Formalizing the Representation of Immune Exposures for Human Immunology Studies

Randi Vita[1], James A. Overton[1], Kei-Hoi Cheung[2],
Steven H. Kleinstein[3], Bjoern Peters[1]

[1]Division for Vaccine Discovery
La Jolla Institute for Allergy and Immunology,
La Jolla, California, U.S.A.

[2]Department of Emergency Medicine and Yale
Center for Medical Informatics, Yale School of
Medicine, New Haven, CT, USA.

[3]Department of Pathology, Yale School of Medicine,
New Haven, Connecticut, U.S.A. and
Interdepartmental Program in Computational
Biology and Bioinformatics, Yale University, New
Haven, Connecticut, USA

***Abstract*—Human immunology studies typically examine how immune exposures associated with vaccinations, infectious, allergic or autoimmune diseases, or transplantations perturb the immune system with the goal to develop diagnostic tools and therapeutic interventions. While there are established approaches to formally represent the experimental data generated in such studies, which often comprises gene expression data, flow cytometry data, or serology data, the description of the immune exposures themselves is not well standardized. We here present a formal approach to represent immune exposures at a high level of granularity. We capture the exposure process (e.g. 'vaccination' or 'occurrence of allergic disease'), exposure material (e.g. 'Tdap vaccine' or 'House dust mite'), and the associated disease name and stage (e.g. 'allergic rhinitis' and 'chronic'). This representation scheme has been used successfully in the IEDB and an extended version has been adopted by HIPC to capture studies in ImmPort. We are reporting here on this scheme, our ongoing attempts to map the terms used to existing ontologies, and the challenges encountered.**

Keywords—immune exposure; modeling; HIPC; ontology

I. Introduction

The Immunology Database and Analysis Portal (ImmPort) [1] is the primary resource to capture human immunology studies funded by the National Institute of Health, Division of Allergy, Immunology and Transplantation. ImmPort provides structured data fields to capture a variety of different experimental data and free-text fields to store meta-data on cohorts from which subjects where recruited. This free-text cohort description data typically contains a description of immune exposures that are expected to perturb the immune system. While free-text allows for a detailed account how a given study is conducted and a cohort is defined, without standardization, such descriptions are difficult to query and compare across many studies in a large database such as ImmPort.

In particular, ImmPort is the designated repository for data from studies performed by the Human Immunology Project Consortium (HIPC) [2], a collaboration between a number of centers aimed at performing large scale human immunology studies with a focus on profiling the human immune response to natural infection and vaccination. A key goal of the HIPC consortium is to cross-compare results from different centers. To facilitate this, we set out to develop a standardized representation of immune exposures for HIPC studies that can be stored in ImmPort to represent their central elements in a structured format.

The need to represent immune exposures extends beyond the HIPC program. Most human immunology studies examine how the immune system responds to perturbations. Subjects are compared across cohorts and/or at defined time points that are intended to isolate the effect of immune exposures. The Immune Epitope Database (IEDB) [3] implemented a structured representation of immune exposures that has been applied to model over one million experiments in which human samples were tested for T cell or B cell reactivity to specific epitopes. The IEDB representation of exposures is decoupled from the epitope mapping experiments, so we decided to test if it could be utilized as a basis to describe immune exposures for the HIPC program. By adapting the IEDB model for HIPC, we have developed an even more general representation of immune exposures that can be used by the wider scientific community.

II. APPROACH

A. *Semi-formal Immune Exposure Representation*

All HIPC centers funded by the middle of 2017 were asked to supply textual descriptions of study designs that they planned on submitting to ImmPort. We then examined the immune exposures that were part of these study designs and how they would be entered into the IEDB format. As a result of this process, we found that the broader scope of HIPC compared to the IEDB required extension of the IEDB structured representation. In the following, we present the resulting expanded schema to represent immune exposures for HIPC, of which the IEDB immune exposures are a subset. This schema has been implemented by adding columns to the

'Human Subject Template' spreadsheet that is used to submit information to ImmPort.

We consider four elements critical to the description of an immune exposure, as listed as the column headers in Table I. The 'Exposure process' identifies the type of process through which a host was exposed and the type of evidence for that exposure to have happened, which are tightly intertwined. This is the only element of the four that was deemed mandatory. Based on the choice made for 'Exposure process', other elements are required or not applicable as listed in Table I. The 'Exposure material' describes what substance(s) the host was exposed to and/or developed immune reactions to as part of the exposure process. The 'Disease name' indicates the specific disease of the host associated with the exposure being described and lastly, the 'Disease stage' provides a broad classification of how the disease progressed at the time of the study.

Exposure process	Exposure material	Disease name	Disease stage
administration	required	X	X
vaccination	required	X	X
infectious challenge	required	optional	X
transplant/transfusion	required	X	X
disease	X	required	required
infectious disease	required	required	required
allergic disease	required	required	required
autoimmune disease	X	required	required
cancer	X	required	required
exposure(without disease)	required	X	X
asymptomatic infection/colonization	required	X	X
exposure with immune reactivity	required	X	X
exposure with documentation	required	X	X
exposure to endemic/ubiquitous agent	required	X	X
no exposure	X	X	X
unknown	X	X	X

TABLE I. Four structured elements to describe immune exposures.

To illustrate how this representation was used in practice, Table II shows three examples of studies by actual HIPC centers that involved immune exposures, described in free text (first column to the left), and how these were modeled using the four elements of the exposure scheme (columns to the right). These examples illustrate the three main types of exposure processes, namely 'administration', 'disease', and 'exposure without disease'.

Free-text description of immune exposure	Exposure process	Exposure material	Disease name	Disease stage
"Adults receiving a Varicella-zoster shot"	vaccination	Varicella-zoster virus vaccine (VO:0000669)		
"Hospitalized patients with Hemorraghic Dengue fever"	infectious disease	Dengue virus (NCBITaxon:12637)	Dengue hemorrhagic fever (DOID:12206)	acute / recent onset
"Subjects from endemic area that tested positive for antibodies against Dengue 2 based on serology"	exposure with immune reactivity	Dengue virus 2 (NCBITaxon:11060)		

TABLE II. Three examples of immune exposures modeled in this schema.

Thus, "Adults receiving a Varicella-zoster shot" would be the result of a vaccination 'Exposure process' which delivered the 'Exposure material' that was the Varicella-zoster virus vaccine. No disease resulted from this immune exposure.

B. *Ontology Mapping*

Our intent is to map each of the four data elements described above to ontology terms with textual and logical definitions, ideally derived from established ontologies covering the various domain. For 'Exposure process', all allowed values are listed in the first column of Table I. This collection of options has been assembled by the IEDB team over the past 13 years and has been proven to be robust and stable, with minimal modifications occurring in the last 5 years. Each of the options come with a definition and rules when it should be applied. These terms will be mapped to formal external ontology terms, as initiated in Supplementary Table S1 (https://doi.org/10.6084/m9.figshare.6741791.v1). The main challenge in this process is that terms for e.g. 'vaccination', 'infectious disease' and 'transplantation' come from different external ontologies, and presenting users their definitions side-by-side is not helpful. We are planning to engage representatives of different ontology communities, and harmonize their definitions. Until this is done, we proceeded with implementation of temporary terms for this immune exposure model in ONTIE [5], which we intend on replacing/merging with new or edited terms in the appropriate external ontologies.

In addition to the main three categories of immune exposure (administration, disease, exposure without disease) and their subtypes, there are two options (no exposure and unknown) which are not actual types of exposures but rather values to signify two different reasons why it is not possible or meaningful to fill out the exposure type for a given study subject. The value 'no exposure' is intended to be used for subjects that are enrolled as negative controls, and indicates specifically that these subjects are *not* be exposed to something. The value 'unknown' is used when samples are from subjects for which no relevant exposure information is available. This is applicable when, for example, a study utilizes samples from anonymous blood bank donors in order to establish a 'normal range'.

For 'Exposure material', the vast majority of HIPC studies submitted to us required specifying an organism that was either the causative agent of an infection, exposure without infection, or utilized to vaccinate to protect against future infection. Organisms can be specified by the broadly utilized NCBI Taxonomy [6], which has the key advantage of linking organism specifications to sequence information in NCBI. All taxa from the NCBI Taxonomy are valid entries for Exposure material, and can be looked up at https://www.ncbi.nlm.nih.gov/taxonomy. One potential concern with this choice is that NCBI does not assign new taxa to every organism isolate identified, which in some cases is desirable, such as in the case of drug resistant *M. tuberculosis* isolates, where it is of interest to relate even single nucleotide differences to efficacy of drug treatments. We expect that going forward, there will be a developing community consensus on how to handle this, along the lines of grouping different isolates based on their NCBI GenBank ID under their closest parent taxon.

Not all 'Exposure materials' in HIPC studies submitted to us were whole organisms. In the case of vaccinations, specific antigens are often utilized over whole organisms such as in the case of subunit vaccines. Also, in the case of multi-valent vaccines, multiple organisms or antigens of organisms are combined into one vaccine. We plan to specify vaccines through the Vaccine Ontology (http://www.violinet.org/vaccineontology/) [7]. It may be necessary to add new entries to the Vaccine Ontology to capture new experimental vaccines, but as vaccines administered to humans have to go through a stringent approval process, this will not overwhelm the Vaccine Ontology development team.

To specify the 'Disease name', the IEDB utilizes values from the Disease Ontology (DO) (http://disease-ontology.org/) [8], which has the advantage of providing mappings to most of the other vocabularies that could be considered such as ICD10, SNOMED CT, MESH and UMLS. The IEDB has been successful in mapping the disease terms encountered in the literature to DO terms. In addition, the Disease Ontology is part of the OBO Foundry [9] and thus more compatible with other basic research ontologies, providing explicit definitions and links to basic research domains, such as clarifying which infectious agent is causative for a given disease. Thus, our immune exposure model will continue to use DO, which was incorporated into ImmPort submission templates via requiring submitters to enter DO terms to describe the diseases of the study subjects.

In terms of 'Disease stage', the IEDB has defined three values that in combination with disease name clarify some typical major distinctions how a disease manifests in different study subjects: (1) 'acute/recent onset' is utilized for subjects that currently have symptomatic disease and may or may not clear it. (2) 'chronic' is utilized for subjects that persistently have a disease and it is not considered highly likely that they will soon clear the disease without intervention. (3) 'post' is utilized for subjects that have cleared a disease which they had in the past. So far, these broad categories have proven sufficient to also describe HIPC needs, although more detailed description of disease specific stages could be desirable in the future and we are open to further discussion.

III. Challenges and Conclusions

The ability to formalize what otherwise would be free-text is a significant accomplishment to improve the integration of data across HIPC studies. More importantly, as this model was adopted by HIPC by adding columns to the Human Subject data submission template, all studies submitted to ImmPort can now include the same fields to describe immune exposures, the HIPC studies will be better connected to other studies in ImmPort. To ease data entry for these fields and others into ImmPort spreadsheet templates, work is ongoing through the CEDAR [10] effort and others to create interactive forms that will ensure that only valid terms are entered.

Now that newly entered data will be formalized, improved query and comparisons will be possible due to standardized terminology. We fully expect that as more data gets submitted to ImmPort using this scheme for HIPC, questions will continue to arise, and based on our experience with the IEDB, we expect to handle them by consulting domain expects for the disease of interest. Controversial cases will be presented to the Clinical Subcommittee, to ensure that decisions are made uniformly across the HIPC program. Overall, it has to be stressed that the structured representation of immune exposures is not intended to fully represent every nuance of each study, but rather achieve its intended function to enable a computable high level comparison of immune exposures across studies. Reassessment of how well this model meets the needs of the community and how it improves the quality of the data after several months of use would be beneficial.

Acknowledgements

This work was supported by the National Institute of Allergy And Infectious Diseases of the National Institutes of Health under Award Number NIH U19 AI118610 and U19AI089992. It would not have been possible without strong support by the ImmPort team, and Patrick Dunn in particular.

References

[1] S. Bhattacharya, S. Andorf, L. Gomes, et al, "ImmPort: disseminating data to the public for the future of immunology," Immunol Res. 58(2-3), pp. 234-239, May 2014.

[2] https://www.immuneprofiling.org/hipc/page/show (accessed 6/1/2018).

[3] R. Vita, J.A. Overton, J.A. Greenbaum, et al, "The immune epitope database (IEDB) 3.0," Nucleic Acids Res. 43(Database issue):D, pp. 405-412, October 2014.

[4] A. Bandrowski, R. Brinkman, M. Brochhausen, et al, "The Ontology for Biomedical Investigations," PLoS One 29;11(4). Apr 2016.

[5] J.A. Greenbaum, R. Vita, L. Zarebski, et al, "ONTology of Immune Epitopes (ONTIE) Representing the Immune Epitope Database in OWL," The 12th Annual Bio-Ontologies Meeting, ISMB, pp. 45–48, 2009.

[6] E.W. Sayers, T. Barrett, D.A. Benson, et al, "Database resources of the National Center for Biotechnology Information," Nucleic Acids Res. 37, pp. D5–D15, May 2009.

[7] Y. He, L. Cowell, A.D. Diehl, "VO: Vaccine Ontology," The 1st International Conference on Biomedical Ontology (ICBO 2009), Buffalo, NY, USA. Nature Precedings. 2009.

[8] W.A. Kibbe, C. Arze, V. Felix, et al, "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," Nucleic Acids Res. 43(Database issue):D, pp. 1071-1078, January 2015.

[9] B. Smith, M. Ashburner, C. Rosse, et al, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," Nat Biotechnol. 25(11), pp. 1251-1255, November 2007.

[10] M.A. Musen, C.A. Bean, K.H. Cheung, et al, "The center for expanded data annotation and retrieval," J Am Med Inform Assoc. 22(6), pp. 1148-52, November 2015.