**Background on Topic Modelling**

Topic modelling is a probabilistic approach to identifying clusters of related words within a corpus of documents. Documents can be any textual data (for e.g., U.S. Congressional Record speeches1; tweets2; and electronic medical records3) making topic modelling a flexible tool in reducing dimensionality. While there are multiple approaches to topic modelling (see for Barde & Bainwad, 2017, for overview)4, the approach taken in this paper employs term-frequency inverse-document-frequency indices to discover sets of terms that efficiently identify latent structure (i.e. topics) within a corpus of documents. Topic modelling uses a generative procedure called Latent Dirichlet Allocation (LDA) that iteratively tunes probabilities that candidate topics are associated with terms within the observed data.5,6 Due to this underlying generative procedure topic modelling is particularly well-suited to highly sparse discrete data; a common data format within the epidemiology of psychiatric disorders in which co-morbidity is rife, symptoms are not always indicative, and where it can be difficult to collect data in sufficient quantities due to low population incidence.

1 Quinn KM, Monroe BL, Colaresi M, Crespin MH, Radev DR. How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science* 2010; **54**: 209–228.

2 Lau J, Collier N, Baldwin T. On-line trend analysis with topic models: #twitter trends detection topic model online. *Proceedings of COLING 2012, C12-1093* 2012; : 1519–1534.

3 McCoy TH, Castro VM, Snapper L, Hart K, Januzzi JL, Huffman JC *et al.* Polygenic loading for major depression is associated with specific medical comorbidity. *Transl Psychiatry* 2017; **7**: e1238–e1238.

4 Barde BV, Bainwad AM. An overview of topic modeling methods and tools. In: *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*. IEEE: Madurai, 2017, pp 745–750.

5 Griffiths TL, Steyvers M. Finding scientific topics. *Proceedings of the National Academy of Sciences* 2004; **101**: 5228–5235.

6 Blei D, Ng A, Jordan M. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 2001; **3**: 601–608.

**Supplementary Figure 1.** Pseudo-EMRs constructed in preparation for topic modelling by latent Dirichlet allocation

Numbers 1-6 indicate unique subjects with ID masked. Observed psychiatric and sociodemographic features have been transformed into a pseudo-EMR for each subject.

**Supplementary Figure 2.** Evaluation of various model diagnostics in relation to selecting the optimal number of topics to model given the observed data

Top left: Held-out likelihood indicates perplexity (higher is better). Top right: Lower bound indicates model convergence (higher is better). Bottom left: Residuals indicate model saturation (lower is better). Bottom right: Semantic coherence indicates co-occurrence of probable terms per topic (higher is better).



**Supplementary Table 1.** Descriptive Table of Counts and Percentages for Clinical Features included in Topic Modeling

|  |  |  |
| --- | --- | --- |
| Clinical label | *n* | % |
| high-tactile | 450 | 63.6 |
| low-acoustic | 446 | 63.1 |
| handed-right | 427 | 60.4 |
| abnorm-sensory | 401 | 56.7 |
| stereotypies | 390 | 55.2 |
| low-pain | 350 | 49.5 |
| low-visual | 350 | 49.5 |
| hearing-behavioural | 330 | 46.7 |
| diet-preferences | 325 | 46.0 |
| symptoms-ocd | 299 | 42.3 |
| symptoms-odd | 275 | 38.9 |
| symptoms-sleep | 273 | 38.6 |
| current-alcohol | 269 | 38.0 |
| past-tobacco | 269 | 38.0 |
| past-alcohol | 268 | 37.9 |
| current-drugs | 268 | 37.9 |
| past-drugs | 268 | 37.9 |
| current-tobacco | 267 | 37.8 |
| symptoms-adhd | 264 | 37.3 |
| delay-bladder-nocte | 263 | 37.2 |
| delay-sit | 242 | 34.2 |
| diet-picky | 225 | 31.8 |
| abnorm-gastro | 213 | 30.1 |
| allergies | 200 | 28.3 |
| abnorm-coordination | 198 | 28.0 |
| symptoms-anxiety | 189 | 26.7 |
| sleep-insomnia | 187 | 26.4 |
| hearing-assessment | 185 | 26.2 |
| diet-carbohydrates | 172 | 24.3 |
| diet-texture | 168 | 23.8 |
| sleep-interrupted | 164 | 23.2 |
| abnorm-skin | 161 | 22.8 |
| delay-phrases | 152 | 21.5 |
| delay-bowel-diem | 145 | 20.5 |
| delay-bladder-diem | 141 | 19.9 |
| test-karyotype | 134 | 19.0 |
| delay-bowel-nocte | 134 | 19.0 |
| test-other | 131 | 18.5 |
| abnorm-gait | 127 | 18.0 |
| past-diet | 121 | 17.1 |
| diagnosis-adhd | 115 | 16.3 |
| test-metabolic | 103 | 14.6 |
| delay-walk | 102 | 14.4 |
| diet-pica | 100 | 14.1 |
| delay-words | 90 | 12.7 |
| hearing-baer | 89 | 12.6 |
| gluten-free | 85 | 12.0 |
| symptoms-depression | 84 | 11.9 |
| asthma | 81 | 11.5 |
| diet-salt | 81 | 11.5 |
| diet-sugar | 79 | 11.2 |
| abnorm-ear-pe | 75 | 10.6 |
| vision | 73 | 10.3 |
| abnorm-movement | 69 | 9.8 |
| casein-free | 69 | 9.8 |
| abnorm-teeth | 68 | 9.6 |
| symptoms-bpd | 65 | 9.2 |
| abnorm-respiratory | 63 | 8.9 |
| abnorm-growth | 59 | 8.3 |
| seizures-other | 58 | 8.2 |
| past-hospital | 55 | 7.8 |
| delay-roll | 54 | 7.6 |
| dairy-free | 48 | 6.8 |
| abnorm-ear | 45 | 6.4 |
| diagnosis-odd | 44 | 6.2 |
| infant-irritable | 42 | 5.9 |
| sleep-terrors | 42 | 5.9 |
| abnorm-genital | 41 | 5.8 |
| infant-floppy | 40 | 5.7 |
| abnorm-eye | 40 | 5.7 |
| abnorm-joint | 39 | 5.5 |
| hearing | 32 | 4.5 |
| abnorm-bone | 31 | 4.4 |
| sleep-fidgeting | 30 | 4.2 |
| diagnosis-anxiety | 29 | 4.1 |
| diagnosis-ocd | 28 | 4.0 |
| abnorm-craniofacial | 26 | 3.7 |
| abnorm-heart | 25 | 3.5 |
| seizures-febrile | 24 | 3.4 |
| diagnosis-sleep | 21 | 3.0 |
| diagnosis-depression | 20 | 2.8 |
| abnorm-renal | 19 | 2.7 |
| infant-stiff | 16 | 2.3 |
| abnorm-cerebral | 15 | 2.1 |
| infant-lethargic | 10 | 1.4 |
| seizures-intractable | 10 | 1.4 |
| endocrine-growth | 9 | 1.3 |
| endocrine-sex | 9 | 1.3 |
| diagnosis-bmd | 9 | 1.3 |
| endocrine-thyroid | 8 | 1.1 |
| test-imaging | 5 | 0.7 |
| endocrine-diabetes | 3 | 0.4 |
| egg-free | 2 | 0.3 |

*Note.* Each count represents the presence of the token psychiatric feature for a single subject.

**Supplementary Table 2.** Association between all topics and polygenic risk score (PRS) tranches of autism spectrum disorder (ASD). Seven tranches of SNPs (1×10−2, 1×10−3, 1×10−4, 1×10−5, 1×10−6, 1×10−7, 5×10−8) are labelled as S2–S8.

|  |  |
| --- | --- |
| **Trait** | **ASD PRS Trenches** |
| **S2** | **S3** | **S4** | **S5** |
| *Beta* | *SE* | *P* | *Beta* | *SE* | *P* | *Beta* | *SE* | *P* | *Beta* | *SE* | *P* |
| *Topic 1* | 172.233 | 485.528 | 0.723 | -121.616 | 1217.723 | 0.920 | 260.379 | 2427.650 | 0.915 | 5354.636 | 4734.450 | 0.259 |
| *Topic 2* | 516.265 | 481.227 | 0.284 | 1895.451 | 1205.217 | 0.116 | 176.934 | 2408.610 | 0.941 | 2002.122 | 4702.410 | 0.670 |
| *Topic 3* | -107.757 | 488.015 | 0.825 | 879.714 | 1223.251 | 0.472 | 1230.124 | 2439.311 | 0.614 | -3009.802 | 4762.496 | 0.528 |
| *Topic 4* | 153.631 | 485.456 | 0.752 | -415.478 | 1217.383 | 0.733 | 2721.559 | 2424.222 | 0.262 | 1359.814 | 4739.257 | 0.774 |
| *Topic 5* | -108.248 | 485.983 | 0.824 | 167.355 | 1218.762 | 0.891 | -195.763 | 2429.755 | 0.936 | -2420.770 | 4743.325 | 0.610 |
| *Topic 6* | -926.047 | 485.405 | 0.057 | -2571.968 | 1216.277 | 0.035 | -5953.031 | 2421.017 | 0.014 | -14839.312 | 4709.537 | 0.002 |
| *Topic 7* | -827.747 | 487.130 | 0.090 | -806.447 | 1224.575 | 0.510 | -1012.808 | 2441.949 | 0.678 | -3383.638 | 4766.761 | 0.478 |
| *Topic 8* | -92.369 | 486.588 | 0.850 | 510.458 | 1220.074 | 0.676 | -1420.269 | 2431.939 | 0.559 | -3337.325 | 4748.066 | 0.482 |
| *Topic 9* | -608.856 | 483.374 | 0.208 | 274.520 | 1214.030 | 0.821 | -61.888 | 2420.412 | 0.980 | 3187.722 | 4724.145 | 0.500 |
| *Topic 10* | 451.947 | 484.661 | 0.352 | 977.912 | 1215.678 | 0.422 | 1209.988 | 2424.534 | 0.618 | -1775.132 | 4734.838 | 0.708 |
| *Topic 11* | -2.234 | 486.422 | 0.996 | -1768.340 | 1217.275 | 0.147 | -1271.892 | 2431.184 | 0.601 | -3994.960 | 4745.261 | 0.400 |
| *Topic 12* | 619.120 | 479.903 | 0.198 | 1045.106 | 1204.566 | 0.386 | -629.205 | 2403.057 | 0.794 | -809.413 | 4692.571 | 0.863 |
| *Topic 13* | 380.204 | 487.377 | 0.436 | 494.217 | 1222.760 | 0.686 | 2168.882 | 2436.173 | 0.374 | 4604.506 | 4756.373 | 0.333 |
| *Topic 14* | -261.684 | 482.938 | 0.588 | 894.445 | 1210.784 | 0.460 | 257.320 | 2415.102 | 0.915 | 8586.476 | 4700.397 | 0.068 |
| *Topic 15* | -76.892 | 487.349 | 0.875 | -864.137 | 1221.572 | 0.480 | -3091.688 | 2432.635 | 0.204 | -3470.810 | 4755.250 | 0.466 |
| *Topic 16* | 541.183 | 485.491 | 0.266 | 198.051 | 1218.961 | 0.871 | 3239.184 | 2425.890 | 0.182 | 3869.329 | 4742.226 | 0.415 |
| *Topic 17* | -61.643 | 487.211 | 0.899 | -2462.132 | 1216.881 | 0.044 | -5080.260 | 2425.271 | 0.037 | -6614.400 | 4747.226 | 0.164 |
| *Topic 18* | 250.365 | 485.859 | 0.607 | 259.092 | 1218.681 | 0.832 | 1084.662 | 2429.191 | 0.655 | -2164.388 | 4743.379 | 0.648 |

|  |  |
| --- | --- |
| **Trait** | **ASD PRS Trenches** |
| **S6** | **S7** | **S8** |
| *Beta* | *SE* | *P* | *Beta* | *SE* | *P* | *Beta* | *SE* | *P* |
| *Topic 1* | 616.098 | 8588.841 | 0.943 | -8908.612 | 14966.803 | 0.552 | -8908.612 | 14966.803 | 0.552 |
| *Topic 2* | 7072.687 | 8515.631 | 0.407 | 8866.437 | 14849.292 | 0.551 | 8866.437 | 14849.292 | 0.551 |
| *Topic 3* | -1797.712 | 8631.851 | 0.835 | 10043.452 | 15040.954 | 0.505 | 10043.452 | 15040.954 | 0.505 |
| *Topic 4* | -824.277 | 8587.318 | 0.924 | 11642.097 | 14960.475 | 0.437 | 11642.097 | 14960.475 | 0.437 |
| *Topic 5* | -29.585 | 8596.289 | 0.997 | -3259.756 | 14984.274 | 0.828 | -3259.756 | 14984.274 | 0.828 |
| *Topic 6* | -23831.077 | 8550.878 | 0.006 | -26259.611 | 14974.760 | 0.080 | -26259.611 | 14974.760 | 0.080 |
| *Topic 7* | -404.406 | 8640.832 | 0.963 | 21910.061 | 15030.946 | 0.146 | 21910.061 | 15030.946 | 0.146 |
| *Topic 8* | -7775.528 | 8599.891 | 0.366 | -23380.881 | 14967.173 | 0.119 | -23380.881 | 14967.173 | 0.119 |
| *Topic 9* | 12285.892 | 8545.646 | 0.151 | 18464.483 | 14904.552 | 0.216 | 18464.483 | 14904.552 | 0.216 |
| *Topic 10* | -1103.284 | 8579.744 | 0.898 | -6761.168 | 14953.346 | 0.651 | -6761.168 | 14953.346 | 0.651 |
| *Topic 11* | -4473.570 | 8601.315 | 0.603 | -22476.810 | 14964.250 | 0.134 | -22476.810 | 14964.250 | 0.134 |
| *Topic 12* | -6112.541 | 8497.988 | 0.472 | -9016.880 | 14815.785 | 0.543 | -9016.880 | 14815.785 | 0.543 |
| *Topic 13* | 2744.836 | 8624.861 | 0.750 | -9580.928 | 15030.229 | 0.524 | -9580.928 | 15030.229 | 0.524 |
| *Topic 14* | 17820.554 | 8507.463 | 0.037 | 22493.041 | 14860.877 | 0.131 | 22493.041 | 14860.877 | 0.131 |
| *Topic 15* | -6263.427 | 8615.705 | 0.468 | 4930.336 | 15025.106 | 0.743 | 4930.336 | 15025.106 | 0.743 |
| *Topic 16* | 2369.588 | 8597.107 | 0.783 | 206.053 | 14987.535 | 0.989 | 206.053 | 14987.535 | 0.989 |
| *Topic 17* | -3244.903 | 8616.499 | 0.707 | 5410.359 | 15020.389 | 0.719 | 5410.359 | 15020.389 | 0.719 |
| *Topic 18* | -2217.438 | 8595.371 | 0.797 | 8351.233 | 14979.743 | 0.577 | 8351.233 | 14979.743 | 0.577 |

 **Table 2.** (*Cont.*)