Ala-Korpela, M. (2019). Data-driven subgrouping in epidemiology and medicine. *International Journal of Epidemiology*, *48*(2), 374-376. [dyz040]. https://doi.org/10.1093/ije/dyz040

Data-driven subgrouping in epidemiology and medicine

Mika Ala-Korpela

1. Systems Epidemiology, Baker Heart and Diabetes Institute, Melbourne, VIC, Australia
2. Computational Medicine, Faculty of Medicine, University of Oulu and Biocenter Oulu, Oulu, Finland
3. NMR Metabolomics Laboratory, School of Pharmacy, University of Eastern Finland, Kuopio, Finland
4. Medical Research Council Integrative Epidemiology Unit at the University of Bristol, Bristol, UK
5. Population Health Science, Bristol Medical School, University of Bristol, Bristol, UK
6. Department of Epidemiology and Preventive Medicine, School of Public Health and Preventive Medicine, Faculty of Medicine, Nursing and Health Sciences, The Alfred Hospital, Monash University, Melbourne, VIC, Australia

Professor Mika Ala-Korpela
Baker Heart and Diabetes Institute, Systems Epidemiology
75 Commercial Road, Melbourne, Victoria 3004, Australia
E-mail: mika.ala-korpela@baker.edu.au
Mobile: +61 452 392 966

No abstract

Main text: 986 words

Figures: 0

Tables: 0

References: 10

There is a growing interest in multiple disciplines for subgrouping people. This would be relevant in integrated health systems to improve the overall health of populations by better targeted and more effective healthcare services.[1] Subgrouping would also be pertinent for scientific applications, for example, finding genetically and metabolically distinct individuals and patients,[2,3] complex disease subtyping[4] and disentangling subgroup specific risk factors and risk assessment.[5,6]

In all the abovementioned applications the statistical approach on subgrouping shares similar characteristics. The amount of data available is huge, e.g., due to increasing use of electronic health care registries and large collections of multiple omics data combined with extensive clinical information in large biobanks. The data are typically continuous and do not instinctively represent subgroups. However, intuitive thinking would be that elaborate utilisation of the extensive multivariate data would lead to better understanding of the complexities in each situation and also potentially better applications. Importantly, also in the translational domain the abovementioned disciplines share fundamental similarities – while individual protocols are not feasible (e.g., in population health care[1]) or fundamentally not reliable (e.g., for the risk prediction of a complex disease[7]) – data-driven holistic approaches, leading to a decent number of characteristic subgroups, pose an interesting and potentially sensible way to improve current one-size-fits-for-all tactics.

In this Issue Mäkinen and co-workers present a Software Application Profile for an open-source R library, titled Numero, which would be a versatile and powerful tool for the abovementioned types of subgrouping needs.[8] Numero provides a three-step framework that combines the self-organizing map (SOM) algorithm, permutation analyses for statistical evidence and an expert-driven final subgrouping decision.[5,8] Numero can handle both continuous and categorical variables in situations where there is no intrinsic clustering in the data and it creates data-driven statistically validated two-dimensional visualisations without explicit boundaries for the subgroups. This adds unavoidable complexity to the use of the SOM algorithm for subgrouping and calls for substance experts in each application for balanced interpretations. Nonetheless, the subgrouping can also be done based on certain pertinent quantitative rules; as an example, in the case of a SOM application to define metabolic risk for subclinical atherosclerosis, the median of carotid intima-media thickness

was taken to define the borderline between high and low risk and then the high risk area in the SOM was concomitantly divided into subgroups based on the 90th percentile of serum triglycerides and the upper tertile of low-density lipoprotein cholesterol to arrive at three metabolic subgroups representing high risk.[9]

Open-source, easy-to-use R library for SOM analyses in epidemiology and medicine is timely since intriguing subgrouping applications in various areas of medicine have started to appear in the literature. I will use diabetes as an exemplar here.

In fact, Mäkinen *et al.* have recently also presented a reassuring application of SOM analysis in the area of type 1 diabetes.[6] In this new work they revisited their own data and SOM analyses they published earlier in 2008.[5] In their new paper[6] they analysed new clinical endpoints and looked at the predictive characteristics of the earlier derived[5] six data-driven metabolic subtypes after seven years of additional follow-up. They successfully replicated an epidemiological multivariable model, based on the SOM algorithm, across two time periods by using pre-defined data-driven baseline subtypes with new prospective data.[6] This study focused on epidemiological aspect of data-driven subtyping and thus the authors strongly cautioned translational implications. Notably, regardless of statistical methods, prediction of vascular endpoints at an individual level from general biochemical data is not possible.[7,10] This is not only because of analytical and standardisation issues between study centres and countries, but importantly also due to natural heterogeneity and continuity of polygenetic diseases and outcomes,[10] as well as due to inherent randomness of physiological events.[7] Nevertheless, there is considerable overall value in applying the SOM algorithm in large-scale population cohorts to reveal subgroups that may not be covered by traditional approaches or accurately addressed by current treatment guidelines. Mäkinen *et al.*[6] demonstrated this for patients with type 1 diabetes: baseline metabolic subtypes, derived from an array of quantitative biomarkers, contained a wealth of diagnostic and prognostic information of potential value for public health interventions, e.g., informing on a subset of women with high high-density lipoprotein cholesterol who would still be at high risk for cardiovascular endpoints.

In another recent work, Groop and co-workers[3] applied k-means and hierarchical clustering analyses of six diabetes-related variables measured at diagnosis in adults with newly diagnosed diabetes, and arrived at five different data-driven clusters of patients that had different characteristics and risk of diabetic complications. These included a cluster of very insulin-resistant individuals with higher risk of diabetic kidney disease than the other clusters, a cluster of relatively young insulin deficient individuals with poor metabolic control, and a large group of elderly patients with the most benign disease course.[3] In an independent work, Florez *et al.*[2] utilised a Bayesian clustering analysis to genetic variants associated with type 2 diabetes. These analyses also identified five clusters with distinct trait associations, two relating to insulin deficiency and three to insulin resistance. The authors showed that these clusters were differentially enriched for relevant tissue-specific enhancers and promoters. Florez *et al.*[2] discussed their findings with respect to those by Groop *et al.*[3] and suggested some correspondence between the independently specified data-driven clusters even though the data used were fundamentally different, i.e., genetic loci[2] versus clinical and biomarker data at the time of diabetes diagnosis.[3]

The examples of SOM analyses demonstrated by Mäkinen and co-workers,[5,6,8] as well as the recent other clustering applications in diabetes;[2,3] suggest alluring potential for data-driven subgroup analyses in epidemiology and medicine. These types of approaches are likely to inform on the fundamental genetic and metabolic variation defining the complexity of polygenic diseases. Through increased molecular understanding of the complexities, and realistic expert-guided assessments, this route of population and patient subgrouping may lead to improved risk assessment and treatment models. Numero[8] should definitely encourage endeavours towards this direction in large-scale epidemiological data sets.

Acknowledgements

MAK works in a Unit that is supported by the University of Bristol and UK Medical Research Council (MC_UU_12013/1). The Baker Institute is supported in part by the Victorian Government[1]s Operational Infrastructure Support Program.

Disclosure

The author reports no conflicts of interest.

References

1. Yan S, Kwan YH, Tan CS, Thumboo J, Low LL. A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Med Res Methodol* 2018;**18**:68.

2. Udler MS, Kim J, Grotthuss von M, *et al*. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: A soft clustering analysis. *PLoS Med* 2018;**15**:e1002654.

3. Ahlqvist E, Storm P, Käräjämäki A, *et al*. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *Lancet Diabetes Endocrinol* 2018;**6**:361–9.

4. Mäkinen V-P, Kangas AJ, Soininen P, Würtz P, Groop P-H, Ala-Korpela M. Metabolic phenotyping of diabetic nephropathy. *Clin Pharmacol Ther* 2013;**94**:566–9.

5. Mäkinen V-P, Forsblom C, Thorn LM, *et al*. Metabolic phenotypes, vascular complications, and premature deaths in a population of 4,197 patients with type 1 diabetes. *Diabetes* 2008;**57**:2480–7.

6. Lithovius R, Toppila I, Harjutsalo V, *et al*. Data-driven metabolic subtypes predict future adverse events in individuals with type 1 diabetes. *Diabetologia* 2017;**60**:1234–43.

7. Davey Smith G. Epidemiology, epigenetics and the 'Gloomy Prospect': embracing randomness in population health research and practice. *Int J Epidemiol* 2011;**40**:537–62.

8. Gao S, Mutter S, Casey A, Mäkinen V-P. Numero: a statistical framework to define multivariable subgroups in complex population-based datasets. *Int J Epidemiol* 2018; in press.

9. Würtz P, Soininen P, Kangas AJ, *et al*. Characterization of systemic metabolic phenotypes associated with subclinical atherosclerosis. *Mol Biosyst* 2011;**7**:385–93.

10. Ala-Korpela M, Davey Smith G. Metabolic profiling-multitude of technologies with great research potential, but (when) will translation emerge? *Int J Epidemiol* 2016;**45**:1311–8.