

Using Latent Class Analysis to characterise clusters of multimorbid patients using CPRD

Yajing Zhu^{1,3}, Duncan Edwards¹, Jonathan Mant¹, Rupert A Payne², Steven J Kiddle¹

Affiliation: ¹ University of Cambridge, ² University of Bristol. ³ F.Hoffmann - La Roche Ltd

Email: yajing.zhu@roche.com

Statistical methods for Cluster Analysis applied to multimorbidity workshop

26 Nov 2020

Agenda

1. Intro & study background
2. Data
3. Methodology
4. Results
5. Discussion

Introduction

- Data scientist in Neuroscience @ Personalised healthcare, PD, Roche (RWD + Advanced analytics)
- Postdoctoral researcher at MRC Biostatistics Unit, precision medicine (CPRD, VitalPAC - clinical monitoring system in hospitals, NHS)

Background: MM

- Multimorbidity (MM, co-existence of 2+ chronic conditions) is increasingly common in ageing societies.
- 25% English adults (~14 million people) have multimorbidity.
- 30% of people with 4+ conditions are under 65s.
- Challenges on the single-disease-centred treatment /caring framework
 - Heavy users of medications
 - Greater mortality, higher NHS service use

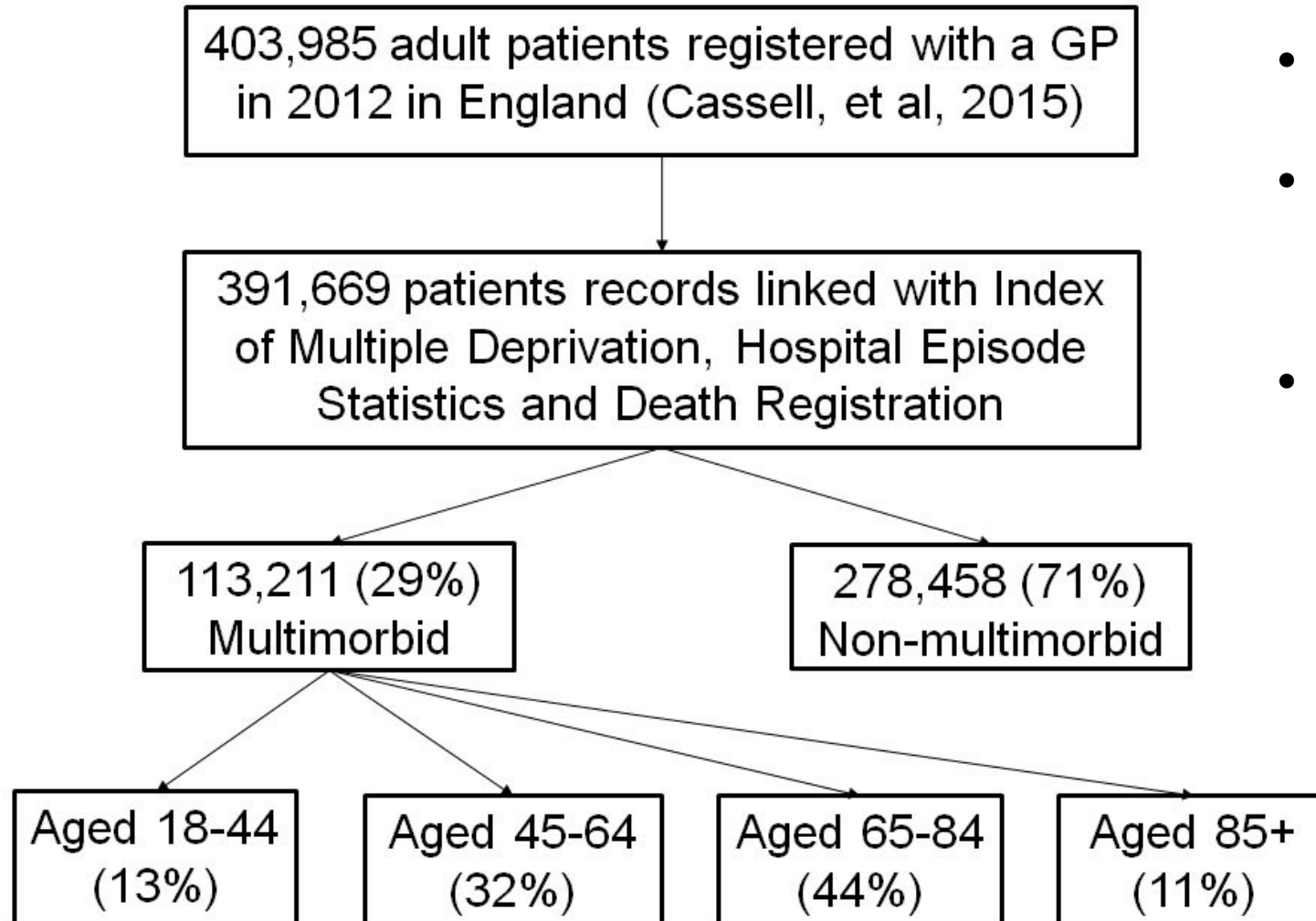
Background: limitations of previous work

- By 2020, large MM studies used UK Biobank data (**healthier**, less socioeconomically deprived)
- Mostly **focused on 60+ populations**
- **Not age-stratified** → scarce evidence for the younger MM population
- Profiling MM groups mostly **focused on 2 conditions only**
- Most studies focused on **grouping diseases, not patients**
- **Lack of validity and generalisability** for patient-centered policy-making and care

Goal

- Which diseases co-occur?
- To provide a comprehensive MM profiles across age groups:
What is their distributions across age groups?
- What are the social patterns of multimorbidity clusters?
- Highlight combinations that lead to the highest mortality and service use.

Data: UK primary care electronic health records (CPRD-GOLD)

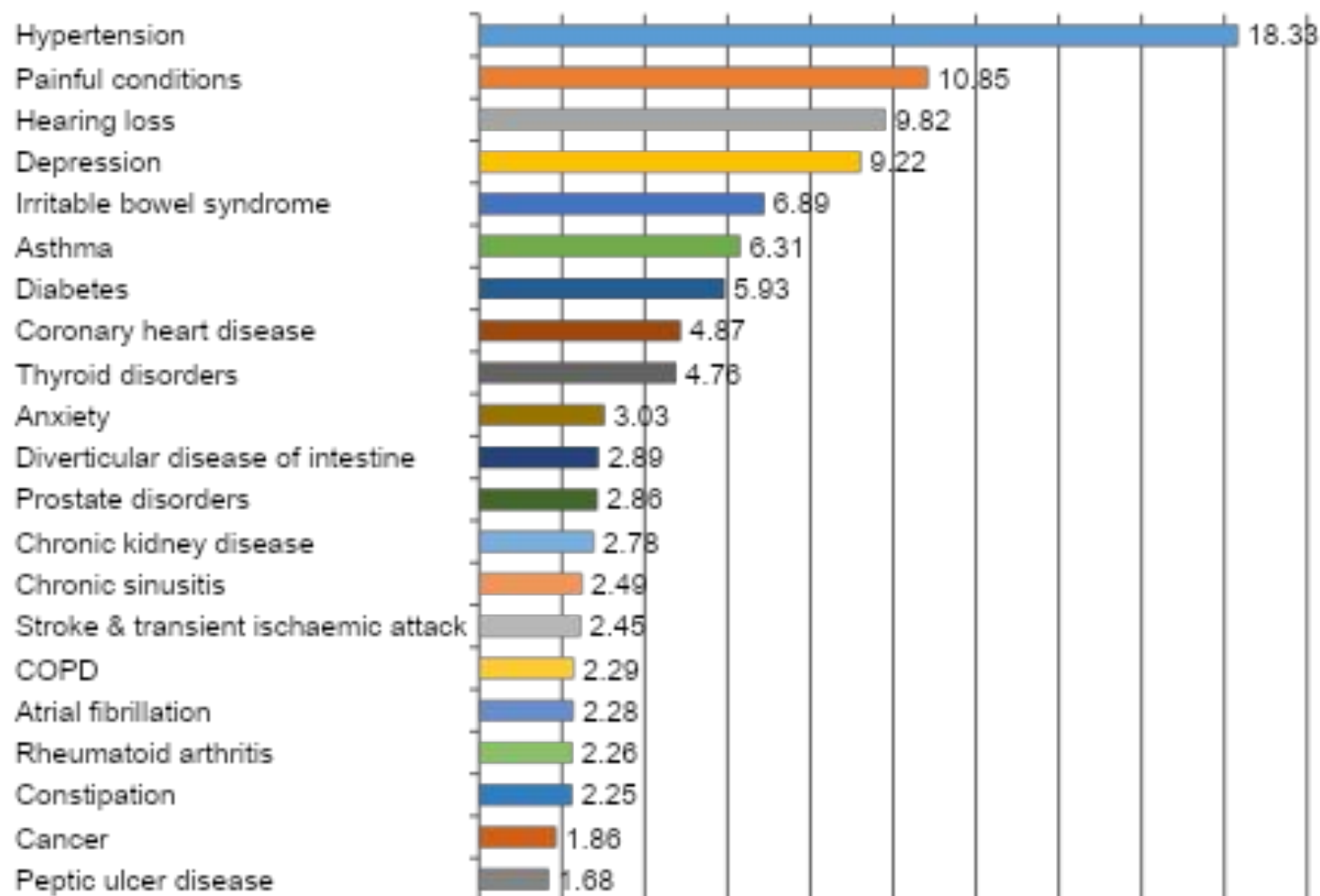


- ONS: all-cause mortality data
- IMD: area-based (1000–1500 people per area around patients' home)
- Study protocol 16_057RA2

Data: definition of 38 LTCs

- Binary(present or not): classification of LTCs in primary care developed by Barnett et al. (2012, Lancet)
- Readcode and Procode system. Cambridge-CPRD codelist
https://www.phpc.cam.ac.uk/pcu/cprd_cam/codelists/v11/
- Distribution of LTCs match other large-sample-size UK multimorbidity studies

Data: distribution of LTCs (snapshot)



Data: definition of patient characteristics

- Gender
- Age groups (stratified into 18–44, 45–64, 65–84 and 85+ years) in 2012
- Last recorded pre-2012
 - body mass index (BMI)
 - smoking status (current, never and ex-smokers)
- Socioeconomic deprivation - quintiles of IMD across the UK (1 for the least socioeconomically deprived quintile of areas and 5 for the most).

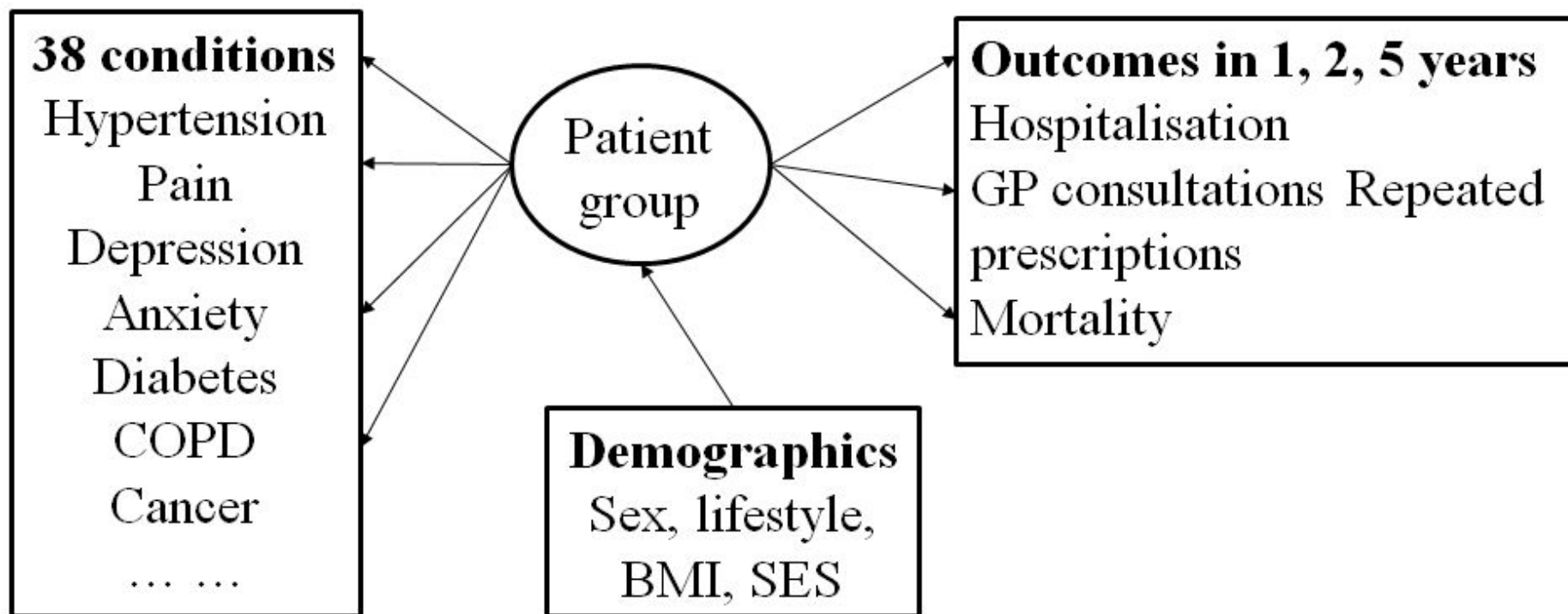
Data: definition of outcomes

- All-cause mortality at 2 and 5 years (ONS)
- NHS service utilisation or treatment burden in 1 year:
 - **primary care consultations** (with any clinician in the primary care team),
 - # all-type **hospitalisation spells** (defined by discharge dates)
 - the count of **regular medications** (regular = 4+ repeated prescriptions in a year by counting the unique British National Formulary (BNF) codes).

Methodology

Step 1: Latent class analysis (LCA)

Step 2: Relate patient groups to health outcomes and patient demographics



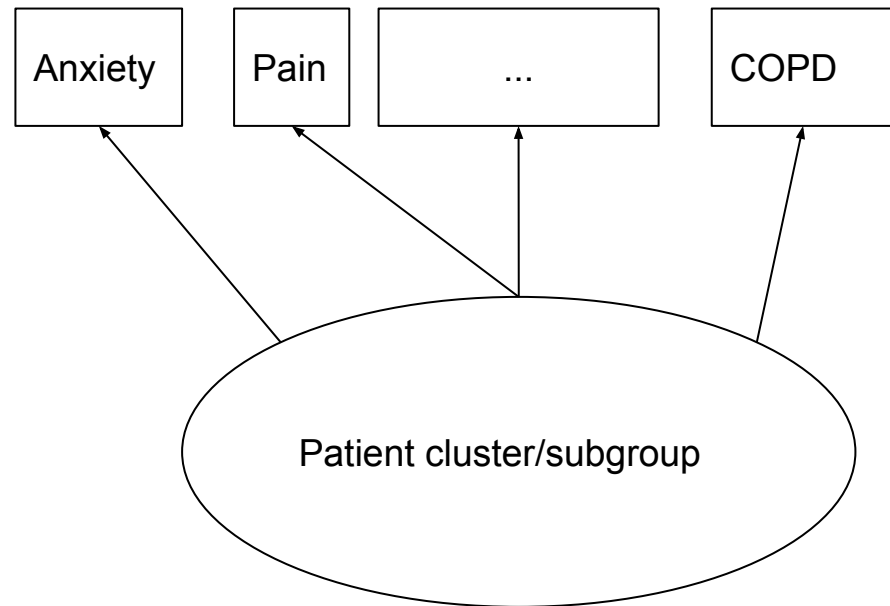
Methodology: details

- Latent class analysis (latent variable model - finite mixture model) to assign all patients to **non-overlapping clusters** (i.e. each patient is assigned to only one cluster)
- Model-based probabilistic clustering approach & **data-driven**
- Stratified by **age groups**
- Identify multimorbidity clusters using a random set of 80% of the multimorbid patients, with **consistency** of results checked in the remaining 20%
- Relationship with external quantities: descriptive & generalised linear models

Latent class analysis: technical summary

- Parameters
 - Prevalence of cluster → defines cluster size
 - Conditional probabilities → defines cluster profile
- Assumptions
 - Conditional independence** of items → local independence, ensures likelihood function is neat.
 - Finite mixtures** → likelihood function derived using component likelihood from each mixture.
 - MAR**
(Full-information-maximum-likelihood)
- Model selection statistics**
BIC, SABIC, Lo-Mendell-Rubin Test, Bootstrapped parametric likelihood ratio test, relative entropy

$$E = 1 + \frac{1}{N \log(k)} \left(\sum_{i=1}^N \sum_{k=1}^K P(C = k | U_i) \log(P(C = k | U_i)) \right)$$



Extension:

Continuous items → Latent profile analysis

Patient risk score → Factor analysis

Predictors of patient clusters → LCR

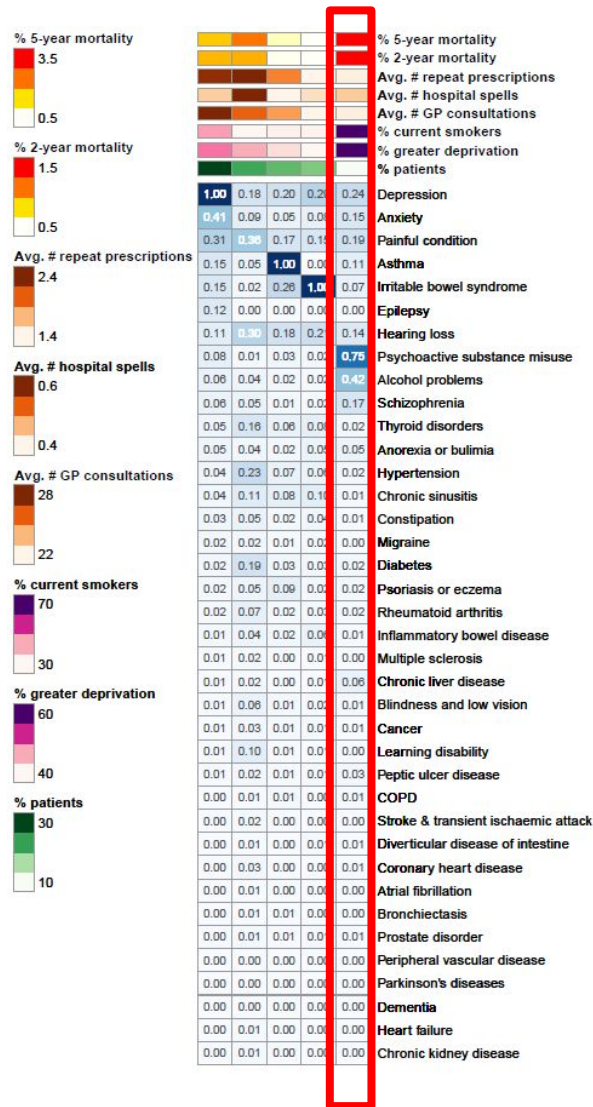
Which item distinguishes clusters the most → IRT

¹Hagenaars, J. A., & McCutcheon, A. L. (Eds.). (2002). *Applied latent class analysis*. Cambridge University Press.

²Prados-Torres, A., Calderón-Larranaga, A., Hanco-Saavedra, J., Poblador-Plou, B., & van den Akker, M. (2014). Multimorbidity patterns: A systematic review. *Journal of Clinical Epidemiology*, 67(3), 254–266

³Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(4), 535–569.

Clustering solution - example

















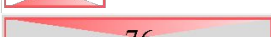
18-44 age strata

- Top panel = service use / mortality outcomes & key patient characteristics. Darker shades = higher values
- Bottom panel (blue) = probabilistic profile for each subgroup (5 in total)
 - Column = subgroup, row = disease
 - Cell value = probability (prevalence) of each disease in each group
 - e.g. column 1 = “depression” group, column 3 = “Asthma” group.
 - Last column = smallest in size but with highest mortality & service use

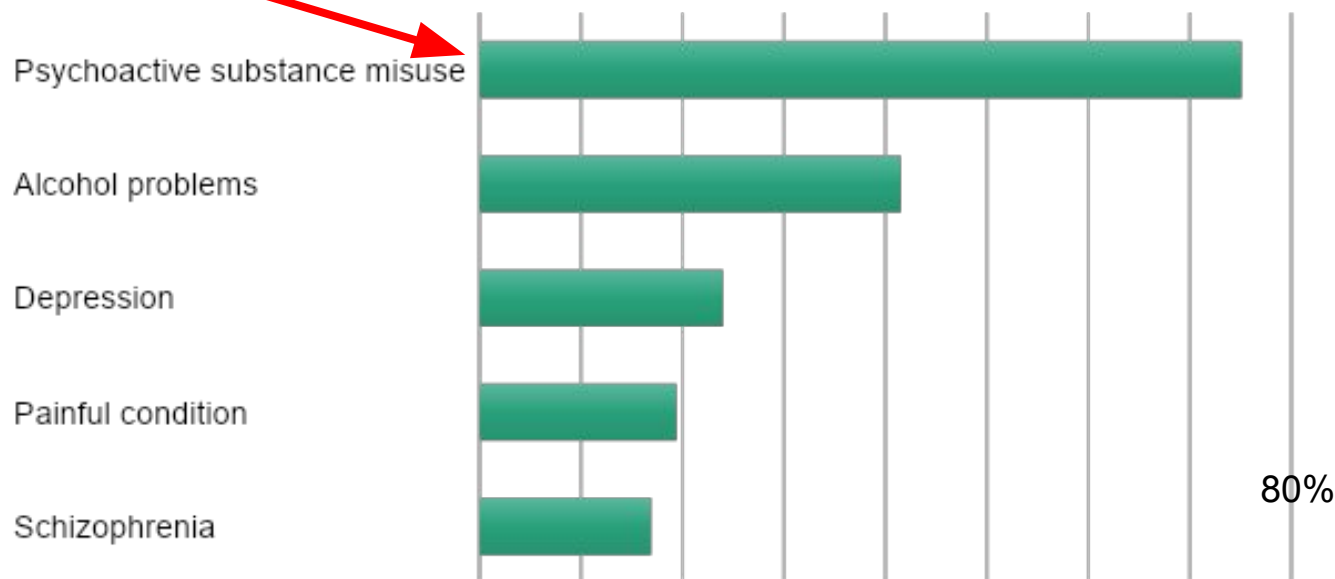
Clustering solution

Three key conditions (prevalence)		Patients (%)	No. of morbidities (median [Q1–Q3])	Female (%)	Greater deprivation (%)	Current smokers (%)
Lead condition (%)	Subsidiary conditions (%)					
Age 18–44 years						
Depression (100%)	Anxiety (41%), pain (31%)	32	2 [2–3]	66	50	46
Pain (36%)	Hearing loss (30%), hypertension (23%)	23	2 [2–3]	52	46	27
Asthma (100%)	IBS (26%), depression (20%)	20	2 [2–3]	63	41	29
IBS (100%)	Depression (29%), hearing loss (21%)	18	2 [2–3]	77	37	28
PSM (75%)	Alcohol (42%), depression (24%)	7	2 [2–3]	28	63	76
Age 45–64 years						
Hypertension (76%)	Diabetes (37%), pain (25%)	37	2 [2–3]	42	38	20
IBS (40%)	Hearing loss (29%), pain (28%)	24	2 [2–3]	64	29	20
Depression (93%)	Pain (53%), anxiety (31%)	22	3 [2–5]	68	46	35
Asthma (100%)	Pain (24%), COPD (16%)	12	2 [2–3]	61	35	20
Alcohol (62%)	PSM (42%), pain (28%)	4	3 [2–4]	31	57	63
Age 65–84 years						
Hypertension (100%)	Diabetes (31%), pain (27%)	41	3 [2–4]	54	30	10
Hearing loss (40%)	Prostate disorder (21%), IBS (3%)	22	3 [2–4]	48	25	9
Depression (56%)	Pain (56%), anxiety (23%)	14	4 [3–5]	72	33	15
CHD (54%)	Diabetes (32%), atrial fibrillation (29%)	11	4 [3–5]	30	33	12
COPD (57%)	Asthma (49%), pain (33%)	8	3 [2–5]	50	40	24
Pain (81%)	CHD (53%), depression (45%)	5	7 [7–9]	54	43	16
Age 85+ years						
Hypertension (72%)	Hearing loss (39%), diabetes (18%)	58	3 [2–4]	61	30	5
Pain (64%)	Depression (41%), constipation (24%)	23	5 [4–6]	80	30	5
CHD (61%)	Atrial fibrillation (53%), heart failure (49%)	11	7 [6–8]	60	30	4
Asthma (48%)	COPD (48%), pain (44%)	8	5 [4–6]	59	30	8

Results - deep dive :18-44 year olds

Lead condition	Multimorbid patients	Greater deprivation	Current smokers	5-year mortality	# GP contacts in a year
(%)	(%)	(%)	(%)	(%)	(Median [Q1-Q3])
Non-multimorbid				0.2	1 [0-5]
Depression (100%)	 32	 50	 46	1.8	12 [5-20]
Pain (36%)	 23	 46	 27	2.7	9 [3-17]
Asthma (100%)	 20	 41	 29	0.6	9 [4-16]
IBS (100%)	 18	 37	 28	0.4	8 [3-15]
PSM (75%)	 7	 63	 76	3.9	7 [1-16]

Most prevalent 5 conditions in the **highest** mortality cluster (PSM)

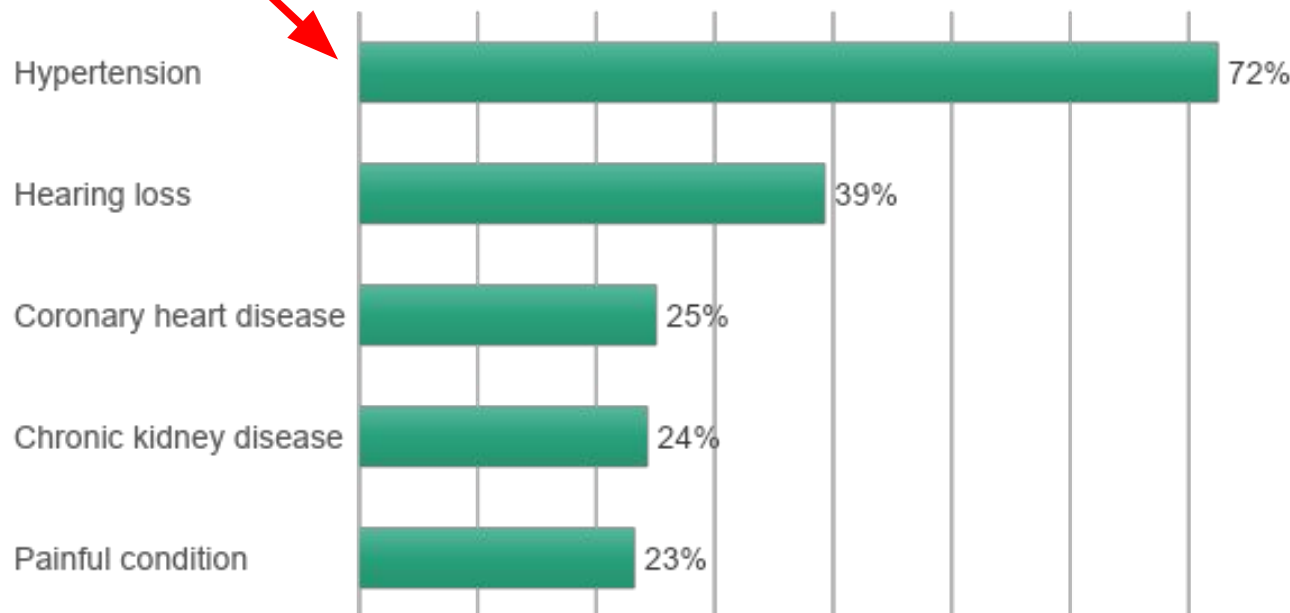


Survival effect?

Results - deep dive :85+ year olds

Lead condition	Multimorbid patients	Greater deprivation	Current smokers	5-year mortality	# GP contacts in a year
(%)	(%)	(%)	(%)	(%)	(Median [Q1-Q3])
Non-multimorbid				36.0	0 [0-2]
Hypertension (72%)	58	30	5	49.5	4 [0-6]
Pain (64%)	23	30	5	62.9	7 [4-10]
CHD (61%)	11	30	4	70.8	8 [5-11]
Asthma (48%)	8	30	8	56.5	7 [4-10]

Most prevalent 5 conditions in the **lowest** mortality cluster (HYP)



Validation of results

- Cannot directly validate clustering solutions (clusters are “unobserved”)
- **Indirect approach:** Similarity of multimorbidity clusters (training vs test sets)
 - similar **relationship** between clusters and patient demographics and outcomes
 - consistency of the **quality of cluster solutions** (using entropy)
 - consistency between **cluster profiles** (using Jensen-Shannon distance & Pearson’s correlation coefficient)

Discussion

- Novel and comprehensive mapping of multimorbidity cluster profiles across age spectrum
- Validated cluster solutions in a representative English multimorbid population
- Evidence-based policy implications for highlighted patient groups:
 - Justified the **push for parity of physical and mental health** within the healthcare system
 - Improve outcomes of **younger multimorbid patients** with **psychoactive substance misuse** given that risk factors (drug use, smoking, deprivation) are amenable to intervention
 - **Chronic pain** may be better managed within the context of multimorbidity rather than in its own right
- Results may be further strengthened by validation in external databases (CPRD Aurum)

Appendix: stability of MM clusters

We employed three methods to indirectly validate our cluster solutions (a direct approach was not possible as clusters were unobserved). First, to check the consistency between disease profiles for 38 LTCs in the training and test sets, each cluster in the test set was matched (using two criteria for robustness) with a corresponding cluster in the training set. Matched cluster pairs were selected such that Jensen–Shannon distance [27] (JSD; a measure of the divergence between disease profiles) is the smallest and the bivariate Pearson’s correlation coefficient [28] (the degree to which two disease profiles co-vary) is the highest (Additional file 2: tables 4a, b). Second, entropy measures [25] (for classification quality) computed in the training and test sets were expected to be similar. Finally, stability was further assessed by observing in the training and test sets similar associations (in terms of size, direction and statistical significance) between clusters, patient demographics and outcome variables. For more details, see Additional file 2: section 4.