Disaggregating asthma: Big investigation versus big data

Danielle Belgrave, PhD,[a] John Henderson, MD,[b] Angela Simpson, MD, PhD,[c] Iain Buchan, MD, PhD,[d] Christopher Bishop, PhD,[e] and Adnan Custovic, MD, PhD, FAAAI[a] *London, Bristol, Manchester, and Cambridge, United Kingdom*

We are facing a major challenge in bridging the gap between identifying subtypes of asthma to understand causal mechanisms and translating this knowledge into personalized prevention and management strategies. In recent years, "big data" has been sold as a panacea for generating hypotheses and driving new frontiers of health care; the idea that the data must and will speak for themselves is fast becoming a new dogma. One of the dangers of ready accessibility of health care data and computational tools for data analysis is that the process of data mining can become uncoupled from the scientific process of clinical interpretation, understanding the provenance of the data, and external validation. Although advances in computational methods can be valuable for using unexpected structure in data to generate hypotheses, there remains a need for testing hypotheses and interpreting results with scientific rigor. We argue for combining data- and hypothesis-driven methods in a careful synergy, and the importance of carefully characterized birth and patient cohorts with genetic, phenotypic, biological, and molecular data in this process cannot be overemphasized. The main challenge on the road ahead is to harness bigger health care data in ways that produce meaningful clinical interpretation and to translate this into better diagnoses and properly personalized prevention and treatment plans. There is a pressing need for cross-disciplinary research with an integrative approach to data science, whereby basic scientists, clinicians, data analysts, and epidemiologists work together to understand the heterogeneity of asthma. (J Allergy Clin Immunol 2017;139:400-7.)

Key word: *Asthma, endotypes, machine learning, big data, birth cohorts*

A major obstacle to realizing precision (stratified or personalized) medicine in asthmatic patients is the lack of consensus in defining the disease, which is, at least in part, a consequence of "asthma" being an aggregated diagnosis comprising several different diseases.[1-4] It is now well established that both asthma[3,5-8] and allergic sensitization[9-12] are umbrella terms (or syndromes) incorporating a variety of underlying endotypes sharing common symptoms and phenotypic characteristics.[13,14] Although by definition each endotype has unique pathophysiology and hence genetic and environmental associations,[13,14] it is likely that some mechanisms overlap 1 or more endotypes.[15] This underlying heterogeneity is also reflected in responses to treatment. For example, a therapeutic agent might be specific for a pathway that is primarily responsible for the patient's asthma subtype, and therapeutic response can be predicted reasonably well by using relevant biomarkers,[16,17] such as the number of eosinophils in peripheral blood or sputum for mepolizumab[18] or periostin levels for lebrikizumab.[19] Alternatively, a therapeutic agent might be relatively nonspecific and target broad mechanisms shared between different asthma endotypes, in which case patients across different endotypes might display a spectrum of responses, which is likely the case with inhaled corticosteroids.

Across different disease areas, a vast number of genetic studies have initially raised expectations over "significant hits" that later delivered neither meaningful clinical diagnostic tools nor useful insights into disease pathogenesis.[20] Genetic studies have thus far explained little of the heritability of complex diseases.[21] Associated genetic variants generally have small effect sizes, and for many of these genetic variants, there is a lack of clear functional implication. In addition to gene-environment interactions,[22] gene-environment correlations,[23] and epigenetic mechanisms,[24] the use of aggregated definitions of disease can also contribute to inconsistent findings between studies investigating genetic components of asthma. However, by using more specific phenotyping, a recent genome-wide association study identified an association of a specific asthma subtype characterized by early-life onset and recurrent severe exacerbations at preschool age, with a functional variant in the novel susceptibility gene *CDHR3* (rs6967330, C529Y).[25] This genetic variant was associated with a greater risk of asthma hospitalizations in 2 birth

From [a]the Department of Paediatrics, Imperial College, London; [b]the School of Social and Community Medicine, Faculty of Health Sciences, University of Bristol; [c]the Division of Infection, Immunity and Respiratory Medicine, and [d]Health Informatics, Faculty of Biology, Medicine and Health, University of Manchester; and [e]Microsoft Research, Cambridge.

D.B. is supported by Medical Research Council Career Development Fellowship MR/M015181/1. The STELAR consortium is funded by Medical Research Council grant MR/K002449/1.

Disclosure of potential conflict of interest: D. Belgrave receives grant support from Medical Research Council Career Development Fellowship MR/M015181/1, serves as a consultant for GlaxoSmithKline, and receives payment for lectures from GlaxoSmithKline. I. Buchan receives grant support from MRN and Microsoft. C. Bishop is an employee of Microsoft Research and has stock options with Microsoft Research. A. Custovic serves as a consultant for Novartis, Regeneron/Sanofi, and ALK-Abelló and received speaker fees from Bayer and Thermo Fisher. The rest of the authors declare that they have no relevant conflicts of interest.

Received for publication September 30, 2016; revised November 7, 2016; accepted for publication November 9, 2016.

Available online November 18, 2016.

Corresponding author: Adnan Custovic, MD, PhD, FAAAI, Imperial College London, Department of Paediatrics, St Mary's Campus Medical School, London W2 1PG, United Kingdom. E-mail: a.custovic@imperial.ac.uk.

The CrossMark symbol notifies online readers when updates have been made to the article such as errata or minor corrections

0091-6749

http://dx.doi.org/10.1016/j.jaci.2016.11.003

Abbreviation used
STELAR: Study Team for Early Life Asthma Research

cohorts, but there was no association with an aggregated definition of "doctor-diagnosed asthma." Subsequent studies have shown that expression of human *CDHR3* facilitates rhinovirus C binding and replication and that a coding single nucleotide polymorphism in *CDHR3*, which was linked with asthma hospitalizations in birth cohort studies, mediates enhanced rhinovirus C binding and increased progeny yields *in vitro*.[26] It is also of note that when asthma was disaggregated into subtypes, much stronger associations were observed for some of the genetic variants previously identified in genome-wide association studies, such as those in the 17q21 locus.[25] The value of focusing on specific subgroups has been demonstrated in a study that showed that variants at 17q21 were associated with asthma but only in children who had rhinovirus-induced wheezing illness.[27] Similarly, the risk of transient early wheeze, but not persistent wheeze, increases with the number of chronic obstructive pulmonary disease–associated alleles.[28] Most of the genetic studies that used more precise phenotypes showed higher relative risk estimates than the modest effect sizes of genetic hits that were identified by using a simple binary trait definition of asthma, highlighting the need for a more refined subtyping of asthma to accurately identify genetic variants of clinical importance.[29]

Many environmental exposures are implicated in the development of asthma and in determining its severity.[30,31] As with genetic associations, there have been many inconsistent reports about the role of environmental exposures in asthmatic patients. We and others have shown that different phenotypes of childhood wheezing have different environmental associations.[2,8,32-38] Similarly, different subtypes of atopic sensitization differ in their environmental risk factors; for example, endotoxin exposure is protective for multiple early but not multiple late sensitizations.[39] It is likely that the effect of most environmental factors varies across subjects with different genetic predispositions, but the precise nature of most gene-environment interactions remains unclear.[22] One of the most replicated findings of gene-environment interactions in the development of allergic sensitization is between *CD14* variants and environmental endotoxin exposure.[40] Several studies have reported that high endotoxin exposure can protect against sensitization but only among subjects with a specific genetic predisposition (C allele homozygotes of rs2569190).[40,41] However, in the same genotype group the effect of endotoxin exposure differed by phenotype, decreasing the risk of atopic sensitization and eczema but increasing the risk of nonatopic (but not atopic) wheezing.[41] Other examples that the nature of gene-environment interactions can differ between different wheeze phenotypes include the finding that day care attendance can have opposite effects on atopic wheezing among subjects with different genetic variants in the Toll-like receptor 2 gene (being protective in some but increasing the risk in others),[42] with no such effect being observed for nonatopic wheezing.[42] This suggests that replication of gene-environment interactions can be improved through a more precise definition of the outcome of interest.[43] The lessons for intervention studies aimed at personalized prevention is that individual genetic predisposition must be taken into account when seeking the environmental protective/susceptibility factors amenable to intervention[30] and that interventions that might be effective in one subtype of wheezing might not necessarily work for other subtypes.

One area that has been relatively more successful is the identification of biomarkers[16] for more targeted treatment strategies.[17] A recent review Berry and Busse[44] identified 4 main biomarkers that might help optimize treatment strategies for different asthma phenotypes. These biomarkers are generally limited to T2 mechanisms: eosinophils, exhaled nitric oxide, periostin, and IgE. However, biomarker assessment has not as yet become an integral part of clinical practice, nor is it reflected in current asthma guidelines. Validation steps are necessary, and acknowledgement in asthma guidelines would prompt application of such information in clinical practice. The identification of non-T2 biomarkers is an important area of research that needs to be exploited[44] with biomarker identification for asthma and allergic diseases still in its embryonic stages. Furthermore, although biomarker identification has indeed led to more targeted asthma treatment strategies, there are currently no biomarkers that reflect the underlying causal mechanisms, which could predict disease onset or progression.

Although phenotypic heterogeneity of asthma is now widely accepted, we are still scratching the surface of identifying the different endotypes of asthma and understanding their unique underlying pathophysiologic mechanisms, which is a prerequisite for precision medicine.[15] Although there is general consensus that there are different asthma endotypes and different phenotypes of wheezing during childhood, there is no consensus on how best to define them. A more refined endotypic definition of asthma and allergic diseases can drive more targeted research to identify distinct molecular, genetic, environmental, and demographic characteristics that might allow us to predict causality of distinct endotypes with greater accuracy.[45]

One approach used in a number of studies has been to investigate temporal patterns of symptoms over time. The common labels across most studies have been transient early wheeze, late-onset wheeze, and persistent wheeze.[46] However, different studies reported different numbers of childhood wheeze phenotypes (eg, ranging between 2 and 6).[2,46,47] One of the challenges in current research aimed at defining subgroups of patients based on the natural history of wheezing is the lack of consistency in definition of these phenotypes and what they represent. The inconsistency in defining wheeze phenotypes based on longitudinal profiles of symptoms over time across different studies might merely reflect inconsistencies in the nature and timing of questions used (eg, physician-confirmed wheezing[8,34] vs parentally reported wheezing[6,36]). Thus although the definition of subtypes based on profiles of symptoms over time is better than that based on a single time point, variability in input variables has an effect on the accuracy of defining subtypes and identifying predictive models.[2,47-49]

CAN "BIG DATA" PROVIDE SOLUTIONS?

Big data refers not only to the ready availability of large volumes of routine health care data being rapidly generated but also to the complexity of these data, which is evident in the amplified scale of biological, genetic, environmental, and phenotypic data. The scale of these data often makes handling, management, and analysis challenging with the use of standard