

Accepted Manuscript

Disaggregating Asthma: Big Investigation vs. Big Data

Danielle Belgrave, PhD, John Henderson, MD, Angela Simpson, MD PhD, Iain Buchan, MD PhD, Christopher Bishop, PhD, Adnan Custovic, MD PhD FAAAAI



PII: S0091-6749(16)31345-8

DOI: [10.1016/j.jaci.2016.11.003](https://doi.org/10.1016/j.jaci.2016.11.003)

Reference: YMAI 12457

To appear in: *Journal of Allergy and Clinical Immunology*

Received Date: 30 September 2016

Revised Date: 7 November 2016

Accepted Date: 9 November 2016

Please cite this article as: Belgrave D, Henderson J, Simpson A, Buchan I, Bishop C, Custovic A, Disaggregating Asthma: Big Investigation vs. Big Data, *Journal of Allergy and Clinical Immunology* (2016), doi: 10.1016/j.jaci.2016.11.003.

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 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.

Disaggregating Asthma: Big Investigation vs. Big Data

Danielle Belgrave PhD¹, John Henderson MD², Angela Simpson MD PhD³, Iain Buchan MD PhD⁴,
Christopher Bishop PhD⁵ and Adnan Custovic MD PhD FAAAAI¹

¹Department of Paediatrics, Imperial College, London, UK

²School of Social and Community Medicine, Faculty of Health Sciences, University of Bristol, UK

³Division of Infection, Immunity and Respiratory Medicine, Faculty of Biology, Medicine and Health,
University of Manchester, UK

⁴Health Informatics, Faculty of Biology, Medicine and Health, University of Manchester, UK

⁵Microsoft Research Ltd, 21 Station Road, Cambridge, CB1 2FB, UK

Corresponding author: Adnan Custovic MD PhD FAAAAI, Clinical Professor of Paediatric Allergy, Imperial
College London, Department of Paediatrics, St Mary's Campus Medical School, London W2 1PG, UK

Tel: +44 20 7594 3274, email: a.custovic@imperial.ac.uk

Declaration of funding: Dr Danielle Belgrave is supported by the Medical Research Council Career
Development Fellowship MR/M015181/1: STELAR consortium is funded by the Medical Research
Council grant MR/K002449/1

Unstructured Abstract:

We are facing a major challenge in bridging the gap between identifying subtypes of asthma, to
understanding 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 healthcare; 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 healthcare data and computational tools for data analysis is that the process of data mining may 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-driven 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' healthcare 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.

Disaggregating asthma for precision science and medicine

A major obstacle to realising precision (stratified or personalised) medicine in asthma 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 diseases¹⁻⁴. It is now well-established that both asthma^{3, 5-8} and allergic sensitization⁹⁻¹² are umbrella terms (or syndromes) incorporating a variety of underlying endotypes sharing common symptoms and phenotypic characteristics^{13, 14}. Whilst, by definition, each endotype has unique pathophysiology, and hence genetic and environmental associations^{13, 14}, it is likely that some mechanisms overlap one or more endotypes¹⁵. This underlying heterogeneity is also reflected in responses to treatments. For example, a therapeutic agent may be specific for a pathway which is primarily responsible for the patient’s asthma subtype, and therapeutic response may be predicted reasonably well using relevant biomarkers^{16, 17} (e.g. number of eosinophils in peripheral blood or sputum for mepolizumab¹⁸, or periostin for lebrikizumab¹⁹). Alternatively, a therapeutic agent may be relatively non-specific and target broad mechanisms shared between different asthma endotypes, in which case patients across different endotypes may display a spectrum of responses - which is likely the case with inhaled corticosteroids.

Across different disease areas, blizzards of genetic studies have initially raised expectations over ‘significant hits’ that later delivered neither meaningful clinical diagnostic tools nor useful insights into disease pathogenesis²⁰. Genetic studies have thus far explained little of the heritability of complex diseases²¹. 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 interactions²², gene-environment correlations²³, and epigenetic mechanisms²⁴, the use of aggregated definitions of disease may also contribute to inconsistent findings between studies investigating genetic components of asthma. However, by using more specific phenotyping, a recent genome-wide association study (GWAS) identified an association of a specific asthma subtype, characterized by early-life onset and recurrent, severe exacerbations in pre-school age, with a functional variant in a novel susceptibility gene *CDHR3* (rs6967330, C529Y)²⁵. This genetic variant was associated with a greater risk of asthma hospitalizations in two birth 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 SNP 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 to subtypes, much

stronger associations were observed for some of the genetic variants previously identified in GWAS (such as those in 17q21 locus)²⁵. The value of focussing on specific subgroups has been demonstrated in a study which showed that variants at 17q21 were associated with asthma, but only in children who had rhinovirus-induced wheezing illness²⁷. Similarly, the risk of transient early wheeze, but not of persistent wheezing, increases with the number of COPD alleles²⁸. Most of the genetic studies which used more precise phenotypes showed higher relative risk estimates than the modest effect sizes of genetic hits that were identified using a simple binary trait definition of asthma, highlighting the need for a more refined subtyping of asthma in order to accurately identify genetic variants of clinical importance²⁹.

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 asthma. We and others have shown that different phenotypes of childhood wheezing have different environmental associates^{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 for multiple late sensitization)³⁹. It likely that the effect of most environmental factors varies across individuals with different genetic predispositions, but the precise nature of most gene-environment interactions remains unclear²². One of the most replicated findings of gene-environment interactions in the development of allergic sensitization is between *CD14* variants and environmental endotoxin exposure⁴⁰. Several studies have reported that high endotoxin exposure may protect against sensitization, but only among individuals 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 non-atopic (but not atopic) wheezing⁴¹. Other examples that the nature of gene-environment interactions may differ between different wheeze phenotypes include the finding that daycare attendance may have opposite effects on *atopic* wheezing among individuals with different genetic variants in the *TLR2* gene (being protective in some, but increasing the risk in others)⁴², with no such effect being observed for *non-atopic* wheezing⁴². This suggests that replication of gene-environment interactions may be improved through more precise definition of the outcome of interest⁴³. 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 intervention³⁰, and that interventions which may be effective in one subtype of wheezing may not necessarily work for other subtypes.

One area which has been relatively more successful is the identification of biomarkers¹⁶ for more targeted treatment strategies¹⁷. In a recent review, Berry and Busse identified four main biomarkers which may help to optimise treatment strategies for different phenotypes of asthma⁴⁴. These biomarkers are generally limited to T2 mechanisms: eosinophils, exhaled nitric oxide, periostin and IgE. However, assessment of biomarker has not as yet become an integral part of clinical practice, and neither 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 which needs to be exploited⁴⁴, 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 which 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 pathophysiological mechanisms, which is a prerequisite for precision medicine¹⁵. Whilst 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 which may allow us to predict causality of distinct endotypes with greater accuracy⁴⁵.

One approach utilized 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 wheeze⁴⁶. However, different studies reported different number of childhood wheeze phenotypes (e.g. ranging between two and six)^{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 may merely reflect inconsistencies in the nature and timing of questions used (e.g. 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 impact 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 healthcare data being rapidly generated, but also to the complexity of this data, evident in the amplified scale of biological, genetic, environmental and phenotypic data. The scale of this data often makes handling, management and analysis challenging with the use of standard statistical methods. The evolution of powerful computational tools to analyse such high-dimensional, large datasets 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 which is inherent to its volume. Further, the vast increase in the quantity of data which is generated has made it impossible at times to know what we are looking for, 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) *via* 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 in the case 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 (such as 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 healthcare; 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 vs. 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 in order to make predictions and identify patterns. It is used commonly by computer scientists for problem-solving in a variety of fields, and is increasingly used 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 mathematical and statistical foundations have been in existence since the beginning of the 20th century⁵⁰⁻⁵³. Machine learning as a new discipline is rather a result of the exponential growth in computational power which has enabled implementation of the mathematical groundwork that was initiated decades earlier^{54, 55}.

One of the (somewhat artificial) distinctions between machine learning and conventional statistical approaches is that whereas machine learning focuses on prediction models and trying 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 disease and to understand uncertainty in the modelling assumptions. Model assumptions refer to our framework for representing the data or research questions related to that data. As these are assumptions, we are uncertain about these assumptions and need some way of testing whether or not 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 vs. Frequentist approach to understanding disease aetiology

We now introduce the reader to two different approaches to hypothesis-testing and hypothesis-generating: the Bayesian and Frequentist approaches. The aim of this discussion is to provide a conceptual framework which is currently commonly used in statistical and machine learning and can be applied to both big and small datasets in healthcare research. An understanding of these two paradigms is formative for a team approach to understanding disease aetiology in healthcare.

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, emphasises the frequency of the data. The Bayesian approach on the other hand 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 explicitly incorporated with the data at hand in the analytical model to formulate a posterior distribution which takes account of both the observed data and prior knowledge⁵⁶. 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 aetiology of asthma and allergic disease, Bayesian models provide a flexible, unified framework for understanding the probability of disease manifestation and co-manifestation, incorporating evidence from the literature or hypotheses from medical experts⁵ through the explicit quantification of this evidence. However, although Bayesian methods provide an intuitive, unified platform for carrying out statistical research, the results are often computationally intractable, and resolving these is active area of research⁵⁷⁻⁶⁰. The exponential increase in computational power and the increasing availability of tools which can handle large-scale data has facilitated the use of Bayesian methods, and it is important that we capitalise on these relatively complex tools in order to improve human health.

The use of Bayesian statistics is relatively uncommon in 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 problems⁶¹ by assigning explicit quantities to scientific assumptions. This may 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 these priors and quantifying expert knowledge^{62, 63}, mainly due to the training and time necessary to develop “informative” priors. 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 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 priors based on expert knowledge may be overstated. The Frequentist approach can be used to compare different model assumptions based on expert knowledge, which may 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 parsimony⁶⁴.

Away from methodological polemics towards 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(s) for the problem/question. Different methodologies are not mutually exclusive; indeed, an ensemble of methods may 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 where Bayesian and Frequentist paradigms compete (or complement) is latent variable modelling⁶⁵. This section highlights the importance of latent variable modelling as a generalised framework for hypothesis-generating and dimensionality reduction. Dimensionality reduction is an important tool for analysis of “big data” where we have multiple clinical, molecular, genetic, environmental and phenotypic elements (i.e. we have high-dimensions). As the name suggests, in dimensionality reduction, the aim is to reduce the dimension of the dataset 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 dataset, but also to identify subgroups of patients based on patterns within these variables. Latent variable models are increasingly cited in the medical literature⁶⁶⁻⁶⁸ for classifying different phenotypes and sub-phenotypes 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 may 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 datasets, where associations between a set of observed variables is supplemented with additional latent variables. Latent variable modelling therefore allows us to move from hypothesis testing to hypothesis generating⁶⁹. A further advantage of using latent variable modelling is that it is easier to represent high-dimensional parameters⁷⁰⁻⁷³ 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 are representative of a larger dataset. 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 generalised 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 pathophysiological mechanisms (endotypes) which drive these distinct subtypes¹.

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 data⁷⁴. Birth cohort studies are an interesting parallel, as cradle-to-grave healthcare records can be thought of in this way. However, unlike routine healthcare 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) consortium¹⁵. STELAR combines data from five UK birth cohorts aimed at understanding the development of asthma and allergic diseases through the lifecourse. The cohorts include The Avon Longitudinal Study of Parents and Children (ALSPAC), Ashford cohort, The Isle of Wight cohort, Manchester Asthma and Allergy Study (MAAS) and the Aberdeen Study of Eczema and Asthma to Observe the Effects of Nutrition (SEATON). STELAR has data on >14000 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 healthcare records because they have more selected/biased observations sampled later in the natural history of disease. Figure 1 summarises the challenges in understanding asthma and allergic disease which will drive future research in STELAR consortium.

An important area where 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 which 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 aetiology⁷⁸. An important, largely-unanswered question is how best to translate findings between case and birth cohorts, i.e. between clinical and general populations, in order to inform better prevention and early intervention strategies⁷⁹. The case has been argued for automated methods to update disease models in real-time⁸⁰⁻⁸⁴. 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 characterised 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.

Conclusions: 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 two contexts), to understanding 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 aetiologies⁸⁶. From this perspective, significant investment needs to be made in advancing statistical and computational tools to solve healthcare 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 rigour and testing hypotheses which arise from this process. One of the dangers of ready accessibility of healthcare data and computational tools for analysing these is that the process of data mining may become uncoupled from the scientific process of clinical interpretation, understanding the provenance of the data, and external validation⁸⁷. 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. Figure 2 illustrates a data cycle encompassing the problem – an integrative approach to data science whereby basic scientists, clinicians, bio-statisticians 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 re-assess systems and criteria for promotion that still, in many cases, do not give sufficient credit for perceived "non-leadership" roles.

We need to ensure that we harness 'bigger' healthcare data in ways that produce meaningful clinical interpretation, and to translate the findings into better diagnoses, biomarkers and properly personalised 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 regimen¹⁷. 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 diseases. The direction needs to move away from looking at average effects (which is a strategy commonly employed in randomised clinical trials which 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 healthcare professionals work towards transferring evidence-based information to better patient care. Clinical practice should therefore be aware of the need to treat asthma and other heterogeneous diseases in a more personalised manner, and be ready to incorporate the discovered stratified medicine strategies in a timely fashion.

LEGEND FOR FIGURES

Figure 1: Roadmap of Challenges to understanding Asthma and Allergic Diseases within the STELAR Consortium.

Figure 2: Data Cycle: An integrative approach to understanding disease endotypes

REFERENCES

1. Belgrave D, Simpson A, Custovic A. Challenges in interpreting wheeze phenotypes: the clinical implications of statistical learning techniques. *American journal of respiratory and critical care medicine* 2014; **189**(2): 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 review of clinical immunology* 2013; **9**(10): 921-36.
3. Prosperi MC, Sahiner UM, Belgrave D, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. *American journal of respiratory and critical care medicine* 2013; **188**(11): 1303-12.
4. Depner M, Fuchs O, Genuneit J, et al. Clinical and epidemiologic phenotypes of childhood asthma. *American journal of respiratory and critical care medicine* 2014; **189**(2): 129-38.
5. Belgrave DC, Granell R, Simpson A, et al. Developmental profiles of eczema, wheeze, and rhinitis: two population-based birth cohort studies. *PLoS Med* 2014; **11**(10): e1001748.
6. Henderson J, Granell R, Heron J, 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**(11): 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. *Clinical & Experimental Allergy* 2003; **33**(5): 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. *New England Journal of Medicine* 1995; **332**(3): 133-8.
9. Custovic A, Sonntag H-J, Buchan IE, Belgrave D, Simpson A, Prosperi MC. Evolution pathways of IgE responses to grass and mite allergens throughout childhood. *Journal of Allergy and Clinical Immunology* 2015; **136**(6): 1645-52. e8.
10. Lazic N, Roberts G, Custovic A, et al. Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. *Allergy* 2013; **68**(6): 764-70.
11. Simpson A, Lazic N, Belgrave DC, et al. Patterns of IgE responses to multiple allergen components and clinical symptoms at age 11 years. *Journal of Allergy and Clinical Immunology* 2015; **136**(5): 1224-31.
12. Simpson A, Tan VY, Winn J, et al. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. *American journal of respiratory and critical care medicine* 2010; **181**(11): 1200-6.
13. Anderson GP. Endotyping asthma: new insights into key pathogenic mechanisms in a complex, heterogeneous disease. *Lancet* 2008; **372**(9643): 1107-19.
14. Lotvall J, Akdis CA, Bacharier LB, et al. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. *The Journal of allergy and clinical immunology* 2011; **127**(2): 355-60.
15. Custovic A, Ainsworth J, Arshad H, 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; thoraxjnl-2015-206781.
16. Teach SJ, Gergen PJ, Szeffler SJ, et al. Seasonal risk factors for asthma exacerbations among inner-city children. *Journal of Allergy and Clinical Immunology* 2015; **135**(6): 1465-73. e5.
17. Teach SJ, Gill MA, Togias A, et al. Preseasonal treatment with either omalizumab or an inhaled corticosteroid boost to prevent fall asthma exacerbations. *Journal of Allergy and Clinical Immunology* 2015; **136**(6): 1476-85.

18. Pavord ID, Korn S, Howarth P, et al. Mepolizumab for severe eosinophilic asthma (DREAM): a multicentre, double-blind, placebo-controlled trial. *Lancet* 2012; **380**(9842): 651-9.
19. Corren J, Lemanske RF, Hanania NA, et al. Lebrikizumab treatment in adults with asthma. *The New England journal of medicine* 2011; **365**(12): 1088-98.
20. Laprise C, Bouzigon E. The genetics of asthma and allergic diseases: pieces of the puzzle are starting to come together. *Current opinion in allergy and clinical immunology* 2013; **13**(5): 461-2.
21. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009; **461**(7265): 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**(3): 301-8.
23. Semic-Jusufagic A, Belgrave D, Pickles 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. *The Lancet Respiratory medicine* 2014; **2**(8): 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**(3): 304-11.
25. Bønnelykke K, Sleiman P, Nielsen K, et al. A genome-wide association study identifies CDHR3 as a susceptibility locus for early childhood asthma with severe exacerbations. *Nature genetics* 2014; **46**(1): 51-5.
26. Bochkov YA, Watters K, Ashraf S, et al. Cadherin-related family member 3, a childhood asthma susceptibility gene product, mediates rhinovirus C binding and replication. *Proceedings of the National Academy of Sciences of the United States of America* 2015; **112**(17): 5485-90.
27. Çalışkan M, Bochkov YA, Kreiner-Møller E, et al. Rhinovirus wheezing illness and genetic risk of childhood-onset asthma. *New England Journal of Medicine* 2013; **368**(15): 1398-407.
28. Kerkhof M, Boezen HM, Granell R, et al. Transient early wheeze and lung function in early childhood associated with chronic obstructive pulmonary disease genes. *Journal of Allergy and Clinical Immunology* 2014; **133**(1): 68-76. e4.
29. Wjst M, Sargurupremraj M, Arnold M. Genome-wide association studies in asthma: what they really told us about pathogenesis. *Current opinion in allergy and clinical immunology* 2013; **13**(1): 112-8.
30. Custovic A, Simpson A. Environmental allergen exposure, sensitisation and asthma: from whole populations to individuals at risk. *Thorax* 2004; **59**(10): 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, 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. *Pediatric pulmonology* 2013; **48**(7): 683-92.
33. Grad R, Morgan WJ. Long-term outcomes of early-onset wheeze and asthma. *Journal of Allergy and Clinical Immunology* 2012; **130**(2): 299-307.
34. Belgrave DC, Simpson A, Semic-Jusufagic A, et al. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing. *Journal of Allergy and Clinical Immunology* 2013; **132**(3): 575-83. e12.
35. Spycher BD, Kuehni CE. Asthma phenotypes in childhood: conceptual thoughts on stability and transition. *European Respiratory Journal* 2016; **47**(2): 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**(10): e48359.
37. Lodge CJ, Zoloumis S, Lowe AJ, et al. Early-life risk factors for childhood wheeze phenotypes in a high-risk birth cohort. *The Journal of pediatrics* 2014; **164**(2): 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. *European Respiratory Journal* 2016; **47**(2): 499-509.
39. Custovic A, Lazic N, Simpson A. Pediatric asthma and development of atopy. *Curr Opin Allergy Clin Immunol* 2013; **13**(2): 173-80.
40. Simpson A, Martinez FD. The role of lipopolysaccharide in the development of atopy in humans. *Clin Exp Allergy* 2010; **40**(2): 209-23.
41. Simpson A, John SL, Jury F, et al. Endotoxin exposure, CD14, and allergic disease: an interaction between genes and the environment. *Am J Respir Crit Care Med* 2006; **174**(4): 386-92.
42. Custovic A, Rothers J, Stern D, 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. *The Journal of allergy and clinical immunology* 2011; **127**(2): 390-7 e1-9.
43. Sordillo JE, Kelly R, Bunyavanich S, et al. Genome-wide expression profiles identify potential targets for gene-environment interactions in asthma severity. *Journal of Allergy and Clinical Immunology* 2015; **136**(4): 885-92. e2.
44. Berry A, Busse WW. Biomarkers in asthmatic patients: Has their time come to direct treatment? *Journal of Allergy and Clinical Immunology* 2016; **137**(5): 1317-24.
45. Saria S, Goldenberg A. Subtyping: What it is and its role in precision medicine. *IEEE Intelligent Systems* 2015; **30**(4): 70-5.
46. Howard R, Rattray M, Prosperi M, Custovic A. Distinguishing Asthma Phenotypes Using Machine Learning Approaches. *Curr Allergy Asthma Rep* 2015; **15**(7): 38.
47. Deliu M, Sperrin M, Belgrave D, Custovic A. Identification of Asthma Subtypes Using Clustering Methodologies. *Pulmonary Therapy* 2016; **2**(1): 19-41.
48. Manolio TA. Bringing genome-wide association findings into clinical use. *Nature Reviews Genetics* 2013; **14**(8): 549-58.
49. Agustí A, Antó JM, Auffray C, et al. Personalized respiratory medicine: exploring the horizon, addressing the issues. Summary of a BRN-AJRCCM workshop held in Barcelona on June 12, 2014. *American journal of respiratory and critical care medicine* 2015; **191**(4): 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): 1-106.
51. Gelman A, Shalizi CR. Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology* 2013; **66**(1): 8-38.
52. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Statistics and computing* 2004; **14**(3): 199-222.
53. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning: Springer series in statistics Springer, Berlin; 2001.
54. Bishop CM. Model-based machine learning. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 2013; **371**(1984): 20120222.
55. Murphy KP. Machine learning: a probabilistic perspective: MIT press; 2012.
56. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis: Chapman & Hall/CRC Boca Raton, FL, USA; 2014.
57. Bishop CM. Neural networks for pattern recognition: Oxford university press; 1995.
58. Williams CK, Barber D. Bayesian classification with Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1998; **20**(12): 1342-51.
59. Gibbs MN, MacKay DJ. Variational Gaussian process classifiers. *IEEE Transactions on Neural Networks* 2000; **11**(6): 1458-64.
60. Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS-a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing* 2000; **10**(4): 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): 1-73.
62. Berger J. The case for objective Bayesian analysis. *Bayesian analysis* 2006; **1**(3): 385-402.
63. Gilboa I, Marinacci M. Ambiguity and the Bayesian paradigm. *Readings in Formal Epistemology*: Springer; 2016: 385-439.
64. Jordan MI. Are You a Bayesian or a Frequentist? *Summer School Lecture, Cambridge* 2009.
65. Dunson DB. Commentary: practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology* 2001; **153**(12): 1222-6.
66. Bentler P, Stein J. Structural equation models in medical research. *Statistical Methods in Medical Research* 1992; **1**(2): 159-81.
67. Muthén BO. Beyond SEM: General latent variable modeling. *Behaviormetrika* 2002; **29**(1): 81-117.
68. Skrondal A, Rabe-Hesketh S. Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models: 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. *Journal of psychiatric research* 1986; **20**(1): 57-67.
70. Bishop CM. Pattern recognition. *Machine Learning* 2006; **128**.
71. Hofmann T. Unsupervised learning by probabilistic latent semantic analysis. *Machine learning* 2001; **42**(1-2): 177-96.
72. Lawrence ND. Gaussian process latent variable models for visualisation of high dimensional data. *Advances in neural information processing systems* 2004; **16**(3): 329-36.
73. Bernardo J, Bayarri M, Berger J, et al. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian statistics* 2003; **7**: 733-42.
74. van Staa T-P, 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, et al. Patterns of Growth and Decline in Lung Function in Persistent Childhood Asthma. *New England Journal of Medicine* 2016; **374**(19): 1842-52.
76. Belgrave DC, Buchan I, Bishop C, Lowe L, Simpson A, Custovic A. Trajectories of lung function during childhood. *American journal of respiratory and critical care medicine* 2014; **189**(9): 1101-9.
77. Han M, Ortega VE, Dransfield MT, et al. Cluster Analysis Of Chronic Obstructive Pulmonary Disease (COPD) Related Phenotypes In The SubPopulations And Intermediate Outcome Measures In COPD Study (SPIROMICS). B43 COPD: PHENOTYPES AND CLINICAL OUTCOMES: Am Thoracic Soc; 2016: A3509-A.
78. Bechhofer S, Buchan I, De Roure D, et al. Why linked data is not enough for scientists. *Future Generation Computer Systems* 2013; **29**(2): 599-611.
79. Schneeweiss S. Learning from big health care data. *New England Journal of Medicine* 2014; **370**(23): 2161-3.
80. Velasquez G, Kain J, Villamizar M, et al. ESP" Smart Flow" Integrates Quality and Control Data for Diagnostics and Optimization in Real Time. SPE Digital Energy Conference; 2013: 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.
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. *International Diabetes Nursing* 2015; **12**(1): 27-33.
83. Gardy J, Loman NJ, Rambaut A. Real-time digital pathogen surveillance—the time is now. *Genome Biology* 2015; **16**(1): 1.

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**(10): e1003256.
85. Velikova M, van Scheltinga JT, Lucas PJ, Spaanderman M. Exploiting causal functional relationships in Bayesian network modelling for personalised healthcare. *International Journal of Approximate Reasoning* 2014; **55**(1): 59-73.
86. Glass TA, Goodman SN, Hernán MA, Samet JM. Causal inference in public health. *Annual review of public health* 2013; **34**: 61.
87. Williams SM, Moore JH. Big Data analysis on autopilot? *BioData mining* 2013; **6**(1): 1.



