Supplementary Tables (see Excel file)

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

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

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

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

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

Supplementary Table 6. Number of diagnoses assigned to each subtypes for 52 diseases. We listed the number of diagnoses assigned to each disease subtypes by the diagnosis-specific topic probability, for the 52 diseases that have at least two subtypes with >500 diagnosis.

Supplementary Table 7. Average age at diagnosis for each subtypes of the 52 diseases. We listed the age at diagnosis across all diagnoses within each disease subtype, for the 52 diseases that have at least two subtypes with >500 diagnosis.

Supplementary Table 8. Excess PRS in cases for all topics across 10 diseases (selected by heritability z-score). We report the estimated changes in s.d. of PRS per unit changes in the patient topic weight, which is estimated through regression across disease diagnoses. The PRS

was estimated using BOLT-LMM and all the cases of British Isle Ancestry. Numbers correspond to Figure 5B and Supplementary Figure 17.

Supplementary Table 9. Excess genetic correlations. Columns are: Phecode of the first disease (trait1), Phecode of the second trait (trait2), subtype of the first trait (topic1), subtype of the second trait (topic2), the z-score of genetic correlation between two disease subtypes (subtype.rho.zscore), estimate of genetic correlation between two disease subtypes (subtype.rho.est), standard error of genetic correlation between two disease subtypes (subtyp.rho.err), the z-score of genetic correlation between two diseases (all.rho.zscore), estimate of genetic correlation between two diseases (all.rho.est), standard error of genetic correlation between two diseases (all.rho.err), z-score of excess genetic correlation (diff.zscore), estimate of excess genetic correlation (diff.rg), absolute value of excess genetic correlation (diff.rg.zscore.abs), p-values for excess genetic correlation (P), FDR for excess genetic correlation (FDR), name of the first disease (phenotype.1), and name of the second disease (phenotype.2). We only reported p-values of excess genetic correlation when both genetic correlation estimation has standard error <0.1 and at least one of the genetic correlation has |z-score|>4. Numbers correspond to Figure 6 and Supplementary Figure 19.

Supplementary Table 10. Heritability estimation for disease subtypes. For each disease we list heritability estimates for two subtypes using LDSC. Topic.x refer to the subtype with the highest heritability, topic.y refer to subtype with the lowest heritability. We report the point estimate, standard error, z-scores of both disease subtypes. The z-score of heritability differences between the two subtypes are also reported. Note we used a different sample threshold 1000 (due to the power of LDSC), which includes 26 of the 52 diseases that have subtypes.

Supplementary Table 11. Excess F_{ST} estimation across disease subtypes. We report the estimate of excess F_{ST} (computed as the F_{ST} across subtypes subtracted by the F_{ST} from controls with matched topic weights). The p-values are for excess $F_{ST}>0$, which is computed from 1000 randomly sampled control sets.

Supplementary Table 12. GxTopic interaction tests across independent GWAS SNPs. Each row represents one SNP x topic weight paris (disease subtype). OR, SE, STAT, and P represent the odds ratio, standard error, test statistics, and p-value of the main effects. SNPxTopic OR, SNPxTopic STAT, SNPxTopic P, SNPxTopic FDR represent the odds ratio, test statistics, p-value, and genome-wide FDR of testing the interaction effect in model 2 of Supplementary Figure 22. We use Studywise FDR which adjusts for multiple testing across GWAS SNPs of all disease subtypes.

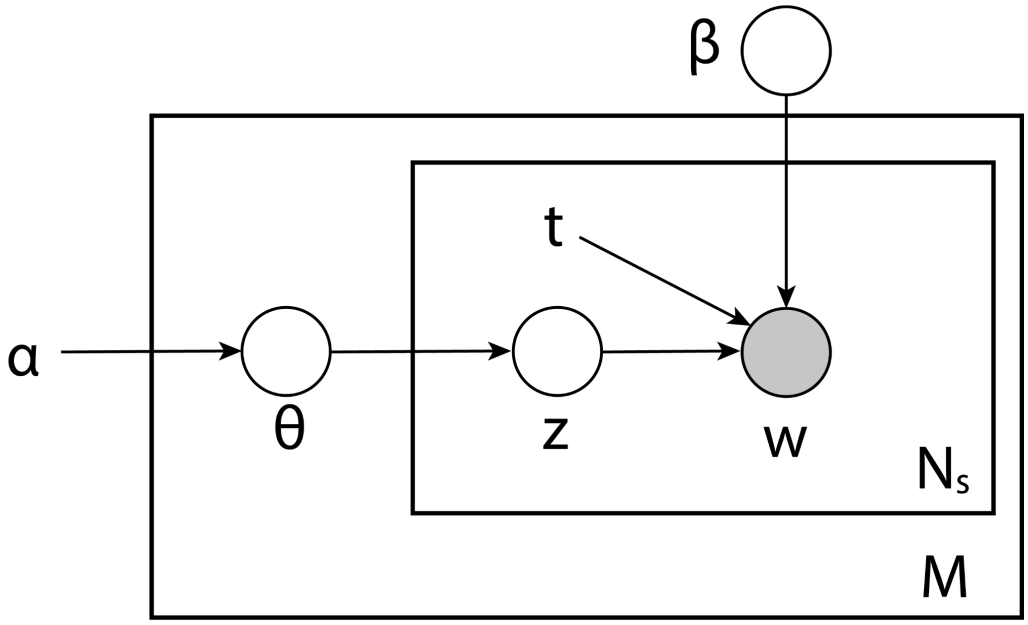
Supplementary Table 13. Significant SNP x topic interactions. Same table as supplementary Table 12, but filtered to SNP-topic pairs with interaction effect passing FDR<0.1. Reported SNP-topic pairs were selected for topic weights specific effect estimation in Figure 7 and Supplementary Figure 24.

Supplementary Table 14. Effect size estimation across topic weight quartiles for significant SNP x topic interactions. Quartile, mean_effect, se_effect, refer to the quartile of topic weight, estimate of effect sizes of the SNP using case-controls from this quartile, and standard error of the effect size estimation. We also reported the nearest genes reported by GWAS Catalog. The last column reports the P-value of effect size being different between the top and bottom quartiles.

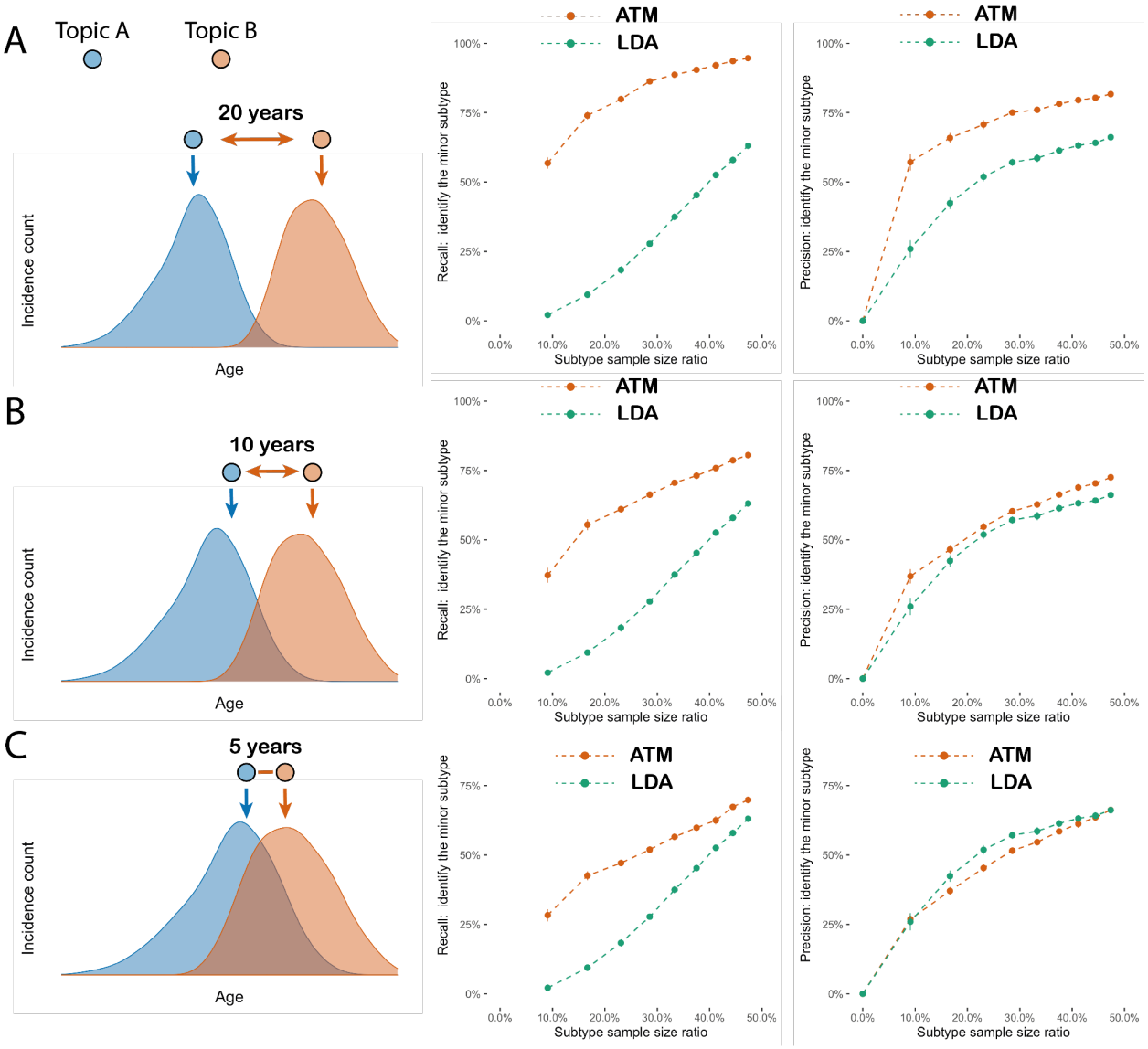
Supplementary Table 15. Literature search of disease subtypes identified by ATM. We searched on Pubmed using the description (ignoring conjunctions) AND "subtype" in title/abstract and manually screened the top 10 relevant results between 2012 and 2022. The studies that mentioned subtypes of the searched diseases are included in the "Published references" column. If there is no reference of target disease subtypes among the top 10 search results, we use "NA". We note our search is not exhaustive but nevertheless provides information on whether subtypes of the target disease are described in studies involving target disease and subtypes.

Supplementary Table 16. ATM running time. We tested the running time on the UK Biobank data using ATM of varying topic number and parametric form of topic loadings. Degrees of freedom from 2 to 7 represent linear, quadratic polynomial, cubic polynomial, spline with one knot, spline with two knots, and spline with three knots. Note a few models with 50 topics did not converge.

Supplementary Figures

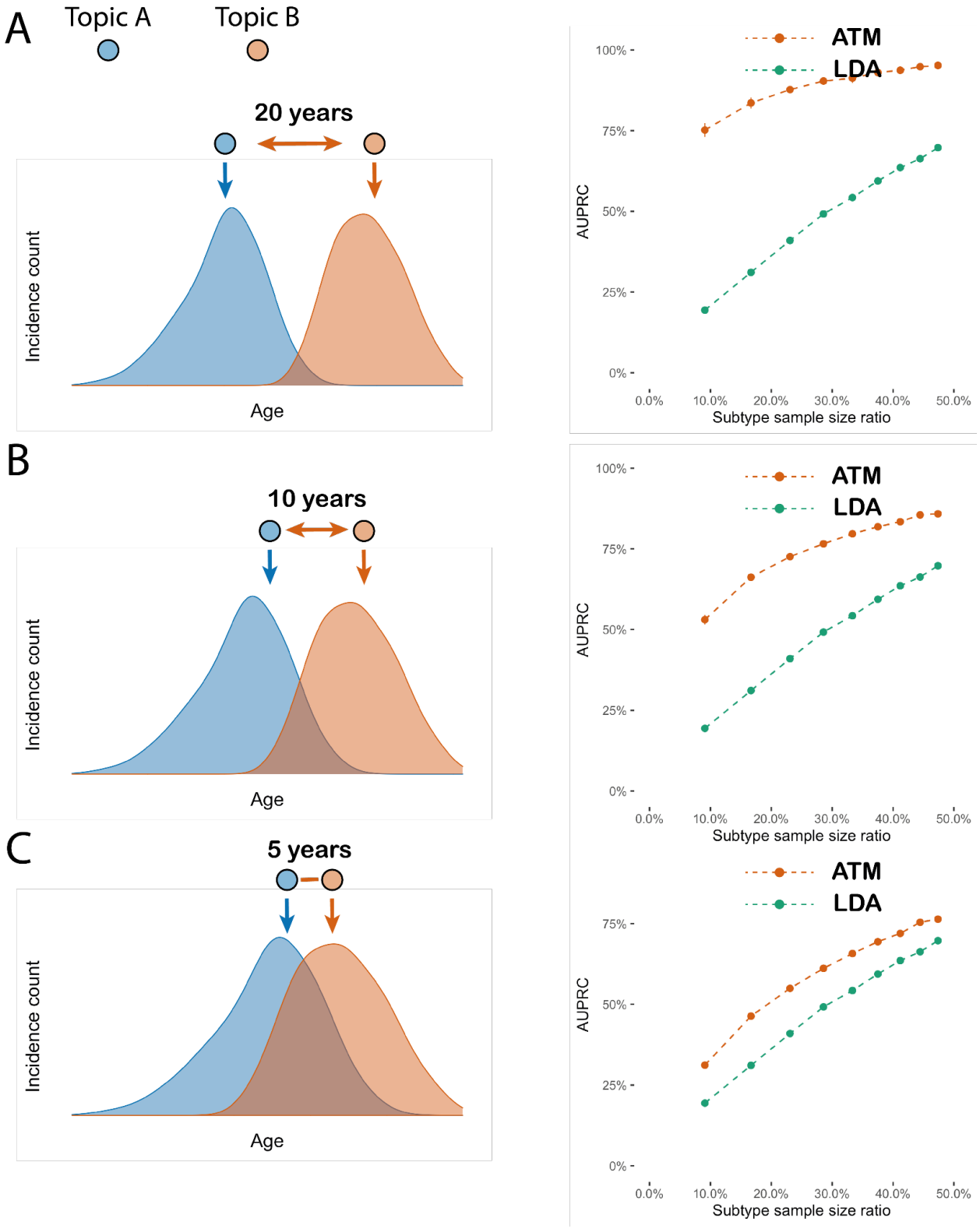


Supplementary Figure 1. Plate notation of ATM generative model. M is the number of subjects; N_s is the number of records within s^{th} subject. All plates (circles) are variables in the generative process, where the plates with shade w is the observed variable and plates without shade are unobserved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



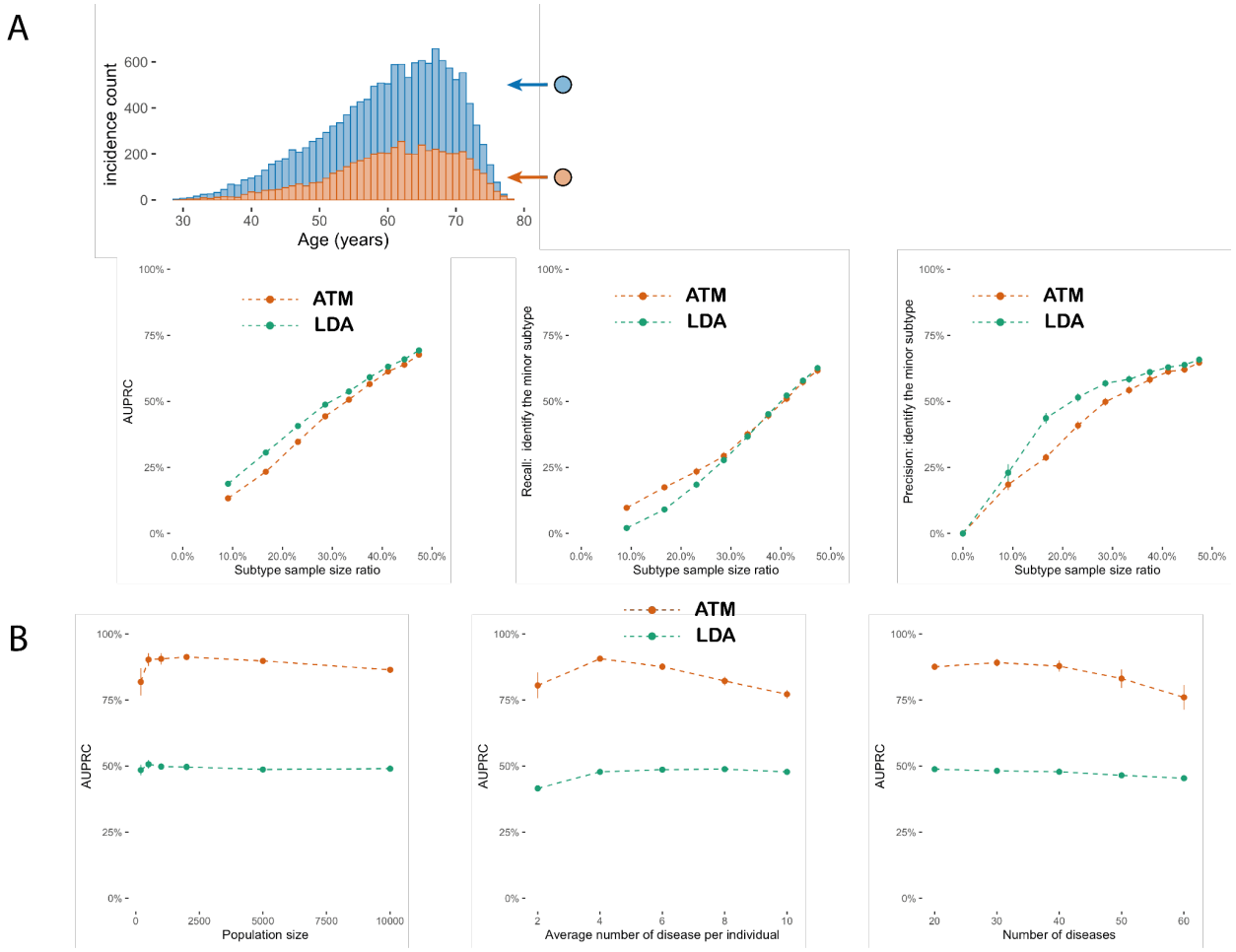
Supplementary Figure 2. Additional simulation studies established the power of the method to identify comorbidity. The precision and recall rate to correctly assign incident disease to correct comorbidity profiles using Latent Dirichlet Allocation (LDA) and our method (ATM). X-axis refers to the proportion of cases that belong to the small 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.(a) Scenario where two subtypes are simulated with 20 years of difference in age at diagnosis. (b) Scenario where two subtypes are simulated with 10 years of difference in age at diagnosis. (c) Scenario where two subtypes are simulated with 5

years of age difference.



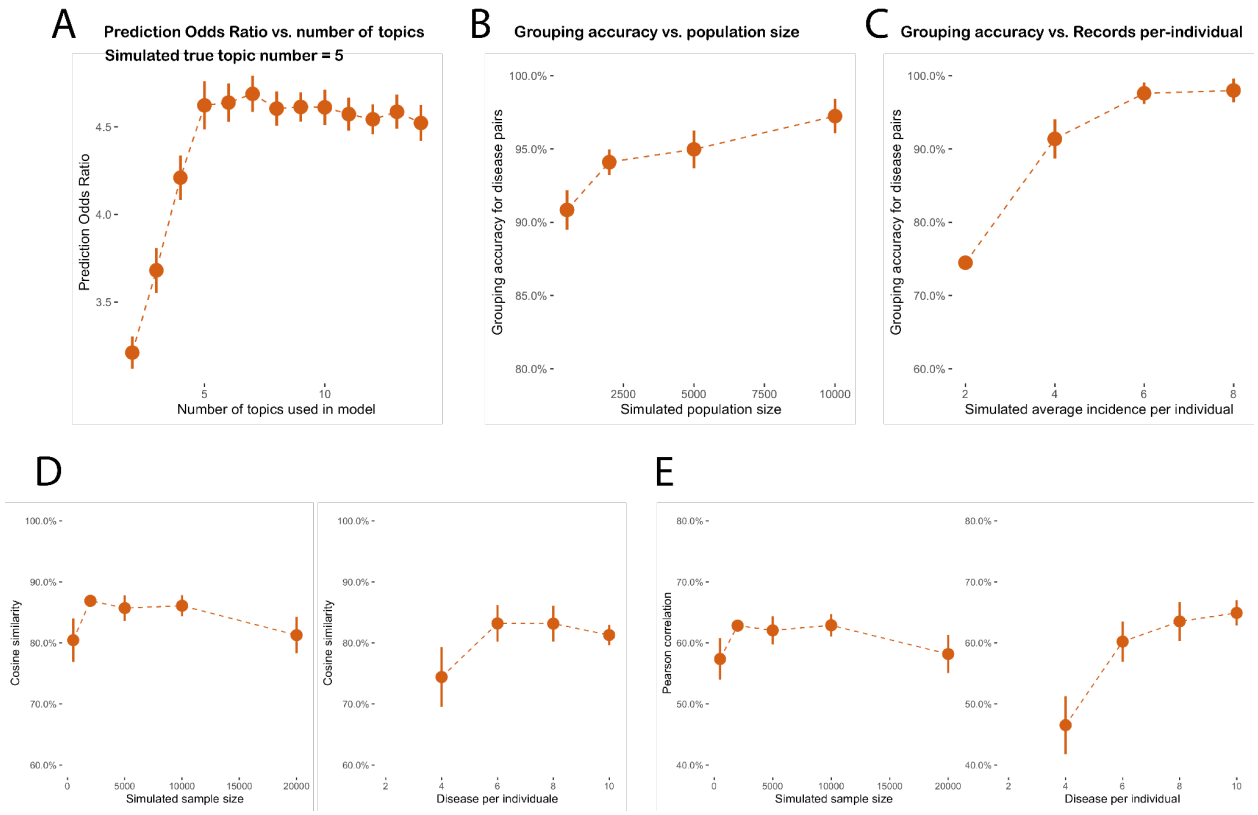
Supplementary Figure 3. Same analysis as in Figure 2 but simulating the smaller subtype to have older age at diagnosis. The area under precision and recall curve (AUPRC) to correctly

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



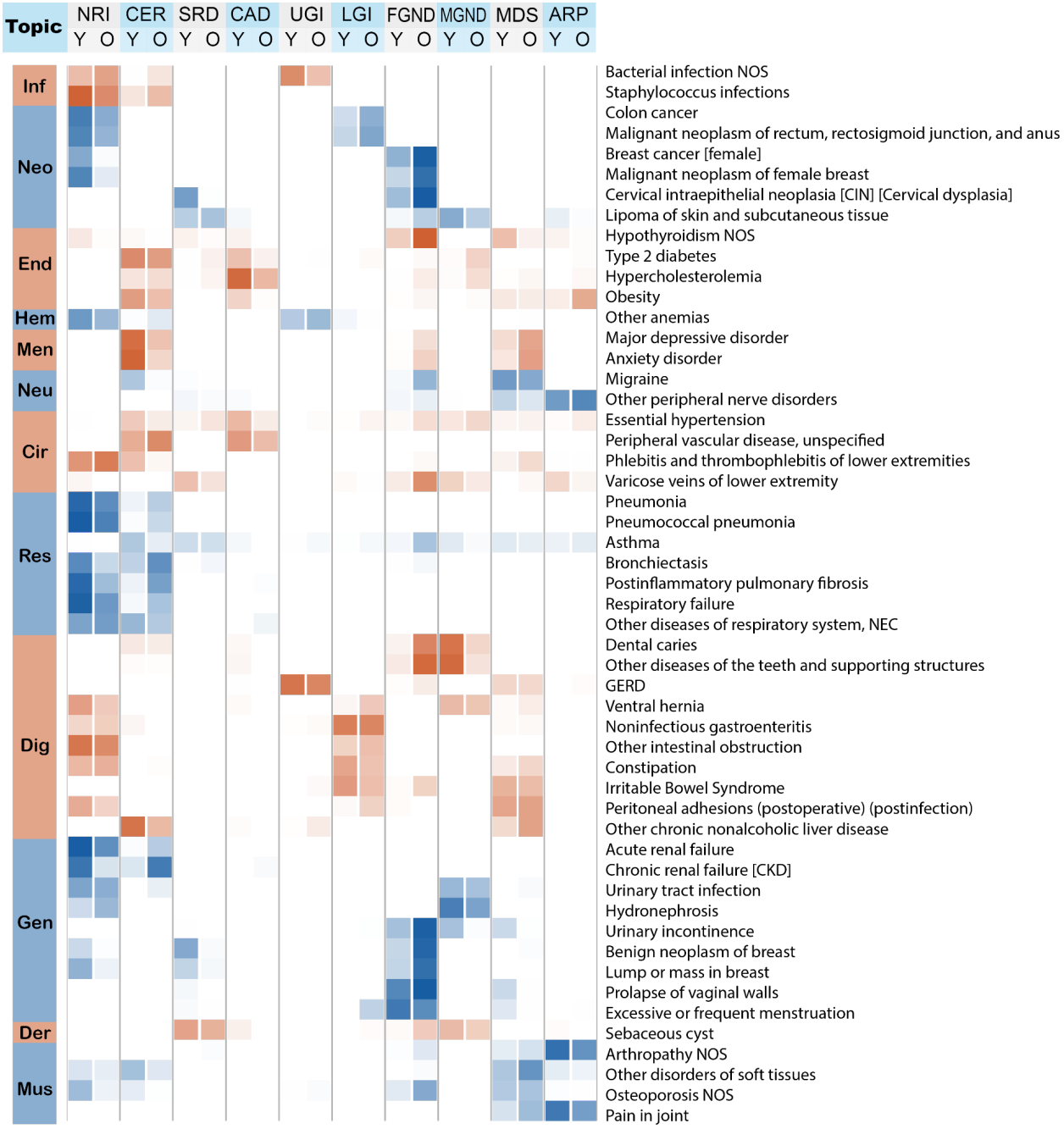
Supplementary Figure 4. Additional simulation studies established the power of the method to identify comorbidity. (a) Same analysis as Figure 2 but simulated subtypes with same age at diagnosis distribution. LDA outperforms ATM slightly as we have additional regularisation when modelling topic loading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

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

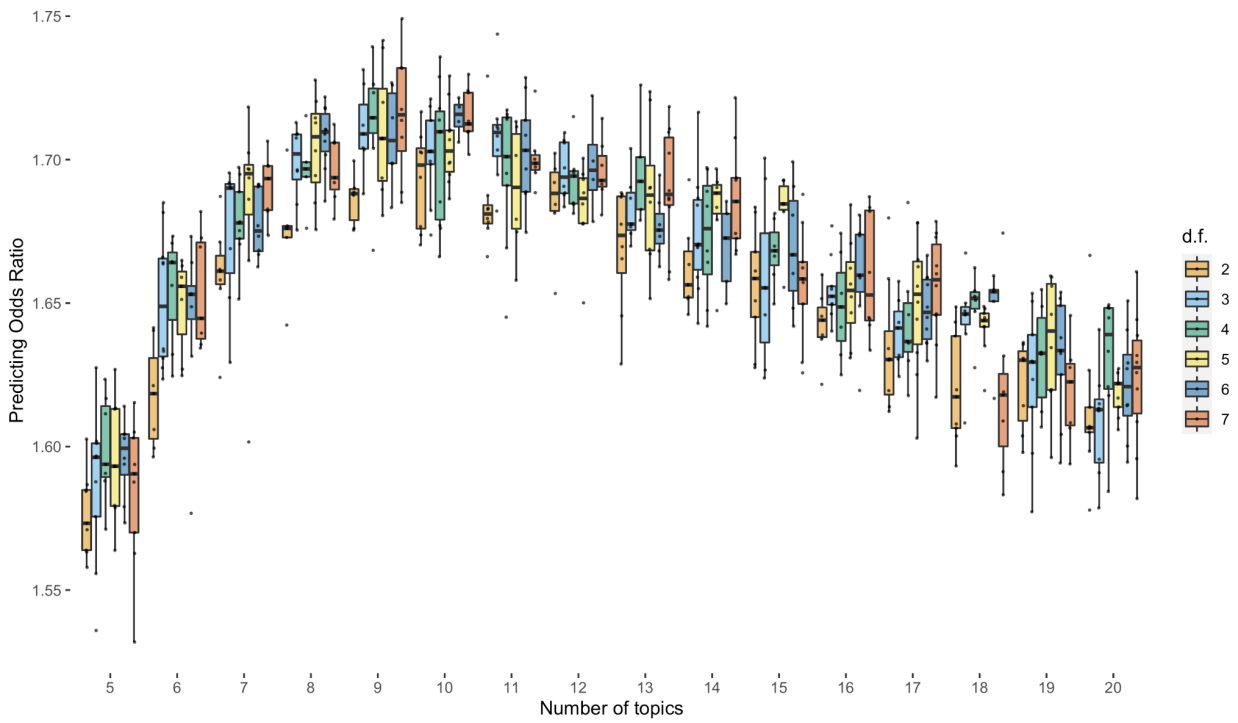


Supplementary Figure 5. Simulations confirming that ATM could accurately recover topic loadings and topic weights. (a) We simulated data using 5 topics while fitting models of varying topic numbers. To compute prediction odds ratios (see Methods), we used 80% of data as training data to fit ATM and computed 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

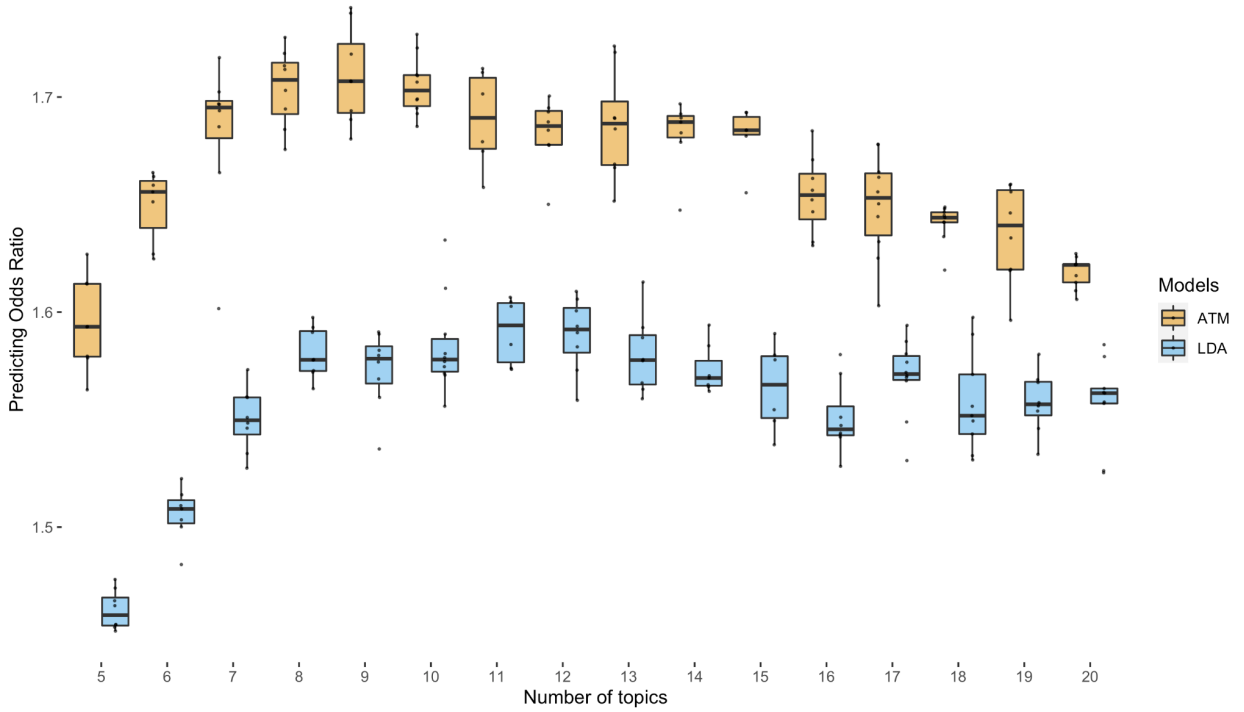
topic weight inference by computing the correlation of inferred topic weights and with the underlying truth. The ordering of topics uses the same strategy as in panel d.



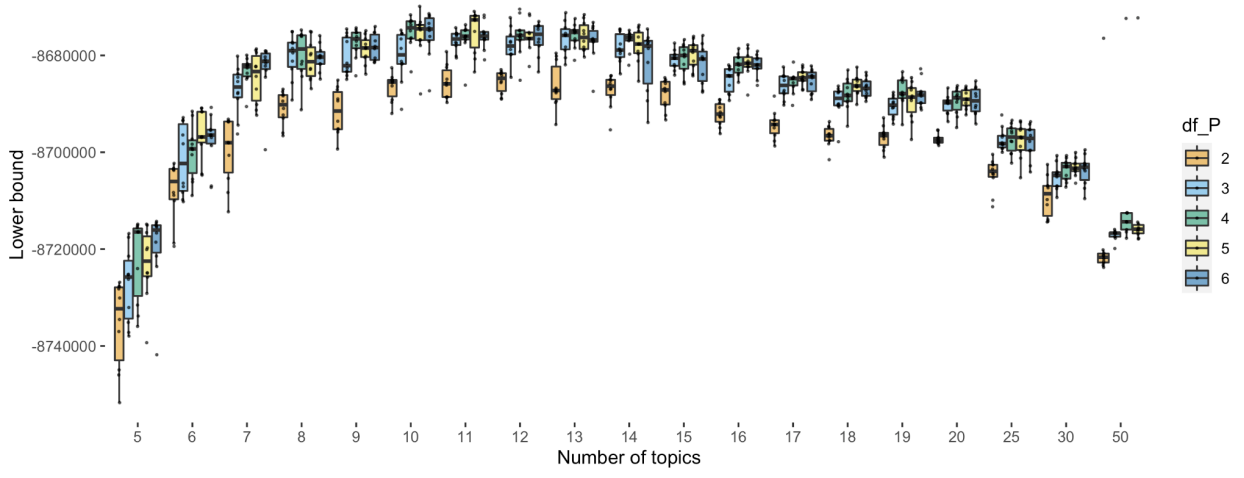
Supplementary Figure 6. Posterior topic distributions are different between age groups for diseases that have subtypes. The figure has the same legends as Figure 3A but focusing on 52 diseases that have a subtype with at least 500 incidences.



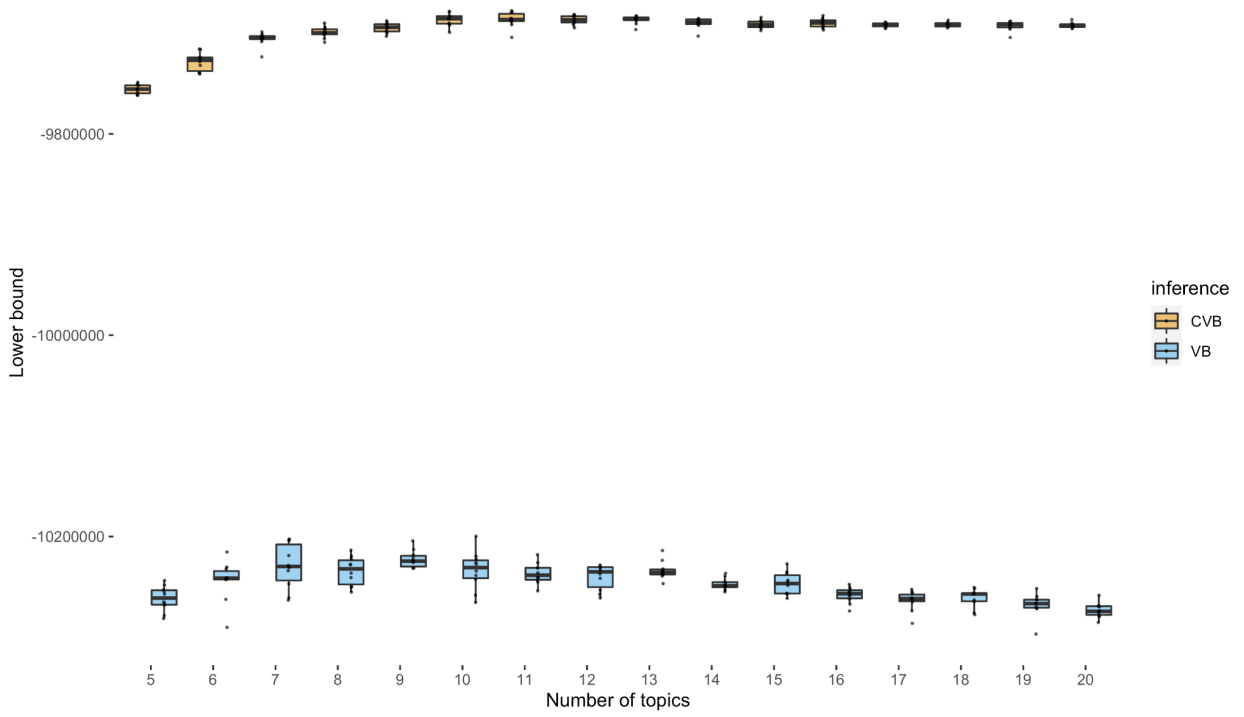
Supplementary Figure 7. Prediction odds ratio across different model configurations. Each dot represents one inference on a random training and testing split of the UK Biobank individuals. The models are run with different topic numbers and parametric configurations of topic loadings. Degrees of freedom (d.f.) from 2 to 7 represent linear, quadratic polynomial, cubic polynomial, spline with one knot, spline with two knots, and spline with three knots.The prediction odds ratios are computed on the testing data using topic loadings inferred from the training data and topic weights inferred using previous diseases of testing individuals. The odds ratios are between the odds that target diseases are within model-predicted top percentile disease set versus the odds that target diseases are within the prevalence-ordered top percentile disease set.



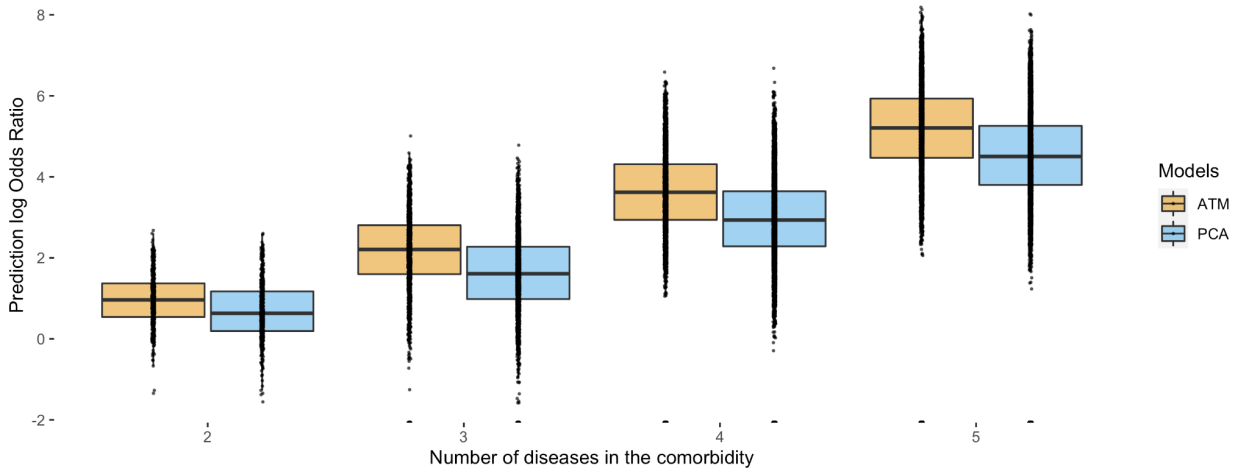
Supplementary Figure 8. Comparison of prediction odds ratio between LDA and ATM. Each dot represents results from running either ATM or LDA on the same random training and testing split. The models were run with different topic numbers and we chose a cubic spline with one knot for configuring ATM topic loadings. The prediction odds ratios are computed on the testing data using topic loadings inferred from the training data and topic weights inferred using previous diseases of testing individuals. The odds ratios are between the odds that target diseases are within model-predicted top percentile disease set versus the odds that target diseases are within the prevalence-ordered top percentile disease set.



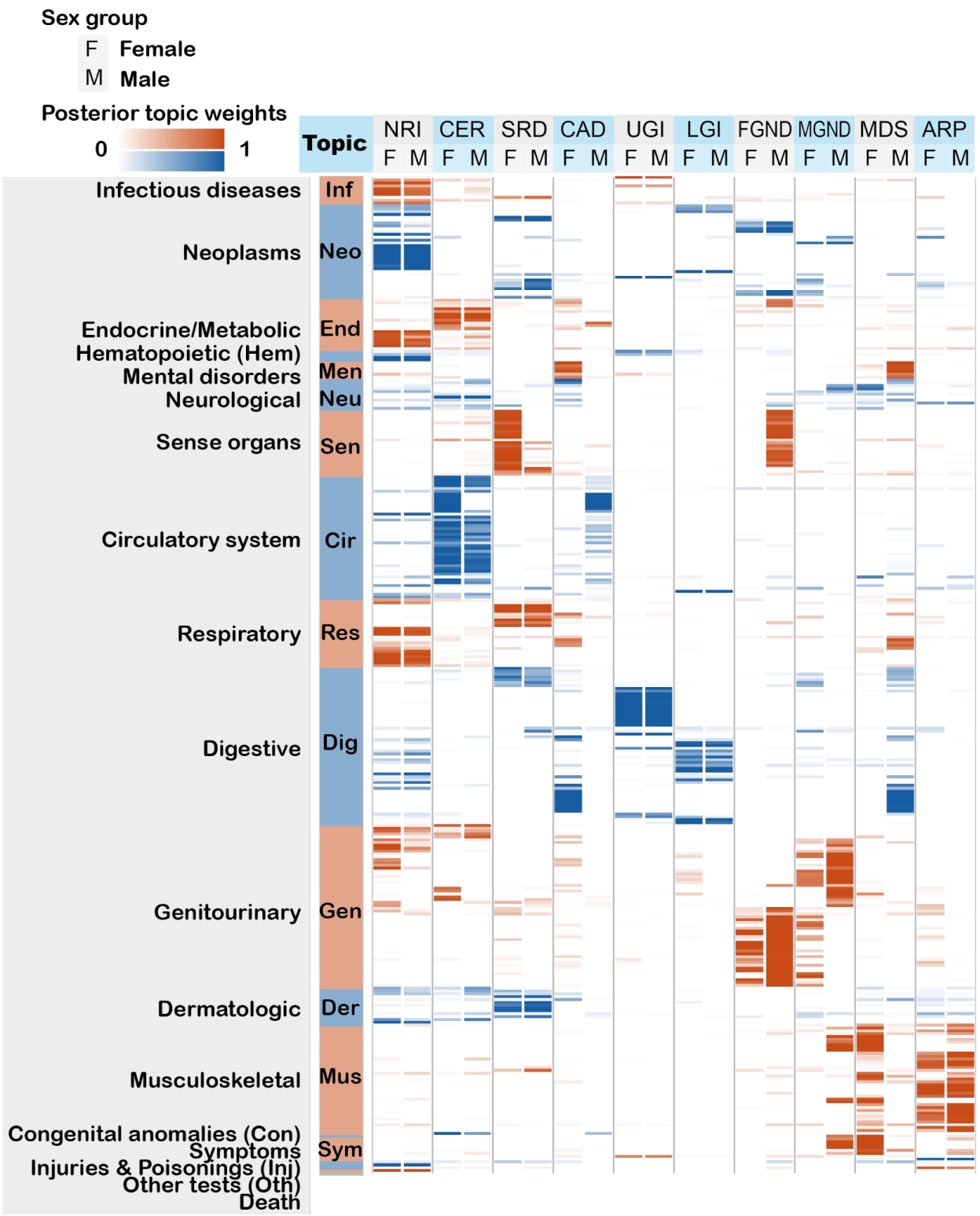
Supplementary Figure 9. Evidence lower bound (ELBO) of different model configurations on the entire dataset. Each dot represents one inference on a random training and testing split of the UK Biobank individuals. The models are run with different topic numbers and parametric configurations of topic loadings.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



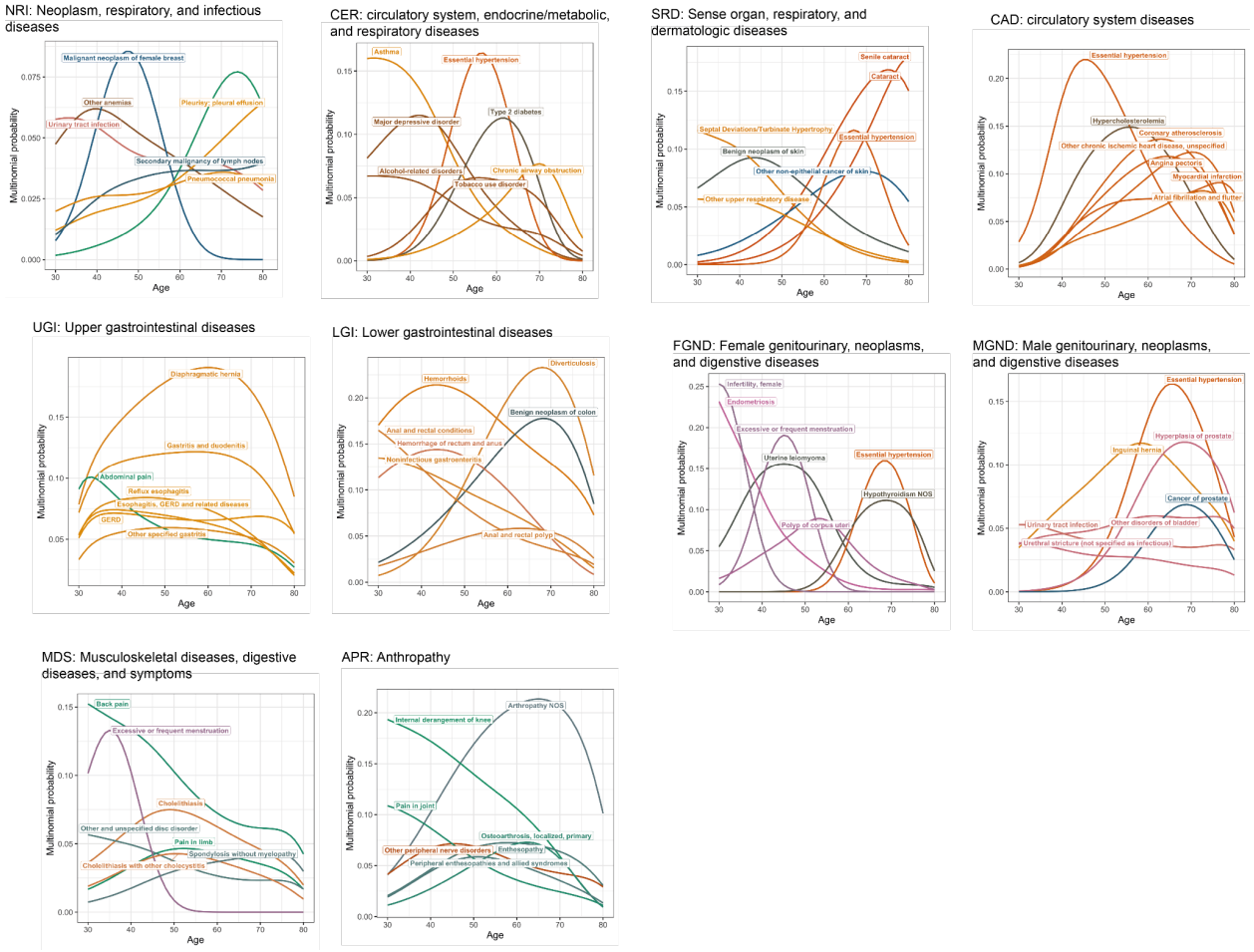
Supplementary Figure 10. Comparison of ELBO for collapsed variational inference and mean-field variational inference. ELBO is computed by fitting the ATM using two inference methods on the entire UK Biobank dataset, where the topic loadings are configured as cubic polynomials. Models of different numbers of topics are fitted with 10 random initialisations for both CVB and the VB (mean-field variational inference, which is a more commonly used inference method for Bayesian models). The ELBO of an inference methods is a lower bound that approximates the evidence function, which depends on the number of topics and parametric form of topic loading, but not the inference methods; higher ELBO means better inference accuracy.



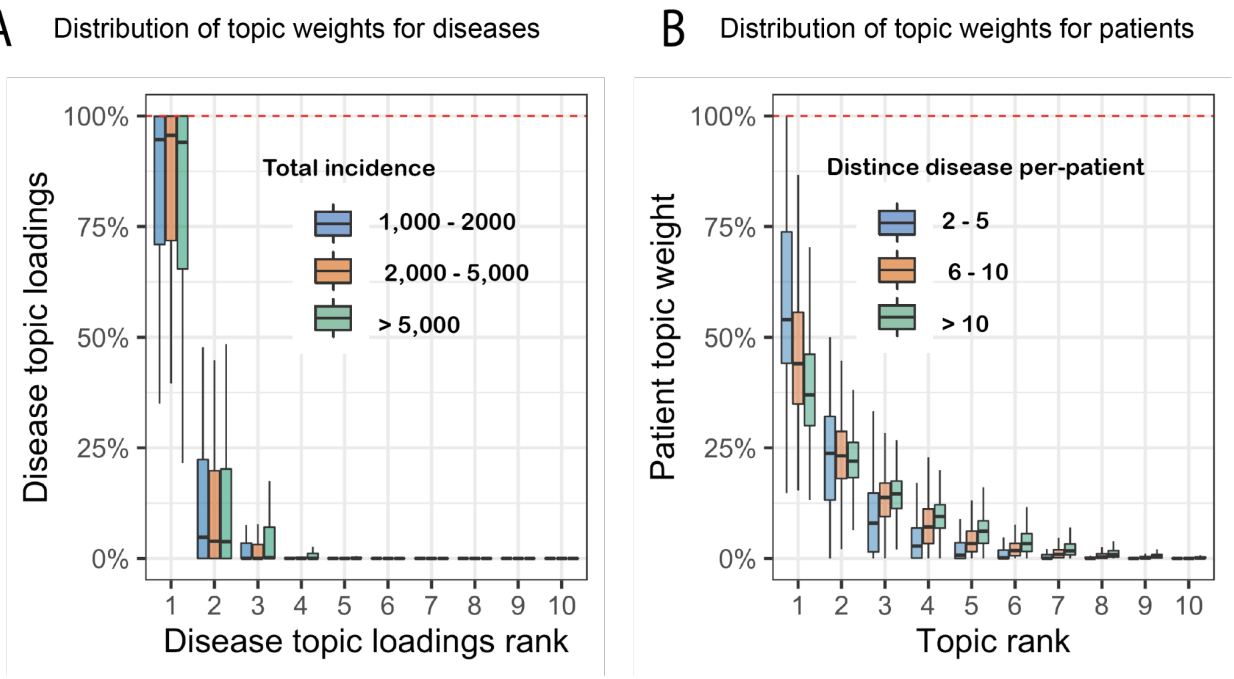
Supplementary Figure 11. Prediction log odds ratio of comorbidities. Diseases combinations (comorbidities) are extracted from topics loadings that are trained on a training set, using max value of average topic assignments (Methods). Roughly, the odds ratio of each disease combination is computed by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



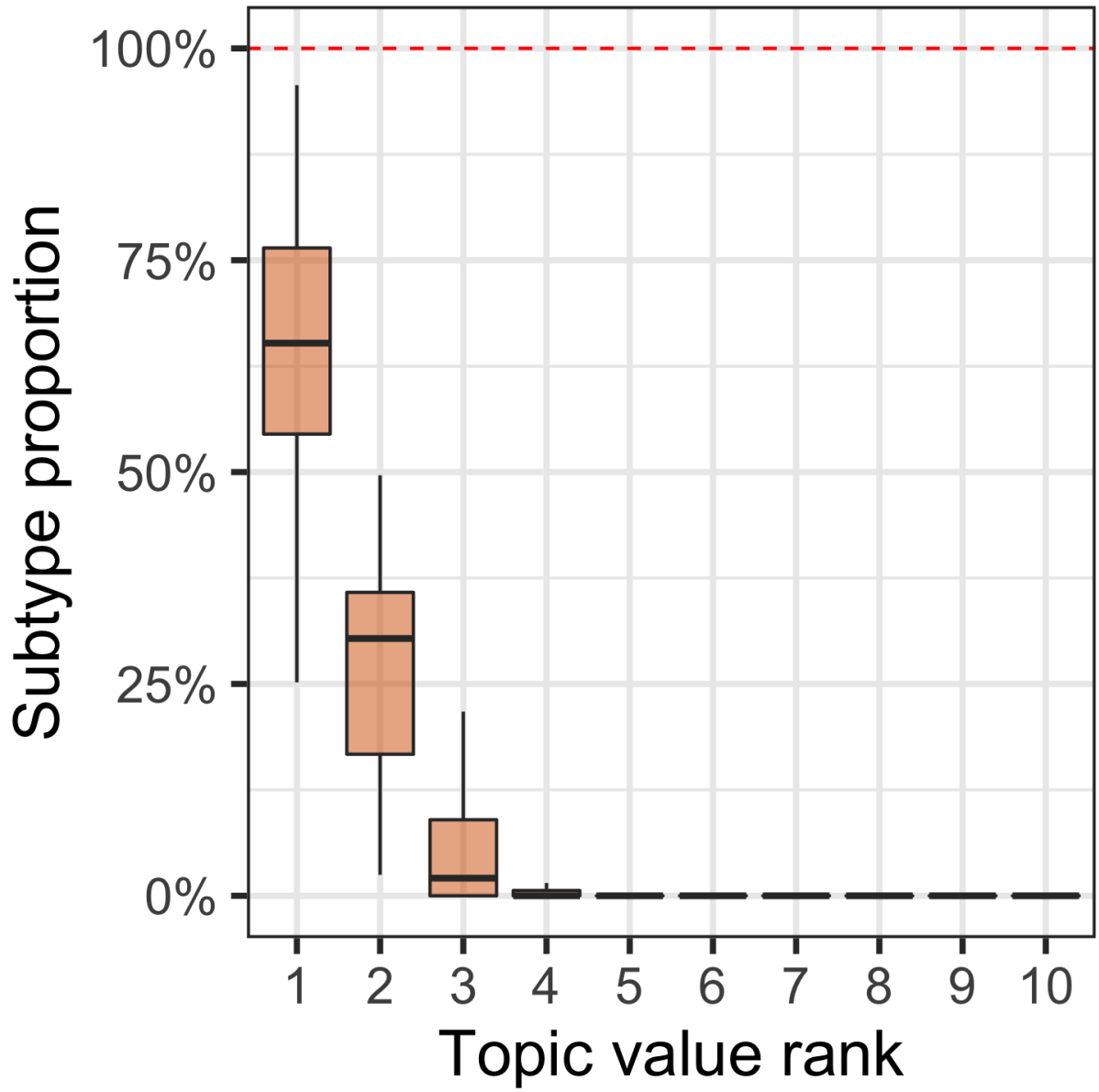
Supplementary Figure 12: Posterior topic distributions of female and male populations. The figure is the same as Figure 3A but comparing the topics that are inferred from female and male populations separately.



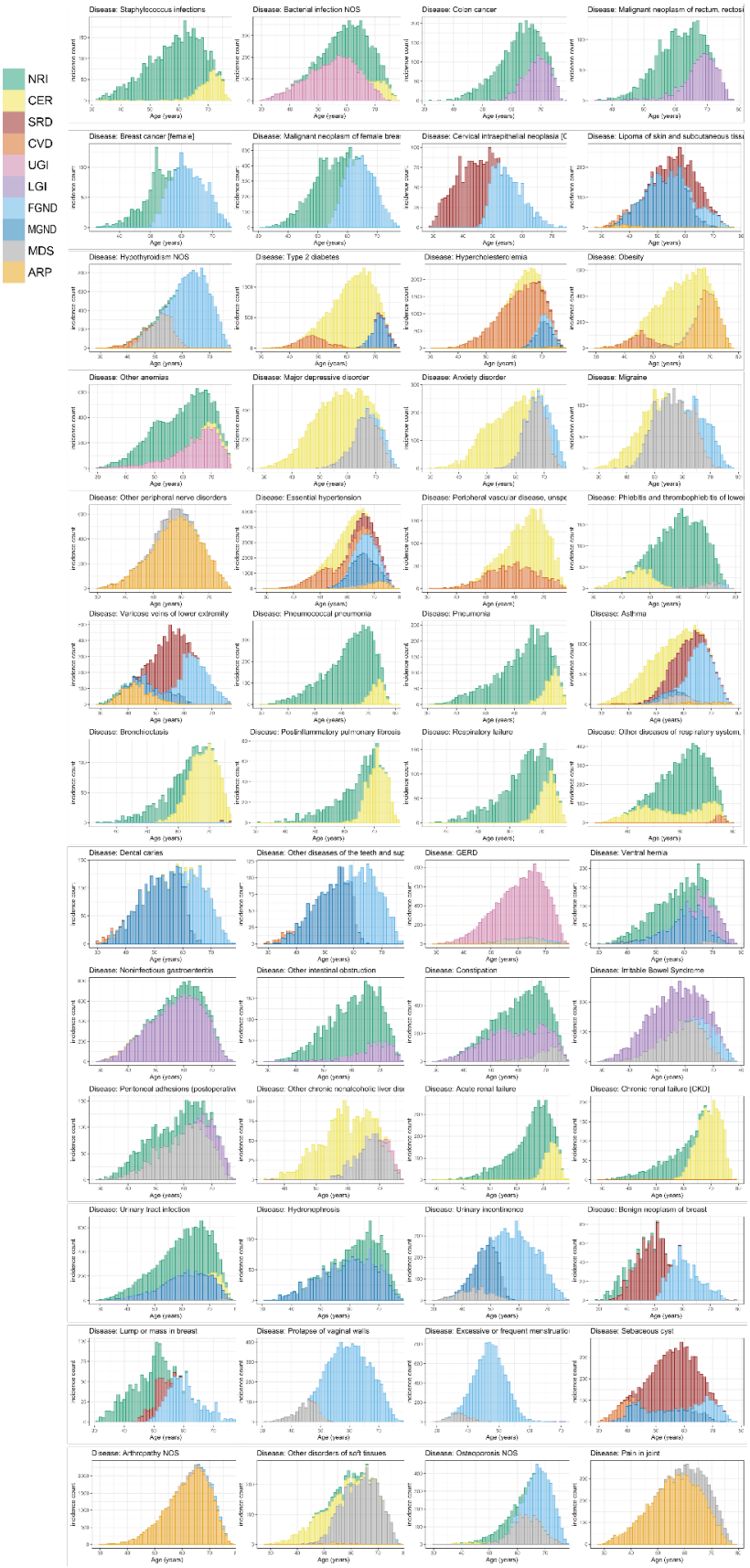
Supplementary Figure 13. Top seven diseases in each comorbidity topic. Seven diseases that have highest loading within the topic are shown for each comorbidity topic. Colour of the curves reflect the ordering of Phecodes. We chose seven for best visual presentation. Numerical results are reported in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



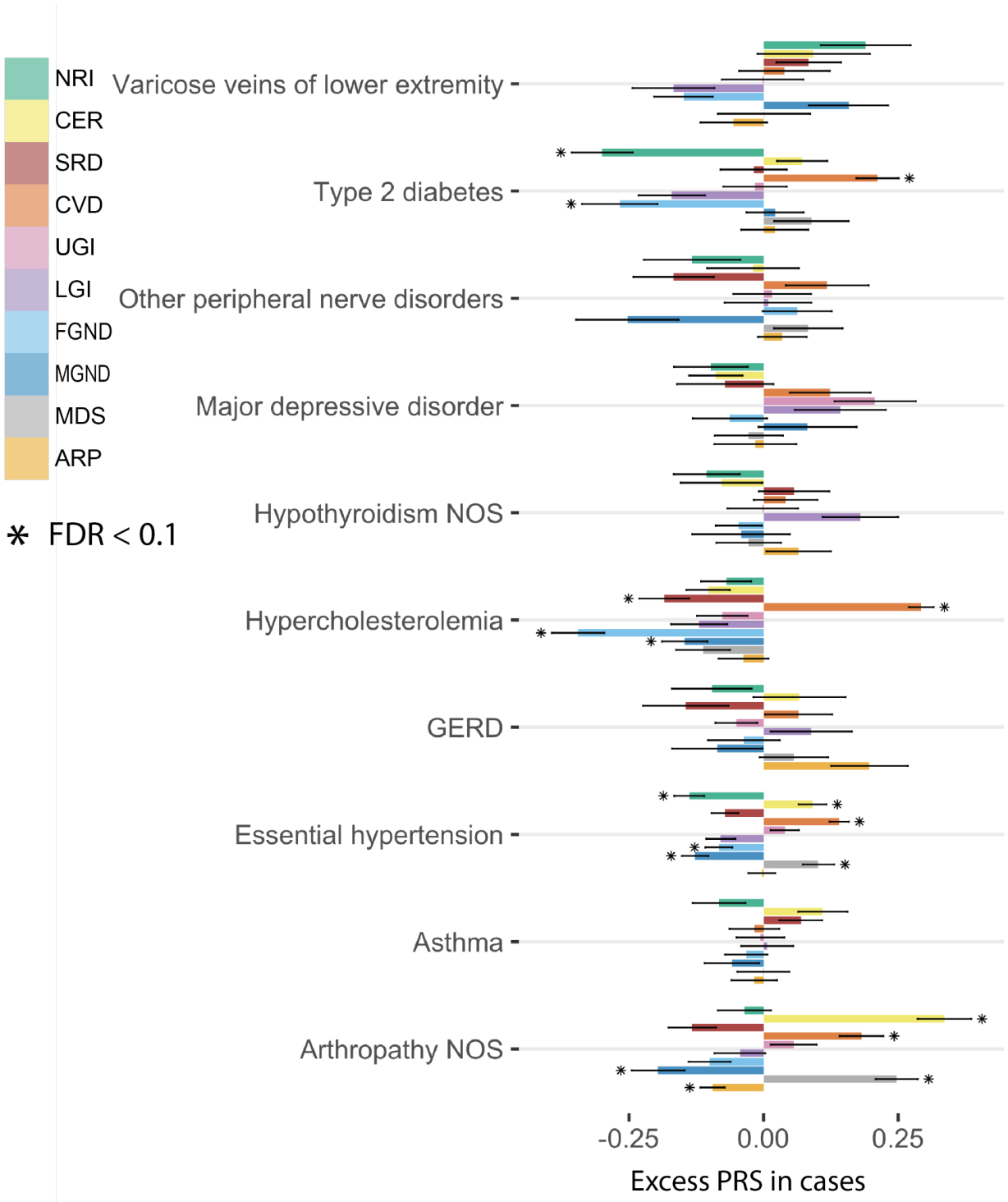
Supplementary Figure 14. Additional topic sparsity analysis. (a) Sparsity of disease topic loadings. Box plot shows the distribution of topic loading for disease of different incidence numbers. (b) Sparsity of patient topic weights. Box plot shows the topic weight distribution in decreasing order for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



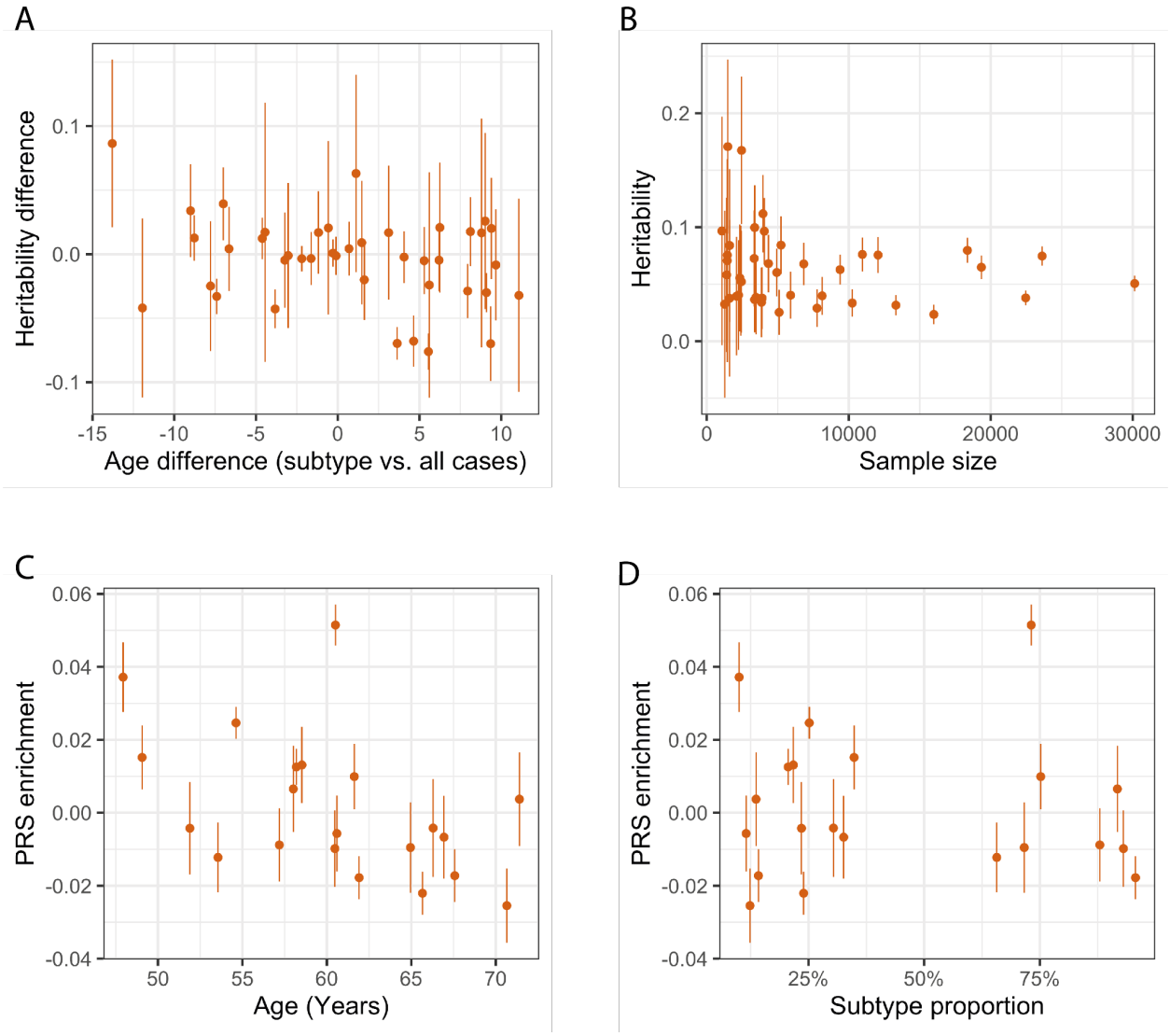
Supplementary Figure 15. Topic distribution for the 52 diseases that have at least 500 cases assigned to distinct topics. For each disease of the 52 diseases, we computed the proportion of diagnoses assigned to each subtype. The box plot shows the distribution of the subtype proportion from the largest (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: Malignant neoplasm of rectum, rectosi
Disease: Breast cancer [female]
Disease: Malignant neoplasm of female brea
Disease: Cervical intraepithelial neoplasia [C
Disease: Lipoma of skin and subcutaneous tissu
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: Peripheral vascular disease, unspe
Disease: Phlebitis and thrombophlebitis of lowe
Disease: Varicose veins of lower extremity
Disease: Pneumococcal pneumonia
Disease: Pneumonia
Disease: Asthma
Disease: Bronchiectasis
Disease: Postinflammatory pulmonary fibrosis
Disease: Respiratory failure
Disease: Other diseases of respiratory system,
Disease: Dental caries
Disease: Other diseases of the teeth and sup
Disease: GERD
Disease: Ventral hernia
Disease: Noninfectious gastroenteritis
Disease: Other intestinal obstruction
Disease: Constipation
Disease: Irritable Bowel Syndrome
Disease: Peritoneal adhesions (postoperative
Disease: Other chronic nonalcoholic liver dise
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: Excessive or frequent menstruatio
Disease: Sebaceous cyst
Disease: Arthropathy NOS
Disease: Other disorders of soft tissues
Disease: Osteoporosis NOS
Disease: Pain in joint
Age (years)
incidence count

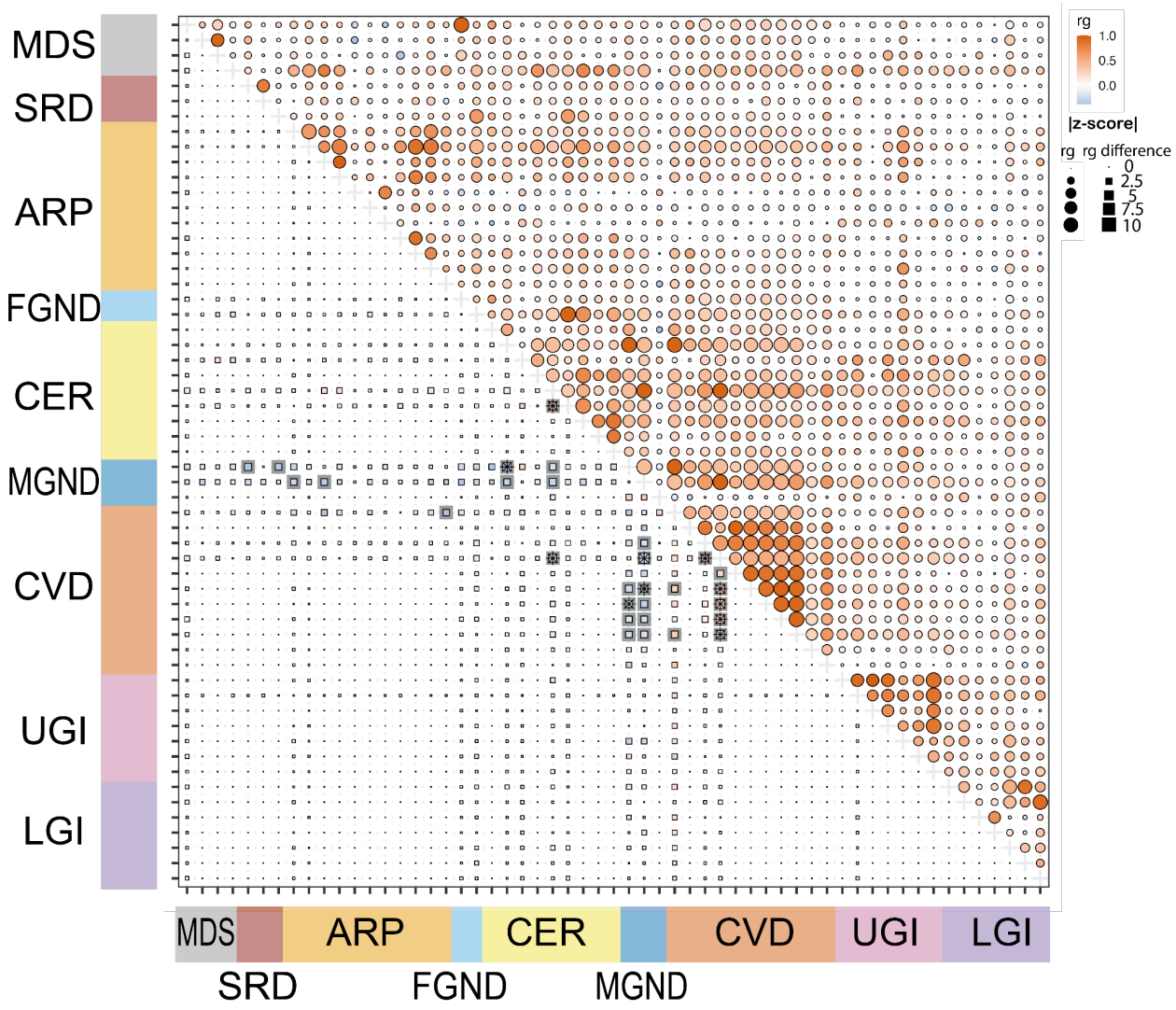
Supplementary Figure 16. Comorbidity subtype distribution over age for 52 diseases. Diseases shown are ordered by 13 phecode systems: infectious diseases, neoplasms, endocrine/metabolic, hematopoietic, mental disorders, neurological, circulatory system, respiratory, digestive, neoplasms, genitourinary, dermatologic, and musculoskeletal. Numerical results are reported in Supplementary Table 7.



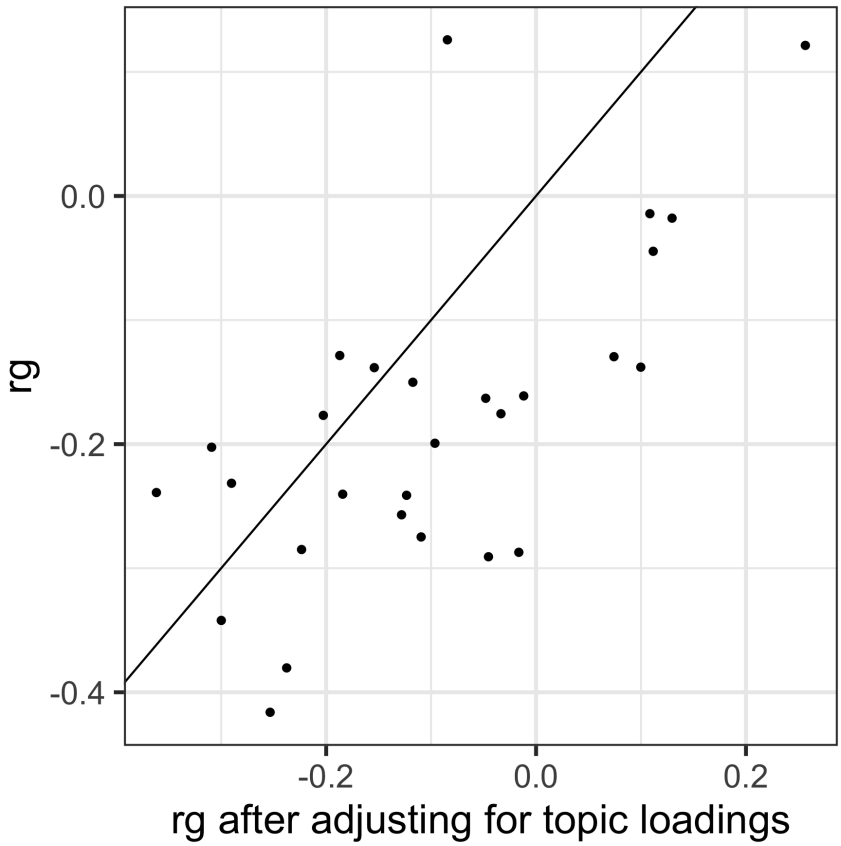
Supplementary Figure 17. Excess PRS analysis for all topics across 10 diseases (selected by heritability z-score). The bar plot shows the estimated changes in s.d. of PRS per unit changes in the patient topic weight in disease cases. The PRS is estimated using all the cases in British Isle Ancestry. The stars show disease-topic pairs that are significant at FDR = 0.1. Numerical results are reported in Supplementary Table 8.



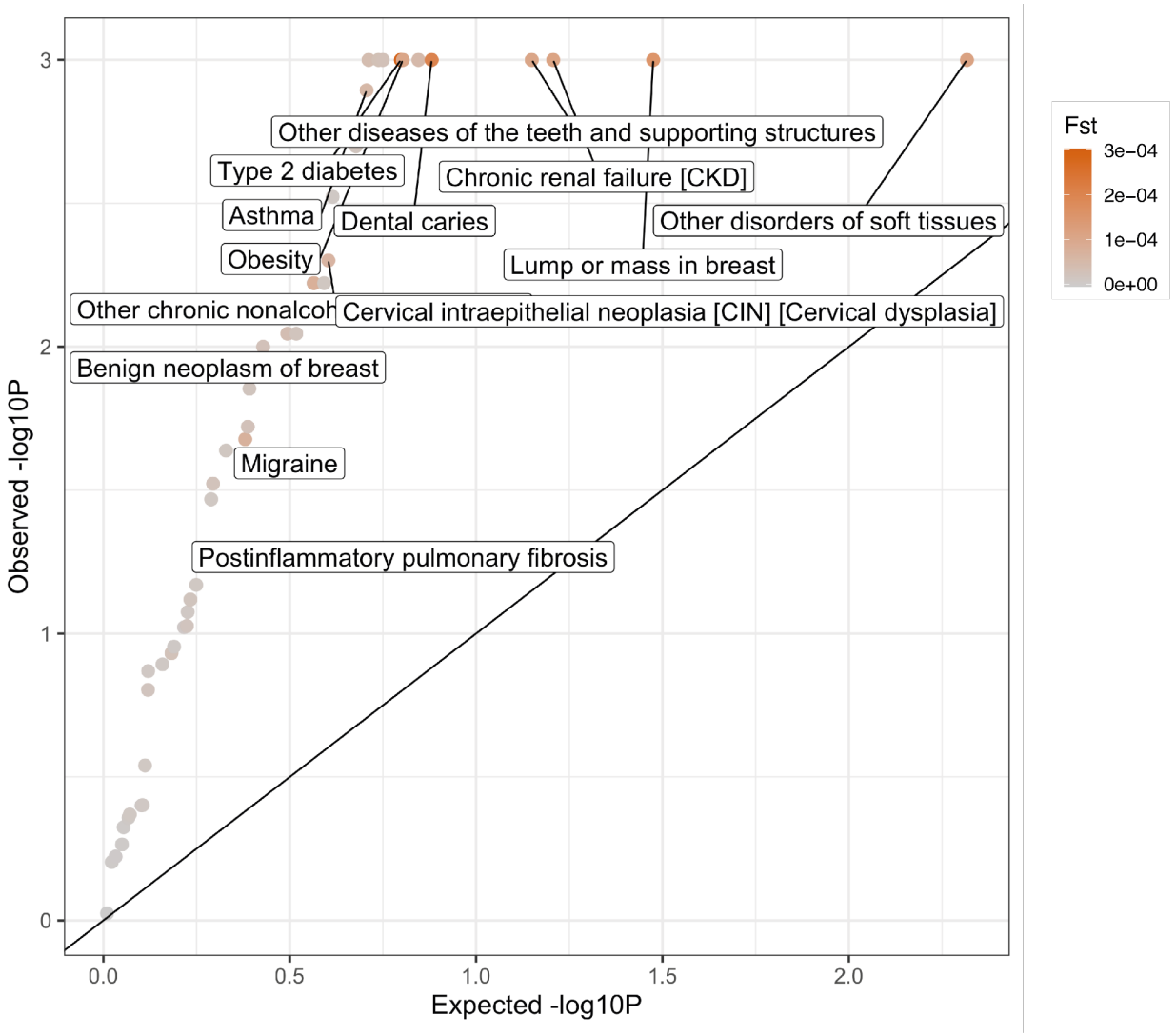
Supplementary Figure 18. Heritability and PRS are not associated with age and subtype size. (a) We plot heritability deviation from all cases versus age deviation from all cases of 41 subtypes (from 14 diseases that have heritability z-score above 5). The heritability is estimated by first performing mixed-effect association analysis using BOLT-LMM on imputed SNPs from the British Isle Ancestry then using LDSC. (b) Heritability for the subtypes plotted with the sample size of the subtype. (c-d) Excess PRS from Figure 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.



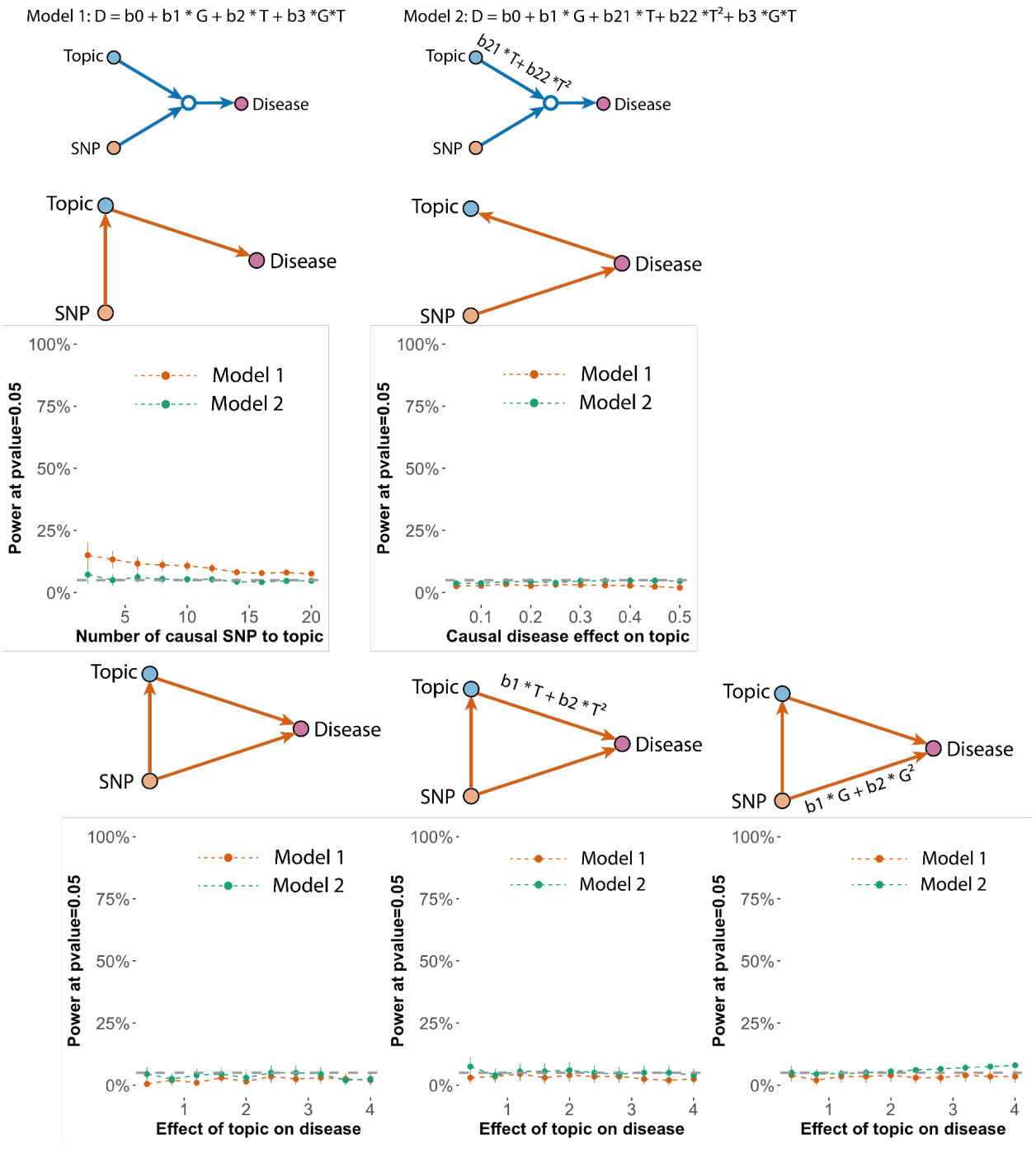
Supplementary Figure 19. Excess genetic correlation (r_g) between disease subtypes across disease subtypes. Lower left panel shows the r_g of disease-subtype or subtype-subtype pairs subtracted by the r_g of corresponding disease-disease pairs. Each row or column represents a disease or subtype. For disease-disease pairs, the excess r_g is not defined in the lower left panel, since the difference is 0. The upper right panel shows r_g of corresponding disease-disease pairs, where values could be duplicated as the same disease could have multiple subtypes. The 89 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. We kept 57 rows and columns for better visualisation by removing 32 diseases that have subtypes included. A star means FDR < 0.1, while a shade means a nominal statistical significance at P = 0.05.



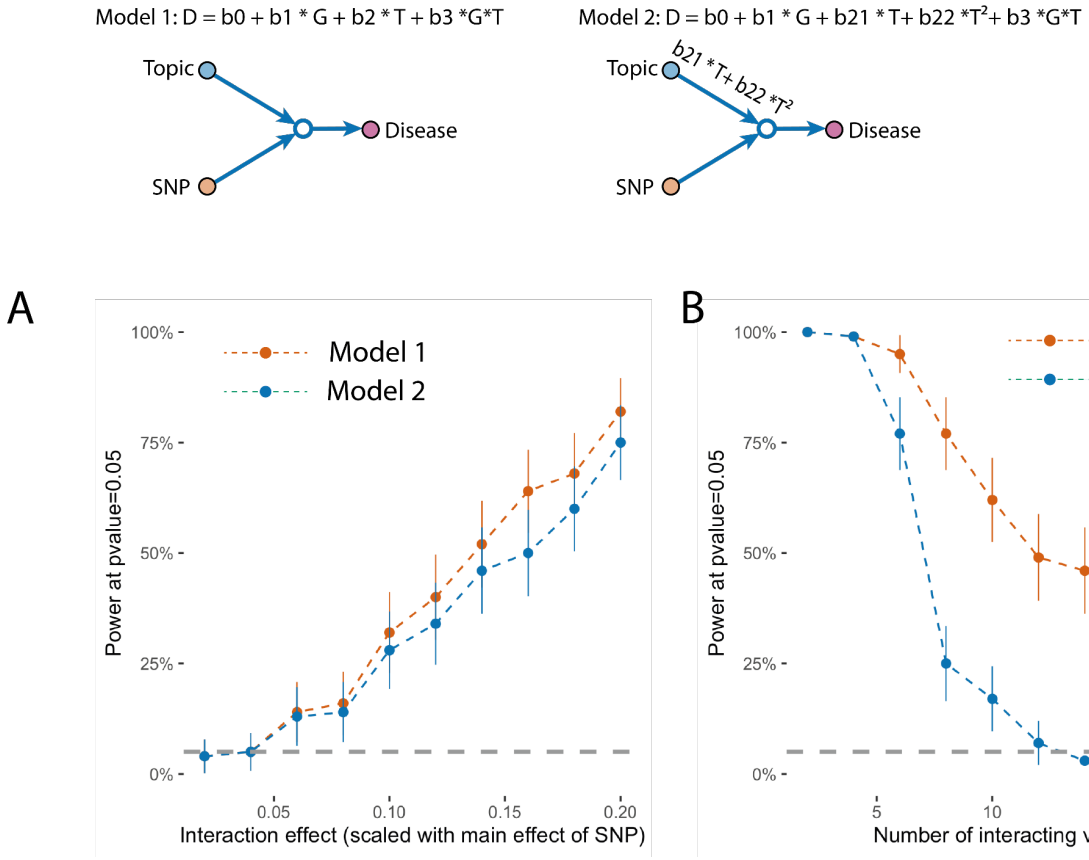
Supplementary Figure 20. Comparison of excess genetic correlation (r_g) between subtypes using summary statistics from case-control matched by topic weights (x-axis) and not matched by topic weights. The analysis is performed on subtypes of four diseases: type 2 diabetes, hypercholesterolemia, hypertension, and asthma. The excess r_g (shown in panel (a) of Figure 6) of two case-control matching strategies across 28 subtype pairs are compared and the effect lies along the diagonal line. The excess genetic correlation attenuated slightly when topic weights are matched, while it can not explain all the excess r_g.



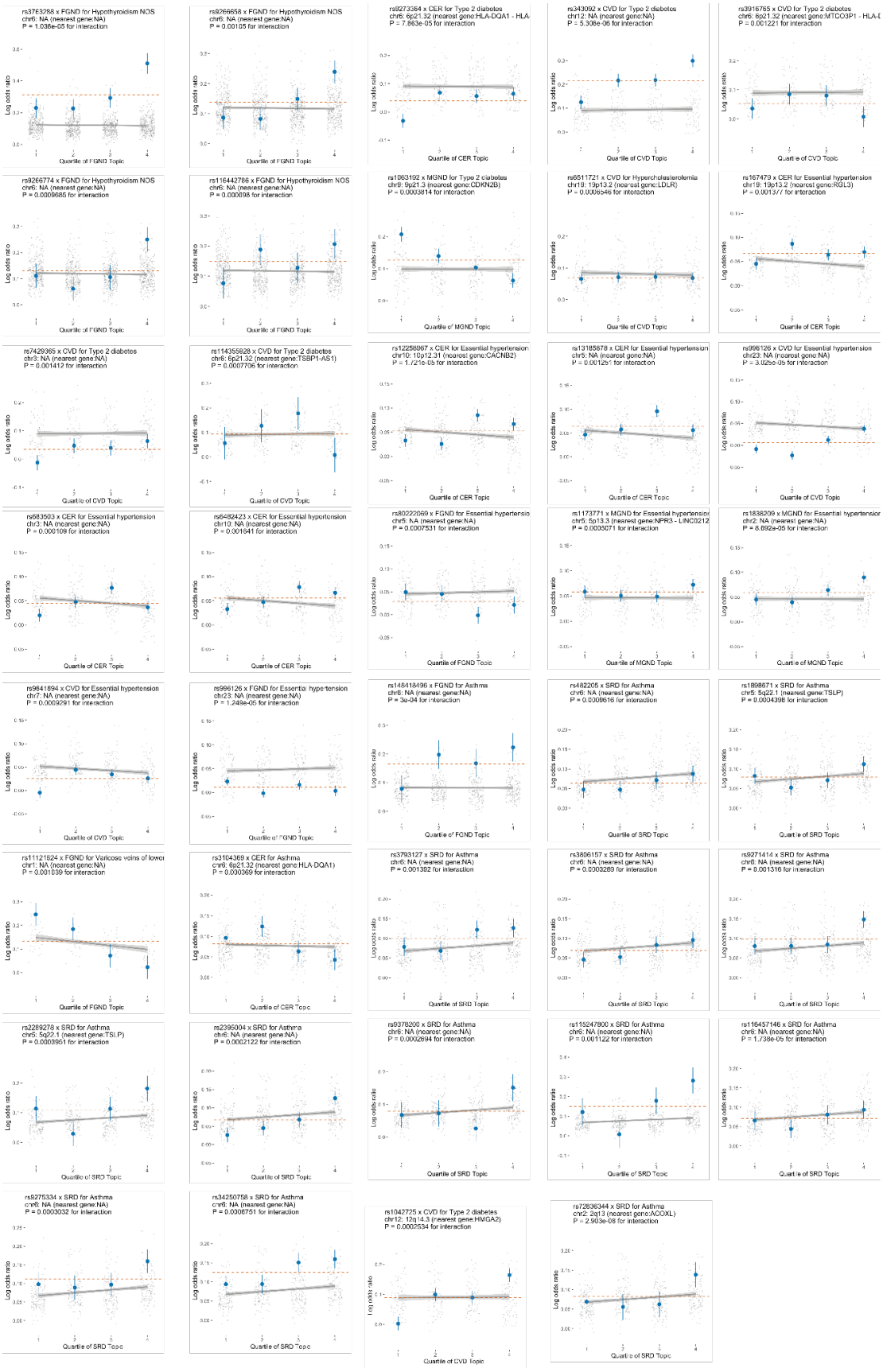
Supplementary Figure 21. Excess F_{ST} of disease subtypes compared with controls with matched topic weights. P-values are for testing case-F_{ST} significantly higher than controls of similar topic weight distribution. The permutation controls are sampled for 1,000 times with the same topic weights distribution and sample size to the disease subtypes. We focus on 49 of the 52 diseases which have more than one subgroup of at least 500 cases. Subtypes are defined based on the max value of the diagnosis-specific topic probability. Three diseases ("hypertension", "hypercholesterolemia", and "arthropathy") are excluded as there are not enough controls that match the topic weights of cases. The colour shows the value of F_{ST} across subtypes.



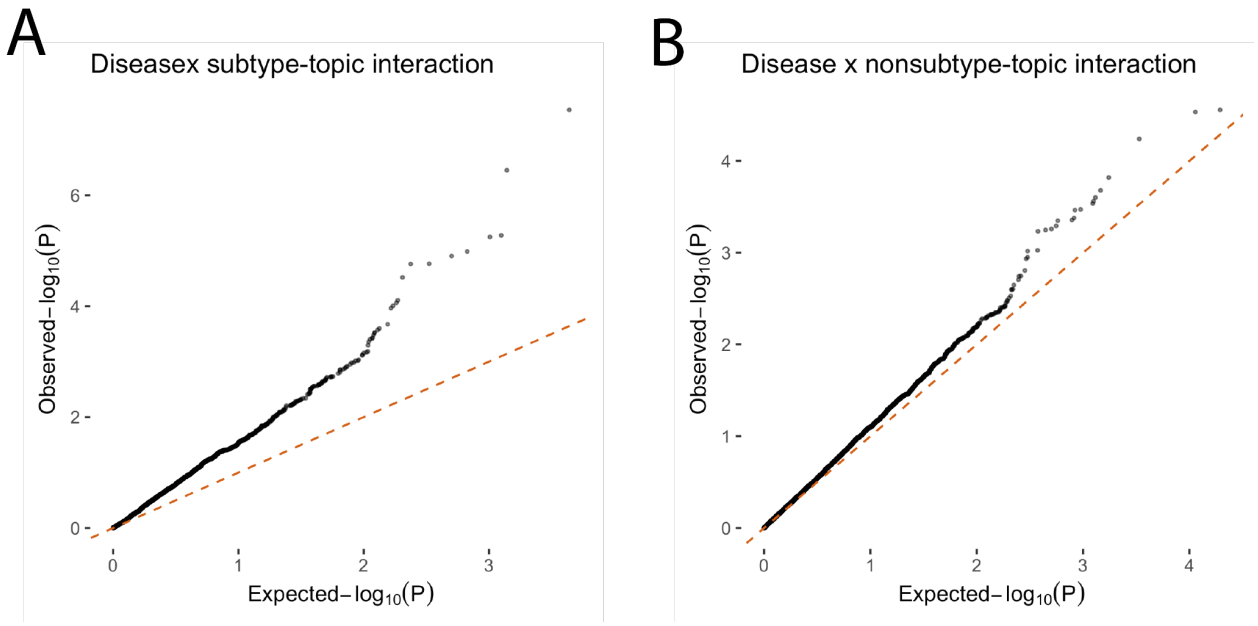
Supplementary Figure 22. Simulation analysis verified the SNP x topic interaction tests are calibrated when no actual interaction exists. We show the false positive rate of two models under five simulated model structures where no actual interaction exists (Methods). The false positive rate is computed as the power to detect interaction effects using model 1 (red; linear model) and model 2 (green; with non-linear main effect term) under P-value=0.05. 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.



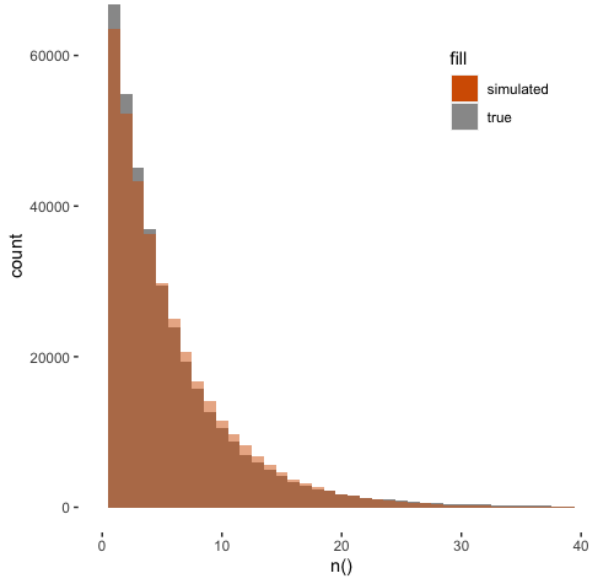
Supplementary Figure 23: Power to detect true SNP x topic interaction effect under simulation. (a) We simulated data with 10,000 individuals, topic-to-disease main effect size equal to 2, interaction effect from 0.04 to 0.4, and SNP-to-disease main effect size proportional to the interaction effect (0.02 to 0.2); disease diagnoses are generated using gaussian liability with top 20 percentile as cases. We tested for the SNP x topic interaction 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 SNP x Topic interaction in model 1 and model 2 with one variant at a time, which is the same strategy as most GWAS interaction tests. 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 22).


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 = 0.001221 for interaction
rs9266774 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 = 0.000698 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 = 0.00137 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
rs6468423 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 = 0.0005071 for interaction
rs1838209 x MGND for Essential hypertensior
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
rs996126 x FGND for Essential hypertension
chr23: NA (nearest gene:NA)
P = 1.249e-05 for interaction
rs145418496 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:AOC4L)
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

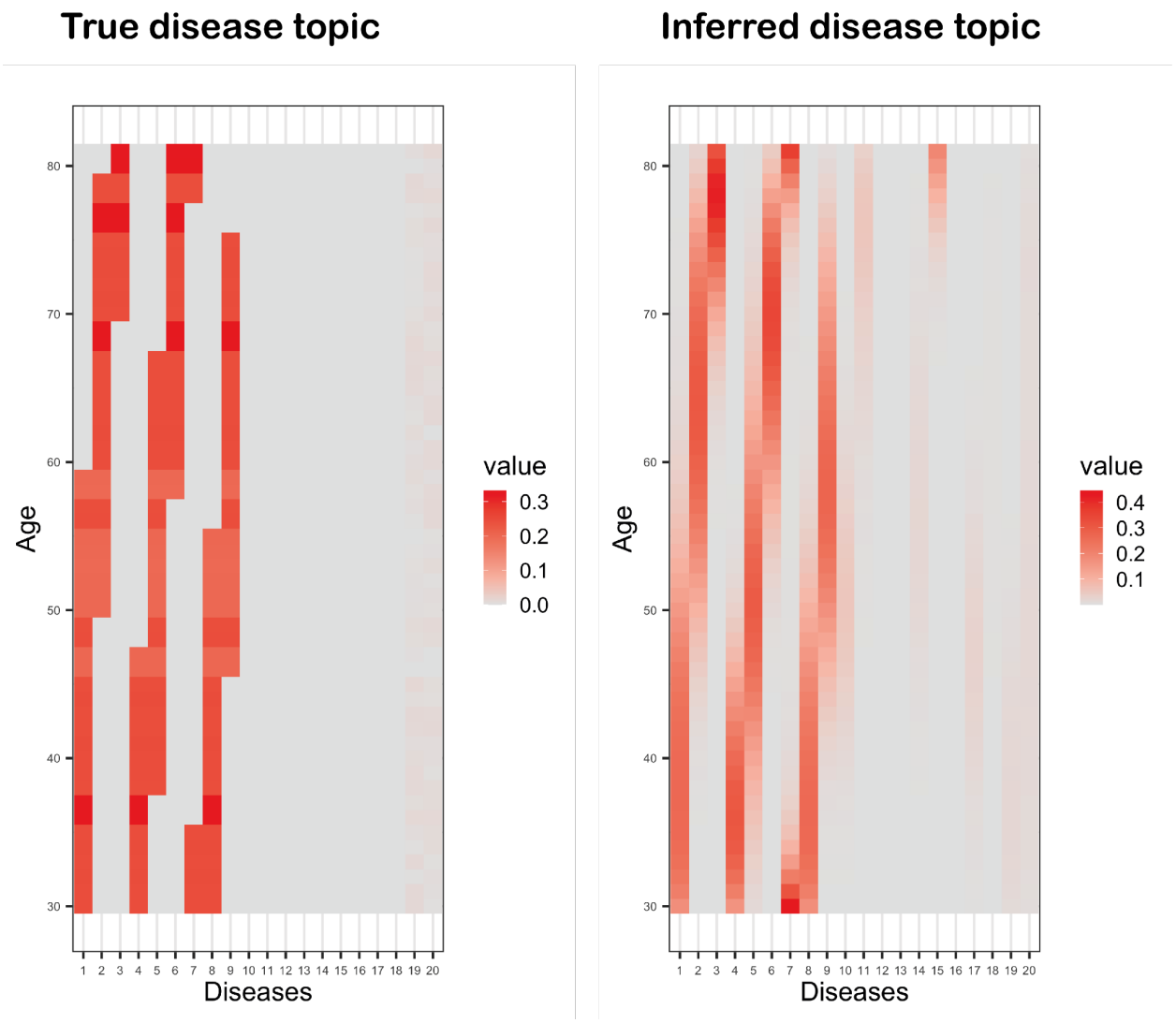
Supplementary Figure 24. Additional 39 SNPs (mapped genes in the parentheses) that have different effect sizes in different quantiles of topic weights. Blue dots are the effect sizes of the target SNP within each topic weights quartile; grey dots are the effect sizes of background SNPs which are genome-wide 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. The topic weights (of the topic being tested) are matched for both case and controls within each quartile. Numerical results are reported in Supplementary Table 14.



Supplementary Figure 25: QQ plot of SNP x topic 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 for testing the interaction effects with nonlinear topic-to-disease main effects (Model 2 in Supplementary Figure 22). (b) As a control to show the calibration of the tests, we plot the QQ-plot over the same set of GWAS SNPs, but over the topic that are not identified as subtypes of the disease by ATM.



Supplementary Figure 26. Comparison of simulated diagnosis per individual versus true UK Biobank data. 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 vs. inferred topic loadings from ATM. Left panel shows the topic loadings used to simulated 10,000 individuals; right panel shows the inferred topic loadings using topic loadings parametrized as cubic polynomials.

Supplementary Note

Xilin Jiang

November 25, 2022

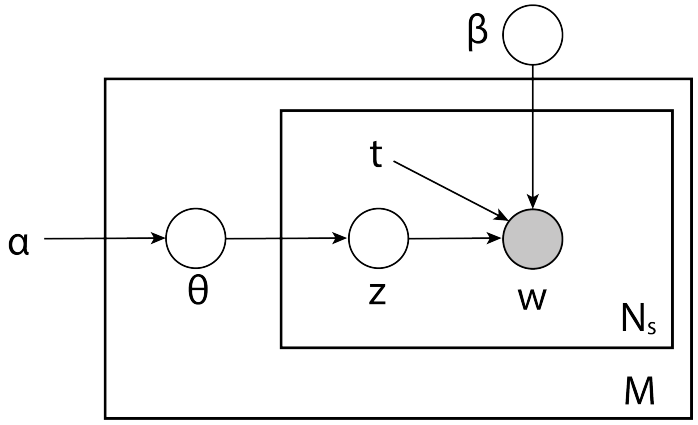
Contents

1 Usage instruction for the supplementary notes

The purpose of this supplementary note is to provide a self-contained explanation of the mathematical basis underlying our topic based model. Therefore, the materials are not meant to provide new discoveries but to help readers to derive all of our inference methods without referring to external materials (though we do listed references to text wherever appropriate). As a consequence, we made no efforts to condense steps, and opt to expand with more details when we feel it is necessary.

2 Generative process of a curve topic model

We constructed a Bayesian hierarchical model to infer latent risk profiles for common diseases. In summary, the model assumes there exist a few disease topics that underlie



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

many common diseases. Each topic is age-evolving and contain risk trajectories for all diseases considered. An individual's risk for each diseases is determined by the weights of all topics. The indices in this note are as follows:

$$s = 1, ..., M;$$

$$n = 1, ..., N_s;$$

$$i = 1, ..., K;$$

$$j = 1, ..., D;$$

where M is the number of subjects, N_s is the number of records within s^{th} subject, K is number of topics, and D is the total number of diseases we are interested in. The generative process (Supplementary Figure 1) is as follows:

- $\theta \in \mathcal{R}^{M \times K}$ is the topic weight for all individuals, each row of which ($\in \mathcal{R}^K$) is assumed to be sampled from a Dirichlet distribution with parameter α. α is set as a hyper parameter.

$$\theta_s \sim Dir(\alpha).$$

- $\mathbf{z} \in \{1, 2, ..., K\}^{\sum_s N_s}$ is the topic assignment for each diagnosis $\mathbf{w} \in \{1, 2, ..., D\}^{\sum_s N_s}$. 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 parameter equal to s^{th} individual topic weight.

$$z_{sn} \sim Multi(\theta_s).$$

- $\beta(t) \in \mathcal{F}(t)^{K \times D}$ is the topic which is $K \times D$ functions of age t. $\mathcal{F}(t)$ 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 $\mathcal{F}(t)$ to be spline or polynomial functions:

$$\beta_{ij}(t) = \frac{\exp(\boldsymbol{p}_{ij}^T \phi(t))}{\sum_{j=1}^D \exp(\boldsymbol{p}_{ij}^T \phi(t))},$$

  where $\boldsymbol{p}_{ij} = \{p_{ijd}\},\ d = 1, 2, ..., P$; P is the degree of freedom than controls the smoothness; $\phi(t)$ is polynomial and spline basis for age t.

- $w \in \{1, 2, ..., D\}^{\sum_s N_s}$ are observed diagnoses. The n^{th} diagnosis of s^{th} individual w_{sn} is sampled from the topic $\beta_{z_{sn}}(t)$ chosen by z_{sn}:

$$w_{sn} \sim Multi(\boldsymbol{\beta}_{z_{sn}}(t_{sn})),$$

  here t_{sn} is the age of the observed age-at-onset of the observed diagnosis w_{sn}.

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

In the first step we could work on the likelihood function fore each patient to estimate posterior distributions of patient specific variables θ and z. The likelihood function for s^{th} individual is as follows:

$$\begin{aligned}
&\ln p(\mathbf{w}, \mathbf{z}, \theta | \alpha, \beta) = \ln p(\theta|\alpha) + \sum_{n=1}^{N_s} \{\ln p(z_n|\theta) + \ln p(w_n|z_n, \beta)\}, \\
&p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_i^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i - 1}, \\
&p(z_n|\theta) = \theta_{1(i=z_n)}, \\
&p(w_n|z_n, \beta(t_n)) = \beta_{1(i=z_n), 1(j=w_n)}(t_n).
\end{aligned} \tag{1}$$

Due to the computational cost of simultaneously modelling hundreds of diseases in the biobank and the inference accuracy consideration (which we will explain in section 4), we adopted a collapsed variational methods for this step. The method is motivated by [1]. Detailed explanation on why we chose this rather sophisticated methods rather than the commonly used mean filed methods is discussed in section 4, for those interested.

In the second step, we treated the β as parameter of the model and seek to maximise the evidence function $p(\mathbf{w}|\alpha, \beta)$ (obtained by integrate out θ and z from the likelihood function). Directly working on the evidence function is implausible, therefore we work on the evidence lower bound, where we made use of the posterior distribution $q(\mathbf{z}, \theta)$ estimated in previous step.

$$\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha) = \mathbf{E}_q\{\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) - \ln q(\mathbf{z}, \theta)\}. \tag{2}$$

We will see this is still not easily achieved, therefore we applied a local variational method to find an approximate solution. For an easy introduction to local variational inference, see chapter 10.5 of [2]. Details of the inference will be explained in section 3.

3 Inference of model posterior distribution and parameters

The model inference will be performed by alternation an E-step and a M-step. The EM algorithm will guarantee good convergence properties. For both steps, variational methods will be used to approximate the distribution, though the techniques are very different. We have tested under realistic parameters, these approximated distribution are close to the true distribution.

3.1 Collapsed variational inference to estimate patient-level posterior distribution $q(z, \theta)$

The variational inference aims to approximate posterior $p(\mathbf{z}, \theta|\mathbf{w}, \alpha, \beta)$ using variational distributions $q(\mathbf{z}, \theta)$ that has some constraint on, which makes them easier to estimate. The most widely used form of variational distribution is the factorised ones, where we assume target posterior distributions are independently distributed, i.e. $q(\mathbf{z}, \theta) = q(\mathbf{z})q(\theta)$. We will derive the inference using this assumption and compare it with the collapsed variational inference in section 4.

The latent variable model using Dirichlet distribution is typically designed to model text, where a document is equivalent as a patient in our model. A document will have thousands of words (equivalent of our diagnoses), which provides strong information to tolerate stronng assumptions on $q(\mathbf{z}, \theta)$. To estimate a model with less diagnoses per patient, flexible variational distributions are preferred for approximation accuracy. Here we adopted a collapsed variational method, which is is more accurate than the mean-field variational inference method. [1] The idea is to only assume a factorization over $q(\mathbf{z})$, but

not between $\mathbf{z}$ and θ. Therefore the assumptions and lower bound of evidence became (note we are considering likelihood function for only the s^{th} patient from now on, as all of patients are independend conditional on α, β):

$$\begin{aligned} q(\mathbf{z}, \theta) &= q(\theta|\mathbf{z}) \prod_n q(z_n), \\ \mathcal{L}(\mathbf{z}, \theta, \beta, \alpha) &= \mathbf{E}_q\{\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) - \ln q(\mathbf{z}) - \ln q(\theta|\mathbf{z})\} \\ &= \mathbf{E}_{q(z)}\{\mathbf{E}_{q(\theta|z)}\{\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) - \ln q(\theta|\mathbf{z})\} - \ln q(\mathbf{z})\} \end{aligned} \tag{3}$$

Maximise $\mathbf{E}_{q(\theta|z)}\{\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) - \ln q(\theta|\mathbf{z})\}$ with respect to $q(\theta|\mathbf{z})$ will give us $q(\theta|\mathbf{z}) = p(\theta|\mathbf{w}, \mathbf{z}, \alpha, \beta)$. The maximisation is achieved similarly to the mean-field approximation where the evidence is decomposed into a lower bound and KL divergence, where lower bound is maximised when KL divergence is 0. Using methods provided in Teh et al, the lower bound could then be simplified to:

$$\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha) = \mathbf{E}_{q(\mathbf{z})}\{\ln p(\mathbf{w}, \mathbf{z}|\alpha, \beta) - \ln q(\mathbf{z})\}$$

The optimisation of this lower bound is same to those provided in collapsed gibbs sampling, where we first marginalise over θ.

$$\begin{aligned} p(\mathbf{w}, \mathbf{z}|\alpha, \beta) &= \int_\theta \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_i^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i + \sum_n z_{ni} - 1} \cdot \prod_{i=1}^K \prod_{j=1}^D \beta_{ij}^{\sum_n z_{ni} w_{nj}} \\ &= \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_i^K \Gamma(\alpha_i)} \frac{\prod_i^K \Gamma(\alpha_i + \sum_n z_{ni})}{\Gamma(\sum_{i=1}^K \alpha_i + N_s)} \cdot \prod_{i=1}^K \prod_{j=1}^D \beta_{ij}^{\sum_n z_{ni} w_{nj}}. \end{aligned} \tag{4}$$

From this marginal complete data likelihood, we could derive the conditional distribution $p(z_{n'} = k|\mathbf{z}_{\neg \mathbf{n}'}, \mathbf{w}, \alpha, \beta)$ (as in collapsed gibbs sampling) to evaluate the dependency within $\mathbf{z}$. Here $\neg n'$ refer to indices of all words excluding n'.

$$\begin{aligned} p(z_{n'}|\mathbf{z}_{\neg \mathbf{n}'}, \mathbf{w}, \alpha, \beta) &= \frac{p(z_{n'}, \mathbf{z}_{\neg \mathbf{n}'}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{z}_{\neg \mathbf{n}}, \mathbf{w}|\alpha, \beta)} \\ &= \frac{p(\mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{z}_{\neg \mathbf{n}}, w_{\neg n}|\alpha, \beta) p(w_{n'}|\alpha, \beta)} \\ &\propto \frac{\prod_i^K \Gamma(\alpha_i + \sum_n z_{ni}) \prod_{i=1}^K \prod_{j=1}^D \beta_{ij}^{\sum_n z_{ni} w_{nj}}}{\prod_i^K \Gamma(\alpha_i + \sum_{\neg n'} z_{ni}) \prod_{i=1}^K \prod_{j=1}^D \beta_{ij}^{\sum_{\neg n'} z_{ni} w_{nj}}} \\ &\propto \prod_i^K (\alpha_i + \sum_{n \in \neg n'} z_{ni})^{z_{n'i}} \prod_{i=1}^K \prod_{j=1}^D \beta_{ij}^{z_{n'i} w_{n'j}}. \end{aligned} \tag{5}$$

For a large N_s, $(\alpha_i + \sum_{n\in\neg n'} z_{ni})$ will be approximately the same across n', therefore $z_{n'}$ will be less dependent on $\mathbf{z}_{\neg n'}$.

$$\lim_{N_s\to\infty} p(z_{n'}|\mathbf{z}_{\neg\mathbf{n}'}, \mathbf{w}, \alpha, \beta) \propto \prod_i^K \Big[(\alpha_i + N_s\theta_i)\prod_{j=1}^D \beta_{ij}^{w_{n'j}}\Big]^{z_{n'i}},$$

where distribution of $\mathbf{z}$ factorises over n within a single subjects. Therefore, the $q^*(\mathbf{z})$ in equation 17 which factorises over n could approximate $p(\mathbf{z}|\mathbf{w}, \alpha, \beta)$ accurately. However, N_s is likely to be small in the patient dataset, therefore the mean-field approximation in equation 17 would be less accurate as it does not include any dependency between $\mathbf{z}_n$ and $\mathbf{z}_{\neg\mathbf{n}'}$. We therefore adopt the strategies proposed by Teh et al to use variatioanal distribution to approximate the marginal distribuion in equation 4:

$$\begin{aligned}
\ln q^*(z_{n'}) &= \mathbf{E}_{q(\mathbf{z}_{\neg\mathbf{n}'})}\{\ln p(\mathbf{w}, \mathbf{z}|\alpha, \beta)\} \\
&= \mathbf{E}_{q(\mathbf{z}_{\neg\mathbf{n}'})}\{\sum_{i=1}^K \ln\Gamma(\alpha_i + \sum_n z_{ni}) + \sum_{i=1}^K\sum_{j=1}^D z_{n'i}w_{n'j}\ln\beta_{ij}\} + const(z_{n'}) \\
&= \sum_{i=1}^K z_{n'i}\Big(\mathbf{E}_{q(\mathbf{z}_{\neg\mathbf{n}'})}\{\ln(\alpha_i + \sum_{n\in\neg n'} z_{ni})\} + \sum_{j=1}^D w_{n'j}\ln\beta_{ij}\Big) + const(z_{n'})
\end{aligned} \tag{6}$$

We now have the form of multinomial distribution of $z_{n'}$. The key lies in how to estimate $\mathbf{E}_{q(\mathbf{z}_{\neg\mathbf{n}'})}\{\ln(\alpha_i + \sum_{n\in\neg n'} z_{ni})\}$. Teh et al [1] proposed a Gaussian approximation which could improve computation efficiency by magnitudes. We first expand $\ln(\alpha_i + \sum_{n\in\neg n'} z_{ni})$ by Taylor expansion:

$$\ln(\alpha_i + \sum_{n\in\neg n'} z_{ni}) = \ln(\alpha_i + n_0) + \frac{\sum_{n\in\neg n'} z_{ni} - n_0}{(\alpha_i + n_0)} - \frac{(\sum_{n\in\neg n'} z_{ni} - n_0)^2}{2(\alpha_i + n_0)^2},$$

where we included only first two terms. If setting $n_0 = \mathbf{E}_{q(\mathbf{z}_{\neg\mathbf{n}'})}\{\sum_{n\in\neg n'} z_{ni}\} = \sum_{n\in\neg n'} \mathbf{E}_q\{z_{ni}\}$, we get:

$$\mathbf{E}_{q(\mathbf{z}_{\neg\mathbf{n}'})}\{\ln(\alpha_i + \sum_{n\in\neg n'} z_{ni})\} = \ln(\alpha_i + n_0) - \frac{Var_q[\sum_{n\in\neg n'} z_{ni}]}{2(\alpha_i + n_0)^2}. \tag{7}$$

where, $Var_q[\sum_{n\in\neg n'} z_{ni}] = \sum_{n\in\neg n'}(1 - \mathbf{E}_q\{z_{ni}\})\mathbf{E}_q\{z_{ni}\}$ .Plugging this into equation 6 and notice the normalization of multinomial distribution, we get:

$$z_{n'i} \sim \mathbf{Cat}\Big(\frac{(\alpha_i + n_0)\exp\big(-\frac{Var_q[\sum_{n\in\neg n'} z_{ni}]}{2(\alpha_i+n_0)^2} + \sum_{j=1}^D w_{n'j}\ln\beta_{ij}\big)}{\sum_{i=1}^K(\alpha_i + n_0)\exp\big(-\frac{Var_q[\sum_{n\in\neg n'} z_{ni}]}{2(\alpha_i+n_0)^2} + \sum_{j=1}^D w_{n'j}\ln\beta_{ij}\big)}\Big). \tag{8}$$

For prediction tasks, the posterior θ could be evaluated using distribution of $\mathbf{z}$.

$$\theta \sim \mathbf{Dir}(\alpha + \sum_{n=1}^{N_s} \mathbf{E}_q\{z_n\})$$

The evidence lower bound over all subjects is as follows:

$$\begin{aligned}
\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha) =& \mathbf{E}_{q(\mathbf{z})}\{\ln p(\mathbf{w}, \mathbf{z}|\alpha, \beta) - \ln q(\mathbf{z})\} \\
=& \sum_{s=1}^{M} \Big(\ln \Gamma(\sum_{i=1}^{K} \alpha_i) - \sum_{i}^{K} \ln \Gamma(\alpha_i) - \ln \Gamma(N_s + \sum_{i=1}^{K} \alpha_i) + \\
& \sum_{i=1}^{K} \mathbf{E}_{q(\mathbf{z})}\{\ln \Gamma(\alpha_i + \sum_{n} z_{ni})\} + \\
& \sum_{n=1}^{N_s} \sum_{i=1}^{K} \mathbf{E}\{z_{sni}\} \sum_{j=1}^{D} w_{snj} \ln \beta_{ij} \Big) - \\
& \sum_{s=1}^{M} \Big(\sum_{n=1}^{N_s} \sum_{i=1}^{K} \mathbf{E}\{z_{ni}\} \ln \mathbf{E}\{z_{ni}\} \Big),
\end{aligned} \tag{9}$$

where we need to approximate $\mathbf{E}_{q(\mathbf{z})}\{\ln \Gamma(\alpha_i + \sum_n z_{ni})\}$. Making use of the Stirling's approximation, we found $\ln \Gamma(z) = (z - \frac{1}{2}) \ln(z) - z + \frac{1}{12z} + \frac{1}{2} \ln(2\pi)$ could approximate $\ln \Gamma(z)$ accurately for $z > 1$. Therefore, by plugging in Stirling's approximation and reuse

equation 7 we could approximate this expectation:

$$\begin{aligned}
\mathbf{E}_{q(\mathbf{z})}\{\ln\Gamma(\alpha_i+\sum_n z_{ni})\} =&\mathbf{E}_{q(\mathbf{z})}\{(\alpha_i+\sum_n z_{ni})\ln(\alpha_i+\sum_n z_{ni}) \\
&-\frac{1}{2}\ln(\alpha_i+\sum_n z_{ni})-(\alpha_i+\sum_n z_{ni})+\frac{1}{12(\alpha_i+\sum_n z_{ni})}+\frac{1}{2}\ln(2\pi)\} \\
=&\mathbf{E}_{q(\mathbf{z})}\{(\alpha_i+n_0)\ln(\alpha_i+n_0)+\frac{(\sum_n z_{ni}-n_0])^2}{2(\alpha_i+n_0)} \\
&-\frac{1}{2}\ln(\alpha_i+n_0)+\frac{(\sum_n z_{ni}-n_0])^2}{4(\alpha_i+n_0)2} \\
&-(\alpha_i+\sum_n z_{ni}) \\
&+\frac{1}{12(\alpha_i+n_0)}+\frac{(\sum_n z_{ni}-n_0])^2}{12(\alpha_i+n_0)^3}+\frac{1}{2}\ln(2\pi)\} \\
=&(\alpha_i+n_0)\ln(\alpha_i+n_0)-\frac{1}{2}\ln(\alpha_i+n_0)-(\alpha_i+n_0)+\frac{1}{12(\alpha_i+n_0)} \\
&+Var_q[\sum_n z_{ni}]\big(\frac{1}{2(\alpha_i+n_0)}+\frac{1}{4(\alpha_i+n_0)2}+\frac{1}{12(\alpha_i+n_0)^3}\big) \\
&+\frac{1}{2}\ln(2\pi),
\end{aligned} \tag{10}$$

Note here the first order terms in Taylor expansion are cancelled after taking the expectation and setting $n_0 = \sum_n \mathbf{E}_q\{z_{sni}\}$. The variance are computed by applying the independent assumption over $q(z_n)$: $Var_q[\sum_n z_{ni}] = \sum_n(1-\mathbf{E}_q\{z_{ni}\})\mathbf{E}_q\{z_{ni}\}$.

3.2 Estimate topic profiles $\beta(t)$

In the conventional topic modeling, the topic values could be estimated by directly maximising the evidence lower bound with a constraint $\sum_{j=1}^{D}\beta_{ij}=1$, which is described in section 4. Here we estimate the topic as functions of age by parameterising each β_{ij} as a function of age. The only related term in likelihood function (equation 1) is:

$$\ln p(w_n|z_n,\beta(t_n)) = \sum_{i=1}^{K} z_{sni}\sum_{j=1}^{D} w_{snj}\ln\pi(\beta_{ij}(t_n)),$$

where we use softmax function to ensure topics are each a multinomial distribution:

$$\pi(\beta_{ij}(t_n)) = \frac{\exp(\boldsymbol{\beta}_{ij}^T\phi(t_n))}{\sum_{j=1}^{D}\exp(\boldsymbol{\beta}_{ij}^T\phi(t_n))}.$$

We used spline/polynomial functions to model age. The goal is to estimate spline/polynomial coefficients $\boldsymbol{\beta}_{ij} = \{\beta_{ijd}\},\ d = 1, 2, ..., P$, where P is the degree of freedom that controls the smoothness. $\phi(t_n)$ is polynomial or spline basis. Notice here the scale of $\boldsymbol{\beta}_{ij} = \{\beta_{ijd}\}$ does not matter, as we could subtract same intercept from the exponential in both numerator and denominator to change in the scale. However, in practice we put a prior $\mathcal{N}(\boldsymbol{\beta}_{ij}|0, \sigma_0^2\mathbf{I})$ on β_{ij} to regularise the search space of the gradient descent optimization described below. Here we choose a non-informative prior with large variance, $\sigma_0^2 = 100$.

To maximise the evidence lower bound, we notice that $\ln(\cdot)$ is a concave function and by Taylor expansion:

$$\ln(\sum_{j=1}^{D} \exp(\boldsymbol{\beta}_{ij}^T \phi(t_{sn}))) \leq \ln \zeta + \zeta^{-1}(\sum_{j=1}^{D} \exp(\boldsymbol{\beta}_{ij}^T \phi(t_{sn})) - \zeta).$$

Therefore, by introducing a variational variable ζ, we find following lower bound of the ELBO function $\mathcal{L}$ with respect to $\boldsymbol{\beta}_{ij}$:

$$\begin{aligned}
\mathcal{L}_{[\beta]} &= \sum_{s=1}^{M}\sum_{n=1}^{N_s} \mathbf{E}_q\{\ln p(w_n|z_n, \beta(t_{sn}))\} \\
&= \sum_{s=1}^{M}\sum_{n=1}^{N_s}\sum_{i=1}^{K}\sum_{j=1}^{D} \Big(\boldsymbol{\beta}_{ij}^T\phi(t_{sn}) - \ln(\sum_{j'=1}^{D}\exp\{\boldsymbol{\beta}_{ij'}^T\phi(t_{sn})\})\Big)\mathbf{E}\{z_{sni}\}w_{snj} \geq \\
&\sum_{s=1}^{M}\sum_{n=1}^{N_s}\sum_{i=1}^{K}\sum_{j=1}^{D} \big(\boldsymbol{\beta}_{ij}^T\phi(t_{sn}) - \zeta_{sni}^{-1}\sum_{j'=1}^{D}\exp\{\boldsymbol{\beta}_{ij'}^T\phi(t_{sn})\} - \ln\zeta_{sni} + 1\big)\mathbf{E}\{z_{sni}\}w_{snj}.
\end{aligned} \tag{11}$$

We could then apply a method called local variational inference to maximise the right hand side of equation 11. We do this by updating $\boldsymbol{\beta}$ and ζ in turn. Take derivative with respect to ζ_{sni}, we obtained following update:

$$\zeta_{sni} = \sum_{j=1}^{D}\exp\{\boldsymbol{\beta}_{ij}^T\phi(t_{sn})\} \tag{12}$$

In order to update the lower bound with respect to $\boldsymbol{\beta}$, we separate the terms containing $\boldsymbol{\beta}_{ij}$:

$$\mathcal{L}_{[\beta_{ij}]} = \sum_{s=1}^{M}\sum_{n=1}^{N_s}\mathbf{E}\{z_{sni}\}w_{snj}\boldsymbol{\beta}_{ij}^T\phi(t_{sn}) - \sum_{s=1}^{M}\sum_{n=1}^{N_s}\mathbf{E}\{z_{sni}\}\zeta_{sni}^{-1}\exp\{\boldsymbol{\beta}_{ij}^T\phi(t_{sn})\}.$$

There is no analytical solution for $\boldsymbol{\beta}_{ij}$, but the lower bound is convex so we could maximize the lower bound using following gradient information:

$$\nabla_{\boldsymbol{\beta}_{ij}}\mathcal{L}_{[\beta]} = \sum_{s=1}^{M}\sum_{n=1}^{N_s}\Big(\mathbf{E}\{z_{sni}\}w_{snj} - \mathbf{E}\{z_{sni}\}\zeta_{sni}^{-1}\exp\{\boldsymbol{\beta}_{ij}^T\phi(t_{sn})\}\Big)\phi(t_{sn}) \tag{13}$$

The gradient information of $\mathcal{L}_{[\beta_{ij}]}$ allows efficient numeric estimation of $\boldsymbol{\beta}_{ij}$. However, evaluating $\mathcal{L}_{[\beta_{ij}]}$ and $\nabla_{\boldsymbol{\beta}_{ij}}\mathcal{L}_{[\beta]}$ is computational expensive due to $\exp\{\boldsymbol{\beta}_{ij}^T\phi(t_{sn})\}$, which require looping through s, n (the entire records set over all subjects!). The gradient descent methods for estimating $\boldsymbol{\beta}_{ij}$ requires evaluating$\mathcal{L}_{[\beta_{ij}]}$ and $\nabla_{\boldsymbol{\beta}_{ij}}\mathcal{L}_{[\beta]}$ at each gradient step, which prohibit scaling up the model to large data set. To solve this problem in practice we discretise t_{sn} into years which allows us to pre-compute the sum of $\mathbf{E}\{z_{sni}\}\zeta_{sni}^{-1}$ over all incidences that happened at each age year. For each new $\boldsymbol{\beta}_{ij}$, we could then sum over all years, which reuses the sums computed. This trick significant reduced the computation cost of evaluating $\mathcal{L}_{[\beta_{ij}]}$ and $\nabla_{\boldsymbol{\beta}_{ij}}\mathcal{L}_{[\beta]}$ which makes the estimation of age topics over the entire UK Biobank HES possible. In conclusion, we could update $\boldsymbol{\beta}$ using following psuedo-code:

Algorithm 1: Maximize local variationl lower bound

```
initialization;
for i ← 1 to K do
    for j ← 1 to D do
        Update ζ_sni = Σ_{j=1}^{D} exp{β_ij^T φ(t_sn)} ;
        Update β_ij to maximize L_[β_ij] ;
    end
end
```

Note here we need to update ζ_{sni} for each j, while in practice we only update ζ_{sni} once for each optimization of $\boldsymbol{\beta}$ to allow parallel computation over j.

The above computation provide a point estimate for $\boldsymbol{\beta}$, which we adopted when applying our methods to empirical data. We also provide mathematical derivation for posterior distributions of $\boldsymbol{\beta}$, though we did not present results using this method to empirical data. We use a Gaussian prior for β_{ijd} and perform full variational inference on β:

$$\boldsymbol{\beta}_{ij} \sim \mathcal{N}(0, \sigma^2\mathbf{I}).$$

Here sigma is a hyperparameter that encourages sparsity.

$$
\begin{aligned}
\mathcal{L}_{[\beta]} &= \sum_{s=1}^{M}\sum_{n=1}^{N_s}\mathbf{E}_q\{\ln p(w_n|z_n,\beta(t_{sn}))\} + \sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\Big(\mathbf{E}_q\{\ln p(\boldsymbol{\beta}_{ijd})\} - \mathbf{E}_q\{\ln q(\boldsymbol{\beta}_{ijd})\}\Big) \\
&= \sum_{s=1}^{M}\sum_{n=1}^{N_s}\sum_{i=1}^{K}\sum_{j=1}^{D}\big(\mathbf{E}_q\{\boldsymbol{\beta}_{ij}\}^T\phi(t_{sn}) - \mathbf{E}_q\{\ln(\sum_{j=1}^{D}\exp(\boldsymbol{\beta}_{ij}^T\phi(t_{sn})))\}\big)\mathbf{E}\{z_{sni}\}w_{snj} - \\
&\frac{1}{2\sigma^2}\sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\mathbf{E}_q\{\boldsymbol{\beta}_{ijd}^2\} - \sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\mathbf{E}_q\{\ln q(\boldsymbol{\beta}_{ijd})\} \geq \\
&\sum_{s=1}^{M}\sum_{n=1}^{N_s}\sum_{i=1}^{K}\sum_{j=1}^{D}\big(\mathbf{E}_q\{\boldsymbol{\beta}_{ij}\}^T\phi(t_{sn}) - \zeta_{sni}^{-1}\sum_{j=1}^{D}\mathbf{E}_q\{\exp(\boldsymbol{\beta}_{ij}^T\phi(t_{sn}))\} - \ln\zeta_{sni} + 1\big)\mathbf{E}\{z_{sni}\}w_{snj} - \\
&\frac{1}{2\sigma^2}\sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\mathbf{E}_q\{\boldsymbol{\beta}_{ijd}^2\} - \sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\mathbf{E}_q\{\ln q(\boldsymbol{\beta}_{ijd})\}.
\end{aligned}
\tag{14}
$$

Following [3], we assumed an independent variational Gaussian distribution for each β_{ijd}:

$$\beta_{ijd} \sim \mathcal{N}(\lambda_{ijd}, \nu_{ijd}^2),$$

and observe the moment-generating function of Guanssian distribution is:

$$\mathbf{E}_q\{\exp(\boldsymbol{\beta}_{ijd}\phi_d(t_{sn}))\} = \exp\Big(\phi_d(t_{sn})\lambda_{ijd} + \frac{\phi_d^2(t_{sn})\nu_{ijd}^2}{2}\Big),$$

We obtain a tractable lower bound with respect to the variational parameters $\{\zeta_{sni}, \lambda_{ijd}, \nu_{ijd}^2\}$:

$$
\begin{aligned}
\mathcal{L}_{\zeta,\lambda,\nu^2} = \sum_{s=1}^{M}\sum_{n=1}^{N_s}\sum_{i=1}^{K}\sum_{j=1}^{D}\big(\lambda_{ij}^T\phi(t_{sn}) - \zeta_{sni}^{-1}\sum_{j=1}^{D}\exp\{\sum_{d=1}^{P}\Big(\phi_d(t_{sn})\lambda_{ijd} + \frac{\phi_d^2(t_{sn})\nu_{ijd}^2}{2}\Big)\} - \\
\ln\zeta_{sni} + 1\big)\cdot\mathbf{E}\{z_{sni}\}w_{snj} - \sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{d=1}^{P}\big(\frac{1}{2\sigma^2}\nu_{ijd}^2 + \frac{1}{2}\ln\nu_{ijd}^2\big).
\end{aligned}
\tag{15}
$$

4 Comparison of collapsed variational inference and mean field variational inference

A vast number of inference methods have been developed for models based on original Latent Dirichlet Allocation. The most prominet of which are collapsed gibbs sampling and mean field variational inference. For inference of model with exchangeable variables

using extremely large and noisy data set, it is desirable to have a deterministic method such as variational inference. Collapsed variational inference makes less assumptions for approximation, therefore the inferred distributions are strictly closer to the true posterior distributions than the mean-field variational Bayesian methods. We will explain why accuracy would be importance for the data we are considering in section 4.2.

4.1 Mean field variational inference to estimate patient-level posterior distribution $q(z, \theta)$

Please note this section is just a replication of [4] using our notation, which is provided to make the note self-contained. We assume variational distributions for latent variables θ and $\mathbf{z}$ are independent of each other, then we could get the variational lower bound for the log likelihood of a single subject:

$$\begin{aligned} &q(\mathbf{z}, \theta) = q(\mathbf{z})q(\theta), \\ &\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha) = \mathbf{E}_q\{\ln p(\mathbf{w}, \mathbf{z}, \theta|\alpha, \beta) - \ln q(\mathbf{z}, \theta)\}. \end{aligned} \tag{16}$$

It is straightforward to estimate $q(\mathbf{z})$ and $q(\theta)$ that maximise the lower bound $\mathcal{L}(\mathbf{z}, \theta, \beta, \alpha)$:

$$\begin{aligned} \ln q^*(\theta) &= \ln p(\theta|\alpha) + \mathbf{E}_{q(\mathbf{z})}\{\sum_{n=1}^{N_s} \ln p(z_n|\theta)\} + const \\ &= \sum_{i=1}^{K}(\alpha_i + \sum_{n=1}^{N_s} \mathbf{E}\{z_{ni}\} - 1)\ln\theta_i + const, \\ \ln q^*(\mathbf{z}) &= \mathbf{E}_{q(\theta)}\{\sum_{n=1}^{N_s} \ln p(z_n|\theta)\} + \sum_{n=1}^{N_s} \ln p(w_n|z_n, \beta) + const \\ &= \sum_{n=1}^{N_s}\sum_{i=1}^{K} z_{ni}\Big(\mathbf{E}\{\ln\theta_i\} + \sum_{j=1}^{D} w_{nj}\ln\beta_{ij}\Big) + const. \end{aligned} \tag{17}$$

We see that $q(\theta)$ factorises over i and $q(\mathbf{z})$ factorises over n, i. Therefore, we get the variational distribution for $\mathbf{z}$ and θ:

$$\theta_i \sim \mathbf{Dir}(\alpha_i + \sum_{n=1}^{N_s} \mathbf{E}\{z_{ni}\})$$

$$z_{ni} \sim \mathbf{Cat}\Big(\frac{\exp\big(\mathbf{E}\{\ln\theta_i\} + \sum_{j=1}^{D} w_{nj}\ln\beta_{ij}\big)}{\sum_{i=1}^{K}\exp\big(\mathbf{E}\{\ln\theta_i\} + \sum_{j=1}^{D} w_{nj}\ln\beta_{ij}\big)}\Big)$$

We then has the $(m+1)^{th}$ E-step as follows:

$$\begin{aligned}\mathbf{E}^{m+1}\{\ln\theta_i\} &= \Psi(\alpha_i + \sum_{n=1}^{N_s}\mathbf{E}^m\{z_{ni}\}) - \Psi(\sum_{i=1}^{K}\big(\alpha_i + \sum_{n=1}^{N_s}\mathbf{E}^m\{z_{ni}\}\big)),\\ \mathbf{E}^{m+1}\{z_{ni}\} &= \frac{\exp\big(\mathbf{E}^{m+1}\{\ln\theta_i\} + \sum_{j=1}^{D} w_{nj}\ln\beta_{ij}^m\big)}{\sum_{i=1}^{K}\exp\big(\mathbf{E}^{m+1}\{\ln\theta_i\} + \sum_{j=1}^{D} w_{nj}\ln\beta_{ij}^m\big)},\end{aligned} \tag{18}$$

where $\mathbf{E}^m$ and β^m refers to the estimation of previous step (m^{th} step); Ψ is the digamma function.

To perform the M-step, we maximize the lower bound $\mathcal{L}$ in equation 2 for the entire population.

$$\begin{aligned}\mathcal{L}(\mathbf{z},\theta,\beta,\alpha) = &\sum_{s=1}^{M}\Big(\ln\Gamma(\sum_{i=1}^{K}\alpha_i) - \sum_{i}^{K}\ln\Gamma(\alpha_i) + \sum_{i=1}^{K}(\alpha_i-1)\mathbf{E}\{\ln\theta_{si}\} +\\ &\sum_{n=1}^{N_s}\sum_{i=1}^{K}(\mathbf{E}\{z_{sni}\}\mathbf{E}\{\ln\theta_{si}\}) +\\ &\sum_{n=1}^{N_s}\sum_{i=1}^{K}\mathbf{E}\{z_{sni}\}\sum_{j=1}^{D}w_{snj}\ln\beta_{ij}\Big) -\\ &\sum_{s=1}^{M}\Big(\ln\Gamma(\sum_{i=1}^{K}\big(\alpha_i + \sum_{n=1}^{N_s}\mathbf{E}\{z_{ni}\}\big)) - \sum_{i=1}^{K}\ln\Gamma(\alpha_i + \sum_{n=1}^{N_s}\mathbf{E}\{z_{ni}\}) +\\ &\sum_{i=1}^{K}(\alpha_i + \sum_{n=1}^{N_s}\mathbf{E}\{z_{ni}\} - 1)\mathbf{E}\{\ln\theta_{si}\} +\\ &\sum_{n=1}^{N_s}\sum_{i=1}^{K}\mathbf{E}\{z_{ni}\}\ln\mathbf{E}\{z_{ni}\}\Big)\end{aligned} \tag{19}$$

For β, we take terms in $\mathcal{L}$ and add Lagrange multipliers:

$$\mathcal{L}_{[\beta]} = \sum_{i=1}^{K}\sum_{j=1}^{D}\ln\beta_{ij}\sum_{s=1}^{M}\sum_{n=1}^{N_s}\mathbf{E}\{z_{sni}\}w_{snj} + \sum_{i=1}^{K}\lambda_i(\sum_{j=1}^{D}\beta_{ij} - 1).$$

Set the derivative of $\mathcal{L}_{[\beta]}$ with respect β to zero, we could get the $(n+1)^{th}$ update for *beta*:

$$\beta_{ij}^{n+1} = \frac{\sum_{s=1}^{M}\sum_{n=1}^{N_s}\mathbf{E}^{n+1}\{z_{sni}\}w_{snj}}{\sum_{j=1}^{D}\sum_{s=1}^{M}\sum_{n=1}^{N_s}\mathbf{E}^{n+1}\{z_{sni}\}w_{snj}}$$

The terms in lower bound that contains α are:

$$\mathcal{L}_{[\alpha_i]} = \sum_{s=1}^{M}\Big(\ln\Gamma(\sum_{i=1}^{K}\alpha_i) - \ln\Gamma(\alpha_i) + \sum_{i=1}^{K}(\alpha_i-1)\mathbf{E}^{n+1}\{\ln\theta_{si}\}\Big).$$

Take the derivatives with respect to α:

$$\frac{\partial \mathcal{L}_{[\alpha]}}{\partial \alpha_i} = M \cdot \Big(\Psi(\sum_{i=1}^{K} \alpha_i) - \Psi(\alpha_i)\Big) + \sum_{s=1}^{M} \mathbf{E}^{n+1}\{\ln \theta_{si}\}.$$

And the Hessian:

$$\nabla^2_{\alpha} \mathcal{L}_{[\alpha]} = M \cdot \mathrm{diag}(-\Psi^1(\alpha_i)\Big) + M \cdot \Psi^1(\sum_{i=1}^{K} \alpha_i),$$

where Ψ^1 is the Trigamma function. We use the Newton-Raphson method the find the maximal of α as described in [4]. In practice, we used $\alpha = 1$ to put an uninformative prior robust optimization.

4.2 Patient with a few diseases versus documents with many words

In section 3.1, we briefly explained why we chose to use collapsed variational inference over a simpler mean-filed variational inference method. We will focus on the difference between equation 17 and equation 5. For the mean-field variational distribution:

$$\ln q^*(\mathbf{z}) = \sum_{n=1}^{N_s} \sum_{i=1}^{K} z_{ni} \Big(\mathbf{E}\{\ln \theta_i\} + \sum_{j=1}^{D} w_{nj} \ln \beta_{ij}\Big) + const,$$

which factorised over each of the N diagnosis. Therefore, the inferred distribution for each z_n is conditional i.i.d.

$$q(z_{n'} | \mathbf{z}_{\neg \mathbf{n}'}, \mathbf{w}, \alpha, \beta, \theta) = q(z_{n'} | \mathbf{w}, \alpha, \beta, \theta),$$

Here $\neg n'$ refer to indices of all diagnoses excluding n'. However, for collapsed VB, conditional distribution depends on other diagnoses of the same patient:

$$q(z_{n'} | \mathbf{z}_{\neg \mathbf{n}'}, \mathbf{w}, \alpha, \beta) \propto \prod_{i}^{K} (\alpha_i + \sum_{n \in \neg n'} z_{ni})^{z_{n'i}} \prod_{i=1}^{K} \prod_{j=1}^{D} \beta_{ij}^{z_{n'i} w_{n'j}}$$

The impact of the the dependency on the accuracy of posterior approximation depends on the data structure. Most of topic modelling modes were designed for text modelling, where each document have a large word number N_s. In this case, $(\alpha_i + \sum_{n \in \neg n'} z_{ni})$ will be approximately the same across n':

$$\lim_{N_s \to \infty} p(z_{n'} | \mathbf{z}_{\neg \mathbf{n}'}, \mathbf{w}, \alpha, \beta) \propto \prod_{i}^{K} \Big[(\alpha_i + N_s \theta_i) \prod_{j=1}^{D} \beta_{ij}^{w_{n'j}}\Big]^{z_{n'i}},$$

where θ_i is the topic value for the s^{th} document. We see $\mathbf{z}_{\neg \mathbf{n}'}$ no longer exist and $q^*(\mathbf{z})$ in equation 17 could approximate $p(\mathbf{z}|\mathbf{w}, \alpha, \beta)$ accurately. However, each patient on average have 6.1 distinct diagnoses in UK Biobank HES, making N_s small for the mean-field approximation. Note, we do not need to assume independence of z_n, it is a consequence of assume independence between $q(\theta)$ and $q(\mathbf{z})$, which is called induced factorisation in some cases. (section 10.2.5 in [2]) In this cases collapsed VB models the dependency between $\mathbf{z}_n$ and $\mathbf{z}_{\neg \mathbf{n}'}$ and have better accuracy at approximating posterior distribution.

References

[1] Teh, Y. W., Newman, D. & Welling, M. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. Tech. Rep., CALIFORNIA UNIV IRVINE SCHOOL OF INFORMATION AND COMPUTER SCIENCE (2007).

[2] Bishop, C. M. & Nasrabadi, N. M. *Pattern recognition and machine learning* (Springer, 2006).

[3] Blei, D. M., Lafferty, J. D. *et al.* A correlated topic model of science. *The annals of applied statistics* **1**, 17–35 (2007).

[4] Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research* **3**, 993–1022 (2003).