Topological data analysis reveals genotype–phenotype relationships in primary ciliary dyskinesia

Amelia Shoemark[1,2,30], Bruna Rubbo[3,4,30], Marie Legendre[5,6], Mahmoud R. Fassad[7,8], Eric G. Haarman[9], Sunayna Best[7,10], Irma C.M. Bon[9], Joost Brandsma[4], Pierre-Regis Burgel[11,12], Gunnar Carlsson[13], Siobhan B. Carr[1], Mary Carroll[3,4], Matt Edwards[14], Estelle Escudier[5,6], Isabelle Honoré[11], David Hunt[15], Gregory Jouvion[5,6], Michel R. Loebinger[16,17], Bernard Maitre[18,19], Deborah Morris-Rosendahl[14], Jean-Francois Papon[20,21,22,23], Camille M. Parsons[24], Mitali P. Patel[7], N. Simon Thomas[25,26], Guillaume Thouvenin[4,27,28], Woolf T. Walker[3,4], Robert Wilson[16], Claire Hogg[1], Hannah M. Mitchison[7,29,31] and Jane S. Lucas[3,4,31]

[1]PCD Diagnostic Centre and Dept of Paediatric Respiratory Medicine, Royal Brompton and Harefield NHS Trust, London, UK. [2]Division of Molecular and Clinical Medicine, University of Dundee, Ninewells Hospital and Medical School, Dundee, UK. [3]Primary Ciliary Dyskinesia Centre, University Hospital Southampton NHS Foundation Trust, Southampton, UK. [4]School of Clinical and Experimental Sciences, University of Southampton Faculty of Medicine, Southampton, UK. [5]Département de Génétique Médicale, Hôpital Trousseau, Assistance Publique–Hôpitaux de Paris (AP–HP), Paris, France. [6]Sorbonne Université, Institut National de la Santé et de la Recherche Médicale (INSERM) U933, Hôpital Trousseau, Paris, France. [7]Genetics and Genomic Medicine Dept, University College London, UCL Great Ormond Street Institute of Child Health, London, UK. [8]Dept of Human Genetics, Medical Research Institute, Alexandria University, Alexandria, Egypt. [9]Dept of Pediatric Pulmonology, Emma Children's Hospital, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. [10]Leeds Institute of Medical Research, Faculty of Medicine and Health, University of Leeds, Leeds, UK. [11]Service de Pneumologie, Hôpital Cochin, Assistance Publique–Hôpitaux de Paris (AP–HP), Paris, France. [12]Université de Paris, Institut National de la Santé et de la Recherche Médicale (INSERM) U1016, Institut Cochin, Paris, France. [13]Dept of Mathematics, Stanford University, Stanford, CA, USA. [14]Clinical Genetics and Genomics, Royal Brompton and Harefield NHS Foundation Trust, London, UK. [15]Wessex Clinical Genetics Service, University Hospitals Southampton, Princess Anne Hospital, Southampton, UK. [16]Host Defence Unit, Dept of Respiratory Medicine, Royal Brompton and Harefield NHS Foundation Trust, London, UK. [17]National Heart and Lung Institute (NHLI), Imperial College, London, UK. [18]Service de Pneumologie, DHU A-TVB, Centre Hospitalier Intercommunal de Créteil, Université Paris Est, Créteil, France. [19]Université Paris Est, Institut National de la Santé et de la Recherche Médicale (INSERM) U955, Institut Mondor de Recherche Biomédicale (IMRB), Créteil, France. [20]Service d'ORL et Chirurgie Cervico-Faciale, Hôpital Kremlin-Bicêtre, Assistance Publique–Hôpitaux de Paris (AP–HP), Le Kremlin-Bicêtre, France. [21]Faculté de Médecine, Université Paris-Saclay, Le Kremlin-Bicêtre, France. [22]Centre national de la recherche scientifique (CNRS) ERL 7240, Créteil, France. [23]Institut National de la Santé et de la Recherche Médicale (INSERM) U955, Créteil, France. [24]Medical Research Council (MRC) Lifecourse Epidemiology Unit, University of Southampton, Southampton, UK. [25]Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury, UK. [26]Human Genetics and Genomic Medicine, University of Southampton Faculty of Medicine, Southampton, UK. [27]Service de Pneumologie Pédiatrique, Hôpital Trousseau, Assistance Publique–Hôpitaux de Paris (AP–HP), Paris, France. [28]Sorbonne Université, Institut National de la Santé et de la Recherche Médicale (INSERM) U938, Centre de Recherche Saint-Antoine, Paris, France. [29]National Institute for Health Research (NIHR) Great Ormond Street Hospital Biomedical Research Centre, London, UK. [30]Equal first author contribution. [31]H.M. Mitchison and J.S. Lucas contributed equally to this article as lead authors and supervised the work.

Corresponding author: Jane Lucas (jlucas1@soton.ac.uk)

Shareable abstract (@ERSpublications)
Topological data analysis of 396 primary ciliary dyskinesia patients shows genetic mutations of worse (*CCDC39*), variable (*DNAH5*) and milder (*DNAH11*) effects on lung function, offering the potential for more accurately targeted disease management https://bit.ly/3oL5r64

Cite this article as: 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 [DOI: 10.1183/13993003.02359-2020].

Copyright ©The authors 2021. For reproduction rights and permissions contact permissions@ersnet.org

Abstract

Background Primary ciliary dyskinesia (PCD) is a heterogeneous inherited disorder caused by mutations in approximately 50 cilia-related genes. PCD genotype–phenotype relationships have mostly arisen from small case series because existing statistical approaches to investigating relationships have been unsuitable for rare diseases.

This article has supplementary material available from erj.ersjournals.com

This article has an editorial commentary: https://doi.org/10.1183/13993003.00392-2021

Received: 16 June 2020
Accepted: 24 Dec 2020

Methods We applied a topological data analysis (TDA) approach to investigate genotype–phenotype relationships in PCD. Data from separate training and validation cohorts included 396 genetically defined individuals carrying pathogenic variants in PCD genes. To develop the TDA models, 12 clinical and diagnostic variables were included. TDA-driven hypotheses were subsequently tested using traditional statistics.

Results Disease severity at diagnosis, measured by forced expiratory volume in 1 s (FEV_1) z-score, was significantly worse in individuals with *CCDC39* mutations (compared to other gene mutations) and better in those with *DNAH11* mutations; the latter also reported less neonatal respiratory distress. Patients without neonatal respiratory distress had better preserved FEV_1 at diagnosis. Individuals with *DNAH5* mutations were phenotypically diverse. Cilia ultrastructure and beat pattern defects correlated closely to specific causative gene groups, confirming these tests can be used to support a genetic diagnosis.

Conclusions This large scale, multi-national study presents PCD as a syndrome with overlapping symptoms and variations in phenotype according to genotype. TDA modelling confirmed genotype–phenotype relationships reported by smaller studies (*e.g.* FEV_1 worse with *CCDC39* mutation) and identified new relationships, including FEV_1 preservation with *DNAH11* mutations and diversity of severity with *DNAH5* mutations.

Introduction

Primary ciliary dyskinesia (PCD) is clinically and genetically heterogeneous. Symptoms relate to dysfunction of multiple motile cilia and can include neonatal respiratory distress syndrome (NRDS), wet cough, recurring upper and lower respiratory tract infections, otitis media, bronchiectasis, infertility, *situs inversus* and congenital heart disease (CHD) [1]. Mutations in 50 ciliary genes have been described so far [2, 3].

Understanding of genotype–phenotype relationships informs diagnostic decisions and treatment; however, due to the rarity (≈1:10 000) and diversity of PCD, and the constraints of traditional statistical methods, a large patient cohort has never been studied for these relationships. Evidence for clinically relevant genotype–phenotype associations is mostly limited to small case series for a specific gene or clinical characteristic. For example, individuals with variants in *HYDIN* (a central pair projection gene), or in multiciliogenesis genes like *MCIDAS* and *CCNO*, are unlikely to have *situs inversus* as nodal cilia are not affected [4–7]. Using traditional statistical approaches, cohort studies have been underpowered to investigate by single gene and have instead combined functionally similar genes for analysis. A North American study of 137 children reported worse lung disease in those patients with central apparatus or microtubular disorganisation (MTD) (with inner dynein arm (IDA) ultrastructural defects), most of whom had *CCDC39* and *CCDC40* variants, than in those patients with outer dynein arm (ODA) defects caused by *DNAH5* variants [8, 9].

Topological data analysis (TDA) allows for the visual exploration of data without establishing *a priori* hypotheses [10]. It can be used to explore the underlying patterns in complex datasets by generating clusters of individuals with similar features in multiple dimensions, in an unsupervised manner, as extensively validated in several clinical studies [11–13]. TDA can be used to highlight small groups of interest in large or complex datasets, which could otherwise be overlooked when applying traditional clustering methods that are typically more constrained by a requirement for pre-selection of parameters (*e.g.* definition of the number of clusters) to drive data analyses [10, 14]. In doing so, TDA can uncover patient subgroups more likely to benefit from a particular therapeutic intervention [12, 15–17]. It thereby provides a promising approach to investigate genotype–phenotype associations in heterogeneous patients with rare diseases.

Our aim was to investigate relationships between clinical, diagnostic and genetic data, hypothesising that different subgroups of PCD patients, with particular clinical and diagnostic phenotypes, could be identified according to their underlying genotypes.

Methods

Ethics

Local and national research approvals and ethical approvals were obtained and adhered to (NRES Committee South Central Hampshire Ethics 06/Q1702/109, London Bloomsbury Research Ethics Committee 08/H0713/82 and Ile-de-France Ethics Committee CPP07729).

Study Design

Clinical and diagnostic data were retrospectively collected from patients with a confirmed genetic diagnosis of PCD (*i.e.* carrying autosomal bi-allelic variants or an X-linked variant classified as pathogenic

according to international guidelines) [18, 19]. The data coding for the clinical characteristics included in the study is shown in supplementary table E1.

The study design was based on previous TDA studies and is outlined in figure 1 [15]. TDA was performed in order to generate hypotheses, which could then be tested using more traditional statistical methods. TDA was applied to a discovery cohort of 199 patients (cohort details and genetics can be found in supplementary tables E2–E4) and validated using a second cohort of 197 patients (cohort details and genetics can be found in supplementary figure E1 and supplementary tables E5 and E6). An overview of the PCD genes affected by mutations in the full study population is shown in supplementary figure E2.

Topological data analysis

Topological models were developed using a licensed version of TDA software through the Symphony AyasdiAI cloud-based platform, version 2.0 (Ayasdi Inc., Menlo Park, CA, USA; www.ayasdi.com). More details of TDA are available in the supplementary material.

The phenotypic data used for clustering were body mass index (BMI), forced expiratory volume in 1 s (FEV_1) z-score, forced vital capacity (FVC) z-score, neonatal respiratory distress (NRDS), wet cough, rhinitis, glue ear, cardiac situs, congenital heart disease (CHD), nasal nitric oxide (nNO), ciliary beat pattern (CBP) and transmission electron microscopy (TEM). Genetic data were not used to generate the topological models, as these were the study's main variable of interest; however, genes of interest were later mapped onto the models to develop hypotheses regarding genotype–phenotype associations.

Models were generated using an automated analysis option. Locally-linear embedding (LLE) is a non-linear dimensionality reduction technique, in which highly complex data are summarised and compressed into smaller representations of their variability. The topological model with the best-defined clusters upon visual inspection used two LLE lenses and the correlation distance as a metric (*i.e.* a distance function). These identical parameters were applied to develop the discovery and validation models.

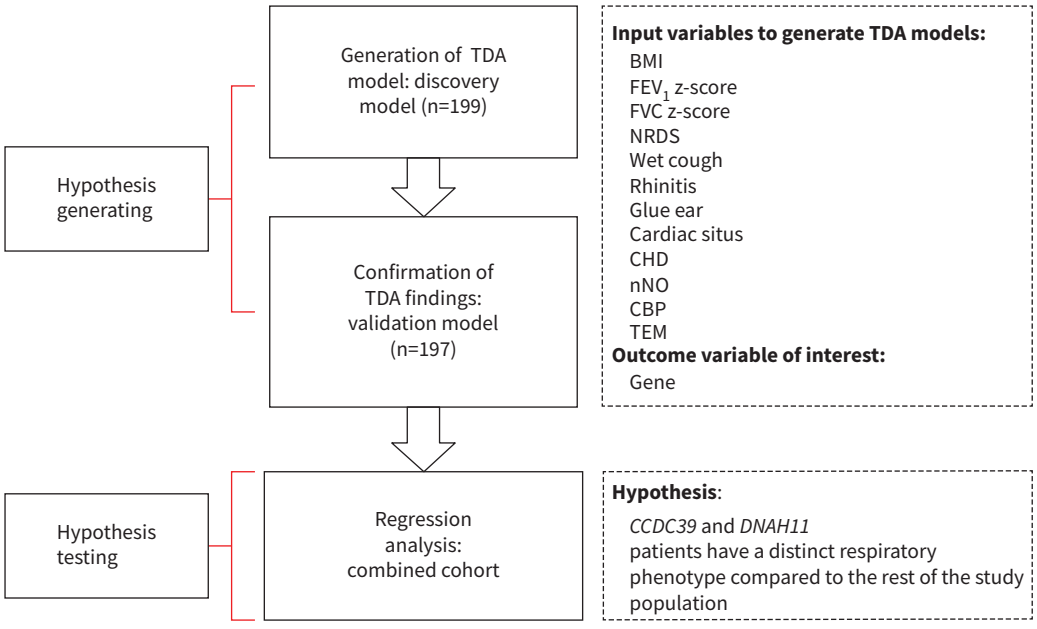

FIGURE 1 Study design. Topological data analysis (TDA) models were used to identify clusters of clinical and diagnostic characteristics. Gene groups and individual genes were mapped onto these clusters to develop hypotheses, which could subsequently be tested using traditional statistical approaches such as ANOVA. Without the use of TDA, comparison of forced expiratory volume in 1 s (FEV_1) across more than 20 genes would require multiple comparisons and statistical power would be lost, whereas using this method we were able to directly test a single directed hypothesis. FVC: forced vital capacity; BMI: body mass index; NRDS: neonatal respiratory distress syndrome; CHD: congenital heart disease; nNO: nasal nitric oxide; CBP: ciliary beat pattern; TEM: transmission electron microscopy.

The Mapper algorithm was used to identify coherent groups of samples [20]. Each node of the topology model constitutes patients who have combinations of features that are similar between them, with connecting lines (edges) representing data points that are shared between nodes. The size of the node represents the number of subjects with that specific combination of features. Genotypes were mapped onto the model to visualise hypothesised associations between genotype and phenotypic clusters. Validation of hypotheses suggested by the TDA was then performed using standard statistical analysis techniques. Generating hypotheses using TDA prevented the requirement for multiple comparisons and loss of statistical power.

TDA is an effective method to apply in clinical studies as it can allow for missing data [21]. A more detailed explanation of TDA can be found in the supplementary material.

Statistical analysis

Selection of variables for hypothesis testing was guided by the topological models to limit the number of comparisons. Further methodological details are provided in the supplementary material.

The derived hypotheses were tested through statistical analyses of the whole dataset and of the validation dataset alone. Where the same outcome was tested twice, p-values were adjusted using the Bonferroni

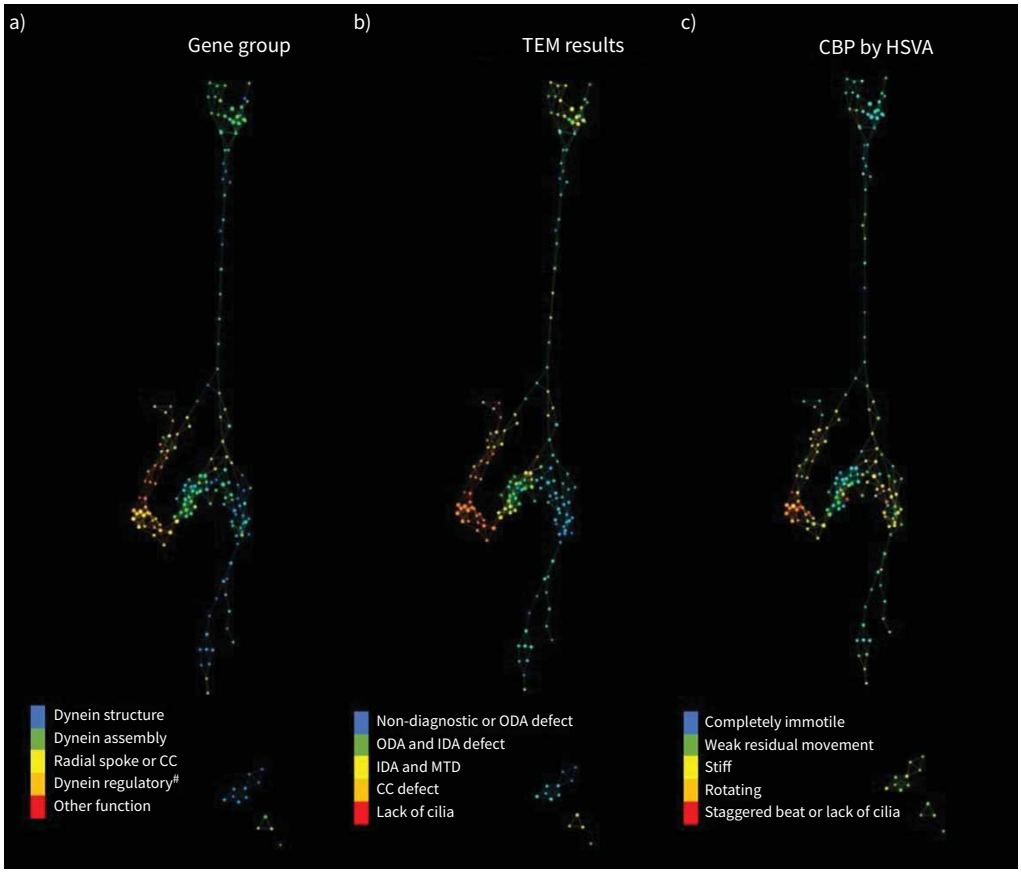

FIGURE 2 Topological discovery model. Topology analysis display of the results of unbiased clustering of several levels of data, here showing the connections amongst the patients according to their underlying gene defect and the resulting cilia structure and motility defect. Each node represents combinations of features. The size of the nodes represents the number of subjects. The connections represent that there are patients shared between the two nodes. Models are coloured by the following features: a) gene group; b) transmission electron microscopy (TEM) results; and c) ciliary beat pattern (CBP) by high-speed video analysis (HSVA). Within each of the three models, patients are grouped according to five different classes of gene, TEM and CBP, respectively. CC: central complex; ODA: outer dynein arm; IDA: inner dynein arm; MTD: microtubular disorganisation. #: the nexin–dynein regulatory complex (N-DRC)/molecular ruler group.

correction ($p \leq 0.049$ was found to be significant). Continuous data were compared using t-tests, ANOVA and Kruskal–Wallis tests, while categorical data were compared using Chi-squared or Fisher's exact tests. Tukey's test was used for pairwise comparisons following ANOVA and Dunn's test with Holm–Sidak adjustment was used following Kruskal–Wallis tests. Multiple regression models were used to model FEV_1 z-scores, adjusting for age at diagnosis, history of NRDS and presence of CHD. Normality of residuals was investigated using kernel density estimations, and visual inspection of histograms and residuals *versus* fits graph plots. Number of observations (n), regression coefficients (r) with 95% confidence intervals (CIs) and a model's goodness-of-fitness (adjusted R^2) were reported for each model. Data were analysed in STATA version 14.0 (StataCorp, College Station, TX, USA).

Results

Data-driven genotype–phenotype associations using TDA in a discovery group of 199 PCD patients

Genotype and diagnostic test phenotype associations

TEM defect and CBP visually mapped very closely to the corresponding gene group (figure 2).

Genotype and FEV_1 associations

Systematic exploration of each of the features collected for this study showed that patients with defects in the radial spoke/central complex (CC) and nexin–dynein regulatory complex (N-DRC)/molecular ruler gene functional groups had worse FEV_1 z-scores at diagnosis (as indicated by dark blue coloured nodes in figure 3b) than those with dynein structural gene mutations (which had higher FEV_1 z-scores, as indicated by white coloured nodes in figure 3b). Interestingly, in the cluster with predominantly poor FEV_1 (dark blue in figure 3b), which corresponds to N-DRC or molecular ruler genes (*CCDC39*, *CCDC40*, *CCDC65*, *DRC1*) (figure 3a), there was a defined group showing absence of history of rhinitis (supplementary figure E3b).

The group with predominantly preserved lung function at diagnosis (white in figure 3b) corresponds to a cluster of individuals with absence of NRDS (white in figure 3c) and an area associated with gene defects of dynein structure (blue in figure 3a). Further exploration of the topological model showed that, within this dynein structural defects group, it was predominantly *DNAH11* patients that had preserved lung function at diagnosis and absence of NRDS (green in figure 3e).

In contrast, individuals with variants in *DNAH5* (the most common genetic cause of PCD and the predominant patient group in the cohort) were a phenotypically diverse group regarding lung function, with no clear cluster observed (figure 3f).

Genotype and other clinical phenotype associations

The model showed a group of patients with CC and N-DRC/molecular ruler gene mutations without *situs inversus* but with increased likelihood of glue ear (yellow and orange in supplementary figure E3a, as well as red in supplementary figure E3c) [7, 22]. In addition, there was a lack of laterality defects associated with *MCIDAS* and *CCNO* in the other function gene group (red in supplementary figure E3d) [6, 23]. Conversely, TDA revealed a cluster of patients with absence of glue ear. This was a genetically diverse group of individuals with dynein structural and assembly defects (blue and green in supplementary figure E3a and white in supplementary figure E3c).

Validation using TDA in a replication group of 197 PCD patients

A topological validation model was generated by analysis of a replication cohort of 197 additional patients (61 from the UK, 28 from The Netherlands and 108 from France) (supplementary tables E5 and E6). This confirmed the discovery group findings, with *CCDC39* mutation patients clustering in an area of the structure with lower FEV_1 z-scores at diagnosis (dark blue in figure 4b and green in figure 4d) and a higher proportion of reported NRDS (red in figure 4c). Meanwhile, *DNAH11* mutation patients clustered in an area with higher FEV_1 z-scores (green in figure 4e and light blue and white in figure 4b) and less reported NRDS (red and white in figure 4c). The model also confirmed the absence of a clear cluster of patients with *DNAH5* mutations (green in figure 4f). Additional features of the validation cohort are shown in supplementary figure E4.

When analysing gene groups, those with mutations in the dynein regulatory/molecular ruler gene category had worse FEV_1 z-scores (orange in figure 4a and dark blue in figure 4b) and less rhinitis (data not shown) at diagnosis, as seen in the discovery model. The cluster with preserved lung function was mostly formed by patients with dynein structure gene variants (light blue and white in figure 4b and blue in figure 4a), particularly *DNAH11* (green in figure 4e). However, we could not confirm the inverse association between

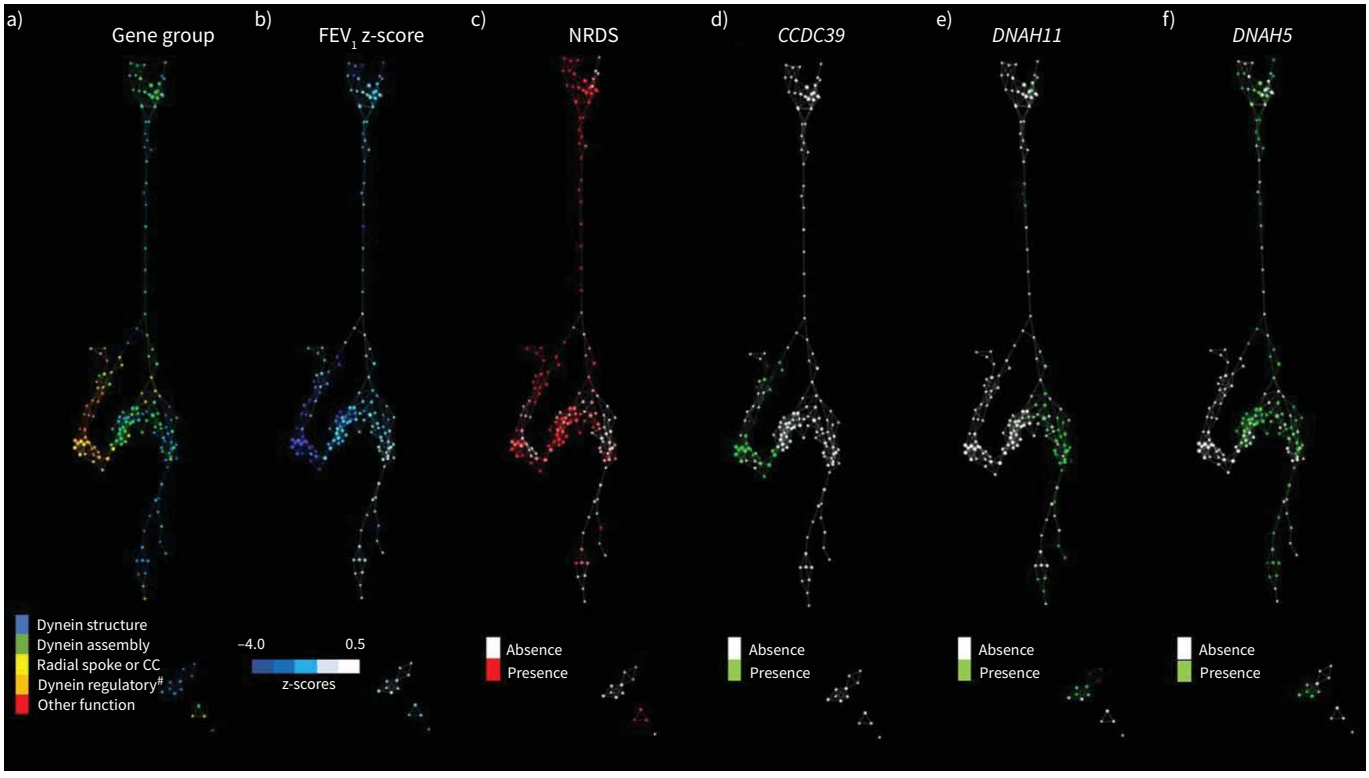

FIGURE 3 Topological discovery model. Each node represents a combination of features. The size of the nodes represents the number of subjects. The connections represent that there are patients shared between the two nodes. Models are coloured by the following features: a) gene group; b) forced expiratory volume in 1 s (FEV_1) z-score; c) neonatal respiratory distress syndrome (NRDS); d) *CCDC39* mutation; e) *DNAH11* mutation; and f) *DNAH5* mutation. CC: central complex. #: the nexin–dynein regulatory complex (N-DRC)/molecular ruler group.

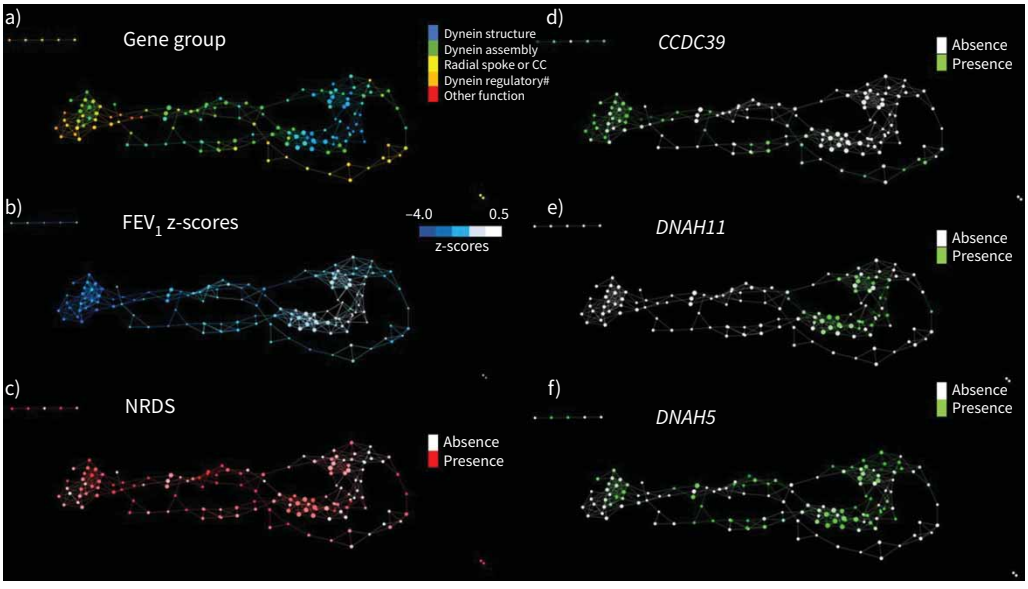

FIGURE 4 Topological validation model. Each node represents a combination of features. The size of the nodes represents the number of subjects. The connections represent that there are patients shared between the two nodes. Models are coloured by the following features: a) gene group; b) forced expiratory volume in 1 s (FEV_1) z-score; c) neonatal respiratory distress syndrome (NRDS); d) *CCDC39* mutation; e) *DNAH11* mutation; and f) *DNAH5* mutation. CC: central complex. #: the nexin–dynein regulatory complex (N-DRC)/molecular ruler group.

upper airway disease (rhinitis and glue ear) and lower airway disease (FEV_1 and NRDS) observed in the discovery model (supplementary figure E4).

The distribution of gene variants in the sum total of 396 patients from both cohorts, in 31 PCD genes, is shown in figure 5 and the clinical and diagnostic characteristics are given in supplementary tables E7 and E8.

Validation of hypothesis suggested by TDA using standard statistical analysis

Two genes, *CCDC39* and *DNAH11*, fulfilled the criteria for further hypothesis-driven statistical analysis. This required the identification of clearly defined clusters of patients, with mutations in each gene, showing distinct features in both the hypothesis-driving discovery (figure 3) and validation (figure 4) topological models. In addition, sufficient patients were required in each phenotype to allow standard statistical approaches (n=35 and n=48, respectively) (figure 5). These two genes clustered in areas with extreme values of FEV_1 z-score in both topological models, leading to the hypothesis that *CCDC39* and *DNAH11* patients had a distinct respiratory phenotype compared to the rest of the study population.

Testing these hypotheses using traditional statistical analysis techniques, *CCDC39* mutation patients had significantly lower FEV_1 z-scores at diagnosis compared to all other patient genotypes grouped together when adjusted for age at diagnosis, NRDS and CHD (r=−1.2, 95% CI −1.88 to −0.55; adjusted R^2=8.0%; p<0.001; n=205). Conversely, those with *DNAH11* mutation had significantly higher FEV_1 z-scores at diagnosis (r=0.09, 95% CI 0.27–1.53; adjusted R^2=5.8%; p=0.003; n=205) and reported less NRDS compared to patients with mutations in any of the other genes (41.03% *versus* 63.91%; p=0.008).

In contrast, there were no statistically significant differences in NRDS for patients with *CCDC39* mutations (67.86% *versus* 60.29% for any of the other genes). Nor were there statistically significant differences in upper airway symptoms (*i.e.* rhinitis and glue ear) for patients with *CCDC39* mutations (96.77%) compared to muations in other genes (93.44%) or in *DNAH11* mutations (97.67%) compared to other genes (93.18%).

Discussion

This is the first large-scale study to systematically investigate associations between genotype and phenotype in the genetically heterogeneous disorder PCD. It demonstrates the use of a new methodology for the visualisation of data and the generation of hypotheses that complements more traditional statistical

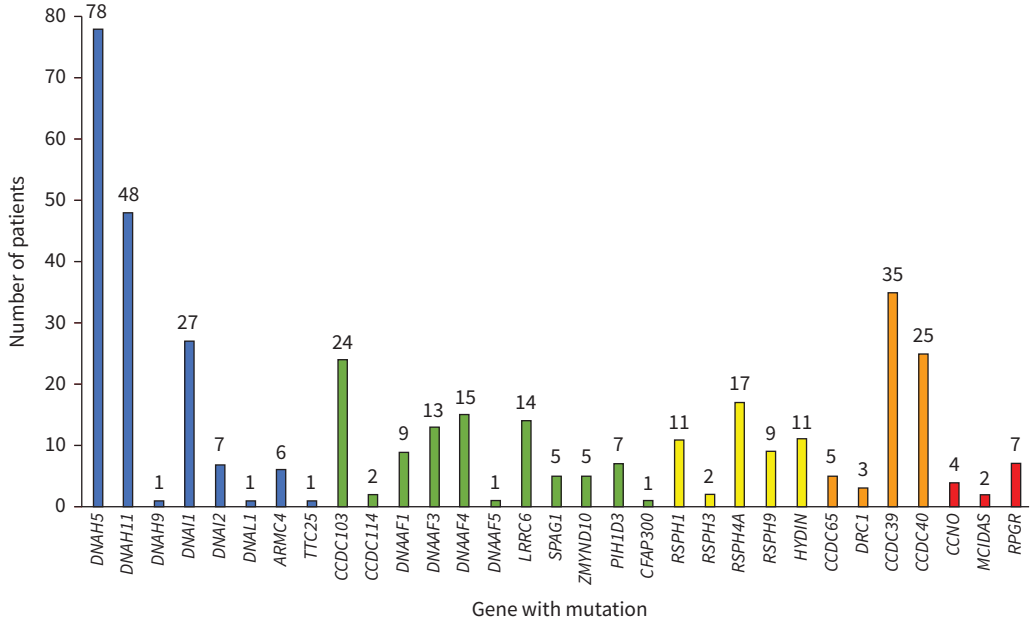

FIGURE 5 Total patient population according to genotype (n=396). Mutations in 31 primary ciliary dyskinesia (PCD) genes were included for analysis. Bars are coloured according to gene group, as follows: dynein structure (blue), dynein assembly (green), radial spoke and central complex (CC) (yellow), nexin–dynein regulatory complex (N-DRC)/molecular ruler (orange) and other functions (*e.g.* ciliogenesis) (red).

approaches, which, where used alone, would not be sufficiently powered even in multinational cohorts. TDA cluster modelling, in nearly 400 individuals from three European countries, identified several previously unknown genotype–phenotype relationships, in addition to confirming previously reported genetic associations [7, 22, 24]. PCD, a disease with many well-defined features and 50 causal genes, leant itself to TDA and machine learning for the identification of distinct phenotypic clusters that might share an underlying genetic mutation. TDA was able to identify clinical patterns amongst relatively small numbers of patients (less than 40) with mutations in a particular gene. We suggest the approach might be beneficial for similar rare diseases, where traditional statistical methods are not suitable.

The TDA model confirmed well-established associations between diagnostic tests (*e.g.* TEM and CBP) and genetics, as seen by the similar colour patterns in the topological models (figure 2), where TEM defect and CBP mapped very closely visually to corresponding gene group. This confirms a strong association that is in agreement with the published PCD literature [2, 21]. Distinct genetic findings were also associated with disease severity. We found *CCDC39* patients had significantly worse lung function at diagnosis (by FEV_1 z-score) when compared to all other groups, as has previously been observed in individuals with microtubular defects [8, 9, 25, 26]. Furthermore, modelling identified other findings not previously reported, including that individuals with *DNAH11* mutations were significantly less likely to have NRDS and, in turn, that the absence of NRDS is associated with better lung function at diagnosis. These findings were consistent between both the discovery and validation groups, and when using traditional statistical approaches.

The underlying pattern of the discovery group topological model data suggests that patients with compromised lower airways at diagnosis (*i.e.* decreased lung function and a history of NRDS) reported fewer upper airway symptoms (*i.e.* a history of glue ear and rhinitis). However, these findings could not be verified in the validation model as they may result from its over-fitting and this requires independent validation in an adequately powered independent dataset.

Comparison to previous literature

Our findings confirm and add to the evidence from other PCD genotype–phenotype studies. The largest of these have been two cross-sectional, longitudinal studies from the USA and Canada (the Genetic Disorders of Mucociliary Clearance Consortium) which also showed, based on ultrastructural phenotype and limited genotype information, that patients with microtubular defects had worse lung function [8, 9]. We also confirmed associations previously described in smaller studies, such as the absence of *situs inversus* in individuals with radial spoke, CC and N-DRC/molecular ruler gene mutations [4, 5, 22, 27, 28].

A previous study, using lung clearance index as a more sensitive measure of lung function, showed preserved lung function in a small group of patients from our cohort with normal ultrastructure (of which the majority had *DHAH11* defects) [26]. We have further confirmed that this genotype is associated with milder lung disease by showing that these patients clustered in an area with higher values of FEV_1 z-score. Traditional statistics also showed better preserved lung function in patients with *DNAH11* variants compared to those with mutations in any of the other genes.

Notably, patients carrying mutations in *DNAH5* were phenotypically diverse. The reasons for this are unclear, but may likely be connected to the variety of different mutations within this large gene. *DNAH5* was the gene found to have the widest spectrum of gene variants in our overall cohort. This diversity and high number of different mutations is in line with *DNAH5* being the most common overall genetic cause of PCD and most frequently mutated gene in affected individuals (with at least 100 different pathogenic mutations recorded worldwide) [29]. In PCD, it is likely that there will be patient phenotypic differences associated not just with the specific gene, but also the nature and location of the mutations within that gene. These genotype-related differences are already emerging on a smaller scale. For example, diagnostic results in *DNAH5* are known to vary somewhat depending on the mutation type (*e.g.* premature stop codon (nonsense) *versus* missense) [30]. Differences are also associated with missense *versus* truncation mutations in *CCDC103*, where a milder diagnostic and clinical phenotype has been described in individuals with p.His154Pro missense mutations [18].

Strengths and weaknesses

This is the largest study to date investigating genotype–phenotype associations in PCD. Using a new methodology of hypothesis-free TDA to examine underlying patterns in the dataset, genotype–phenotype patterns were identified from relatively few patients, something that would be difficult with usual clustering methods. The use of temporally and geographically distinct training and validation groups is highly recommended for such topological clustering approaches [31]. Initial UK discovery findings were

validated in the mixed internal and external dataset, including by replication of several important associations published previously, suggesting that these results are generalisable to other PCD populations.

The major weakness of our study remains the statistical power required to tease out relationships in a heterogeneous rare condition. To avoid problems with multiple comparisons and loss of statistical power, TDA-led hypothesis testing was performed for only two genes (*CCDC39* and *DNAH11*). This required combining the discovery and validation datasets, and a multinational dataset larger than any existing cohort will be required to ascertain further differences, especially to analyse whether variant types (*e.g.* stop-gain, frameshift, splicing, missense and copy number variants) explain some of the differences seen in the phenotypic data.

Another limitation of our study was potential recall bias for neonatal and early life events, due to a reliance on parental memory to report symptoms at the time of diagnosis. Not all medical records were complete and therefore "missing data" was recorded for some of these variables; however, TDA is particularly robust to missing data (see supplementary material for more information) [14]. Finally, we acknowledge that TDA is not completely hypothesis free, as we chose variables to enter into the models. Furthermore, there may be confounding variables affecting our models that have not yet been identified.

Potential impact for clinical management and research

A better understanding of genotype–phenotype associations from studies such as these should inform education and counselling for PCD patients and their families, and will alter disease management in the future. Identifying patients that may require more aggressive or personalised treatment due to underlying genetics will allow for better and more targeted care (*e.g.* high-risk groups, such as patients with *CCDC39* mutations, might benefit from more intense and targeted therapies).

The identification of mutations in known PCD-causative genes confirms a diagnosis of PCD. The topological models highlighted previously describe links between affected genes, TEM defects and CBPs from high-speed video analysis (HSVA), indicating that diagnostic TEM and HSVA tests can play an important supportive role in the classification of novel gene variants (likely of causal nature) and variants of uncertain clinical significance [2, 19]. These tests can also direct genetic testing to target a specific subset of genes.

By including longitudinal parameters such as lung function in the model, our approach for exploring genotype–phenotype associations might be useful for future longitudinal trials in PCD. It is a model-generating approach that could also usefully be applied to other rare diseases and to more common conditions. More accurate mapping of clinical characteristics, including severity, will allow a more targeted approach to treatment with an associated improvement in patient outcomes.

Overall, these clinically important findings can be useful in counselling parents and when considering prognosis and ongoing therapeutic interventions.

Acknowledgements: We thank the patients and their families for participating in the study and acknowledge the PCD Family Support Group. Borislav Dimitrov (University of Southampton) led initial discussions regarding statistical and TDA approaches for exploring genotype–phenotype relationships. He sadly died before the analyses began. We thank the following for their clinical and laboratory contributions to data used in this manuscript: Lucy Jenkins, Thomas Cullup, Alexandros Onoufriadis (University College London and Great Ormond Street, UK), Patricia Goggin, Claire L. Jackson, Janice Coles, James Thompson, Amanda Harris, Amanda Friend (University of Southampton and University Hospital Southampton, UK), Mellisa Dixon, Sarah Ollosson, Andrew V. Rogers, Emily Frost, Charlotte Richardson, Farheen Daudvohra, Paul Griffin and Thomas Burgoyne (Royal Brompton Hospital, UK). The researchers are supported by the Better Evidence to Advance Therapeutic options for PCD (BEAT-PCD) network (COST Action 1407 and European Respiratory Society (ERS) Clinical Research Collaboration). Several authors of this publication are members of the European Reference Network for Rare Respiratory Diseases (ERN-LUNG) (project ID number 739546).

Author contributions: Concept and design of the study: J.S. Lucas, C. Hogg, H.M. Mitchison, A. Shoemark and B. Rubbo. Genotyping: H.M. Mitchison, M.R. Fassad, M.P. Patel, N.S. Thomas, D. Hunt, M. Edwards, D. Morris-Rosendahl and M. Legendre. Clinical characterisation: J.S. Lucas, C. Hogg, W.T. Walker, M. Carroll, S.B. Carr, M.R. Loebinger, R. Wilson, E.G. Haarman, J-F. Papon, B. Maitre, G. Thouvenin, P-R. Burgel and I. Honoré. TDA models: B. Rubbo, J. Brandsma and G. Carlsson. Data collection: B. Rubbo, A. Shoemark, S. Best, W.T. Walker, C. Hogg, J-F. Papon, B. Maitre, G. Thouvenin, P-R. Burgel, I. Honoré, E.G. Haarman, I.C.M. Bon, E. Escudier,

G. Jouvion and M. Legendre. Planned and performed the statistical analyses: B. Rubbo, A. Shoemark, C.M. Parsons and J.S. Lucas. Laboratory analyses and data collection: E. Escudier, G. Jouvion, A. Shoemark, M. Legendre, S. Best, H.M. Mitchison and M.R. Fassad. Interpretation of data analyses: B. Rubbo, A. Shoemark, J.S. Lucas, H.M. Mitchison and C. Hogg. Drafted the manuscript: A. Shoemark, B. Rubbo, J.S. Lucas, H.M. Mitchison and C. Hogg. Revised the manuscript: J.S. Lucas, C. Hogg, H.M. Mitchison, A. Shoemark, B. Rubbo, N.S. Thomas, M. Legendre, M.R. Fassad, W.T. Walker, S.B. Carr, I. Honoré and E. Escudier. All authors read and approved the final manuscript. H.M. Michison (h.mitchison@ucl.ac.uk) was principal investigator for the genetics aspect of the work, and J.S. Lucas (jlucas1@soton.ac.uk) was clinical and epidemiological principal investigator. H.M. Mitchison and J.S. Lucas had full access to all data and take final responsibility for the decision to submit for publication.

Conflict of interest: B. Rubbo has nothing to disclose. M. Legendre has nothing to disclose. M.R. Fassad has nothing to disclose. E.G. Haarman has nothing to disclose. S. Best has nothing to disclose. I.C.M. Bon has nothing to disclose. J. Brandsma has nothing to disclose. P-R. Burgel reports personal fees for lectures and advisory board work from AstraZeneca, Boehringer Ingelheim, Chiesi, Novartis, Teva and Vertex, as well as personal fees for lectures from GSK, Pfizer and Zambon, outside the submitted work. G. Carlsson has nothing to disclose. S.B. Carr reports non-financial support and other (advisory board, lecture fee, travel, steering committee) from Vertex Pharmaceuticals, other (advisory board) from Profile Pharmaceuticals and other (lectures) from Teva Pharmaceuticals, as well as non-financial support and other (advisory board, travel, accommodation) from Chiesi Pharmaceuticals, outside the submitted work. M. Carroll has nothing to disclose. M. Edwards has nothing to disclose. E. Escudier has nothing to disclose. I. Honoré has nothing to disclose. D. Hunt has nothing to disclose. G. Jouvion has nothing to disclose. M.R. Loebinger reports personal fees for advisory board work and consultancy from AstraZeneca, Insmed, Polyphor, Bayer and Griffols, outside the submitted work. B. Maitre has nothing to disclose. D. Morris-Rosendahl has nothing to disclose. J-F. Papon has nothing to disclose. C.M. Parsons has nothing to disclose. M.P. Patel has nothing to disclose. N.S. Thomas has nothing to disclose. G. Thouvenin has nothing to disclose. W.T. Walker has nothing to disclose. R. Wilson has nothing to disclose. C. Hogg has nothing to disclose. H.M. Mitchison has nothing to disclose. J.S. Lucas has nothing to disclose. A. Shoemark has nothing to disclose.

Support statement: The PCD Centres in Southampton and London, the Wessex Regional Genetics Laboratory and Wessex Clinical Genetics Service are funded by the National Health Service for England (NHSE). Clinical research in Southampton was supported by the National Institute for Health Research (NIHR) Southampton Respiratory Biomedical Research Centre (BRC) and NIHR Southampton Wellcome Trust Clinical Research Facility. H.M. Mitchison acknowledges support from Action Medical Research, the Great Ormond Street Children's Charity and the NIHR Great Ormond Street Hospital (GOSH) BRC. M.R. Fassad was also supported by the NIHR GOSH BRC and a PhD studentship from the British Council Newton–Mosharafa Fund and the Ministry of Higher Education in Egypt. In France, this work was supported by the Institut National de la Santé et de la Recherche Médicale (INSERM), the RaDiCo funded by the French National Research Agency under the specific programme "Investments for the Future" (cohort grant agreement ANR-10-COHO-0003) and the legs Poix grant from La Chancellerie des Universités of the Sorbonne Universités de Paris. The funders had no role in the writing of the manuscript or the decision to submit it for publication. No payment was received to write this article.

References

1 Goutaki M, Meier AB, Halbeisen FS, *et al.* Clinical manifestations in primary ciliary dyskinesia: systematic review and meta-analysis. *Eur Respir J* 2016; 48: 1081–1095.
2 Lucas JS, Davis SD, Omran H, *et al.* Primary ciliary dyskinesia in the genomics age. *Lancet Respir Med* 2020; 8: 202–216.
3 Wallmeier J, Nielsen KG, Kuehni CE, *et al.* Motile ciliopathies. *Nat Rev Dis Primers* 2020; 6: 77.
4 Olbrich H, Schmidts M, Werner C, *et al.* Recessive *HYDIN* mutations cause primary ciliary dyskinesia without randomization of left-right body asymmetry. *Am J Hum Genet* 2012; 91: 672–684.
5 Castleman VH, Romio L, Chodhari R, *et al.* Mutations in radial spoke head protein genes *RSPH9* and *RSPH4A* cause primary ciliary dyskinesia with central-microtubular-pair abnormalities. *Am J Hum Genet* 2009; 84: 197–209.
6 Boon M, Wallmeier J, Ma L, *et al. MCIDAS* mutations result in a mucociliary clearance disorder with reduced generation of multiple motile cilia. *Nat Commun* 2014; 5: 4418.
7 Best S, Shoemark A, Rubbo B, *et al.* Risk factors for situs defects and congenital heart disease in primary ciliary dyskinesia. *Thorax* 2019; 74: 203–205.
8 Davis SD, Ferkol TW, Rosenfeld M, *et al.* Clinical features of childhood primary ciliary dyskinesia by genotype and ultrastructural phenotype. *Am J Respir Crit Care Med* 2015; 191: 316–324.
9 Davis SD, Rosenfeld M, Lee HS, *et al.* Primary ciliary dyskinesia: longitudinal study of lung disease by ultrastructure defect and genotype. *Am J Respir Crit Care Med* 2019; 199: 190–198.

10 Lum PY, Singh G, Lehman A, *et al.* Extracting insights from the shape of complex data using topology. *Sci Rep* 2013; 3: 1236.
11 Nielson JL, Paquette J, Liu AW, *et al.* Topological data analysis for discovery in preclinical spinal cord injury and traumatic brain injury. *Nat Commun* 2015; 6: 8581.
12 Frattini V, Pagnotta SM, Fan JJ, *et al.* A metabolic function of *FGFR3-TACC3* gene fusions in cancer. *Nature* 2018; 553: 222–227.
13 Bruno JL, Romano D, Mazaika P, *et al.* Longitudinal identification of clinically distinct neurophenotypes in young children with fragile X syndrome. *Proc Natl Acad Sci U S A* 2017; 114: 10767–10772.
14 Offroy M, Duponchel L. Topological data analysis: a promising big data exploration tool in biology, analytical chemistry and physical chemistry. *Anal Chim Acta* 2016; 910: 1–11.
15 Nicolau M, Levine AJ, Carlsson G. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proc Natl Acad Sci U S A* 2011; 108: 7265–7270.
16 Li L, Cheng WY, Glicksberg BS, *et al.* Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci Transl Med* 2015; 7: 311ra174.
17 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: 1457–1468.
18 Lucas JS, Barbato A, Collins SA, *et al.* European Respiratory Society guidelines for the diagnosis of primary ciliary dyskinesia. *Eur Respir J* 2017; 49: 1601090.
19 Richards S, Aziz N, Bale S, *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 2015; 17: 405–424.
20 Singh G, Mémoli F, Carlsson F, *et al.* Topological methods for the analysis of high dimensional data sets and 3D object recognition. *In*: Botsch M, Pajarola R, Chen B, Zwicker M, eds. Eurographics symposium on point based graphics. The Eurographics Association, 2007; pp. 91–100.
21 Glushakov S, Kotenko I, Rekalo A. Paper ML07: handling missing data in clinical trials using topological data analysis. Pharmaceutical Users Software Exchange 2018 (PHUSE 2018). November 4–7, 2018, Frankfurt, Germany. www.lexjansen.com/phuse/2018/ml/ML07.pdf
22 Pruliere-Escabasse V, Coste A, Chauvin P, *et al.* Otologic features in children with primary ciliary dyskinesia. *Arch Otolaryngol Head Neck Surg* 2010; 136: 1121–1126.
23 Wallmeier J, Al-Mutairi DA, Chen CT, *et al.* Mutations in *CCNO* result in congenital mucociliary clearance disorder with reduced generation of multiple motile cilia. *Nat Genet* 2014; 46: 646–651.
24 Davis SD, Ferkol TW, Rosenfeld M, *et al.* Clinical features of childhood primary ciliary dyskinesia by genotype and ultrastructural phenotype. *Am J Respir Crit Care Med* 2015; 191: 316–324.
25 Shah A, Shoemark A, MacNeill SJ, *et al.* A longitudinal study characterising a large adult primary ciliary dyskinesia population. *Eur Respir J* 2016; 48: 441–450.
26 Irving S, Dixon M, Fassad MR, *et al.* Primary ciliary dyskinesia due to microtubular defects is associated with worse lung clearance index. *Lung* 2018; 196: 231–238.
27 Knowles MR, Ostrowski LE, Leigh MW, *et al.* Mutations in *RSPH1* cause primary ciliary dyskinesia with a unique clinical and ciliary phenotype. *Am J Respir Crit Care Med* 2014; 189: 707–717.
28 Vallet C, Escudier E, Roudot-Thoraval F, *et al.* Primary ciliary dyskinesia presentation in 60 children according to ciliary ultrastructure. *Eur J Pediatr* 2013; 172: 1053–1060.
29 Landrum MJ, Lee JM, Benson M, *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 2018; 46: D1062–D1067.
30 Kispert A, Petry M, Olbrich H, *et al.* Genotype-phenotype correlations in PCD patients carrying *DNAH5* mutations. *Thorax* 2003; 58: 552–554.
31 Moons KG, Altman DG, Reitsma JB, *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med* 2015; 162: W1–W73.