# *Ontology representation and ANOVA analysis of vaccine protection investigation*

*Yongqun He[1*], Zuoshuang Xiang[1], Thomas Todd[1], Melanie Courtot[2], Ryan Brinkman[2], Jie Zheng[3], Chris Stoeckert[3], James Malon[4], Philippe Rocca-Serra[4], Susanna-Assunta Sansone[4], Jennifer Fostel[5], Larisa N. Soldatova[6], Bjoern Peters[7], Alan Ruttenberg[8]*

*[1] University of Michigan, Ann Arbor, USA; [2]British Columbia Cancer Agency, Vancouver, Canada; [3]Center for Bioinformatics, Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA, USA; [4]The European Bioinformatics Institute, Cambridge, UK; [5]Global Health Sector, SRA International, Inc, Durham, NC, USA; [6]Aberystwyth University, Wales, UK; [7]La Jolla Institute for Allergy and Immunology, La Jolla, CA, USA; [8]Science Commons, Cambridge, MA, USA.*

**ABSTRACT**

**Motivation:** It is still challenging to represent statistical analysis of experimental data in a semantical framework. As a first step towards this goal, ontological representation of statistical ANOVA analysis is proposed. In a vaccine protection use case, 151 instance data of *Brucella* vaccine protection investigation were collected from the literature and analyzed using ANOVA. Out of 16 parameters, 10 were found statistically significant in contributing to the protection. The careful study of these instances led to building and validating an OBI-based semantic framework to represent ANOVA formally. An ontology-based representation and statistical analysis of biomedical data allows data consistency checking and data sharing in Semantic Web.

**Contact**: yongqunh@med.umich.edu

## 1 INTRODUCTION

The Ontology for Biomedical Investigations (OBI) is being developed to address the need for a common, integrated ontology for the description of biological and clinical investigations. OBI has been used in experimental investigations in different communities, for example, Bioinvindex (http://www.ebi.ac.uk/bioinvindex), isa-tools (http://isatab.sourceforge.net/), and IEDB (http://www.immuneepitope.org/). In our recent study, we used OBI and other ontologies to represent an investigation of vaccine protection against influenza viral infection (Brinkman et al, 2010). The vaccine protection investigation measures how efficient a vaccine or vaccine candidate induces protection against virulent pathogen infection *in vivo*.

While ontology representation of experimental assays in terms of material inputs and data outputs provide a foundation for further data sharing and semantic web studies of specific domains, it is still challenging to apply semantic frameworks to statistical analysis of instance data. OntoDM is a newly proposed ontology of data mining that provides a framework and describes entities from the domain of data mining and knowledge discovery. OntoDM is aligned with OBI. The updated OBI has included many statistical terms (e.g., ANOVA, F-test, t-test) and relevant supports that facilitate statistical analysis.

The community-based Vaccine Ontology (VO; http://www.violinet.org/vaccineontology/) is biomedical ontology that covers the vaccine domain (He et al, 2009). Development of VO has emphasized classification of vaccines and vaccine components, vaccination investigation, and host responses to vaccines. The VO development follows the OBO Foundry principles [Smith *et.al*., 2007]. VO uses the Basic Formal Ontology (BFO) [Grenon *et.al*, 2004] as the top-level ontology. OBI is used as another upper level ontology for vaccine investigation. VO uses relations defined by primarily the Relation Ontology (RO) [Smith *et.al*., 2005] and also by OBI and the Information Artifact Ontology (IAO) ontologies. The close association with these ontologies facilitates data integration and automated reasoning.

In this report, we first introduce our ontology representation of the ANOVA statistical analysis, then apply it to investigate the *Brucella* vaccine protection results curated from the literature. *Brucella* is an intracellular bacterium that causes brucellosis, the most common zoonotic disease worldwide. In this study, we hypothesized that some experimental variables significantly contribute to *Brucella* vaccine protection efficacy while others do not. Our study indicates that relying on a semantic framework such as OBI and OntoDM is a useful approach to support biomedical statistical data analyses.

## 2 METHODS

The following methods were applied in this study:

**Ontology representation of ANOVA Statistical analysis**: The analysis of variance (ANOVA) was modeled primarily in OBI. A design pattern was generated. The use case in this study is ANOVA in terms of a linear model.

* To whom correspondence should be addressed.

**Ontology-based representation of vaccine protection investigation:** All variables in this use case are represented using different ontologies as needed. The main ontologies used include VO, OBI, and IAO.

**Literature curation of individual *Brucella* vaccine protection data:** Peer-reviewed *Brucella* vaccine protection research papers were obtained from PubMed search. These papers were manually curated to identify variables and extract values taken by these variables potentially important for vaccine protection efficacy investigation. The data were stored in an OWL file.

**Ontology-based ANOVA analysis of *Brucella* vaccine protection results:** ANOVA was applied to study the *Brucella* vaccine protection investigation instance data. The results were also represented in ontology.

## 3 RESULTS

We will first introduce how ANOVA is modeled in OBI. The ontology representation of vaccine protection investigation using VO and OBI is then described. Using literature curated data, we will last introduce how the vaccine protection results are analyzed by ANOVA and modeled using ontology.

### 3.1 Ontology design pattern of ANOVA data analysis

The analysis of variance (ANOVA) provides a statistical test of whether or not the means of several groups are all equal. In statistics, ANOVA includes a collection of statistical models (e.g., linear models), and their associated procedures, in which the observed variance is partitioned into components due to different explanatory variables. The ontology-based ANOVA data analysis design pattern is illustrated in Fig. 1. *ANOVA* is a subclass of *data transformation* process in OBI. F-test is part of ANOVA process. ANOVA has specified input of data item. The individual data items come from two sources. The data items are possibly the output of individual processes (e.g., CFU reduction assay). Alternatively, a data item can be an output of a *discretization process* that descretizes non-measurable data (e.g., mouse age) into categorized measurement data (e.g., 1 for young mouse, 2 for middle-aged mouse, and 3 for old mouse). One approach to obtain the data items necessary for ANOVA analysis is through *data item extraction from journal article* (IAO_0000443). In this case, the input is journal article, and the output is data. The ANOVA output is a *p-value data set*, which includes a set of p-value results for an independent variable data set that is predefined.

*ANOVA* is concretization of *ANOVA protocol*. The ANOVA protocol includes a predictive model that specifies a testable hypothesis model (Fig. 1).

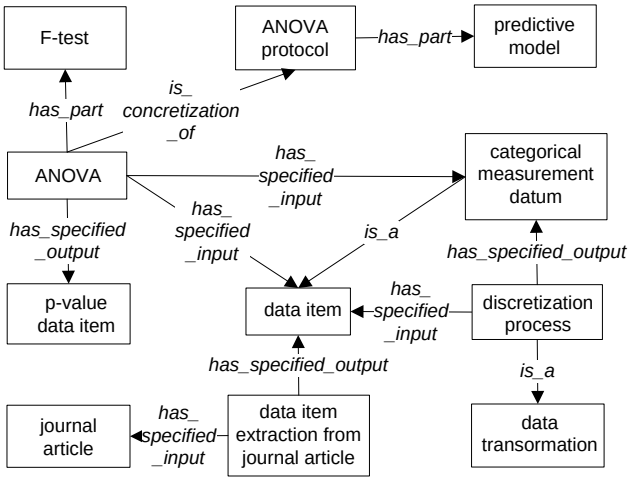


Fig. 1. Representation of ANOVA analysis process.

### 3.2 Ontology representation of *Brucella* vaccine protection investigation

A *vaccine protection investigation* includes three processes (or steps): *vaccination*, *pathogen challenge*, and *vaccine protection efficacy assessment*. For those pathogens that kill a model animal (e.g., mouse), survival assessment is used for assessing vaccine protection efficacy (Brinkman et al, 2010). Since virulent *Brucella* does not kill mice, the survival of pathogen challenged mice is not a useful method to assess *Brucella* vaccine efficacy. Instead, a colony forming unit (CFU) reduction assay is used to determine the difference of live bacterial recovery from vaccinated mice and non-vaccinated mice (Schurig et al., 1991).

To prove vaccine protection efficacy, a vaccine protection investigation using a specific animal model is often required. In this process, many variables may affect the outcomes. We summarized 17 variables that are described in typical vaccine protection studies. The ontology terms of these 17 variables are summarized in Table 1.

As an example of this *Brucella* vaccine protection investigation, *Brucella abortus* cattle vaccine RB51 was used in a typical vaccine protection study as reported in reference (Schurig et al., 1991). In this typical mouse experiment, live RB51 ($1 \times 10^8$ CFU) was used to vaccinate Balb/C mice, and the mice were challenged with *B. abortus* strain 2308 ($1 \times 10^5$ CFU) 8 weeks later. CFU reduction in mouse spleen was then counted to determine the vaccine protection. An ontology representation of this example is shown in Fig. 2.

The experimental hypothesis is "Some experimental variables statistically significantly contribute to *Brucella* vaccine protection efficacy". This hypothesis can be laid out as an instance of the *hypothesis entity text*.

**Table 1.** Ontology terms for 17 variables in this use case.

| # | Classes / ANOVA variables | Sources & term IDs |
|---|---|---|
| 1 | vaccine protection efficacy | VO: VO_0000456 |
| 2 | vaccine strain | VO: VO_0001180 |
| 3 | vaccine viability | VO: VO_0001139 |
| 4 | vaccine protective antigen | VO: VO_0000457 |
| 5 | mutated gene in vaccine strain | VO: VO_0001195 |
| 6 | vaccination mouse strain | VO: VO_0001189 |
| 7 | vaccination dose specification | VO: VO_0001160 |
| 8 | pathogen strain for challenge | VO: VO_0001194 |
| 9 | pathogen challenge (subclass) | OBI: OBI_0000712 |
| 10 | CFU per volume | UO: UO_0000212 |
| 11 | CFU reduction | VO: VO_0001164 |
| 12 | IL-12 vaccine adjuvant | VO: VO_0001147 |
| 13 | biological sex | PATO: PATO_0000047 |
| 14 | vaccination (subclass) | VO: VO_0000002 |
| 15 | animal age at vaccination | VO: VO_0000897 |
| 16 | vaccination-challenge interval | VO: VO_0001191 |
| 17 | challenge dose specification | VO: VO_0001161 |

Note: The first variable is dependent variable, and the others are independent variables. The last six variables did not contribute to the vaccine protection (p-value < 0.05).

### 3.3 ANOVA analysis of *Brucella* vaccine protection results from literature curation

*Brucella* vaccine research is an active research area with more than 1,000 peer-reviewed papers stored in PubMed. To determine which variables play significant roles in changing the *Brucella* vaccine protection efficacy, more than 40 papers were manually curated to get instance data that correspond to these variables. In total, 151 instance data were collected from the literature. In this study, we only focused on mice as the animal model. Different mouse strains were analyzed in our use case investigation. Each instance of vaccine protection investigation has individual values for all 17 variables (Table 1).

To analyze which variables contribute to the vaccine protection, the significance of vaccine protection (three values: no protection, protection, enhanced protection) is set as a dependent variable, and the other 16 variables are independent variables. An ANOVA analysis was performed and indicated that six variables do not statistically significantly contribute to the protection (p-value > 0.05). These six variables include IL-12 vaccine adjuvant, mouse sex, vaccination route, mouse age at vaccination, vaccination-challenge interval, and challenge dose. The other 10 parameters statistically significantly contribute to the vaccine protection (p-value < 0.05).

The predictive model is "Protection_Significance ~ ." indicating we are testing how each other variable affects the protection significance. This linear model representation can be understood and processed by statistical software programs such as R programming.

This use case was used to derive an instance level representation based on the formal semantic representation of ANOVA analysis (Fig. 1 and 2, Table 1). Specifically, to represent this use case ANOVA data analysis using ontology, we defined a 'vaccine protection ANOVA' (VO_0000572) under 'ANOVA'. This ANOVA has vaccine protection efficacy as dependent variable and 16 other independent variables (Table 1). All values for individual variables were obtained from literature curation. A hypothesis was also generated as an instance of the 'hypothesis textual entity'. The 151 instance data of this use case study was represented in OWL format. Each set of instance data is defined under an instance of '*vaccine protection investigation*'. The ANOVA output is a p-value data set that corresponds to a list of p-values for different independent variables.

## DISCUSSION

The advantage of ontology-based statistical analysis is that the results can be potentially shared and used worldwide through semantic explicit representation. Also, ontology based approach facilitates data consistency checking. For a specific variable (e.g., vaccine strain) from a biomedical investigation, specific instances are generated and match to the variable (e.g., RB51 as an instance of vaccine strain). In our use case, many subclasses also act as instances for parent class variables. For example, RB51 is a subclass of vaccine strain. If a vaccine strain instance does not belong to a vaccine strain, it may indicate the data is not right. Existing OWL reasoners, such as Pellet (http://clarkparsia.com/pellet) and FACT++ (http://owl.man.ac.uk/factplusplus/), can be effectively leveraged to detect inconsistencies in statistical analysis representation.

There are still many challenges in modeling statistical analyses using ontology. For example, there is no consistent representation of null hypothesis in statistical analysis yet. However, the example we described in this report provides a first demonstration that it is feasible and provides more powerful features than traditional statistical analysis without ontology and semantic support. However, ANOVA has been chosen in the first place, as it is such an important tool in life science. ANOVA is a special case of linear model analysis, so experience gained from applying formal semantics to ANOVA could be beneficial for some more advanced representation of such linear models.

Ontology representation of vaccine protection study provides an advanced approach to represent and mine vaccine-induced protection experimental processes. More than 400 vaccines and the data of protection studies with these vaccines have been manually curated and stored in the VIOLIN vaccine database system (Xiang *et.al.*, 2008). To make full use of the VIOLIN vaccine data for advanced query and integration with data from other data sources, we plan to apply the ontology-based approach learned from this *Brucella* study to other vaccine protection data available in VIOLIN.

Our method of ontology-based representation and statistical analysis is applicable for other ontology-based statistical studies. The logical definitions of the ontology entities involved allow computers to unambiguously understand and integrate different biological data with the help of an OWL reasoner. We anticipate that more statistical analyses will be represented in ontology, and ontology-based statistical methods will be applied for shared data analysis, data exchange, and automatic reasoning. Various new software programs will most likely be developed in the future to take advantage of this novel semantic framework.

## ACKNOWLEDGEMENTS

This research is supported by NIH grants R01AI081062 and U54-DA-021519.

## REFERENCES

Brinkman RR, Courtot M, Derom D, et al. (2010) Modeling biomedical experimental processes with OBI. *Journal of Biomedical Semantics*. In press.

He Y, Cowell L, Diehl AD, et al. (2009) VO: Vacccine Ontology. International Conference on Biomedical Ontology (ICBO), 24 July 2009. *Nature Precedings*. Available at web site: http://precedings.nature.com/documents/3552/version/1.

Panov P, Soldatova LN, Dzeroski S. (2009) Towards an Ontology of Data Mining Investigations. *Proceedings of the 12th International Conference on Discovery Science*, Porto, Portugal.

Schurig GG, Roop RM, Bagchi T, et al. (1991) Biological properties of RB51; a stable rough strain of Brucella abortus. *Vet Micobiol*, 28(2) : 171-188.

Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C (2005) Relations in Biomedical Ontologies. *Genome Biology*, **6**: R46

Smith et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature Biotechnology*, **25**: 1251 - 1255.

Xiang Z, Todd T, Ku KP, Kovacic BL, Larson CB, et al. (2008) VIOLIN: vaccine investigation and online information network. *Nucleic Acids Res*. **36** (Database issue): D923-8.

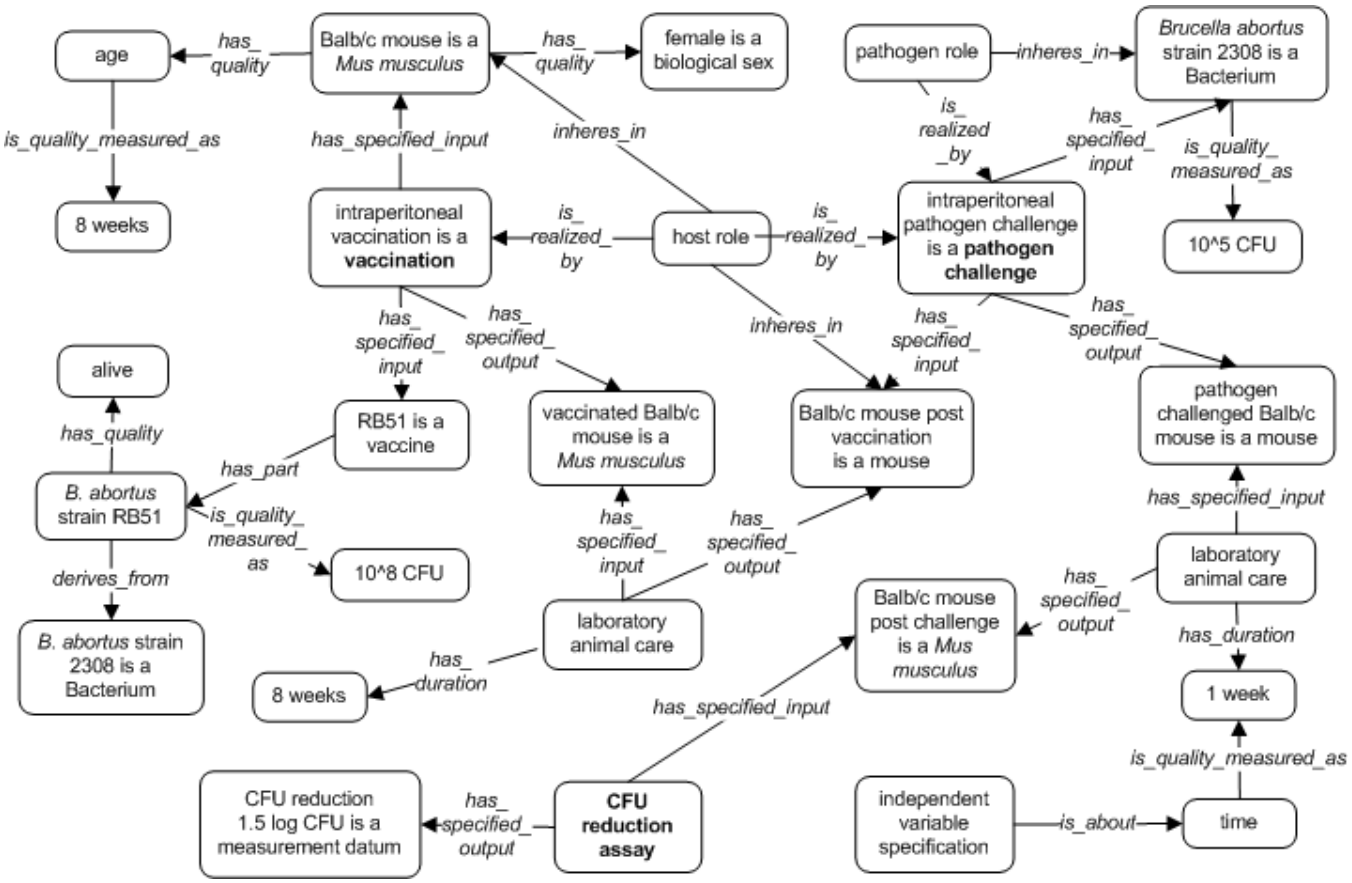


Fig. 2: Representation of a protection assay with *Brucella* vaccine RB51 (Schurig et al, 1991). Boxes represent OWL individuals. Terms from different ontologies (e.g, OBI, VO, IAO) are used. Italicized text in the middle of arrows represents relations. The bold terms represent three major processes in the vaccine protection investigation as described in the text.