MRC Biostatistics Unit UNIVERSITY OF CAMBRIDGE

Recent progress on multimorbidity clustering research

Yajng Zhu

MRC Biostatistics Unit

December 9, 2019

Outline

- Summary of current work
  (doi: https://doi.org/10.1101/19000422)
  Characteristics, service use and mortality of multimorbidity patients across the age spectrum
  Joint work: Dr. Duncan Edwards, Prof. Jonathan Mant, Dr. Rupert Payne, Dr. Steven Kiddle

- (Work in progress) methodological work
  Impact of local dependence on mixture models: relating clusters to later outcomes
  Joint work: Dr. Robert Goudie, Prof. Irini Moustaki, (Dr. Brian Tom), Dr. Steven Kiddle

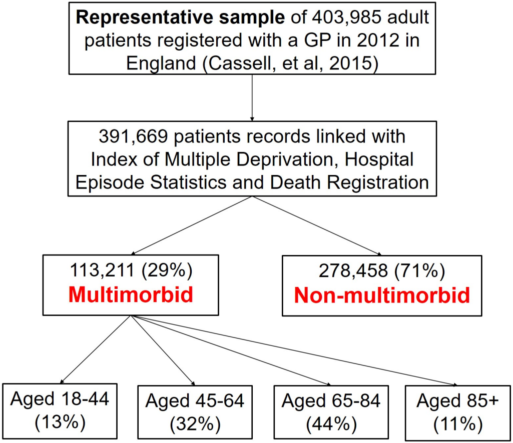
Multimorbidity clusters: RQ

Across the age spectrum

- Which diseases often co-occur (i.e. multimorbidity clusters)?
- Epidemiological profiles associated with each cluster?
- Service use (# GP consultations, # hospital consultations, polypharmacy), 2-year and 5-year mortality?

Multimorbidity clusters: Data

Linked routine primary care data (CPRD-GOLD)
38 chronic conditions (Cambridge codelist, also used in Barnett et al., 2012, Lancet)



Representative sample of 403,985 adult patients registered with a GP in 2012 in England (Cassell, et al, 2015)

391,669 patients records linked with Index of Multiple Deprivation, Hospital Episode Statistics and Death Registration

113,211 (29%) **Multimorbid**

278,458 (71%) **Non-multimorbid**

Aged 18-44 (13%)

Aged 45-64 (32%)

Aged 65-84 (44%)

Aged 85+ (11%)

Multimorbidity clusters: Previous work vs our contribution

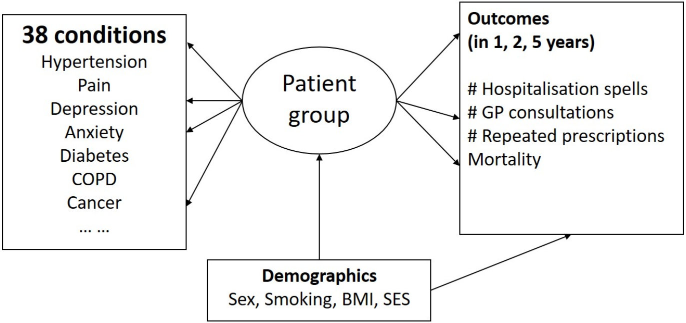
Limitation of previous work

1. Focused on 65+ populations & limited inclusion of long-term conditions (<20)
2. Mostly focused on grouping conditions, not patients
3. Factor analysis (EFA, PCA), hierarchical clustering: factor rotation / subjective choices of ”distance measures”
4. Lack of validation of solutions

Our solutions

1. Age-stratified analysis (18-44, 45-64, 65-84, 85+)
2. Patient-centred and model-based clustering approach: latent class analysis (LCA)
3. Develop clusters using 80% of sample, check cluster stability on 20% of sample
4. Three indirect validation schemes

Multimorbidity clusters: modelling strategy



38 conditions
Hypertension
Pain
Depression
Anxiety
Diabetes
COPD
Cancer
… …

Patient group

Outcomes
(in 1, 2, 5 years)

Hospitalisation spells
GP consultations
Repeated prescriptions
Mortality

Demographics
Sex, Smoking, BMI, SES

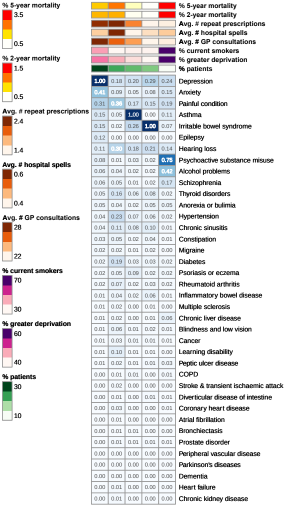
Multimorbidity clusters: validation strategy

Across the training and test set, we checked

1. Consistency between disease probabilistic profiles: can clusters in the test set (fewer clusters, due to fewer patterns) be matched to a cluster in the training set (more clusters, used comprehensive patterns)?
   - Jensen-Shannon divergence measure (divergence between profiles)
   - Bivariate Pearson's correlation coefficient (the degree to which two disease profiles are associated)
2. Similar associations (in terms of size, direction and statistical significance) between clusters and patient demographics?
3. Between cluster and outcomes (service use, mortality)?

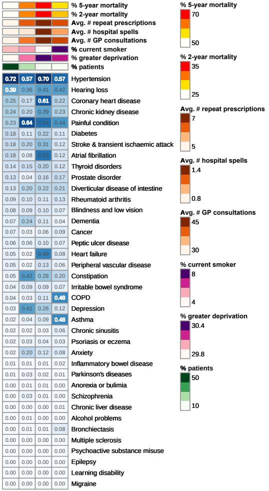
Multimorbidity clusters: examples of full results

18-44 year olds



1.00	0.18	0.20	0.29	0.24	Depression
0.41	0.09	0.05	0.08	0.15	Anxiety
0.31	0.36	0.17	0.15	0.19	Painful condition
0.15	0.05	1.00	0.00	0.11	Asthma
0.15	0.02	0.26	1.00	0.07	Irritable bowel syndrome
0.12	0.00	0.00	0.00	0.00	Epilepsy
0.11	0.30	0.18	0.21	0.14	Hearing loss
0.08	0.01	0.03	0.02	0.75	Psychoactive substance misuse
0.06	0.04	0.02	0.02	0.42	Alcohol problems
0.06	0.05	0.01	0.02	0.17	Schizophrenia
0.05	0.16	0.06	0.08	0.02	Thyroid disorders
0.05	0.04	0.02	0.05	0.05	Anorexia or bulimia
0.04	0.23	0.07	0.06	0.02	Hypertension
0.04	0.11	0.08	0.10	0.01	Chronic sinusitis
0.03	0.05	0.02	0.04	0.01	Constipation
0.02	0.02	0.01	0.02	0.00	Migraine
0.02	0.19	0.03	0.03	0.02	Diabetes
0.02	0.05	0.09	0.02	0.02	Psoriasis or eczema
0.02	0.07	0.02	0.03	0.02	Rheumatoid arthritis
0.01	0.04	0.02	0.06	0.01	Inflammatory bowel disease
0.01	0.02	0.00	0.01	0.00	Multiple sclerosis
0.01	0.02	0.00	0.01	0.06	Chronic liver disease
0.01	0.06	0.01	0.02	0.01	Blindness and low vision
0.01	0.03	0.01	0.01	0.01	Cancer
0.01	0.10	0.01	0.01	0.00	Learning disability
0.01	0.02	0.01	0.01	0.03	Peptic ulcer disease
0.00	0.01	0.01	0.00	0.01	COPD
0.00	0.02	0.00	0.00	0.00	Stroke & transient ischaemic attack
0.00	0.01	0.00	0.01	0.01	Diverticular disease of intestine
0.00	0.03	0.00	0.00	0.01	Coronary heart disease
0.00	0.01	0.00	0.00	0.00	Atrial fibrillation
0.00	0.01	0.01	0.00	0.00	Bronchiectasis
0.00	0.01	0.01	0.01	0.01	Prostate disorder
0.00	0.00	0.00	0.00	0.00	Peripheral vascular disease
0.00	0.00	0.00	0.00	0.00	Parkinson's diseases
0.00	0.00	0.00	0.00	0.00	Dementia
0.00	0.01	0.00	0.00	0.00	Heart failure
0.00	0.01	0.00	0.00	0.00	Chronic kidney disease

85+ year olds

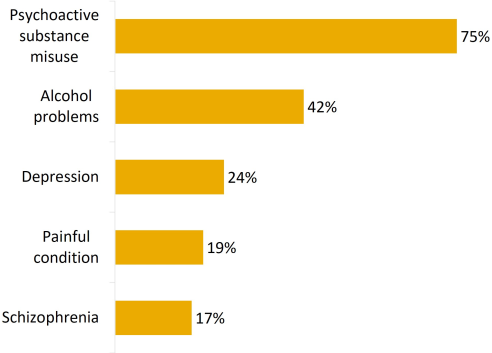


0.72	0.57	0.70	0.57	Hypertension
0.39	0.36	0.41	0.42	Hearing loss
0.25	0.17	0.61	0.22	Coronary heart disease
0.24	0.20	0.39	0.23	Chronic kidney disease
0.23	0.64	0.54	0.44	Painful condition
0.18	0.11	0.22	0.11	Diabetes
0.18	0.20	0.31	0.12	Stroke & transient ischaemic attack
0.18	0.08	0.53	0.12	Atrial fibrillation
0.14	0.15	0.20	0.12	Thyroid disorders
0.13	0.04	0.16	0.17	Prostate disorder
0.13	0.20	0.22	0.21	Diverticular disease of intestine
0.09	0.10	0.11	0.13	Rheumatoid arthritis
0.08	0.09	0.10	0.07	Blindness and low vision
0.07	0.24	0.11	0.04	Dementia
0.07	0.03	0.06	0.09	Cancer
0.06	0.06	0.10	0.07	Peptic ulcer disease
0.05	0.02	0.49	0.08	Heart failure
0.05	0.02	0.13	0.06	Peripheral vascular disease
0.05	0.42	0.28	0.20	Constipation
0.04	0.09	0.09	0.07	Irritable bowel syndrome
0.04	0.03	0.11	0.48	COPD
0.03	0.41	0.26	0.12	Depression
0.02	0.04	0.09	0.48	Asthma
0.02	0.02	0.03	0.06	Chronic sinusitis
0.02	0.03	0.04	0.03	Psoriasis or eczema
0.02	0.20	0.12	0.08	Anxiety
0.01	0.01	0.01	0.02	Inflammatory bowel disease
0.01	0.03	0.02	0.01	Parkinson's diseases
0.00	0.01	0.01	0.00	Anorexia or bulimia
0.00	0.03	0.01	0.00	Schizophrenia
0.00	0.00	0.00	0.01	Chronic liver disease
0.00	0.01	0.00	0.00	Alcohol problems
0.00	0.01	0.01	0.08	Bronchiectasis
0.00	0.00	0.00	0.00	Multiple sclerosis
0.00	0.00	0.00	0.00	Psychoactive substance misuse
0.00	0.00	0.00	0.00	Epilepsy
0.00	0.00	0.00	0.00	Learning disability
0.00	0.00	0.00	0.00	Migraine

Multimorbidity clusters: 18-44 year olds

5-year mortality of non-multimorbid peers: 0.2%

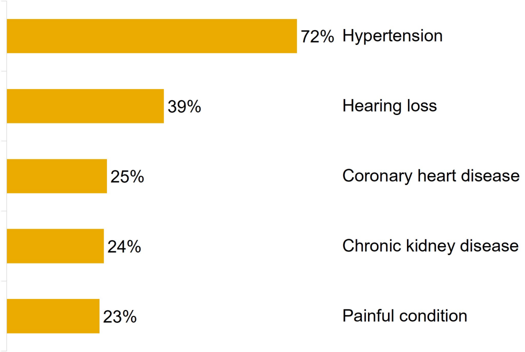
Lead condition	Multimorbid patients	Greater deprivation	Current smokers	5-year mortality
(%)	(%)	(%)	(%)	(%)
Depression (100%)	32	50	46	1.8
Pain (36%)	23	46	27	2.7
Asthma (100%)	20	41	29	0.6
IBS (100%)	18	37	28	0.4
PSM (75%)	7	63	76	3.9



Multimorbidity clusters: 85+ year olds

5-year mortality of non-multimorbid peers: 36%

Lead condition	Multimorbid patients	Greater deprivation	Current smokers	5-year mortality
(%)	(%)	(%)	(%)	(%)
Hypertension (72%)	58	30	5	49.5
Pain (64%)	23	30	5	62.9
CHD (61%)	11	30	4	70.8
Asthma (48%)	8	30	8	56.5



Multimorbidity clusters: Policy implications

- Supports the push for parity of physical and mental health within the healthcare system
- Unmet need to improve outcomes of younger multimorbid patients with psychoactive substance misuse given that risk factors (drug use, smoking, deprivation) are amenable to intervention
- The majority of 85+ year old multimorbid patients have relatively low service use and mortality
- Pain features in 13/20 clusters - treatment of pain should be put in the context of multimorbidity

Relating clusters to later outcomes

Common approach: modal-class approach

- Develop cluster models for patients
- Assign patients to most likely clusters
- Compare profiles of risk factors and outcomes

Accurate clustering $\neq$ accurate prediction of outcomes

Impact of local dependence on mixture models: relating clusters to later outcomes

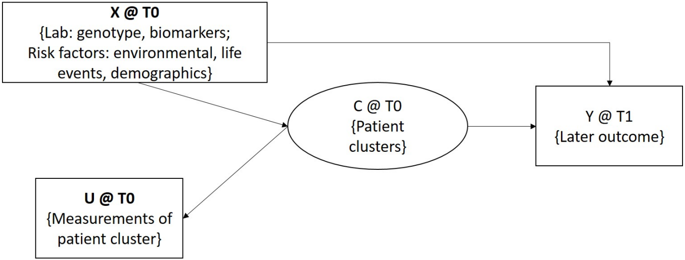
Notation: vector of measurements ($\mathbf{U} = \{U_1, U_2, ..., U_p\}$), patient cluster($C$), baseline covariate set ($\mathbf{X}$), outcome (Y, e.g. death), cluster-specific outcome ($Y|C$), patient-specific outcome ($Y|\mathbf{U}, \mathbf{X}$).

Problem

- $\mathbf{U} \perp\!\!\!\perp Y|C$ rarely holds in real data → Misclassification/biased coefficients/poor prediction.
- Inference ($Y|C$ & C) & prediction goals (Y) are often mixed
- Ambiguity in the definition of clusters:
  - What is a ”subpopulation”, does it really exist?
  - How do we want to use the derived clusters?
    (just a proxy for population heterogeneity/carry clinical meaning?)

Relating clusters to later outcomes: simplification of the real world

Data-generating mechanism


X @ T0
{Lab: genotype, biomarkers;
Risk factors: environmental, life events, demographics}
C @ T0
{Patient clusters}
Y @ T1
{Later outcome}
U @ T0
{Measurements of patient cluster}

Assumptions on the framework

- *C* exists in the true world
- We want *C* to reduce heterogeneity in disease patterns (i.e. measurements), not in ”people”
- There is a temporal order between quantities

Relating clusters to later outcomes: RQs (1)

All these questions are consistent with the data-generating mechanism

1. (Inference) Recover true C @ T0 and correct $Y|C$. C @ T0 should not be influenced by later outcomes
   - $C =$ baseline frailty, allows for target-treat each homogeneous group at baseline
   - $C =$ actual state of a condition (e.g. depression), not easily defined using a single measurement
   - Once baseline C is correctly recovered, policy wants to target on demographic factors in the associated Xs
   - $C =$ clusters of co-occuring diseases that share common biological pathways, improve understanding of disease development

   Modal-class, bias-corrected 3-step approaches

Relating clusters to later outcomes: RQs (2)

2. (Inference) Recover ”a type of patient cluster” based on how they respond to treatment (outcome-guided clustering). Patients share within-cluster-homogeneity in terms of risk factors for Y.
   - $C =$ ”prone-to-death” or ”low-risk” groups for targeted treatments

   1-step approach
3. (Prediction) Do not care about C even when it exists, only cares about Y.

   1-step approach, regression
4. (Inference & Prediction) Care about both the correctness of C, $Y|C$ and Y

   1-step/bias-corrected 3-step/modal class approaches

Relating clusters to later outcomes: summary of methods

- 1-step approach (RQ2,RQ3, RQ4)
  - $C \sim f(\mathbf{U}, Y)$
- Modal class approach (RQ1, RQ4)
  - $C \sim f(\mathbf{U})$, assign to class W, $Y \sim W$
- Bias-corrected 3-step (RQ1,RQ4)
  - $C \sim f(\mathbf{U})$, assign to class W
  - $Y \sim C$; weight observations by inverse of classification error

$$\begin{aligned}\log L_{BCH} &= \sum_{i=1}^{N}\sum_{k=1}^{K}\sum_{s=1}^{K}\omega_{is}d_{sk}\log P(C_i = k, Y = y_i),\\ \omega_{is} &= P(W_i = s|\mathbf{U}_i = \mathbf{u}_i),\\ d_{sk} &= [P(W = s|C = k)]^{-1}_{KxK}\end{aligned}$$

- Regression (RQ3)
  - $Y \sim f(\mathbf{U})$

Relating clusters to later outcomes: simulation study

- Entropy (distinguishablility of the cluster) → $\mathbf{U}|C$

$$E_k = 1 - \frac{\sum_1^N \sum_{k=1}^K [-\hat{p}_{ik} log(\hat{p}_{ik})]}{NlogK},$$
$$\hat{p}_{ik} = p(C_i = k|\mathbf{U}_i)$$

- Cluster distribution (2-class): 60/40 vs 95/5
- CIA holds vs mild violation: $(U_5, Y) \sim f(C, \eta)$

Scenarios

- N=10,000, high (0.9)/low (0.5) entropy, balanced/imbalanced cluster = 4 settings
- Set A: binary Us, CIA holds
- Set B: binary Us, CIA does not hold
- Set C: mixed types of Us, CIA does not hold

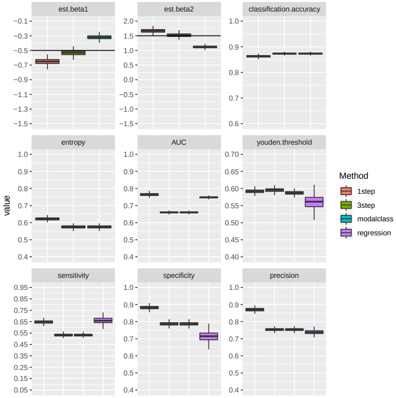
Relating clusters to later outcomes: consequence of local dependence (1)

Performance metrics

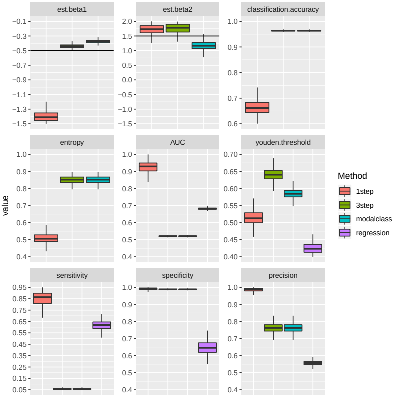
- (Inference) $\hat{\beta}_1$, $\hat{\beta}_2$, classification accuracy
- (Prediction) AUC, sensitivity, specificity, precision, youden's threshold

Relating clusters to later outcomes: consequence of local dependence (2)

Mixed-type Us, low entropy, balanced *C*, CIA fails



Mixed-type Us, low entropy, imbalanced *C*, CIA fails



Relating clusters to later outcomes: consequence of local dependence (3)

Findings

- Across all scenarios, 1-step gives best predictive performance
- Across all scenarios, bias-corrected 3-step gives best inference (mostly unbiased, good coverage probability)
- What if we want both inference/predictive goals? Local dependence correction!

Relating clusters to later outcomes: correction of local dependence (1)

General approaches

- Uniform association model (categorical Y only)

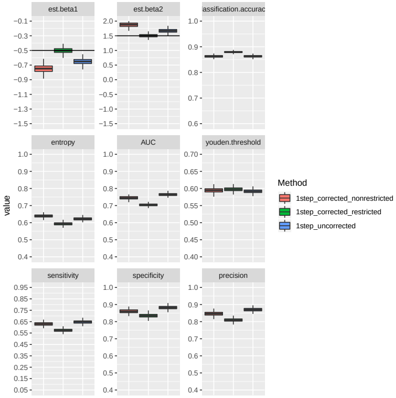
$$P(U_1 = a_1, U_2 = a_2, \dots Y = a_{p+1} | C = k) = \frac{\exp(\sum_i \tau_{i,a_i,k} + \sum_{i<j} \beta_{ij,k} a_i a_j)}{\sum_{a_1,a_2,\dots,a_{p+1}} \exp(\sum_i \tau_{i,a_i,k} + \sum_{i<j} \beta_{ij,k} a_i a_j)}$$

- Latent variable approach (flexible, integration)

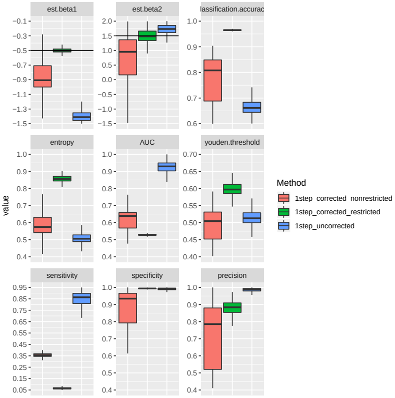
$$P(U_1 = a_1, U_2 = a_2, \dots Y = a_{p+1} | C = k, \eta)$$

Relating clusters to later outcomes: correction of local dependence (2)

Mixed-type Us, low entropy, balanced *C*, CIA fails



Mixed-type Us, low entropy, imbalanced *C*, CIA fails



Relating clusters to later outcomes: correction of local dependence (3)

Findings

- Educated (restricted) correction works best
  - 1-step $C \sim f(\mathbf{U}, Y)$
  - Categorical Y: check pairwise bivariate residuals & correct for ”naughty pairs”

$$\frac{(O - E)}{\sqrt{E(1 - E/O)}}$$

  - Continuous Y: fit non-restricted latent variable model, LRT, fit restricted model.
- After correction: minor sacrifice on prediction, substantial gain on inference.

Final recommendations

Stop using modal class approach for either inference/prediction goals! As CIA is likely to fail:

- RQ1: to recover baseline C and correct $Y|C$
  - bias-corrected 3-step, check for local dependence (needs future developments)
- RQ2,4: to recover outcome-guided cluster/good inference & prediction
  - 1-step with restricted correction
- RQ3: only interested in accurate Y
  - 1-step approach/regression

Next steps

1. Application: multimorbidity (under RQ1-4)
2. Hanging question: what if C does not exist?
   - Philosophical understanding of the world
   - What is the clinical goal?

References:

- Bias-correction stepwise approaches: Vermunt (2010); Asparouhov & Muthen & Bakk (2013, 2014, 2015,2016,2018)
- Residual association: Goodman (1979); Hagennars (1988), Madison & Vermunt (2004); Asparouhov & Muthen (2014)

Thanks for listening!

MRC Biostatistics Unit

UNIVERSITY OF CAMBRIDGE

yajing.zhu@mrc-bsu.cam.ac.uk

www.mrc-bsu.cam.ac.uk