

Machine learning in asthma research: moving toward a more integrated approach

Sara Fontanella, Alex Cucco & Adnan Custovic

To cite this article: Sara Fontanella, Alex Cucco & Adnan Custovic (2021) Machine learning in asthma research: moving toward a more integrated approach, Expert Review of Respiratory Medicine, 15:5, 609-621, DOI: [10.1080/17476348.2021.1894133](https://doi.org/10.1080/17476348.2021.1894133)

To link to this article: <https://doi.org/10.1080/17476348.2021.1894133>



© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.



Published online: 16 Mar 2021.



Submit your article to this journal [↗](#)



Article views: 2490



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 9 View citing articles [↗](#)

REVIEW



Machine learning in asthma research: moving toward a more integrated approach

Sara Fontanella, Alex Cucco and Adnan Custovic

National Heart and Lung Institute, Imperial College London, UK

ABSTRACT

Introduction: Big data are reshaping the future of medicine. The growing availability and increasing complexity of data have favored the adoption of modern analytical and computational methodologies in every area of medicine. Over the past decades, asthma research has been characterized by a shift in the way studies are conducted and data are analyzed. Motivated by the assumptions that 'data will speak for themselves', hypothesis-driven approaches have been replaced by data-driven hypotheses-generating methods to explore hidden patterns and underlying mechanisms. However, even with all the advancement in technologies and the new important insight that we gained to understand and characterize asthma heterogeneity, very few research findings have been translated into clinically actionable solutions.

Areas covered: To investigate some of the fundamental analytical approaches adopted in the current literature and appraise their impact and usefulness in medicine, we conducted a bibliometric analysis of big data analytics in asthma research in the past 50 years.

Expert opinion: No single data source or methodology can uncover the complexity of human health and disease. To fully capitalize on the potential of 'big data', we will have to embrace the collaborative science and encourage the creation of integrated cross-disciplinary teams brought together around technological advances.

ARTICLE HISTORY

Received 31 October 2020
Accepted 19 February 2021

KEYWORDS

Asthma; big data; data analytics; machine learning; statistics

1. Introduction

Asthma is a common chronic disorder which has long been recognized as a heterogeneous disease [1]. It is characterized by the variable symptom expression, airway inflammation, and therapeutic responses, making the clinical diagnosis challenging and long-term prognosis uncertain [2,3]. In the era of big data, technological advances and an ever-increasing quantity of available data in healthcare offer a unique opportunity to improve our understanding of this complex disease [4]. Studies of asthma increasingly involve the collection of large and diverse datasets that include various sources of data (e.g. clinical, environmental, omics, imaging, electronic health records, etc.), and modern big data analytics offer a powerful tool to mine knowledge from them [5]. In this context, classical hypothesis-driven approaches have in recent years been replaced by complex data-driven analytical techniques that interrogate the data in an unbiased fashion to discover hidden patterns and associations in a quest to mine knowledge from big data [6]. Such hypothesis-generating techniques have been extensively applied to the discovery of asthma subtypes and their pathophysiological mechanism [7–18], its relation with allergic sensitization [19–23] and environmental exposures [24–29], the discovery of biomarkers [30–40] and asthma prediction [41–46]. However, many challenges remain unmet, and very few research findings have been translated into

clinical practice. For example, there is a lack of universally accepted diagnostic tools, especially in childhood asthma [47]. Moreover, the definition of asthma subtypes remains controversial [48]: while the existence of asthma endotypes is universally acknowledged, there is no consensus on how best to define them [3,49].

In this review, we aim to analytically investigate the methodological shift toward data-driven analyses in asthma research and critically appraise the impact of modern analytical techniques. In particular, we wish to highlight the advantages and pitfalls of their application in the current literature and provide our interpretation on the difficulty in deriving clinically meaningful insights about the disease development and its pathophysiological mechanism. To this end, we conducted a bibliometric analysis highlighting the relevant characteristics such as publication trends and topics that influence the research in asthma. On the premise that their principal goals differ significantly, we distinguish between studies that adopted classical Multivariate Statistics (MS) and Bayesian Statistics (BS) versus more recent Machine Learning (ML) and Artificial Intelligence (AI) methodologies. In Sections 2 and 3, we briefly introduce the reader to different methodologies. In Section 4 we describe the analytical approach adopted to perform the bibliometric analysis. The main results are discussed in Section 5, and our commentary and view on the current status of the field are presented in Section 6.

Article highlights

- Studies of asthma increasingly involve collection of large and diverse datasets that include various sources of data (clinical, environmental, omics, imaging, electronic health records, etc.), and modern statistics, machine learning and artificial intelligence methodologies offer a powerful tool to mine knowledge from them
- However, this is a challenging framework for discovering useful knowledge, and despite their undoubted potential, real-world deployments of machine learning algorithms and artificial intelligence in clinical practice are rare
- Modern computational methods can be valuable for identifying unexpected structure in data to generate hypotheses; however, these should not be accepted without a critical appraisal of the results and corroboration of the derived hypotheses
- There has been a growing concern around accuracy, reliability, security and privacy related to the use of big data analytics, and a mounting necessity to define a framework to describe how to undertake effective and ethical research in ML and AI in healthcare
- One of the dangers of ready accessibility of healthcare data, and models for their analysis, is that the process of data mining can become uncoupled from understanding the provenance of the data, external validation the scientific process of interpretation
- We wish to highlight the potential danger of ill-considered use of readily available computational software for data analysis: the modeling choices made by inexperienced researchers can be flawed and draw inconsistent conclusions
- To fully harness the potential of ML, AI and modern statistical approaches in shaping the future clinical decision-making system, technological and methodological advances will need to be coupled with clinical and data analytical expertise.
- To fully capitalize on the potential of 'big data', we will have to encourage the creation of cross-disciplinary teams which are increasingly and by necessity cross-institutional and cross the geographical boundaries.
- To facilitate this process, we need to develop and implement incentives for "team scientists), including clear pathways for a progression in their academic careers, as current pathways favor 'solo' scientists.

2. Harnessing the power of statistics and machine learning

According to the HACE theorem [50], 'big data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data'. These characteristics make large complex medical 'big data' a challenging framework for discovering useful knowledge [6]. Therefore, it is not surprising that, despite the extraordinary availability and accessibility, only a portion of the data is integrated, understood, and analyzed [51]. Modern statistics, ML and AI methodologies offer unique tools to harness high volumes of data, integrate them from various sources, and mine them to extract meaningful information.

AI and ML have become trendy terms in medicine and healthcare, and often they are used interchangeably. However, although they are closely related, there are important differences. AI is a computer science field that aims to produce computer systems and machines that can mimic human intelligence, thereby performing intellectual tasks such as decision-making and problem-solving. ML is a subfield of AI, where machines are trained to learn how to process and mine the data to identify features and relations to make predictions and identify patterns. Using mathematical, statistical and computational tools, ML enables machines to

automatically learn and improve from past data and experiences without being explicitly programmed. It is of note that whilst ML as a discipline is fairly new, the statistical and mathematical foundations have been in existence for more than a century; the recent exponential growth in computational power enabled implementation of this statistical/mathematical groundwork [6]. Most of the statistical concepts used in ML are drawn from the Multivariate statistics (MS) framework, a collection of statistical models that deal with the analysis of multivariate data, for which observations are measured on more than one variable and where there is an inherent interdependence between the variables [52]. Our perspective on the relationship between different methodologies/fields is depicted in (Figure 1).

One of the main differences between machine learning and traditional multivariate statistics techniques lies in their purpose. Machine learning focuses on prediction models and on making the prediction as accurate as possible. It uses algorithms to learn rules from available data and create models able to generalize to new unseen instances. For example, in medicine, extant data can be used to train models that can then be used to predict future disease states or development of symptoms in later life. Statistics and multivariate analysis, instead, primarily seek to infer the relationship between variables. They use present observations to build models that can enhance the understanding of the phenomena at hand. In medical research, their focus is on the investigation of causality and associations to explain the disease and understand uncertainty in the modeling assumptions [6].

In medicine, traditional regression-based models often involve a small number of variables to predict the value of an outcome or the probability of an event, and are often used to infer association rather than to establish a prediction rule [53]. Their popularity is motivated by their high degree of interpretability: results are easy to understand, interpret and contextualize in a clinical setting, and underlying biological mechanisms can be discovered. In contrast, the interpretation of results from ML can be challenging. The lack of interpretability is apparent in artificial neural networks (ANNs), which are highly parameterized, non-linear models with sets of processing units called neurons [54] that can approximate and learn the relationship between predictors and outcomes in a complex system. Despite their remarkable predictive power, the fitted models do not provide insights into the relative importance of the predictors or covariates, their underlying relationships and structures, in relation to the modeled outcomes [54]. On the other hand, ML has many appealing properties. While the application of classical statistical models relies on the definition of several modeling assumptions, such as additivity, linearity and distributional assumptions, which are seldom met in clinical practice, ML models are free from prior distributional assumptions. These computational techniques can easily incorporate and model complex interactions between a high number of variables, which are difficult to assert in a statistical model. Moreover, while most multivariate statistics techniques incur computational issues in the presence of high-dimensional data – e.g., the number of predictors is often larger than the number of observations – ML can easily handle this large-scale data. Another critical advantage of ML lies in its ability to accommodate and combine different types of data: structured

MS - Multivariate statistics

ML - Machine learning

AI - Artificial intelligence

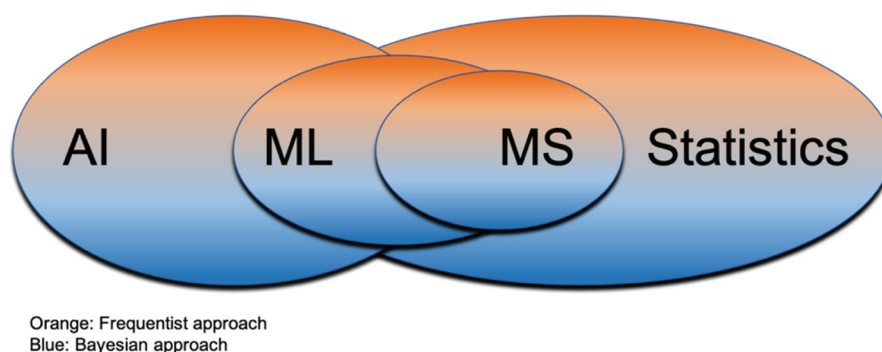


Figure 1. The relationship between statistical and computer science disciplines.
MS – Multivariate statistics ML – Machine learning AI – Artificial intelligence

and unstructured data can be integrated into investigations of the prediction of disease diagnosis, prognosis and response to treatment.

Recently, the adoption of ML and AI models has been accelerated by the increasing availability and accessibility of Electronic Health Records (EHRs). Given that a great proportion of data are stored in textual format, the field of Natural Language Processing (NLP) is of particular relevance, and the correct use of NLP can help in many practical clinical purposes, from diagnosis to management [55]. NLP refers to a specific area of ML in which the input is represented by a written or spoken information. While in a classical framework, the researcher can work with 'structured' data characterized by a precise organization of the information, in the NLP context, the information is 'unstructured', i.e. there is no common structure in which the information is stored and presented (e.g. different length of the audio or text, different use of language, etc.). In this context, pre-processing of data is essential: the documents, words, phrases or letters need to be translated into numbers/vectors. Technically speaking, it means to project each document, word, phrase, or letter into a vector-space. With a numerical (vector) representation of text, researchers can rely on structured data sets that can be used as any other sources of information. Consequently, supervised or unsupervised ML or MS models can be applied. For an in-depth analysis of clinical applications of text mining, we refer the reader to a comprehensive review on the application of NLP in asthma [55] and an open access book on clinical text mining [56].

3. A philosophical debate: Bayesian versus frequentist approach

Without the intention of being exhaustive, the purpose of this section is to introduce two opposing philosophies of statistical analysis, namely the Bayesian and the frequentist. Both are

valid, although they differ methodologically and perhaps philosophically [57].

In the frequentist approach, the scientist uses models that hold under specific assumptions and extract all the information from the data, without introducing any external information. The intention is to 'let the data speak for themselves', under the belief that the experimental design completely respects the assumptions needed to use a specific model. In this setting, the unknown population value, for example disease prevalence, is considered as a fixed, unvarying quantity, without a probability distribution [58]. The aim is to obtain point or confidence intervals estimates for such quantity from the available data or test hypotheses around these parameters.

In contrast, the Bayesian approach uses a probabilistic framework for obtaining such estimates. This approach is based on the Bayes' theorem, which underpins the subjectivist approach to statistics and logic of evidential support, and allows for the use of some knowledge or belief that we already have (commonly referred to as the prior) to help us calculate the probability of a related event. Bayesian methods are formulated on the premise that unknown parameters have probability distributions that can be constructed using our prior knowledge, before adding the information which comes from the data. In this framework, prior knowledge can be incorporated explicitly in the analytic model and combined with the data at hand to formulate a posterior distribution over the parameters of interests. This is one of the most characterizing aspects of Bayesian statistics and the feature that clearly distinguishes it from the frequentist perspective, where the prior knowledge cannot be explicitly included, but only drawn from the available data and translated into the assumptions we make about the statistical models for the data. However, the key question is not which of the two approaches is 'better', but rather, which

method is best suited for the scientific question being addressed [6].

4. Literature review

4.1. Search strategy

We systematically searched the literature up to 2020 using Scopus database to capture papers relevant to the application of MS, BS and ML and AI in asthma research. Scopus is an online bibliometric database developed by Elsevier that boasts a broad coverage of scientific resources and high accuracy [59]. We built three separate queries to investigate the impact of each discipline in asthma research. The term 'asthma' was included in all searches combined with different search terms for the three individual categories:

- Multivariate statistics (MS): 'principal component' OR 'discriminant analysis' OR 'correspondence analysis' OR 'canonical correlation' OR 'Markov models' OR 'factor analysis' OR 'structural equation' OR 'latent variable' OR 'multidimensional scaling' OR 'clustering' OR 'latent class' OR 'cluster analysis' OR 'latent profile' OR 'profile regression' OR 'mixture models'.
- Machine learning and artificial intelligence (ML&AI): 'artificial neural networks' OR 'deep learning' OR 'supervised learning' OR 'unsupervised learning' OR 'support vector machine' OR 'SVM' OR 'decision trees' OR 'classification trees' OR 'regression trees' OR 'random forest' OR 'machine learning' OR 'artificial Intelligence'.
- Bayesian statistics (BS): 'Bayesian' OR 'Bayes'.

We wish to point out that although we appreciate that several of the listed methodologies are nowadays included in both MS and ML&AI, we assigned them according to the field in which they were initially developed. However, to capture this overlap to some extent, we included in the ML query, broad terms such as 'unsupervised' and 'supervised learning'. The description of the individual methodologies is beyond the scope of this review, and we refer the readers to well-known texts on MS [52,60], ML&AI [60–63] and BS [64]. We omitted the term 'big data' in the search as we wished to highlight both a 'big-data' context and a 'small sample size' scenario. Specifically, while some of the methodologies of interest for this review require a big data set, other models are also appropriate in a small sample context.

4.2. Bibliometric analysis

To understand the state of the literature and to investigate the possible differences among the considered methodologies – MS, BS and ML&AI – we performed a bibliometric analysis on the set of articles returned from the search queries using the *bibliometrix* [65] package in R, which is a language and environment for statistical computing and graphics and provides readily available codes to implement a wide variety of statistical models. Bibliometrics is an analytical discipline that utilizes statistical, and ML approaches to analyze scientific production in a given research field. From bibliographic data

(e.g. authors, words, journals, countries, keywords), descriptive statistics can be used to evaluate author and country production, journal sources, and co-word analysis to identify significant themes in asthma research. On the premise that article keywords provide a suitable description of documents' content, the analysis of keywords co-occurrences can reveal possible links between the topics covered in the articles. This analysis can provide insights into the knowledge structure of a scientific field. Keyword co-occurrence network and strategic thematic map [66,67] were used to investigate patterns and themes in the research field of asthma. Networks provide a representation of the relationships occurring in a system. A graph is generated assuming keywords as a set of nodes and a set of links established using their co-occurrences. This approach provides a schematic representation of the research themes and their relations. More details can, then, be extrapolated in the thematic map, where themes are embedded in a two-dimensional representation defined by the centrality and density indices. The centrality index captures the degree of interaction among networks; in other words, it is 'a measure of the importance of a theme in the development of the entire research field analysed' [66]. Instead, the density measures the internal strength of the network, and it captures the level of development of a specific theme. The interpretability of the results relies on the definition of four quadrants in the two-dimensional embedding space, and the themes' relevance can be interpreted based on their position with respect to the quadrants.

4.3. Probabilistic topic modeling

To further explore the knowledge structure of the asthma research field, we applied Latent Dirichlet Allocation (LDA) on title, keywords, and abstract [68,69]. LDA is an unsupervised ML approach that allows uncovering potentially hidden topic information in document sets by revealing recurring clusters of co-occurring words. LDA is a probabilistic topic model that represents each document as a random mixture over latent topics, and each topic is characterized by a distribution over words [68]. The topic distributions in all documents share a common Dirichlet prior, and the word distributions of topics share a common Dirichlet prior as well. LDA aims at inferring the latent structure that is likely to have generated the observed corpus [70].

Given a collection of documents, D , where each document is considered as a sequence of N words $w_d = \{w_{d,i}\}_{i=1,2,\dots,N}$ and each word is an item from a vocabulary indexed by $\{1, \dots, V\}$, we assume that there are K latent topics, $\beta_{1:K}$, defined as a distribution over the vocabulary. Then, the conditional distribution, known as the *posterior*, can be defined as [69]:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D} | w_{1:K}) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D})}{p(w_{1:D})}$$

where $\theta_{1:D}$ are the per-corpus topic proportions, $z_{1:D}$ the per-corpus topic assignments and $w_{1:D}$ the set of observed words. Both the topic proportions and the topic distributions over the vocabulary follow a Dirichlet distribution. As the posterior

Table 1. Summary information on retrieved asthma studies, 1965–2020.

Description	
Years	1965–2020
Documents	2395
Authors	12,607
Authors of single-authored documents	95
Authors of multi-authored documents	12,512
Documents per Author	0.19
Authors per Document	5.26
Coauthors per Documents	7.83
Collaboration Index	5.46
Average citations per documents	30.38
Sources	775
Keywords Plus (ID)	11,450
Author's Keywords (DE)	4546

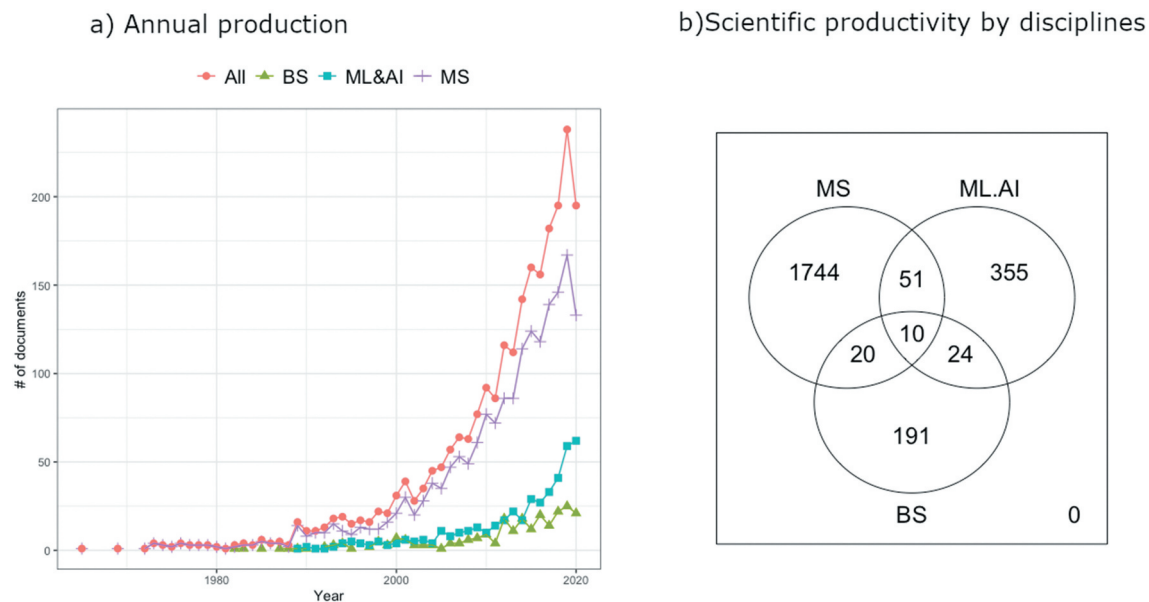
distribution is intractable for exact inference, inference algorithms need to be used to approximate it. Hence, topic models rely on sampling-based algorithms or variational algorithms [68]. The analysis was performed with ad-hoc routines [70] and the *Text Analytics Toolbox* of MATLAB [71].

5. Asthma research and themes

Three separated searches, using the search strategy described above, were launched on Scopus and the bibliographic data were extracted by applying the queries to the contents of title, abstract and keywords. The search was limited to the subject area of Medicine, and erratum or undefined documents were excluded from the analysis. Articles were then manually downloaded in the BibTeX file format. All available information was retrieved. The searches resulted in 1922, 501 and 255 for MS, ML&AI and BS, respectively. After combining these three lists and removing the articles without abstract, we found 2395 unique articles.

5.1. The publication trends

The retrieved studies were conducted between 1965 and 2020 and involved 12,607 authors, with 0.19 article per author and 5.26 authors per article, and a collaboration index of 5.46 (Table 1). Only 95 authors produced single-authored articles, while the remaining 12,512 authors were involved in multi-author publications, and the average number of citations per article was 30.38. (Figure 2a) shows the evolution of the asthma research field and the impact of modern analytical approaches. To evaluate the changes in scientific production over time, and we adopted the three research stages defined in Aparicio et al. [72]. Accordingly, we identified i) the *initial stage*, from 1965 to 1988, with 56 articles published, with a rate of three articles per year; ii) the *development stage*, which spanned 1989–1999, with 179 articles and an average of 16 articles per year, and iii) the *expansive stage*, from 2000 to 2020, with 2160 articles published (with publication rate of almost 103 articles per year). In this bibliographic collection, MS techniques made their first appearance in 1965 and had remarkable growth throughout the whole period. Among the three disciplines, MS has the highest contribution to the scientific production of this field. The first paper utilizing BS methodologies dates back to 1982, while more modern ML techniques have seen their first application in asthma research in 1989 and have since gained increasing popularity. 1744 articles (72.8%) adopted MS techniques only, 355 articles (14.8%) utilized only ML&AI techniques, while BS models were applied to 191 (8.0%), exclusively. In 95 articles, the three analytical methodologies were used in pairs, while 10 papers mentioned all three of them (Figure 2b). (Table 2) shows the 10 most influential authors in the field. Research output for the top 10 most productive countries is shown in (Table 3). Considering productivity in terms of the number of

**Figure 2.** Annual publication production.

Number of publications on asthma per year for the three scientific disciplines are shown separately and jointly. MS – Multivariate statistics; ML&AI – Machine learning and artificial intelligence; BS – Bayesian statistics

Table 2. The 10 most productive authors in the application of Multivariate statistics (MS), Machine learning and artificial intelligence (ML&AI) and Bayesian statistics (BS) to asthma research.

Author	Number of articles	Times Cited	Year
Custovic A	34	1992	2008
Sterk P	30	1765	2004
Weiss S	30	882	2000
Chung K	27	2345	2007
Bousquet J	26	1173	1996
Lee S	26	174	2003
Simpson A	26	1473	2008
Bisgaard H	22	536	2009
Kim H	21	234	2012
Kim S	19	323	2009

articles published, the USA ranked first (599, 34.9%), followed by the UK (163, 9.5%) and the Netherlands (82, 4.8%). The frequency of publication varied among the top countries, from 3.1% to 34.9%. With productivity defined as the number of citations, only the USA and the UK remained in the same positions. Other countries that made up the top 10 were Switzerland (2785) and Georgia (1537).

5.2. Conceptual framework

In the co-occurrence network (Figure 3a), the top 50 keywords are shown, and the Louvain community detection algorithm [73] identified 4 key research themes, that we qualitatively labeled as:

1) *Asthma heterogeneity* (green); this theme comprised studies related to disentangling asthma heterogeneity using unsupervised data-driven approaches on clinical and biological data to uncover different 'phenotypes' of asthma and their underlying pathophysiological mechanisms.

2) *Atopic diseases* (red), which referred to the investigation of the relationship between asthma and atopic diseases and allergic sensitization.

3) *Respiratory diseases diagnosis* (blue). This theme is characterized by the use of unsupervised and supervised machine learning techniques to classify and predict respiratory diseases, evaluate novel diagnostic tools and identify risk factors.

4) *Environmental exposures and quality of life* (yellow), where we found references to the investigation of the effect of environmental exposure on asthma and the relationship between social and psychological factors and asthma.

These themes were further characterized on the thematic map (Figure 3b), which captures a more detailed definition of the knowledge structure, highlighting the importance of the

themes in the research field. Themes appearing on the upper right panel have high centrality and high density and are known as *motor themes*; these are well-developed themes and play a central role in the conceptual structure of the scientific field. In this quadrant, we found themes related to unsupervised MS techniques applied to asthma and allergic diseases. This is not surprising, given that in the past decades, a large number of studies have adopted these data-driven approaches in an attempt to characterize the heterogeneity of asthma and allergic diseases. The upper left quadrant is characterized by high density and lower centrality; themes in this quadrant are well developed but present a high degree of specialization and are marginal in the overall field. The theme in this area concerns epidemiological studies that aim at investigating asthma diagnosis with a focus on the genetic component. Themes that appear on the lower left part are emerging or declining themes characterized by low density and centrality. The emerging themes here concerned the application of modern ML and deep learning methodologies to asthma research, with a focus on the prediction of the disease, and the study of multimorbidity. Finally, themes in the lower right quadrant have low density and high centrality, meaning that they are essential for the field's scientific structure, but are not well-developed. Here, we found the theme related to the assessment of the relationship between respiratory diseases and COPD with quality of life.

5.3. Research topics in asthma literature

In LDA, the topics are assumed to be latent variables, which need to be meaningfully interpreted. To this end, we examined, for each topic, the top 50 keywords, depicted in (Figure 4a) where word relevance is measured normalizing the posterior word probabilities per topic by the geometric mean of the posterior probabilities for the word across all topics [70]. The eight identified topics reveal important areas of asthma research. Topics 1, 2, 3, 4, 5 and 7 confirmed the themes discovered in the previous analyses and provided additional insights into their structure. In particular, Topic 1, including terms such as 'phenotypes', 'cluster', 'asthma', 'lung', 'COPD', and Topic 4, with words 'sensitization', 'allergic', 'wheeze', 'phenotypes', 'eczema', 'rhinitis', dealt with the application of unsupervised statistical and ML methodologies to understand and characterize asthma heterogeneity and its relation to allergic sensitization, and to investigate the hypothesis related to the atopic march.

Table 3. The 10 most productive countries in the application of Multivariate statistics (MS), Machine learning and artificial intelligence (ML&AI) and Bayesian statistics (BS) to asthma research SCP: single country publications; MCP: multiple country publications.

Rank by # of articles	Country	Articles	Frequency	SCP	MCP	Rank by # of citation	Country	Total Citations	Average Article Citations
1	USA	599	34.9%	596	3	1	USA	23,828	39.8
2	United Kingdom	163	9.5%	163	0	2	United Kingdom	9101	55.8
3	the Netherlands	82	4.8%	82	0	3	Canada	3161	50.2
4	France	77	4.5%	77	0	4	Australia	3034	44.0
5	China	70	4.1%	70	0	5	Switzerland	2785	174.1
6	Australia	69	4.0%	68	1	6	the Netherlands	2751	33.6
7	Canada	63	3.7%	63	0	7	France	2087	27.1
8	Korea	58	3.4%	58	0	8	Italy	1951	36.1
9	Italy	54	3.1%	54	0	9	Sweden	1599	50.0
10	Germany	53	3.1%	53	0	10	Georgia	1537	512.3

Table 4. Distribution of the major methodologies across the selected articles.

Discipline	Methodology	Relative frequency (%)
Multivariate statistics (MS)	<i>Cluster analysis</i>	849/2935 (28.9%)
	<i>Factor analysis</i>	372/2935 (12.7%)
	<i>Principal component analysis</i>	244/2935 (8.3%)
	<i>Discriminant analysis</i>	149/2935 (5.1%)
	<i>Latent Class/Transition models</i>	126/2935 (4.3%)
	<i>SEM</i>	78/2935 (2.7%)
	<i>Markov models</i>	67/2935 (2.3%)
	<i>Mixture models</i>	24/2935 (0.8%)
	<i>Canonical Correlation</i>	19/2935 (0.6%)
	<i>Correspondence analysis</i>	17/2935 (0.6%)
	<i>Latent Variable models</i>	11/2935 (0.4%)
	<i>Multidimensional Scaling</i>	8/2935 (0.3%)
	<i>Profile Regression</i>	1/2935 (0.0%)
Machine learning and artificial intelligence (ML&AI)	<i>Decision Trees</i>	159/2935 (5.4%)
	<i>Machine Learning</i>	145/2935 (4.9%)
	<i>Artificial Neural Networks</i>	76/2935 (2.6%)
	<i>Artificial intelligence</i>	74/2935 (2.5%)
	<i>Unsupervised Learning</i>	66/2935 (2.2%)
	<i>Random Forest</i>	50/2935 (1.7%)
	<i>Support Vector Machine</i>	42/2935 (1.4%)
	<i>Deep Learning</i>	25/2935 (0.9%)
	<i>Supervised Learning</i>	18/2935 (0.6%)
Bayesian statistics (BS)	<i>Bayesian Models</i>	156/2935 (5.3%)
	<i>Spatial and Temporal models</i>	39/2935 (1.3%)
	<i>Bayesian Networks</i>	35/2935 (1.2%)
	<i>Naïve Bayes</i>	4/2935 (0.1%)
	<i>Bayesian Hierarchical models</i>	28/2935 (1.0%)
	<i>Bayesian meta-analysis</i>	27/2935 (0.9%)
	<i>Bayesian Analysis</i>	25/2935 (0.9%)

(Figure 5). Interestingly, 50% of the studies involving ML&AI are linked to Topic 2 (asthma diagnosis), compared to only 0.5% to studies aimed at questionnaires construction and validation. 42.4% of articles adopting Bayesian approaches were concerned with asthma diagnosis, 20.0% investigated environmental exposures, and 10.6% evaluated the efficacy of asthma treatments. Multivariate statistical techniques are evenly spread across the different topics, with higher prevalence in Topics 1, 4, 7 and 8. Clustering approaches and dimensionality reduction techniques, such as principal component analysis (PCA) and factor analysis (FA), were among the most commonly applied methods, followed by supervised machine learning algorithms and different Bayesian modeling approaches (Table 4).

6. Expert opinion

The exciting potential of modern analytical approaches has been welcomed across many different areas of medicine. They promise new opportunities to extract useful information from previously untapped and rapidly growing data sources that can be translated into a more refined disease diagnosis and prognosis, understanding of disease endotypes and etiology, and the development of personalized treatment plans.

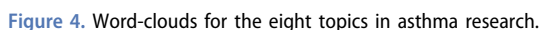
Our analysis of 2935 academic publications offers potentially essential insights. The temporal evolution of academic production followed three significant stages. The initial and development stages were followed by the expansive stage, which started in 2000. In these past two decades, the number of articles using complex analytical techniques in asthma increased exponentially. The predominant themes concerned

the investigation of disease subtyping and etiology, improving diagnostic accuracy, and targeting treatments.

From the year 2000, multivariate statistical models have appeared more and more frequently in asthma research and have covered all the main topics. By contrast, ML and AI techniques have become popular only in the last decade, and they have focused around a few key topics, such as disease diagnosis and treatment evaluation.

These findings offer evidence of a preference toward the inferential goal in asthma research. The popularity of MS methods draws attention to the long-lasting interest in precision medicine and disease subtyping. The extensive use of these techniques echoes the medical scientific community's need to obtain, more than accurate predictions, interpretable solutions that can be used to derive useful insights for the definition of asthma endotypes and their underlying pathophysiological mechanisms. In this context, cluster analysis and unsupervised MS models that, by integrating different types of data, can explore different domains of the disease, represent a flexible and powerful analytical tool to identify homogeneous subpopulations of similar patients that can guide treatment decisions for a given individual.

By contrast, in recent years, the advent of modern large-scale experimentations (such as microarrays and advanced Omics technologies) and the accessibility of new data sources (such as electronic patients' records) have provoked a surge of studies involving cutting-edge ML and AI prediction models in an attempt to obtain automated systems that can accurately diagnose asthma by modeling increasingly large datasets. The volume, velocity and variety of big data often demand complex analytical approaches to enable their analysis. This is



However, despite their potential, real-world implementations of machine learning and AI algorithms in clinical practice are rare. Furthermore, in recent years, there has been a growing concern around accuracy, reliability, security and privacy related to big data analytics in healthcare in general, and asthma specifically [6,74–79]. In a similar fashion, we posit that, among others, there are two main aspects that can explain the growing distrust in big data analytical approaches. One factor is inherent to the methodological characteristics of ML algorithms, while the other can be considered as an extrinsic factor, related to how these models are adopted and contextualized in healthcare. The former aspect can be scrutinized in the context of the long-standing debate on the difference between machine learning and statistics and their effectiveness and applicability. In his seminal paper, Breiman identified two different cultures: the ‘data modeling’ (statistics), and the ‘algorithmic modeling’ (ML and AI) cultures, roughly analogous to the explanatory/predictive dichotomy [80]. In contrast to Breiman’s claim that algorithmic models are more flexible and accurate, especially when dealing with complex big data scenarios, many statisticians have contended that, even though these models accurately predict the outcome of interest, the lack of an explicit data-

We believe that the argument on the incompatibility of these two approaches to investigation is not helpful and

Topic1	17.9	12.3	5.7
Topic2	5.8	49.1	42.4
Topic3	10	5.7	10.6
Topic4	15	6.6	7.3
Topic5	9.5	6.4	8.2
Topic6	8.7	10.7	20
Topic7	18.4	0.5	0
Topic8	14.1	7.7	5.3
	MS	ML	BS

Figure 5. Distribution of the big data analytics approaches in the eight topics.
MS – Multivariate statistics; ML – Machine learning; BS – Bayesian statistics

could be detrimental to scientific progress. We would argue that different emphases pursued by the two ‘cultures’ can be an asset in the attempt to characterize multifaceted heterogeneous long-term conditions. There is no single best model that can address different research questions. Combination of predictive and explanatory models can offer insights on different domains and aspects of the disease that could not be gathered by a single approach, and can be more effective for identifying distinct disease subtypes and explain their etiology. Probabilistic approaches and Bayesian inference should not be regarded with suspicion but should be explored and considered a beneficial modeling framework in which valuable prior knowledge about likely causal mechanisms can be explicitly incorporated.

The ‘extrinsic’ aspect that affects the failure of modern ML and AI approaches in clinical practise, has been extensively discussed [3,6,79,90] and relates to the practical use and implementation of these techniques. Modern computational and analytical methods can help identifying unexpected structure in data to generate hypotheses; however, these should not be accepted without a critical appraisal of the results and corroboration of the derived hypotheses [6]. One of the dangers of ready accessibility of healthcare data, and analytical models for their analysis, is that the process of data mining can become uncoupled from the scientific process of interpretation, understanding the provenance of the data, and external validation [6,90].

We wish to highlight the potential danger of ill-considered use of readily available computational software for data analysis. These have favored the application of complex

methodologies without expressly demanding the knowledge of the methodological properties underpinning the models. Most statistical analyses have assumptions, which often involve properties of the sample, variables, and models. Analytical methodologies have been criticized for the lack of reproducibility and practical usefulness. However, too often, this is the consequence of the incorrect use of these methods: the modeling choices made by inexperienced researchers can be flawed and draw inconsistent conclusions.

To fully harness the potential of ML, AI and modern statistical approaches in shaping the future clinical decision-making system, technological and methodological advances will need to be coupled with clinical and data analytical expertise. There is a mounting necessity to define a unified framework to describe how to undertake effective and ethical research in ML and AI in healthcare [73]. Data scientists and medical professionals should engage in continuous collaboration and conversation to guarantee that the application of data analytics respects both the theoretical and methodological foundations, and the inherent characteristics of the data, the context and the research question under investigation. Not only findings but also methodological choices need to be carefully evaluated to assure relevant scientific discoveries. Bigger healthcare data need to be analyzed in ways that produce meaningful clinical interpretation to ensure that the findings can be translated into better diagnoses and properly personalized prevention and treatment plans [6].

To achieve this, we need to accept that no single source of data or methodology can uncover the underlying complexity of human health and disease. To fully capitalize on the

potential of 'big data', we will have to enhance the effectiveness of collaborative science and encourage the creation of integrated cross-disciplinary teams brought together around technological advances. Such teams are by necessity cross-institutional (and increasingly cross the geographical boundaries), and we need to develop and implement incentives for academic researchers engaged in team science, including clear pathways for a progression in their academic careers, including tenure and promotion, which currently favor 'solo' scientists.

Declaration of interest

A Custovic reports personal fees from Novartis, personal fees from Thermo Fisher Scientific, personal fees from Philips, personal fees from Sanofi, personal fees from Stallergenes Greer, outside the submitted work. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

Reviewer disclosures

Peer reviewers on this manuscript have no relevant financial or other relationships to disclose.

Funding

This paper was not funded.

References

- Papers of special note have been highlighted as either of interest (*) or of considerable interest (**) to readers.
1. Rackemann FM. A working classification of asthma. *Am J Med*. 1947;3:601–606.
2. Pavord ID, Beasley R, Agusti A, et al. After asthma: redefining airways diseases. *Lancet*. 2018;391:350–400.
- **Lancet commission on asthma highlighting the importance of understanding asthma heterogeneity.**
3. Oksel C, Haider S, Fontanella S, et al. Classification of pediatric asthma: from phenotype discovery to clinical practice. *Front Pediatr*. 2018;6. DOI:10.3389/fped.2018.00258
4. Haider S, Custovic A. Breaking down silos in asthma research: the case for an integrated approach. *EMJ Innov*. 2019;3:82–92.
5. Custovic A, Custovic D, Kljaic Bukvic B, et al. Atopic phenotypes and their implication in the atopic march. *Expert Rev Clin Immunol*. 2020;16:9, 873–881.
- **Recent review summarizing the evidence on the multimorbidity of eczema, asthma, and rhinitis, and the implication of results of data-driven analyses on the concept framework of atopic march.**
6. Belgrave D, Henderson J, Simpson A, et al. Custovic A: Disaggregating asthma: big investigation versus big data. *J Allergy Clin Immunol*. 2017;139:400–407.
7. 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;9:921–936.
8. Prosperi MC, Sahiner UM, Belgrave D, et al. Challenges in identifying asthma subgroups using unsupervised statistical learning techniques. *Am J Respir Crit Care Med*. 2013;188:1303–1312.
9. 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;63:974.
10. Oksel C, Granell R, Haider S, et al. Distinguishing wheezing phenotypes from infancy to adolescence. A pooled analysis of five birth cohorts. *Ann Am Thoracic Soc*. 2019;16:868–876.
11. Martinez FD, Wright AL, Taussig LM, et al. Asthma and wheezing in the first six years of life. The group health medical associates. *N Engl J Med*. 1995;332(3):133–138.
12. Belgrave DCM, Simpson A, Semic-Jusufagic A, et al. Joint modeling of parentally reported and physician-confirmed wheeze identifies children with persistent troublesome wheezing. *J Allergy Clin Immunol*. 2013;132(575–583):e512.
13. Belgrave DCM, Granell R, Simpson A, et al. Developmental profiles of Eczema, Wheeze, and Rhinitis: two population-based birth cohort studies. *PLoS Med*. 2014;11.. DOI:10.1371/journal.pmed.1001748.
- **Important study which used Bayesian machine learning to jointly model eczema, wheeze and rhinitis across two population-based birth cohorts, which revealed eight distinct disease profiles - latent classes of allergic diseases.**
14. Belgrave DCM, Buchan I, Bishop C, et al. Trajectories of lung function during childhood. *Am J Respir Crit Care Med*. 2014;189:1101–1109.
15. Duijts L, Granell R, Sterne JAC, et al. Childhood wheezing phenotypes influence asthma, lung function and exhaled nitric oxide fraction in adolescence. *Euro Respir J*. 2016;47(2):510–519.
16. Fitzpatrick AM, Bacharier LB, Guilbert TW, et al. Phenotypes of recurrent wheezing in preschool children: identification by latent class analysis and utility in prediction of future exacerbation. *J Allergy Clin Immunol: In Pract*. 2019;7(915–924):e917.
17. Krautenbacher N, Flach N, Böck A, et al. A strategy for high-dimensional multivariable analysis classifies childhood asthma phenotypes from genetic, immunological, and environmental factors. *Allergy: Eur J Allergy Clin Immunol*. 2019;74:1364–1373.
18. Weinmayr G, Keller F, Kleiner A, et al. Asthma phenotypes identified by latent class analysis in the ISAAC phase II Spain study. *Clini Exp Allergy*. 2013;43:223–232.
19. Fontanella S, Frainay C, Murray CS, et al. Machine learning to identify pairwise interactions between specific IgE antibodies and their association with asthma: a cross-sectional analysis within a population-based birth cohort. *PLoS Med*. 2018;15(11): e1002691.
- **The study which used machine learning to identify a key role of the pairwise interactions between specific IgE antibodies to individual allergenic molecules in the classification of asthma.**
20. Howard R, Belgrave D, Papastamoulis P, et al. Evolution of IgE responses to multiple allergen components throughout childhood. *J Allergy Clin Immunol*. 2018;142:1322–1330.
21. Lazic N, Roberts G, Custovic A, et al. Multiple atopy phenotypes and their associations with asthma: similar findings from two birth cohorts. *Allergy: Eur J Allergy Clin Immunol*. 2013;68:764–770.
22. Roberts G, Fontanella S, Selby A, et al. Connectivity patterns between multiple allergen specific IgE antibodies and their association with severe asthma. *J Allergy Clin Immunol*. 2020;146:821–830.
- **This study used machine learning and network analyses to identify a networks of specific IgE antibodies to individual allergenic molecules in mild/moderate and severe asthma in adults and children.**
23. Simpson A, Tan VYF, Winn J, et al. Beyond atopy: multiple patterns of sensitization in relation to asthma in a birth cohort study. *Am J Respir Crit Care Med*. 2010;181:1200–1206.
24. Bacharier LB, Beigelman A, Calatroni A, et al. Longitudinal phenotypes of respiratory health in a high-risk urban birth cohort. *Am J Respir Crit Care Med*. 2019;199:71–82.
25. Boudier A, Chanoine S, Accordini S, et al. Data-driven adult asthma phenotypes based on clinical characteristics are associated with asthma outcomes twenty years later. *Allergy: Eur J Allergy Clin Immunol*. 2019;74:953–963.
26. Carreiro-Martins P, Viegas J, Papoila AL, et al. CO2 concentration in day care centres is related to wheezing in attending children. *Eur J Pediatrics*. 2014;173:1041–1049.

27. Carreras G, Chellini E, Blangiardo M. A Bayesian model for studying urban air pollution and respiratory symptoms in children. *Inter J Environ Health*. 2012;6(2):125–140.
 28. O' Lenick CR, Chang HH, Kramer MR, et al. Ozone and childhood respiratory disease in three US cities: evaluation of effect measure modification by neighborhood socioeconomic status using a Bayesian hierarchical approach. *Environ Health: A Global Access Sci Source*. 2017;16(36).
 29. Virot E, Godet J, Khayath N, et al. Cluster analysis of indoor environmental factors associated with symptoms of mite allergy. *Ann Allergy, Asthma Immunol*. 2019;123:280–283.
 30. Broadbent L, Manzoor S, Zarcone MC, et al. Comparative primary paediatric nasal epithelial cell culture differentiation and RSV-induced cytopathogenesis following culture in two commercial media. *PLoS One*. 2020;15(3):1–12.
 31. Abdel-Aziz MI, Brinkman P, Vijverberg SJH, et al. Sputum microbiome profiles identify severe asthma phenotypes of relative stability at 12 to 18 months. *J Allergy Clin Immunol*. 2020;147(1):123–134.
 32. Clark H, Granell R, Curtin JA, et al. Paternoster L: Differential associations of allergic disease genetic variants with developmental profiles of eczema, wheeze and rhinitis. *Clini Exper Allergy*. 2019;49(11):1475–1486.
 33. Custovic A, Belgrave D, Lin L, et al. Cytokine responses to rhinovirus and development of asthma, allergic sensitization, and respiratory infections during childhood. *Am J Respir Crit Care Med*. 2018;197:1265–1274.
 34. Wu J, Prosperi MCF, Simpson A, et al. Relationship between cytokine expression patterns and clinical outcomes: two population-based birth cohorts. *Clini Exper Allergy*. 2015;45:1801–1811.
 35. Dong Z, Ma Y, Zhou H, et al. Integrated genomics analysis highlights important SNPs and genes implicated in moderate-to-severe asthma based on GWAS and eQTL datasets. *BMC Pulm Med*. 2020;270.
 36. Moss LC, Gauderman WJ, Lewinger JP, et al. Using Bayes model averaging to leverage both gene main effects and G $\times$ E interactions to identify genomic regions in genome-wide association studies. *Genet Epidemiol*. 2019;43:150–165.
 37. Ober C, McKennan CG, Magnaye KM, et al. Expression quantitative trait locus fine mapping of the 17q12-21 asthma locus in African American children: a genetic association and gene expression study. *Lancet Respir Med*. 2020;8:482–492.
 38. Modena BD, Bleecker ER, Busse WW, et al. Gene expression correlated with severe asthma characteristics reveals heterogeneous mechanisms of severe disease. *Am J Respir Crit Care Med*. 2017;195:1449–1463.
 39. Su MW, Lin WC, Tsai CH, et al. Childhood asthma clusters reveal neutrophil-predominant phenotype with distinct gene expression. *Allergy: Eur J Allergy Clin Immunol*. 2018;73:2024–2032.
 40. Yeh YL, Su MW, Chiang BL, et al. Genetic profiles of transcriptomic clusters of childhood asthma determine specific severe subtype. *Clin Exp Immunol*. 2018;48:1164–1172.
 41. Goto T, Camargo CA, Faridi MK, et al. Machine learning approaches for predicting disposition of asthma and COPD exacerbations in the ED. *Am J Emerg Med*. 2018;36:1650–1654.
 42. Himes BE, Dai Y, Kohane IS, et al. Prediction of Chronic Obstructive Pulmonary Disease (COPD) in asthma patients using electronic medical records. *J Am Med Inf Assoc*. 2009;16:371–379.
 43. Kothalawala DM, Kadalayil L, Weiss VBN, et al. Prediction models for childhood asthma: a systematic review. *Pediatr Allergy Immunol*. 2020;31:616–627.
 44. Ullah R, Khan S, Ali H, et al. A comparative study of machine learning classifiers for risk prediction of asthma disease. *Photodiagnosis Photodyn Ther*. 2019;28:292–296.
 45. Spyroglou II, Spöck G, Chatzimichail EA, et al. A Bayesian logistic regression approach in asthma persistence prediction. *Epidemiol Biostatistics Public Health*. 2018;15:e12777-12771-e12777–12714.
 46. Tibble H, Tsanas A, Horne E, et al. Predicting asthma attacks in primary care: protocol for developing a machine learning-based prediction model. *BMJ Open*. 2019;9(7).
 47. Latzin P, Fuchs O. Asthma diagnosis in children: more evidence needed. *Lancet Child Adolesc Health*. 2017;1:83–85.
 48. Oksel C, Granell R, Mahmoud O, et al. Causes of variability in latent phenotypes of childhood wheeze. *J Allergy Clin Immunol*. 2019;143(1783–1790):e1711.
 49. Custovic A. “Asthma” or “Asthma Spectrum Disorder”? *J Allergy Clin Immunol Pract*. 2020;8:2628–2629.
 50. Tamhane D, Sayyad S. Big data using HACE theorem. *Int J Advanc Res Comput Eng Technol*. 2015;4:18–23.
 51. Rajula HSR, Verlato G, Manchia M, et al. Comparison of conventional statistical methods with machine learning in medicine: diagnosis, drug development, and treatment. *Medicina (B Aires)*. 2020;56:455.
 52. Mardia KV, Kent JT, Kent JT, et al. *Multivariate Analysis*. London ; New York : Academic Press; 1979.
 53. Azzolina D, Baldi I, Barbati G, et al. Machine learning in clinical and epidemiological research: Isn't it time for biostatisticians to work on it?. *Epidemiology Biostatistics and Public Health*. 2019;16:4.
 54. Zhang Z, Beck MW, Winkler DA, et al. opening the black box of neural networks: methods for interpreting neural network models in clinical applications. *Ann Transl Med*. 2018;6(11):216.
 55. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. *J Allergy Clin Immunol*. 2020;145:463–469.
- **A review article summarizing the literature on the use of electronic health record data for clinical research in the fields of allergy, asthma, and immunology, highlighting the potential, challenges, and implications of natural language processing techniques.**
56. Dalianis H. clinical text mining: secondary use of electronic patient recordsnull. Springer Nature, Switzerland: Springer; 2018.
 57. Gill CJ, Sabin L, Schmid CH. Why clinicians are natural Bayesians. *BMJ*. 2005;330:1080–1083.
 58. Bland JM, Altman DG. Bayesians and frequentists. *BMJ*. 1998;317:1151–1160.
 59. Falagas ME, Pitsouni EI, Malietzis GA, et al. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *Faseb J*. 2008;22:338–342.
 60. Hastie T, Tibshirani R, Friedman JH. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer; 2009.
 61. Mitchell TM. *Machine Learning*. New York, NY United States: McGraw-Hill; 1997.
 62. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. Cambridge, MA, USA: MIT Press; 2016.
 63. Murphy KP. *Machine Learning: a probabilistic perspective*. Cambridge, MA, USA: MIT Press; 2012.
 64. Gelman A, Carlin JB, Stern HS, et al. *Bayesian data analysis*. Boca Raton: CRC Press; 2013.
 65. Aria M, Cuccurullo C. bibliometrix: an R-tool for comprehensive science mapping analysis. *J Informetrics*. 2017;11:959–975.
 66. Cobo MJ, López-Herrera AG, Herrera-Viedma E, et al. An approach for detecting, quantifying, and visualizing the evolution of a research field: a practical application to the Fuzzy Sets Theory field. *J Informetrics*. 2011;5:146–166.
 67. Cobo MJ, Martínez MA, Gutiérrez-Salcedo M, et al. Herrera-Viedma E: 25 years at Knowledge-Based Systems: a bibliometric analysis. *Knowledge-Based Syst*. 2015;80:3–13.
 68. Blei DM, Ng AY, Jordan MI: Latent dirichlet allocation. *J Mach Learn Res*. 2003;3:993–1022.
 69. Blei DM. Probabilistic topic models. *Commun ACM*. 2012;55:77–84.
 70. Tontodimamma A, Nissi E, Sarra A, et al. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*. 2021; 126:157–179.
 71. MATLAB: **version 9.9.0.1467703 (R2020b)**. Edited by: The MathWorks Inc.; 2020.

72. Aparicio G, Iturralde T, Maseda A: Conceptual structure and perspectives on entrepreneurship education research: a bibliometric review. *Eur Res Manage Business Econ.* **2019**;25:105–113.
73. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E: Fast unfolding of communities in large networks. *J Stat Mech Theor Exper.* **2008**;2008:P10008.
74. Vollmer S, Mateen BA, Bohner G, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ.* **2020**;368:l6927.
75. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **2019**;17:195.
- **The article exploring the main challenges and limitations of AI in healthcare highlighting the steps required to translate these potentially transformative technologies from research to clinical practice.**
76. Jones LD, Golan D, Hanna SA, et al. Artificial intelligence, machine learning and the evolution of healthcare. *Bone Joint Res.* **2018**;7:223–225.
77. Wachter RM. *The digital doctor: hope, hype, and harm at the dawn of medicine's computer age.* New York: McGraw-Hill Education, **2015**.
78. Dhindsa K, Bhandari M, Sonnadara RR: What's holding up the big data revolution in healthcare? *BMJ.* **2018**;363:k5357.
79. Belgrave D, Custovic A: The importance of being earnest in epidemiology. *Acta Paediatrica.* **2016**;105:1384–1386.
80. Breiman L. Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statist Sci.* **2001**;16:199–231.
81. Mukhopadhyay S, Wang K. Breiman's "Two Cultures" Revisited and Reconciled. *ArXiv.* **2005**.13596.
- **This article presents establishes a link between the two statistical cultures and describes the challenges and gains of the integrated statistical thinking.**
82. Efron B. Prediction, Estimation, and Attribution. *J Am Stat Assoc.* **2020**;115:636–655.
83. Guidotti R, Monreale A, Ruggieri S, et al. A survey of methods for explaining black box models. *ACM Comput Surv.* **2018**;51 (Article):93.
84. Murdoch WJ, Singh C, Kumbier K, et al. Definitions, methods, and applications in interpretable machine learning. *Proc Nat Acad Sci.* **2019**;116:22071–22080.
85. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Mach Intell.* **2019**;1:206–215.
86. Hand DJ. Classifier technology and the illusion of progress. *Statist Sci.* **2006**;21:1–14.
87. Duin RPW. A note on comparing classifiers. *Pattern Recognit Lett.* **1996**;17:529–536.
88. Holte RC. Very simple classification rules perform well on most commonly used datasets. *Mach Learn.* **1993**;11:63–90.
89. Christodoulou E, Ma J, Collins GS, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* **2019**;110:12–22.
90. Williams SM, Moore JH. Big data analysis on autopilot? *BioData Min.* **2013**;6:22.