Published in final edited form as:
J Invest Dermatol. 2020 February ; 140(2): 477–479. doi:10.1016/j.jid.2019.08.432.

Identifying Phenotypes of Atopic Dermatitis in a Longitudinal US Cohort Using Unbiased Statistical Clustering

Ronald Berna, BS[1,2], **Nandita Mitra, PhD**[1,3], **Ole Hoffstad, MA**[1,2], **Joy Wan, MD MSCE**[1,2], **David J. Margolis, MD PhD**[1,2,3]

[1] Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2] Department of Dermatology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3] Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

TO THE EDITOR:

Machine learning is a form of artificial intelligence which utilizes mathematical and computational algorithms to identify patterns or predict outcomes within large datasets (Safran et al., 2018). Such algorithms perform comparably to humans on a number of different tasks, including the classification of skin cancer (Esteva et al., 2017; Safran et al., 2018). Cluster analysis is a form of machine learning which uses individual-specific attributes to group individuals such that individuals within the same group are more similar to each other than to individuals in other groups (Ding et al., 2018). Since cluster analysis

Corresponding author: Ronald Berna BS, Perelman School of Medicine, University of Pennsylvania, ronald.berna@pennmedicine.upenn.edu, 3400 Civic Center Boulevard, Philadelphia, PA 19104, 609-304-6024.

CRediT STATEMENT
Conceptualization: Berna R, Margolis DJ
Data Curation: Berna R, Hoffstad O
Formal Analysis: Berna R, Margolis DJ
Funding Acquisition: Margolis DJ
Investigation: Berna R, Margolis DJ
Methodology: Berna R, Margolis DJ, Mitra N
Project Administration: Berna R, Margolis DJ
Resources: Margolis DJ
Software: Berna R, Hofffstad O, Margolis DJ
Supervision: Margolis DJ, Mitra N
Validation: Berna R
Visualization: Berna R, Margolis DJ
Writing—Original Draft: Berna R, Margolis DJ
Writing—Review & Editing: Berna R, Margolis DJ, Mitra N, Wan J

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

CONFLICT OF INTEREST
David Margolis receives research funding as the principal investigator via the Trustees of the University of Pennsylvania (R01-AR060962, R01-AR070873, R01-DK116199) and from the NIH and Valeant Pharmaceuticals (PEER study) and Sunovion Pharmaceuticals. None of this funding was used for this study and these projects are not associated with the outcomes of this study. He has consulting activities primarily as a member of data monitoring boards or scientific advisory boards with Leo, Johnson and Johnson, Pfizer, Sanofi, Kerecis, and Cell Constructs. None of these activities are associated with the outcomes of this study. Joy Wan receives research funding from Pfizer, Inc for work unrelated to this study. No other authors state financial conflicts of interest with respect to this investigation.

can detect latent patterns within a group, it may identify phenotypes of diseases which would not be otherwise apparent.

Atopic dermatitis (AD) is a common chronic inflammatory skin disease for which disease severity and persistence vary with respect to patient phenotype and genotype (Garrett et al., 2013; Margolis et al., 2014b; Thorsteinsdottir et al., 2019). AD is increasingly recognized as a heterogeneous disease, so there is great interest in distinguishing the clinically relevant endotypes, which will be pivotal in the selection of future personalized therapies (Czarnowicki et al., 2019; Paternoster et al., 2018).

Previous studies have used outcome-informed clustering to subcategorize AD based on disease activity (Abuabara et al., 2018; Paternoster et al., 2018; Thijs et al., 2017). Our objective was to use an *unsupervised* clustering approach to identify phenotypically distinct subgroups within a cohort of children with AD and determine their associations with longitudinal AD clearance.

Data were obtained from the Pediatric Eczema Elective Registry (PEER) study, which has been previously described (Margolis et al., 2014a; Margolis et al., 2014b). To create clusters in an unbiased fashion, we used computational techniques which were hypothesis-independent. These techniques construct groups such that children within a group are more similar to each other than to children in other groups with respect to baseline covariates. Variables were of mixed types, so Gower’s similarity coefficient was used as the similarity index for clustering (Ding et al., 2018; Gower et al., 1971). Clustering was implemented with the Partitioning Around Medoids (PAM) algorithm (Kaufman et al., 1987). PAM is analogous to k-means, another approach to clustering, but it is appropriate for use with non-Euclidian distance metrics like Gower’s.

Clustering on all subject variables yielded poorly differentiated clusters as measured by silhouette width, a measure of cluster differentiation. Therefore, a subset of seven variables was used for clustering: race, age of AD onset, disease control at enrollment, AD symptom clearance in the six months prior to enrollment, medication allergies, maternal AD, and paternal AD. These seven variables were identified with a genetic algorithm (implemented via the rbga.bin function in R) which iteratively selected subsets of our study variables to identify a subset for which silhouette width was maximized. All variables were weighted equally to maintain the hypothesis-independent nature of the methods. White, African American, and Hispanic race/ethnicity were treated as independent binary variables.

To assess the robustness of the generated clusters, the PEER cohort was randomly divided into two subsets. Subset A (n = 4007) and Subset B (n = 4008) were similar in all variables studied. The PAM algorithm was independently implemented on each subset, and replication of clusters across subsets was assessed. Five clusters were identified in both subsets. Given the strong similarities between clusters in Subset A and Subset B, similar clusters were combined such that each subject was placed into one of five clusters (Table).

Disease clearance was defined using a self-reported outcome of whether or not a child’s skin was AD symptom free during the previous 6 months. To account for repeated outcome measures for each subject, we used generalized estimating equations (GEE) for binary outcomes to determine the association between the clusters and these outcomes, assuming an independence working correlation structure with empirical standard errors. All analyses were implemented in Stata, version 15.1 and R, version 3.5.3.

Cluster 1 (African American Active at Enrollment) described a predominantly African American group, with little AD clearance in the 6 months prior to enrollment, few medication allergies, and little maternal history of AD. This cluster was associated with the poorest long-term outcomes (Table). Cluster 2 (African American Indolent at Enrollment) described a predominantly African American cohort, with late AD onset, relatively high rate of AD clearance in the 6 months prior to enrollment, and no maternal history of AD. These individuals had high odds of AD clearance over time. Cluster 3 (African American with Early Onset and Atopic Family History) described a predominantly African American group with earlier AD onset and maternal AD. These individuals had moderate odds of AD clearance relative to the other two predominantly African American clusters. Cluster 4 (White without Atopic Family History) included only white individuals with no maternal AD and was associated with moderate AD clearance relative to the other clusters. Cluster 5 (White with High Income and Atopic Family History) was predominantly white with early AD onset, highest incidence of allergies, maternal history of AD, and highest income. This group, together with Cluster 2, had the highest odds of AD clearance in the long-term. All covariates studied were statistically significantly different across clusters (Table).

Similar to previous studies (Abuabara et al., 2018; Paternoster et al., 2018), we identified subgroups of AD characterized by varying disease clearance. Severe disease at outset, atopic family history, race, and low family income were associated with more persistent AD, consistent with the literature on AD (Abuabara et al., 2018; Kim et al., 2019; Margolis et al., 2014b; Paternoster et al., 2018; Thorsteinsdottir et al., 2019).

While previous reports used long-term outcomes to retrospectively identify subphenotypes of AD, we constructed our clusters with only data available at enrollment, and examined outcomes thereafter (Abuabara et al., 2018; Paternoster et al., 2018; Thijs et al., 2017). Our clusters thus represent unbiased groupings of individuals with AD. We found that there is an African American subgroup with disease persistence comparable to whites (Cluster 2), a finding not observed in other reports (Abuabara et al., 2018; Kim et al., 2019). In addition, we uncovered an early onset and highly allergic group (Cluster 5) with high odds of disease clearance, suggesting that early AD onset and comorbid allergies are not necessarily predictive of more severe disease course. Together, these findings suggest that predicting AD activity based on race, AD onset age, or allergic burden alone may be inadequate, and a more holistic approach accounting for multiple phenotypic variables is warranted.

There are several limitations of this report. First, the PEER cohort is not a random sample of individuals with AD, so it is possible that our clusters are not generalizable to the greater AD population. The PEER population has slightly more males than the general AD population (46.7% vs. 38% in US AD population), is enriched for individuals with asthma (45.8% vs. approximately 30% in US AD population) and may be skewed towards individuals with more severe AD (Fuxench et al., 2019; Garrett et al., 2013; Margolis et al., 2014b). Second, all children enrolled in PEER had to receive at least 6 weeks of pimecrolimus cream in the 6 months prior to enrolling in the study (Margolis et al., 2014). It is possible that this therapy and any other concomitant treatments for AD may have had differential effects on the subgroups we identified, thus modifying the association between different subgroups and disease outcomes. However, 45% of PEER participants were no longer using pimecrolimus by the third year of follow-up, so we think this is unlikely (Margolis et al., 2014a).

To our knowledge, this is one of the first studies to use hypothesis-independent statistical techniques to identify distinct clusters of patients with long-term AD outcome differences in a data set of more than 8000 children. Because our subgroups are defined by phenotypic characteristics available early in the AD course, they may be useful for disease prognostication. These clusters could also be used in genetic analyses to reduce the number of relevant variants examined, by removing variants that are evenly distributed across clusters. This study demonstrates that hypothesis-independent clustering can be used to differentiate patterns in the presentation and clinical outcomes of children with AD. Similar methodology may be valuable in the study of other skin diseases.

DATA AVAILABILITY

The PEER data (source) is not currently publically available. The PEER study is an ongoing study sponsored by Valeant in response to a post marketing commitment with the FDA.

ACKNOWLEDGMENTS

This study was funded by support from the National Institutes of Health National Institute of Arthritis and Musculoskeletal and Skin Diseases R01-AR069062, P30-AR069589, and T32-AR007465. The sponsor did not have a role in the design of the study, the collection, analysis, or interpretation of the data, the preparation of this report, or the decision to submit this report for publication.

Abbreviations:

AD	atopic dermatitis
CI	confidence interval
PEER	Pediatric Eczema Elective Registry
PAM	Partitioning around medoids
GEE	generalized estimating equations

REFERENCES

Abuabara K, Hoffstad O, Troxel AB, Gelfand JM, McCulloch CE, Margolis DJ. Patterns and predictors of atopic dermatitis disease control past childhood: an observational cohort study. J Allergy Clin Immunol. 2018;141(2):778–779. [PubMed: 28629748]

Czarnowicki T, He H, Krueger JG, Guttman-Yassky E. Atopic dermatitis endotypes and implications for targeted therapeutics. J Allergy Clin Immunol. 2019;143(1):1–11. [PubMed: 30612663]

Ding L, Wathen M, Altaye M, Mersha TB. African ancestry is associated with clusterbased childhood asthma subphenotypes. BMC Med Genom. 2018; 11(51):1–11.

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017; 542:115–118. [PubMed: 28117445]

Fuxench ZCC, Block JK, Boguniewicz M, Boyle J, Fonacier L, Gelfand JM, et al. Atopic Dermatitis in America Study: A Cross-Sectional Study Examining the Prevalence and Disease Burden of Atopic Dermatitis in the US Adult Population. J Invest Dermatol. 2019;139(3):583–590. [PubMed: 30389491]

Garrett JP-D, Apter AJ, Hoffstad O, Spergel JM, Margolis DJ. Asthma and frequency of wheeze: risk factors for the persistence of atopic dermatitis in children. Ann Allergy Asthma Immunol. 2013;110(3):146–149. [PubMed: 23548521]

Gower JC. A general coefficient of similarity and some of its properties. Biometrics. 1971;27(4):857–871.

Kaufman L, Rousseeuw PJ. Clustering by means of medoids In: Dodge Y, ed. Statistical data analysis based on the L1 norm and related methods. Amsterdam, North Holland: Elsevier; 1987:405–416.

Kim Y, Blomberg M, Rifas-Shiman SL, Camargo CA Jr, Gold DR, Thyssen JP, et al. Racial/Ethnic Differences in Incidence and Persistence of Childhood Atopic Dermatitis. J Invest Dermatol. 2019;139(4):827–834. [PubMed: 30414911]

Margolis DJ, Kim B, Apter AJ, Gupta J, Hoffstad O, Papadopoulos M, et al. Thymic Stromal Lymphopoietin Variation, Filaggrin Loss of Function, and the Persistence of Atopic Dermatitis. JAMA Dermatol. 2014a;150(3):254–259. [PubMed: 24401911]

Margolis JS, Abuabara K, Bilker W, et al. Persistence of mild to moderate atopic dermatitis. JAMA Dermatol. 2014b; 150(6):593–600. [PubMed: 24696036]

Paternoster L, Savenije OEM, Heron J, Evans DM, Vonk JM, Brunekreef B, et al. Identification of atopic dermatitis subgroups in children from 2 longitudinal birth cohorts. J Allergy Clin Immun. 2018; 141(3):964–971. [PubMed: 29129583]

Safran T, Viezel-Mathieu A, Corban J, Kanevsky A, Thibaudeau S, Kanevsky J. Machine learning and melanoma: the future of screening. J Am Acad Dermatol. 2018; 78(3):620–621. [PubMed: 28989109]

Thijs JL, Strickland I, Bruijnzeel-Koomen CAFM, Nierkens S, Giovannone B, Csomor E, et al. Moving toward endotypes in atopic dermatitis: identification of patient clusters based on serum biomarker analysis. J Allergy Clin Immun. 2017;140(3):730–737. [PubMed: 28412391]

Thorsteinsdottir S, Stokholm J, Thyssen JP, Norgaard S, Thorsen J, Chawes BL, et al. Genetic, Clinical, and Environmental Factors Associated With Persistent Atopic Dermatitis in Childhood. JAMA Dermatol. 2019;155(1):50–57. [PubMed: 30427975]

Table:

Composition of clusters obtained after combining like clusters from Subsets A and B. Table entries for dichotomous variables indicate number of individuals followed by percent in that cluster. Values for non-binary variables (age of onset, disease control, and income) represent the mean followed by standard deviation in that cluster.

Variables	Cluster 1: African American Active at Enrollment	Cluster 2: African American Indolent at Enrollment	Cluster 3: African American with Early onset and Atopic Family History	Cluster 4: White without Atopic Family History	Cluster 5: White with High Income and Atopic Family History	p value
	N=2054	**N=1931**	**N=975**	**N=2325**	**N=730**	
	n (%)	**n (%)**	**n (%)**	**n (%)**	**n (%)**	
male sex	885 (43.1)	862 (44.6)	435 (44.6)	1199 (51.6)	360 (49.3)	<0.001
African American race	1755 (85.4)	1636 (84.7)	781 (80.1)	87 (3.7)	41 (5.6)	<0.001
white race ‡	0 (0)	2 (0.1)	100 (10.3)	2325 (100)	724 (99.2)	<0.001
Presence of flexural AD	1107 (53.9)	818 (42.4)	560 (57.4)	1557 (67)	512 (70.1)	<0.001
age at AD onset ‡, years, mean (SD)	2.196 (2.99)	2.696 (3.26)	1.918 (2.84)	2.399 (3.26)	2.123 (2.80)	<0.001
AD disease control at enrollment ‡*, mean (SD)	3.019 (0.59)	2.071 (0.50)	2.658 (0.73)	2.453 (0.75)	2.332 (0.69)	<0.001
complete AD clearance in 6 months prior to enrollment ‡	237 (11.5)	1593 (82.5)	415 (42.6)	1220 (52.5)	494 (67.7)	<0.001
history of wheezing, asthma, or whistling	810 (39.4)	841 (43.6)	489 (50.2)	1097 (47.2)	434 (59.5)	<0.001
seasonal allergies	1067 (51.9)	1128 (58.4)	630 (64.6)	1524 (65.5)	585 (80.1)	<0.001
pet allergies	302 (14.7)	196 (10.2)	205 (21)	723 (31.1)	255 (34.9)	<0.001
medication allergies ‡	148 (7.2)	125 (6.5)	82 (8.4)	360 (15.5)	190 (26)	<0.001
food allergies	409 (19.9)	294 (15.2)	234 (24)	656 (28.2)	227 (31.1)	<0.001
child’s mother has eczema ‡	3 (0. 1)	0 (0)	975 (100)	0 (0)	730 (100)	<0.001
child’s father has eczema ‡	243 (11.8)	160 (8.3)	168 (17.2)	299 (12.9)	146 (20)	<0.001
child’s mother has asthma	256 (12.5)	273 (14.1)	320 (32.8)	391 (16.8)	283 (38.8)	<0.001
family income ♦, USD, mean (SD)	1.444 (0.88)	1.387 (0.86)	1.643 (1.07)	2.498 (1.38)	2.822 (1.45)	<0.001
Odds Ratio (95% CI) of AD clearance °	reference	3.52 (3.20, 3.87)	1.65 (1.48, 1.84)	2.54 (2.33, 2.77)	3.22 (2.84, 3.66)	

‡variable used for clustering

*measured on a scale of 1 to 4, with 1 denoting “complete disease control” and 4 denoting “uncontrolled disease”

♦measured on a 5 point scale, with 1 = “<24,000 USD”, 2 = “25,000–49,999 USD”, 3 = “50,000–74,999 USD”, 4 = “75,000–99,999 USD”, 5 = “>100,000 USD”

[o] Generalized estimating equations comparing longitudinal clearance of AD among individuals within a cluster to that among individuals in Cluster 1.