Supplementary Material

Identifying Comorbidity Patterns in People with and without Alzheimer's Disease Using Latent Dirichlet Allocation

Study Cohort Details

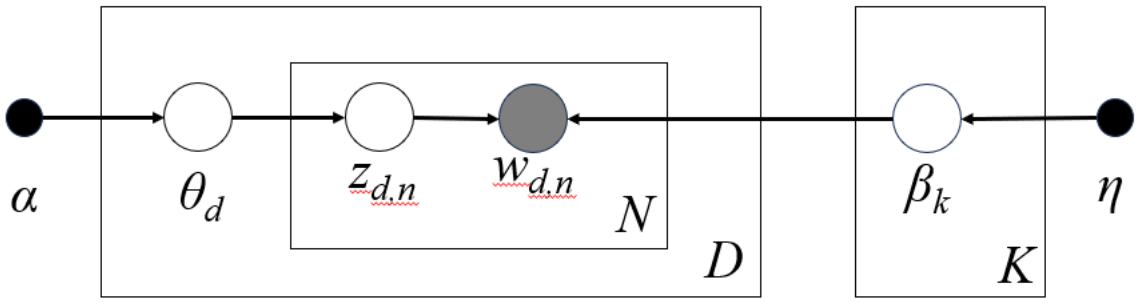
Those with a clinically verified AD diagnosis were identified from the Finnish Special Reimbursement Register (FSRR), which is maintained by the Social Insurance Institution of Finland (SII) as described in Tolppanen et al.[1] The FSRR contains records of all people who are eligible for higher reimbursement of medications due to certain chronic diseases, such as AD. For a person to be eligible for the FSRR for AD, they need a verified diagnosis of AD written in a medical statement by their physician and submitted to SII. The medical statement must include that the patient has: 1) symptoms consistent with AD, 2) a decrease in social capacity over a period ≥ 3 months, 3) neuroanatomical changes consistent with AD, confirmed by computed tomography (CT)/ magnetic resonance imaging scan (MRI) 4) possible alternative diagnoses excluded, and 5) received confirmation of the diagnosis by a registered neurologist or geriatrician. The findings from CT/MRI, laboratory tests, cognitive tests, and statements from the patient and their family need to be included. Each case is systematically reviewed by a geriatrician/ neurologist to confirm whether pre-specified criteria are met. The AD diagnosis is based mainly on the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association's (NINCDS-ADRDA)[2] and Diagnostic and Statistical Manual of Mental Disorders, 4th edition (DSM-IV)[3] criteria for Alzheimer's disease. People with AD were matched 1:1 to people without AD based on age (± 1 year), sex, and hospital district region. The median age was 81.1 years (interquartile range 76.5-85) and 65% were female in both cohorts.
Ethics committee approval or informed consent was not required according to the Finnish legislation because only de-identified, routinely collected register data was used, and the study participants were not contacted. The MEDALZ study protocol was approved by the register maintainers (Statistics Finland, SII, and National Institute of Health and Welfare).

People were excluded from the control (CTRL) cohort if they did not have any ICD-10 codes reported (Fig. 1, main document). Codes for AD or dementia (F00, F01, F02, F03, and G30) were not included in the analysis (Supplementary Table 2). We also excluded "Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified" (R codes) and "factors

influencing health status and contact with health services" (Z codes) which may be recorded prior to a more appropriate diagnosis. "R" codes are unspecified symptoms and in the later clinical process may be diagnosed as a disease. An example of an "R" code is R51 *Headache* which could be a stress reaction or in the diagnostic process as a sign of serious intracerebral process. "Z" codes are used for encounters with health services for circumstances other than disease, injury, or external causes, like prophylactic vaccinations or health screenings. An example of a "Z" code is Z12 *Encounter for screening for malignant neoplasms* which may be documented when a routine check for skin cancer is done. Additional diagnoses would be recorded if any abnormalities were documented.

Latent Dirichlet Allocation Model

Latent Dirichlet Allocation (LDA) is a probabilistic graphical model for processing text documents, using Bayesian statistics.[4,5] The purpose is to group a predefined number of K topics over a collection of words. It is assumed that the K topics are drawn from a Dirichlet distribution, $\beta_k \sim$ Dirichlet(η). The topics then defines a multinomial distribution over the collection of words. LDA assumes that for each document d, the topics are generated by first drawing the topics over the distribution $\theta_d \sim$ Dirichlet(α) , from which each word i in d a topic weight is drawn $z_{di} \sim \theta_d$ where $z_{di} \in \{1,\ldots,K\}$ are topic indices. Finally, the observed word w_{di} is drawn from a selected topic $w_{di} \sim \beta_{z_{di}}$ (see Supplementary Figure 1). Because of the observational nature of our study, we are assuming equiprobable priors for both words in topics $\vec{\beta}$ and topics $\vec{\theta}$.



Supplementary Figure 1. Probabilistic graphical model representation of LDA. Each document $d \in D$ is assumed to be generated by the distribution θ_d $Dirichlet(\alpha)$, and each K topic is assumed to be drawn from β_k $Dirichlet(\eta)$. The topics defines a multinomial distribution over the collection of words, from the sample documents D.

By modelling the word-topic assignments with LDA, the resulting posterior distribution reveals a latent structure which can be used for data exploration. The posterior distribution

cannot be computed directly, so it has to be approximated. Here we are using the Online Variational Inference approximation suggested by Hoffman et al.[5]

Full derivation of the Online Variational Bayes Inference can be found in Hoffmann et al.[5] The goal is to approximate the posterior distribution by using a simpler distribution $q(\vec{z}, \vec{\theta}, \vec{\beta})$ by optimizing the Evidence Lower Bound (ELBO) criterion

$$logp(\vec{w}|\alpha,\eta) \geq L(\vec{w},\vec{\psi},\vec{\gamma},\vec{\lambda}) \equiv E_q[logp(\vec{w},\vec{z},\vec{\theta},\vec{\beta}|\alpha,\eta)] - E_q[logq(\vec{z},\vec{\theta},\vec{\beta})], \quad (1)$$

where maximizing ELBO is the same as minimizing the Kullback-Leibler divergence between q($\vec{z}$, $\vec{\theta}$, $\vec{\beta}$) and the posterior p($\vec{z}$, $\vec{\theta}$, $\vec{\beta}$ | $\vec{w}$, α, η). Hoffman's implementation follows the same factorized distribution of the distribution q() as in Blei et al. [6], that is,

$$q(z_{di} = k) = \phi_{dw_{di}k}; q(\theta_d) = \text{Dirichlet}(\theta_d; \gamma_d); q(\beta_k) = \text{Dirichlet}(\beta_k; \lambda_k).$$

The variational objective function relies only how many times n_{wd} a particular word *w* appears in a document *d*, which allows the following variational inference so summarize word counts:

$$L(\vec{w},\vec{\phi},\vec{\gamma},\vec{\lambda}) = \sum_d \sum_w n_{dw} \sum_k \phi_{dwk}(E_q[log\theta_{dk}] + E_q[log\beta_{kw}] - log\phi_{dwk}) \\ - log\Gamma\left(\sum_k \gamma_{dk}\right) + \sum_k (\alpha - \gamma_{dk})E[log\theta_{dk}] + log\Gamma(\gamma_{dk}) + \\ , \quad (3)$$

where *W* is the total collection of words and *D* is the total collection of words. Equation (3) is then solved using coordinate ascent over the parameters ϕ, Γ, λ:

$$\phi_{dwk} \propto expE[log\theta_{dk}] + E_q[log\beta_{kw}]; \gamma_{dk} = \alpha + \sum_w n_{dw}\phi_{dwk}; \lambda_{kw} \\ = \eta + \sum_d n_{dw}\phi_{dwk} \quad (4)$$

where

$$E_q[log\theta_{dk}] = \Psi(\gamma_{dk}) - \Psi\left(\sum_{i=0}^{K} \gamma_{di}\right); E_q[log\beta_{kw}] = \Psi(\lambda_{kw}) - \Psi\left(\sum_{i=0}^{W} \lambda_{ki}\right), \quad (5)$$

Ψ denotes the first derivative of the logarithm in the gamma function.

If the variational inference (3) is solved using batches, it would require to pass the entire collection of words on each iteration. The online variational inference solves for $\vec{\lambda}$ (by keeping it fixed) over the topic distributions $\vec{\beta}$,

$$L(\vec{n}, \vec{\lambda}) \equiv \sum_{d} l(n_d, \gamma(n_d, \vec{\lambda}), \phi(n_d, \vec{\lambda}), \vec{\lambda}) \quad (6)$$

The goal is to maximize the *d*th documents $l(n_d, \gamma(n_d, \vec{\lambda}), \phi(n_d, \vec{\lambda}), \vec{\lambda})$contribution in the variational inference (3). The algorithm for the online variational inference can be found in Hoffman et al.[5]

The implementation used in the study was from Scikit-learn.[7] The number of maximum iterations and learning offset (controlling the speed of early iterations in the variational inference) were selected by trial-and-error. All other hyperparameters were left as default in the Scikit-Learn model definition. The selection of the number of topics is discussed in the Results section. The Bag of Words representation was used, by counting the occurrences of words each document.

Convergence analysis can be found in Hoffman et al. [5], which we will not consider here. Because the nature of the study is qualitative, finding an optimal solution (that is, getting the model to converge) is not appropriate here. In qualitative work, the solutions are obtained by the joint work of experts working on the problem.

REFERENCES

1. Tolppanen AM, Taipale H, Koponen M, et al. Cohort profile: the Finnish Medication and Alzheimer's disease (MEDALZ) study. *BMJ Open* 2016; 6: 012100.
2. McKhann G, Drachman D, Folstein M, et al. Clinical diagnosis of Alzheimer's disease: *Neurology* 1984; 34: 939–944.
3. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, 4th ed. (DSM-IV)*. American Psychiatric Publishing, https://ajp.psychiatryonline.org/doi/10.1176/ajp.152.8.1228 (1994, accessed 14 November 2018).
4. Hoffman MD, Blei DM, Wang C, et al. Stochastic variational inference. *J Mach Learn Res* 2013; 14: 1303–1347.
5. Hoffman M, Bach F and Blei D. Online Learning for Latent Dirichlet Allocation. In: *Advances in Neural Information Processing Systems*. Curran Associates, Inc., https://papers.nips.cc/paper_files/paper/2010/hash/71f6278d140af599e06ad9bf1ba03cb0-Abstract.html (2010, accessed 8 September 2023).

6. Blei DM, Ng AY and Jordan MI. Latent Dirichlet Allocation. *J Mach Learn Res* 2003; 3: 993–1022.
7. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python.

Supplementary Table 1. List of ICD-10 blocks used in study

A00-A09	Intestinal infectious diseases
A15-A19	Tuberculosis
A20-A28	Certain zoonotic bacterial diseases
A30-A49	Other bacterial diseases
A50-A64	Infections with a predominantly sexual mode of transmission
A65-A69	Other spirochetal diseases
A70-A74	Other diseases caused by chlamydia
A75-A79	Rickettsioses
A80-A89	Viral infections of the central nervous system
A90-A99	Arthropod-borne viral fevers and viral hemorrhagic fevers
B00-B09	Viral infections characterized by skin and mucous membrane lesions
B15-B19	Viral hepatitis
B20-B24	Human immunodeficiency virus [HIV] disease
B25-B34	Other viral diseases
B35-B49	Mycoses
B50-B64	Protozoal diseases
B65-B83	Helminthiases
B85-B89	Pediculosis, ascariasis and other infestations
B90-B94	Sequelae of infectious and parasitic diseases
B95-B98	Bacterial, viral and other infectious agents
B99-B99	Other infectious diseases (updated)
C00-C14	Malignant neoplasm of lip, oral cavity and pharynx
C15-C26	Malignant neoplasm of digestive organs
C30-C39	Malignant neoplasm of respiratory and intrathoracic organs
C40-C41	Malignant neoplasm of bone and articular cartilage
C43-C44	Melanoma and other malignant neoplasms of skin
C45-C49	Malignant neoplasms of mesothelial and soft tissue
C50-C50	Malignant neoplasm of breast
C51-C58	Malignant neoplasm of female genital organs
C60-C63	Malignant neoplasms of male genital organs
C64-C68	Malignant neoplasm of urinary tract
C69-C72	Malignant neoplasms of eye, brain and other parts of central nervous system
C73-C75	Malignant neoplasm of thyroid and other endocrine glands
C76-C80	Malignant neoplasms of ill-defined, other secondary and unspecified sites
C81-C97	Malignant neoplasms of lymphoid, hematopoietic and related tissue
D00-D09	In situ neoplasms
D10-D36	Benign neoplasms
D37-D48	Neoplasms of uncertain or unknown behavior
D50-D53	Nutritional anemias
D55-D59	Hemolytic anemias
D60-D64	Aplastic and other anemias
D65-D69	Coagulation defects, purpura and other hemorrhagic conditions
D70-D77	Other diseases of blood and blood-forming organs
D80-D89	Certain disorders involving the immune mechanism
E00-E07	Disorders of thyroid gland
E10-E14	Diabetes mellitus
E15-E16	Other disorders of glucose regulation and pancreatic internal secretion
E20-E35	Disorders of other endocrine glands
E40-E46	Malnutrition
E50-E64	Other nutritional deficiencies
E65-E68	Obesity and other hyper alimentation
E70-E90	Metabolic disorders

F00-F09 [a]	Organic, including symptomatic, mental disorders
F10-F19	Mental and behavioral disorders due to psychoactive substance use
F20-F29	Schizophrenia, schizotypal and delusional disorders
F30-F39	Mood [affective] disorders
F40-F48	Neurotic, stress-related and somatoform disorders
F50-F59	Behavioral syndromes associated with physiological disturbances and physical factors
F60-F69	Disorders of adult personality and behavior
F70-F79	Mental retardation
F80-F89	Disorders of psychological development
F90-F98	Behavioral and emotional disorders with onset usually occurring in childhood and adolescence
F99-F99	Unspecified mental disorder
G00-G09	Inflammatory diseases of the central nervous system
G10-G13	Systemic atrophies primarily affecting the central nervous system
G20-G26	Extrapyramidal and movement disorders
G30-G32[a]	Other degenerative diseases of the nervous system
G35-G37	Demyelinating diseases of the central nervous system
G40-G47	Episodic and paroxysmal disorders
G50-G59	Nerve, nerve root and plexus disorders
G60-G64	Polyneuropathies and other disorders of the peripheral nervous system
G70-G73	Diseases of myoneural junction and muscle
G80-G83	Cerebral palsy and other paralytic syndromes
G90-G99	Other disorders of the nervous system
H00-H06	Disorders of eyelid, lacrimal system and orbit
H10-H13	Disorders of conjunctiva
H15-H22	Disorders of sclera, cornea, iris and ciliary body
H25-H28	Disorders of lens
H30-H36	Disorders of choroid and retina
H40-H42	Glaucoma
H43-H45	Disorders of vitreous body and globe
H46-H48	Disorders of optic nerve and visual pathways
H49-H52	Disorders of ocular muscles, binocular movement, accommodation and refraction
H53-H54	Visual disturbances and blindness
H55-H59	Other disorders of eye and adnexa
H60-H62	Diseases of external ear
H65-H75	Diseases of middle ear and mastoid
H80-H83	Diseases of inner ear
H90-H95	Other disorders of ear
I00-I02	Acute rheumatic fever
I05-I09	Chronic rheumatic heart diseases
I10-I15	Hypertensive diseases
I20-I25	Ischemic heart diseases
I26-I28	Pulmonary heart disease and diseases of pulmonary circulation
I30-I52	Other forms of heart disease
I60-I69	Cerebrovascular diseases
I70-I79	Diseases of arteries, arterioles and capillaries
I80-I89	Diseases of veins, lymphatic vessels and lymph nodes, not elsewhere classified
I95-I99	Other and unspecified disorders of the circulatory system
J00-J06	Acute upper respiratory infections
J09-J18	Influenza and pneumonia
J20-J22	Other acute lower respiratory infections
J30-J39	Other diseases of upper respiratory tract
J40-J47	Chronic lower respiratory diseases

J60-J70	Lung diseases due to external agents
J80-J84	Other respiratory diseases principally affecting the interstitium
J85-J86	Suppurative and necrotic conditions of lower respiratory tract
J90-J94	Other diseases of pleura
J95-J99	Other diseases of the respiratory system
K00-K14	Diseases of oral cavity, salivary glands and jaws
K20-K31	Diseases of esophagus, stomach and duodenum
K35-K38	Diseases of appendix
K40-K46	Hernia
K50-K52	Noninfective enteritis and colitis
K55-K63	Other diseases of intestines
K65-K67	Diseases of peritoneum
K70-K77	Diseases of liver
K80-K87	Disorders of gallbladder, biliary tract and pancreas
K90-K93	Other diseases of the digestive system
L00-L08	Infections of the skin and subcutaneous tissue
L10-L14	Bullous disorders
L20-L30	Dermatitis and eczema
L40-L45	Papulosquamous disorders
L50-L54	Urticaria and erythema
L55-L59	Radiation-related disorders of the skin and subcutaneous tissue
L60-L75	Disorders of skin appendages
L80-L99	Other disorders of the skin and subcutaneous tissue
M00-M25	Arthropathies
M30-M36	Systemic connective tissue disorders
M40-M54	Dorsopathies
M60-M79	Soft tissue disorders
M80-M94	Osteopathies and chondropathies
M95-M99	Other disorders of the musculoskeletal system and connective tissue
N00-N08	Glomerular diseases
N10-N16	Renal tubulo-interstitial diseases
N17-N19	Renal failure
N20-N23	Urolithiasis
N25-N29	Other disorders of kidney and ureter
N30-N39	Other diseases of urinary system
N40-N51	Diseases of male genital organs
N60-N64	Disorders of breast
N70-N77	Inflammatory diseases of female pelvic organs
N80-N98	Noninflammatory disorders of female genital tract
N99-N99	Intraoperative and postprocedural complications and disorders of genitourinary system, not elsewhere classified
O00-O08	Pregnancy with abortive outcome
O10-O16	Edema, proteinuria and hypertensive disorders in pregnancy, childbirth and the puerperium
O20-O29	Other maternal disorders predominantly related to pregnancy
O30-O48	Maternal care related to the fetus and amniotic cavity and possible delivery problems
O60-O75	Complications of labor and delivery
O80-O84	Delivery
O85-O92	Complications predominantly related to the puerperium
O94-O99	Other obstetric conditions, not elsewhere classified
P00-P04	Fetus and newborn affected by maternal factors and by complications of pregnancy, labor and delivery
P05-P08	Disorders related to length of gestation and fetal growth

P10-P15	Birth trauma
P20-P29	Respiratory and cardiovascular disorders specific to the perinatal period
P35-P39	Infections specific to the perinatal period
P50-P61	Hemorrhagic and hematological disorders of fetus and newborn
P70-P74	Transitory endocrine and metabolic disorders specific to fetus and newborn
P75-P78	Digestive system disorders of fetus and newborn
P80-P83	Conditions involving the integument and temperature regulation of fetus and newborn
P90-P96	Other disorders originating in the perinatal period
Q00-Q07	Congenital malformations of the nervous system
Q10-Q18	Congenital malformations of eye, ear, face and neck
Q20-Q28	Congenital malformations of the circulatory system
Q30-Q34	Congenital malformations of the respiratory system
Q35-Q37	Cleft lip and cleft palate
Q38-Q45	Other congenital malformations of the digestive system
Q50-Q56	Congenital malformations of genital organs
Q60-Q64	Congenital malformations of the urinary system
Q65-Q79	Congenital malformations and deformations of the musculoskeletal system
Q80-Q89	Other congenital malformations
Q90-Q99	Chromosomal abnormalities, not elsewhere classified
S00-S09	Injuries to the head
S10-S19	Injuries to the neck
S20-S29	Injuries to the thorax
S30-S39	Injuries to the abdomen, lower back, lumbar spine and pelvis
S40-S49	Injuries to the shoulder and upper arm
S50-S59	Injuries to the elbow and forearm
S60-S69	Injuries to the wrist and hand
S70-S79	Injuries to the hip and thigh
S80-S89	Injuries to the knee and lower leg
S90-S99	Injuries to the ankle and foot
T00-T07	Injuries involving multiple body regions
T08-T14	Injuries to unspecified part of trunk, limb or body region
T15-T19	Effects of foreign body entering through natural orifice
T20-T32	Burns and corrosions
T33-T35	Frostbite
T36-T50	Poisoning by drugs, medicaments and biological substances
T51-T65	Toxic effects of substances chiefly nonmedicinal as to source
T66-T78	Other and unspecified effects of external causes
T79-T79	Certain early complications of trauma
T80-T88	Complications of surgical and medical care, not elsewhere classified
T90-T98	Sequelae of injuries, of poisoning and of other consequences of external causes
U00-U49	Provisional assignment of new diseases of uncertain etiology
V01-X59	Accidents
X60-X84	Intentional self-harm
X85-Y09	Assault
Y10-Y34	Event of undetermined intent
Y35-Y36	Legal intervention and operations of war
Y40-Y84[b]	Therapeutic and surgical interventions
Y85-Y89	Sequelae of external causes of morbidity and mortality
Y90-Y98	Supplementary factors related to causes of morbidity and mortality classified elsewhere

List of codes modified from World Health Organization ICD-10 2010 version (https://apps.who.int/classifications/apps/icd/ClassificationDownload/DLArea/Download.aspx)
[a] F00, F01, F02, F03, and G30 codes removed from blocks; [b] Code used in Finland

Supplementary Table 2. List of ICD-10 blocks removed in study

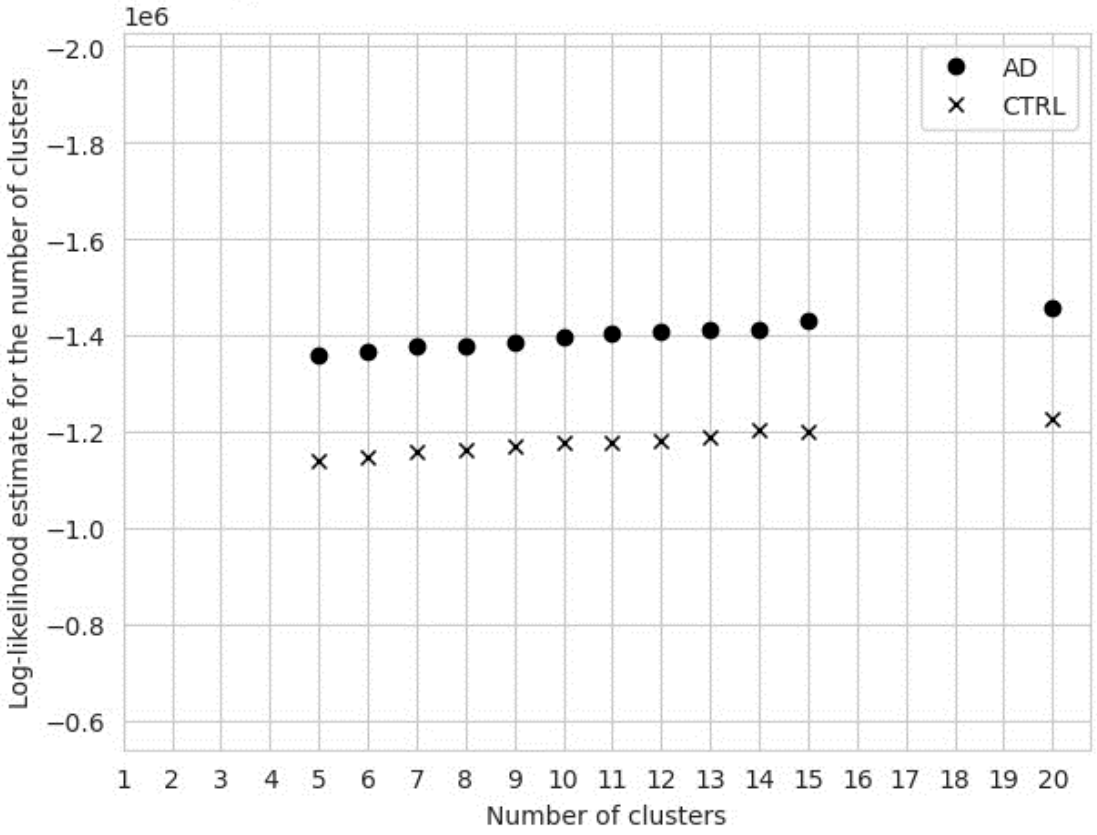
R00-R09	Symptoms and signs involving the circulatory and respiratory systems
R10-R19	Symptoms and signs involving the digestive system and abdomen
R20-R23	Symptoms and signs involving the skin and subcutaneous tissue
R25-R29	Symptoms and signs involving the nervous and musculoskeletal systems
R30-R39	Symptoms and signs involving the urinary system
R40-R46	Symptoms and signs involving cognition, perception, emotional state and behavior
R47-R49	Symptoms and signs involving speech and voice
R50-R69	General symptoms and signs
R70-R79	Abnormal findings on examination of blood, without diagnosis
R80-R82	Abnormal findings on examination of urine, without diagnosis
R83-R89	Abnormal findings on examination of other body fluids, substances and tissues, without diagnosis
R90-R94	Abnormal findings on diagnostic imaging and in function studies, without diagnosis
R95-R99	Ill-defined and unknown causes of mortality
Z00-Z13	Supplementary factors related to causes of morbidity and mortality classified elsewhere
Z20-Z29	Persons with potential health hazards related to communicable diseases
Z30-Z39	Persons encountering health services in circumstances related to reproduction
Z40-Z54	Persons encountering health services for specific procedures and health care
Z55-Z65	Persons with potential health hazards related to socioeconomic and psychosocial circumstances
Z70-Z76	Persons encountering health services in other circumstances
Z80-Z99	Persons with potential health hazards related to family and personal history and certain conditions influencing health status

Supplementary Table 3. 30 most frequent ICD-10 blocks for males in AD and CTRL cohorts

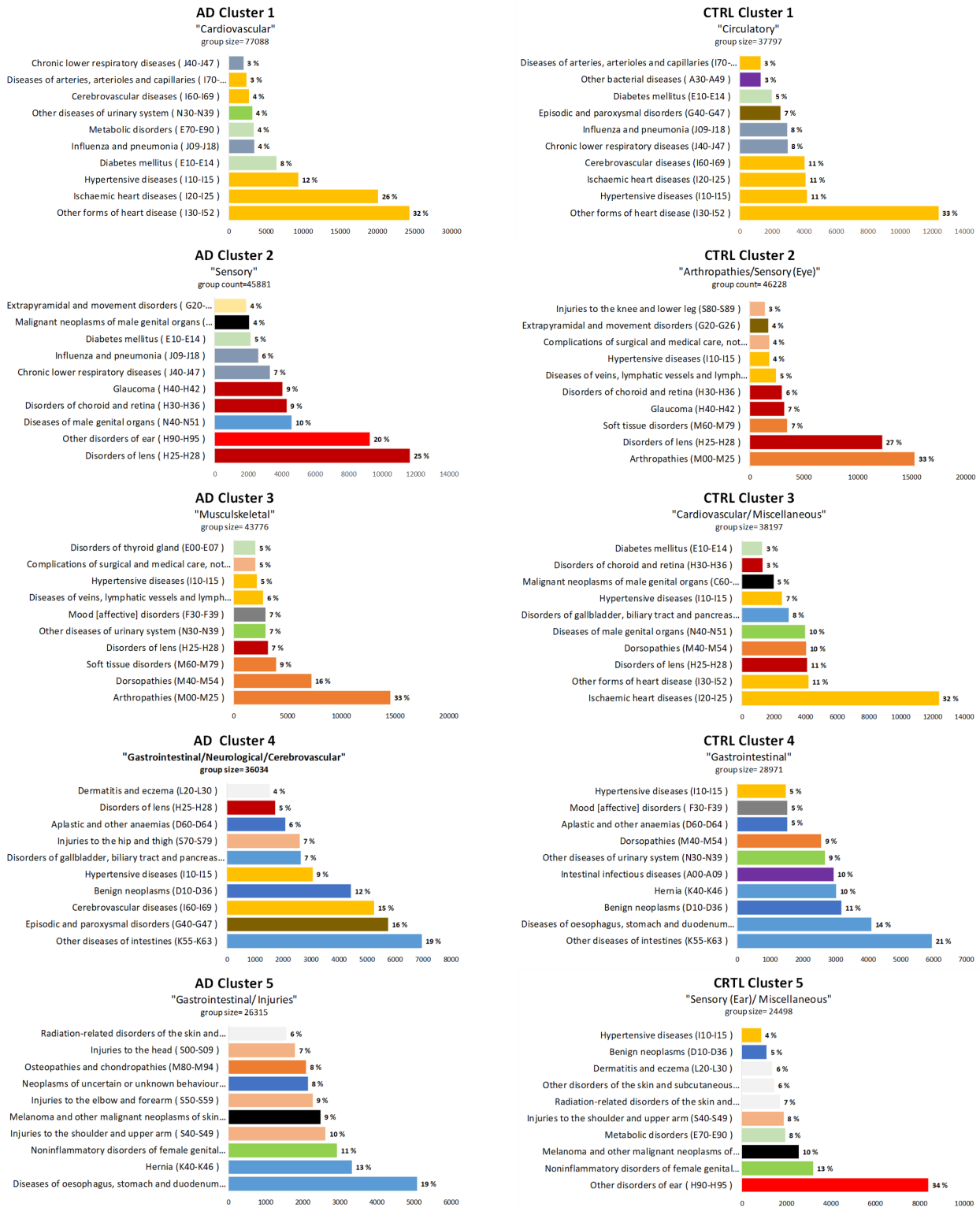
	AD cohort			CTRL cohort		
	ICD-10 Block	**n**	**% of cohort with diagnosis block**	**ICD-10 Block**	**n**	**% of cohort with diagnosis block**
1	I30-I52	9121	8%	I20-I25	7699	8%
2	I20-I25	8503	7%	I30-I52	7384	8%
3	H25-H28	4893	4%	H25-H28	4518	5%
4	I10-I15	4804	4%	M00-M25	4258	5%
5	N40-N51	4521	4%	I10-I15	4059	4%
6	M00-M25	4310	4%	N40-N51	3862	4%
7	H90-H95	3663	3%	H90-H95	3250	3%
8	I60-I69	3592	3%	E10-E14	2408	3%
9	E10-E14	3546	3%	I60-I69	2401	3%
10	J09-J18	2831	2%	J40-J47	2302	2%
11	G40-G47	2576	2%	J09-J18	2168	2%
12	M40-M54	2361	2%	M40-M54	2047	2%
13	K55-K63	2349	2%	C60-C63	1955	2%
14	J40-J47	2348	2%	K55-K63	1819	2%
15	C60-C63	2097	2%	G40-G47	1771	2%
16	K40-K46	1952	2%	K40-K46	1732	2%
17	K20-K31	1827	2%	D10-D36	1576	2%
18	S00-S09	1700	1%	K20-K31	1479	2%
19	N30-N39	1677	1%	I70-I79	1441	2%
20	D10-D36	1641	1%	H30-H36	1408	1%
21	I70-I79	1624	1%	M60-M79	1371	1%
22	A30-A49	1505	1%	A30-A49	1194	1%
23	H30-H36	1497	1%	H40-H42	1180	1%
24	M60-M79	1405	1%	N30-N39	1073	1%
25	F00-F09	1362	1%	K80-K87	1071	1%
26	E70-E90	1305	1%	C43-C44	1066	1%
27	H40-H42	1210	1%	I80-I89	1026	1%
28	I80-I89	1174	1%	E70-E90	986	1%
29	A00-A09	1157	1%	S00-S09	908	1%
30	K80-K87	1048	1%	D37-D48	881	1%

Supplementary Table 4. 30 most frequent ICD-10 blocks for females in AD and CTRL cohorts

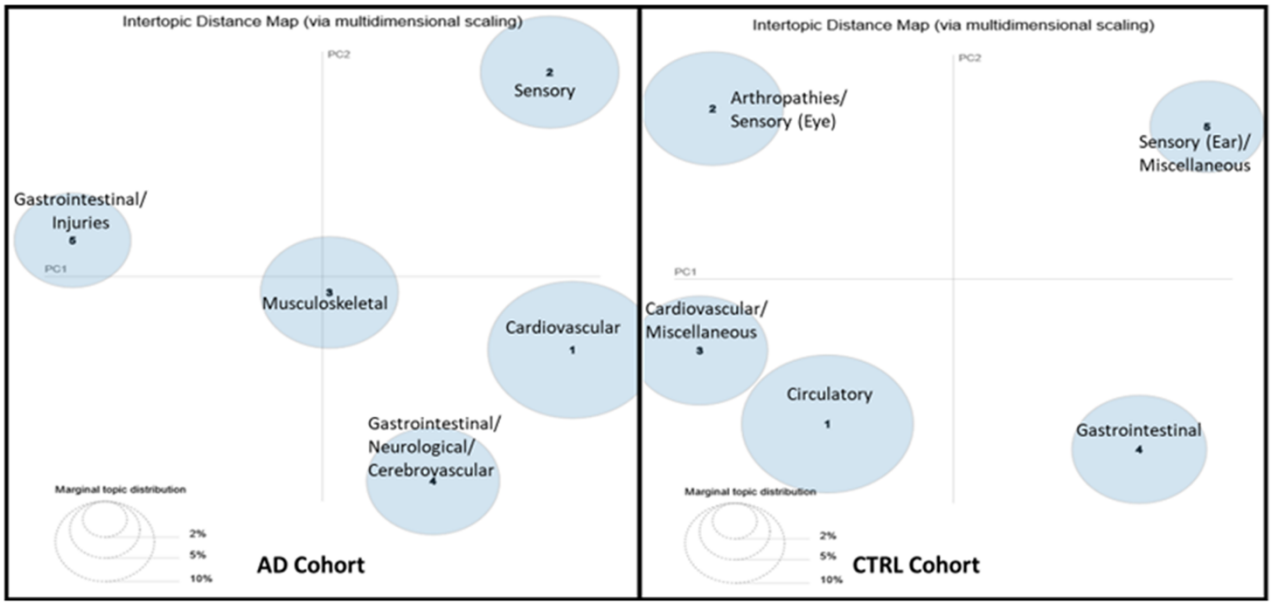
	AD cohort			CTRL cohort		
	ICD-10 Block	**n**	**% of cohort with diagnosis block**	**ICD-10 Block**	**n**	**% of cohort with diagnosis block**
1	I30-I52	15304	8%	I30-I52	13433	8%
2	H25-H28	12549	6%	H25-H28	12215	7%
3	I20-I25	11612	6%	M00-M25	11011	6%
4	I10-I15	11254	6%	I20-I25	10449	6%
5	M00-M25	10389	5%	I10-I15	9970	6%
6	N30-N39	7167	4%	H90-H95	5229	3%
7	H90-H95	5614	3%	M40-M54	4600	3%
8	E10-E14	5042	2%	N30-N39	4455	3%
9	M40-M54	4901	2%	K55-K63	4113	2%
10	K55-K63	4669	2%	E10-E14	4096	2%
11	I60-I69	4411	2%	I60-I69	3403	2%
12	G40-G47	3750	2%	N80-N98	3226	2%
13	S70-S79	3312	2%	H40-H42	3015	2%
14	K20-K31	3217	2%	H30-H36	2946	2%
15	E70-E90	3213	2%	J40-J47	2879	2%
16	J09-J18	3174	2%	J09-J18	2840	2%
17	N80-N98	2933	1%	G40-G47	2787	2%
18	J40-J47	2884	1%	D10-D36	2773	2%
19	H30-H36	2839	1%	K20-K31	2640	2%
20	H40-H42	2832	1%	M60-M79	2612	2%
21	S00-S09	2815	1%	I80-I89	2511	1%
22	I80-I89	2804	1%	E70-E90	2336	1%
23	D10-D36	2783	1%	A00-A09	2170	1%
24	A00-A09	2752	1%	S70-S79	2056	1%
25	F30-F39	2579	1%	A30-A49	1890	1%
26	A30-A49	2569	1%	K80-K87	1853	1%
27	M60-M79	2565	1%	N10-N16	1758	1%
28	N10-N16	2460	1%	C43-C44	1559	1%
29	F00-F09	2341	1%	T80-T88	1524	1%
30	S50-S59	2002	1%	S00-S09	1496	1%



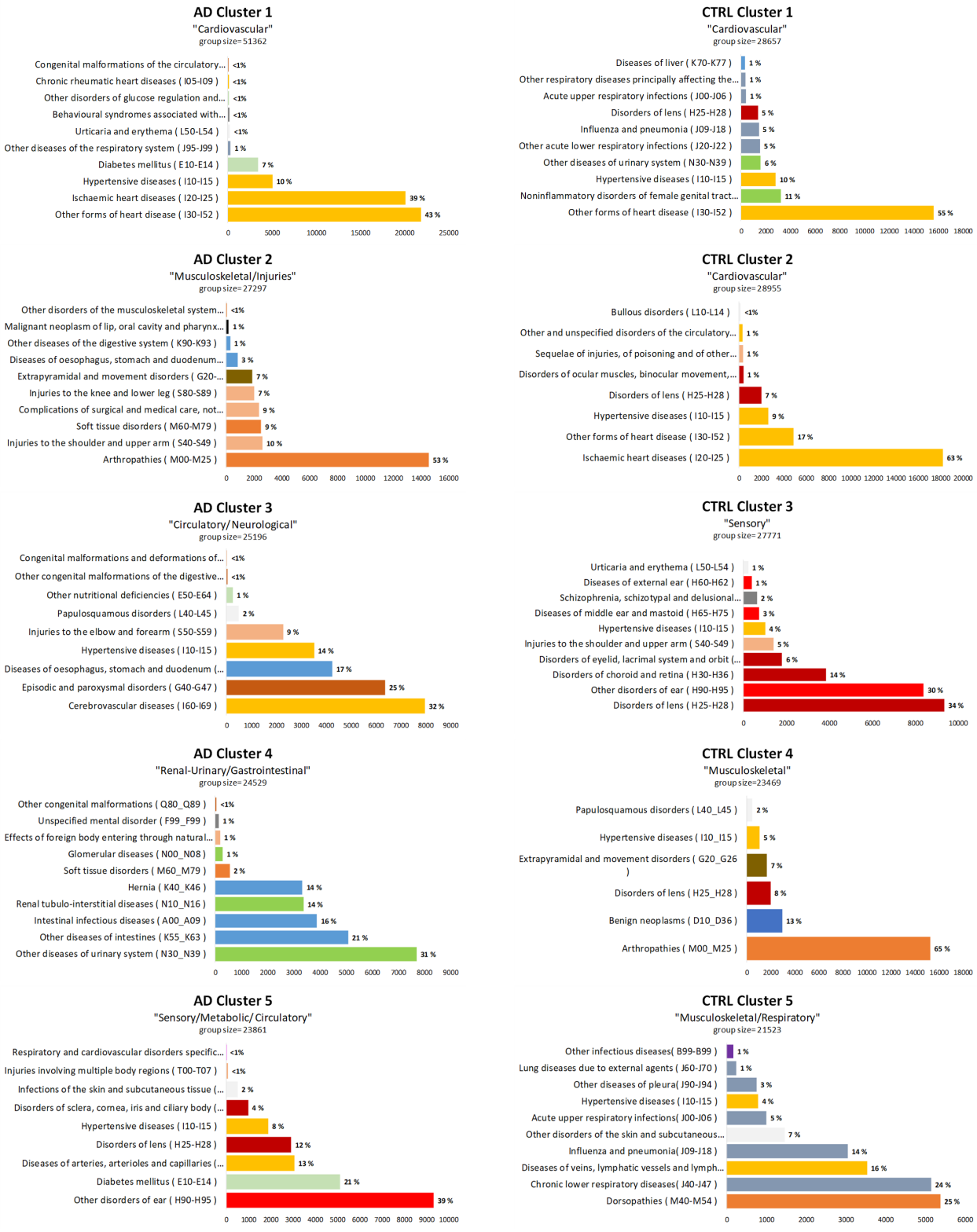
Supplementary Figure 2. Cohort-wise log-likelihood estimates of Latent Dirichlet Allocation with different number of clusters. AD, Alzheimer's Disease; CTRL, Control



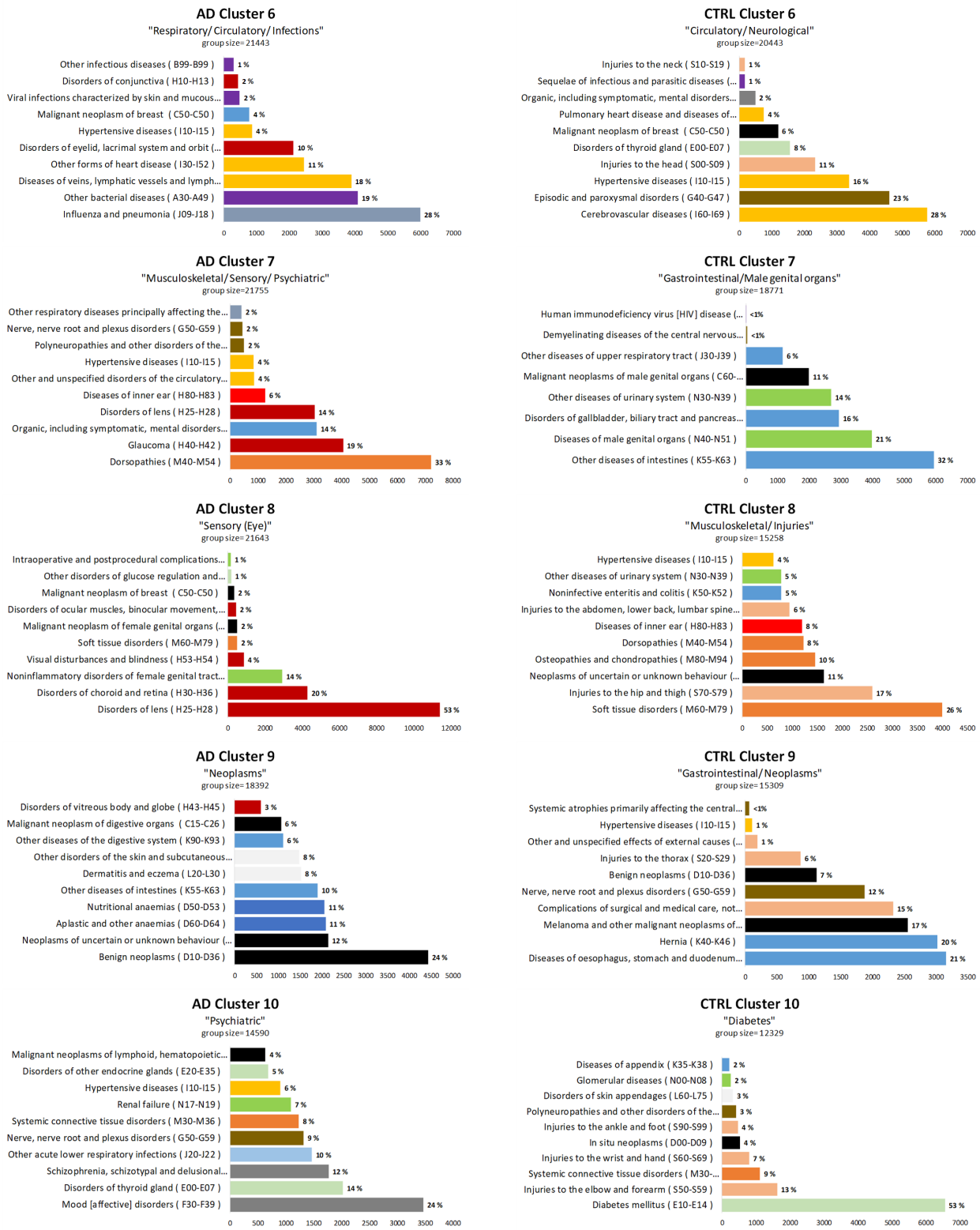
Supplementary Figure 3. 5-Cluster model with ICD-10 Block distribution for AD and CTRL cohorts.



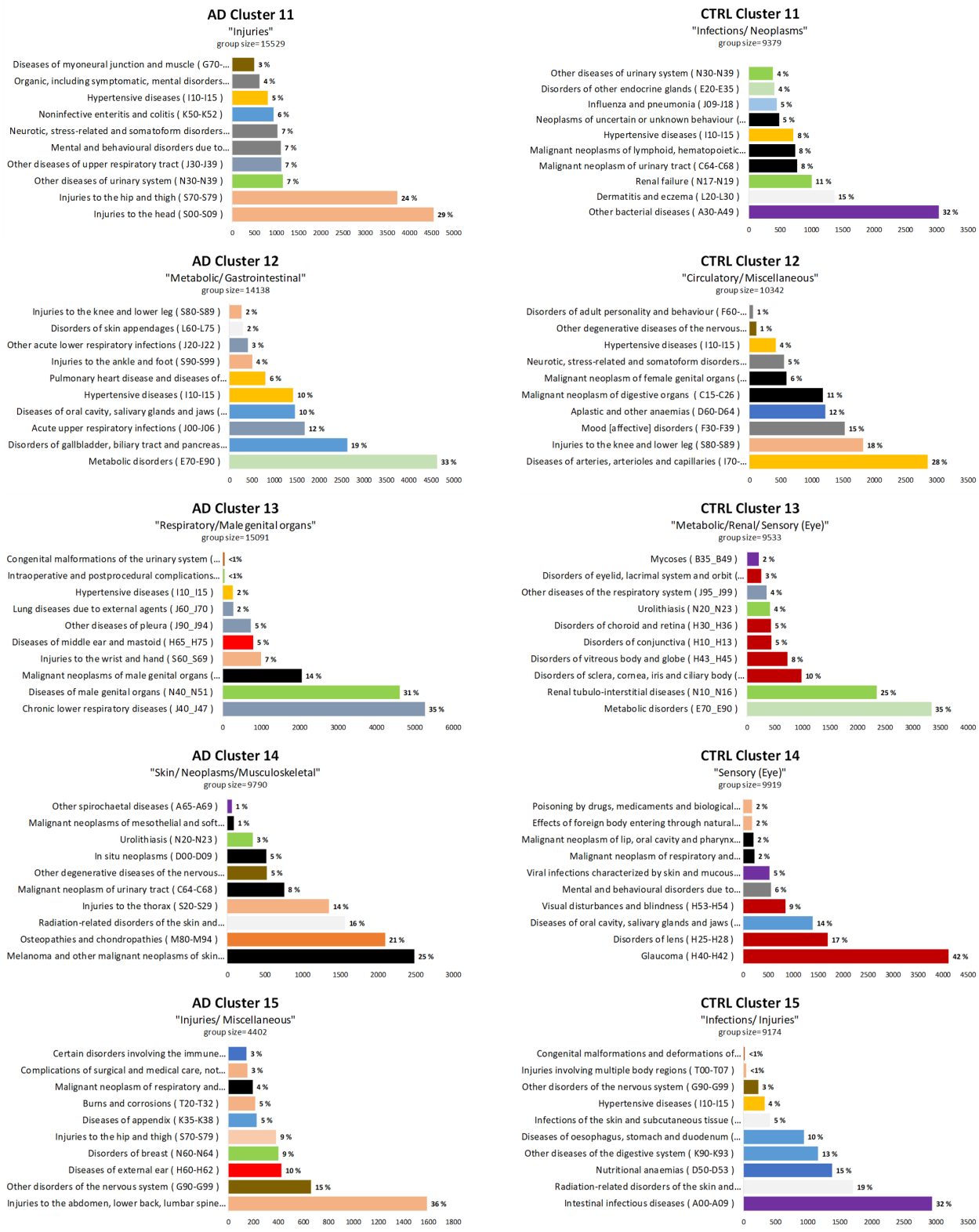
Supplementary Figure 4. Inter-topic distance map for the 5-cluster AD and comparison cohorts via multidimensional scaling. A two-dimensional visualization of all the clusters where the size of the bubble represents the percentage of the blocks (words) in the corpus that the cluster contains. The distance between the clusters shows how similar the clusters are to each other. (The axes are not interpretable and come from the multidimensional scaling algorithm). AD, Alzheimer's disease; CTRL, Control.



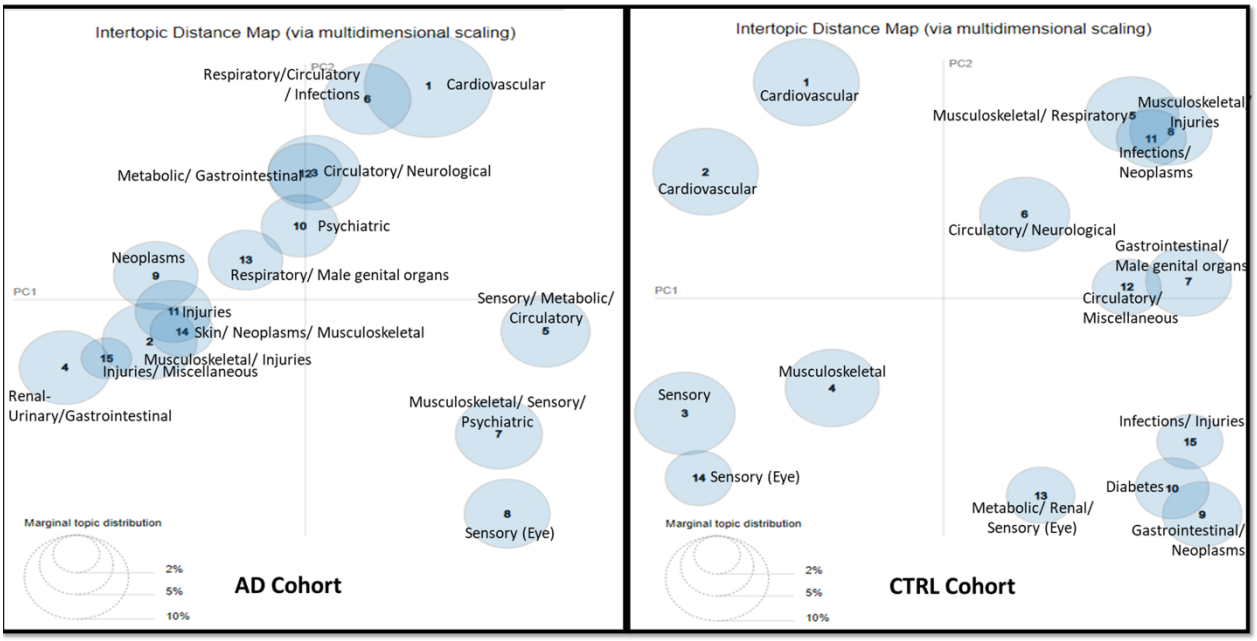
Supplementary Figure 5. 15-Cluster model with ICD-10 Block distribution for AD and CTRL cohorts (page 1/3; continued on next page).



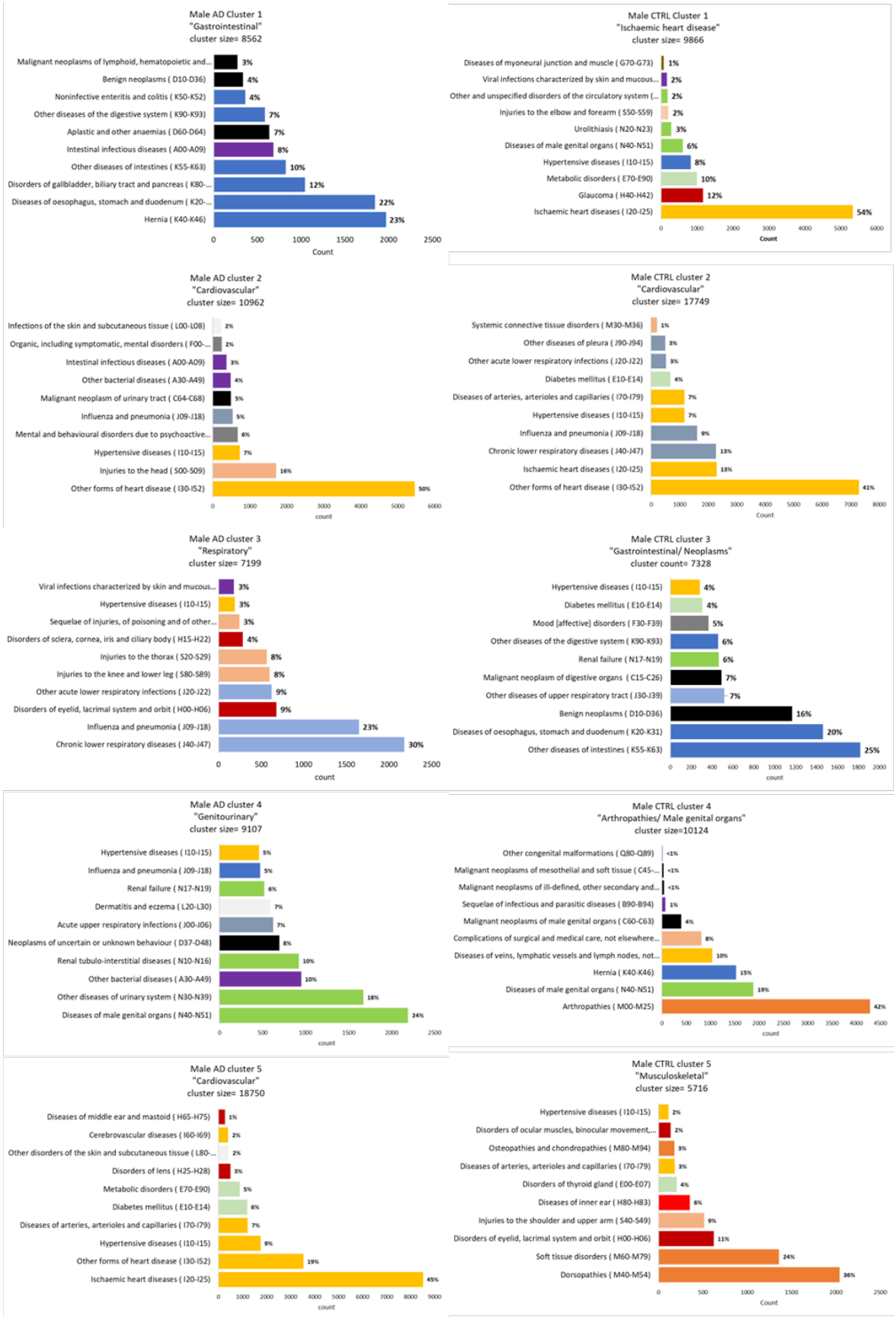
Supplementary Figure 5. 15-Cluster model with ICD-10 Block distribution for AD and CTRL cohorts (page 2/3; continued on next page).



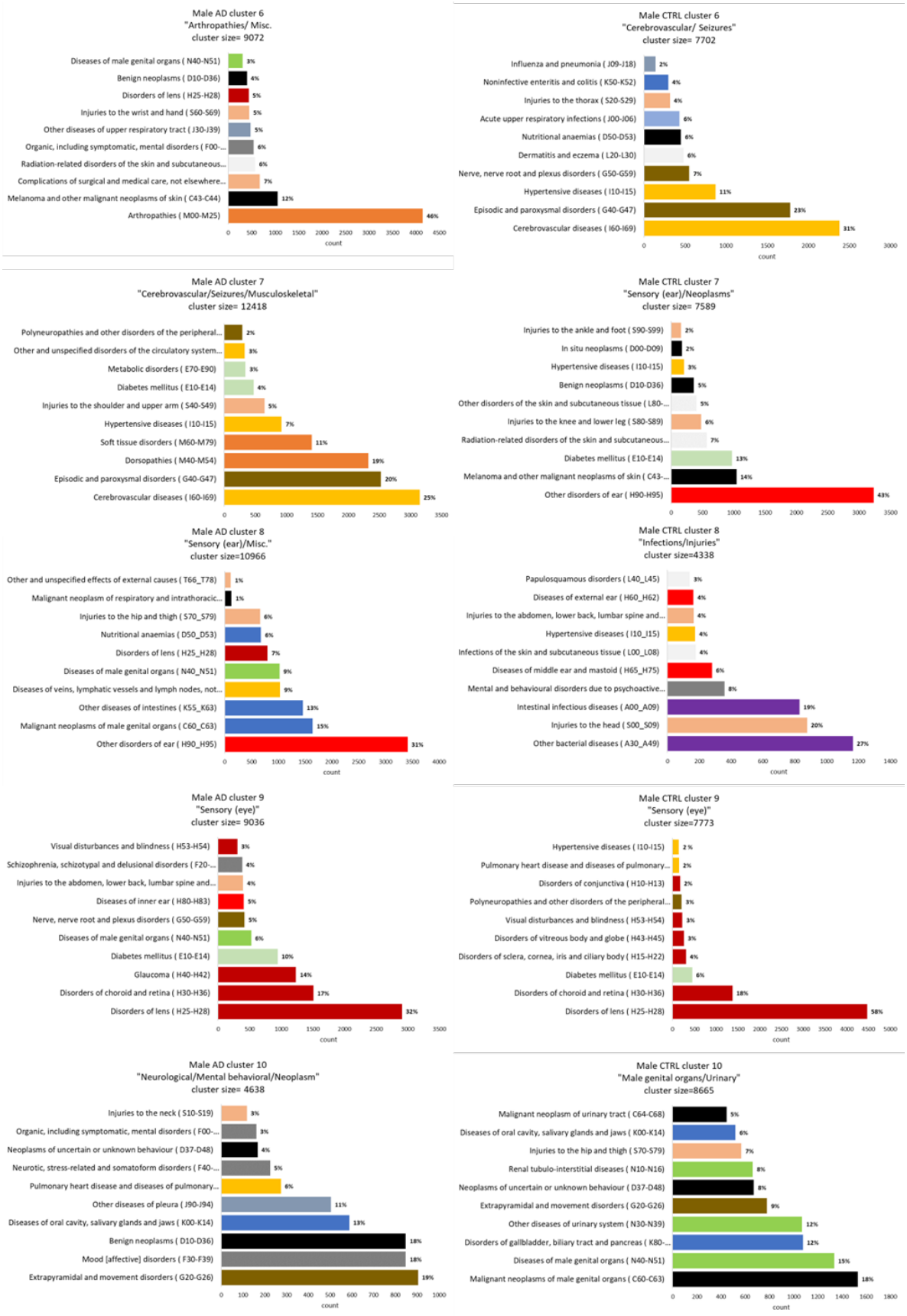
Supplementary Figure 5. 15-Cluster model with ICD-10 Block distribution for AD and CTRL cohorts (page 3/3).



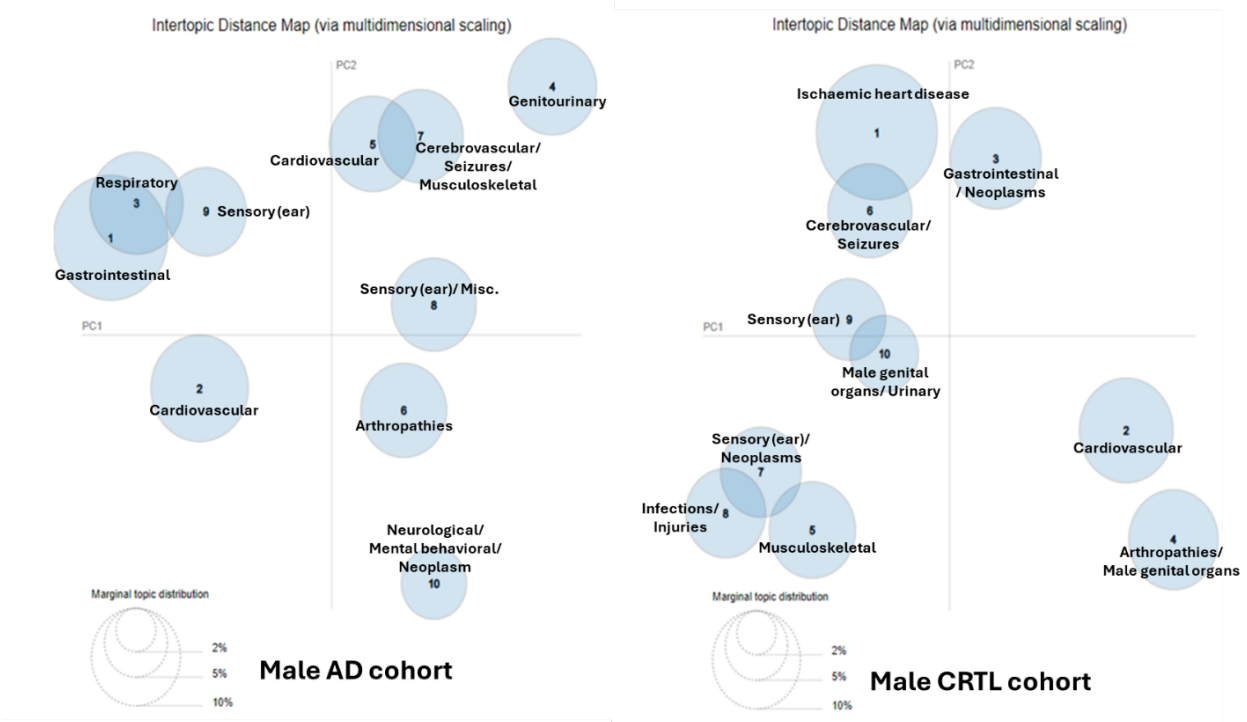
Supplementary Figure 6. Inter-topic distance map for the 15-cluster AD and comparison cohorts via multidimensional scaling. A two-dimensional visualization of all the clusters where the size of the bubble represents the percentage of the blocks (words) in the corpus that the cluster contains. The distance between the clusters shows how similar the clusters are to each other. (The axes are not interpretable and come from the multidimensional scaling algorithm). AD, Alzheimer's disease; CTRL, Control; Misc., Miscellaneous.



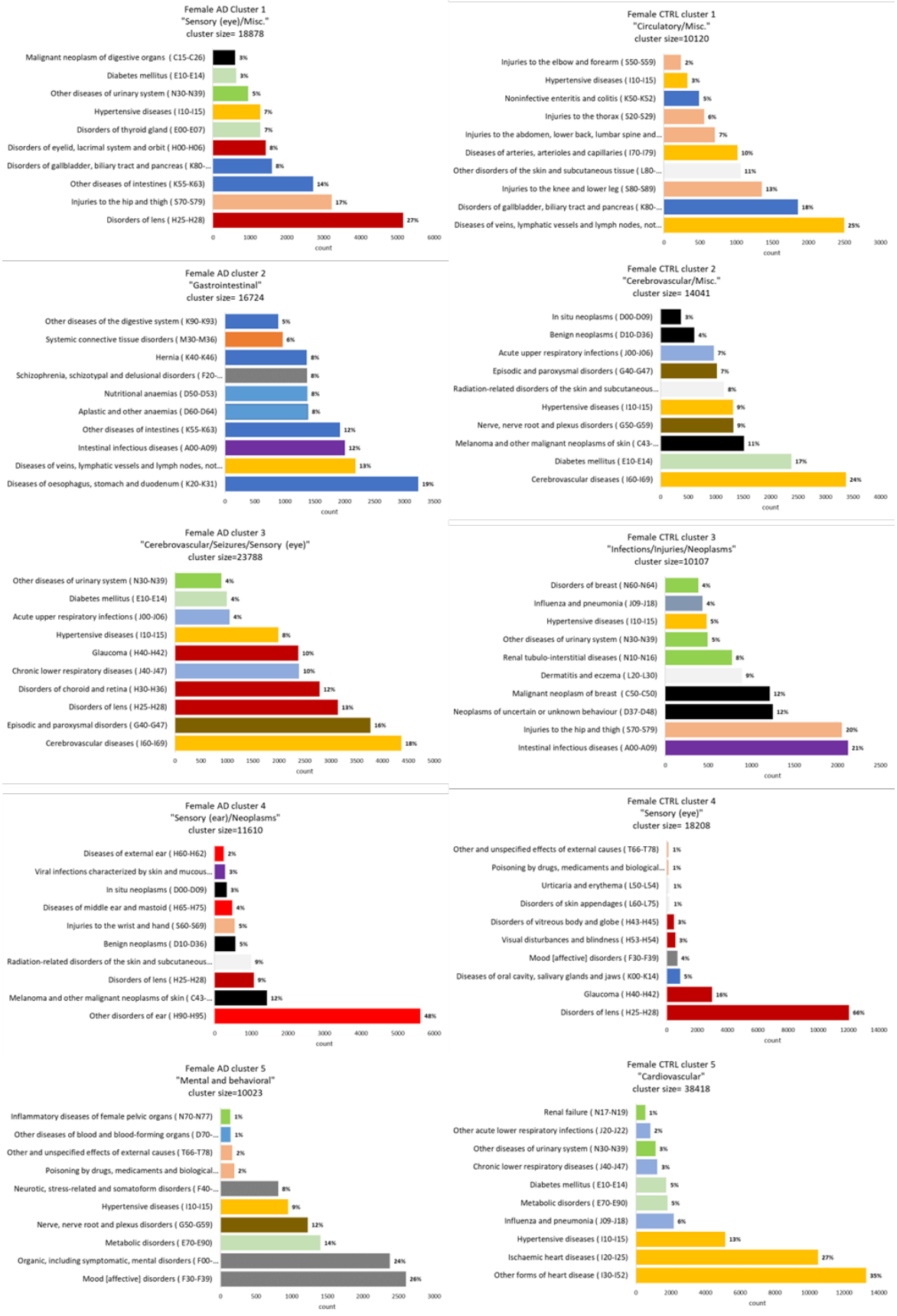
Supplementary Figure 7. 10-Cluster model with ICD-10 Block distribution for males in the AD and comparison cohort (continued on next page).



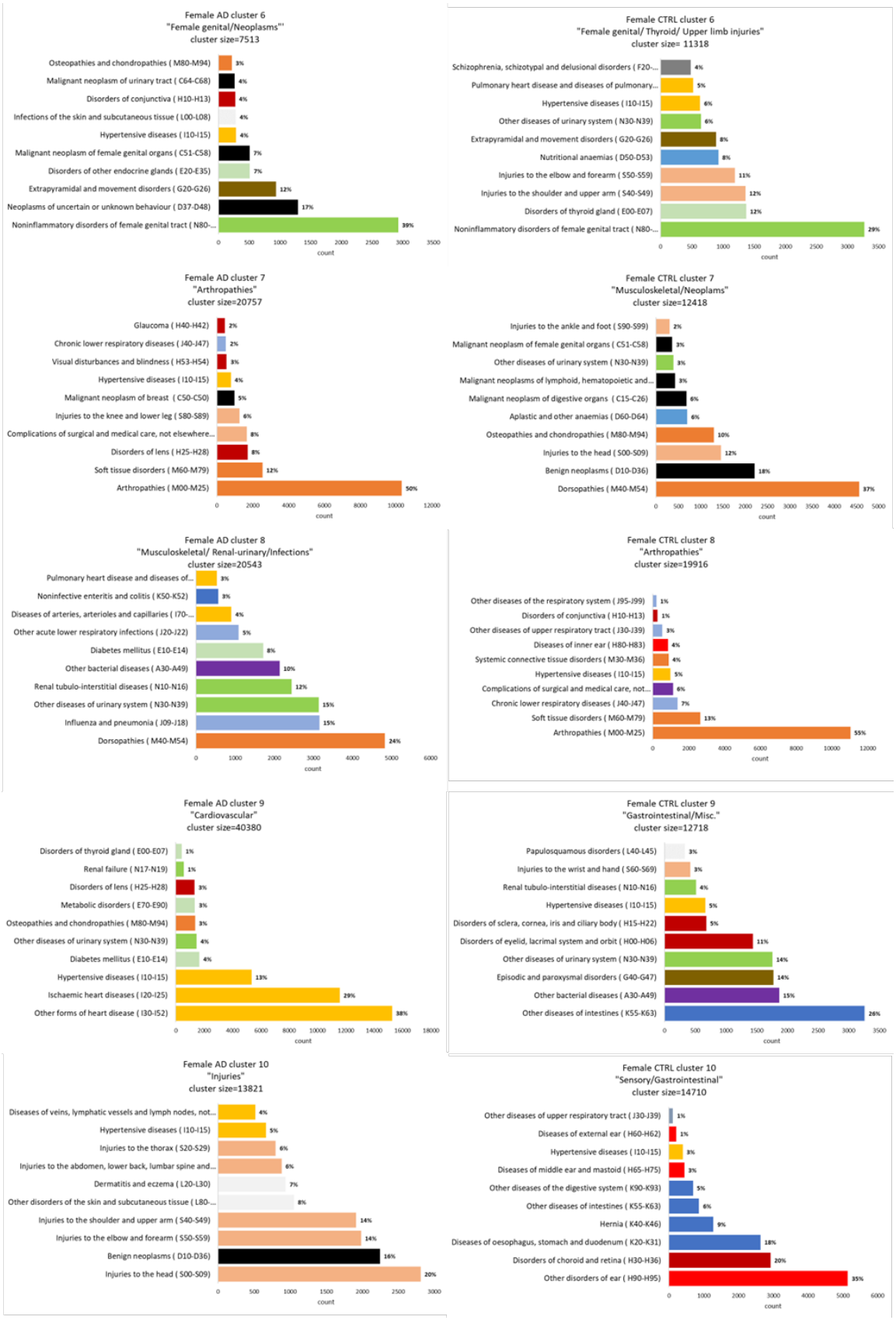
Supplementary Figure 7. 10-Cluster model with ICD-10 Block distribution for males in the AD and comparison cohort.



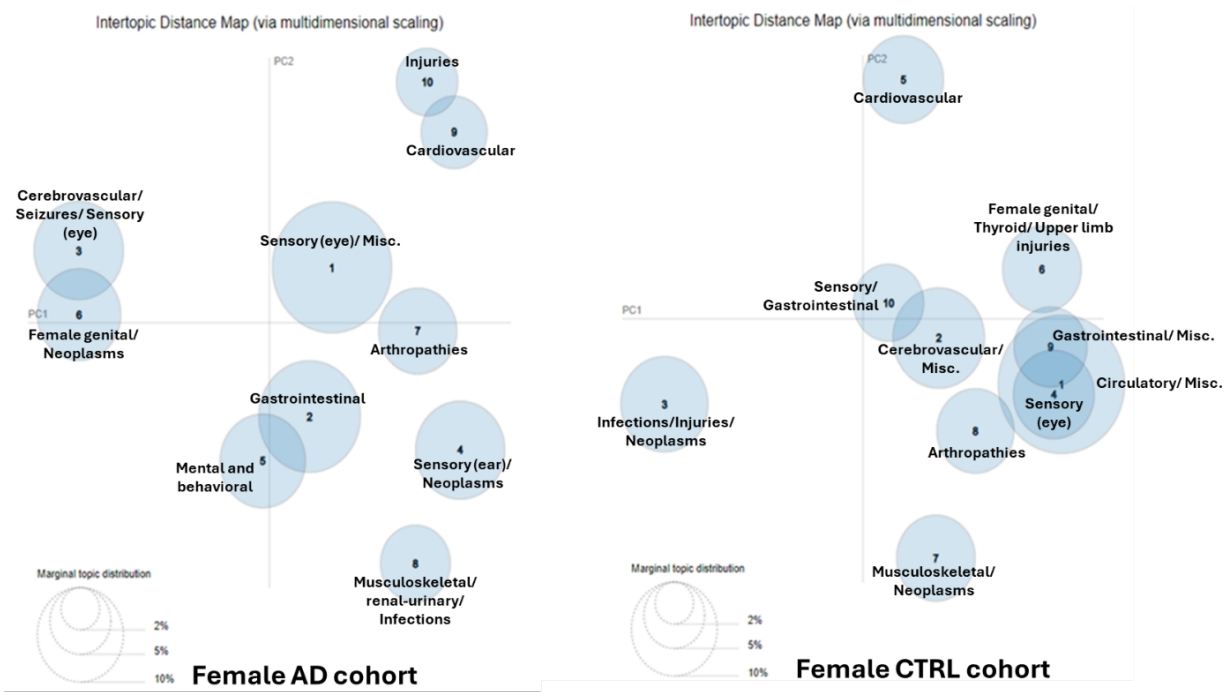
Supplementary Figure 8. Male inter-topic distance map for the 10-cluster AD and comparison cohorts via multidimensional scaling. A two-dimensional visualization of all the clusters where the size of the bubble represents the percentage of the blocks (words) in the corpus that the cluster contains. The distance between the clusters shows how similar the clusters are to each other. (The axes are not interpretable and come from the multidimensional scaling algorithm). AD, Alzheimer's disease; CTRL, Control; Misc., Miscellaneous.



Supplementary Figure 9. 10-Cluster model with ICD-10 Block distribution for females in the AD and comparison cohort (continued on next page).



Supplementary Figure 9. 10-Cluster model with ICD-10 Block distribution for females in the AD and comparison cohort.



Supplementary Figure 10. Female inter-topic distance map for the 10-cluster AD and comparison cohorts via multidimensional scaling. A two-dimensional visualization of all the clusters where the size of the bubble represents the percentage of the blocks (words) in the corpus that the cluster contains. The distance between the clusters shows how similar the clusters are to each other. (The axes are not interpretable and come from the multidimensional scaling algorithm). AD, Alzheimer's disease; CTRL, Control; Misc., Miscellaneous.