Current state and prospects of artificial intelligence in allergy

Merlijn van Breugel[1], Rudolf S. N. Fehrmann[2], Marnix Bügel[3], Faisal I. Rezwan[4], John Holloway[4], Martijn Nawijn[5], Sara Fontanella[6], A Custovic[6], and Gerard Koppelman[1]

[1]Universitair Medisch Centrum Groningen Beatrix Kinderziekenhuis
[2]Universitair Medisch Centrum Groningen Afdeling Oncologie
[3]MIcompany
[4]University of Southampton School of Human Development and Health
[5]Universitair Medisch Centrum Groningen Groningen Research Institute for Asthma and COPD
[6]Imperial College London National Heart and Lung Institute

April 21, 2023

Abstract

The field of medicine is witnessing an exponential growth of interest in Artificial Intelligence (AI), which enables new research questions and the analysis of larger and new types of data. Nevertheless, applications that go beyond proof of concepts and deliver clinical value remain rare, especially in the field of allergy and immunology. This narrative review provides a fundamental understanding of the core concepts of AI and critically discusses its limitations and open challenges, such as data availability and bias, along with potential directions to surmount them. We provide a conceptual framework to structure AI applications within this field and discuss forefront case examples. Most of these applications of AI and machine learning in allergy concern supervised learning and unsupervised clustering, with a strong emphasis on diagnosis and subtyping. A perspective is shared on guidelines for good AI practice to guide readers in applying it effectively and safely, along with prospects of field advancement and initiatives to increase clinical impact. We anticipate that AI can further deepen our knowledge of disease mechanisms and contribute to precision medicine in allergy.

Current state and prospects of artificial intelligence in allergy

Merlijn van Breugel[1,2,3], Rudolf S. N. Fehrmann[4], Marnix Bügel[3], Faisal I. Rezwan[6,7], John Holloway[6,8], Martijn C. Nawijn[2,5], Sara Fontanella[9,10], Adnan Custovic[9,10], Gerard H. Koppelman[1,2,*]

1. University of Groningen, University Medical Center Groningen, Beatrix Children's Hospital, Department of Pediatric Pulmonology and Pediatric Allergology, Groningen, the Netherlands

2. University of Groningen, University Medical Center Groningen, Groningen Research Institute for Asthma and COPD (GRIAC), Groningen, the Netherlands

3. MIcompany, Amsterdam, the Netherlands

4. Department of Medical Oncology, University Medical Center Groningen, University of Groningen, Groningen, the Netherlands

5. Department of Pathology & Medical Biology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

6. Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, United Kingdom

7. Department of Computer Science, Aberystwyth University, Aberystwyth, United Kingdom

8. National Institute for Health and Care Research Southampton Biomedical Research Centre, University Hospitals Southampton NHS Foundation Trust, Southampton, United Kingdom

9. National Heart and Lung Institute, Imperial College London, United Kingdom

10. National Institute for Health and Care Research Imperial Biomedical Research Centre (BRC)

*Corresponding author:

Prof. dr. Gerard H. Koppelman

Department of Pediatric Pulmonology and Pediatric Allergology

Beatrix Children's Hospital

University Medical Center Groningen

PO Box 30.001

9700 RB Groningen, the Netherlands

Email: g.h.koppelman@umcg.nl

Phone: + 31 50 3611036

Fax: + 31 50 3614235

Word count main text: 6260 [limit 6500]

Figures: Figure 1-5, Supplementary Figure 1

Table: Glossary, Table 1

Author contributionsM.v.B. and G.H.K. devised and set out the overall outline of this review. R.S.N.F., M.B., and M.C.N. have provided critical input on the outline and figure drafts for improved positioning and relevance for the readership. F.I.R., J.H., S.F., and A.C. have been invited to the author group to bring in additional domain expertise, for which they have written and/or sharpened the respective sections. All authors have provided feedback on the first complete draft of the manuscript. M.v.B. and G.H.K. wrote this first draft and have revised it based on inputs from all authors. All authors have read and approved the final version of this manuscript.

Conflict of Interest statement All authors declare to have no conflict of interests.

Abstract

The field of medicine is witnessing an exponential growth of interest in Artificial Intelligence (AI), which enables new research questions and the analysis of larger and new types of data. Nevertheless, applications that go beyond proof of concepts and deliver clinical value remain rare, especially in the field of allergy and immunology. This narrative review provides a fundamental understanding of the core concepts of AI and critically discusses its limitations and open challenges, such as data availability and bias, along with potential directions to surmount them. We provide a conceptual framework to structure AI applications within this field and discuss forefront case examples. Most of these applications of AI and machine learning in allergy concern supervised learning and unsupervised clustering, with a strong emphasis on diagnosis and subtyping. A perspective is shared on guidelines for good AI practice to guide readers in applying it effectively and safely, along with prospects of field advancement and initiatives to increase clinical impact. We anticipate that AI can further deepen our knowledge of disease mechanisms and contribute to precision medicine in allergy.

Keywords (max 5): Artificial Intelligence; Deep Learning; Machine Learning; Diagnosis; Precision Medicine

List of abbreviations

AI - Artificial intelligence

CE - Conformité Européene

EHR - Electronic health records

FDA - US Food & Drug Administration

GIGO - Garbage in garbage out

ML - Machine learning

RCT - Randomized control trial

Glossary

Insert Glossary Table

Introduction

As of February 2023, the US Food & Drug Administration (FDA) has approved 521 medical applications that utilize Artificial Intelligence (AI) and Machine Learning (ML). Most of these (75%) are in radiology, followed by cardiology, hematology and neurology. Similar trends are observed in Conformité Européene (CE)-marked medical devices incorporating AI within the European Union. Currently, no registered AI and ML-based applications are being utilized in the field of allergy and immunology. One can therefore question if this field is missing out on new research opportunities and clinical applications either because of insufficient access to AI applications or a lack of awareness of potential applications. However, given the rapid pace of technological advances, it can be anticipated that AI and ML algorithms will be increasingly applied in allergy and immunology soon.

Over the past decade, medicine has witnessed an exponential growth of interest in AI and the yearly number of scientific articles on AI has increased tenfold since 2012. This trend is fueled by the explosion of (bio)medical data, including multi-omics, image data, and digital electronic health records (EHRs), along with advancements in computing power. These developments have paved the way for advanced analytical approaches to address new research questions on large-scale datasets. Traditional analytical techniques are no longer adequate to handle such data complexity, volume, and structure. The introduction of accessible software and methodological advancements in AI have further promoted the use of AI in the (bio-)medical field. Most importantly, ML and AI can identify complex patterns in vast amounts of data, such as images, text, or audio and deliver superior predictive power, often surpassing traditional statistical methods.

This review provides a fundamental understanding of ML and AI's core concepts. A framework is presented to structure the broad umbrella term AI, and an overview of several state-of-the-art applications of AI in medicine and allergy and immunology, specifically, is provided. The focus is on applications that preferably adhere to any, and ideally multiple, of the following conditions: (1) are externally or prospectively validated, (2) demonstrate a positive effect on clinically relevant patient outcomes, (3) FDA and/or CE approval, (4) outperform traditional methods, and (5) answer research questions where traditional analytical techniques fail. Additionally, we critically discuss the limitations and open challenges of AI applications and share an outlook on good practices of AI and ML in allergy and immunology.

The fundamentals and terminology of artificial intelligence

Machine learning and deep learning

AI is the discipline in computer science that develops computer systems that can simulate human intelligence and perform tasks that generally require humans (see Glossary for key concepts). This discipline can be further narrowed down into ML and its subdiscipline *deep learning*(Figure 1). ML can be described as an algorithm that learns from data by automatically mapping input data to an output prediction. While this draws a parallel to traditional regression methods such as ordinary least squares, most machine learning techniques have the advantage of inherently modeling complex patterns, including non-linearity and interaction effects. ML concentrates primarily on prediction and finding patterns in vast amounts of data without making prior assumptions about the distribution of these data. The predictive performance generally improves when more (high-quality) data is fed to the algorithm.

Insert Figure 1

Over the past years, the ML subfield of *deep learning* has gained tremendous popularity, as it has yielded superior results in analyzing*unstructured data* such as medical images, text data, and audio data. This technique is based on large artificial neural networks (ANNs). ANNs form networks inspired by the biological animal brain, consisting of multiple layers of processing units called neurons. Deep learning methods can detect complex data relationships by automatically compressing data and distilling relevant features in various levels of abstraction. This makes it different from statistical approaches such as regression methods, which require explicitly defining independent variables and making assumptions about their relationship to the outcome variable. Another advantage of deep learning is its ability to continue learning and improve performance with larger datasets. Besides applications in computer vision, which is the ability to interpret image data by an AI system, deep learning has also propelled natural language processing (NLP) forward, which is the capacity of a computer to understand written and spoken human language. We refer to recent, extensive reviews that cover the subfields of deep learning and its applications in medicine. A state-of-the-art example is the detection of diabetic retinopathy from retinal images, for which the IDx-DR deep-learning-based software has been FDA-approved and validated in a clinical setting. Relatedly, deep learning approaches have outperformed trained physicians in breast cancer detection using imaging data, with currently nine applications FDA-approved. Some of the latest AI breakthroughs that ignited the general public's interest are based on deep learning approaches. These involve generative models that are trained to create new data. Examples include deep fakes, DALL-E (an OpenAI application that creates figures and art based on written descriptions), and most recently, ChatGPT, an AI tool that generates highly realistic written text based on user prompts. Figure S1 displays two AI-generated illustrations of this review's topic.

Learning strategies

A useful categorization of AI is made on the learning strategy, which defines how an algorithm learns from data. Three different approaches are distinguished: supervised, unsupervised, and reinforcement learning. We provide a conceptual framework to structure AI applications based on learning strategy, learning goal, data modality, and medical domain in Figure 2.

Insert Figure 2

Supervised Machine Learning

Most machine learning applications concern *supervised learning* , where a model is trained to predict a known outcome, called the target variable, label or dependent variable. Supervised learning often requires manual labeling of the target variable. Supervised learning can be applied in almost all medical domains, such as disease diagnosis, treatment outcome prediction, or classifications in medical imaging. One such example is the FDA-approved *Koios DS for Breast* application. This tool supports clinicians in breast cancer diagnosis by classifying ultrasound images into benign, probably benign, suspicious, and probably malignant. It has

been shown to improve assessment performance compared to the clinician's assessment in a retrospective study. Supervised machine learning has also surged in screening, predicting, contact and tracing, and drug development. For example, during the COVID-19 pandemic, supervised ML was used to predict which potential drug compounds could be effective against SARS-CoV-2 targets by developing prediction models for the drug-likeliness of candidate compounds from chemical libraries based on chemical descriptors. Within supervised ML, gradient-boosted decision tree methods have been among the most popular and performant, with the leading algorithms being Random Forest, XGBoost, and LightGBM (Glossary).

Unsupervised Machine Learning

In *unsupervised learning,* the aim is to learn groupings in data or reduce their dimensionality. Contrary to its supervised counterpart, there are no known labels to predict. Unsupervised learning is often used for clustering analysis. Here, the algorithm aims to describe the data in a limited number of clusters or groups, where goodness-of-fit tests determine the most parsimonious model. An example[30] is the discovery of asthma phenotypes based on longitudinal wheezing patterns or clinical variables. Techniques for unsupervised learning are latent class analysis (LCA), k-means clustering, principal component analysis (PCA), and Multidimensional Scaling (MDS). Recently, also *semi-supervised learning* has grown in popularity, which aims to overcome the lack of sufficiently large, labeled datasets and the tedious task of manual labeling. It leverages a dataset of yet unlabeled data to improve the performance of a model that is initially trained on labeled data.

Reinforcement learning

Reinforcement learning has recently delivered breakthroughs in the biomedical field. This strategy is characterized by an iterative process that aims to take actions that deliver maximum reward based on a defined objective function. This is comparable to nature, where animals have learned to interpret signs such as hunger as negative, whereas satiety after food intake is seen as positive reinforcement. When animals learn how to behave to gain optimal positive reinforcement, they show reinforcement learning. Applications within medicine have indicated its potential. For example, an "AI Clinician" algorithm has been developed to improve the treatment of sepsis by suggesting the personalized treatment of intravenous fluids and vasopressors. While still requiring prospective validation, an independent validation cohort was used to assess this algorithm, showing that mortality was lowest when clinicians' treatment policy was close to AI recommended policy and higher when deviating from it.

Challenges and pitfalls for AI application in medicine

AI is not without pitfalls, and serious challenges must be overcome to deliver its full potential. The most critical challenges are described below, with potential directions to surmount them. For a more in-depth discussion of AI's current most pressing issues, the reader is referred to several excellent reviews.

Data

AI systems and models are as good as the data they learn from. This relates to the data's (1) quality and quantity, (2) suitability, and (3) availability. The first challenge refers to data quality and quantity. Low quality of input data, leading to biased outcomes, is often referred to as the GIGO ('garbage in garbage out') principle. Data quantity also remains challenging, since AI models are extremely 'data hungry', especially for deep learning methods. The availability and quality of data labels are critical, as label inaccuracies directly impair model reliability. In particular for images, manual labeling of images is time-consuming. Combining and harmonizing multiple datasets is increasingly used to overcome these data limitations. The use of synthetic data may also help, where additional data is generated by simulating from a known data distribution, which has been shown to improve model performance. Similarly, in image analysis, data augmentation is often used to (fictively) increase the data sample size by applying data transformations on existing (non-synthetic

data points). Another strategy to improve model reliability on relatively small datasets is *transfer learning* , especially popular in NLP and image analysis. This technique enables researchers to train a complex model on relatively small datasets by recalibrating existing parameters of known models.

Data suitability poses a second challenge. Akin to traditional analytical methods, AI approaches need adequate study designs to yield reliable outcomes, from data collection to the appropriate analytical strategy. Training algorithms based on unsuitable data may lead to biased outcomes. For example, it is increasingly clear that AI and ML algorithms can engrain racial bias when models are trained using racially imbalanced data sets.

Data availability may pose a third challenge, as data is often siloed within individual institutions, and curated, publicly accessible clinical datasets remain rare. The reason for this includes patient privacy, lack of data-sharing infrastructure, and competition among institutions. In immunology, efforts are being made to break open silos and democratize datasets. Examples include the National Institutes of Health (NIH)-curated resources on open-access COVID-19 data, or the European Health Data Space for the safe exchange and reuse of health data. These developments are aided by novel data sharing and integration approaches, such as federated learning, where a model is centrally trained while the data are kept locally. Recently Swarm Learning was introduced, a decentralized machine-learning approach that does not require central coordination. The researchers demonstrate that the model outperforms individual sites in disease classification while retaining complete confidentiality.

Explainability

The lack of explainability of AI algorithms hampers clinical implementation. Unlike statistical methods such as regression, which are inherently explainable, the learned patterns of AI models are more complex, and their estimated parameters are not directly interpretable (Figure 3).

Insert Figure 3

Deep learning models suffer from this due to their hidden learning behavior and up to billions of parameters. One recent study made this vividly clear. When researchers trained a model to distinguish COVID-19 patients with pneumonia from those with other respiratory diseases based on chest radiographs, the algorithm based its prediction on the printed dates on the radiological images; it found a shortcut and classified all patients dated since 2020 as COVID-19 cases. Thus, there is a growing demand for 'white box' approaches, referring to methods and models that are easy to explain and interpret. This need is further amplified when the aim is to bring applications to clinical practice, which has many technical, medical, legal, and ethical dimenions**.** The urgent need for explainability has accelerated methodological innovations to 'open the black box'. Relevant examples are SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and CAM (Class Activation Maps). For example, SHAP was recently implemented to describe the contribution of features selected for inclusion in asthma prediction models. These analytical methods calculate how each input feature contributes to each prediction, providing detailed insights into the learning patterns of the AI model.

Validation and generalizability

A structured modeling process is essential in developing an ML prediction model to create a reliable model and establish confidence in its outcomes. There are many ML algorithms, and it is difficult to tell which will perform best beforehand. This is called the *no-free lunch* theorem, which emphasizes the need to develop and evaluate ML models iteratively. Thus, multiple ML methods should be applied to the data and their performance compared. Figure 4 depicts the steps to build a supervised learning prediction model for disease risk. The steps needed for unsupervised learning overlap to a large extent. Skipping or mismanaging these steps poses a risk to model reliability, for example, by not properly separating the training and validation data, which may lead to overfitting of the prediction model.

Insert Figure 4.

One of AI's significant benefits is its ability to scale intelligence at an unprecedented pace. In a time in which a clinician could diagnose a single patient, an AI system could analyze unlimited number of patients, at least in theory. However, the same scalability holds for mistakes and faulty diagnoses, and validation is of the utmost importance to prevent the lack of generalizability of AI models. AI models tend to 'overfit' on the training data, which results in a model that works seemingly well on the training population but poorly predicts future or other patient outcomes, especially in high-dimensional models. One example was from IBM's Watson, which recommended unsafe cancer treatments, because it was trained on a sample size too limited for its dimensionality. For models to be more broadly applicable and generalizable to other populations, diligent validation and replication (in external datasets) are paramount. Unfortunately, this is often insufficient or altogether missing, respectively. Even FDA-approved AI applications fall short in this domain: Only 11/118 FDA applications (up until 2021) reported a validation set of more than 1,000 samples, and only 19/118 reported a multi-reader, multicenter validation study. Site-specific recalibration or retraining on multiple datasets are solutions to adapt a model to another context, although caution is required to avoid spurious learning patterns.

Randomized control trials (RCTs) or prospective validations are scarce in medical AI[80,81]. Most applications are only tested on retrospective data and have not passed prospective validation in an independent dataset. New guidelines have been emerging for reporting and evaluating RCTs with an AI intervention component in the past two years, such as the CONSORT-AI standard and SPIRIT-AI. A systematic review from 2022 reported that none of the 41 assessed RCTs adhered to this standard and suggested that AI applications with FDA approval do not always prove efficacy. Thus, the clinical utility and safety remain uncertain, providing a clear direction for future research to confidently implement AI in clinical medicine.

Ethical considerations

AI systems often rely on and are trained on confidential personal data, such as health records, imaging, or genomic data. The more voluminous these data become, especially with integrating multiple data sources and unlocking new data sources, the more critical privacy becomes. The EU's General Data Protection Regulation (GDPR) already provides a 'right to explanation' when decisions are based on "automated processing" such as AI. There is a complicated relationship between privacy and trust. If the mechanisms of algorithms remain hidden for privacy reasons, this could also impede trust in the solution and slow down adoption by patients and clinicians. Furthermore, being overprotective of privacy in data collection, usage, and sharing can also hinder potential patient benefits from using these data to drive AI solutions for novel diagnostic or therapeutic options. Novel approaches are emerging that preserve privacy without slowing down innovation, such as the generation of synthetic data. Rather than (pseudo)-anonymizing samples, AI-generated synthetic data samples can be used for safe data sharing or even new model development.

While AI systems are not moral agents, their decisions can have ethical consequences. Especially bias and fairness are two key concepts in this context, and various cases of embedded biases exist in developed models. The 2021 AI action plan from the FDA, warns that biases in healthcare systems, such as racial or gender biases, can be inadvertently introduced to algorithms. This will lead to research conclusions and applications biased toward specific populations while overlooking others. If they are not corrected, this could further reinforce biases and exacerbate health inequalities experienced by certain underrepresented populations by excluding them from AI-driven medical innovations. Therefore, researchers need to ensure that the training sample is diverse and represents any future population to which the AI model will be applied.

While the above risks are important, it is essential to realize that humans are not free from implicit biases. For instance, cardiologists are trained to recognize symptoms of coronary artery disease more frequently in men, resulting in underdiagnosis in women. The advantage of data and algorithms is that biases may be detected, corrected, or prevented. From the study's outset, during the data collection phase, investigators should strive for a representative training dataset that resembles the data distribution the algorithm would encounter once

deployed. Before model development, guidelines have been defined to assess the risk of algorithmic bias, such as the PROBAST tool. Likewise, new techniques for the modeling phase are emerging that can help to mitigate bias, such as adversarial debiasing. Lastly, dedicated tools have been developed to evaluate the fairness of algorithms along a variety of fairness definitions, like the open-source Python library AI Fairness 360.

When the above considerations are not managed adequately, an AI system may make mistakes. This raises the intricate question of (moral) accountability, which becomes increasingly pressing with more clinical applications in place. However, the traditional notion of accountability is problematic in the context of an AI system. It is questionable whether a clinician can be held responsible for such a system's decisions. Furthermore, the system's complexity can make it infeasible for the clinician, and sometimes even the designer, to understand precisely why certain decisions are made. Therefore, we anticipate that the introduction of AI in clinical medicine will first be limited to decision support systems, with the final clinical decision to be made by the caring physician.

Clinical implementation

Despite exciting showcases, AI has been criticized for underdelivering tangible clinical impact. Translating solid AI models to effective action remains an open challenge and actual clinical use is still nascent. Recently, even with the surge of COVID-19-related AI research, the clinical value of AI applications remained limited. Important challenges for clinical implementation include questionable clinical advantages, inadequate reporting, and adoption and integration in clinical practices.

Developers of algorithms are also urged to be transparent and complete in their reporting to provide a fair view of improving patient care. RISE criteria (Regulatory aspects, Interpretability, Interoperability, Structured Data, and Evidence) can support overcoming major pitfalls in developing AI applications for clinical practice. Recently, the DECIDE-AI guideline has been introduced as a reporting checklist of AI-based (early-stage) clinical evaluation of decision support systems. In addition, clinicians and patients must adapt to working with and trusting new AI systems, and such behavioral change is notoriously hard. There is a need for (better) AI education for clinicians that will need to adapt to new roles and tools to support them in their decision-making. To smoothen this transition, integration into the medical education system has been proposed. The recent American Academy of Asthma, Allergy and Immunology workgroup has underscored a knowledge and an educational gap in the allergy and immunology field. Furthermore, interoperability of AI systems is vital to ensure that they can be integrated with existing clinical and technical workflows across sites and health systems.

Current state of AI in the allergy research field

AI applications within the allergy field can be broadly categorized into three domains: clinical research, fundamental research, and drug and therapy development (Figure 5).

Insert Figure 5.

Many studies leveraging AI have been published in the research setting. Virtually all applications concern supervised learning and unsupervised clustering, whereas semi-supervised learning and reinforcement learning are mainly absent. Overviews of the use of ML in asthma and eczema research over the last seven decades have recently been published.

Clinical research

Diagnosis of allergic diseases

The diagnosis or classification of allergic disease has been the area in which AI has been applied most, an exemplary case of supervised learning[105–109]. ML has used a wide range of data sources to improve allergy or asthma diagnosis: *text* data from electronic health records (EHRs), *sound* data of wheezes,*image* data from lung CT scans, or large-scale *multi-omics*data. The extraction of relevant clinical features from EHRs using NLP has successfully diagnosed (childhood) asthma in discovery and replication cohorts. In a study of a US birth cohort study, Seol*et al.* (2020) applied an AI algorithm to define asthma using established predictive and diagnostic criteria in 8196 children. Of all patients that met those criteria, 30% did not have a physician diagnosis of asthma, signifying the potential for early disease identification and population management with EHRs.

Additionally, several studies have investigated the potential of omics data for diagnosis. One study developed an ML model that diagnosed IgE-sensitized allergic disease in 16-year-old children based on nasal cell DNA-methylation of only three CpG sites. External validation in an independent cohort indicated the prospect of reproducible epigenetic tests for diagnosis. Alag *et al.* (2019) pursued a similar approach to diagnosing food allergy, where neural networks were trained on blood epigenetic markers. The predictive markers were subsequently associated with a 13-gene profile linked to immune response. This study highlights the potential of novel diagnostic approaches to food allergy.

ML-based modelling of the component-resolved diagnostic multiplex array data has shown that component-specific IgE responses to multiple allergenic proteins are functionally coordinated and co-regulated, and that the networks of interactions are associated with asthma diagnosis and severity. Machine learning has also been used to predict disease risk or persistence. In a prospective study of 704 children aged 2 to 13 months, unsupervised clustering on 16S rRNA data was used to identify profiles of longitudinal changes in nasal airway microbiota that were significantly associated with asthma risk at age seven. These results affirm that the microbiome plays a vital role in the early development of asthma and show promise for early identification and prevention strategies. In another study, a supervised machine learning model was able to predict asthma persistence in almost 10,000 patients diagnosed before age 5 for persistence by age 10. The XGBoost algorithm delivered the most robust performance (AUC=0.86), using clinically relevant features such as the number of (non) asthma-related visits before age five and noninvasive pulse oximetry data. The study was not independently replicated, which is essential in pursuing clinical support tools. Kothalawala *et al.* (2021) used data from birth cohorts to train and validate two predictive models, CAPE and CAPP, to predict the likelihood of asthma at school-age using predictors from 0-2 and 0-4 years of age, respectively. Predictive performance was externally validated in the Manchester Asthma and Allergy Study (MAAS) cohort. Support Vector Machine (SVM) algorithms provided the best performance for both the CAPE (AUC=0.71) and CAPP (AUC=0.82) models, and both demonstrated good generalisability in the replication cohort, performing better than previous regression-based models.

AI guided image analysis has been performed to diagnose eczema. One study developed a classifier of atopic dermatitis in multiphoton tomography images, reaching over 97% accuracy through transfer learning. Highlighted areas of interest in the images could support clinicians in faster diagnosis.

*Prediction of asthma exacerbations and hospitalizations*Asthma exacerbations are related to increased morbidity, mortality, and healthcare use, yet these are challenging to predict. Several studies have applied ML to predict exacerbations. In a large study involving EHR data from 60.000 patients, researchers used different ML techniques in a supervised setup to predict three exacerbation outcomes: oral glucocorticoid bursts, ED visits, and hospitalization. The study achieved a ROC AUC of 0.88 on the latter outcome, which is significantly higher than the results of previous studies (AUC of 0.77); this was replicated in an independent cohort. Important predictors for hospitalization included oral glucocorticoid burst, inhaled corticosteroid, and blood creatinine, the latter being unexpected. Another study used self-reported daily home monitoring data of asthma symptoms and peak expiratory flow, which were reduced in dimensionality using PCA and then fed to various supervised ML methods. The best model achieved a sensitivity of 90% and specificity of 83%, predicting severe asthma exacerbations on the same day or up to three days in the future. A more extensively validated example is Asthma-Guidance and Prediction System (a-GPS), an AI tool to optimize

asthma management. A-GPS uses NLP on open text from EHRs to provide clinicians with the most relevant clinical information. In a randomized control trial, the tool significantly reduced the time for reviewing EHRs (11.3 to 3.5 min), but no significant change in clinical outcome (i.e., exacerbations) was observed. Sensor data from an electronic multi-dose dry powder inhaler (eMDPI), such as inhalation volume and duration, has also been utilized to predict exacerbations with a ROC AUC of 0.83.

Disease management

Medication non-adherence in allergic diseases is common in clinical practice and can negatively impact disease control. To address this issue, researchers explored ML approaches for disease management and medication adherence. One such approach involves using ML to provide early warnings for loss of control in the Asthma Mobile Health Study data of 5,875 patients, containing over 75,000 daily surveys on symptoms and medicine use, medical history, demographics, location and EuroQol 5D questionnaire. The supervised classifier obtained an AUC of 0.87, but peak flow readings did not further enhance its performance. External or prospective validation is strongly needed.

In addition to early warning systems, chatbots have also been proposed to support disease management by providing personalized advice to patients and tracking medication compliance. One example is KBot, an early prototype of a chatbot for asthma that utilizes contextual information (such as high pollen triggers) and NLP for dialogue processing. AI can also leverage the capabilities of wearables and mHealth technologies to monitor disease outside clinical contexts. A recent study tested a prototype application for real-time counting of coughs using a deep learning model on ambient sound recorded by mobile phone . This yielded accurate and real-time cough count with a specificity of 92% and a specificity of 98%. Another study applied ML to analyze the sounds of asthma inhalers to predict adequate usage and drug actuations. Recorded sound on mobile devices has also been proposed to monitor lung function in asthmatics. While requiring further validation, these techniques could be used to develop future telehealth solutions including smartphone-based applications, which have the potential to aid decision-making and self-monitoring in asthma.

Fundamental research AI can provide insights into disease classification, pathophysiology, and the underlying biological mechanisms, by clustering large numbers of data points into interpretable patterns.

Heterogeneity and endotype discovery

There is an increasing awareness that allergic diseases (asthma, eczema, rhinitis, food allergy) are umbrella terms of subtypes characterized by distinct disease mechanisms (endotypes). Developments in ML techniques provide new ways to capture the heterogeneity in longitudinal patterns of the development of distinct symptoms of allergic diseases in individual patients. For example, childhood wheezing illness has been extensively investigated using ML approaches to derive more homogenous groups for genetic, mechanistic, and therapeutic studies. Most studies modelled repeated measurements of wheeze through the life-course to derive classes. These different symptom patterns may indicate distinct biological mechanisms, and their discovery may facilitate stratified treatment, but this is not certain (i.e., the classes may not directly translate to endotypes). The derived classes . However, recent studies from the US CREW and UK UNICORN consortia demonstrated that LCA using binary information on wheezing might classify individuals imprecisely, and children with identical wheezing patterns can be assigned to different phenotypes. Recently, a novel data-driven method suggested a potential way to improve assignment to wheeze "phenotypes". Repeated observations of current wheezing were transformed to derive multidimensional indicators of wheezing spells (reflecting duration, temporal sequencing, and the extent of persistence/recurrence). Clustering these indicators resulted in a structure that was much more robust to data imputation, and with a remarkably high agreement between cluster assignment of individual children when using complete or imputed data.

Similarly, over the past five years, longitudinal data on eczema was clustered using data-driven approaches. There were notable differences in the estimated prevalence of each phenotype, and inconsistent associations with the filaggrin (FLG) genotype.

Bayesian machine learning approach has been used to model the development of eczema, wheeze, and rhinitis

from birth to school-age. The developmental profiles were heterogeneous, and the progression of the symptoms fitting the atopic march profile was rare among those with atopic comorbidities . The findings revealed eight latent profiles of symptom development, each with different temporal patterns of their co-manifestation, and distinct genetic associates. Further studies indicated that atopic march, as initially described, occurs rarely, that most 2-disease combinations occur by chance, but that there is a very important cluster of multimorbidity (affecting ˜8% of the population that have a high disease burden).

Numerous studies have applied ML clustering to identify asthma subtypes too. Different endotypes may have a specific response to treatment, making this differentiation potentially clinically significant. Using k-means clustering, researchers identified four distinct clusters of asthma patients in the Severe Asthma Research Program with different responses to corticosteroids (CS). One cluster involves patients, that despite severe baseline airflow limitations, have the lowest response to CS with almost no improvement in lung function, suggesting that this group would benefit from alternative treatment options. The authors also show that the variables that characterize the clusters robustly predict cluster assignment in an independent test set.

A hypothesis-generating unbiased analysis which included data on lower airway inflammation and infection from bronchoalveolar lavage in preschool children with severe wheeze revealed four distinct pathophysiological clusters of approximately equal size: (1) Atopic; (2) Non-atopic, low infection rate; (3) Non-atopic, high infection rate; and (4) Non-atopic, low infection rate, no inhaled corticosteroids (ICS), with marked differences in BAL microbial profiles between the clusters. In a multicenter prospective study, authors used clustering on integrated clinical, virus, and serum proteome data to identify a cluster in children with bronchiolitis with a significantly higher risk of developing asthma by age six. Multi-omics has also been employed in this domain, such as the novel and open-source method Merged Affinity Network Association Clustering (MANAclust), which provides an automated pipeline to integrate clinical and omics data. The authors identified clinically and molecularly distinct asthma clusters that responded differently to treatment, and substantial heterogeneity in healthy controls. In another recent study, researchers used unsupervised clustering on proteomics data of infants hospitalized with bronchiolitis. They identified two distinct clusters with dysregulated pathways and a higher risk for developing asthma. ML approaches have also shown utility for clustering exhaled volatile organic compounds (VOCs) in exhaled breath (breathomics), an exciting non-invasive biomarker for airway disease sensitive to inflammation.

*Pathways and disease mechanisms*Multi-omics and system biology are comprehensive approaches expected to increase insight into the complex biological mechanisms underlying allergic and immunological diseases. The level of detail of such studies can be increased further using single-cell methods, analyzing gene expression profiles, chromatin accessibility, CpG methylation, or the proteome in thousands of cells individually[195,196]. A fully integrated reference atlas has recently been released for the lung, with consensus annotations for 61 cell types based on data from more than 100 healthy tissue donors. Using a trained model of this fully integrated healthy lung cell atlas, the dataset was expanded by projection and transfer learning using scArches to a dataset of more than 2.4 million cells from more than 480 individuals. This illustrates the use of deep learning in biology, to define cell types and states. This extended Lung Cell Atlas allowed direct comparison of cell types across datasets based on consensus labels, leading to the identification of disease-associated cell states common to multiple lung diseases[197,198].

Drug and therapy development and precision medicine AI has the potential to accelerate drug discovery and development throughout the whole pipeline and contribute to precision medicine. Precision medicine promises to enable personalized and more effective treatments based on an individual's genetic variability, environmental exposures, and lifestyle. We here highlight promising examples for treatment response analysis and drug repurposing.

*Treatment response*In a pediatric cohort, asthma control after six months of medication could be accurately predicted using an AdaBoost classification algorithm, outperforming traditional logistic regression.[187] Wu *et al.* (2022) developed a supervised ML model to predict low response to dupilumab in atopic dermatitis patients. The authors identified various indicators of nonresponse, including a high Quan-Charlson Comorbidity Index value, a claim for ibuprofen, or no claims for prednisone medication before dupilumab

initiation. Similar approaches have been pursued to analyze nonresponse to Type 2-directed biologics in asthma patients.

*Drug repurposing*Artificial intelligence has been used extensively in drug repurposing to overcome the immense time and investments required for new drug development. AI has been applied for virtual drug screening, treatment combination optimization, and drug-target interaction predictions. Patrick *et al.* (2019) developed a workflow to model drug-disease relationships using unsupervised text analysis and supervised classification for cutaneous diseases, including atopic dermatitis. They created word embeddings – a dimensionality reduction method that creates a lower-dimension projection of high-dimensional text data – from 20 million abstracts in PubMed. Some of the strongest identified associations were not directly mentioned in any research article, demonstrating how the analysis of large-scale textual data can unveil novel repurpose opportunities. Despite promising results in other medical fields, we identify a research gap in target discovery and clinical trial optimization applied in the allergy and immunology domain. Also, of over 10,000 clinical studies related to allergic diseases, we could only identify five with a fundamental role for artificial intelligence (search ClinicalTrial.gov, performed March 13, 2023).

Future prospects

Successfully translating AI proof of concepts into clinical practice remains pivotal for fully realizing AI's impact.

While practical guidelines and best practices are emerging in medical AI, they are not always adhered to and require frequent reassessment due to the pace at which the AI field is moving forward. When implementing AI, it is strongly recommended to verify available guidelines to ensure applications are reliable and provide meaningful outcomes. We here propose a set of minimal requirements for good practice in AI (Table 1) based on published guidelines of the FDA, literature on best-practice model development in biomedicine, or expert-based checklists for developing and reporting algorithms (e.g., STARD-AI, TRIPOD checklist, and awaited TRIPOD-AI adaptation).

In the allergy and immunology field, research beyond proofs-of-concept is relatively scarce, let alone meaningful clinical applications. We provide an expert outlook on noteworthy AI trends. Firstly, the ever-increasing accessibility, automation, and transferability of*ML tooling* are expected to drive AI adoption further, enabling non-specialized researchers to apply novel techniques. Secondly, we expect an increase in the use of *unstructured data.* Innovations such as AI-based image analysis, NLP, and generative AI are at the forefront of academic efforts in computer science while being underutilized in our field. For clinicians, an AI clinical assistant akin to readily available 'home assistants', could quietly listen in on consults and subsequently support in documentation in EHRs, diagnosis, and therapy suggestions. Clinical solutions that leverage speech recognition are entering the market, aiming to improve the clinical workflow and efficiency, although adoption and showcases of tangible impact are still limited. Thirdly, the emerging trend of multi-modal learning can open new research avenues by integrating multiple data sources and modalities in a singular analytical approach, hereby creating more holistic models and insights.

The largest future impact from AI is expected when current proofs-of-concept are translated successfully to clinical practice. The US and the EU are making steps towards developing AI and algorithm regulations, to facilitate updates and improve privacy, security and transparency.

The developers of algorithms play a role in clinical translation, and clinicians would need to adapt to the integration of AI within healthcare. While most AI systems are designed as a support mechanism rather than a replacement, it will change their work and role. Clinician training in the fundamentals of AI is needed to gain trust in these systems and work with them effectively. One of the common concerns regarding AI is that these systems will replace humans in their installment. While many studies position their analytical solution in a head-to-head comparison with humans, most clinical applications are designed as decision-support tools that strengthen and assist experts in their profession rather than replacing them[41]. Lastly, we

foresee further developments in dynamic learning systems, which continuously evolve based on clinical usage. Such approaches are rare, and FDA-approved tools are generally 'locked', referring to a fixed algorithm state. The FDA is working on an action plan to better assess and support such applications.

In conclusion, the potential of AI to transform clinical medicine is evident, but the steps from a proof of concept to clinical applications are not easily made. Innovations from the field of AI can address many important open questions in allergy; we anticipate that good future utilization of AI (Table 1) will deepen our knowledge of disease mechanisms and contribute to precision medicine in allergy.

References

Tables

Glossary

TERM	DEFINITION
Artificial neural network (ANN)	Technique that is build up of a network of interconnected nodes (neurons) that pr
Convolutional neural network (CNN)	Class of neural networks that is charactericed by convolution filters that slide over
Decision tree	Among the most popular ML algorithms that learns to split data on certain condi
Generative adverserial network (GAN)	Class of deep neural networks for the generation of new data samples. GANs has f
Gradient boosting	Machine learning model type that uses an ensemble of weak prediction models (oft
k-means	Unsupervised clustering method that aims to partition observations into k clusters
LightGBM	Popular ML algorithm of relatively recent origin (2016), similar performance to XG
Natural language processing (NLP)	The discipline in AI involved in the understanding of written and spoken human la
Overfitting	A model that captures the training data too closely, hereby hindering generalizatio
Principal component analysis (PCA)	Dimensionaliry reduction technique that uses linear transformation to map data to
Random forest	Popular ML algorithm that builds an ensemble of decision trees, improving on the
Support vector machine (SVM)	Supervised model that aims to find the optimal hyperplane that best seperates dif
Tabular data	Data that is organized in a table with rows and column
Transfer learning	Technique to improve model learning by leveraging knowledge gained on a related
Unstructured data	Data that has an internal structure but one that is not represented in a row-colum
XGBoost	Popular ML algorithm that uses gradient boosting and builds decision trees iterati

Table 1. Guidelines for good AI and ML usage.

CATEGORY	GUIDING PRINCIPLE
Purpose & relevance	P1. Disclose which clinical problem the model addresses and how it fits in a clinical workflow
	P2. Collect modeling data in a consistent, clinically relevant and generalizable manner that ali
	P3. Benchmark performance to existing clinical standards of care or previous AI studies or pro
Model development	M1. Design a conceptual model with a definition of the predicted outcome and its presumed re
	M2. Safeguard appropriate separation between training, validation, and test datasets
	M3. Ensure proper documentation and execution of model optimization steps
	M4. Determine the evaluation procedure, metrics and rationale up-front, before starting the m
Replicability	R1. Evaluate model performance in a prospective study, randomized trial, or at least an indep
	R2. Perform sensitivity and robustness checks to assess whether the system is impartial to cha
	R3. Disclose data preprocessing and the way in which data quality is assessed and ensured
Explainability	E1. Determine and provide appropriate levels of interpetability, depending on use case and use
	E2. Leverage interpretability toolkits and libraries for black box models
System design & usage	S1. Focus on multi-displinary collaboration during the full AI solution lifecyle

CATEGORY	GUIDING PRINCIPLE
	S2. Invest in the instruction of users on how to interact with the system and predictions
	S3. Set up monitoring processes to track technical and analytical performance
	S4. Set up a feedback flow to facilitate iterative system improvement
Risks & ethics	R1. Define and evaluate the ethical considerations of the system, e.g. algorithmic fairness
	R2. Assess the potential risks involved in the system and outline approach to manage and mit

Caption Table 1. Guidelines for good AI and ML usage. Best practices are coherent with existing frameworks, where the overlap is annotated with Roman numbers as I - CONSORT-AI; II - FUTURE-AI; III - DECIDE-AI; IV - TRIPOD; V - TRIPOD-AI (in development); VI - STARD-AI (in development); VII - MI-CLAIM; VIII - MINIMAR; IX - FDA checklist; X - RISE criteria. Reporting guidelines (I-VIII) are generally broader than just the AI and ML component, with substantial overlap with standard protocols for academic reporting, such as encouragement of code availability and well-described participant characteristics. Our focus is on the unique and new elements when applying AI in a biomedical context. Components are referenced to TRIPOD-AI when the Delphi round 1 consensus was larger than 80% for being essential and desirable to include as reporting.

Main figure captions

Figure 1. Hierarchy of disciplines of artificial intelligence. The discipline of AI is often categorized as part of computer science, although it also builds upon other fields, such as mathematics and cognitive sciences. ML is a subdiscipline of AI, whereas deep learning is a further specialization within ML, generally characterized by large-scale artificial neural networks consisting of many layers, hence the term *deep.* The right part of the figure displays typical terms that one encounters within the (sub)discipline.

Figure 2. A conceptual framework for AI applications in the biomedical domain. The framework is structured by learning strategy, learning goal and data modality. The included studies are selected as illustrations of how AI is used within the medical field and how its applications can be conceptually categorized. They are not necessarily selected based on the inclusion criteria stated in the introduction. References for the shown studies are included in the Supplementary References.

Figure 3. Difference between ordinary least squares (OLS), machine learning (ML), and deep learning (DL) methodology. (a) In OLS, features (or predictors) are modeled manually, and their relationship is assumed linear to the output variable unless specified differently. Interpretation of the model and learned patterns (inference) is straightforward. (b) A similar procedure is followed with ML, but the algorithm can learn more complex patterns from the provided features. Nevertheless, thorough feature engineering by the practitioner is a critical step for delivering a performant model. (c) With DL, especially when applied to unstructured data, feature engineering is an inherent behavior of the interconnected neural network layers. The relationship between input features (tabular data fields, image pixels, text snippets, etc.) to the predicted output is more opaque and harder to interpret.

Figure 4. Workflow of developing a machine learning model to predict disease risk. The best practice in machine learning modeling is using distinct training, validation (or tuning), and test datasets. The modeling steps till testing are generally executed in sequential order, where it is common to iterate multiple times based on validation results that inform model improvements. It is discouraged to assess and improve test performance iteratively, as this can lead to overfitting. The steps preceding model development are excluded, primarily consisting of problem definition and study design, data collection, and preprocessing.

Figure 5. Application domains of artificial intelligence within the allergy field. Domains identified as currently most active and discussed in the main text. Other areas, such as clinical trial optimization, have been excluded due to the limited number of impactful applications.

Figure 1. Hierarchy of disciplines of artificial intelligence

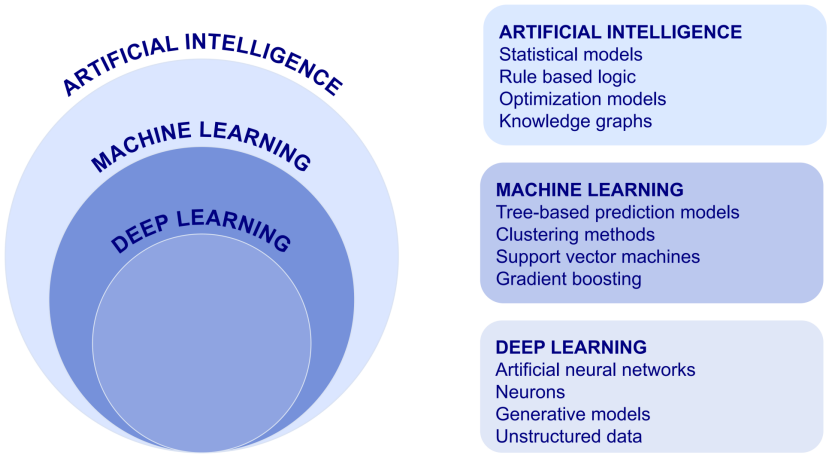

Figure 2. Conceptual framework for AI applications in biomedical domain

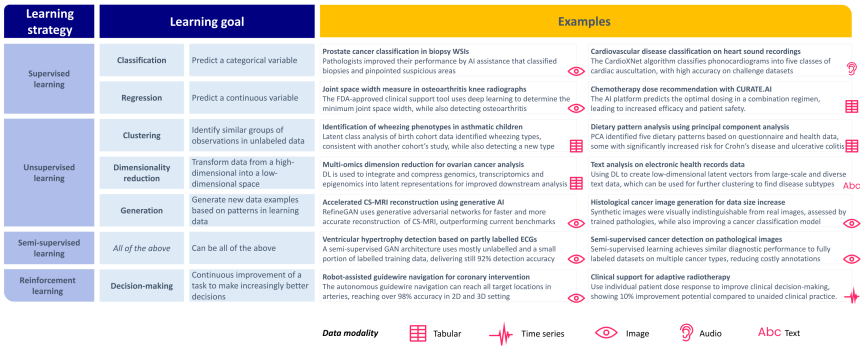

Figure 3. Difference between ordinary least squares (OLS), machine learning (ML), and deep learning (DL) methodology

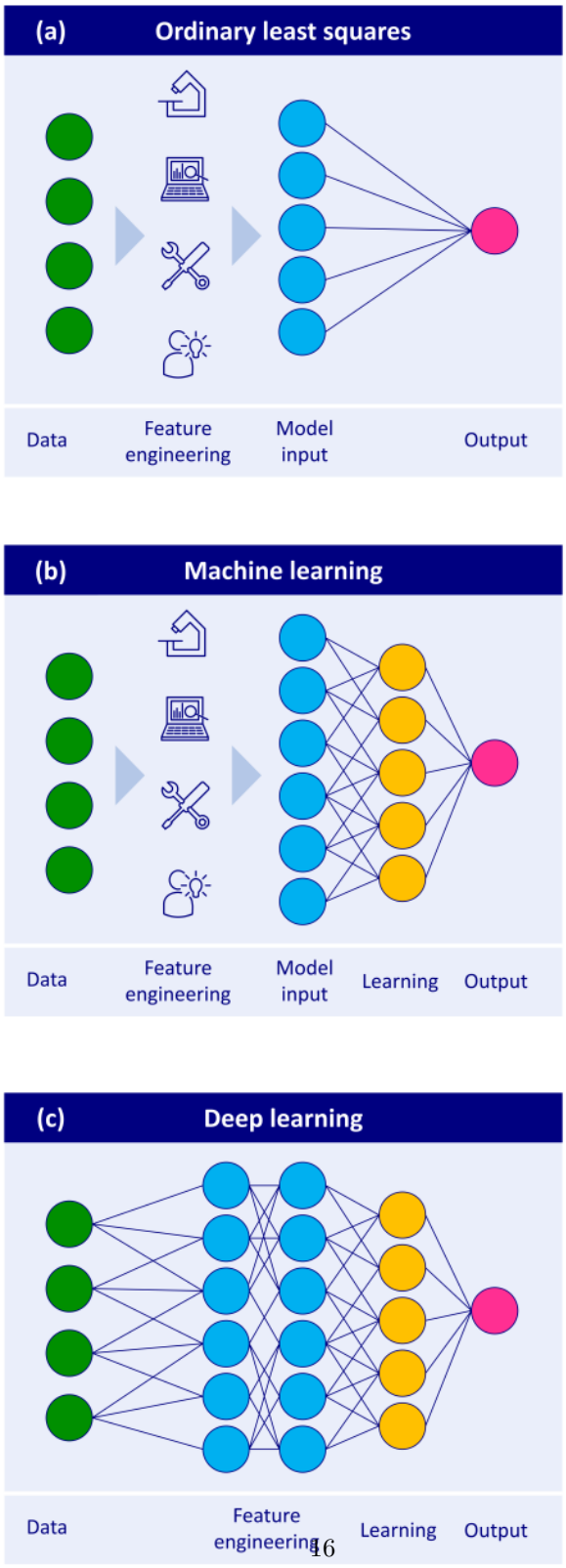

Figure 4. Workflow of developing a machine learning model to predict disease risk

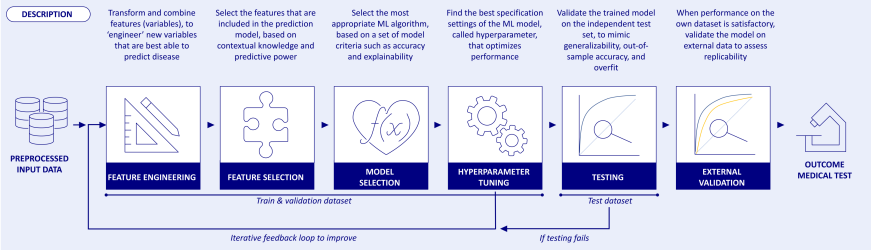

Figure 5. Application domains of artificial intelligence within the allergy field