



Published in final edited form as:

J Allergy Clin Immunol Pract. 2022 July ; 10(7): 1757–1762. doi:10.1016/j.jaip.2022.04.016.

Novel Analysis Methods to Mine Immune-Mediated Phenotypes and Find Genetic Variation Within the Electronic Health Record (Roadmap for Phenotype to Genotype: Immunogenomics)

Matthew S. Krantz, MD^a, V. Eric Kerchberger, MD^{a,b}, Wei-Qi Wei, MD, PhD^b

^aDivision of Allergy, Pulmonary and Critical Care Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, Tenn

^bDepartment of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tenn

Abstract

The field of immunogenomics has the opportunity for accelerated genetic discovery aided by the maturation of electronic health records (EHRs) linked to DNA biobanks. Novel analysis methods in deep phenotyping of EHR data allow the full realization of the paired and increasingly dense genetic/phenotypic information available. This enables researchers to uncover genetic risk factors for the prevention and optimal treatment of immune-mediated diseases and immune-mediated adverse drug reactions. This article reviews the background of EHRs linked to DNA biobanks, potential applications to immunogenomic discovery, and current and emerging techniques in EHR-based deep phenotyping.

Keywords

Immunogenomics; Biomedical informatics; Deep phenotyping; Electronic health records; DNA biobanks; Immune-mediated diseases; Immune-mediated adverse drug reactions

INTRODUCTION

In the 2 decades since the completion of the Human Genome Project in 2003, dense and inexpensive genotyping technology linked to electronic health records (EHRs) has emerged as a novel domain in biomedical research, leading to numerous discoveries in genomic medicine.¹ However, although EHRs contain a vast array of clinical data that can be leveraged for medical research, using EHRs for research presents novel challenges because these systems were designed for clinical care and not specifically to facilitate research. The EHR data exist in both structured and unstructured formats hampering their ready use in research applications.² Clinical data are often fragmented across health care systems with limited ability for exchange of EHR information because patients visit multiple medical providers, have changes in insurance, and move.^{3,4} Further, variance in clinical

Corresponding author: Matthew S. Krantz, MD, T-1218 Medical Center North, 1161 21st Ave. S., Nashville, TN 37232-2650. matthew.s.krantz@vumc.org.

Conflicts of interest: The authors declare that they have no relevant conflicts of interest.

data gathering, clinician experience, documentation practices, or coding misspecification can introduce error into the EHR during any clinical encounter.⁵ Therefore, significant effort is expended on the development of phenotyping algorithms for accurate identification of cases and controls, which requires involvement of experts both in biomedical informatics and clinicians with domain knowledge relevant to the clinical question of interest. Overall, this process can be time-intensive and costly, require iterative algorithm refinement along with manual chart review, and run the risk of generating overfitted models that lack generalizability for replication in other settings. In this review, we provide an overview of large-scale EHR biobanks, current methods to mine phenotypes from these resources for genetic association studies, and a roadmap for emerging novel analysis approaches with a particular focus on the field of immunogenetics (Figure 1).

ELECTRONIC HEALTH RECORDS LINKED TO DNA BIOBANKS

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 triggered widespread adoption of EHRs in the United States, generating clinical data at scales previously unattainable with paper medical records.⁶ Anticipating the potential secondary value of the vast phenotypic data amassing in EHRs, health care organizations started collecting DNA biospecimens from volunteer patients, either through prospective collection or using discarded blood samples obtained during the course of routine clinical care. Examples of current efforts to utilize clinical data for genetic discovery include single-institute collections (eg, Vanderbilt BioVU and Mount Sinai BioMe) and national datasets (eg, the United Kingdom Biobank, the Estonian Biobank, the Million Veteran Program, and the All of Us Research Program).^{7–12} Collaborations such as the Electronic Medical Records and Genomics Network (eMERGE) and Integrating Biology and the Bedside (i2b2) develop tools for researchers to leverage EHRs for genetic studies.^{13,14} Clinical data incorporated into these biobanks vary but broadly fall into 2 categories: structured data or unstructured data. Structured EHR data include billing codes such as the International Classification of Diseases (ICD-9 or ICD-10), symptoms (Systematized Nomenclature of Medicine—Clinical Terms [SNOMED-CT]), current procedural terminology (CPT) codes, laboratory test results (Logical Observation Identifier Names and Codes [LOINC]), and medication prescription data (RxNorm).^{15,16} Clinical notes and reports from radiology or pathology studies represent the majority of unstructured EHR data and necessitate use of contextual natural language processing (NLP) to transform them into an output useful for phenotyping. Radiology and pathology reports can be further complicated if stored in external databases or entered in the EHR as a portable document format. Although clinical notes have this limitation, they can provide opportunities to capture health information not specifically or reliably captured in structured billing codes, such as drug infusions reactions and anaphylaxis to penicillin.

EXTRACTING RESEARCH-GRADE PHENOTYPES FROM THE EHR

Strategies for phenotype algorithm development for case and control identification can be constructed in 2 main ways: (1) manual feature selection by informaticians and domain experts and (2) high-throughput methods, including grouping similar diagnosis codes, automated feature selection from public medical knowledge sources, and machine learning (ML) approaches, both supervised and unsupervised.^{17–20} Domain experts have the clinical

phenotyping skills for the most salient features for case and control identification and how these features are documented in the EHR. Clinical informaticians have the skills to extract these features from the EHR and transform them into data for analysis. In an iterative process, domain experts and clinical informaticians collaborate to identify and extract features for a phenotype definition until sufficient sensitivity is achieved.²¹ In either manual feature selection or high-throughput methods, a clinician-reviewed, gold-standard training dataset must be employed for algorithm validation and determination of accuracy²².

ADVANTAGES AND DISADVANTAGES OF EHR BIOBANKS IN BIOMEDICAL RESEARCH

The EHR-based research affords distinct advantages compared with traditional prospective cohort studies. This includes greater cost efficiency because patient recruitment is broadly applied across a population rather than targeted to specific conditions.²³ Once the initial upfront costs of obtaining genetic data are expended, serial use for diverse clinical questions comes with marginal additional cost, primarily for developing and executing phenotypic algorithms.²³ Further, EHR biobanks can dramatically increase available sample sizes with potentially thousands to hundreds of thousands of genotyped patients, increasing study power and making possible detection of rare disease-associated variants.²⁴ The EHR-based studies often have much shorter development cycles while still yielding similar genetic association results to those found in traditional cohort studies.²³ Whereas genetic associations for immune-mediated diseases have been replicated using EHR-based studies, immune-mediated adverse drug reactions (ADRs) present a significant challenge despite their high effect size with class I human leukocyte antigen (HLA) alleles. These challenges include poor precision of phenotypes and low prevalence compared with immune-mediated diseases, such as type 1 diabetes or hypothyroidism. This limitation is highlighted by a phenome-wide association study (PheWAS) by Karnes et al²⁵ in which many known immune-mediated disease HLA associations were replicated yet no ADR associations were replicated. Longitudinal clinical information can be gleaned from EHRs; however, it may be incomplete owing to fragmentation across disparate health care systems with separate EHRs.²⁶ Data available for extraction are limited and biased toward what is entered in real-time during routine clinical care, which often is not standardized nor comprehensive for all phenotypes of interest. Expertise in EHR phenotyping, although growing, is mainly restricted to academic centers with National Library of Medicine biomedical informatics trainings programs, making both the number and the time of researchers a potential rate-limiting step for studies.²⁷

GENETIC ASSOCIATIONS WITH IMMUNE-MEDIATED DISEASES AND ADRS

Immunogenomics is the study of genetic associations with immune-mediated diseases and ADRs, including severe cutaneous adverse drug reactions (SCARs) such as Stevens-Johnson syndrome and toxic epidermal necrolysis (SJS/TEN), which is associated with an overall mortality of 20%.^{28,29} The primary objective of immunogenomics is delivering precision medicine to patients by predicting, preventing, and identifying novel therapeutic targets for immune-mediated diseases or ADRs. Most genetic associations with immune-mediated

diseases and ADRs localize to a region of chromosome 6 encoding the HLA genes. The HLA region is the most polymorphic region of the human genome and contains 3 classic HLA class I alleles (-A, -B, and -C), which are expressed on all nucleated cells, and 3 classic HLA class II alleles (-DP, -DQ, and -DR), which are expressed only by professional antigen-presenting cells. Through HLA, processed protein antigens are presented to specific T-cell receptors for the required signal 1 of T-cell activation. Examples of autoimmune diseases associated with HLA include HLA-B*27 with ankylosing spondylitis, HLA-DRB1*03:01 with systemic lupus erythematosus, and the HLA-DQ2.5 haplotype (HLA-DQA1*05:01/DQB1*02:01) and the DQ8 haplotype (HLA-DQA1*03:01/DQB1*03:02) with celiac disease.²⁵ The SCARs encompass drug reaction with eosinophilia and systemic symptoms (DRESS), SJS/TEN, and acute generalized exanthematous pustulosis. The HLA risk allele associations with specific drugs have been identified for SJS/TEN and DRESS.³⁰ These SCARs and HLA associations include examples such as vancomycin with HLA-A*32:01, allopurinol with HLA-B*58:01, and HLA-B* 15:02 and HLA-A*31:01 with carbamazepine.^{31–33} However, these associated HLA alleles are often necessary but not sufficient for development of SCARs, dampening their clinical utility for determining risk of SCARs following exposure to a specific drug.^{31,34} The positive predictive value of HLA risk alleles are affected by the prevalence of disease, frequency of the risk allele in a given population, and the underlying biological complexity of immune-mediated ADRs. The negative predictive value of a HLA risk allele determines the ultimate safety of the test, meaning the proportion of patients with an absent risk allele who do not develop an immune-mediated ADR. Variable antigen presentation and T-cell activation are not restricted to HLA alleles alone but further involve differences in metabolism and proteasome processing, trimming of peptides by endoplasmic reticulum aminopeptidases (ERAP), T-cell receptor specificity, and costimulatory signals received through killer cell immunoglobulin-like receptors (KIR), which can lead to either greater cytotoxicity or tolerance.²⁸

MODERN IMMUNOTHERAPY IS REVOLUTIONIZING TREATMENT OF IMMUNE-MEDIATED DISEASES

Downstream of the effects of genetic and ecological factors, functional correlates from patient biospecimens have led to the development of targeted treatments in cancer, autoimmune disorders, and allergic diseases, primarily against immune-related molecules such as cytokines, chemokines, and their receptors. Immunotherapies have revolutionized cancer treatment with checkpoint inhibitors, including monoclonal antibodies against antiprogrammed cell death protein (PD)-1, anti-PD ligand-1, and anticytotoxic T-lymphocyte—associated protein 4, as well as several chimeric antigen receptor T-cell therapies all approved in the last decade.³⁵ Autoimmune disorders including inflammatory bowel disease, rheumatoid arthritis, psoriasis, and systemic lupus erythematosus now have a variety of steroid-sparing therapies that leverage their skewed cell-mediated and cytokine profiles, including monoclonal antibodies against tumor necrosis factor, interleukin (IL)-12/IL-23, and B-cell activating factor, as well as small molecule Janus kinase (JAK) inhibitors.^{36,37} In allergic diseases such as asthma and atopic dermatitis, biologic therapies against the type 2 immune response have also ushered in a therapeutic paradigm shift, targeted IL-4 receptor alpha, IL-5, IL5 receptor, and thymic stromal lymphopoietin.^{38,39}

Although currently no immunotherapies are approved for immune-mediated ADRs, a recent case report of prolonged and refractory DRESS, single-cell transcriptomic analysis from lesional skin biopsy demonstrated significant upregulation of the JAK-STAT (single transducer and activator of transcription) pathway.⁴⁰ This prompted treatment with a JAK inhibitor (tofacitinib) and led to resolution of DRESS, highlighting the promise for precision medicine and targeted therapies for SCARs.

IDENTIFYING IMMUNE-MEDIATED PHENOTYPES IN THE EHR

Currently, several challenges remain between connecting the genetic drivers of immune-mediated diseases and ADRs and selecting the optimal therapy for each patient. The EHR biobanks provide a unique opportunity to address these challenges. High-throughput phenotyping methods are needed to accurately identify appropriate cases and controls for these phenotypes. This remains an area of active research utilizing multiple approaches. Wolfson et al⁴¹ recently reported on the feasibility of a semi-automated NLP algorithm to identify DRESS cases, a relatively rare drug reaction, using the allergy list of a large multihospital EHR, and their findings were consistent with data derived from U.S. Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS).⁴² Yang et al⁴³ reported an ML method to detect allergic reactions from free-text narratives of hospital safety event reports, which has the potential to permit real-time surveillance and significantly reduce the number of cases that require manual review.

CONNECTING IMMUNE-MEDIATED PHENOTYPES TO GENOTYPES

Once accurate phenotypes of immune-mediated diseases and ADRs are developed in an EHR biobank, these phenotypes must be analyzed in the context of genetic and transcriptomic data available in the biobank, which may be in either low- or high-resolution genomic coverage formats. Modern genotyping arrays sample from several hundred thousand to a few million single nucleotide polymorphisms (SNPs), insertions, or deletions across the human genome, and are currently the most cost-efficient and scalable format for obtaining genetic data. Although these platforms are powerful, they are considered low-resolution because they sample only a small fraction of the over 3 billion nucleotides constituting the human genome. Additional SNPs at the whole-genome level can be imputed with use of software packages like Beagle and IMPUTE.^{44,45} Importantly, these platforms typically do not specifically provide HLA typing nor identify all HLA variants, potentially limiting their application to immunogenomics. These challenges may be ameliorated with use of imputation programs such as SNP2HLA and HIBAG that can impute HLA alleles from SNP data with reasonably high accuracy in most populations.⁴⁶ Imputation programs have been developed for KIR alleles, such as Pulling Out Natural Killer cell Genomics, and ERAP as well.^{47,48} Tissue-specific gene expression can also be imputed with programs, such as PrediXcan, that importantly instead predict mRNA expression based on a patient's known genotypes.⁴⁹ These tools enable studies that would otherwise require acquisition of high-resolution genetic or expression data either through dedicated allele sequencing or whole-genome sequencing or obtaining tissue-specific mRNA sequencing. Notably, full allelic sequencing remains the gold standard for class I and II HLA typing. By using KIR imputation programs, this has led to discoveries in associations of KIR with atopic

dermatitis and psoriasis.^{50,51} In addition, epistatic variation of ERAP polymorphisms in those carrying the HLA-B*57:01 risk allele, which is necessary but not sufficient for development of abacavir hypersensitivity syndrome, has been shown to either be protective against or confer greater chance of development of abacavir hypersensitivity syndrome owing to differences in antigen processing into immunodominant epitopes.^{34,52}

EMERGING NOVEL APPROACHES FOR ADVANCED EHR PHENOTYPING OF IMMUNE-MEDIATED DISEASES AND ADRS

Because research-grade genetic data are generated by automated methods and stored in structured formats, often the rate-limiting step needed for immunogenomic studies in EHRs lies in generating comparably research-grade phenotyping.⁵³ A common and efficient method of high-throughput phenotyping is the use of phecodes, which are clinically meaningful groupings of related ICD codes for specific diseases.^{17,54} Phecodes repurpose ICD codes for research that were initially designed to bill a patient's insurance. Phecode-based algorithms assist in identifying both phenotype cases and clinically meaningful phenotype controls that lack any of the relevant ICD codes bundled into a phecode. Genome-wide association studies and PheWAS have been successfully performed using phecodes to replicate known findings from prior literature and discover previously unidentified associations between HLA alleles and immune-mediated diseases or ADRs.²⁵ The most current version of phecode mappings are available from <https://www.vumc.org/wei-lab/phecode>.

Deep phenotyping of immune-mediated ADRs by ICD codes or phecodes alone has several particular challenges. There is a bias toward under-entry of relevant ICD codes for immune-mediated ADRs other than at the time of index reaction or at time of consultation referral to an allergist/immunologist. Therefore, if the index reaction occurred in another health care system or no subsequent assessment of the immune-mediated ADR was made, the patient may never have a relevant ICD code entered into the EHR, leading to the under-recognition and underdiagnosis of immune-mediated ADRs. To overcome the potential misclassification of controls in studies, a control may be defined more stringently by having both prior medication prescription data and absence of an allergy label for the drug of interest.

Further, ICD codes have limited specificity for immune-mediated ADRs; thus, they may not allow a researcher to accurately identify the immune-mediated ADR. For example, an ICD code was not available for “drug rash with eosinophilia and systemic symptoms syndrome” (D72.12) until ICD-10.¹⁵ Consequently, clinicians using ICD-9 would instead enter single or multiple nonspecific ICD-9 codes to approximate a DRESS diagnosis, such as “dermatitis due to drugs and medicines” (693.0), “eosinophilia” (288.3), “hepatitis, unspecified” (573.3), or “enlargement of lymph nodes” (785.6).¹⁵ In order to overcome this under-representation of immune-mediated ADRs in ICD codes, alternative frameworks have leveraged the allergy list from clinical notes. The allergy list exists in narrative text as a semistructured format and typically includes at least 2 key data elements: both the allergen (eg, offending medication name such as “penicillin”) and the description of the reaction (eg, “urticaria”). Because of this consistent format, the allergy list field can be mined in

a high-throughput fashion without advanced NLP. A proof-of-concept study by Wu et al⁵⁵ evaluated using the EHR allergy list to perform drug-drug interaction wide association study (DDIWAS), replicating over 30 known drug-drug interactions and describing 3 potential novel interactions. Zheng et al⁵⁶ also recently described a high-throughput framework utilizing the allergy list to efficiently identify ADRs in the EHR for genetic analyses. Because the format of the allergy list is often standardized across EHRs, the DDIWAS approach can be rapidly deployed to other institutions for detection and validation of drug-drug interactions.

In addition to structured ICD codes and the semistructured allergy list, many other sources of structured data and unstructured narrative text exist in the EHR that may provide the basis for future immunogenomic phenotypes. Harnessing clinical data (eg, laboratory tests, medications) in tandem with ICD codes has been shown to improve the quality of phenotyping but face challenges on how to streamline process of high-throughput development.⁵⁷ Traditional phenotyping methods using such unstructured notes used NLP text scanning algorithms to extract words and text features that were curated by clinician domain experts.^{22,57} Although this process has led to development of high-quality validated algorithms, such as those available from the eMERGE network, it is time-burdensome and requires onerous manual chart review and sequential algorithm adjustment for each disease of interest.^{22,58} The recently published PheMap algorithm overcomes some of these challenges by leveraging multilevel knowledge available from online resources, such as Mayo Clinic and Wikipedia. PheMap scans online articles on specific diseases and conditions to extract the conditions, medications, laboratory tests, and other facts mentioned in the article text and map those features to standard medical terminologies including ICD-9 and ICD-10, SNOMED-CT, CPT, LOINC, and RxNorm.¹⁸ The algorithm then uses the frequency of various EHR features among each online article to determine the strength of associations between EHR features and diseases. Finally, PheMap examines patients' medical records for the presence of each EHR feature to assign a final probability for having a phenotype. As an example: the presence of insulin in a patient's medication record might indicate a strong probability that the patient has diabetes, whereas the presence a urinalysis test could be a less strong indicator for diabetes because a urinalysis can be used to screen diabetic patients for proteinuria but is also used for diagnosis of many other nondiabetic conditions. When compared with current phenotyping approaches, including eMERGE-based phenotypes, eXtraction of Phenotypes from Records using Silver Standards (XPRESS), and phecodes, PheMap performed comparably or better for calculating an individual's probability of having type 2 diabetes mellitus, dementia, or hypothyroidism.^{18,22,59,60}

The final frontier of deep EHR phenotyping lies in utilizing the rich unstructured data contained in clinical notes. Whereas ML techniques to extract meaningful phenotypic data from unstructured notes in the EHR have emerged,⁶¹ many of these approaches currently rely upon supervised ML algorithms and, therefore, require efforts by clinical domain experts to provide case-control labels for training data and/or assist with model feature selection.²⁰ Furthermore, the inner workings of the most complex artificial intelligence algorithms are often uninterpretable to humans; therefore, ML models run the risk of

overfitting and lacking generalizability to other institutions for replication owing to differences such as cohort demographics and clinician diagnostic patterns.^{62,63}

DISCUSSION

With the continued maturation of EHR biobanks, immunogenomic discoveries will be further advanced as emerging deep phenotyping methods enable large-scale studies of immune-mediated phenotypes and their genetic drivers. Current algorithms are most focused on defining genetic associations with phenotypes; however, these genetic discoveries will ultimately provide opportunities for integrated approaches for the prediction, prevention, diagnosis, and treatment of immune-mediated diseases and ADRs in at-risk patients.

Acknowledgments

V. E. Kerchberger receives funding from National Institutes of Health (NIH; grant number K01HL157755-01) and the American Thoracic Society. W. Q. Wei receives funding from multiple National Institutes of Health (NIH; grant numbers R01 GM139891, R01AG069900, U01 HG011181-01, R01 HL133786-01A1) and the American Heart Association (AHA).

Abbreviations used

ADRs	Adverse drug reactions
DRESS	Drug reaction with eosinophilia and systemic symptoms
EHRs	Electronic health records
eMERGE	Electronic Medical Records and Genomics Network
ERAP	Endoplasmic reticulum aminopeptidases
HLA	Human leukocyte antigen
ICD	International Classification of Diseases
IL	Interleukin
JAK	Janus kinase
KIR	Killer cell immunoglobulin-like receptors
ML	Machine learning
NLP	Natural language processing
SCARs	Severe cutaneous adverse drug reactions
SJS/TEN	Stevens-Johnson syndrome and toxic epidermal necrolysis
SNP	Single nucleotide polymorphisms

REFERENCES

1. Abul-Husn NS, Kenny EE. Personalized medicine and the power of electronic health records. *Cell* 2019;177:58–69. [PubMed: 30901549]
2. Wei W-Q, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. *Genome Med* 2015;7:41. [PubMed: 25937834]
3. Wei W-Q, Leibson CL, Ransom JE, Kho AN, Caraballo PJ, Chai HS, et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J Am Med Inform Assoc* 2012;19:219–24. [PubMed: 22249968]
4. Wei W-Q, Leibson CL, Ransom JE, Kho AN, Chute CG. The absence of longitudinal data limits the accuracy of high-throughput clinical phenotyping for identifying type 2 diabetes mellitus subjects. *Int J Med Inform* 2013;82:239–47. [PubMed: 22762862]
5. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res* 2005;40:1620–39. [PubMed: 16178999]
6. Adler-Milstein J, Jha AK. HITECH act drove large gains in hospital electronic health record adoption. *Health Aff (Millwood)* 2017;36:1416–22. [PubMed: 28784734]
7. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, et al. Secondary use of clinical data: the Vanderbilt approach. *J Biomed Inform* 2014;52:28–35. [PubMed: 24534443]
8. Icahn School of Medicine. BioMe BioBank Program. Accessed January 24, 2022. <https://icahn.mssm.edu/research/ipm/programs/biome-biobank>.
9. UK Biobank. UK Biobank. Accessed January 17, 2022. <https://www.ukbiobank.ac.uk/>
10. EIT Health Scandinavia. The Estonian Biobank. Accessed January 17, 2022. <https://www.eithealth-scandinavia.eu/biobanks/the-estonian-biobank/>.
11. U.S. Department of Veterans Affairs. Million Veteran Program (MVP). 2021. Accessed January 17, 2022. <https://www.research.va.gov/mvp/>.
12. National Institutes of Health (NIH). All of Us Research Program. 2020. Accessed January 17, 2022. <https://allofus.nih.gov/future-health-begins-all-us>.
13. eMERGE. Welcome to eMerge > Collaborate. Accessed January 17, 2022. <https://emerge-network.org/>.
14. i2b2 TranSMART Foundation. i2b2: Informatics for Integrating Biology & the Bedside. Accessed January 17, 2022. <https://www.i2b2.org/>.
15. Centers for Disease Control and Prevention. International Classification of Diseases (ICD). 2019. Accessed January 21, 2022. <https://www.cdc.gov/nchs/icd/icd9.htm>.
16. Bodenreider O, Cornet R, Vreeman D. Recent developments in clinical terminologies—SNOMED CT, LOINC, and RxNorm. *Yearb Med Inform* 2018;27:129–39. [PubMed: 30157516]
17. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM codes to Phecodes: workflow development and initial evaluation. *JMIR Med Inform* 2019;7:e14325. [PubMed: 31553307]
18. Zheng NS, Feng Q, Kerchberger VE, Zhao J, Edwards TL, Cox NJ, et al. PheMap: a multi-resource knowledge base for high-throughput phenotyping within electronic health records. *J Am Med Inform Assoc* 2020;27:1675–87. [PubMed: 32974638]
19. Zheng NS, Kerchberger VE, Borza VA, Eken HN, Smith JC, Wei W-Q. An updated, computable MEDication-Indication resource for biomedical research. *Sci Rep* 2021;11:18953. [PubMed: 34556781]
20. Zhang Y, Cai T, Yu S, Cho K, Hong C, Sun J, et al. High-throughput phenotyping with electronic medical record data using a common semi-supervised approach (PheCAP). *Nat Protoc* 2019;14:3426–44. [PubMed: 31748751]
21. Pathak J, Kho AN, Denny JC. Electronic health records—driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc* 2013;20:e206–11. [PubMed: 24302669]
22. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record—based phenotyping algorithms: results and lessons learned from the eMERGE network. *J Am Med Inform Assoc* 2013;20:e147–54. [PubMed: 23531748]

23. Bowton E, Field JR, Wang S, Schildcrout JS, Van Driest SL, Delaney JT, et al. Biobanks and electronic medical records: enabling cost-effective research. *Sci Transl Med* 2014;6:234cm3.
24. Henderson GE, Cadigan RJ, Edwards TP, Conlon I, Nelson AG, Evans JP, et al. Characterizing biobank organizations in the U.S.: results from a national survey. *Genome Med* 2013;5:3. [PubMed: 23351549]
25. Karnes JH, Bastarache L, Shaffer CM, Gaudieri S, Xu Y, Glazer AM, et al. Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants. *Sci Transl Med* 2017;9:eaai8708. [PubMed: 28490672]
26. Bourgeois FC. Patients treated at multiple acute health care facilities: quantifying information fragmentation. *Arch Intern Med* 2010;170:1989. [PubMed: 21149756]
27. U.S. National Library of Medicine. NLMs University-based Biomedical Informatics and Data Science Research Training Programs. Accessed January 26, 2022. <https://www.nlm.nih.gov/ep/GrantTrainInstitute.html>.
28. Deshpande P, Hertzman RJ, Palubinsky AM, Giles JB, Karnes JH, Gibson A, et al. Immunopharmacogenomics: mechanisms of HLA-associated drug reactions. *Clin Pharmacol Ther* 2021;119:607–15.
29. Redwood AJ, Pavlos RK, White KD, Phillips EJ. HLAs: key regulators of T-cell—mediated drug hypersensitivity. *HLA* 2018;91:3–16. [PubMed: 29171940]
30. Blumenthal KG, Peter JG, Trubiano JA, Phillips EJ. Antibiotic allergy. *Lancet* 2019;393:183–98. [PubMed: 30558872]
31. Konvinse KC, Trubiano JA, Pavlos R, James I, Shaffer CM, Bejan CA, et al. HLA-A*32:01 is strongly associated with vancomycin-induced drug reaction with eosinophilia and systemic symptoms. *J Allergy Clin Immunol* 2019;144:183–92. [PubMed: 30776417]
32. Jarjour S, Barrette M, Normand V, Rouleau JL, Dubé M-P, de Denus S. Genetic markers associated with cutaneous adverse drug reactions to allopurinol: a systematic review. *Pharmacogenomics* 2015;16:755–67. [PubMed: 25965122]
33. Shi Y-W, Min F-L, Qin B, Zou X, Liu X-R, Gao M-M, et al. Association between HLA and Stevens-Johnson syndrome induced by carbamazepine in Southern Han Chinese: genetic markers besides B*1502? *Basic Clin Pharmacol Toxicol* 2012;111:58–64. [PubMed: 22348435]
34. Sousa-Pinto B, Pinto-Ramos J, Correia C, Gonçalves-Costa G, Gomes L, Gil-Mata S, et al. Pharmacogenetics of abacavir hypersensitivity: a systematic review and meta-analysis of the association with HLA-B*57:01. *J Allergy Clin Immunol* 2015;136:1092–1094.e3. [PubMed: 25934581]
35. Waldman AD, Fritz JM, Lenardo MJ. A guide to cancer immunotherapy: from T cell basic science to clinical practice. *Nat Rev Immunol* 2020;20:651–68. [PubMed: 32433532]
36. Burmester GR, Bijlsma JWJ, Cutolo M, McInnes IB. Managing rheumatic and musculoskeletal diseases—past, present and future. *Nat Rev Rheumatol* 2017;13:443–8. [PubMed: 28615732]
37. Tanaka Y, Luo Y, O'Shea JJ, Nakayamada S. Janus kinase—targeting therapies in rheumatology: a mechanisms-based approach. *Nat Rev Rheumatol* 2022;18:133–45. [PubMed: 34987201]
38. van de Veen W, Akdis M. The use of biologics for immune modulation in allergic disease. *J Clin Invest* 2019;129:1452–62. [PubMed: 30882368]
39. Wu AC, Busse WW. Biologic therapy in allergy practice: a new era in treatment has begun. *J Allergy Clin Immunol Pract* 2021;9:1118–20. [PubMed: 33685607]
40. Kim D, Kobayashi T, Voisin B, Jo J-H, Sakamoto K, Jin S-P, et al. Targeted therapy guided by single-cell transcriptomic analysis in drug-induced hypersensitivity syndrome: a case report. *Nat Med* 2020;26:236–43. [PubMed: 31959990]
41. Wolfson AR, Zhou L, Li Y, Phadke NA, Chow OA, Blumenthal KG. Drug Reaction with Eosinophilia and Systemic Symptoms (DRESS) syndrome identified in the electronic health record allergy module. *J Allergy Clin Immunol Pract* 2019;7:633–40. [PubMed: 30176295]
42. Bluestein SB, Yu R, Stone CA, Phillips EJ. Reporting of drug reaction with eosinophilia and systemic symptoms from 2002–2019 in the FDA Adverse Event Reporting System (FAERS). *J Allergy Clin Immunol Pract* 2021;9:3208–32011.e1. [PubMed: 34033979]

43. Yang J, Wang L, Phadke NA, Wickner PG, Mancini CM, Blumenthal KG, 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. [PubMed: 33196805]
44. Browning BL, Zhou Y, Browning SR. A one-penny imputed genome from next-generation reference panels. *Am J Hum Genet* 2018;103:338–48. [PubMed: 30100085]
45. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009;5:e1000529. [PubMed: 19543373]
46. Karnes JH, Shaffer CM, Bastarache L, Gaudieri S, Glazer AM, Steiner HE, et al. Comparison of HLA allelic imputation programs. *PLoS One* 2017;12:e0172444. [PubMed: 28207879]
47. Harrison GF, Leaton LA, Harrison EA, Viken MK, Shortt J, Gignoux CR, et al. Allele imputation for the killer cell Immunoglobulin-like receptor KIR3DL1/S1. *PLoS Comput Biol* 2022;18:e1009059. [PubMed: 35192601]
48. Reeves E, Edwards CJ, Elliott T, James E. Naturally occurring ERAP1 haplotypes encode functionally distinct alleles with fine substrate specificity. *J Immunol* 2013;191:35–43. [PubMed: 23733883]
49. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nat Genet* 2015;47:1091–8. [PubMed: 26258848]
50. Margolis DJ, Mitra N, Hoffstad OJ, Kim BS, Monos DS, Phillips EJ. Association of KIR genes and MHC class I ligands with atopic dermatitis. *J Immunol* 2021;207:1522–9. [PubMed: 34408014]
51. Ahn R, Vukcevic D, Motyer A, Nititham J, Squire DMcG, Hollenbach JA, et al. Large-scale imputation of KIR copy number and HLA alleles in North American and European psoriasis case-control cohorts reveals association of inhibitory KIR2DL2 with psoriasis. *Front Immunol* 2021;12:684326. [PubMed: 34177931]
52. Pavlos R, Deshpande P, Chopra A, Leary S, Strautins K, Nolan D, et al. New genetic predictors for abacavir tolerance in HLA-B*57:01 positive individuals. *Hum Immunol* 2020;81:300–4. [PubMed: 32173028]
53. Weng C, Shah NH, Hripcsak G. Deep phenotyping: embracing complexity and temporality—towards scalability, portability, and interoperability. *J Biomed Inform* 2020;105:103433. [PubMed: 32335224]
54. Wei W-Q, Bastarache LA, Carroll RJ, Marlo JE, Osterman TJ, Gamazon ER, et al. Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record. *PLoS One* 2017;12:e0175508. [PubMed: 28686612]
55. Wu P, Nelson SD, Zhao J, Stone CA Jr, Feng Q, Chen Q, et al. DDIWAS: high-throughput electronic health record-based screening of drug-drug interactions. *J Am Med Inform Assoc* 2021;28:1421–30. [PubMed: 33712848]
56. Zheng NS, Stone CA, Jiang L, Shaffer CM, Kerchberger VE, Chung CP, et al. High-throughput framework for genetic analyses of adverse drug reactions using electronic health records. *PLoS Genet* 2021;17:e1009593. [PubMed: 34061827]
57. Wei W-Q, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc* 2016;23:e20. [PubMed: 26338219]
58. Kirby JC, Speltz P, Rasmussen LV, Basford M, Gottesman O, Peissig PL, et al. PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability. *J Am Med Inform Assoc* 2016;23:1046–52.
59. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics* 2010;26:1205–10. [PubMed: 20335276]
60. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. *J Am Med Inform Assoc* 2016;23:1166–73. [PubMed: 27174893]

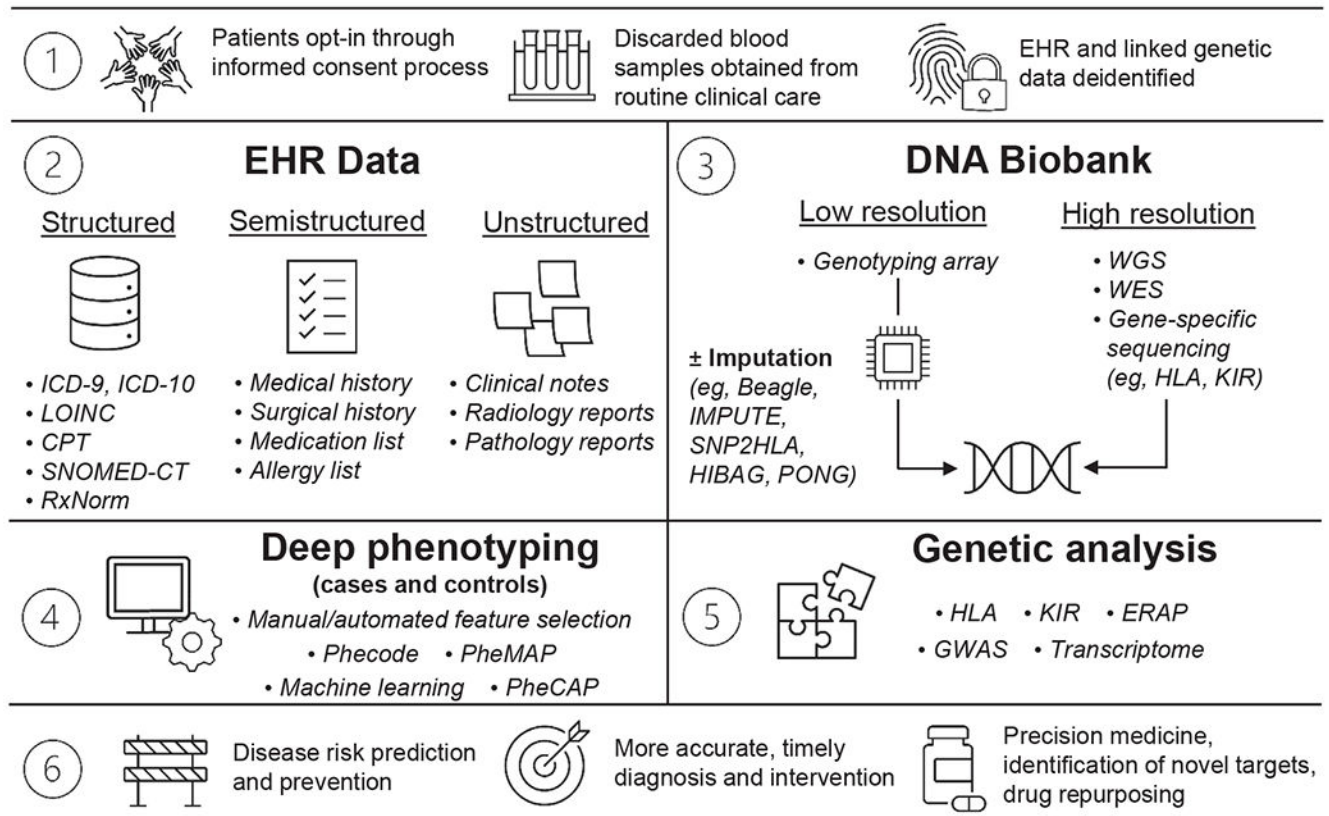
61. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med* 2019;380:1347–58. [PubMed: 30943338]
62. Song Y, Skinner J, Bynum J, Sutherland J, Wennberg JE, Fisher ES. Regional variations in diagnostic practices. *N Engl J Med* 2010;363:45–53. [PubMed: 20463332]
63. Andaur Navarro CL, Damen JAA, Takada T, Nijman SWJ, Dhiman P, Ma J, et al. Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 2021;375:n2281. [PubMed: 34670780]

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

**FIGURE 1.**

Phenotype to genotype roadmap for immunogenomics. (1) Patient enrollment process and acquisition of samples for DNA biobank. (2) Composition of key data structures in the EHR. (3) Quality of genetic data available in DNA biobanks and preprocessing for analysis. (4) Models for deep phenotyping algorithm development. (5) Select specific alleles of interest in immunogenomics and genetic analysis methods. (6) Translation of immunogenomic discoveries to immune-mediated diseases and ADRs. *CPT*, Current procedural terminology; *GWAS*, genome-wide association study; *HIBAG*, *HLA* Genotype Imputation with Attribute Bagging; *LOINC*, Logical Observation Identifier Names and Codes; *PONG*, Pulling Out Natural Killer cell Genomics; *SNOMED-CT*, Systematized Nomenclature of Medicine—Clinical Terms; *WES*, whole-exome sequencing; *WGS*, whole-genome sequencing.