BMC Pulmonary Medicine

RESEARCH

Open Access

The clinical value and most informative threshold of polygenic risk score in the Quebec City Case-Control Asthma Cohort

Martin Pariès[1,2], Stéphanie Bougeard[2], Aida Eslami[3,4], Zhonglin Li[3], Michel Laviolette[3], Louis-Philippe Boulet[3], Evelyne Vigneau[1] and Yohan Bossé[3,5*]

Abstract

Genome-wide association studies (GWAS) have identified genetic variants robustly associated with asthma. A potential near-term clinical application is to calculate polygenic risk score (PRS) to improve disease risk prediction. The value of PRS, as part of numerous multi-source variables used to define asthma, remains unclear. This study aims to evaluate PRS and define most informative thresholds in relation to conventional clinical and physiological criteria of asthma using a multivariate statistical method. Clinical and genome-wide genotyping data were obtained from the Quebec City Case-Control Asthma Cohort (QCCCAC), which is an independent cohort from previous GWAS. PRS was derived using LDpred2 and integrated with other asthma phenotypes by means of Principal Component Analysis with Optimal Scaling (PCAOS). PRS was considered using 'ordinal level of scaling' to account for non-linear information. In two dimensional PCAOS space, the first component delineated individuals with and without asthma, whereas the severity of asthma was discerned on the second component. The positioning of high vs. low PRS in this space matched the presence and absence of airway hyperresponsiveness, showing that PRS delineated cases and controls at the same extent as a positive bronchial challenge test. The top 10% and the bottom 5% of the PRS were the most informative thresholds to define individuals at high and low genetic risk of asthma in this cohort. PRS used in a multivariate method offers a decision-making space similar to hyperresponsiveness in this cohort and highlights the most informative and asymmetrical thresholds to define high and low genetic risk of asthma.

Keywords Asthma, Threshold, Polygenic risk score, Multivariate method, Optimal scaling, Cohort study

*Correspondence:
Yohan Bossé
yohan.bosse@criucpq.ulaval.ca
[1]Oniris, INRAE, StatSC, Nantes 44300, France
[2]Anses (French Agency for Food, Environmental and Occupational Health and Safety), Ploufragan 22440, France
[3]Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval, Quebec City, Canada
[4]Department of Social and Preventive Medicine, Université Laval, Quebec City, Canada
[5]Department of Molecular Medicine, Université Laval, Quebec City, Canada

Introduction

Asthma is a heterogeneous disease that can be caused by a combination of genetic and environmental factors [1]. Major progress to identify genetic variants associated with asthma has been made during the past decade by genome-wide association studies (GWAS) [2, 3]. Individually, genetic loci associated with asthma have small effect sizes but collectively they can be grouped into a polygenic risk score (PRS) to delineate a subgroup of individuals at higher or lower genetic risk of asthma [4, 5]. A PRS is thus arguably the most near-term clinical application of the new genetic knowledge derived from asthma GWAS. However, the clinical value of PRS in asthma, relative to conventional risk factors, remains largely unknown. In addition, the specific threshold to define an individual at high or low genetic risk remains to be established.

The Quebec City Case-Control Asthma Cohort (QCCCAC) is a new resource to study clinical and genetic factors implicated in asthma [6]. QCCCAC was not part of the previous GWAS on asthma and thus represents an independent dataset to evaluate the clinical value of PRS. QCCCAC consists of individuals well-characterized for asthma and related phenotypes including demographic characteristics, pulmonary functions, smoking status, blood biomarkers, allergies as well as genetics. QCCCAC was first analyzed to identify subgroups of clinically similar asthma individuals [6, 7]. However, no analysis of all the gathered variables was carried out, particularly between phenotypes and genome-wide genotyping data. In addition, the analysis of such data raises methodological challenges: analyzing simultaneously a set of variables of heterogeneous nature (numeric, ordinal, nominal).

Principal Component Analysis (PCA) is a common statistical method to explore the relationships among variables gathered on a cohort [8, 9]. This method reduces the complex variable's space into an optimally low-dimensional space based on the first components. Limitations of PCA is that it can only be applied to numeric variables and assumes that relationships between variables and components are linear. However, clinical variables usually have different natures: some are numeric (e.g., body mass index), some are nominal with unordered categories (e.g., never, former, and current smoker) or ordinal with ordered categories (e.g., symptoms of a disease with no/light/moderate/severe symptoms). To deal with such complex data, various solutions were proposed in the statistical literature [10–14]. However, the integration of ordinal variables is usually done without further consideration, either assumed to be numeric or nominal, none of these options being fully satisfactory. To deal with heterogeneous variables including ordinal ones, Optimal Scaling (OS) methods, based on variable's quantification, were proposed by the Dutch School of data analysis [15, 16]. OS aims to explore the relationships among numerous variables in a reduced and interpretable space while taking into account their specific nature.

The aim of this study was to evaluate the value of PRS in the complex space of variables that define asthma and the best possible PRS thresholds in the QCCCAC. Several statistical challenges have to be addressed: explore the relationships among all these variables in an interpretable (thus reduced) space, while taking into account the heterogeneous nature of the variables. Moreover, we took advantage of the ordinal level of scaling, and the resulting quantification to investigate the clinical value of PRS and define its most informative thresholds.

Materials and methods

Data

Data come from the Quebec City Case-Control Asthma Cohort (QCCCAC). It contains 1,585 French Canadian white subjects over 18 years of age with and without asthma. Details on data collection are given in [6]. The study protocol for the QCCCAC was approved by the Research Ethics Board of the *Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval* (#20273). All participating subjects signed an informed consent approved by the REB. Subjects are de-identified using a code number for confidentiality. Access to data is protected using the data management structure approved by the REB.

Among the available variables in the QCCCAC, fourteen were selected, related to the demographic characteristics (age, sex, body mass index), pulmonary functions (spirometry measurements: forced expiratory volume in one second (FEV1) and forced vital capacity (FVC); airway responsiveness to methacholine challenge measured using the 2-min tidal method [17]), smoking status, blood biomarkers (blood cell counts of neutrophils and eosinophils), allergies (positive skin-prick test), polygenic risk score, asthma status (case/control; physician diagnosis based on clinical symptoms, lung function and airway responsiveness) and asthma severity symptoms according to the Canadian Asthma Guidelines [18] (Supplementary Table S1). These variables are of heterogeneous nature: eight are numeric, three are binary (nominal with two categories) and two are ordinal. These variables for $n = 1{,}352$ individuals (see the 'Handling missing data' section for the number of individuals) are described in Table 1. Clinical characteristics of QCCCAC are presented in Table 2, and differences between cases and controls were assessed with analysis of variance for numeric variables and with chi-square tests for categorical variables.

Table 1 Description of the selected variables of the Quebec City Case-Control Asthma Cohort ($n = 1,352$). Descriptive statistics are mean (standard deviation) for numeric variables or frequency (%) for nominal and ordinal variables

Name	Nature	Descriptive statistics	Details
Sex	Nominal (binary)	***Female*** F (61%) ***Male*** M (39%)	
Age	Numeric	***Mean***: 38.6 (16.3)	***Age*** at evaluation (year).
BMI	Numeric	***Mean***: 26.1 (5.4)	***Body mass index***: weight of an individual divided by its squared height (kg/m^2).
Smoking status	Nominal (binary)	***Non-smoker*** NS (95%) ***Smoker*** S (5%)	
FVC (% predicted)	Numeric	***Mean***: 105.8 (16.6)	***Pulmonary function*** measured by spirometry: forced expiratory volume in 1 s (FEV1) and forced vital capacity (FVC), expressed as percentage of predicted value.
FEV1 (% predicted)	Numeric	***Mean***: 94.4 (19.6)	
Eosinophils	Numeric	***Mean***: 0.03 (0.03)	***Blood cell counts*** for neutrophils and eosinophils ($10e^9.L^{-1}$).
Neutrophils	Numeric	***Mean***: 0.6 (0.09)	
IgE	Numeric	***Mean***: 1.9 (0.6)	***Total serum immunoglobulin E***: Measured with enzyme immunofluorometry ($UI.mL^{-1}$).
AHR	Nominal (binary)	***Hyperresponsive*** AHR+ (56%) ***Non-hyperresponsive*** AHR- (44%)	***Airway hyperresponsiveness***: Measured using the 2-min tidal method [17] (the cut-off value of the test is 8 $mg.mL^{-1}$).
Nb allergens	Numeric	***Mean***: 7.0 (6.8)	***Skin-prick tests*** with 25 standard allergens performed to measure the allergic status of each individual. The number of positive responses (wheal diameter of at least 3 mm) is summed up.
Asthma symptoms	Ordinal with 4 categories	***No symptoms*** (36.2%) ***Light*** (32.6%) ***Moderate*** (20.6%) ***Severe*** (10.7%)	***Severity of asthma*** determined according to the Canadian Asthma Guidelines [18]. Individuals with similar severity of symptoms were grouped (Supplementary Table S1).
Asthma status	Nominal binary	***Asthma*** (63.8%) ***Control*** (36.2%)	***Asthma status*** confirmed by physicians (L.-P. Boulet and M. Laviolette) based on clinical symptoms, lung function and airway responsiveness.
PRS	Discretized numeric variable (ordinal)	***PRS.1*** (5%) ***PRS.2*** (5%) … ***PRS.20*** (5%)	***Polygenic risk score***: A single value estimate of an individual's genetic liability to a trait or disease [19].

Genotyping and quality control

Genotyping of the QCCCAC was performed using the Illumina Global Screening Array (GSA) version 3 BeadChip with the multi-disease (MD) drop-in panel. Quality controls (QC) were performed excluding low quality genetic variants with 10th percentile of Illumina GenCall score ≤ 0.1, call rate < 0.97%, Hardy-Weinberg equilibrium $P < 1E\text{-}7$, minor allele frequency (MAF) < 1%, or duplicate variants sharing the same base pair coordinate. Low quality DNA samples were also filtered out after consideration for the genotype completion rate < 95%, genotypic and phenotypic sex mismatch, unexpected duplicates and genetic relatedness (first-degree relatives) evaluated by identity-by-state using PLINK, outliers based on the inbreeding coefficients (F > 10 standard deviation from the mean), and genetic background outliers detected by principal component analysis with HapMap subjects as population reference panel.

Development and coding strategy of the polygenic risk score

We are specifically interested in deciphering the value of an asthma-Polygenic Risk Score (PRS) beyond conventional clinical features and expiratory airflow limitations. The PRS is a numerical score which summarizes the effect of a large number of genetic variants on an individual's phenotype (asthma here) [19]. It is calculated using genome-wide genotyping data and relevant Genome-Wide Association Study (GWAS) summary statistics (effect sizes). The GWAS identified several genetic variants, mostly single-nucleotide polymorphisms (SNPs), associated with asthma [2, 3]. The PRS is calculated as a weighted sum of the trait-associated alleles. To calculate the PRS, we applied the LDpred2 function with the automatic mode in the R package *bigsnpr* (version 1.10.7). LDpred2 is based on a Bayesian method which aims at estimating the average posterior effect size using a linkage disequilibrium matrix and summary statistics by assuming a prior distribution on the real effect sizes [20]. Summary statistics were extracted from a European

Table 2 Clinical characteristics of the Quebec City Case-Control Asthma Cohort ($n = 1,352$). Means and standard deviation of numeric variables and frequencies for categorical variables according to the asthma status. *P*-values are obtained from analysis of variance (numeric variables) or Chi-square test (categorical variables)

	***p*-value**	**Cases (*N* = 863)**	**Controls (*N* = 489)**	**Overall (*N* = 1352)**
Sex				
Female	0.484	522 (60.5%)	306 (62.6%)	828 (61.2%)
Male		341 (39.5%)	183 (37.4%)	524 (38.8%)
Age				
Mean (SD)	0.001	39.7 (16.1)	36.8 (16.4)	38.6 (16.3)
Median [Min, Max]		36.0 [18.0, 86.0]	30.0 [18.0, 77.0]	34.0 [18.0, 86.0]
BMI				
Mean (SD)	2.91e-09	26.8 (5.72)	25.0 (4.49)	26.1 (5.37)
Median [Min, Max]		25.6 [14.9, 52.6]	24.2 [16.7, 43.1]	25.1 [14.9, 52.6]
Smoking status				
Non-smoker	0.834	817 (94.7%)	465 (95.1%)	1282 (94.8%)
Smoker		46 (5.3%)	24 (4.9%)	70 (5.2%)
FVC				
Mean (SD)	< 2.22e-16	102 (17.0)	112 (14.0)	106 (16.6)
Median [Min, Max]		104 [55.0, 145]	111 [78.0, 165]	106 [55.0, 165]
FEV1				
Mean (SD)	< 2.22e-16	87.7 (19.4)	106 (13.2)	94.4 (19.6)
Median [Min, Max]		90.0 [25.0, 180]	106 [72.0, 159]	96.0 [25.0, 180]
Eosinophils				
Mean (SD)	< 2.22e-16	0.0358 (0.0293)	0.0236 (0.0185)	0.0314 (0.0266)
Median [Min, Max]		0.0280 [0, 0.320]	0.0180 [0, 0.138]	0.0240 [0, 0.320]
Neutrophils				
Mean (SD)	0.010	0.601 (0.0858)	0.588 (0.0864)	0.597 (0.0862)
Median [Min, Max]		0.599 [0.290, 0.950]	0.597 [0.0100, 0.782]	0.598 [0.0100, 0.950]
IgE				
Mean (SD)	< 2.22e-16	2.05 (0.600)	1.52 (0.548)	1.86 (0.634)
Median [Min, Max]		2.03 [0.556, 4.13]	1.45 [0.690, 3.43]	1.85 [0.556, 4.13]
AHR				
Non hyperresponsive	< 2.22e-16	136 (15.8%)	458 (93.7%)	594 (43.9%)
Hyperresponsive		727 (84.2%)	31 (6.3%)	758 (56.1%)
Nb.allergies				
Mean (SD)	< 2.22e-16	8.65 (6.96)	4.10 (5.42)	7.00 (6.81)
Median [Min, Max]		8.00 [0, 24.0]	2.00 [0, 22.0]	5.00 [0, 24.0]
Asthma symptoms				
No symptom	< 2.22e-16	0 (0%)	489 (100%)	489 (36.2%)
Light		441 (51.1%)	0 (0%)	441 (32.6%)
Moderate		278 (32.2%)	0 (0%)	278 (20.6%)
Severe		144 (16.7%)	0 (0%)	144 (10.7%)

ancestry GWAS meta-analysis on asthma (19,954 cases and 107,715 controls) published by the Trans-National Asthma Genetic Consortium (TAGC) [3]. The statistics reported by TAGC are the result of a meta-analysis of 66 GWAS where asthma status is based on physician diagnosis and standardized questionnaires. For the current study, the PRS includes a total 931,818 genetic variants which are in common between our asthma GWAS in QCCCAC and the GWAS summary statistics from TAGC after restricting to HapMap3 variants as recommended [20]. LDpred2 generates a genome-wide PRS and is not limited to genome-wide significant variants or preselected variants by the authors.

From a statistical point of view, the PRS is a numeric variable with a low value representing a low liability, and a high value representing a high liability to the disease, here asthma. In the following, this score is transformed into an ordinal variable using a quantile-based discretization. Several discretizations were tested and twenty categories were finally chosen as a compromise between interpretation and accuracy (see Results section). Accordingly, each ordered category corresponds to a fixed fraction (5%) of

the observed distribution function. They are labeled from 'PRS.1' (lowest PRS) to 'PRS.20' (highest PRS). The performance of the PRS was assessed by means of logistic regression explaining the asthma status by the PRS variable (recoded in categories), and adjusted for covariates including age, sex and 10 ancestry-based principal components. Ten is the number of principal components that we are using for genetic association studies in our European ancestry population [21, 22]. Body mass index (BMI) was also evaluated as a covariate considering the statistically significant difference between asthma cases and controls. The effect size estimates of PRS on asthma were highly similar in models with and without BMI and it was thus not kept in the final model.

Missing data

Out of the 1,585 individuals in the cohort, 233 were removed from the analysis as they have missing values for all the variables in a group, such as lung functions (FEV1, FVC) or biomarkers (eosinophils, neutrophils and IgE). These missing data cannot be reliably imputed. However, the remaining missing values (1% of the data) were considered at random and imputed with the K-nearest neighbors' method (K = 5 by default, using the R package *VIM* [23]). Because of the heterogeneous nature of the variables, the general coefficient of similarity proposed by Gower [24] was considered to compute distances between individuals.

Multivariate statistical analysis (PCAOS)

The exploratory multivariate analysis with variables of heterogeneous nature was achieved by means of a Principal Component Analysis with Optimal Scaling (PCAOS) [16, 25, 26]. The PCAOS method achieved both dimensionality reduction of data and quantification of the variables. This was performed using an Alternating Least Squares with Optimal Scaling (ALSOS) algorithm which alternates two steps: (1) quantification and (2) components estimation, until the minimization of a least square loss function [27]. For the 'optimal scaling step' (1), each variable is quantified so that each category is associated with a numeric value (interested readers may refer to [28, 29]). The 'component estimation step' (2) consisted of a Principal Components Analysis applied to the quantified variables.

Relationships among variables are studied by plotting variable loadings that represent correlations between quantified variables and components. Supplementary variables can be considered; they do not participate to component building but can be projected onto them. This variable-plot helps to interpret the graphical display of observations directly given by components. The appropriate number of components to be interpreted is selected after computing models with different dimensions (e.g., from 1 to 6) and study gains in explained inertia for solutions with H + 1 components compared to H components. As categories from categorical (nominal and ordinal) variables are quantified with a single numeric value, quantification plots with categories on the x-axis and single quantification on the y-axis can be drawn. These graphs allow to display non-linear relations between categorical variables and components.

The R package *PCA.OS* was developed and is available on GitHub (https://github.com/martinparies/PCA.OS). Sensitivity of the quantification for the multivariate analysis was performed on 1,000 bootstrapped data.

Results

QCCCAC

The clinical characteristics of cases and controls are described in Table 2. A total of 863 (64%) asthma cases and 489 (36%) controls were enrolled with an average age at recruitment of 38.6 years. All participants were of white European ancestry confirmed at genotyping. A larger majority were women (61%) and a small fraction were current-smokers (5%). Asthma patients were in large part atopic (81% with at least one positive skin prick test). As expected, lung function was lower in asthma (FVC p-value < 2.22e-16; FEV1 p-value < 2.22e-16) and the proportion of hyperresponsiveness was higher in cases compared to controls (AHR p-value < 2.22e-16). Blood eosinophils (p-value < 2.22e-16), neutrophils (p-value = 0.010) and IgE (p-value < 2.22e-16) were also higher in cases.

Polygenic risk score performance

The distribution of the PRS in asthma cases and controls is illustrated in Supplementary Figure S1. Moreover, results from the logistic regression of the PRS (recoded as three categories: bottom 20%; 20-80% and top 20%, which is a common practice in the field [30]), explaining the asthma status are available in Supplementary Figure S2. Individuals in the top 20% category had a 3.2-fold odds of asthma compared to the reference group consisting of the 20% of individuals with the lowest PRS. The corresponding area under the receiver operating characteristic curve was of 64.7% (CI 95%, 61.6-67.8%).

PCAOS

The PCAOS algorithm was performed on the variables described in Table 1. The 'asthma severity symptoms' was considered as supplementary information, thus 13 variables were used to build the model. For the PRS, several discretizations were tested and twenty categories were selected as a compromise between interpretation and accuracy (Supplementary Figure S3). The PCAOS algorithm converged in four iterations for a model with two components. The first component explained 24.0% of the

variance of the quantified variables and the second component 14.4% (for a total of 38% of variance explained) (Supplementary Figure S4).

Relationships among variables

The loadings of the quantified variables are represented in a two-dimensional space where their relationships can be highlighted (Fig. 1). Loadings values and correlation matrix between quantified variables are in Supplementary Table S2 and Figure S5. The first component (PC1) made it possible to differentiate individuals with (right-hand side of PC1) and without (left-hand side of PC1) asthma. The non-asthma status (left-hand side of PC1) was associated with non-hyperresponsiveness to methacholine (94% of control individuals were AHR-), good pulmonary functions (high values of FEV1 and FVC) and lowest PRS categories (PRS.1 and to a lesser extent PRS.2 to PRS.4). The asthma status (right-hand side of PC1) was associated with hyperresponsiveness (85% of individuals with asthma were AHR+), high numbers of allergens, high values of IgE and eosinophils and highest PRS categories (PRS.19 and 20 and to a lesser extent PRS.18). Extreme PRS categories (PRS.19 and PRS.20 categories, on the one hand, and PRS.1 category, on the other hand) were both linked to 'asthma status' and to 'AHR'. More specifically, the proportion of asthmatic and AHR + individuals increases with PRS categories; for low PRS category (PRS.1) 43% of individuals are asthmatic and 37% are AHR+, and for high PRS categories (PRS 19 and PRS.20) 80% of individuals are asthmatic and 70% are AHR+.

The negative side of the second component (bottom-hand side of PC2) showed that the oldest individuals with the highest BMI have higher values of neutrophils and lower eosinophils in circulation, as well as relatively low pulmonary capacities (FEV1, FVC) and low levels of IgE.

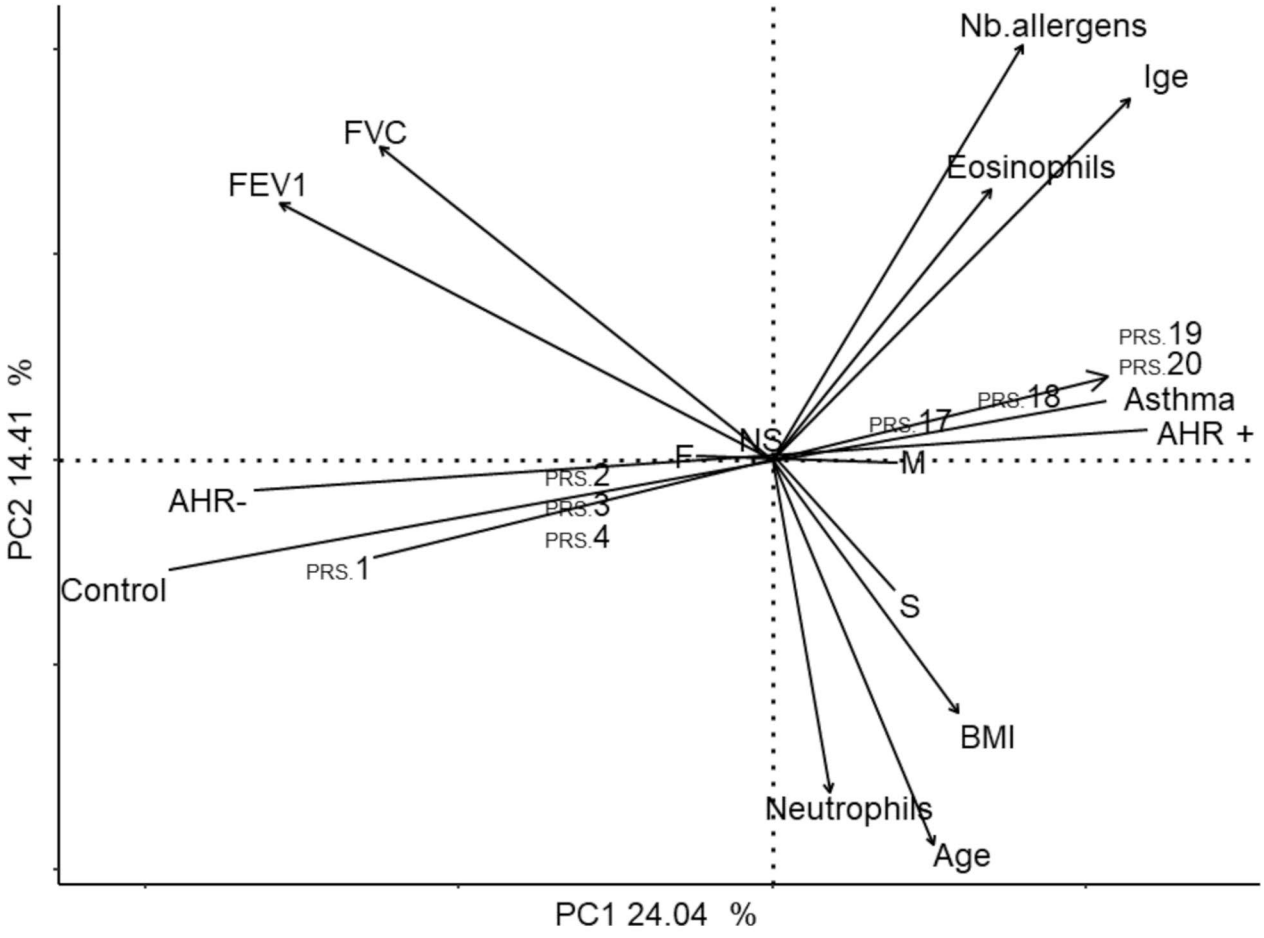

Fig. 1 Graphical display of the 13 variables from the Quebec City Case-Control Asthma Cohort ($n = 1,352$) in the two dimensional PCAOS space (38% of total variance explained). Numeric variables are represented by vectors starting from the origin. Nominal and ordinal variables are represented by their categories (connected by a line for better readability). For ease of reading, only extreme PRS categories with the lower and higher values are plotted. AHR, airway hyperresponsiveness; BMI, body mass index; F, female; FEV1, forced expiratory volume in one second; FVC, forced vital capacity; IgE, immunoglobulin E; M, male; Nb.allergens, number of allergens; NS, non-smoker; PRS, polygenic risk score; S, smoker

The positive side of PC2 (top-hand side of PC2) showed individuals with high numbers of allergens, IgE levels and eosinophils. In contrast to other variables, smoking and sex had a reduced link with asthma.

Polygenic risk score in multivariate analysis (PCAOS)

As mentioned, the numeric PRS was transformed into an ordinal variable to be able to highlight possible non-linear relationship between this variable and the others. Means and confidence intervals of the 1000 quantifications of the PRS variable, obtained with bootstrap simulation procedure, are illustrated in Fig. 2. Each confidence interval contains 95% of the 1000 quantifications. The value attributed to each PRS category reflects the influence of that category in the construction of the components.

Confidence intervals show that the overall shape of the curve is constant. Moreover, the curve clearly reveals that the gap between two consecutive categories is not constant, extreme categories having a much greater impact in seeking the PCAOS components. In practice, this highlights a non-linear relationship between PRS and other variables. Individuals with extreme values of PRS (in absolute values) exert more influence in building the components. In other words, the subset made of 5% of the individuals having the lowest PRS (Low risk = PRS.1) and the subset made of 10% of the individuals having the highest PRS (High risk = PRS.19, PRS.20) can be well differentiated into the two-dimensional space model (Fig. 1). Moreover, using logistic regression, it was found that high genetic risk individuals had 5.16 (CI 95% 2.69–9.93) increased odds of asthma compared to low genetic risk individuals, as indicated in Supplementary Table S3.

Graphical display of the individuals

The 1,352 individuals are plotted according to their coordinates on the two PCAOS components (Fig. 3). Each individual is colored according to its 'asthma symptoms' category (no/light/moderate/severe symptoms) considered as a supplementary variable in the analysis. Individuals with similar severity of symptoms are identifiable with more or less overlap. Control individuals ('no symptom') are clearly positioned on the left-hand side of Fig. 3

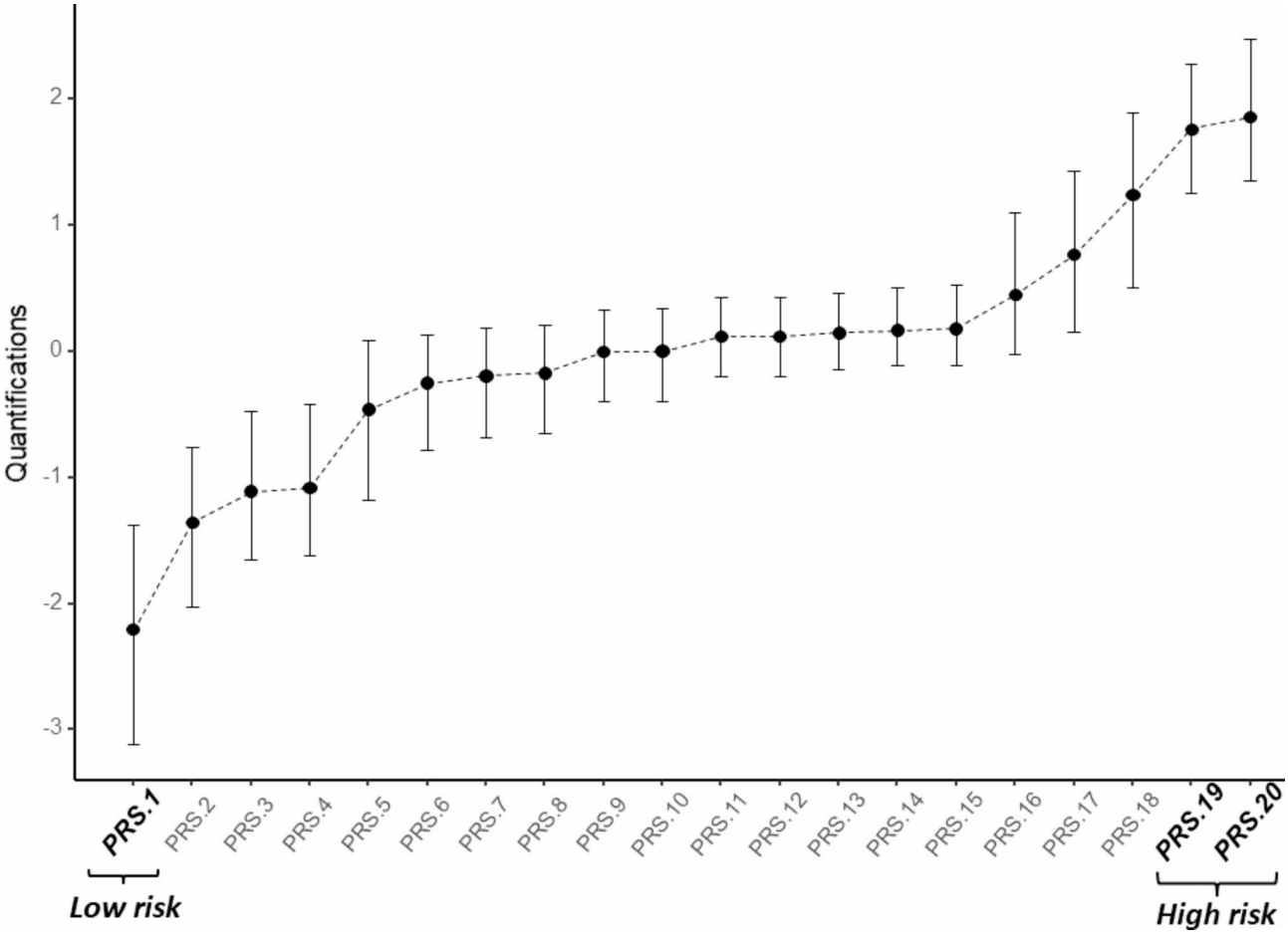

Fig. 2 Mean of PCAOS quantifications for the Polygenic Risk Score (PRS) variable obtained from bootstrapped data in the Quebec City Case-Control Asthma Cohort ($n = 1,352$). Each confidence interval contains 95% of the 1000 quantifications

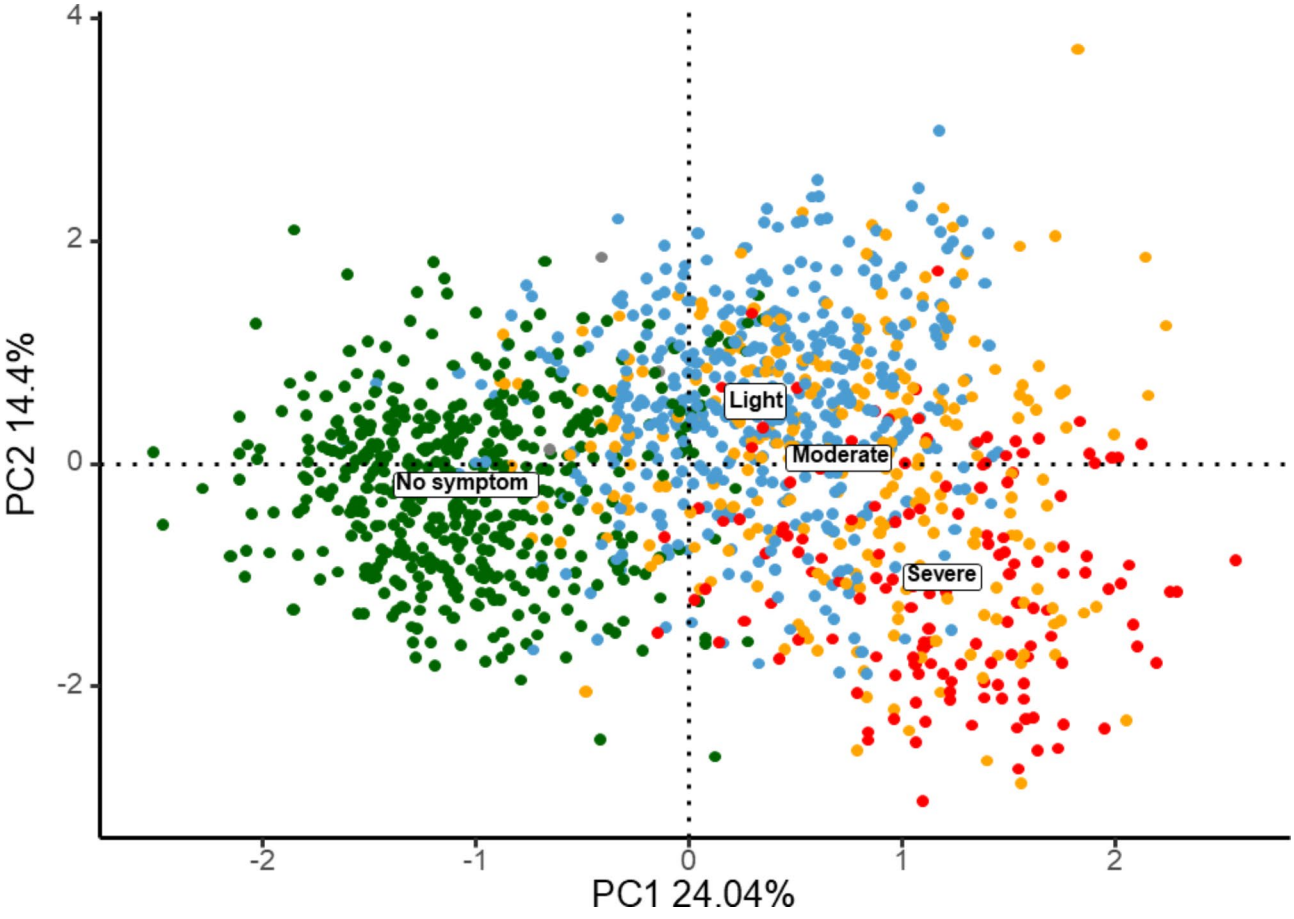

Fig. 3 Factorial representation on the first two PCAOS components of the individuals colored according to their asthma symptoms severity (green = 'no asthma'; blue = 'light'; orange = 'moderate', red = 'severe'). Categories of the 'asthma symptoms severity' are located as the average coordinates of individuals sharing the same category. Quebec City Case-Control Asthma Cohort ($n = 1,352$)

(low PC1 values) whereas individuals with asthma having light to severe symptoms are positioned on the right-hand side (high PC1 values). This is consistent with the location of the other variables (Fig. 1). Furthermore, individuals with light or moderate symptoms are separated from those with severe symptoms along the second component (PC2). It turned out that individuals with severe asthma symptoms could be identified as older individuals with high BMI, while individuals with light and moderate symptoms seemed to have a higher number of positive allergens.

Discussion

In this study, we used a relevant multivariate statistical method, PCAOS, to assess and integrate PRS as part of key variables to define asthma. A model with two components was interpreted with the first component delineating patients with and without asthma. Interestingly, the same axis was associated with hyperresponsiveness and high PRS categories, suggesting that PRS has clinical value similar to hyperresponsiveness in defining asthma. In addition, the multivariate PCAOS approach highlighted the non-linear nature of PRS where the extremities of the PRS distribution are more influential in shaping the model. In the QCCCAC cohort, the model established the upper and lower boundaries of PRS at the top 10% for high genetic risk and the bottom 5% for low genetic risk. This indicates asymmetrical PRS thresholds to define low vs. high-genetic risk of asthma.

Asthma is a heterogeneous airways disease. The definition of asthma relies on a number of criteria documenting the variable extent of airflow limitation, respiratory symptoms, hyperresponsiveness, inflammation as well as allergy [31]. The diagnosis must be performed by experienced physicians on the basis of one or more criteria depending on whether they have access or not to complex investigations. In our clinical research setting, we were able to assess asthma with gold standard methods. This had allowed us to study the relationships among clinical features of asthma, but most importantly evaluate how PRS performs in this complex mix of variables.

Interestingly, the PRS was found on the same dimension as asthma status and AHR. More specifically, high PRS categories were associated to two categories:

'asthma' from the asthma status variable and hyperresponsiveness to methacholine challenge. Thus, individuals with high values of PRS are prone to develop asthma characteristics such as hyperresponsiveness. Having asthma status and AHR on the same dimension is not surprising considering the weight of a positive bronchial challenge test on asthma diagnosis. However, to have PRS on the same dimension is novel and insightful. In fact, a genetic component to asthma has long been established [1], but it is only recently that we can capture this component into a PRS. The PRS calculated in our cohort was derived from genetic variants associated with asthma in previous GWAS. It should be emphasized that the QCCCAC was not part of previous GWAS of asthma and thus represents an independent cohort. It may be considered intuitive to delineate asthma cases from controls in QCCCAC using a PRS aggregating the effects of genetic variants associated with asthma in previous GWAS, but to do so at the same scale as AHR is surprising considering the small fraction of the total heritability explained by genetic variants, e.g., the SNP-heritability was estimated at 11.3% in UK Biobank [2]. These results are thus encouraging in terms of clinical application for the PRS considering foreseeable improvements in genetics of asthma and new methods to model polygenic architecture [32]. Concerning the severity of asthma, it seems that a severe form of asthma is not associated with the current asthma PRS, but more to other characteristics of the individual such as age, BMI and blood neutrophils. Additional studies will be needed to evaluate PRS derived from genetic variants specifically associated with asthma severity.

The non-linear information provided by the PRS and the non-symmetrical thresholds to define low vs. high-genetic risk of asthma have implications for other complex diseases. A common practice in the field is to define high, intermediate, and low genetic risk groups, for instance, based on the bottom 20%, 20-80% and the top 20% of the PRS, respectively [30]. However, thresholds vary widely across studies. For example, the first study showing the clinical utility of PRS in the field of lung cancer classified the genetic risk of individuals based on cutoffs at 5% and 95% [33]. There is a critical need to establish PRS thresholds that are clinically meaningful and to integrate PRS within the specific context of each disease. The PCAOS method used in this study can potentially provide a data-driven method to obtain thresholds that are more clinically relevant. The non-symmetrical thresholds obtained using this method is of particular interest and is likely to be suitable for other diseases. Studies will be needed on other diseases with multifactorial causes combining genetic and environmental factors.

Several studies have developed PRS to predict an individual's risk of developing asthma [4, 5, 34–36]. Results across studies are difficult to compare owing to the various methods to develop PRS and the metrics used to report the predictive properties of the PRS. In terms of diagnostic discrimination performance, our genome-wide PRS derived using the LDpred2 method in QCCCAC had similar AUC compared to the recent genome-wide PRS derived using the PRS-CS method in the white British subset of UK Biobank (AUC = 64.7% in QCCCAC vs. 62.3% UK Biobank) [34]. To the best of our knowledge, no previous PRS studies in asthma have tried to identify the most informative PRS thresholds to define individuals at high and low genetic risk of asthma. In this study, we demonstrated that informative PRS thresholds estimated from an asthma-specific multivariate model can improve risk stratification. Individuals with a PRS above the 10% of the distribution were 5.16 times (95% CI = 2.69–9.93) more likely to have asthma than those in the lower 5% of the distribution. In QCCCAC, these asymmetrical thresholds (lower 5% and upper 10%) identify the lower and upper subsets of individuals exerting the more influence in building the model components and differentiating the genetic risk as part of a complex set of asthma-related variables. Whether specific PRS thresholds for asthma can be established will require further investigation in other populations.

This study has limitations, first the method and results obtained, such as the proposed threshold for high and low genetic risks, will need to be assessed and validated in other asthma cohorts. Replication may be difficult as asthma phenotyping, i.e., set of demographic/clinical/genetic variables to define asthma, differs across sites. Second, the sample size of the QCCCAC is relatively small. However, individuals included in this cohort have been extensively phenotyped for asthma (demographic and clinical data, blood cell counts and IgE levels, allergy skin-prick tests as well as physiology evaluation by spirometry and bronchial challenge test). As recently indicated smaller studies with deep phenotyping will be equally important relative to historically large-scale GWAS for PRS development [37]. Third, GWAS summary statistics to develop the PRS were obtained from the European ancestry GWAS meta-analysis on asthma published by TAGC [3]. Larger GWAS on asthma have been reported [38, 39] and may improve the diagnostic discrimination performance of the PRS. Continuing progress in elucidating the genetics of asthma will improve the predictive accuracy of PRS and will require further investigation.

In conclusion, PCAOS in well-phenotyped asthma cohort makes it possible to study relationships among variables of different natures. The integration of the PRS as an ordinal variable highlights the non-linear

relationships between the genetics of individuals and other variables as well as the asymmetrical genetic risk thresholds. The application to the QCCCAC reveals that individuals at the lowest 5% and highest 10% asthma-PRS percentile are at disproportionally lower and higher genetic risk of asthma.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12890-025-03486-3.

Supplementary Material 1

Acknowledgements
We thank the research subjects for participating in this study. The collection of the Quebec City Case-Control Asthma Cohort was supported by grants from the Fondation de l'Institut universitaire de cardiologie et de pneumologie de Québec. The authors would like to thank the staff at the IUCPQ Biobank for their valuable assistance.

Author contributions
M.P., S.B., A.E., E.V., Y.B. contributed to the conception and design of the work. M.P., Z.L. performed data analysis. M.P., S.B., E.V., Y.B. wrote the main manuscript text. M.P. prepared the figures and tables. M.L. and L.-P.B. provided resources and interpretation of data. S.B., A.E., E.V., Y.B. ensured funding acquisition, supervision and project administration. All authors reviewed the manuscript and approved the submitted version.

Funding
This article is part of an ongoing PhD work located in the 'Nantes-Atlantic National College of Veterinary Medicine, Food Science and Engineering, France'. Researches are half funded by the 'Pays de la Loire, France' region and by the 'French Agency for Food, Environmental and Occupational Health and Safety'. Moreover, we would like to thank 'Mitacs Globalink' and the Faculty of Medicine of Université Laval for according a scholarship allowing collaboration between authors. Y.B. holds a Canada Research Chair in Genomics of Heart and Lung Diseases.

Data availability
The datasets generated and/or analysed during the current study are not publicly available in order to comply with the consent form signed by research participants, but are available from the corresponding author on reasonable request.

Declarations

Ethical approval
The study protocol was approved by the Research Ethics Board of the *Institut universitaire de cardiologie et de pneumologie de Québec – Université Laval* (#20273). All participating subjects signed an informed consent approved by the REB. Subjects are de-identified using a code number for confidentiality. Access to data is protected using the data management structure approved by the REB.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 20 October 2023 / Accepted: 7 January 2025
Published online: 15 January 2025

References

1. Bossé Y, Hudson TJ. Toward a comprehensive set of asthma susceptibility genes. Annu Rev Med. 2007;58:171–84.
2. Valette K, Li Z, Bon-Baret V, Chignon A, Berube JC, Eslami A, Lamothe J, Gaudreault N, Joubert P, Obeidat M, et al. Prioritization of candidate causal genes for asthma in susceptibility loci derived from UK Biobank. Commun Biol. 2021;4(1):700.
3. Demenais F, Margaritte-Jeannin P, Barnes KC, Cookson WOC, Altmuller J, Ang W, Barr RG, Beaty TH, Becker AB, Beilby J, et al. Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks. Nat Genet. 2018;50(1):42–53.
4. Namjou B, Lape M, Malolepsza E, DeVore SB, Weirauch MT, Dikilitas O, Jarvik GP, Kiryluk K, Kullo IJ, Liu C, et al. Multiancestral polygenic risk score for pediatric asthma. J Allergy Clin Immunol. 2022;150(5):1086–96.
5. Sordillo JE, Lutz SM, Jorgenson E, Iribarren C, McGeachie M, Dahlin A, Tantisira K, Kelly R, Lasky-Su J, Sakornsakolpat P, et al. A polygenic risk score for asthma in a large racially diverse population. Clin Exp Allergy. 2021;51(11):1410–20.
6. Lavoie-Charland E, Bérubé JC, Laviolette M, Boulet LP, Bossé Y. Multivariate asthma phenotypes in adults: the Quebec City Case-Control Asthma Cohort. Open J Respiratory Dis. 2013;3:133–42.
7. Lavoie-Charland E, Berube JC, Boulet LP, Bossé Y. Asthma susceptibility variants are more strongly associated with clinically similar subgroups. J Asthma. 2016;53(9):907–13.
8. Zhang Z, Castello A. Principal components analysis in clinical studies. Ann Transl Med. 2017;5(17):351.
9. Jolliffe IT. Generalizations and adaptations of principal component analysis. In: Principal component analysis. Springer series in statistics. New York: Springer; 1986. pp. 223–34. https://doi.org/10.1007/978-1-4757-1904-8_12
10. Escofier B. Traitement simultané de variables qualitatives et quantitatives en analyse factorielle. In: *Cahiers de l'Analyse des Données. Volume 4*, edn.; 1979: 137–146.
11. Chavent M, Kuentz-Simonet V, Labenne A, Saracco J. Multivariate Analysis of Mixed Data: The R Package PCAmixdata. 2017.
12. Hill MO, Smith AJE. Principal component analysis of taxonomic data with multi-state discrete characters. Taxon. 1976;25:249–55.
13. Kiers HAL. Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. Psychometrika. 1991;56:197–212.
14. Pagès J. Analyse factorielle de données mixtes. Revue De Statistique Appliquée Volume. 2004;52:93–111.
15. Young FW. Quantitative analysis of qualitative data. Psychometrika. 1981;46:357–88.
16. Gifi A. Nonlinear multivariate analysis. Wiley-Blackwell; 1990.
17. American Thoracic Society. Standardization of spirometry, 1994 update. Am J Respir Crit Care Med. 1995;152(3):1107–36.
18. Lougheed MD, Lemiere C, Dell SD, Ducharme FM, Fitzgerald JM, Leigh R, Licskai C, Rowe BH, Bowie D, Becker A, et al. Canadian thoracic Society Asthma Management Continuum–2010 Consensus Summary for children six years of age and over, and adults. Can Respir J. 2010;17(1):15–24.
19. Choi SW, Mak TS, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. Nat Protoc. 2020;15(9):2759–72.
20. Prive F, Arbel J, Vilhjalmsson BJ. LDpred2: better, faster, stronger. Bioinformatics. 2021;36(22–23):5424–31.
21. Theriault S, Li Z, Abner E, Luan J, Manikpurage HD, Houessou U, Zamani P, Briend M, Estonian Biobank Research T, Boudreau DK, et al. Integrative genomic analyses identify candidate causal genes for calcific aortic valve stenosis involving tissue-specific regulation. Nat Commun. 2024;15(1):2407.
22. Boumtje V, Manikpurage HD, Li Z, Gaudreault N, Armero VS, Boudreau DK, Renaut S, Henry C, Racine C, Eslami A, et al. Polygenic inheritance and its interplay with smoking history in predicting lung cancer diagnosis: a french-canadian case-control cohort. EBioMedicine. 2024;106:105234.
23. Kowarik A, Templ M. Imputation with the R Package VIM. J Stat Softw. 2016;74:1–16.
24. Gower JC. A general coefficient of similarity and some of its properties. Biometrics JSTOR. 1971:857–71.
25. De Leeuw J, Van Rijckevorsel J. Some generalizations of principal components analysis. In: Data analysis and informatics. Amsterdam: North-Holland Publishing Company; 1980. pp. 231–42.
26. Linting M, Meulman JJ, Groenen PJ, van der Koojj AJ. Nonlinear principal components analysis: introduction and application. Psychol Methods. 2007;12(3):336–58.
27. De Leeuw J. History of nonlinear principal component analysis. 2013.

28. De Leeuw J, Hornik K, Mair P. Isotone optimization in R: pool-adjacent-violators algorithm (PAVA) and active set methods. J Stat Softw. 2010;32:1–24.
29. Kruskal JB. Nonmetric multidimensional scaling: a numerical method. Psychometrika. 1964;29:115–29.
30. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nat Rev Genet. 2018;19(9):581–90.
31. Global Initiative for Asthma. Global strategy for asthma management and prevention. 2024. Available from: www.ginasthma.org. In
32. Ma Y, Zhou X. Genetic prediction of complex traits with polygenic scores: a statistical review. Trends Genet. 2021;37(11):995–1011.
33. Dai J, Lv J, Zhu M, Wang Y, Qin N, Ma H, He YQ, Zhang R, Tan W, Fan J, et al. Identification of risk loci and a polygenic risk score for lung cancer: a large-scale prospective cohort study in Chinese populations. Lancet Respir Med. 2019;7(10):881–91.
34. Dapas M, Lee YL, Wentworth-Sheilds W, Im HK, Ober C, Schoettler N. Revealing polygenic pleiotropy using genetic risk scores for asthma. HGG Adv. 2023;4(4):100233.
35. Dijk FN, Folkersma C, Gruzieva O, Kumar A, Wijga AH, Gehring U, Kull I, Postma DS, Vonk JM, Melen E, et al. Genetic risk scores do not improve asthma prediction in childhood. J Allergy Clin Immunol. 2019;144(3):857–e860857.
36. Moll M, Sordillo JE, Ghosh AJ, Hayden LP, McDermott G, McGeachie MJ, Dahlin A, Tiwari A, Manmadkar MG, Abston ED, et al. Polygenic risk scores identify heterogeneity in asthma and chronic obstructive pulmonary disease. J Allergy Clin Immunol. 2023;152(6):1423–32.
37. Kachuri L, Chatterjee N, Hirbo J, Schaid DJ, Martin I, Kullo IJ, Kenny EE, Pasaniuc B. Polygenic Risk Methods in Diverse Populations Consortium Methods Working G, Witte JS et al. Principles and methods for transferring polygenic risk scores across global populations. Nat Rev Genet. 2024;25(1):8–25.
38. Han Y, Jia Q, Jahani PS, Hurrell BP, Pan C, Huang P, Gukasyan J, Woodward NC, Eskin E, Gilliland FD, et al. Genome-wide analysis highlights contribution of immune system pathways to the genetic architecture of asthma. Nat Commun. 2020;11(1):1776.
39. Tsuo K, Zhou W, Wang Y, Kanai M, Namba S, Gupta R, Majara L, Nkambule LL, Morisaki T, Okada Y, et al. Multi-ancestry meta-analysis of asthma identifies novel associations and highlights the value of increased power and diversity. Cell Genom. 2022;2(12):100212.

Publisher's note