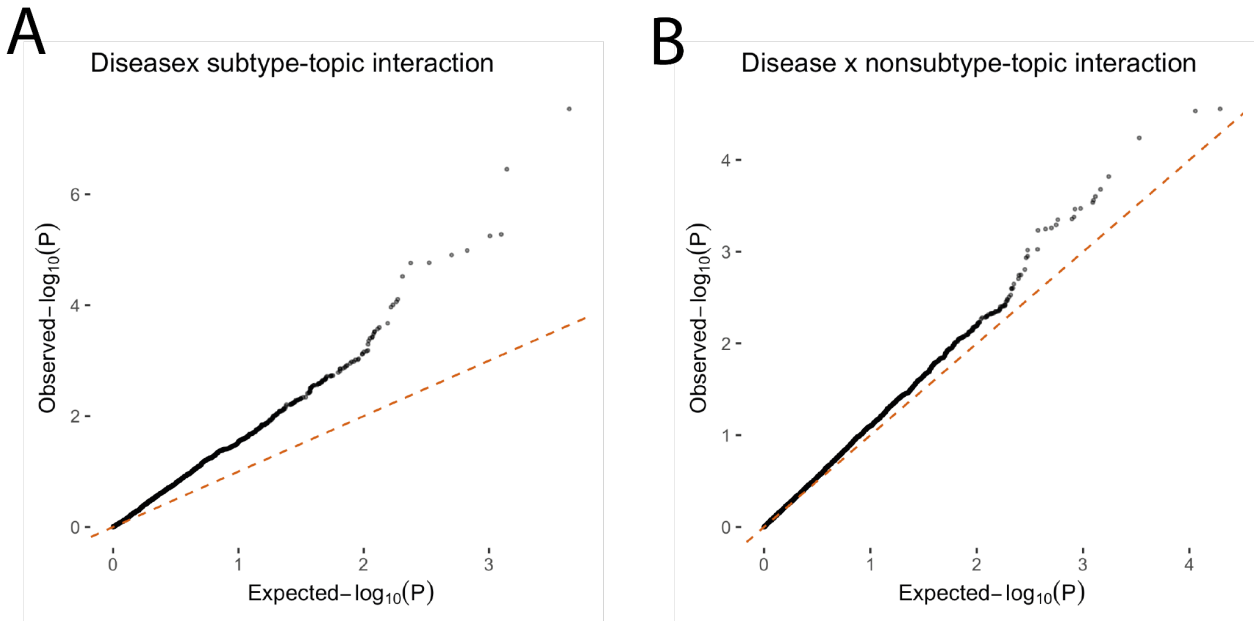
Supplementary Figure 21. Calibration of interaction tests when no actual interaction exists. We show all the structures we simulated that do not have actual interaction (Methods). The power to detect interaction effects using model 1 (red) and model 2 (green) computed under P-value=0.05 is shown. The five structures evaluated are (1) SNP causal to topic and topic causal to disease; (2) SNP causal to disease and disease causal to topic; (3) SNP is causal to both topic and disease; (4) and (5) SNP is causal to both topic and disease with nonlinear effects.



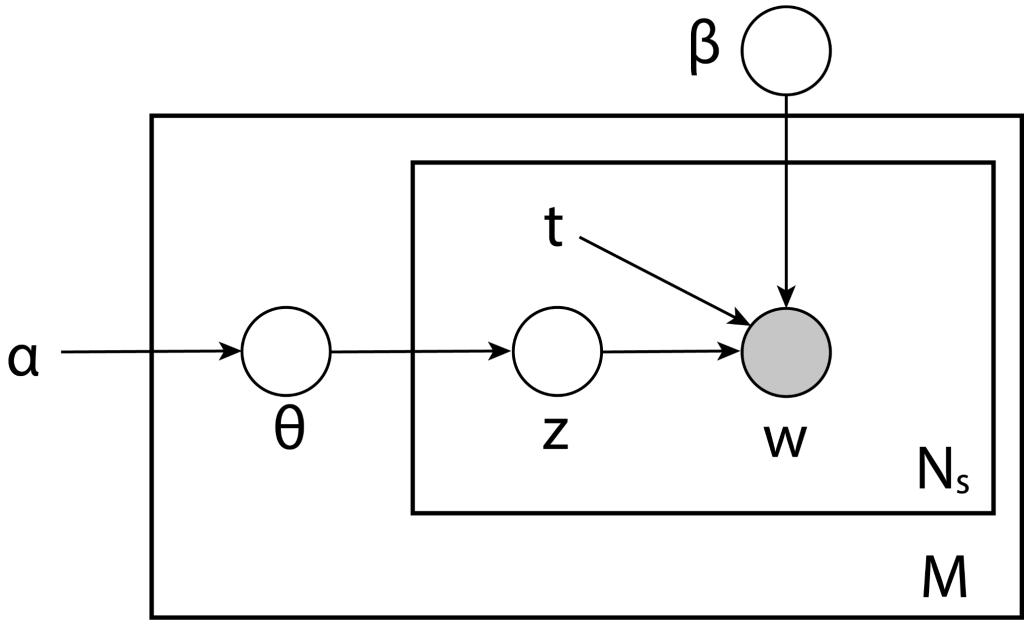
Supplementary Figure 22: Power to detect true interaction effect changes with age information. (a) We simulated data with 10,000 individuals, topic-to-disease main effect size equal to 2, interaction effect from 0.04 to 0.4, and SNP-to-disease main effect size proportional to the interaction effect (0.02 to 0.2); disease diagnoses are generated using gaussian liability with top 20 percentile as cases. We tested the interaction term using model 1 and model 2 and computed the power of discovering the true interaction. (b) We simulated data with 10,000 individuals and an interaction effect equal to 0.4. Instead of simulating a single SNP effect, we simulated 2 to 20 variants that all interact with topic weight. We then test the interaction term in model 1 and model 2 with only one variant at a time, which is the same strategy as GWAS interaction test. We note the power of model 2 is lower than model 1, while we still choose model 2 as it is better calibrated (Supplementary Figure 21).


rs3763268 x FGND for Hypothyroidism NOS
chr6: NA (nearest gene:NA)
P = 1.038e-05 for interaction
rs9266658 x FGND for Hypothyroidism NOS
chr6: NA (nearest gene:NA)
P = 0.00105 for interaction
rs9273364 x CER for Type 2 diabetes
chr6: 6p21.32 (nearest gene:HLA-DQA1 - HLA
P = 7.863e-05 for interaction
rs343092 x CVD for Type 2 diabetes
chr12: NA (nearest gene:NA)
P = 5.308e-06 for interaction
rs3916765 x CVD for Type 2 diabetes
chr6: 6p21.32 (nearest gene:MTCO3P1 - HLA-E
P = 0.0001221 for interaction
rs9266674 x FGND for Hypothyroidism NOS
chr6: NA (nearest gene:NA)
P = 0.0009685 for interaction
rs116442786 x FGND for Hypothyroidism NOS
chr6: NA (nearest gene:NA)
P = 0.000898 for interaction
rs1063192 x MGND for Type 2 diabetes
chr9: 9p21.3 (nearest gene:CDKN2B)
P = 0.0003814 for interaction
rs6511721 x CVD for Hypercholesterolemia
chr19: 19p13.2 (nearest gene:LDLR)
P = 0.0006346 for interaction
rs167479 x CER for Essential hypertension
chr19: 19p13.2 (nearest gene:RGL3)
P = 0.001377 for interaction
rs7429365 x CVD for Type 2 diabetes
chr3: NA (nearest gene:NA)
P = 0.001412 for interaction
rs114355928 x CVD for Type 2 diabetes
chr6: 6p21.32 (nearest gene:TSBP1-AS1)
P = 0.0007706 for interaction
rs12258967 x CER for Essential hypertension
chr10: 10p12.31 (nearest gene:CACNB2)
P = 1.721e-05 for interaction
rs13185878 x CER for Essential hypertension
chr5: NA (nearest gene:NA)
P = 0.001251 for interaction
rs998126 x CVD for Essential hypertension
chr23: NA (nearest gene:NA)
P = 3.025e-05 for interaction
rs983593 x CER for Essential hypertension
chr3: NA (nearest gene:NA)
P = 0.000109 for interaction
rs6482423 x CER for Essential hypertension
chr10: NA (nearest gene:NA)
P = 0.001641 for interaction
rs80222069 x FGND for Essential hypertensio
chr5: NA (nearest gene:NA)
P = 0.0007531 for interaction
rs1173771 x MGND for Essential hypertensio
chr5: 5p13.3 (nearest gene:NPR3 - LINC0212
P = 0.0003071 for interaction
rs1838209 x MGND for Essential hypertensio
chr2: NA (nearest gene:NA)
P = 8.692e-05 for interaction
rs9641894 x CVD for Essential hypertension
chr7: NA (nearest gene:NA)
P = 0.0009291 for interaction
rs998126 x FGND for Essential hypertension
chr23: NA (nearest gene:NA)
P = 1.249e-05 for interaction
rs148418496 x FGND for Asthma
chr6: NA (nearest gene:NA)
P = 3e-04 for interaction
rs482205 x SRD for Asthma
chr6: NA (nearest gene:NA)
P = 0.0008616 for interaction
rs1898671 x SRD for Asthma
chr5: 5q22.1 (nearest gene:TSLP)
P = 0.0004398 for interaction
rs11121824 x FGND for Varicose veins of lower
chr1: NA (nearest gene:NA)
P = 0.001039 for interaction
rs3104369 x CER for Asthma
chr6: 6p21.32 (nearest gene:HLA-DQA1)
P = 0.000369 for interaction
rs3793127 x SRD for Asthma
chr6: NA (nearest gene:NA)
P = 0.001382 for interaction
rs3806157 x SRD for Asthma
chr6: NA (nearest gene:NA)
P = 0.0009289 for interaction
rs9271414 x SRD for Asthma
chr6: NA (nearest gene:NA)
P = 0.001316 for interaction
rs2289278 x SRD for Asthma
chr5: 5q22.1 (nearest gene:TSLP)
P = 0.0003951 for interaction
rs2395004 x SRD for Asthma
chr6: NA (nearest gene:NA)
P = 0.0002122 for interaction
rs9378200 x SRD for Asthma
chr6: NA (nearest gene:NA)
P = 0.0002694 for interaction
rs115247800 x SRD for Asthma
chr6: NA (nearest gene:NA)
P = 0.001122 for interaction
rs116457146 x SRD for Asthma
chr6: NA (nearest gene:NA)
P = 1.738e-05 for interaction
rs9275334 x SRD for Asthma
chr6: NA (nearest gene:NA)
P = 0.0003032 for interaction
rs34250758 x SRD for Asthma
chr6: NA (nearest gene:NA)
P = 0.0006751 for interaction
rs1042725 x CVD for Type 2 diabetes
chr12: 12q14.3 (nearest gene:HMGA2)
P = 0.0002534 for interaction
rs72836344 x SRD for Asthma
chr2: 2q13 (nearest gene:ACOXL)
P = 2.903e-08 for interaction
Log odds ratio
Quartile of FGND Topic
Quartile of CER Topic
Quartile of CVD Topic
Quartile of MGND Topic
Quartile of SRD Topic

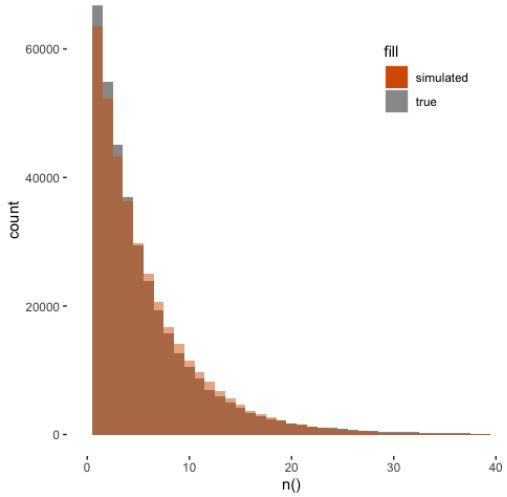
Supplementary Figure 23. Additional 39 SNPs (mapped genes in the parentheses) that have different effect sizes in different quantiles of topic weights. Blue dots are the effect sizes of the target SNP within each topic weights quartile; grey dots are background SNPs which are genome-wide significant for the traits but do not have evidence supporting interaction with topic weights (P>0.05). The grey line shows the regression line of the grey dots and its 95% confidence interval. P-values are for the effect size differences between top and bottom quartiles.



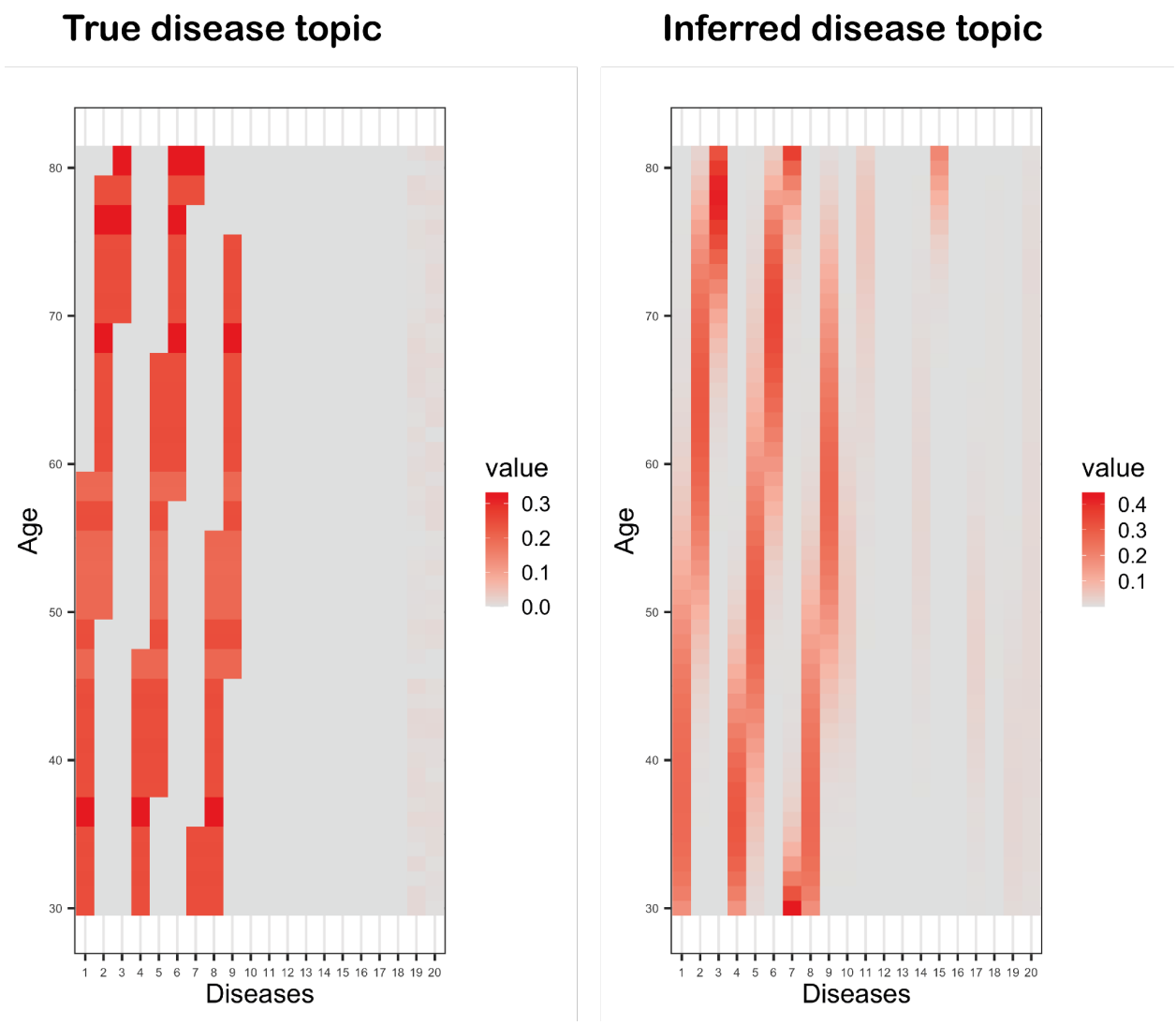
Supplementary Figure 24: QQ plot of GxTopic interaction for all GWAS SNPs (P < 5×10^{-8}). (a) We show the interaction between SNP-topic where the topics define disease subtypes. We focus on the subset of subtypes whose disease have h^2 z-score larger than 4 to ensure there is enough GWAS signal for testing. The P-values are testing the interaction effects with nonlinear topic-to-disease main effects (Model 2 in Supplementary Figure 19). (b) As a control to show the collaboration of the tests, we plot the same tests over the same set of GWAS SNPs, but over the topic that are not subtypes of the disease.



Supplementary Figure 25. Plate presentation of 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. The generative process is described in the text.



Supplementary Figure 26. Histogram of number of distinct diseases per patient from the UK Biobank HES dataset and from the simulated exponential distribution with mean = 6.1.



Supplementary Figure 27. Examples of simulated disease topic loading (left) and corresponding inferred topic loading (right).