How to handle big data for disease stratification in respiratory medicine?

Krasimira Tsaneva-Atanasova [1], Chris Scotton [2]

Increasingly complex datasets of biomedical measurements offer an opportunity for discovering patient endotypes. These represent subtypes of a disease marked by distinct pathomechanisms—which can have enormous implications for prognosis and clinical management. Such datasets often include imaging, genomics and transcriptomics, proteomics, microbiotal composition, allergen/environmental exposures, and immunological data—as well as patient outcomes and routinely collected clinical parameters. Given the sheer volume of data, interpretation is extremely challenging.

Recently, topological data analysis (TDA) has been rapidly gaining in popularity for application to such datasets (see Skaf and Laubenbacher[1] for review). Respiratory medicine is no exception, as topology offers a suite of techniques and tools that could be applied to diverse data. This enables a holistic approach to robustly identify multidimensional properties and relationships within a given multimodal dataset, by using the full range of available clinical and pathobiological data simultaneously. Topological methods also naturally lend themselves to visualisation, rendering them useful for applications that require user interpretation and understanding.

TDA offers a more unbiased and rigorous approach to analysing complex datasets, since it does not depend on prior hypotheses nor focus on pairwise relationships within the data. This contrasts with other established analytical methods, such as supervised clustering and classical association analyses. The Mapper algorithm[2] is a popular technique in TDA that converts a complex dataset with many dimensions into a simpler network representation embedded in a lower number of dimensions. To achieve this, common techniques such as principal components analysis (PCA), t-distributed stochastic neighbour embedding and uniform manifold approximation and projection (UMAP) could be employed to reduce the dimensionality of the data. The latter has certainly gained notoriety in light of the plethora of single cell RNAseq data currently in circulation.

Specifically, the Mapper algorithm starts by applying a projection (eg, UMAP) to the data set and using the projection to bin the data into overlapping bins. This is then followed by clustering each bin using a fixed clustering method (requires data equipped with metric) creating a node for each 'partial cluster' and an edge between any two nodes whose corresponding clusters overlap. In the context of patients' data, each 'node' represents a group of patients comprising subjects 'similar to' each other in the multiple measurements (eg, gene expression profiles) on which the TDA is based. Links or 'edges' in the network represent individual patients that are shared between nodes (groups of patients). Once the TDA network has been created, further clustering and statistical measures and techniques can be applied to investigate network substructures (communities) and/or reveal emergent features of the dataset. Importantly, TDA provides a geometric representation of the data, allowing visualisation and ready identification of meaningful subgroups and complex relationships/signals within the data. Indeed, the combination of TDA and other computational approaches is often more effective in summarising, analysing, quantifying and visualising medical data for complex disease stratification.[3]

TDA is not without limitations, however, including accessibility and validation, dependence on fine-tuning of parameters and robustness as well as computational complexity.[1] The advanced theoretical foundations of TDA make it less accessible for biomedical scientists and clinicians. Consequently, it requires further validation in a broader range of complex biomedical datasets and clinical applications. A related challenge is that in TDA applications, the user needs to carefully choose the parameters depending on the properties of the specific dataset under investigation. These choices depend on the nature of the data and do necessitate understanding of the theoretical basis of TDA. Finally, TDA may be very computationally intensive especially when dealing with very large data sets and higher dimensional projections (in more than two dimensions). In such cases, a careful preprocessing of the data may be required in order to alleviate the computational burden.

In respiratory medicine, TDA has already been applied to conditions such as asthma,[4–11] chronic obstructive pulmonary disease (COPD),[6 12 13] primary ciliary dyskinesia (PCD)[14] and for primary lung tumour classification.[15] Specifically, TDA has enabled robust identification of subgroups of asthmatic patients with distinct clinical and pathological characteristics.[5] Furthermore, as the pathological airway changes in asthma are heterogeneous, TDA is demonstrably useful for helping to identify patient subgroups and characterise transitions between healthy and diseased states.[4] Based on transcriptomic analysis of airway epithelium from the Unbiased Biomarkers for the Prediction of Respiratory Disease Outcomes (U-BIOPRED) study, TDA has also enabled identification of two subtypes of asthma—one with high expression of Th2 cytokines and the other with decreased corticosteroid response.[7] Blood transcriptomic data have also been used with TDA to further reveal phenotypic subtypes in asthma.[11]

In the context of COPD, TDA was recently applied as a validation of COPD patient stratification obtained using hierarchical clustering of fungal allergens, and additionally revealed a subgroup within *Aspergillus*-sensitised COPD of patients with frequent exacerbations.[13] In a large-scale study of primary ciliary dyskinesia patients, TDA has confirmed genotype–phenotype relationships reported by smaller studies and identified new relationships.[14] Based on exhaled breath metabolites, TDA successfully stratified patients presenting with breathlessness due to severe exacerbations of cardiorespiratory aetiology (asthma, COPD, heart failure or pneumonia) and healthy controls,[6] thus demonstrating the potential diagnostic utility. Finally, in the context of imaging data, a TDA analysis involving persistent homology on geometric features extracted from chest CTs was able to classify COPD patients more precisely compared with conventional radiographic measurements.[12] A similar TDA approach applied to classifying primary lung tumours based on CT scan has shown that topological features may improve radiomic-based histology prediction.[15]

The study by Shapanis *et al* in Thorax used the TDA Mapper algorithm for stratification of idiopathic pulmonary fibrosis (IPF) patients.[16] IPF disease progression is known to be heterogeneous,

[1]Mathematics and Statistics and Living Systems Institute, University of Exeter, Exeter, UK
[2]Medical School, University of Exeter, Exeter, UK

Correspondence to Professor Krasimira Tsaneva-Atanasova, Mathematics and Statistics and Living Systems Institute, University of Exeter, Exeter, Devon, UK; k.tsaneva-atanasova@exeter.ac.uk

which significantly hinders precise and timely diagnosis and prognosis, as well as patient-centred treatment choice.[17] It has been hypothesised that the observed heterogeneity in IPF disease progression could be, at least in part, due to different molecular endotypes among the IPF patient population. In their study, TDA served as a tool in support of this hypothesis.

The authors combined publicly available blood transcriptomic gene expression data from five datasets into a single dataset that was labelled using three classes, namely 'IPF' patients, 'healthy' and 'other' for machine learning classification. It is worth noting that this resulted in a rather unbalanced training as well as test dataset, where the number of 'IPF' patients and 'healthy' individuals is approximately four times smaller than 'other'. The proportion of each class relative to the others, however, was roughly preserved in the training and test datasets. Class imbalance is a common problem in real-world datasets that poses a challenge for predictive modelling. This is because most machine learning algorithms used for classification were designed around the assumption of an equal number of samples (in this case individuals) for each class. The predictive performance could, therefore, be compromised in the case of imbalanced classification, specifically for the minority class as the problem is more sensitive to classification errors for the minority class than the majority class. The TDA analysis, however, does not suffer from sensitivity to imbalanced data sets and as shown in the study robustly identifies five IPF subphenotypes with an even distribution of samples from each IPF dataset across the network. In contrast, a recent study by Kraven *et al*[18] using PCA, identified three clusters (subphenotypes) of IPF patients based on six independent whole blood gene expression datasets. Both studies by Shapanis *et al*[16] and Kraven *et al*[18] mapped clinical features on to the clusters and observed phenotypes with significantly different diffusing capacity of the lung for carbon monoxide. The TDA-constructed network identified a cluster (subphenotype 5) characterised by higher likelihood of death/transplant compared with the others. Pathway and upstream regulator analysis was applied to reveal differentially expressed genes between each IPF subphenotype, thus supporting the molecular nature of the classification in general agreement with.[18] Finally, the authors presented evidence in support of a distinct immune cell profile associated with each IPF subphenotype.

The findings reported by Shapanis *et al*[16] and Kraven *et al*[18] are broadly consistent. It is worth noting that both studies had a number of datasets in common—highlighting how extremely similar data may be analysed in different ways for differing purposes; Shapanis *et al*[16] identify a 44 gene transcriptional signature which has diagnostic utility, while Kraven *et al*[18] generated a 13 gene signature with prognostic potential. Employing different methodologies to identify subgroups of IPF patients may, therefore, explain the different numbers of subgroups and specific differentially expressed genes identified. Furthermore, in contrast to PCA, TDA does not assume linear relationships in the data and encodes continuous variation while clustering, by definition, does not. The continuous variation in the data, encoded by TDA, allows for the identification of members (in this case patients) who belong to more than one of the clusters.

Another recent study[19] identified six IPF endotypes based on publicly available gene expression datasets retrieved from bronchoalveolar lavage (BAL) samples. This study was more limited in the availability of longitudinal clinical parameters with which to map against the endotypes, in contrast to.[16] In addition, the validation of the endotypes was carried out using blood transcriptomic datasets since no other BAL transcriptomic datasets were available; two of those datasets were also used by Shapanis *et al*[16]. It is important to note that these tissue compartments (BAL and blood) clearly have distinct cellular compositions, which will have significant influence—ideally the BAL endotypes would benefit from validation in further BAL gene expression datasets. It would also be beneficial to conduct further evaluations of blood transcriptome profiles in patients who have simultaneous BAL gene expression assessment, to demonstrate overlap of endotypes between these tissue compartments, particularly as peripheral blood is readily accessible while BAL is not performed in all patients.

As evidenced by previous work in asthma, COPD, PCD, lung cancer and now IPF, the TDA approach to high-dimensional complex data sets offers great promise for generating highly relevant patient endotypes/subtypes. The identification of distinct subphenotypes in IPF patients by Shapanis *et al*, with characterisation according to differences in both clinical characteristics and molecular composition, represents another stepping stone towards more precise IPF disease prognosis and therapy. Incorporation of such approaches in larger cohorts or clinical trial data may, therefore, benefit our understanding (and prediction) of treatment responses, leading to improvements in clinical management. It is also reassuring that different methods have yielded somewhat consistent results thereby suggesting the potential of employing a combination of these in the future to draw more robust and reliable conclusions. The methodology of choice (and source data) will ultimately depend on which proves to be of most clinical utility (and cost-effectiveness) in the real world.

Twitter Krasimira Tsaneva-Atanasova @KrasiTsaneva and Chris Scotton @isteefer

Contributors KT-A and CS wrote the editorial.

Funding KT-A gratefully acknowledges the financial support of the EPSRC via grant EP/T017856/1. CS is supported by MRC grants MR/V002538/1 and MR/W014491/1.

Competing interests None declared.

Patient consent for publication Not applicable.

Provenance and peer review Commissioned; internally peer reviewed.

To cite Tsaneva-Atanasova K, Scotton C. *Thorax* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/thorax-2023-220138

Accepted 30 March 2023

Thorax 2023;**0**:1–3. doi:10.1136/thorax-2023-220138

ORCID iDs
Krasimira Tsaneva-Atanasova http://orcid.org/0000-0002-6294-7051
Chris Scotton http://orcid.org/0000-0002-9671-9057

REFERENCES

1 Skaf Y, Laubenbacher R. Topological data analysis in biomedicine: a review. *J Biomed Inform* 2022;130:104082.
2 Singh G, Mémoli F, Carlsson GE. Topological methods for the analysis of high dimensional data sets and 3d object recognition. *PBG Eurographics* 2007;2:091–100.
3 De Meulder B, Lefaudeux D, Bansal AT, *et al*. A computational framework for complex disease stratification from multiple large-scale datasets. *BMC Syst Biol* 2018;12:60.
4 Siddiqui S, Shikotra A, Richardson M, *et al*. Airway pathological heterogeneity in asthma: visualization of disease microclusters using topological data analysis. *J Allergy Clin Immunol* 2018;142:1457–68.
5 Hinks TSC, Zhou X, Staples KJ, *et al*. Innate and adaptive t cells in asthmatic patients: relationship to severity and disease mechanisms. *J Allergy Clin Immunol* 2015;136:323–33.

6 Ibrahim W, Wilde MJ, Cordell RL, *et al*. Visualization of exhaled breath metabolites reveals distinct diagnostic signatures for acute cardiorespiratory breathlessness. *Sci Transl Med* 2022;14.
7 Kuo C-HS, Pavlidis S, Loza M, *et al*. A transcriptome-driven analysis of epithelial brushings and bronchial biopsies to define asthma phenotypes in U-BIOPRED. *Am J Respir Crit Care Med* 2017;195:443–55.
8 Östling J, van Geest M, Schofield JPR, *et al*. IL-17-high asthma with features of a psoriasis immunophenotype. *J Allergy Clin Immunol* 2019;144:1198–213.
9 Schofield JPR, Bigler J, Boedigheimer M, *et al*. Topological data analysis (TDA) of U-BIOPRED paediatric peripheral blood gene expression identified asthma phenotypes characterised by alternative splicing of glucocorticoid receptor (GR) mrna. ERS International Congress 2018 abstracts; September 15, 2018:suppl
10 Schofield JPR, Burg D, Nicholas B, *et al*. Stratification of asthma phenotypes by airway proteomic signatures. *Journal of Allergy and Clinical Immunology* 2019;144:70–82.
11 Bigler J, Boedigheimer M, Schofield JPR, *et al*. A severe asthma disease signature from gene expression profiling of peripheral blood from U-BIOPRED cohorts. *Am J Respir Crit Care Med* 2017;195:1311–20.
12 Belchi F, Pirashvili M, Conway J, *et al*. Lung topology characteristics in patients with chronic obstructive pulmonary disease. *Sci Rep* 2018;8:5341.
13 Tiew PY, Narayana JK, Quek MSL, *et al*. Sensitisation to recombinant *aspergillus fumigatus* allergens and clinical outcomes in COPD. *Eur Respir J* 2023;61:2200507.
14 Shoemark A, Rubbo B, Legendre M, *et al*. Topological data analysis reveals genotype-phenotype relationships in primary ciliary dyskinesia. *Eur Respir J* 2021;58:2002359.
15 Vandaele R, Mukherjee P, Selby HM, *et al*. Topological data analysis of thoracic radiographic images shows improved radiomics-based lung tumor histology prediction. *Patterns (N Y)* 2023;4:100657.
16 Shapanis A, Jones MG, Schofield J, *et al*. Topological data analysis identifies molecular phenotypes of idiopathic pulmonary fibrosis. *Thorax* 2023:thorax-2022-219731.
17 Selman M, King TE, Pardo A, *et al*. Idiopathic pulmonary fibrosis: prevailing and evolving hypotheses about its pathogenesis and implications for therapy. *Ann Intern Med* 2001;134:136–51.
18 Kraven LM, Taylor AR, Molyneaux PL, *et al*. Cluster analysis of transcriptomic datasets to identify endotypes of idiopathic pulmonary fibrosis. *Thorax* 2022:thoraxjnl-2021-218563.
19 De Sadeleer LJ, Verleden SE, Schupp JC, *et al*. BAL transcriptomes characterize idiopathic pulmonary fibrosis endotypes with prognostic impact. *Chest* 2022;161:1576–88.