Belgrave, D., Henderson, J., Simpson, A., Buchan, I., Bishop, C., & Custovic, A. (2017). Disaggregating Asthma: Big Investigation vs. Big Data. *Journal of Allergy and Clinical Immunology*, *139*(2), 400-407. https://doi.org/10.1016/j.jaci.2016.11.003

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

statistical methods. The evolution of powerful computational tools to analyze such high-dimensional large data sets has pushed the boundaries of endotype discovery. Such data provide the potential for "learning" patterns or predicting health outcomes and optimal treatment strategies based on prior information. However, one of the major challenges of big data remains the bias inherent to its volume. Furthermore, the vast increase in the quantity of data generated has made it impossible at times to know for what we are looking and what questions need to be asked. As a consequence, data-driven hypothesis-generating approaches to understanding disease are overshadowing traditional hypothesis-based research (hypothesis testing) through carefully constructed questions and observations. In a hypothesis-generating approach to data analysis, we look for structure in the data without necessarily having a specific research hypothesis we want to verify. This is an advantage where, for example, we have measures of multiple biomarkers but are uncertain of the role of these biomarkers in predicting asthma. A hypothesis-generating approach can be used to identify patterns in biomarkers (eg, which ones are similar or which ones modify the effect of other biomarkers) to predict the disease. 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. However, we argue for combining data-driven and hypothesis-driven methods in careful synergy.

ON METHODOLOGIES: UNDERSTANDING REALITY VERSUS PREDICTING THE FUTURE

Machine learning, computational statistics, biostatistics, a traditional approach to epidemiology, and clinical and biological expertise can elucidate different aspects of the same problem. Machine learning is a data-driven approach to identify structure within data to make predictions and identify patterns. It is used commonly by computer scientists for problem solving in a variety of fields and is used increasingly to disaggregate complex disease phenotypes in respiratory medicine and allergy.[1,3,5,10-12] It must be noted that although machine learning as a discipline is fairly new, the mathematic and statistical foundations have been in existence since the beginning of the 20th century.[50-53] Machine learning as a new discipline is a result of the exponential growth in computational power, which has enabled implementation of the mathematic groundwork that was initiated decades earlier.[54,55]

One of the (somewhat artificial) distinctions between machine learning and conventional statistical approaches is that although machine learning focuses on prediction models and attempts to accurately predict future outcomes/events (whether these be future disease states or the development of symptoms in later life), statistics tends to focus on extant observations and constructing models to aid understanding of the data and the current status of disease. Hence statistics tends to focus on causality and associations in an attempt to explain the disease and to understand uncertainty in the modelling assumptions.

Modelling assumptions refers to our framework for representing the data or research questions related to that data. Because these are assumptions, we are uncertain about them and need some way of testing whether these assumptions are true. Machine learning applied to medicine attempts to predict disease states and to get the best estimate of uncertainty analogous to clinical diagnosis. Both approaches combined with epidemiology, which carefully tests hypotheses to infer causality, need to be considered along with medical and biological expertise in a holistic understanding of disease.

BAYESIAN VERSUS FREQUENTIST APPROACH TO UNDERSTANDING DISEASE ETIOLOGY

We now introduce the reader to 2 different approaches to hypothesis testing and hypothesis generation: the Bayesian and frequentist approaches. The aim of this discussion is to provide a conceptual framework that is currently commonly used in statistical and machine learning and can be applied to both big and small data sets in health care research. An understanding of these 2 paradigms is formative for a team approach to understanding disease etiology in health care.

The frequentist paradigm is an unconditional perspective, meaning that it assumes that the observed data are representative of the population with an independent and identical distribution. Thus this paradigm, as the name suggests, emphasizes the frequency of the data. On the other hand, the Bayesian approach uses probability as a principled framework for quantifying our uncertainty of the data and of the true estimated effects in our models, thereby allowing the explicit incorporation of prior scientific knowledge into statistical reasoning. Bayesian models provide a framework whereby prior knowledge and data from previous studies can be incorporated explicitly with the data at hand in the analytic model to formulate a posterior distribution that takes account of both the observed data and prior knowledge.[56] The inherent characteristics of the Bayesian approach to data analysis make this framework more amenable to handling large-scale problems and easily extending the complexity of current models that use classical statistical or frequentist tools. Although the frequentist approach also relies on prior clinical knowledge, the difference is that this approach does not seek to explicitly quantify this knowledge; it only relies on the data at hand for incorporating assumptions we make about the statistical models for the data.

In understanding the etiology of asthma and allergic disease, Bayesian models provide a flexible and unified framework for understanding the probability of disease manifestation and comanifestation, incorporating evidence from the literature or hypotheses from medical experts[5] through the explicit quantification of this evidence. However, although Bayesian methods provide an intuitive and unified platform for carrying out statistical research, the results are often computationally intractable and resolving these is active area of research.[57-60] The exponential increase in computational power and the increasing availability of tools that can handle large-scale data has facilitated the use of Bayesian methods, and it is important that we capitalize on these relatively complex tools to improve human health.

The use of Bayesian statistics is relatively uncommon in the medical literature, in part because of the greater complexity involved in using these models. One of the strengths of the Bayesian approach is that it can be used to enrich our current understanding of disease, with its capacity to elicit robust scientific inference by encouraging the user to think about the underlying statistical and scientific problems[61] by assigning explicit quantities to scientific assumptions. This might provide a powerful tool for extending model complexity to reflect the underlying complexity of the data and the scientific problem

being addressed. Bayesian methods allow the clinician to take an active role in the modelling process by quantifying prior probabilities based on expert assessments. However, one limitation of Bayesian analysis is the difficulty in eliciting this prior knowledge and quantifying expert knowledge,[62,63] mainly because pf the training and time necessary to develop "informative" prior knowledge. In some cases this can be more expensive than collecting more data. This approach is not unique to Bayesian analysis. Prior knowledge can be integrated to a less explicit degree by using a frequentist platform, where, rather than specifying or quantifying expected results, a clinician/topic expert could specify explicit assumptions about expected transitions of allergic disease and symptom profiles, as well as the proportions of patients with different profiles and severities of disease. In this sense the often-acclaimed advantage of Bayesian analysis as being able to incorporate informative prior knowledge based on expert knowledge may be overstated. The frequentist approach can be used to compare different model assumptions based on expert knowledge, which might be a more pragmatic approach than trying to quantify uncertainty surrounding the size of an effect. The important take-home message is that in weighing up the Bayesian and frequentist approach to statistical modelling, the question is not so much which method is best but rather which method is more appropriate for the question being addressed and for encapsulating model complexity with parsimony.[64]

AWAY FROM METHODOLOGICAL POLEMICS TOWARD DATA SCIENCE

This dissonance between machine learning, biostatistics, and epidemiology on the one hand and between the Bayesian and frequentist paradigms on the other presents artificial dichotomies. Beyond the methodological dogma, science needs to be pragmatic, selecting the right method or methods for the problem/question. Different methodologies are not mutually exclusive; indeed, an ensemble of methods might be more effective for identifying distinct subtypes of diseases. Data science must take the path of least inferential resistance, including the use of better ways to incorporate prior knowledge about likely causal mechanisms.

LATENT VARIABLE MODELLING APPROACH TO UNDERSTANDING SUBTYPES OF DISEASE

A general area in which Bayesian and frequentist paradigms compete (or complement) is latent variable modelling.[65] This section highlights the importance of latent variable modelling as a generalized framework for hypothesis generating and dimensionality reduction. Dimensionality reduction is an important tool for analysis of big data, in which we have multiple clinical, molecular, genetic, environmental, and phenotypic elements (ie, high dimensions). As the name suggests, in dimensionality reduction the aim is to reduce the dimension of the data set to a more manageable group of meaningful variables. Latent variable modelling can also be used not just to reduce the dimension of variables within a large data set but also to identify subgroups of patients based on patterns within these variables. Latent variable models are increasingly cited in the medical literature[66-68] for classifying different phenotypes and subphenotypes of diseases based on individual disease profiles. The latent variable model is a statistical model in which the observed association between (manifest) observed variables is regarded as spurious because this observed association can be explained by an indirectly observed, hidden, or latent variable rather than being causally related. This provides a powerful approach to probabilistic modelling and offers a flexible method to investigate substructures within complex data sets, in which associations between a set of observed variables are supplemented with additional latent variables. Therefore latent variable modelling allows us to move from hypothesis testing to hypothesis generation.[69] A further advantage of using latent variable modelling is that it is easier to represent high-dimensional parameters[70-73] on a reduced space with fewer dimensions. For example, using such techniques, we can reduce the dimensionality of multiple continuous variables into a more manageable set with fewer variables (parameters). The reduced number of variables is representative of a larger data set. Reducing dimensionality onto a latent space in turn facilitates the interpretation of multiple correlated continuous factors. The use of Bayesian methods in this context complements the likelihoods from the data with prior hypotheses about the expected distribution of these latent variables. We have successfully used generalized latent variable modelling approaches to identify distinct subtypes of asthma[2,3,6,9-12,15,34,36] and allergic diseases.[5,9,11,12] The key to future discoveries is to uncover underlying pathophysiologic mechanisms (endotypes) that drive these distinct subtypes.[1]

BIG DATA WITH BIG PROMISES: THE CONTRIBUTION OF COHORTS TO OUR UNDERSTANDING OF ASTHMA

The public's expectation that their health data should be used to improve care services has sometimes been stalled by fears over privacy and unregulated commercial uses of the data.[74] Birth cohort studies are an interesting parallel because cradle-to-grave health care records can be thought of in this way. However, unlike routine health care records, birth cohorts make more systematic observations before the onset of disease, facilitating exploration of the natural history of disease. With data from birth cohorts, investigators can follow development of disease over time, which mimics clinicians' diagnoses and follow-up observations but in a more anticipatory way.

One initiative aimed at harnessing data from birth cohorts to understand the development of different endotypes of asthma and allergic diseases is the Study Team for Early Life Asthma Research (STELAR) consortium.[15] STELAR combines data from 5 United Kingdom birth cohorts aimed at understanding the development of asthma and allergic diseases through the life course. The cohorts include the Avon Longitudinal Study of Parents and Children, Ashford cohort, Isle of Wight cohort, Manchester Asthma and Allergy Study, and Aberdeen Study of Eczema and Asthma to Observe the Effects of Nutrition. STELAR has data on more than 14,000 participants with repeated measures on symptoms of asthma and allergy over multiple time points from childhood into adulthood. An important feature is that participants are sampled from the general population, enabling generalizable conclusions about the pathophysiology and development of asthma at large. This would be difficult with routine health care records because they have more selected/biased observations sampled later in the natural history of disease. Fig 1 summarizes the challenges in understanding asthma and allergic disease that will drive future research in the STELAR consortium.

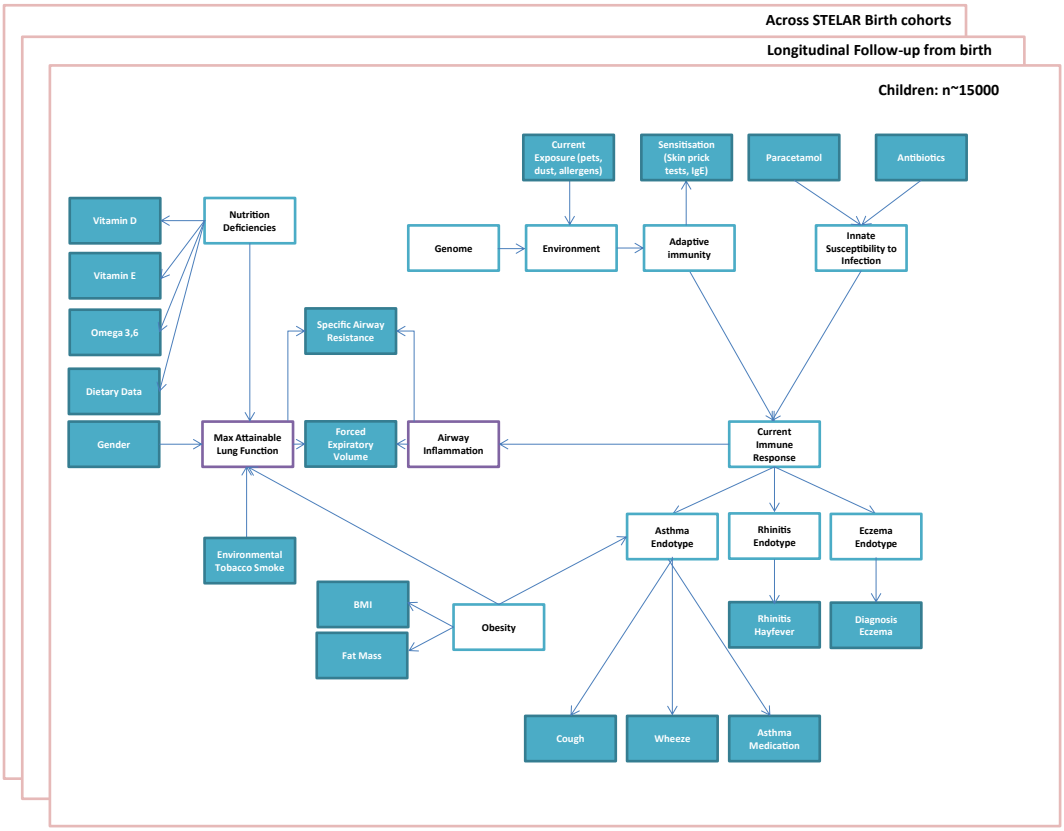

FIG 1. Roadmap of challenges to understanding asthma and allergic diseases within the STELAR consortium.

An important area in which recent cohort studies have elucidated pathways for development of asthma into fixed airway obstruction is in investigating longitudinal profiles of lung function.[75-77] Such profiles can shed light on the causes and consequences of airway obstruction that provide us with an objective marker of airway disease, which can be easily translated into clinical practice.

We consider that clinical/case (patient) cohorts and birth cohorts provide complementary windows on different aspects of understanding disease etiology.[78] An important and largely unanswered question is how best to translate findings between case and birth cohorts (ie, between clinical and general populations) to inform better prevention and early intervention strategies.[79] The case has been argued for automated methods to update disease models in real time.[80-84] The technologies are available, but they have not been applied in this way to accelerate the translation of research findings into clinical practice nor have they been used to enrich research models with emergent clinical phenomena. The importance of carefully characterized birth and patient cohorts with genetic, phenotypic, biological, environmental, and molecular data cannot be overemphasized in the quest to understand asthma and discover its endotypes.

CONCLUSION: THE IMPORTANCE OF TEAM SCIENCE

We are facing a major challenge to bridge the gap between identifying subtypes of asthma in clinical and general populations (and to find ways to translate the findings between these 2 contexts) to understand causal mechanisms of the discovered subtypes and translating this knowledge into better prevention and management strategies.[78,85] To this effect, understanding disease causality within the data analytic framework is fundamental to improve our understanding of asthma endotypes and their distinct etiologies.[86] From this perspective, significant investment needs to be made in advancing statistical and computational tools to solve health care problems. However, although advances in computational methods can be valuable for identifying unexpected structure in data to generate hypotheses, there remains a need for interpreting results with scientific rigor and testing hypotheses that arise from this process. One of the dangers of ready accessibility of health care data and computational tools for analyzing these data 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.[87] There is a pressing need for cross-disciplinary research to avoid the false idol of big data being the single source of truth. A more credible approach is to blend big data with big reasoning, so that prior structure is imposed on the data meaningfully. Fig 2 illustrates a data cycle encompassing the problem: an integrative approach to data science whereby basic scientists, clinicians, biostatisticians, and epidemiologists work together to understand the heterogeneity of asthma and allergic disease. Given that big data takes team science, it would be important for the scientific and academic communities to reassess systems and criteria for promotion that still, in many cases, do not give sufficient credit for perceived nonleadership roles.

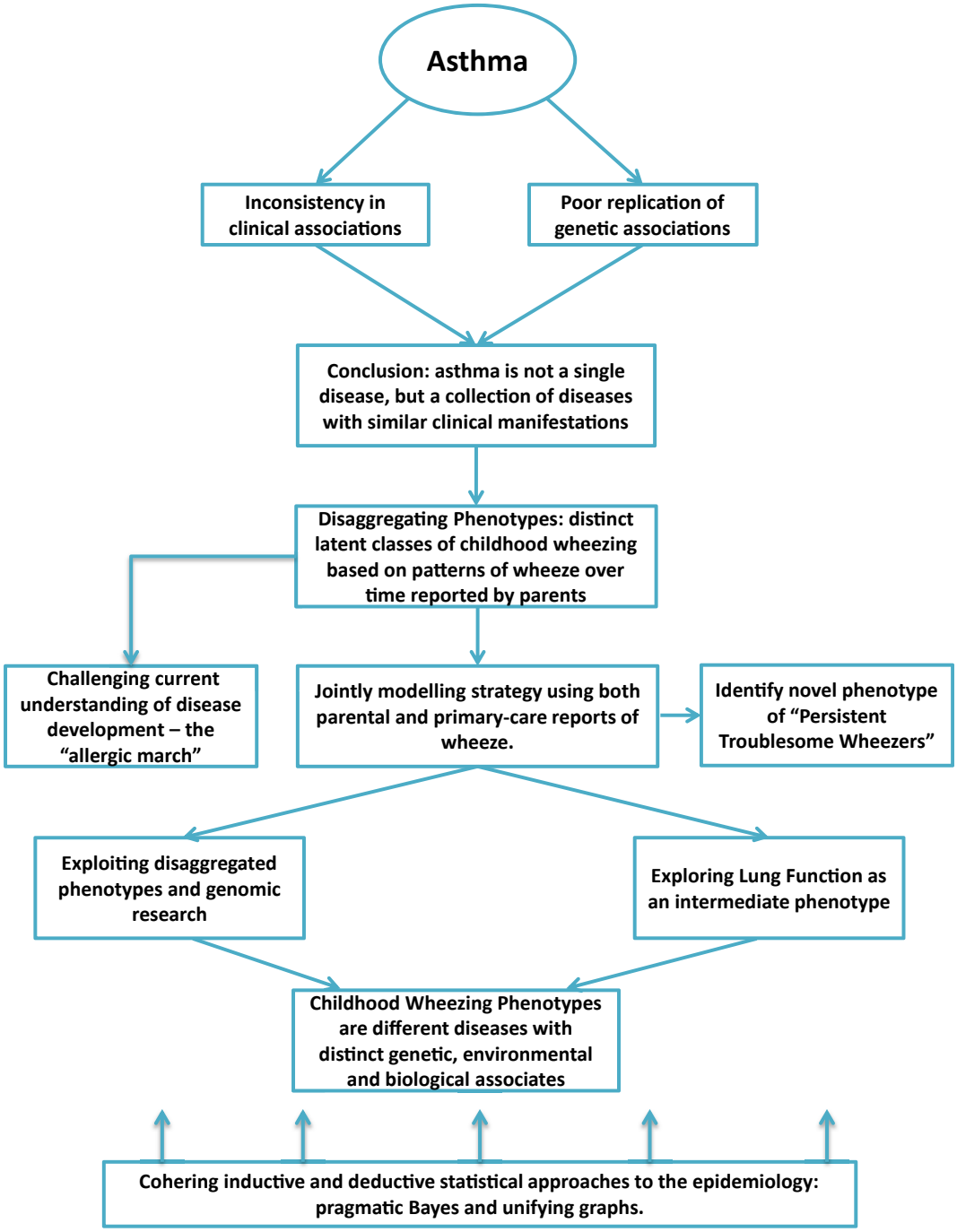

FIG 2. Data cycle: an integrative approach to understanding disease endotypes.

We need to ensure that we harness bigger health care data in ways that produce meaningful clinical interpretation and to translate the findings into better diagnoses, biomarkers, and properly personalized prevention and treatment plans. As an example, big data could be used to identify patients with exacerbations and inadequately controlled asthma and then prompt evaluation of their treatment regimen.[17] One of the advantages of big data is its capacity to change the way we currently do clinical research in asthma through building more robust predictive models to understand subtypes of this complex disease. The direction needs to move away from looking at average effects (which is a strategy commonly used in randomized clinical trials that make use of stringent exclusion criteria as their modelling framework). We advocate that research into causal biomarker identification and optimal management and prevention strategies needs to be anchored in understanding of the underlying disease heterogeneity.

It is important that we, as a community of health care professionals, work toward transferring evidence-based information to better patient care. Therefore clinical practitioners

should be aware of the need to treat asthma and other heterogeneous diseases in a more personalized manner and be ready to incorporate the discovered stratified medicine strategies in a timely fashion.

REFERENCES

1. Belgrave D, Simpson A, Custovic A. Challenges in interpreting wheeze phenotypes: the clinical implications of statistical learning techniques. Am J Respir Crit Care Med 2014;189:121-3.
2. Belgrave DC, Custovic A, Simpson A. Characterizing wheeze phenotypes to identify endotypes of childhood asthma, and the implications for future management. Expert Rev Clin Immunol 2013;9:921-36.
3. Prosperi MC, Sahiner UM, Belgrave D, Sackesen C, Buchan IE, Simpson A, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. Am J Respir Crit Care Med 2013;188:1303-12.
4. Depner M, Fuchs O, Genuneit J, Karvonen AM, Hyvärinen A, Kaulek V, et al. Clinical and epidemiologic phenotypes of childhood asthma. Am J Respir Crit Care Med 2014;189:129-38.
5. Belgrave DC, Granell R, Simpson A, Guiver J, Bishop C, Buchan I, et al. Developmental profiles of eczema, wheeze, and rhinitis: two population-based birth cohort studies. PLoS Med 2014;11:e1001748.
6. Henderson J, Granell R, Heron J, Sherriff A, Simpson A, Woodcock A, et al. Associations of wheezing phenotypes in the first 6 years of life with atopy, lung function and airway responsiveness in mid-childhood. Thorax 2008;63:974-80.
7. Kurukulaaratchy R, Fenn M, Waterhouse L, Matthews S, Holgate S, Arshad S. Characterization of wheezing phenotypes in the first 10 years of life. Clin Exp Allergy 2003;33:573-8.
8. Martinez FD, Wright AL, Taussig LM, Holberg CJ, Halonen M, Morgan WJ. Asthma and wheezing in the first six years of life. N Engl J Med 1995;332:133-8.
9. Custovic A, Sonntag HJ, Buchan IE, Belgrave D, Simpson A, Prosperi MC. Evolution pathways of IgE responses to grass and mite allergens throughout childhood. J Allergy Clin Immunol 2015;136:1645-52.e8.
10. Lazic N, Roberts G, Custovic A, Belgrave D, Bishop CM, Winn J, et al. Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. Allergy 2013;68:764-70.
11. Simpson A, Lazic N, Belgrave DC, Johnson P, Bishop C, Mills C, et al. Patterns of IgE responses to multiple allergen components and clinical symptoms at age 11 years. J Allergy Clin Immunol 2015;136:1224-31.
12. Simpson A, Tan VY, Winn J, Svensén M, Bishop CM, Heckerman DE, et al. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. Am J Respir Crit Care Med 2010;181:1200-6.
13. Anderson GP. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. Lancet 2008;372:1107-19.
14. Lotvall J, Akdis CA, Bacharier LB, Bjermer L, Casale TB, Custovic A, et al. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. J Allergy Clin Immunol 2011;127:355-60.
15. Custovic A, Ainsworth J, Arshad H, Bishop C, Buchan I, Cullinan P, et al. The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together. Thorax 2015;70:799-801.
16. Teach SJ, Gergen PJ, Szefler SJ, Mitchell HE, Calatroni A, Wildfire J, et al. Seasonal risk factors for asthma exacerbations among inner-city children. J Allergy Clin Immunol 2015;135:1465-73.e5.
17. Teach SJ, Gill MA, Togias A, Sorkness CA, Arbes SJ, Calatroni A, et al. Preseasonal treatment with either omalizumab or an inhaled corticosteroid boost to prevent fall asthma exacerbations. J Allergy Clin Immunol 2015;136:1476-85.
18. Pavord ID, Korn S, Howarth P, Bleecker ER, Buhl R, Keene ON, et al. Mepolizumab for severe eosinophilic asthma (DREAM): a multicentre, double-blind, placebo-controlled trial. Lancet 2012;380:651-9.
19. Corren J, Lemanske RF, Hanania NA, Korenblat PE, Parsey MV, Arron JR, et al. Lebrikizumab treatment in adults with asthma. N Engl J Med 2011;365:1088-98.
20. Laprise C, Bouzigon E. The genetics of asthma and allergic diseases: pieces of the puzzle are starting to come together. Curr Opin Allergy Clin Immunol 2013;13:461-2.
21. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.
22. Custovic A, Marinho S, Simpson A. Gene-environment interactions in the development of asthma and atopy. Expert Rev Respir Med 2012;6:301-8.
23. Semic-Jusufagic A, Belgrave D, Pickles A, Telcian AG, Bakhsoliani E, Sykes A, et al. Assessing the association of early life antibiotic prescription with asthma exacerbations, impaired antiviral immunity, and genetic variants in 17q21: a population-based birth cohort study. Lancet Respir Med 2014;2:621-30.
24. Curtin JA, Simpson A, Belgrave D, Semic-Jusufagic A, Custovic A, Martinez FD. Methylation of IL-2 promoter at birth alters the risk of asthma exacerbations during childhood. Clin Exp Allergy 2013;43:304-11.
25. Bønnelykke K, Sleiman P, Nielsen K, Kreiner-Møller E, Mercader JM, Belgrave D, et al. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. Nat Genet 2014;46:51-5.
26. Bochkov YA, Watters K, Ashraf S, Griggs TF, Devries MK, Jackson DJ, et al. Cadherin-related family member 3, a childhood asthma susceptibility gene product, mediates rhinovirus C binding and replication. Proc Natl Acad Sci U S A 2015;112:5485-90.
27. Çalışkan M, Bochkov YA, Kreiner-Møller E, Bønnelykke K, Stein MM, Du G, et al. Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. N Engl J Med 2013;368:1398-407.
28. Kerkhof M, Boezen HM, Granell R, Wijga AH, Brunekreef B, Smit HA, et al. Transient early wheeze and lung function in early childhood associated with chronic obstructive pulmonary disease genes. J Allergy Clin Immunol 2014;133:68-76.e4.
29. Wjst M, Sargurupremraj M, Arnold M. Genome-wide association studies in asthma: what they really told us about pathogenesis. Curr Opin Allergy Clin Immunol 2013;13:112-8.
30. Custovic A, Simpson A. Environmental allergen exposure, sensitisation and asthma: from whole populations to individuals at risk. Thorax 2004;59:825-7.
31. Woodcock A, Custovic A. Role of the indoor environment in determining the severity of asthma. Thorax 1998;53(suppl 2):S47-51.
32. Collins SA, Pike KC, Inskip HM, Godfrey KM, Roberts G, Holloway JW, et al. Validation of novel wheeze phenotypes using longitudinal airway function and atopic sensitization data in the first 6 years of life: evidence from the Southampton Women's survey. Pediatr Pulmonol 2013;48:683-92.
33. Grad R, Morgan WJ. Long-term outcomes of early-onset wheeze and asthma. J Allergy Clin Immunol 2012;130:299-307.
34. Belgrave DC, Simpson A, Semic-Jusufagic A, Murray CS, Buchan I, Pickles A, et al. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing. J Allergy Clin Immunol 2013;132:575-83.e12.
35. Spycher BD, Kuehni CE. Asthma phenotypes in childhood: conceptual thoughts on stability and transition. Eur Respir J 2016;47:362-5.
36. Granell R, Sterne JA, Henderson J. Associations of different phenotypes of wheezing illness in early childhood with environmental variables implicated in the aetiology of asthma. PLoS One 2012;7:e48359.
37. Lodge CJ, Zaloumis S, Lowe AJ, Gurrin LC, Matheson MC, Axelrad C, et al. Early-life risk factors for childhood wheeze phenotypes in a high-risk birth cohort. J Pediatr 2014;164:289-94.e2.
38. Garden FL, Simpson JM, Mellis CM, Marks GB. Change in the manifestations of asthma and asthma-related traits in childhood: a latent transition analysis. Eur Respir J 2016;47:499-509.
39. Custovic A, Lazic N, Simpson A. Pediatric asthma and development of atopy. Curr Opin Allergy Clin Immunol 2013;13:173-80.
40. Simpson A, Martinez FD. The role of lipopolysaccharide in the development of atopy in humans. Clin Exp Allergy 2010;40:209-23.
41. Simpson A, John SL, Jury F, Niven R, Woodcock A, Ollier WE, et al. Endotoxin exposure, CD14, and allergic disease: an interaction between genes and the environment. Am J Respir Crit Care Med 2006;174:386-92.
42. Custovic A, Rothers J, Stern D, Simpson A, Woodcock A, Wright AL, et al. Effect of day care attendance on sensitization and atopic wheezing differs by Toll-like receptor 2 genotype in 2 population-based birth cohort studies. J Allergy Clin Immunol 2011;127:390-7, e1-9.
43. Sordillo JE, Kelly R, Bunyavanich S, McGeachie M, Qiu W, Croteau-Chonka DC, et al. Genome-wide expression profiles identify potential targets for gene-environment interactions in asthma severity. J Allergy Clin Immunol 2015;136:885-92.e2.
44. Berry A, Busse WW. Biomarkers in asthmatic patients: has their time come to direct treatment? J Allergy Clin Immunol 2016;137:1317-24.
45. Saria S, Goldenberg A. Subtyping: What it is and its role in precision medicine. IEEE Int Syst 2015;30:70-5.
46. Howard R, Rattray M, Prosperi M, Custovic A. Distinguishing asthma phenotypes using machine learning approaches. Curr Allergy Asthma Rep 2015;15:38.
47. Deliu M, Sperrin M, Belgrave D, Custovic A. Identification of asthma subtypes using clustering methodologies. Pulm Ther 2016;2:19-41.
48. Manolio TA. Bringing genome-wide association findings into clinical use. Nat Rev Genet 2013;14:549-58.
49. Agustí A, Antó JM, Auffray C, Barbé F, Barreiro E, Dorca J, et al. Personalized respiratory medicine: exploring the horizon, addressing the issues. Summary of a BRN-AJRCCM workshop held in Barcelona on June 12, 2014. Am J Respir Crit Care Med 2015;191:391-401.

50. Bach F, Jenatton R, Mairal J, Obozinski G. Optimization with sparsity-inducing penalties. Foundations and Trends in Machine Learning 2012;4:1-106.
51. Gelman A, Shalizi CR. Philosophy and the practice of Bayesian statistics. Br J Math Stat Psychol 2013;66:8-38.
52. Smola AJ, Schölkopf B. A tutorial on support vector regression. Stat Comput 2004;14:199-222.
53. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning: Springer series in statistics. Berlin: Springer; 2001.
54. Bishop CM. Model-based machine learning. Phil Trans A Math Phys Eng Sci 2013;371:20120222.
55. Murphy KP. Machine learning: a probabilistic perspective. Boston: MIT press; 2012.
56. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis. Boca Raton (FL): Chapman & Hall/CRC; 2014.
57. Bishop CM. Neural networks for pattern recognition. Oxford (UK): Oxford university press; 1995.
58. Williams CK, Barber D. Bayesian classification with Gaussian processes. IEEE Trans Pattern Anal Machine Int 1998;20:1342-51.
59. Gibbs MN, MacKay DJ. Variational Gaussian process classifiers. IEEE Trans Neural Netw 2000;11:1458-64.
60. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 2000;10:325-37.
61. Cosmides L, Tooby J. Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. Cognition 1996;58:1-73.
62. Berger J. The case for objective Bayesian analysis. Bayesian Anal 2006;1:385-402.
63. Gilboa I, Marinacci M. Ambiguity and the Bayesian paradigm. Readings in formal epistemology. Berlin: Springer; 2016:385-439.
64. Jordan MI. Are you a Bayesian or a frequentist? [Summer School Lecture]. Cambridge (United Kingdom); 2009. Video lecture available at http://videolectures.net/mlss09uk_jordan_bfway/.
65. Dunson DB. Commentary: practical advantages of Bayesian analysis of epidemiologic data. Am J Epidemiol 2001;153:1222-6.
66. Bentler PM, Stein JA. Structural equation models in medical research. Stat Methods Med Res 1992;1:159-81.
67. Muthén BO. Beyond SEM: general latent variable modeling. Behaviormetrika 2002;29:81-117.
68. Skrondal A, Rabe-Hesketh S. Generalized latent variable modeling: multilevel, longitudinal, and structural equation models. Boca Raton (FL): CRC Press; 2004.
69. Matthysse S, Holzman PS, Lange K. The genetic transmission of schizophrenia: application of Mendelian latent structure analysis to eye tracking dysfunctions in schizophrenia and affective disorder. J Psych Res 1986;20:57-67.
70. Bishop CM. Pattern recognition. Machine Learn 2006;128.
71. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. Machine Learn 2001;42:177-96.
72. Lawrence ND. Gaussian process latent variable models for visualisation of high dimensional data. Adv Neural Inform Processing Syst 2004;16:329-36.
73. Bernardo J, Bayarri M, Berger J, Dawid A, Heckerman D, Smith A, et al. Bayesian factor regression models in the "large p, small n" paradigm. Bayesian Stat 2003;7:733-42.
74. van Staa TP, Goldacre B, Buchan I, Smeeth L. Big health data: the need to earn public trust. BMJ 2016;354:i3636.
75. McGeachie MJ, Yates KP, Zhou X, Guo F, Sternberg AL, Van Natta ML, et al. Patterns of growth and decline in lung function in persistent childhood asthma. N Engl J Med 2016;374:1842-52.
76. Belgrave DC, Buchan I, Bishop C, Lowe L, Simpson A, Custovic A. Trajectories of lung function during childhood. Am J Respir Crit Care Med 2014;189:1101-9.
77. Han M, Ortega VE, Dransfield MT, Li H, Barr R, Couper DJ, et al. Cluster analysis of chronic obstructive pulmonary disease (COPD) related phenotypes in the SubPopulations And InteRmediate Outcome Measures In COPD Study (SPIROMICS) [abstract]. Am Thorac Soc 2016;193:A3509-A.
78. Bechhofer S, Buchan I, De Roure D, Missier P, Ainsworth J, Bhagat J, et al. Why linked data is not enough for scientists. Future Generation Comp Syst 2013;29:599-611.
79. Schneeweiss S. Learning from big health care data. N Engl J Med 2014;370:2161-3.
80. Velasquez G, Kain J, Villamizar M, Yong Z, Dhar J, Carvajal M, et al. ESP "smart flow" integrates quality and control data for diagnostics and optimization in real time. SPE Middle East Intelligent Energy Conference and Exhibition: Society of Petroleum Engineers; 2013.
81. Vaiciulis A, Peranich L, Mayer U, Zoldi SM, De Zilwa S. Automated entity identification for efficient profiling in an event probability prediction system. Google Patents. 2014. U.S. Patent No. 8,645,301.
82. Katz LB, Stewart LS, Levy BL. Benefits to health care professionals and patients with diabetes of a novel blood glucose meter that provides pattern recognition and real-time automatic messaging compared to conventional paper logbooks. Int Diabetes Nurs 2015;12:27-33.
83. Gardy J, Loman NJ, Rambaut A. Real-time digital pathogen surveillance—the time is now. Genome Biol 2015;16:155.
84. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol 2013;9:e1003256.
85. Velikova M, van Scheltinga JT, Lucas PJ, Spaanderman M. Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare. Int J Approximate Reason 2014;55:59-73.
86. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. Annu Rev Public Health 2013;34:61-75.
87. Williams SM, Moore JH. Big data analysis on autopilot? Biodata Mining 2013;6:22.