Original Article

Development of Food Allergy Data Dictionary: Toward a Food Allergy Data Commons

Shruti Sehgal, MD (Hom), MS[a], Ruchi S. Gupta, MD, MPH[b,c], Mark Wlodarski, MS[a], Lucy A. Bilaver, PhD[b], Firas H. Wehbe, MD, PhD[d], Jonathan M. Spergel, MD, PhD[e], Julie Wang, MD[f], Christina E. Ciaccio, MD, MS[g], Melanie Makhija, MD[h], and Justin B. Starren, MD, PhD, FACMI[d] *Chicago, Ill; Philadelphia, Pa; and New York, NY*

What is already known about this topic? The terminology used to describe food allergy (FA) concepts and data elements is ambiguous and incomplete.

What does this article add to our knowledge? This article highlights the limitations of current FA concept coverage by existing clinical terminologies and describes the development and face validation of the first generation of the Food Allergy Data Dictionary.

How does this study impact current management guidelines? The Food Allergy Data Dictionary can help in limiting the variation in clinical practice by having defined critical FA concepts and data elements and is a pivotal resource for designing structured data collection forms for FA clinical encounters.

BACKGROUND: Food allergy (FA) data lacks a common base of terminology and hinders data exchange among institutions.
OBJECTIVE: To examine the current FA concept coverage by clinical terminologies and to develop and evaluate a Food Allergy Data Dictionary (FADD).
METHODS: Allergy/immunology templates and patient intake forms from 4 academic medical centers with expertise in FA were systematically reviewed, and in-depth discussions with a panel of FA experts were conducted to identify important FA clinical concepts and data elements. The candidate ontology was iteratively refined through a series of virtual meetings. The concepts were mapped to existing clinical terminologies manually with the ATHENA vocabulary browser. Finally, the revised dictionary document was vetted with experts across 22 academic FA centers and 3 industry partners.
RESULTS: A consensus version 1.0 FADD was finalized in November 2020. The FADD v1.0 contained 936 discrete FA concepts that were grouped into 14 categories. The categories included both FA-specific concepts, such as foods triggering reactions, and general health care categories, such as medications. Although many FA concepts are included in existing clinical terminologies, some critical concepts are missing.
CONCLUSIONS: The FADD provides a pragmatic tool that can enable improved structured coding of FA data for both research and clinical uses, as well as lay the foundation for the development of standardized FA structured data entry

[a]Center for Food Allergy and Asthma Research, Institute for Public Health and Medicine, Northwestern University Feinberg School of Medicine, Chicago, Ill
[b]Department of Pediatrics, Northwestern University Feinberg School of Medicine, Chicago, Ill
[c]The Mary Ann & J. Milburn Smith Child Health Outcomes, Research and Evaluation Center, Ann & Robert H. Lurie Children's Hospital of Chicago, Chicago, Ill
[d]Department of Preventive Medicine, Northwestern University Feinberg School of Medicine, Chicago, Ill
[e]Division of Allergy and Immunology, Children's Hospital of Philadelphia, Department of Pediatrics, Perelman School of Medicine at University of Pennsylvania, Philadelphia, Pa
[f]Department of Pediatrics, Jaffe Food Allergy Institute, Icahn School of Medicine at Mount Sinai, New York, NY
[g]Departments of Pediatrics and Medicine, The University of Chicago, Chicago, Ill
[h]Division of Allergy and Immunology, Ann and Robert H. Lurie Children's Hospital of Chicago, Chicago, Ill

Funding: This work was supported by funds from Food Allergy Research and Education (FARE).

Conflicts of interest: R. S. Gupta receives research grant support from the National Institutes of Health (NIH), Food Allergy Research and Education (FARE), Stanford Sean N. Parker Center for Allergy Research, UnitedHealth Group, Thermo Fisher Scientific, Genentech, and the National Confectioners Association (NCA); and has served as a medical consultant/advisor for Aimmune Therapeutics, Genentech, Before Brands, Kaléo, DBV Technologies, ICER, DOTS Technology, and FARE. L. A. Bilaver receives research grant support from the NIH, Thermo Fisher Scientific, FARE, Genentech, NCA, and Before Brands Inc. F. H. Wehbe receives research grant support from the NIH, and FARE. J. M. Spergel receives grant support from the NIH, FARE, Genentech, Novartis, Regeneron, and Sanofi; had consultant agreements with DBV Technologies, Genentech, Novartis, Regeneron, Sanofi, and Allakaos. J. Wang receives research grant support from the NIH, Aimmune, DBV Therapeutics, and Regeneron; and consulting fees from ALK Abello, DBV Therapeutics, FARE, and Genentech. C. E. Ciaccio receives research grant support from the NIH, FARE, Paul and Mary Yovovich, and Takeda; and has served as a medical consultant/advisor for Aimmune Therapeutics, Genentech, Novartis, ALK, DBV Technologies, Siolta, Clostrabio, and FARE. M. Makhija receives research funding from Aimmune therapeutics, DBV Technologies, Regeneron Pharmaceuticals, the NIH, and FARE (for this project). J. B. Starren receives research grant support from the NIH, and FARE. The rest of the authors declare that they have no relevant conflicts of interest.

Received for publication September 29, 2021; revised January 28, 2022; accepted for publication February 10, 2022.
Available online ■■

Corresponding author: Ruchi S. Gupta, MD, MPH, Department of Pediatrics, Northwestern University Feinberg School of Medicine, 750 N. Lake Shore Dr., Suite 680 Chicago, IL 60611. E-mail: r-gupta@northwestern.edu.

2213-2198

© 2022 American Academy of Allergy, Asthma & Immunology

https://doi.org/10.1016/j.jaip.2022.02.024

Downloaded for Anonymous User (n/a) at UNIVERSITY OF PENNSYLVANIA from ClinicalKey.com by Elsevier on May 11, 2022. For personal use only. No other uses without permission. Copyright ©2022. Elsevier Inc. All rights reserved.

Abbreviations used
CDM- Common data model
CPT- Current Procedural Terminology
EHR- Electronic health record
FA- Food allergy
FADD- Food Allergy Data Dictionary
FARE- Food Allergy Research and Education
FDC- Food Allergy Data Commons
ICD- International Classification of Disease
OFC- Oral food challenge
OIT- Oral immunotherapy
OMOP- Observational Medical Outcomes Partnership
SNOMED- Systematized Nomenclature of Medicine

forms. © 2022 American Academy of Allergy, Asthma & Immunology (J Allergy Clin Immunol Pract 2022;■:■-■)

Key words: *Food allergy; Data dictionary; Data commons*

INTRODUCTION

Food allergy (FA) is a significant health problem affecting approximately 2%-8% of children[1] and 2%-10% of adults[2] in the United States and has deleterious effects on health-related quality of life.[3] Despite the high prevalence of FAs, they remain poorly understood for several reasons. The wide variety of food triggers that lead to reactions, the heterogeneity of those reactions, the variability in the response to treatment, and the wide variation in FA clinical practice[4] combine to produce a huge number of potential combinations to be evaluated. This, coupled with the partial inheritance, and the interaction of genetic and environmental factors, means that FAs show all the characteristics of a complex trait.[5] As with most complex traits, to better understand the trait, it is critical to disaggregate the heterogeneous group of patients into groups of subphenotypes.[6]

We have learned from other complex traits that defining subphenotypes requires data from much larger numbers of patients than are available at any single institution and that untangling complex traits frequently requires combining data across multiple sites.[7-10] A standard method for doing this is the creation of a data commons[11]—a centralized data repository in which data from disparate sources will be harmonized to a common data standard (Table E1 for a glossary of informatics related terms; available in this article's Online Repository at www.jaci-inpractice.org) with cohort discovery and analytic capabilities. Currently, there is no data commons that can adequately meet the needs of the FA research community. To address this gap, the Food Allergy Research and Education (FARE)[12] has supported the development of an institution-independent Food Allergy Data Commons (FDC) that will provide a single, cloud-hosted repository bringing together both patient-reported and electronic health record (EHR) data from across many institutions. This would enable the FA researchers to access harmonized data that have been curated using a uniform representation, such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), and processed with common data management pipelines, such that data from a variety of institutional and patient-reported sources could be more easily integrated and analyzed together. The FA community would benefit from this cloud-based data science infrastructure that would connect data sets with analytics tools to allow users to share, integrate, and analyze data[11,13] to drive scientific discovery.

A requisite first step in the creation of any data commons is developing a common terminology or data dictionary[14] that defines what data elements will be stored and how that data will be represented. Without this, it is extremely labor intensive, or impossible, to combine data from different sites or to compare across sites. This need for a common terminology may explain why many of the first clinical data commons efforts were in the domain of cancer, which had already invested decades of effort developing uniform reporting standards for cancer registries. Unfortunately for FA researchers, the terminology used in FA is variable and often ambiguous—even the term food allergy is often applied to a variety of food intolerances,[4] with different, incompletely characterized mechanisms. Terminologies to describe allergic conditions, including FA, tend to be complex. The clinical presentation of FA is highly heterogeneous, affecting different organ systems in different patients and severity of allergenic reactions to food ranges from mild rashes to life-threatening anaphylaxis.[4,15,16] In addition, a large fraction of FA documentation—including allergens, reaction description, and recommendation on food avoidances is recorded as free text within the EHR systems, rather than in structured, coded fields.[17] Thus, combining data across institutions currently requires labor-intensive manual chart abstraction or the implementation and tuning of advanced natural language processing systems at each site. A review of existing clinical terminologies revealed that none had adequate coverage of FA data and practice to serve as a Food Allergy Data Dictionary (FADD). Although it leverages existing terminologies, the FADD needed to be developed *de novo*, relying on existing resources and a panel of FA experts. This paper describes that process to develop and validate the first generation of the FADD, as well as more general observations on the difficulties encountered while mapping FA concepts to the current major clinical terminologies.

METHODS

Establishing design criteria

Establishing a core lexicon and ontology for FA concepts was a multistep process. The first step involved establishing design constraints for the overall FDC and the FADD. We leveraged the criteria previously identified for genomic data commons: (1) modular, composed of functional components with well-specified interfaces; (2) community-driven, created by many groups to foster a diversity of ideas; (3) open, developed under open-source licenses that enable extensibility and reuse, with users able to add custom, proprietary modules as needed; and (4) standards-based.[18]

In line with these principles, the following design criteria were established:

1. The FDC would be structured using the OMOP CDM, in order to maximize interoperability with other ongoing national-scale initiatives.[19-23] (Standards-based).
2. The FADD would leverage existing standard codes to the extent possible and only create new codes if existing codes are inadequate. (Standards-based).
3. While the FDC will utilize OMOP, the FDC should be usable with other schemas. (Modular)
4. The FDC and FADD must be able to capture current FA practice in its variant forms, as opposed to only encoding what might be considered best practice. (Community-based)

Downloaded for Anonymous User (n/a) at UNIVERSITY OF PENNSYLVANIA from ClinicalKey.com by Elsevier on May 11, 2022. For personal use only. No other uses without permission. Copyright ©2022. Elsevier Inc. All rights reserved.

TABLE I. High-level FA category description and type of concepts identified under each category

Serial No.	Category	Description	Type of concepts
1	Events	This concept encompasses occurrences/incidences that we want to know if they happened and when they happened.	Types of clinical encounters—initial visit, follow-up visit, ICU admission, ER/ED visit, FA reaction, disease exacerbations, etc.
2	Medications*	This concept includes, but may not be limited to, current and past medications prescribed by physicians, over-the-counter medicines, and drugs administered as part of in-office procedures, as for an OFC reaction.	Inhalation medications, nasal medications, oral medications, injections, etc.
3	Formal Diagnoses	This concept includes standard medical conditions primarily responsible for the patient's need for treatment or investigation. These are formal diagnoses as opposed to signs and symptoms or ill-defined reactions. It can encompass conditions as admitting diagnosis, final diagnosis, or preliminary diagnosis.	Peanut allergy, pollen-FA syndrome, eosinophilic esophagitis, asthma, etc.
4	Triggers	This concept comprises triggering factors for food-allergic reactions as well as triggers for non−food-allergic conditions including drug allergies. These can represent attributes of a diagnosis or a reaction, challenge food, or therapy food.	Food triggers—egg, cow's milk, peanut, almond, etc. Environmental triggers—animal dander, pollen, tree, weed, etc.
5	Clinical Trials	This concept identifies if a patient is or was enrolled in an FA clinical trial registered with CT.gov, including therapeutic trials, diagnostic trials, combination trials.	Patient enrollment in an FA clinical trial—yes/no.
6.	OIT	This concept consists of the following 4 phases of OIT: (1) an OIT initiation phase, (2) an up-dosing visit or escalation phase, (3) a dose-maintenance phase, (4) OIT completion phase/stop OIT.	Initiation, OIT up-dosing phase, ongoing maintenance, OIT completion, including reason for discontinuation.
7.	Reactions	Clinical manifestations of FA reactions that include signs and symptoms and other reaction attributes.	
7.1	Signs and Symptoms	This concept domain outlines objective and subjective clinical manifestations of FA, including objective, observable evidence indicating possible FA reaction observed by an allergist, and subjective complaints reported by the patient. It also includes physical examination findings.	Signs—rash, erythema, etc. Symptoms—nausea, oral pruritus, etc.
7.2	Other Reaction Attributes	This concept covers other reaction characteristics, as amount of food that triggered the reaction, temporal relation, type of exposure, exposure mode, and environmental location.	Type of exposure—accidental, intentional, etc. Exposure mode—ingestion, contact, inhalation. Location of reaction—home, school, etc.
8	Procedures	This concept encompasses processes ordered by a health care provider that typically have a diagnostic or therapeutic value. It is subdivided into FA diagnostic procedures and other codable procedures performed on a patient.	
8.1	FA Diagnostic Procedures	This concept domain includes physician orders/medical procedures performed to diagnose FA, or to determine course of treatment and monitor the disease.	Skin prick test, blood testing, OFC, contact challenge, inhalation challenge.
8.2	Other Procedures	This concept includes other relevant codable procedures ordered for a patient with FA or to assess any comorbid conditions.	Laboratory tests, endoscopy, etc.
9	Therapeutic Plan	This concept domain focuses on education for patient and family on allergen avoidance, dietary recommendations, and recommendations for follow-up, or referral to other specialists.	FA recommendations, including dietary recommendations, anaphylaxis management plan/emergency action plan, etc.
10	Family History	This concept includes record of health information about a person's close relatives, including parents, brothers, and sisters.	Family history of FA.
11.	History	The history concept captures person's breast-feeding history and personal history.	Breast-feeding history, personal history of other diseases.
12.	Other Observations and Measurements	Observations are outcomes that are routinely collected as part of clinical care but are not specifically tied to a reaction event. Measurements include structured values of a standardized examination or testing performed on a person. Other observations capture economic, environmental, and psychosocial determinants that influence diverse dimensions of FA including disease development, treatment, management, and quality of life.	Other observations—environmental factors, economic determinants, dietary preferences. Measurements—vitals, test result values, such as IgE.

(continued)

Downloaded for Anonymous User (n/a) at UNIVERSITY OF PENNSYLVANIA from ClinicalKey.com by Elsevier on May 11, 2022. For personal use only. No other uses without permission. Copyright ©2022. Elsevier Inc. All rights reserved.

TABLE I. *(Continued)*

Serial No.	Category	Description	Type of concepts
13	Person	The person concept contains records that uniquely identify each patient in the source data.	Race, Data Commons ID, location, external identifiers, etc.
14	Provider	The provider table contains a list of uniquely identified health care providers, including physicians, nurses, behavioral therapists, etc.	NPI, provider specialty, care site, etc.

ED, Emergency department; *ER,* emergency room; *FDA,* U.S. Food and Drug Administration; *ICU,* intensive care unit; *IgE,* immunoglobulin E; *NPI,* National Provider Identifier; *OFC,* oral food challenge.
*Although the FADD medications list encompasses most frequently prescribed medicines in relation to FA and coexisting diseases, other FDA approved medications can also be stored. The OMOP can store any medication that has an Rx norm code irrespective of indication.

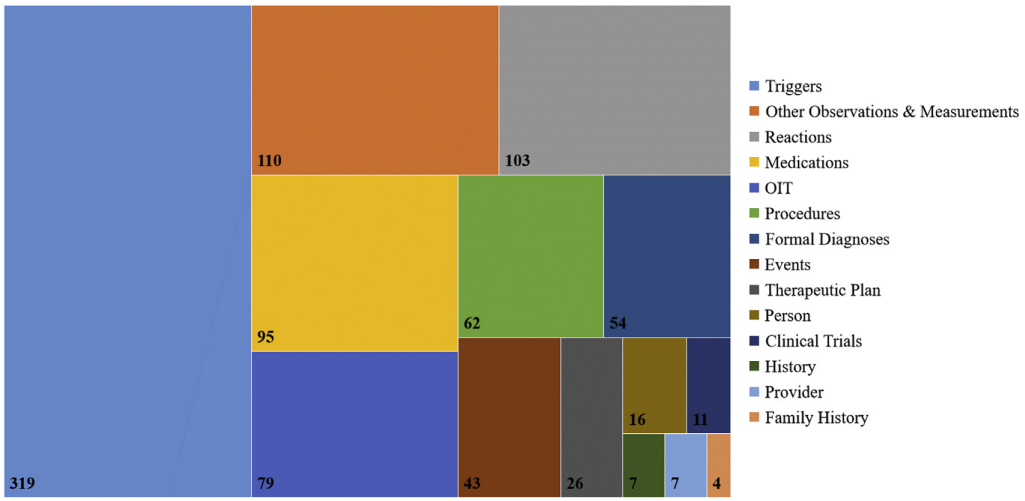

FIGURE 1. The high-level FA categories (n = 14) that have been included in v1.0 of the FADD and the number of concepts under each category. The categories have been arranged in descending order based on the number of concepts under each high-level category.

5. The coverage of the FADD needs to support patient-entered, EHR, and clinical research FA data. (Community-driven)
6. The FADD will focus on FA-specific concepts to leverage standard OMOP coding for nonspecific clinical data (eg, physical examination, medication, laboratory). (Standards-based)

Analysis and development phase

Resource collection, analysis, and expert interviews. The first step toward FADD development was collecting resources from four academic medical centers with expertise in FA (Ann & Robert H. Lurie Children's Hospital of Chicago, Children's Hospital of Philadelphia, The University of Chicago, and Jaffe Food Allergy Institute at Icahn School of Medicine at Mount Sinai). Resources were collected between January 2020 and March 2020. This was followed by a systematic review and manual comparative analysis of the collected clinical documentation tools (ie, patient intake forms for allergy/immunology, templates for documenting allergy visits, clinical notes, and precompiled Epic EHR (Epic Systems Corporation, Verona, Wis) phrases used across these institutions) to identify the concepts and to determine the types of expressions used in clinical practice.[24] In-depth discussions with domain experts from these institutions were conducted, both in-person and virtually, to understand the nuances of current FA clinical workflow and concept prioritization. Overall, 16 documents from these 4 institutions were evaluated over a period of 2 months to compile a comprehensive list of FA concepts (entity/attribute names). After compiling all concepts, common concepts were identified. Next, high-level FA categories were established and defined in a clear and unambiguous manner, and the concepts were collated to develop a candidate ontology (v0.8) that was then reviewed for their uniqueness according to semantic similarity by domain experts.

Face validation phase

Stage I. To evaluate the importance of the categories and concepts in clinical practice and research, the candidate ontology was sequentially evaluated with the consortium over a series of iterations. The consortium met virtually every 2 weeks for 3 months. The goal of the hourly virtual review meetings (12 meetings in total) was to discuss and solicit comments on the candidate ontology. The recommendations of the working group were compiled, and the FADD was reevaluated to look for gaps and remove duplications.

Stage II. During the second stage of this phase, the revised FADD document was disseminated and vetted with FA experts across 22 academic FA centers and 3 industry partners for feedback and review. Prior to dissemination, 2 virtual sessions were organized to help reviewers from these organizations understand the structure and planned function of the FADD. The comments received during this stage were reviewed again with the consortium and revisions with the

Downloaded for Anonymous User (n/a) at UNIVERSITY OF PENNSYLVANIA from ClinicalKey.com by Elsevier on May 11, 2022. For personal use only. No other uses without permission. Copyright ©2022. Elsevier Inc. All rights reserved.

highest agreement were incorporated to create version 1.0 of the FADD.

Evaluation of existing terminologies

At version 0.8 of the FADD, concepts were described in text. Each of these concepts was mapped to existing clinical terminologies manually with the ATHENA vocabulary browser.[25] After completion of version 1.0, the process was repeated for new or modified concepts.

RESULTS

Overview

A consensus on the FADD elements was achieved in November 2020 and included 936 distinct FA concepts. A total of 14 FA categories were identified: Events, Medications, Formal Diagnoses, Triggers, Clinical Trials, Oral Immunotherapy (OIT), Reactions (including signs and symptoms, and other reaction attributes), Procedures (including FA diagnostic procedures, and other procedures), Therapeutic Plan, Family History, History, Other Observations and Measurements (eg, dietary preferences, environmental factors), Person, and Provider (Table I).

Data dictionary structure. The scope of each category was iteratively defined during the development of the FADD. During the development phase, the top-level categories and concepts under each category were drawn from terms frequently used in FA clinical practice, laboratory, pharmacy, radiology, and billing systems. Categories varied in size from the largest, Other Observations and Measurements, which contained 110 concepts, to Family History, which contained 4 concepts (Figure 1). The documentation and usage of certain FA concepts and data elements varied considerably across the academic medical centers, such as the concept of severity of clinical manifestations, regional differences in the prevalence of specific types of food allergies. Thus, the concepts that reached maximum agreement were included in the FADD. Many of the concepts developed from expert opinion were added to the FADD during the face validation phases. Several new data elements related to Formal Diagnoses, Triggers, OIT, Reactions, and Procedures were included as these were deemed important during the decision-making process by our expert panel.

The FADD version 1.0[26] presents the critical concepts to be included in the FDC. Each concept domain discussed in the FADD has a description followed by a tabular representation of entity-attribute relation (Figure 2). These relationships can be hierarchical (parent-child or is-a) or nonhierarchical (is a component-of). The first column in each table of the FADD document represents a unique identification number corresponding to the data element.

Mapping to existing terminologies. Once common data elements were established, existing ontologies including Systematized Nomenclature of Medicine (SNOMED), RxNorm, International Classification of Disease (ICD)-10, Current Procedural Terminology (CPT), Logical Observation Identifiers Names and Codes, and FoodOn, were evaluated for their coverage of the identified FA terminology. The SNOMED, a comprehensive, hierarchical terminology used for clinical documentation and reporting,[27] was examined to gain a better understanding of FA documentation. The RxNorm, a standardized nomenclature for clinical drugs and drug delivery devices, provides a near-complete coding of drugs and medications, both prescription and generic, available on the U.S. market. The ICD-10, a medical classification list maintained by the World Health Organization, contains codes for symptoms, diseases, and other findings and is the most commonly employed diagnostic and reimbursement codes by allergists. The CPT includes standardized codes and terms to code procedures for both medical records and insurance claims.[28] Of the commonly used ontologies that we evaluated, SNOMED provided the most comprehensive coverage of the treatment process and RxNorm covered most medications and FA extracts. The Logical Observation Identifiers Names and Codes and ICD-10 fell behind SNOMED in overall content coverage. For example, ICD codes do not support most specific FA diagnoses, such as allergy to hazelnut. The terminology mapping revealed that that a major area of noncoverage was with respect to foods and FAs. For example, when egg or milk is heated, the protein denatures and exposes different antigens. As a result, food allergists routinely distinguish between milk, to indicate undenatured, and baked milk to indicate denatured protein. Further, a large proportion of patients with a milk allergy tolerate baked milk.[29,30] We were unable to find any clinical vocabulary that contained a concept for baked milk or baked egg. Investigation was also conducted of dedicated food databases, such as FoodOn,[31] a controlled vocabulary representing entities that bear a food role and encompasses materials in natural ecosystems and food webs as well as human-centric categorization of food. None of the food databases evaluated captured the immunologically distinct food forms such as baked egg.

Another area of limited coverage was in the area of FA-specific procedures. As an example, SNOMED uses oral food challenge for the first stage of OIT because it lacks a code for OIT. Codes for specific FAs in SNOMED are created through precoordination (creating a new concept by combining 2 existing concepts) of the FA concept (SNOMED 4188027) with the code for a food. Currently, 79% of the concepts (n = 744) in the FADD are codable within the existing OMOP structure (Table II). Concept areas that are currently not codable include management of FA in the school setting, concepts related to OIT phases, and psychosocial impact of FA on the patient and caregivers.

We also uncovered cases in which, even if concept codes exist, they are used in ways that are in conflict with our current understanding of FAs. Whereas the choice of OMOP CDM for the data schema greatly increases interoperability, the storage of some classes of concepts is not optimized for the FA domain. For example, food allergists distinguish between a reaction to a food, which may be caused by several mechanisms and a diagnosed FA, which requires immunological diagnostic confirmation. However, because reaction to a food is often listed as an admitting diagnosis, codes in the food reaction hierarchy are stored in the diagnosis table, whereas concepts in the FA diagnosis hierarchy are currently stored in the observation table.

DISCUSSION

As any practicing allergist knows, FA clinical documentation and coding is inherently complex because the triggering allergen as well as the resulting allergic reaction need to be adequately represented, including clinical manifestations and severity. Further, combining data from multiple institutions is challenging

Downloaded for Anonymous User (n/a) at UNIVERSITY OF PENNSYLVANIA from ClinicalKey.com by Elsevier on May 11, 2022. For personal use only. No other uses without permission. Copyright ©2022. Elsevier Inc. All rights reserved.

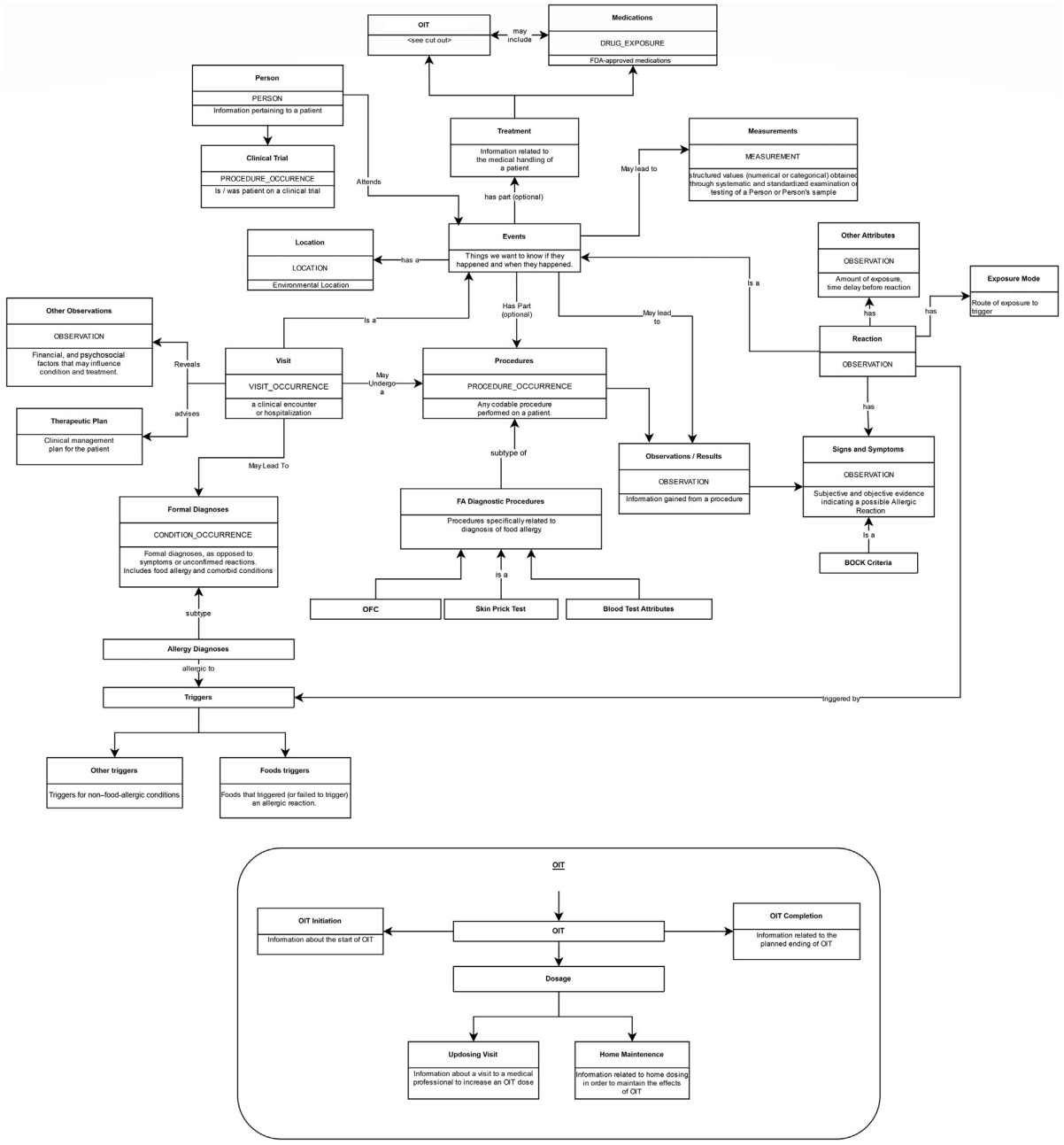

FIGURE 2. Entity relationship diagram details the interactions and workflow between the various concepts identified in the FADD v1.0. FDA, U.S. Food and Drug Administration.

but can be greatly enhanced using standardized language and explicit definitions of data categories.[24,32] We developed a comprehensive FADD consisting of 936 concepts laying the foundation of the FDC project. Although there have been other efforts to improve allergy terminology, such as the recent work to revise the ICD-11 allergy and hypersensitivity diagnosis hierarchy,[33,34] we believe that this is the first attempt to develop a general FADD as part of a standardized, pragmatic data collection tool for research in FA patients.

Because the FDC will ultimately be a shared community resource, engaging FA stakeholders early in the design and development of the FADD process was crucial for defining scope and creating standards to maximize the acceptance and usability of the FADD as well as the FDC, a principle that has been previously utilized in designing other cloud-based platforms,[9] as well as recent initiatives.[35] Further, because the majority of the existing FA data is from a patient's self-report, text-narratives within the EHRs and clinical research data, the FADD

Downloaded for Anonymous User (n/a) at UNIVERSITY OF PENNSYLVANIA from ClinicalKey.com by Elsevier on May 11, 2022. For personal use only. No other uses without permission. Copyright ©2022. Elsevier Inc. All rights reserved.

TABLE II. Current status of category mapping to existing terminologies

Category	Successfully mapped (within the existing OMOP structure)	Total
Events	37	43
Medications	95	95
Formal Diagnoses	54	54
Triggers	315	319
Clinical Trials	3	11
OIT	NA	79
Reactions	92	103
Procedures	38	62
Therapeutic Plan	12	26
Family History	4	4
History	4	7
Other Observations and Measurements	67	110
Person	16	16
Provider	7	7
Total	744	936

NA, Not available.

incorporates FA data elements that will adequately support patient-entered data from FARE's patient registry, physician-entered EHR data, and clinical research data (ongoing effort).

The creation of the FADD can lay the foundation for many advances in FA practice and research. The FADD provides the basis for creation of improved structured data entry forms for FA, which can provide for uniformity and standardization of clinical data entry, and use in templated notes for the EHR more easier.[36] Previously, data dictionaries have been used to build standardized data entry forms in Epic EHR for several conditions, including pediatric epilepsy[37,38] and sickle cell disease.[39] Forms built using the FADD (currently under development) will facilitate best practices in FA data collection; clinical data transport among FA practices in coded form; and integration of data across multiple practices. Improving the granularity of data elements can facilitate complex and meaningful data queries to define FA phenotypes and cohorts for both research and quality improvement. The FADD is even more important for the creation of automated data collection directly from patients. Whereas the training of physicians ensures a level of uniformity in free text clinical notes, the same is not true of patients. The existence of a FADD can enable the creation of automated patient-directed FA history collection tools. Finally, the concepts in the FADD can provide a uniform variable set for FA clinical trials, enabling greater comparison across sites and across studies.

As noted previously, keeping in mind the challenges in integrating multiple reference terminologies for encoding FA information, we chose the OMOP data model because of its wide adoption and support for systematic analysis of disparate observational databases.[40] It also accommodates both administrative claims and EHR data, allowing users to generate evidence from a wide variety of sources.

A critical factor in the selection of the OMOP CDM is the fact that it already supports collection of a wide variety of EHR data, which enabled this project to focus on expanding the FA-specific coverage. In fact, the Observational Health Data Sciences and Informatics (OHDSI) network, [41] which oversees the OMOP CDM, includes EHR data on over 600 million patients. Thus, collection of other EHR data, such as medications and comorbidities, has already been addressed. Although not all concepts in the FADD can be currently mapped to the corresponding OMOP vocabulary, the model does allow the use of local codes to circumvent these limitations. We are working with the OHDSI team to add necessary codes to OMOP terminology so that concepts recorded within the clinical documentation can be adequately represented in the data dictionary. For example, many FAs (ie, the combination concepts for FA and a specific food) had not yet been created. Similarly, we are working to move FA diagnoses to their appropriate location in the condition (ie, diagnosis) table of OMOP. Previously, OIT was handled using the oral food challenge procedure, although this makes it difficult to differentiate between the 2 different procedures. The OMOP has already agreed to add a distinct OIT procedure concept, and we are working with them to cover the other aspects of this procedure.

Limitations

By design, the FADD does not incorporate every concept necessary to encode the entire medical history of an FA patient. Creating the complete picture requires utilizing not only the FADD but also other OMOP concepts. Although FAs and FA procedures can be mapped to existing ICD and CPT codes, both coding systems lack the granularity needed to drive both clinical care and research. Proposing specific additions to ICD and CPT is beyond the scope of this study and deferred to future work. The FADD has been developed through analysis of existing clinical documentation tools and in-depth discussion with experts from 4 leading FA academic centers and vetted with 22 additional academic centers; thus, the current version may not be generalizable to the entire FA community. Similarly, the FADD concepts have not yet been implemented into clinical data collection forms (an ongoing effort) and tested in real-world settings. In addition, v1.0 of the FADD focused on clinical concepts rather than FA research concepts. We fully anticipate that additional concepts need to be added to the FADD over time, driven both by changes in FA practice as well as by the evolution of other clinical terminologies.[34]

Of course, creation of a uniform terminology in a clinical domain does not obviate the well-described limitations of EHR data for research. A full discussion of this topic is beyond the scope of this manuscript, but we will highlight a few points. First, missing clinical documentation, the presence of a term for an FA condition or finding does guarantee that the finding will be documented. Absence of the code may not indicate absence of the finding. In designing EHR forms based on the FDC, considerable effort has focused on cases in which clinicians are prompted to explicitly record the absence of a particular finding (eg, "There was no shortness of breath"). Imperfect clinical documentation practices[42] impact the quality of data in EHR and may result in selection bias. Second, gaps in a patient's record may be a result of loss to follow-up or transition to another care provider or insurer. Third, in domains like FA that often involve referrals to different practices or health care systems, deduplication of records (and the elimination of double counting) can present a significant challenge.[43] Notwithstanding, the creation of a uniform and adequately granular terminology in a clinical domain is an important enabler of multisite research.

Downloaded for Anonymous User (n/a) at UNIVERSITY OF PENNSYLVANIA from ClinicalKey.com by Elsevier on May 11, 2022. For personal use only. No other uses without permission. Copyright ©2022. Elsevier Inc. All rights reserved.

The FADD is available as a human-readable presentation of the critical FA concepts.[26] A separate coding guide[44] maps the FADD concepts to existing codes and coding schemes, where matching codes exist. The coding guide facilitates creation of an OMOP-compliant database for FA data by providing information on the preferred OMOP codes to use for each concept, as well as outline rules defining permitted values for every field of the FADD.

Both FA research and practice have been hampered by the inability to collect FA data in a structured and institution-independent manner. This has been compounded by the limited coverage of FA concepts by existing clinical terminologies. The FADD is a critical first step in addressing this problem. It defines critical FA concepts and common data elements. During the second (ongoing) phase of the FDC, the dictionary will serve as a central resource for designing structured data entry forms to capture FA clinical encounters.

REFERENCES

1. Gupta RS, Warren CM, Smith BM, Blumenstock JA, Jiang J, Davis MM, et al. The public health impact of parent-reported childhood food allergies in the United States. Pediatrics 2018;142:e20181235.
2. Gupta RS, Warren CM, Smith BM, Jiang J, Blumenstock JA, Davis MM, et al. Prevalence and severity of food allergies among US adults. JAMA Netw Open 2019;2:e185630.
3. Walkner M, Warren C, Gupta RS. Quality of life in food allergy patients and their families. Pediatr Clin North Am 2015;62:1453-61.
4. NIAID-Sponsored Expert PanelBoyce JA, Assa'ad A, Burks AW, Jones SM, Sampson HA, et al. Guidelines for the diagnosis and management of food allergy in the United States: report of the NIAID-sponsored expert panel. J Allergy Clin Immunol 2010;126:S1-58.
5. Tsai HJ, Kumar R, Pongracic J, Liu X, Story R, Yu Y, et al. Familial aggregation of food allergy and sensitization to food allergens: a family-based study. Clin Exp Allergy 2009;39:101-9.
6. Morris AP, Lindgren CM, Zeggini E, Timpson NJ, Frayling TM, Hattersley AT, et al. A powerful approach to sub-phenotype analysis in population-based genetic association studies. Genet Epidemiol 2010;34:335-43.
7. Delaney SK, Hultner ML, Jacob HJ, Ledbetter DH, McCarthy JJ, Ball M, et al. Toward clinical genomics in everyday medicine: perspectives and recommendations. Expert Rev Mol Diagn 2016;16:521-32.
8. Rehm HL, Berg JS, Brooks L, Bustamante CD, Evans JP, Landrum MJ, et al. ClinGen: the Clinical Genome Resource. N Engl J Med 2015;372:2235-42.
9. Volchenboum SL, Cox SM, Heath A, Resnick A, Cohn SL, Grossman R. Data commons to support pediatric cancer research. *Am* Soc Clin Oncol Educ Book 2017;37:746-52.
10. Major A, Cox SM, Volchenboum SL. Using big data in pediatric oncology: current applications and future directions. Semin Oncol 2020;47:56-64.
11. Grossman RL, Heath A, Murphy M, Patterson M, Wells W. A case for data commons: toward data science as a service. Comput Sci Eng 2016;18:10-20.
12. Food Allergy Research and Education. Accessed May 20, 2021. https://www.foodallergy.org/
13. Zhang GQ, Cui L, Mueller R, Tao S, Kim M, Rueschman M, et al. The National Sleep Research Resource: towards a sleep data commons. J Am Med Inform Assoc 2018;25:1351-8.
14. Grossman RL. Progress toward cancer data ecosystems. Cancer J 2018;24:126-30.
15. Vogel NM, Katz HT, Lopez R, Lang DM. Food allergy is associated with potentially fatal childhood asthma. J Asthma 2008;45:862-6.
16. Bock SA, Munoz-Furlong A, Sampson HA. Fatalities due to anaphylactic reactions to foods. J Allergy Clin Immunol 2001;107:191-3.
17. Zimmerman CR, Chaffee BW, Lazarou J, Gingrich CA, Russell CL, Galbraith M, et al. Maintaining the enterprisewide continuity and interoperability of patient allergy data. Am J Health Syst Pharm 2009;66:671-9.
18. Denny J, Glazer D, Grossman RL, Paten B, Philippakis A. A Data Biosphere for Biomedical Research. Accessed May 20, 2021. https://medium.com/@benedictpaten/a-data-biosphere-for-biomedical-research-d212bbfae95d
19. Klann JG, Joss M, Embree K, Murphy SN. Data model harmonization for the All Of Us Research Program: Transforming i2b2 data into the OMOP common data model. PLoS One 2019;14:e0212463.
20. Lemke AA, Wu JT, Waudby C, Pulley J, Somkin CP, Trinidad SB. Community engagement in biobanking: experiences from the eMERGE network. Genomics Soc Policy 2010;6:35-52.
21. Fleurence RL, Curtis LH, Califf RM, Platt R, Selby JV, Brown JS. Launching PCORnet, a national patient-centered clinical research network. J Am Med Inform Assoc 2014;2:578-82.
22. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-8.
23. Haendel MA, Chute CG, Bennett TD, Eichmann DA, Guinney J, Kibbe WA, et al. The National COVID Cohort Collaborative (N3C): rationale, design, infrastructure, and deployment. J Am Med Inform Assoc 2021;28:427-43.
24. Tanno LK, Chalmers R, Jacob R, Kostanjsek N, Bierrenbach AL, Martin B, et al. Global implementation of the World Health Organization's International Classification of Diseases (ICD)-11: the allergic and hypersensitivity conditions model. Allergy 2020;75:2206-18.
25. Athena. Observational Health Data Sciences and Informatics − OHDSI. Accessed May 15, 2021. https://athena.ohdsi.org/
26. DigitalHub- Northwestern University. Food Allergy Data Dictionary. Accessed August 8, 2021. https://doi.org/10.18131/g3-9m7w-m310
27. Donnelly K. SNOMED-CT: the advanced terminology and coding system for eHealth. Stud Health Technol Inform 2006;121:279-90.
28. Dotson P. CPT codes: what are they, why are they necessary, and how are they developed? Adv Wound Care (New Rochelle) 2013;2:583-7.
29. Flom JD, Sicherer SH. Epidemiology of cow's milk allergy. Nutrients 2019;11:1051.
30. Upton J, Nowak-Wegrzyn A. The impact of baked egg and baked milk diets on IgE- and non-IgE-mediated allergy. Clin Rev Allergy Immunol 2018;55(2):118-38.
31. FoodOn Consortium. FoodOn: A Farm to Fork Ontology. Accessed July 14, 2020. https://foodon.org/
32. Ziegler P, Dittrich KR. Three decades of data integration—all problems solved?. In: *Building the Information Society*. 18th IFIP World Computer Congress (WCC 2004), 12. Toulouse, France: Kluwer; 2004. p. 3-12.
33. Tanno LK, Calderon MA, Papadopoulos NG, Sanchez-Borges M, Moon HB, Sisul JC, et al. Surveying the new allergic and hypersensitivity conditions chapter of the International Classification of Diseases (ICD)-11. Allergy 2016;71:1235-40.
34. Tanno LK, Sublett JL, Meadows JA, Calderon M, Gross GN, Casale T, et al. Perspectives on the International Classification of Diseases, 11th Revision, developments in allergy clinical practice in the United States. Ann Allergy Asthma Immunol 2017;118:127-32.
35. National Cancer Institute. Childhood Cancer Data Initiative (CCDI). Updated May 24, 2019. Accessed June 14, 2021. https://www.cancer.gov/research/areas/childhood/childhood-cancer-data-initiative
36. Bleeker SE, Derksen-Lubsen G, van Ginneken AM, van der Lei J, Moll HA. Structured data entry for narrative data in a broad specialty: patient history and physical examination in pediatrics. BMC Med Inform Decis Mak 2006;6:29.
37. Grinspan ZM, Patel AD, Shellhaas RA, Berg AT, Axeen ET, Bolton J, et al. Pediatric Epilepsy Learning Healthcare System. Design and implementation of electronic health record common data elements for pediatric epilepsy: foundations for a learning health care system. Epilepsia 2021;62:198-216.
38. Fitzsimons M, Hwang H. Creating the conditions for a learning epilepsy care system. Epilepsia 2021;62(1):217-9.
39. Miller R, Coyne E, Crowgey EL, Eckrich D, Myers JC, Villanueva R, et al. Implementation of a learning healthcare system for sickle cell disease. JAMIA Open 2020;3:349-59.
40. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 2012;19:54-60.
41. Observational Health Data Sciences and Informatics (OHDSI). Accessed August 8, 2021. https://www.ohdsi.org/
42. Gianfrancesco MA, Goldstein ND. A narrative review on the validity of electronic health record-based research in epidemiology. BMC Med Res Methodol 2021;21:234.
43. Kho AN, Cashy JP, Jackson KL, Pah AR, Goel S, Boehnke J, et al. Design and implementation of a privacy preserving electronic health record linkage tool in Chicago. J Am Med Inform Assoc 2015;22:1072-80.
44. Northwestern University, Feinberg School of Medicine. Food Allergy Data Dictionary Coding Guide. Accessed September 13, 2021. https://doi.org/10.18131/g3-q5cy-f023

Downloaded for Anonymous User (n/a) at UNIVERSITY OF PENNSYLVANIA from ClinicalKey.com by Elsevier on May 11, 2022. For personal use only. No other uses without permission. Copyright ©2022. Elsevier Inc. All rights reserved.

ONLINE REPOSITORY

TABLE E1. Brief description of informatics related terms appearing in the manuscript text

Term	Description
Cohort discovery	Cohort discovery may be defined as identification of a population or a set of persons who satisfy 1 or more inclusion criteria for a duration of time, for example, identifying patient populations with certain health care interventions (eg, drug exposure, procedures) and outcomes (eg, conditions, procedures, other drug exposures).
Common data model (CDM)	A convention for representing health care data that allows portability of analysis (the same analysis unmodified can be executed on multiple datasets). The CDM, combined with its standardized content, will ensure that research methods can be systematically applied to produce meaningfully comparable and reproducible results.
Common data standard	A common data standard ensures that data from multiple disparate sources is harmonized, allowing a standardized analytic to be executed on the data.
Data management pipeline	A data management pipeline refers to a series of computer programs that is designed specifically to compose and execute a series of computational or data manipulation steps, to transform or analyze a particular type of data.
RxNorm	It is a standardized nomenclature for clinical drugs (generic and branded drugs), produced by the National Library of Medicine and supports semantic interoperation between drug terminologies and pharmacy knowledge base systems.

Downloaded for Anonymous User (n/a) at UNIVERSITY OF PENNSYLVANIA from ClinicalKey.com by Elsevier on May 11, 2022. For personal use only. No other uses without permission. Copyright ©2022. Elsevier Inc. All rights reserved.