REVIEW

OPEN

Improving machine-learning development in allergology: bridging the gap between open-access and cohort-based databases

Aleix Arnau-Soler[a,b,c], *Jeremy Corriger*[d,e,f], *Yannick Chantran*[e,g,h] *and Julien Goret*[e,i,j,k]

Purpose of review
The advent of high-throughput data generation and artificial intelligence has transformed allergy research. Open-access database (OAD) and cohort-based database (CBD) provide essential resources for machine learning (ML)-driven algorithms for risk stratification and decision support. It is crucial for allergologists to understand their construction, strengths, and limitations. We review recently published databases with a focus on how these datasets can be combined to enhance research.

Recent findings
OAD, including environmental monitoring resources, omics repositories, and electronic health records, offer scale, diversity, and opportunities for new hypotheses, but are often limited by sparse clinical annotation, heterogeneous data generation, and incomplete linkage to patient-level outcomes. CBD provide well-phenotyped patients, longitudinal follow up, and high-quality clinical and immunological measurements, yet face constraints in sample size, population diversity, and data sharing. Studies integrating OAD breadth with CBD label fidelity report improved predictive performance when paired with disciplined evaluation. Federated learning and portable feature specifications are emerging to enable privacy-preserving collaborations.

Summary
Allergologists play a central role in building ML-ready resources. By ensuring rigorous clinical annotation, standardization of data, transparent methods, and independent validation, they can maximize the utility of OAD and CBD and their combination to accelerate progress toward precision allergy medicine.

Keywords
allergy diagnosis, artificial intelligence, electronic health records, machine learning, open-access databases

INTRODUCTION

Allergology is undergoing a methodological shift driven by the exponential growth of biomedical data, the maturation of artificial intelligence (AI) methods, and the clinical need for predictive tools that enable personalized care [1,2]. Classical hypothesis-driven

[a]Max-Delbrück-Center for Molecular Medicine, [b]Clinic for Pediatric Allergy, Experimental and Clinical Research Center of Max-Delbrück-Center for Molecular Medicine and Charité-Universitätsmedizin Berlin, [c]German Center for Child and Adolescent Health (DZKJ) partner site Berlin, Berlin, Germany, [d]Sorbonne Université, Inserm UMR-S1135, Centre d'Immunologie et des Maladies Infectieuses (CIMI-Paris), Paris, [e]e-health and artificial intelligence working group, Société Française d'Allergologie, Montpellier, [f]Department of Allergology, Mercy Hospital, Metz-Thionville Regional Hospital Center, Metz, [g]Environmental Risk Assessment (HERA) Team, UMR261 MERIT, IRD, Inserm 1344, Université Paris Cité, [h]Department of Biological Immunology, Saint-Antoine Hospital, AP-HP Sorbonne University, Paris, [i]Immunoconcept CNRS UMR 5164, University of Bordeaux, [j]Immunology and Immunogenetic Laboratory and [k]Allergy Unit, University Hospital of de Bordeaux, Bordeaux, France

Correspondence to Julien Goret, PharmD-PhD, Immunology and Immunogenetic Laboratory, Reference Medical Laboratory for Allergic Diseases, Groupe Hospitalier Pellegrin – CHU de Bordeaux, Place Amélie Raba Léon, 33076 Bordeaux, France. Tel +33 5 57 82 19 86, +33 5 56 79 56 45; fax: + +33 5 57 82 21 82; e-mail: julien.goret@chu-bordeaux.fr.

Curr Opin Allergy Clin Immunol 2026, 26:000–000

DOI:10.1097/ACI.0000000000001143

KEY POINTS

- Integrating large datasets offers a transformative opportunity for allergy research, enabling machine learning-driven predictive models with translational potential.
- Open-access databases and cohort-based datasets differ in objectives, construction principles, and translational potential.
- Three modes dominated 2025: truly open datasets supporting component resolved diagnostic initiatives; regulated access to national registries or Electronic Health Record; and algorithmic openness.
- Full open-access and federated datasets hold great promise with careful design, standardization, calibration, audit and rigorous clinical evaluation.

clinical and immunological approaches remain essential [3], and they are increasingly complemented by data-driven paradigms designed to capture the complexity, heterogeneity and evolution of allergic diseases across multiple biological and exposomic layers. Within this context, large-scale data resources and purpose-built databases have become essential infrastructure, shaping the questions that can be asked and how answers are generated and translated into practice. Because databases reflect recruitment and measurement choices, basic bias awareness and clear provenance are required for correct interpretation.

Unlike fields where electronic health records (EHR) and imaging repositories dominate AI applications, allergology depends on a diverse ecosystem of data sources [1]. These include clinical phenotypes, immunological assays, molecular allergen characterization, environmental exposure metrics, multi-omics profiles, and longitudinal outcomes. AI systems, principally machine learning (ML) and deep learning (DL), ingest high-dimensional information to perform prediction or classification [3–7]. Models may learn from labeled examples (supervised learning) or discover patterns in unlabeled data (unsupervised learning). Their performances depend on access to large, well-curated datasets and on the relative stability of relationships among variables over time.

In current practice, these data are aggregated through two conceptually distinct categories of database: open-access databases (OAD) and cohort-based datasets (CBD). Both are increasingly used for ML-driven research, but they differ in objectives, construction principles, and translational potential. OAD are designed primarily to maximize data sharing, reuse and interoperability. In allergology, they typically arise from public health agencies [8], EHR resources [9], drug allergy and hypersensitivity surveillance [10], environmental exposure systems, such as pollen monitoring networks [11,12], air-pollution monitoring [13,14▪], and climate or meteorological re-analyses [15], as well as omic repositories [16] and allergen databases [17–20]. Their scale and accessibility make them attractive for unsupervised ML and exposure–response modelling. However, they often lack detailed phenotyping, standardized diagnostic criteria and robust links to individual patient trajectories.

By contrast, CBD are usually created within clearly defined clinical or research frameworks, often centered on academic hospitals, national networks or disease-specific consortia. These datasets are characterized by expert-validated diagnoses, standardized immunological testing, high-quality clinical data, and longitudinal follow-up, and may include integrated omics layers. In allergology, such cohorts [21–24] are well suited to mechanistic questions, biomarkers discovery, and the development of clinically interpretable predictive models, making them ideal for supervised ML approaches. However, CBD face several limitations, including sample sizes, limited population diversity, high operational costs and regulatory and ethical constraints. A practical extension of CBD is the population-based cohort, designed to represent the general population, usually followed over long periods. These cohorts offer larger sample sizes, which improves statistical power, and more diverse genetic, sociodemographic, and environmental backgrounds, which can enhance model generalizability. Trade-offs include variable depth of phenotyping and differences in linkage to specialized testing.

Across both models, the performance and clinical credibility of ML algorithms depend on data quality, representativeness and conceptual consistency. In this review, we focus on studies published in the last 12 months, synthesizing how OAD and CBD have been used to support clinically relevant ML tasks. We discus opportunities, barriers, and next steps, with the aim of clarifying the complementary roles of OAD and CBD and providing a conceptual framework for their open rigorous and effective use in allergy research.

THE RISE OF MACHINE LEARNING MODELS LEVERAGING OPEN DATA TO PREDICT ALLERGY OUTCOMES

OAD aim to follow in the FAIR principles (findable, accessible, interoperable, reusable) [25]. OAD store data in standardized formats (as much as possible) and ontologies and include metadata describing

experimental conditions, analytical pipelines and, where available, the biological or environmental context. Three modes dominated this year: truly open datasets supporting component resolved diagnostic (CRD) initiatives; regulated access to national registries or EHR; and algorithmic openness. These modes align with the integrative pathway summarized in Fig. 1 and the practical steps listed in Table 1. Representative articles illustrating these three modes are summarized in Table 2.

A standout contribution this year was an open challenge that provided standardized immunological data inputs and a strict held-out evaluation set [26▪▪]. Martinroche *et al.* created a nationwide OAD comprising 4271 patients that combined allergen chip (AC) assay results for specific immunoglobulin E (IgE) with demographic factors and 20 expert-selected clinical variables. The dataset underpinned an international ML competition aimed at developing predictive allergy diagnosis algorithms. However, as noted earlier, the retrospective design and phenotype spectrum are limiting. Respiratory and food allergies were well represented, whereas other conditions such as Hymenoptera venom allergies,

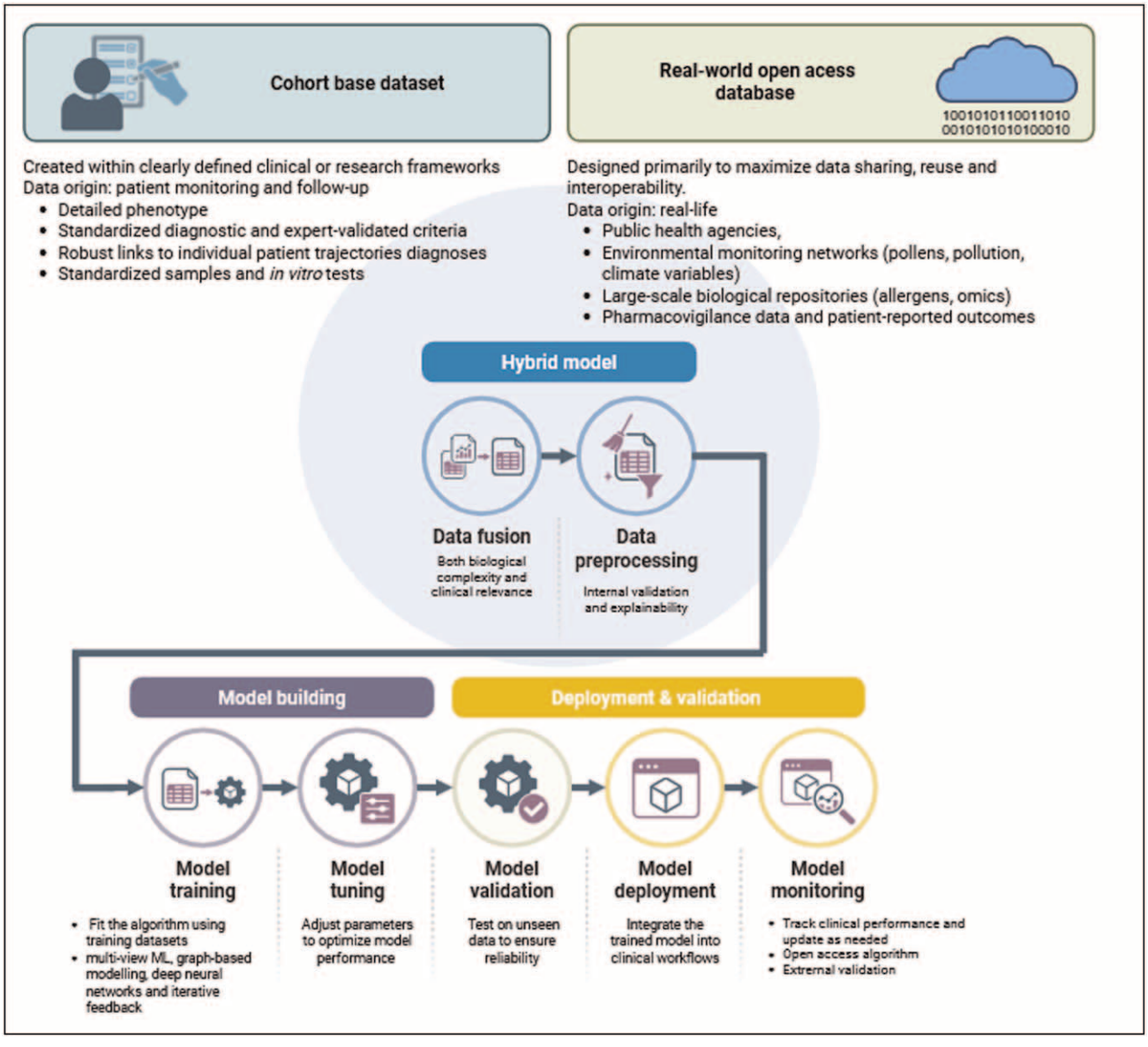

FIGURE 1. Integrating open-access databases with cohort-based resources for allergy prediction. Open-access data streams (environmental monitoring, EHR extracts, omics repositories, allergen knowledge bases) are paired with expert-validated cohort data to create machine-learning (ML) inputs. Hybrid pipelines fuse features, align timelines, and run internal checks before model building. Deployment requires external or temporal validation, calibration, and ongoing monitoring to detect drift and maintain clinical performance. The figure was created with BioRender.com. EHR, electronic health record.

Table 1. Practical recommendations for maximizing the value of combined strategies in allergology

Topic	Description	Clinical and translational relevance
Define clear objectives	Clearly identify whether the primary aim is mechanistic discovery, biomarker identification, predictive modeling, or clinical decision support, as this determines data sources, model design and evaluation plan.	Aligns AI development with clinically meaningful questions and avoids misinterpretation of model outputs.
Prioritize data quality and annotation	Combine high-fidelity cohort data for accurate clinical labels with open-access datasets to increase sample size, diversity, and feature richness; ensure standardized and complete metadata, time stamps and provenance.	Improves model robustness, internal validation, and reproducibility of findings.
Implement robust validation frameworks	Apply cross-validation with time-aware splits, external or temporal validation across independent datasets or sites, and sensitivity analyses; include calibration metrics and decision-curve analysis.	Reduces overfitting and increases confidence in clinical deployment.
Focus on explainability	Use interpretable models or feature-attribution methods to link predictions with biological and clinical reasoning.	Facilitates clinician trust, regulatory acceptance, and clinical adoption.
Encourage collaborative networks	Promote multicenter collaborations, federated learning and portable feature specifications under appropriate governance and harmonization, to increase population diversity while respecting ethical, legal, and privacy constraints.	Enhances external validity and supports scalable, real-world implementation.
Leverage hybrid integration for translation	Integrate open-access data (scale), cohort-based clinical data (precision), and omics (mechanistic insight) through data fusion and transfer learning strategies within unified AI frameworks linked to decision support.	Enables patient stratification, disease trajectory prediction, and personalized therapeutic strategies, advancing precision allergy medicine.

oral syndromes and mastocytosis were poorly represented. Therefore, it lies in hypothesis generation and clustering to identify a minimal informative set of specific IgE components for phenotyping and risk stratification, similar to prior work in asthma endotyping [27].

Structured exposure inputs aligned to temporally precise outcomes has been essential for short-horizon forecasts that generalize across regions and seasons [28]. Sofiev *et al.* recently published a meteorological reanalysis based on ERA5 data and used the System for Integrated Modeling of Atmospheric Composition (SILAM), to predict flowering periods and simulate Europe-wide pollen dispersion patterns [29]. The authors integrated open data from the European Aeroallergen Network (EAN) and designed outputs explicitly as inputs for downstream ML studies. These points demonstrate the value of national open datasets such as those recently made available [30]. In this setting, ML-based pollen assessment complement physical sampling by extending spatial coverage and improving forecasting capabilities [31].

Omics made measured progress. Polygenic risk scores for asthma and atopic disease improved via multi-ancestry training, and initial multi-omic combinations refined endotype definitions [32▪,33–36,37▪]. Translational value strengthened when omics were combined with CRD, clinical variables, and exposures rather than treated as stand-alone predictors. Good practice (pre-registered analyses, hold-out cohorts, and code availability) became more visible, although the field still needs clear decision thresholds that specify when omics-augmented scores should change management.

Digital health data (symptom trackers, mobile apps, ecological momentary assessment) are relevant to improving precision medicine. Tsang *et al.* [38] carried out a 2-phase observational study to monitor asthma using three smart-monitoring devices (peak-flow meter/inhaler and smartwatch), and daily symptom questionnaires on the open-source Mobistudy platform [39]. Combined with localized weather, pollen, and air-quality reports, they collected a rich longitudinal dataset to explore the feasibility of passive monitoring and asthma attack prediction [40] within the next days. The database was built on three previous mobile-health studies in asthma management: the Asthma Mobile Health Study (AMHS) [41,42], myAirCoach [43] and Biomedical Real-Time Health Evaluation (BREATHE) [44].

This year consolidated gradient boosting and regularized Generalized Linear Models (GLMs) as robust baselines for case-finding and risk stratification in asthma. Where national and open EHR

Table 2. Representative 2025 studies illustrating data backbones and openness modes for ML-ready allergy research. Each row summarizes one original study and reports what was done (backbone, design, validation, headline metrics) and why it matters for prediction or decision support

Author (ref)	Mode	What the study did	Why it matters
Martinroche *et al.* [26▪▪]	Open CRD dataset/ challenge	Built a nationwide, open database that links allergen-chip IgE results with 20 clinician-defined variables and 5 demographic factors across 11 French university hospitals, 2014–2023. Final dataset: *n* = 4271 patients, ~700 000 sIgE measurements, chips up to ~295 allergens (ISAC 112i, ISAC E112i, ALEX v2). Labels included confirmed allergy and severity.	Provides the first open national AC+clinical resource with physician-confirmed outcomes, enabling fair, apples-to-apples benchmarking, feature selection, endotype clustering, and calibration studies across respiratory and food allergy. The database is publicly available under an open license, and the paper explicitly calls for external validation and longitudinal extensions to reduce bias and improve generalizability.
Frau *et al.* [64▪]	Regulated-access EHR/registry	Built a framework that maps routine EHR phenotypes in atopic dermatitis to a biomedical KG to connect clinical features with molecular entities and pathways—supporting the discovery of clinically relevant molecular endotypes from real-world records. Proof-of-concept demonstrates how KG-enhanced features can bridge from patient trajectories to mechanistic hypotheses.	Provides a practical path to "algorithmic openness" by enriching EHR signals with curated biological knowledge, improving interpretability and hypothesis generation for endotyping and potential target prioritization in AD. It exemplifies how open algorithms/knowledge resources can add value even when raw patient data cannot be shared.
Bağcı *et al.* [46▪▪]	Regulated-access EHR/registry (single system)	Extracted data on 31 795 asthma patients from a health-system EHR; 1112 met inclusion for analysis. Used PCA to derive 3 principal components (lung function; blood inflammatory markers; systemic corticosteroid receipt) and a Gaussian mixture model to identify 5 subject clusters with distinct clinical/inflammatory profiles.	Shows how time-aware, lab-augmented EHR features can uncover clinically coherent severe asthma phenotypes and support actionable prediction of severity—moving beyond simple code-based case-finding. The combination of unsupervised clustering with a supervised severity model provides a path from discovery to triage, with performance reported (accuracy/precision) and phenotypes interpretable via PCs (physiology, inflammation, steroid exposure).
Pattarakiatjaroen *et al.* [57▪]	Cohort-based dataset (CBD)	Retrospective pediatric OFC cohort (n=179, Thailand; 2014–2022). Built a logistic-regression score using routine variables (female sex, history of anaphylaxis, positive SPT). Internal validation via bootstrapping; ROC = 0.71. Risk groups: low (0–1 points) vs. high (2–3 points) with 6.7% vs. 29.5% reaction rates during OFC, respectively. Reported calibration and classification metrics.	Provides a simple, clinic-ready triage tool for OFC risk using data available at intake; transparent methods and calibration reporting make it easy to replicate and refine externally, and to embed in scheduling/ monitoring workflows where OFC resources and safety planning matter.
Xie *et al.* [45▪▪]	Regulated-access EHR/registry	Large integrated health-system study (Kaiser Permanente Southern California). Built a hybrid NLP pipeline (dictionary/regex + BERT) to identify four asthma-related symptoms (cough, dyspnea, wheeze, chest tightness) from >11 million clinical notes (2013–2018 and 2021–2022). Used double-annotated reference sets (≈9600 notes) with adjudication. The hybrid system achieved PPV ≈96–97%, sensitivity ≈94–99%, F1 > 0.	Shows that hybrid, auditable NLP can convert unstructured EHR text into ML-ready symptom trajectories with high precision/recall—an essential building block for asthma case-finding, trajectory modeling, and short-horizon risk prediction. Strong methods (large scale, double annotation, temporal hold-out, transparent rule components) make it a reusable blueprint for allergy phenotypes in trusted research environments.
Ding *et al.* [52▪]	Regulated-access EHR/registry (multicenter)	Retrospective, multicenter outpatient cohort across three tertiary hospitals (China) using routine clinical + laboratory data to differentiate eczema vs. psoriasis. Trained 8 ML models; XGBoost selected. Performance: AUC 0.891 (train), 0.830 (internal test), 0.812 (external test). Used SHAP for feature attribution (top features: dNLR, neutrophil count, SIRI, RDW, eosinophils).	Shows that non-imaging, routine labs can support an interpretable, deployable classifier for a common dermatology dilemma, with external validation and a usable interface. Provides a portable feature set and a pragmatic pathway (OCR + EHR integration) for wider outpatient decision support beyond image-heavy pipelines.
Turcatel *et al.* [78▪]	Regulated-access EHR/registry	Retrospective EHR study using Optum Panther (USA), adults with physician-diagnosed asthma (2016–2023). Cohort *n* = 1 331 934; 16 279 (1.2%) had ≥1 exacerbation. Baseline windows of 6 months used to predict exacerbation risk in the following 6 months. Compared XGBoost, LSTM, and Transformer models; train/test on independent datasets. Best performance: XGBoost AUROC 0.964, PR-AUC 0.647, precision 0.729, recall 0.529, F1 0.613, accuracy 0.	Demonstrates real-world feasibility of short-horizon exacerbation prediction at national scale with transparent feature attributions. Highlights generalization limits (single commercial EHR, no causal inference) and underlines the need for external/ temporal validation and calibration reporting before clinical deployment.
Liljendahl *et al.* [49▪]	Regulated-access EHR/registry	Retrospective registry study linking multiple Danish national registers (1994–2021). Adults who redeemed dermatology prescriptions were grouped as "Known AD" (ICD-10 L20 in hospital records), "Other skin disease," or "Uncertain AD status." Features: health-service contacts and prescriptions from the prior 2 years; ~8990 candidate variables reduced via variance filtering, univariate screening, and recursive feature elimination.	Provides a pragmatic, transportable workflow to identify mild-to-moderate AD cases that are invisible to hospital-only coding, using routine primary-care and prescription signals. Reports both discrimination and calibration, and compares outputs against a validated pediatric algorithm as a face-validity check—useful for case-finding, burden estimation, and cohort assembly in allergy research.

Table 2 (*Continued*)

Author (ref)	Mode	What the study did	Why it matters
Mora *et al.* [81▪]	Regulated-access EHR/registry	Retrospective, matched case–control study using anonymized administrative health records from Catalonia, Spain. Built a stacked ensemble (logistic regression, decision tree, random forest, XGBoost) with chi-squared feature selection to predict a first recorded diagnosis of anaphylaxis in older adults. Matched dataset size: *n* = 8398 (anaphylaxis *n* = 4199; matched non-anaphylaxis *n* = 4199); broader source population also described.	Demonstrates population-scale risk modeling for first-time anaphylaxis in the elderly using routinely collected data, with transparent ensemble methods and explainability. Results support triage/risk-stratification pathways and highlight interpretable predictors (e.g., healthcare-utilization patterns, socioeconomic proxy, allergy-related codes) that could inform prevention and earlier recognition.
Wong *et al.* [84▪]	Regulated-access EHR/registry (multicenter)	Built and validated an NLP pipeline to measure clinician adherence to NIAID 2017 peanut-prevention guidance using EHR clinical notes and patient instructions from the iREACH cluster-randomized trial (4- and 6-month well-child visits; 30 practices across 3 networks in Illinois). Three-phase development (exploratory → training → validation) with chart-review gold standards. Performance for documenting peanut-introduction recommendations: precision 0.	Provides a scalable, pragmatic method to quantify guideline adherence from unstructured EHR text across health systems, a prerequisite for auditing population-level prevention efforts and embedding decision support. Results also flag documentation gaps (eczema severity), guiding pipeline refinement and improving the reliability of downstream ML models that depend on accurate risk stratification.
Tawfik *et al.* [37▪]	Algorithmic openness (code/model cards/portable features)	Prospective/experimental model paper building a respiratory-sound asthma classifier. Data backbone: two public/benchmark respiratory sound datasets—the Asthma Detection Dataset (ADD) and ICBHI 2017. Proposed E-RespiNet, a triple-stream CNN with LLM-ELECTRA–guided feature extraction and feature fusion. Reported accuracy improvements of 4.82% (ADD) and 0.52% (ICBHI) over prior methods; on cross-institutional validation, accuracy 75.	Demonstrates an audio-signal pipeline for asthma classification using shared datasets and a portable architecture, enabling reproducibility and comparison on common benchmarks. Cross-institution testing highlights domain-shift issues that matter for real-world deployment (performance drop across sites), directly relevant to generalization and monitoring in clinical decision support.
Hu *et al.* [32▪]	Regulated-access EHR/registry (multicenter)	Prospective model-development/validation across multiple hospitals in China. Trained on preschool children with cough (*n* = 14 709) using routine clinical variables plus IOS and FeNO; primary outcome was early asthma diagnosis. Compared models using single modalities (clinical-only, IOS-only, FeNO-only) vs. combined. In an external validation cohort (*n* = 4146), the combined model achieved AUC = 0.90, sensitivity = 0.82, specificity = 0.	Shows that adding IOS and FeNO to basic clinical features improves early asthma prediction in preschoolers and can generalize across sites. Reports calibration and net-benefit analyses, supporting movement from algorithmic performance to decision support in pediatric pathways.
Chushak *et al.* [14▪]	Algorithmic openness (code/model cards/portable features)	Curated two datasets from public sources and literature for respiratory irritation (final set n = 1241; 624 irritants, 627 non-irritants) and respiratory sensitization (final set n = 419; 208 sensitizers, 211 non-sensitizers). Built and compared multiple classifiers (including ASNN, Random Forest, XGBoost, neural networks) using structural-alert fingerprints in OCHEM/SARpy; created consensus models. Reported AUC = 0.931 for irritation and AUC = 0.	Provides portable, feature-based models that can be applied to screen chemicals for respiratory irritation/sensitization before human exposure. Uses publicly sourced data and shareable structural alerts, making the approach transferable to regulatory or occupational settings and complementary to clinical allergy surveillance.
Chen *et al.* [47▪]	Regulated-access EHR/registry	Retrospective cohort, Kaiser Permanente Southern California (USA). Identified 198 873 adults (18–85 y) with mild asthma (2013–2018). Used a validated hybrid NLP algorithm (rule-based + transformer) to extract cough, wheeze, dyspnea, chest tightness from clinical notes in the 12 months before the index visit (t0); prior validation of the NLP tool reported sensitivity > 93% and PPV > 96%.	Demonstrates that routine-note NLP can surface symptom burden at scale and link it to prospective exacerbation risk in a large, real-world system. Supports integration of symptom extraction into EHR-based prediction/triage pipelines for mild asthma, extending beyond coded data while working within a trusted research environment.
Bashir *et al.* [53▪]	Cohort-based dataset (CBD)	Population-based cohort analysis using data from the West Sweden Asthma Study: 3101 adults clinically investigated, of whom 1895 with ever-asthma were included. Forty-four clinical, biological and epidemiological variables were analyzed; missing data were imputed with random forests. Deep Embedded Clustering identified four phenotypes; the number of clusters was selected using NbClust, M3C, and ARI indices in combination with clinical judgment.	Provides a transparent, data-driven phenotyping of adult asthma in a representative population sample, aligning clusters with recognizable clinical patterns (e.g., early-onset atopic, adult-onset high-T2 inflammation). The approach illustrates how unsupervised ML on regulated cohort data can surface clinically interpretable subgroups that may inform future risk-stratification or targeted management pathways; highlights current gaps.

Notes: *Modes*. Open CRD dataset/challenge = truly open component-resolved diagnostics resources or open ML challenges with public inputs and held-out scoring; regulated-access EHR/registry (trusted environment) = national or health-system records and linked registries accessed under approvals; cohort-based dataset (CBD) = prospectively or retrospectively assembled, expert-phenotyped clinical cohorts (often with standardized immunology/omics), shared under consent-governed access; algorithmic openness (code/model cards/portable features) = public code/model cards/feature specs that enable replication or transfer without moving raw patient data. *Selection*. We ran a structured PubMed search limited to the most recent 12 months and original research in English. The operational query combined (i) allergic-disease terms, (ii) AI/ML terms, and (iii) backbone terms (e.g., nationwide, registry, EHR). From 169 records, we extracted a standardized template (disease area, backbone, task, validation, calibration, sample size/diversity, transparency, and a brief clinical/methodological note) and applied a weighted rubric prioritizing external/temporal validation and calibration, dataset scale/openness, clinical actionability, and transparency. We retained 15 exemplars to balance modes, backbones, and clinical coverage.

meaningful signal, Natural language processing improved capture of symptom dynamics and treatment intent in asthma [45▪▪,46▪▪,47▪,48], atopic dermatitis [49▪] and adverse events [50]. Two outstanding articles by Bagci *et al.* [46▪▪] and Xie *et al.* [45▪▪] addressed generalization with external or temporal validation. These time-aware approaches predicted exacerbations, emergency visits, or hospitalization. Gains were most evident when recent medication changes, rescue use, and short-term environmental fluctuations were modeled explicitly rather than averaged away. These most convincing studies went beyond internal cross-validation, validating across hospitals or calendar time, thus assessing the impact of change in clinical coding practice, prevalence, and seasonal baselines on model performance. Crucially, the authors reported calibration (slope, calibration-in-the-large, Brier) and decision-curve analysis, turning statistical discrimination into clinically actionable net benefit at decision thresholds relevant for treatment step-up or targeted follow-up. Operationally, these steps are organized in a validation-and-monitoring pipeline summarized in Fig. 2.

For diagnosing drug allergy, particularly to penicillin, ML-based decision algorithms have been developed using large retrospective databases to facilitate treatment and reduce inappropriate alternative antibiotic use. Ghiordanescu *et al.* assessed the safety and efficacy of four such algorithms in 1884 patients referred for allergy evaluation, using data from the Drug Allergy and Hypersensitivity Database [10]. The work resulted in the development of a fifth algorithm indicating promise for triage.

PROMOTING THE OPEN ACCESS USE OF CBD FOR MACHINE LEARNING ALGORITHM RESEARCH

CBD adopt a fundamentally different approach to data collection than open-access resources. CBD are particularly well suited for ML – especially DL – models aimed at early phenotype discovery [32▪,51,52▪,53▪,54], endotype stratification [55,56], risk assessment [57▪] and treatment-response prediction [58]. To unlock their full value, we encourage models of access that make de-identified CBD data and portable feature specifications available to qualified researchers while respecting consent and governance constraints. Where direct sharing is not feasible, privacy-preserving options (e.g., federated learning) allow multicenter validation without moving data. From an historical perspective, between 1990 and 2010 more than 100 population-based birth cohorts focused on asthma and allergy were initiated worldwide. After 2000, many of these scattered efforts were at least partly coordinated under transnational initiatives, notably within the EU Framework Programmes FP6-FP7, including the Global Allergy and Asthma European Network (GA2LEN) and the Mechanisms of the Development of ALLergy (MeDALL) project [59]. These programs were key milestones that established a federative, collaborative model based on common goals and harmonization. However, no large-scale population-based or patient-based cohort in allergy is yet fully open-access; most operate under controlled-access policies that limit broad reuse.

Population-based longitudinal cohorts are particularly valuable for studying natural history [60], risk factors [61], and early pathogenesis [62] of allergic disease, and for integrating causal frameworks useful for population-level risk prediction, early detection, and public-health interventions. A major opportunity going forward would be to align these cohorts with open-science practices, under appropriate ethical and legal frameworks.

FUTURE DIRECTIONS: BRIDGING THE GAP BETWEEN OPEN-ACCESS DATABASES AND COHORT-BASED RESOURCES

Although OAD and CBD are increasingly used for ML-focused research, their construction and intended use differ. However, their strengths can be combined [63,64▪,65]. OAD scale and diversity make them particularly well-suited to exploratory ML approaches, including unsupervised learning and DL, and hypothesis generation [66,67]. OAD should be viewed as powerful tools for exploration and hypothesis generation. This potential should be promoted because of its central role in precision allergology. The literature also offers practical guidance for clinicians on how to conduct and use such resources [68]. This integrative perspective is summarized in Fig. 1.

Although not easily accessible in open access, the strength of CBD lies in the precision and reliability of their labels. For ML applications, this translates into the ability to develop supervised models that can predict clinically meaningful outcomes, identify disease endotypes, and discover novel biomarkers. Moreover, CBD integrate longitudinal biomarkers and symptom diaries that allow researchers to model the temporal dynamics of allergic inflammation and to evaluate treatment effects. By contrast, models derived exclusively from open data may demonstrate impressive technical performance while remaining disconnected from clinically meaningful endpoints. Despite their strengths, CBD face several intrinsic limitations that constrain their scalability and generalizability [69,70]. First, sample sizes are often

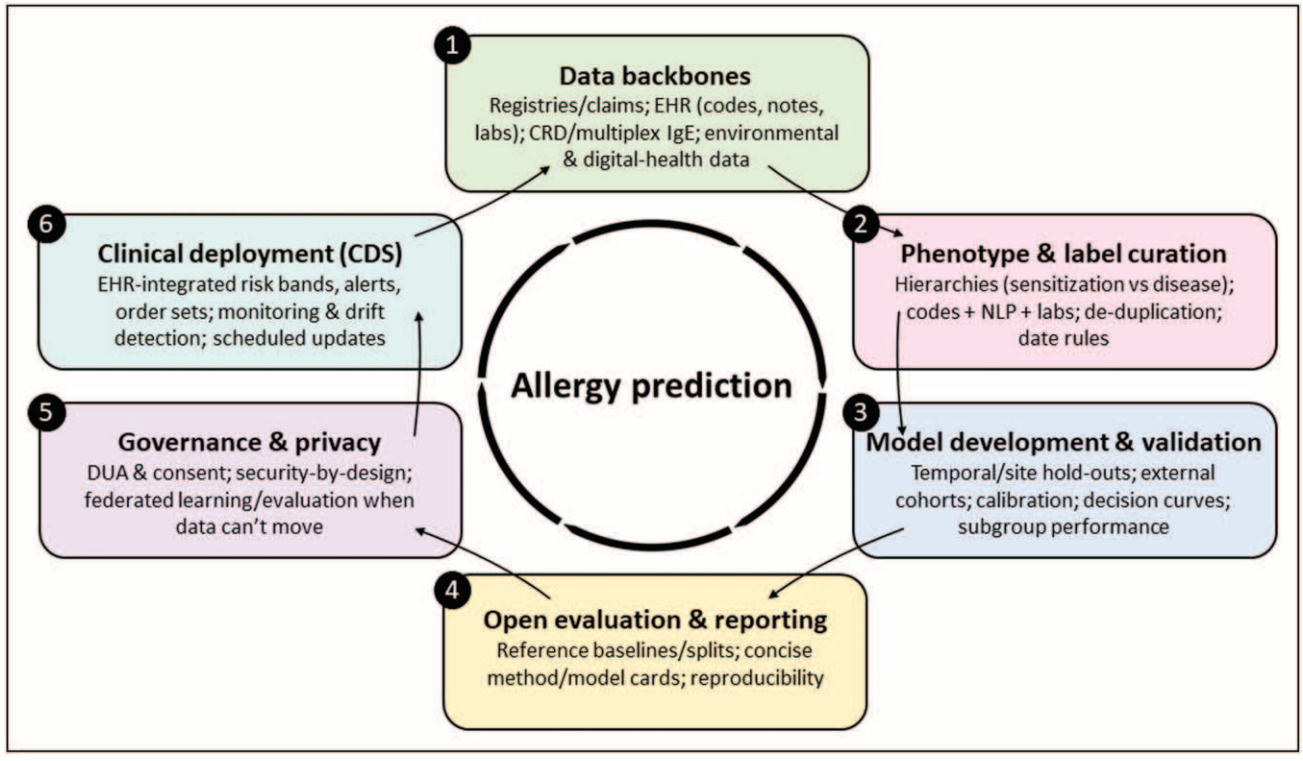

FIGURE 2. Validation and monitoring pipeline from open data to deployable prediction. Circular workflow highlighting leakage-aware cohort construction, clear task definition, transparent baselines, external and temporal validation with subgroup fairness checks, calibration and decision-curve analysis, integration into clinical workflows and post-deployment monitoring with scheduled recalibration. *Abbreviations:* CDS, clinical decision support; EHR, electronic health record; CRD, component-resolved diagnostics; NLP, natural language processing; DUA, data-use agreement.

limited, particularly for rare allergic phenotypes or complex multimorbid conditions [71]. They may be insufficient for training high-dimensional ML models or for capturing the full spectrum of disease variability in diverse populations. Second, population diversity is frequently restricted. Many cohorts are single-center or regionally confined, reflecting the demographic characteristics of the participating hospital or research network. Population cohorts can help to mitigate these constraints. Third, CBD are resource-intensive to maintain.

Currently, too little research data, models and derived resources are freely accessible due to concerns about data misuse, preservation of intellectual property and lack of appropriate acknowledgement for data creators [72,73]. Such open publications and data descriptors should be encouraged, properly valued, and credited in the same way as original research articles. Federated or multicenter cohort initiatives are emerging as solutions to support cross-site model development and validation, although important logistical, regulatory and standardization challenges remain [74–76]. Beyond cohorts, large EHR-based infrastructures developed at local or national levels provide another backbone for large-scale ML tools in asthma [46▪▪,51,77,78▪,79], atopic dermatitis [9] or drug-reaction diagnosis [80,81▪,82,83] and clinician adherence monitoring [84▪,85]. In practice, restricted access to CBD contrasts with the relatively unrestricted nature of OAD and can impede rapid external validation or cross-cohort comparisons.

The scientific validity of ML-based tools depends critically on external validation [86], as recommended [87] and already demonstrated for prediction algorithms in hospital settings [88] and for asthma and eczema [33,89]. Models developed using restricted datasets may exhibit hidden biases, limited generalizability and reduced clinical robustness when applied to new populations [32▪,90]. Broad access to data and models enables peer-driven evaluation, reproducibility of findings and systematic assessment of performance across independent clinical settings, disease phenotypes and demographic contexts [91]. This external validation is needed both to confirm model accuracy and to identify failure modes [92]. Another route to validation is experimental corroboration, as shown by *in vitro* confirmation after a ML prediction of food allergen epitopes [93].

Although OAD and CBD each offer advantages, neither alone is sufficient to realize the full potential

of ML in allergology. Integrative strategies can combine these complementary resources into heterogeneous multimodal data that can be exploited to build unified multimodal models [94]. Several strategies can be implemented, including data fusion (feature-level and model-level), multimodal modeling approaches (multi-view learning, graph-based models, deep neural networks) and transfer learning [95,96]. In transfer learning, high-dimensional open-access datasets (e.g., proteomics) can be used to pre-train models, whose informative features are then applied for supervised training in cohorts. This approach can mitigate the dimensionality – sample size challenge commonly encountered in omics research and may improve predictive performance while preserving interpretability, when appropriate modeling is applied. Hybrid approaches require multi-level validation: internal consistency within cohorts, generalizability across independent datasets and clinical relevance. These strategies may help limit the need for large-scale cohort expansions but require transparency [91], explainability and adequate computational resources. Crucially, their success hinges on close collaboration among allergists, immunologists and data scientists to ensure that models are clinically grounded and robust.

The next months should include the following five key actions: standardize reporting: every study should present at least one external or temporal validation and report calibration metrics; expand open benchmarks: add EHR case-finding tasks, environment-to-symptom forecasting with fixed inputs and public scoring, and reference tasks for multimodal fusion; make federation widely usable: develop standard methodological frameworks for federated or privacy-preserving analytics, so networks can learn across borders while keeping data protected; audit equity by design: pre-specify subgroups, define acceptable performance gaps and plan remediation (recalibration, reweighting and domain adaptation), and report the results; invest in prospective evaluation: embed models in clinical pathways or public-facing tools and measure outcomes that matter to patients and clinicians.

CONCLUSION

Integrating large datasets offers a transformative opportunity for allergy research, enabling ML-driven predictive models with translational potential. Open-access and federated datasets hold great promise, but without careful design, standardization, calibration and rigorous clinical evaluation there is a risk of producing findings that are difficult to translate in routine [97]. Allergologists play a central role in shaping the future of data-driven allergy research by ensuring rigorous methods.

Acknowledgements

None.

Authors contributions: J.G. coordinated the writing of the manuscript. A.A.S., J.C., Y.C. and J.G. wrote the sections of the article. A.A.S. and J.G. formatted the article and edited sections. All the authors proofread the article.

Financial support and sponsorship

None.

Conflicts of interest

J.G. reports speaker and travel support in the past 5 years from ALK, Stallergenes-Greer, Thermo Fisher Scientific, Menarini outside the submitted work. J.C. reports speaker and travel support in the past 5 years from ALK, Stallergenes-Greer, Thermo Fisher Scientific, Menarini outside the submitted work. The authors have no COI relevant to this work.

The Allergen Chip Challenge project mentioned in the article as an open access database promoting open science was conducted by J.G. and in accordance with the recommendations and arguments put forward in this review. J.C. proofread and corrected this article. This is a self-citation[*26*▪▪].

REFERENCES AND RECOMMENDED READING

Papers of particular interest, published within the annual period of review, have been highlighted as:

- ▪ of special interest
- ▪▪ of outstanding interest

1. Proper SP, Azouz NP, Mersha TB. Achieving precision medicine in allergic disease: progress and challenges. Front Immunol 2021; 12:720746.
2. Khoury P, Srinivasan R, Kakumanu S, *et al.* A framework for augmented intelligence in allergy and immunology practice and research—a work group report of the AAAAI Health Informatics, Technology, and Education Committee. J Allergy Clin Immunol Pract 2022; 10:1178–88.
3. Dramburg S, Hilger C, Santos AF, *et al.* EAACI Molecular Allergology User's Guide 2.0. Pediatr Allergy Immunol 2023; 34 (Suppl 28):e13854.
4. Choi RY, Coyner AS, Kalpathy-Cramer J, *et al.* Introduction to machine learning, neural networks, and deep learning. Transl Vis Sci Technol 2020; 9:14.
5. Apprentissage automatique [Internet]. [cité 3 déc 2025]. https://www.cnil.fr/fr/definition/apprentissage-automatique
6. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015; 521:436–444.
7. Apprentissage profond (deep learning) [Internet]. [cité 3 déc 2025]. https://www.cnil.fr/fr/definition/apprentissage-profond-deep-learning
8. European Health Data & Evidence Network (EHDEN) [Internet]. https://www.ehden.eu/
9. Masison J, Lehmann HP, Wan J. Utilization of computable phenotypes in electronic health record research: a review and case study in atopic dermatitis. J Invest Dermatol 2025; 145:1008–1016.
10. Ghiordanescu IM, Ciocănea-Teodorescu I, Molinari N, *et al.* Comparative performance of 4 penicillin-allergy prediction strategies in a large cohort. J Allergy Clin Immunol Pract 2024; 12:2985–2993.
11. European Aeroallergen Network Pollen Database [Internet]. https://ean.polleninfo.eu/Ean/
12. Zentrum Allergie und Umwelt [Internet]. https://www.zaum-online.de/pollen
13. Air Quality Annual Statistics Viewer [Internet]. https://www.eea.europa.eu/en/analysis/maps-and-charts/air-quality-statistics-dashboards

14. Chushak Y, Keebaugh A, Clewell RA. Prediction of respiratory irritation and
■ respiratory sensitization of chemicals using structural alerts and machine learning modeling. Toxics 2025; 13:243.
15. Climate Data Store [Internet]. https://cds.climate.copernicus.eu
16. Fukushima-Nomura A, Kawasaki H, Amagai M. Integrative omics redefining allergy mechanisms and precision medicine. Allergol Int 2025; 74:514–524.
17. Radauer C, Nandy A, Ferreira F, *et al.* Update of the WHO /IUIS alergen nomenclature database based on analysis of allergen sequences. Allergy 2014; 69:413–9.
18. Mari A, Rasi C, Palazzo P, Scala E. Allergen databases: current status and perspectives. Curr Allergy Asthma Rep sept 2009; 9:376–83.
19. Hu X, Li J, Liu T. Alg-MFDL: a multi-feature deep learning framework for allergenic proteins prediction. Anal Biochem 2025; 697:115701.
20. Liu J, Negi SS, Yang C, *et al.* AllergenAI: a deep learning model predicting allergenicity based on protein sequence. BMC Bioinformatics 2025; 26:279.
21. Kauffmann F, Dizier MH. EGEA (Epidemiological study on the Genetics and Environment of Asthma, bronchial hyperresponsiveness and atopy) – design issues. EGEA Co-operative Group. Clin Exp Allergy 1995; 25 (Suppl 2): 19–22.
22. Burney PG, Luczynska C, Chinn S, Jarvis D. The European Community Respiratory Health Survey. Eur Respir J 1994; 7:954–60.
23. Bergmann RL, Bergmann KE, Lau-Schadensdorf S, *et al.* Atopic diseases in infancy. The German multicenter atopy study (MAS-90). Pediatr Allergy Immunol 1994; 5 (Suppl):19–25.
24. Pretolani M, Soussan D, Poirier I, *et al.* COBRA Study Group. Clinical and biological characteristics of the French COBRA cohort of adult subjects with asthma. Eur Respir J 2017; 50:1700019.
25. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016; 3:160018.
26. Martinroche G, Guemari A, Apoil PA, *et al.* Allergen chip challenge: a nation-
■■ wide open database supporting allergy prediction algorithms. J Allergy Clin Immunol 2025; 157:45–55.
27. 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:e1002691.
28. Chen IC, Chen YM, Chen YW, *et al.* Association between high polygenic risk scores and long-term exposure to air pollution in asthma development: a hospital-based case-control study. Environ Health Glob Access Sci Source 2025; 24:49.
29. Sofiev M, Palamarchuk J, Kouznetsov R, *et al.* European pollen reanalysis 2022, for alder, birch, and olive. Sci Data 2024; 11:1082.
30. Données historiques de surveillance des pollens et des moisissures [Internet]. [cité 16 déc 2025]. https://www.data.gouv.fr/datasets/donnees-historiques-de-surveillance-des-pollens-et-des-moisissures/
31. Farooq Q, Oteros J, Galán C. Advancing in the pollen frontier: a comprehensive evaluation and meta-analysis of automatic pollen monitoring systems. Aerobiologia 2025; 41:527–546.
32. Hu Y, Hu Y, Dong M, *et al.* Integrating clinical phenotypes, impulse oscillometry
■ and fractional exhaled nitric oxide: a robust machine learning model for early asthma prediction in preschool children. Respir Med 2025; 247:108290.
33. Owora AH, Jiang B, Shah Y, *et al.* External validation and update of the pediatric asthma risk score as a passive digital marker for childhood asthma using integrated electronic health records. EClinicalMedicine 2025; 84: 103254.
34. Shen K, Lin J. Unraveling the molecular landscape of neutrophil extracellular traps in severe asthma: identification of biomarkers and molecular clusters. Mol Biotechnol 2025; 67:1852–1866.
35. Zhang C, Luo Z, Ji L. Identification of potential diagnostic markers and molecular mechanisms of asthma and ulcerative colitis based on bioinformatics and machine learning. Front Mol Biosci 2025; 12:1554304.
36. Chen Y, Wang J, Zhang Y, *et al.* Investigation of key ferroptosis-associated genes and potential therapeutic drugs for asthma based on machine learning and regression models. Sci Rep 2025; 15:20342.
37. Tawfik M, Fathi IS, Nimbhore SS, *et al.* E-RespiNet: an LLM-ELECTRA driven
■ triple-stream CNN with feature fusion for asthma classification. PLoS One 2025; 20:e0334528.
38. Tsang KCH, Pinnock H, Wilson AM, *et al.* Predicting asthma attacks using connected mobile devices and machine learning: the AAMOS-00 observational study protocol. BMJ Open 2022; 12:e064166.
39. Salvi D, Olsson CM, Ymeri G, *et al.* Mobistudy: mobile-based, platform-independent, multi-dimensional data collection for clinical studies. 2021; ACM, St. Gallen, Switzerland: 219–222. Proceedings of the 11th international conference on the internet of things. https://dl.acm.org/doi/10.1145/3494322.3494363.
40. Tsang KCH. Enhancing asthma self-management with environmental passive-monitoring data and machine learning-based predictions. Stud Health Technol Inform 2024; 316:700–704.
41. Chan YFY, Bot BM, Zweig M, *et al.* The asthma mobile health study, smartphone data collected using ResearchKit. Sci Data 2018; 5:180096.
42. Chan YFY, Wang P, Rogers L, *et al.* The Asthma Mobile Health Study, a large-scale clinical observational study using ResearchKit. Nat Biotechnol 2017; 35: 354–62.
43. Honkoop PJ, Simpson A, Bonini M, *et al.* MyAirCoach: the use of home-monitoring and mHealth systems to predict deterioration in asthma control and the occurrence of asthma exacerbations; study protocol of an observational study. BMJ Open 2017; 7:e013935.
44. Bui AAT, Hosseini A, Rocchio R, *et al.* Biomedical REAl-Time Health Evaluation (BREATHE): toward an mHealth informatics platform. JAMIA Open 2020; 3: 190–200.
45. Xie F, Zeiger RS, Saparudin MM, *et al.* Identifying asthma-related symptoms
■■ from electronic health records using a hybrid natural language processing approach within a large integrated health care system: retrospective study. JMIR 2025; 4:e69132.
46. Bağcı MF, Do T, Spierling Bagsic SR, *et al.* Detection and prediction of real-
■■ world severe asthma phenotypes by application of machine learning to electronic health records. J Allergy Clin Immunol Glob 2025; 4:100473.
47. Chen W, Puttock EJ, Xie F, *et al.* Symptoms of asthma extracted through
■ natural language processing and their associations with acute asthma exacerbation in adults with mild asthma. J Allergy Clin Immunol Pract 2025; 13: 1719–1729.e7.
48. Mehta V, Pardeshi A, Rahman R, *et al.* Prevalence and predictors of asthma among Indian women: a machine learning-based analysis of NFHS-5 data. BMC Public Health 2025; 25:3847.
49. Liljendahl MS, Ibler K, Vestergaard C, *et al.* Identifying mild-to-moderate atopic
■ dermatitis using a generic machine learning approach: a Danish National Health Register Study. Acta Derm Venereol 2025; 105:adv42250.
50. Kim S, Han CH, Chang J, *et al.* Comparative risk for neuropsychiatric events in leukotriene receptor antagonist vs. inhaled corticosteroid in children with asthma: a nationwide observational study with a complementary analysis using natural language processing. Pharmacoepidemiol Drug Saf 2025; 34:e70254.
51. Wu CP, Sleiman J, Fakhry B, *et al.* Novel machine learning identifies 5 asthma phenotypes using cluster analysis of real-world data. J Allergy Clin Immunol Pract 2024; 12:2084–2091.e4.
52. Ding N, Li Y, Zhao Z, *et al.* Differential diagnosis of eczema and psoriasis using
■ routine clinical data and machine learning: development of a web-based tool in a multicenter outpatient cohort. Front Med 2025; 12:1667794.
53. Bashir MBA, Lisik D, Ermis SSO, *et al.* Unsupervised machine learning
■ identifies asthma phenotypes in the population-based West Sweden Asthma Study. Clin Exp Allergy 2026; 56:170–172.
54. Li T, Zhang J, Wu J, *et al.* Asthma detection research based on voice signal processing and machine learning. J Vis Exp 2025.
55. Morgenstern C, Bartosik TJ, Bayer KN, *et al.* Proteomic profiling and machine learning for endotype prediction in chronic rhinosinusitis. J Allergy Clin Immunol 2025; 157:190–202.
56. Fu Y, Zhao J, Wang Y. LASSO regression and Boruta algorithm to explore the relationship between neutrophil percentage to albumin ratio and asthma: results from the NHANES 2001 to 2018. Clin Exp Med 2025; 25:149.
57. Pattarakiatjaroen M, Yuenyongviwat A, Sangsupawanich P. Development of a
■ clinical predictive score for allergic reactions during oral food challenges in pediatric patients. PLoS One 2025; 20:e0322152.
58. Weidinger S, Bewley A, Hong HCH, *et al.* Predicting success with reduced dosing frequency of tralokinumab in patients with moderate-to-severe atopic dermatitis. Br J Dermatol 2025; 192:410–419.
59. Bousquet J, Anto J, Sunyer J, *et al.* Pooling birth cohorts in allergy and asthma: European Union-funded initiatives – a MeDALL, CHICOS, ENRIECO, and GA^{2}LEN joint paper. Int Arch Allergy Immunol 2013; 161:1–10.
60. Peters RL, Guarnieri I, Tang MLK, *et al.* The natural history of peanut and egg allergy in children up to age 6 years in the HealthNuts population-based longitudinal study. J Allergy Clin Immunol 2022; 150:657–665.e13.
61. Kotsapas C, Nicolaou N, Haider S, *et al.* Early-life predictors and risk factors of peanut allergy, and its association with asthma in later-life: Population-based birth cohort study. Clin Exp Allergy 2022; 52:646–657.
62. Asarnoj A, Hamsten C, Lupinek C, *et al.* Prediction of peanut allergy in adolescence by early childhood storage protein-specific IgE signatures: The BAMSE population-based birth cohort. J Allergy Clin Immunol 2017; 140:587–590.e7.
63. Khan M, Banerjee S, Muskawad S, *et al.* The impact of artificial intelligence on allergy diagnosis and treatment. Curr Allergy Asthma Rep 2024; 24: 361–372.
64. Frau F, Loustalot P, Törnqvist M, *et al.* Connecting electronic health records to a
■ biomedical knowledge graph to link clinical phenotypes and molecular endotypes in atopic dermatitis. Sci Rep 2025; 15:3082.
65. Li C, Ying M, Wu F, *et al.* The association between metabolic score for visceral fat and asthma incidence risk: a machine learning analysis based on the NHANES database. Medicine (Baltimore) 2025; 104:e44640.
66. Nakajima S, Nakamizo S, Nomura T, *et al.* Integrating multi-omics approaches in deciphering atopic dermatitis pathogenesis and future therapeutic directions. Allergy sept 2024; 79:2366–2379.
67. Howard R, Fontanella S, Simpson A, *et al.* Component-specific clusters for diagnosis and prediction of allergic airway diseases. Clin Exp Allergy 2024; 54:339–349.
68. Roche N, Reddel H, Martin R, *et al.* Quality standards for real-world research. Focus on observational database studies of comparative effectiveness. Ann Am Thorac Soc 2014; 11 (Suppl 2):S99–S104.

69. Barbiero P, Squillero G, Tonda A. Modeling generalization in machine learning: a methodological and computational study [Internet]. arXiv; 2020 [cité 23 déc 2025]. http://arxiv.org/abs/2006.15680

70. Liu A, Zhang Y, Yadav CP, Chen W. An updated systematic review on asthma exacerbation risk prediction models between 2017 and 2023: risk of bias and applicability. J Asthma Allergy 2025; 18:579–589.

71. Thygesen LC, Ersbøll AK. When the entire population is the sample: strengths and limitations in register-based epidemiology. Eur J Epidemiol aoÛt 2014; 29: 551–558.

72. Gomes M, Turner AJ, Sammon C, *et al.* Acceptability of using real-world data to estimate relative treatment effects in health technology assessments: barriers and future steps. Value Health mai 2024; 27:623–632.

73. Naudet F, Sakarovitch C, Janiaud P, *et al.* Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in *the BMJ* and *PLOS Medicine*. BMJ 2018; k400.

74. Rieke N, Hancox J, Li W, *et al.* The future of digital health with federated learning. Npj Digit Med 2020; 3:119.

75. Xu J, Glicksberg BS, Su C, *et al.* Federated learning for healthcare informatics. J Healthc Inform Res 2021; 5:1–19.

76. Brisimi TS, Chen R, Mela T, *et al.* Federated learning of predictive models from federated electronic health records. Int J Med Inf 2018; 112:59–67.

77. Sagheb E, Wi CI, King KS, *et al.* AI model for predicting asthma prognosis in children. J Allergy Clin Immunol Glob 2025; 4:100429.

78. ■ Turcatel G, Xiao Y, Caveney S, *et al.* Predicting asthma exacerbations using machine learning models. Adv Ther 2025; 42:362–374.

79. Xu J, Talankar S, Pan J, *et al.* Combining federated machine learning and qualitative methods to investigate novel pediatric asthma subtypes: protocol for a mixed methods study. JMIR Res Protoc 2024; 13:e57981.

80. Botero-Aguirre JP, García-Rivera MA. Natural language processing for enhanced clinical decision support in allergy verification for medication prescriptions. Mayo Clin Proc Digit Health 2025; 3:100244.

81. ■ Mora T, Roche D, Muñoz-Cano R. Predicting first-time anaphylaxis in the elderly using stacked machine learning and population registers. Front Allergy 2025; 6:1655662.

82. Kural KC, Mazo I, Walderhaug M, *et al.* Using machine learning to improve anaphylaxis case identification in medical claims data. JAMIA Open 2024; 7: ooae037.

83. Stanekova V, Inglis JM, Lam L, *et al.* Improving the performance of machine learning penicillin adverse drug reaction classification with synthetic data and transfer learning. Intern Med J 2024; 54:1183–1189.

84. ■ Wong AF, Bilaver LA, Jiang J, *et al.* Assessing pediatric clinician adherence to the guidelines for prevention of peanut allergy: a natural language processing study. BMC Med Inform Decis Mak 2025; 26:9.

85. Sagheb E, Wi CI, Yoon J, *et al.* Artificial intelligence assesses clinicians' adherence to asthma guidelines using electronic health records. J Allergy Clin Immunol Pract 2022; 10:1047–1056.e1.

86. Riley RD, Ensor J, Snell KIE, *et al.* External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016; 353:i3140.

87. Shamji MH, Ollert M, Adcock IM, *et al.* EAACI guidelines on environmental science in allergic diseases and asthma – leveraging artificial intelligence and machine learning to develop a causality model in exposomics. Allergy 2023; 78:1742–1757.

88. Yang J, Wang L, Phadke NA, *et al.* Development and validation of a deep learning model for detection of allergic reactions using safety event reports across hospitals. JAMA Netw Open 2020; 3:e2022836.

89. Prosperi MC, Marinho S, Simpson A, *et al.* Predicting phenotypes of asthma and eczema with machine learning. BMC Med Genomics 2014; 7 (Suppl 1):S7.

90. Enzenbach C, Wicklein B, Wirkner K, Loeffler M. Evaluating selection bias in a population-based cohort study with low baseline participation: the LIFE-Adult-Study. BMC Med Res Methodol 2019; 19:135.

91. Wang S, Verpillat P, Rassen J, *et al.* Transparency and reproducibility of observational cohort studies using large healthcare databases. Clin Pharmacol Ther 2016; 99:325–332.

92. Staartjes VE, Kernbach JM. Significance of external validation in clinical machine learning: let loose too early? Spine J 2020; 20:1159–1160.

93. Yu XX, Liu MQ, Li XY, *et al.* Qualitative and quantitative prediction of food allergen epitopes based on machine learning combined with in vitro experimental validation. Food Chem 2023; 405:134796.

94. Hampson LV, Izem R. Innovative hybrid designs and analytical approaches leveraging real-world data and clinical trial data. In: He W, Fang Y, Wang H, editors. Real-world evidence in medical product development. 2023; Springer International Publishing, Cham: 211–232.

95. Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: a mini-review, two showcases and beyond. Inf Fusion 2022; 77:29–52.

96. Reps JM, Williams RD, Schuemie MJ, *et al.* Learning patient-level prediction models across multiple healthcare databases: evaluation of ensembles for increasing model transportability. BMC Med Inform Decis Mak 2022; 22:142.

97. Cabitza F, Campagner A, Balsano C. Bridging the « last mile » gap between AI implementation and operation: « data awareness » that matters. Ann Transl Med 2020; 8:501.

