Data Understanding Report

Utrecht University

Mike Vink

April 23, 2021

Contents

1 Initial data collection

1.1 Technical description data collection

1.1.1 MySQL database set up and data import

By following the guide on the FluPrint Github Repository the MySQL server was set up. In this work the FluPrint github was first added as a submodule. This module provides the php scripts to import raw data csv's into the MySQL database. The operating system and versions of php and MySQL used in this work were OSX "Big Sur" (on Mac Book air 2017), php 7.3.24 (built-in mac version), and MySQL 8.0.23 (homebrew).

In the guide the dependencies to run the php import script were installed first. This was also done in this work, except that the hash-file verification step was skipped.

After the php dependencies were installed the MySQL server was started. By default homebrew recommends to use the `homebrew services [option] [SERVICE]` command to start the MySQL server. However, in this work the server is started using `mysql.server start` which provides a socket that was symlinked using `sudo ln -s /tmp/mysql.sock /var/mysql/mysql.sock`. This was done to prevent an error (StackOverflow: cant connect to local mysql server through socket homebrew) thrown by the php import scripts. Before the import scripts were run a

user was added to the MySQL server and a database was created 1, the password type had to be `mysql_native_password` (how to resolve [SQLSTATEHY000] 2054 the server requested authentication method.).

Listing 1: Adding user and database to sql server

```
1 mysql> CREATE USER 'mike'@'localhost' IDENTIFIED BY ';lkj';
2 mysql> GRANT ALL PRIVILEGES ON * . * TO 'mike'@'localhost';
3 mysql> ALTER USER 'mike'@'localhost' IDENTIFIED WITH
    ↪ mysql_native_password BY 'mike';
4 mysql> CREATE DATABASE fluprint;
```

The databasename, the username, and password were added to the `config/configuration.json` of the FlruPrint github module. At this point the configuration for the php import scripts was finished, and the raw data downloaded in `data/upload` were imported in the MySQL server using `php bin/import.php`.

1.2 Data Requirements

The following subsections will list the information required from the data per data mining goals that are needed to answer the following business questions:

- Which datasets in the FluPrint database are most interesting?
- How do different clinical studies compare?
- What are the differences in efficacy between vaccination types?
- What is the effect of repeat vaccination on vaccine response?
- What immunological factors correlate to a high vaccine response?

1.2.1 Requirements per data mining goal

"Explore and describe SQL queries and corresponding csv tables."

Falling under this data mining objective are the outputs and tasks related to data collection and description. These comprise a report on the initial collection of the data, selection of data, and description of general properties of the data. The data in this case is in a database format, thus here we describe the tables, keys, and attributes in the database, and also include descriptive statistics about the data. The goal is to replicate the description done in A. Tomic, I. Tomic, Dekker, et al., 2019 as well. Using these descriptions we provide insight into which datasets in the database are most interesting, and why in A. Tomic, I. Tomic, Rosenberg-Hasson, et al., 2019 one dataset in particular was chosen.

"Model and visualise the different clinical study populations."

"Model and visualise the difference between vaccination types."

"Model and visualise repeat vaccination effects."

In order to answer the business question "How do clinical studies compare?" subpopulations and groups of attributes need to be visualised and compared across different clinical studies. The data required must have rows corresponding to donors in a particular clinical study and columns that are attributes of tables in the database, these could be biological assay results or information about the donors. Thus we aimed to export one csv from the database per clinical study by querying for different clinical studies.

We aimed to generate csv files of donors corresponding to received vaccine types to answer the business question "What are the differences in efficacy between vaccination types?". One simple method to indicate the difference between vaccines would be to report the proportion of high-reponders across all donors, or to use a simple model to find the best predictor for a high response. These comparisons require one table per vaccine type, with rows corresponding to donors and columns that include the vaccine response classification, in addition to other immune assay and donor attributes.

The objective in question "What is the effect of repeat vaccination on vaccine response?" requires data from long running clinical studies. One dataset that is used by the database authors and was investigated to answer this question was already available, here we aimed to describe and visualise any patterns we could find in this dataset and other long running clinical study datasets. This required data from a subset of clinical studies that spanned multiple years, at this point in the project the data for these clinical studies should have been available, and we just had to choose those that spanned multiple years.

"Apply standard feature selection methods to the most interesting dataset."

"Fit classification models to the most interesting dataset."

These last two data mining objectives were chosen to comprise the data preparation and modelling phases of this project. The authors of fluprint set up an automated machine learning pipeline to investigate the longest running clinical dataset in the database. In this work we use a conventional data mining modelling process to replicate these results.

2 Data description

2.1 Volumetric analysis

In the work of A. Tomic, I. Tomic, Dekker, et al., 2019 data on indiviuals enrolled in influenza vaccine studies at the Stanford-LPCH Vaccine Program was collected, the data was archived at the Stanford Data Miner. This archive was filtered by assays used in influenza studies, resulting in data from 740 healthy donors, enrolled in influenza vaccine studies conducted by the Stanford-LPCH Vaccine Program from 2007 to 2015. These studies are described in the table accompanying the online publication of the fluprint dataset (Table 1).

Importantly in all studies the donors are only vaccinated once, except in the study SLVP015, participants here were vaccinated annually from 2007-2015 (Table 1). In all other studies participants would only be studied within the scope of one influenza season.

From those 740 donors a vaccine response classification was only given for 372 donors (Figure **??**), by a method that will be described later.

Stanford study ID	Name	Description	Vaccines	Data in FluPRINT
SLVP015	Comparison of immune responses to influenza vaccine in adults of different ages (2007-2017)	Who: 18-100yo healthy participants How: immunized annually with the seasonal inactivated influenza vaccines from 2007-2017 When: Blood samples acquired before immunization (Day 0), on days 6-8 and 28 after immunization	2007-2013 Seasonal trivalent, inactivated influenza vaccines (Fluzone) 2014-2015 High Dose trivalent Fluzone for participants *geq* 65yo and quadrivalent Fluzone for younger participants	135 donors Assays: 51-plex Luminex 62-plex Luminex MSD 4plex MSD9plex Other Luminex HAI CMV/EBV Hormones CyTOF phenotype Lyoplate Phospho Cytof pheno Phospho cytof phospho Phosphoflow CBCD
SLVP017	B-cell immunity to influenza (2009-2011 and 2013)	Who: 1-2yo (2013), 8-100yo healthy participants who did not receive the seasonal influenza vaccine in previous years (2010, 2011 and 2013) How: immunized with either seasonal inactivated or live, attenuated influenza vaccines in 2009, 2010, 2011 and 2013 When: Blood samples acquired before immunization (Day 0) and on day 28 after immunization	2009-2011 Seasonal trivalent, inactivated influenza vaccines (Fluzone) or seasonal live, attenuated influenza vaccine (FluMist) 2013 Seasonal trivalent inactivated influenza vaccine- (Fluzone) - pediatric formulation for 1-2yo children	153 donors Assays: 51-plex Luminex 62-plex Luminex HAI CMV/EBV CyTOF phenotype CBCD
SLVP018	T-cell and general immune response to seasonal influenza vaccine (2009-2013)	Who: 1-8yo (2013), 8-100yo healthy participants How: immunized with either seasonal inactivated or live, attenuated influenza vaccines from 2009-2013 When: Blood samples acquired before immunization (Day 0), days 7-10 and 28 after immunization	2009-2010 Seasonal trivalent inactivated influenza vaccine (Fluzone) or seasonal trivalent live attenuated influenza vaccine (FluMist) 2010 High Dose trivalent Fluzone for participants *geq* 65yo 2013 Seasonal trivalent, inactivated influenza Pediatric Dose (Fluzone, 0.25 ml) for 1-3yo children	249 donors Assays: 51-plex Luminex 62-plex Luminex MSD 4plex MSD 9plex HAI CMV/EBV Hormones CyTOF phenotype Lyoplate Phospho Cytof pheno Phospho cytof phospho Phosphoflow CBCD
SLVP021	Plasmablast trafficking and antibody response in influenza vaccination (2011-2014)	Who: 8-34yo healthy participants who did not receive the seasonal influenza vaccine in previous years How: immunized with either seasonal inactivated influenza vaccines, given intramuscularly or intradermally, or live, attenuated influenza vaccines from 2011-2014 When: Blood samples acquired before immunization (Day 0), days 6-8 and 24-32 after immunization	2011-2014 Seasonal trivalent inactivated influenza vaccine (Fluzone) given either intramuscularly or intradermally 2011-2012 Seasonal trivalent live attenuated influenza vaccine (FluMist)	84 donors Assays: 51-plex Luminex 62-plex Luminex HAI CMV/EBV Hormones CyTOF phenotype Phospho Cytof pheno Phospho cytof phospho Phosphoflow CBCD
SLVP024	Protective mechanisms against a pandemic respiratory virus (2012)	Who: 2-9yo healthy participants How: immunized with the seasonal live, attenuated influenza vaccine When: Blood samples only from 18-2yo adults acquired before immunization (Day 0), days 7 and 28 after immunization	Seasonal live, attenuated influenza vaccine (FluMist)	Donors: 8 Assays: HAI Phosphoflow
SLVP028	Genetic and environmental factors in the response to influenza vaccination (2014-2018)	Who: 12-9yo healthy participants How: immunized with either seasonal inactivated or live, attenuated influenza vaccines from 2014-2018 When: Blood samples acquired before immunization (Day 0), days 6-8 and 28 + 7 after immunization	Seasonal quadrivalent inactivated influenza vaccine (Fluzone) or seasonal quadrivalent live attenuated influenza vaccine (FluMist)	Donors: 52 Assays: 62-plex Luminex HAI CMV/EBV Hormones CyTOF phenotype
SLVP029	Innate and acquired immunity to influenza infection and immunization (2014-2017)	Who: 6 mo-49yo healthy participants (who did not receive LAIV in the prior season nor received influenza immunizations in two or more prior seasons) How: immunized with either seasonal inactivated or live, attenuated influenza vaccines from 2014-2017 When: Blood samples acquired before immunization (Day 0), days 7 and 28 after immunization. Children ¡9 yrs received 2 immunizations with the second blood samples acquired 28 days after second immunization	Seasonal quadrivalent inactivated influenza vaccine (Fluzone) or seasonal quadrivalent live attenuated influenza vaccine (FluMist)	Donors: 47 Assays: 62-plex Luminex HAI CMV/EBV Hormones CyTOF phenotype
SLVP030	The role of CD4+ memory phenotype, memory, and effector t cells in vaccination and infection (2014-2019)	Who: 6 mo-10yo healthy participants How: immunized annually with either seasonal inactivated or live, attenuated influenza vaccines from 2014-2019 When: Blood samples acquired before immunization (Day 0), days 7 and 60 after immunization. Children with no prior influenza vaccine received 2 immunizations with the second blood sample acquired 60 days after second immunization	Seasonal quadrivalent inactivated influenza vaccine (Fluzone) or seasonal quadrivalent live attenuated influenza vaccine (FluMist) Seasonal trivalent, inactivated influenza Pediatric Dose (Fluzone, 0.25 ml) for 6-35mo children	Donors: 12 Assays: 62-plex Luminex HAI CMV/EBV Hormones CyTOF phenotype

Table 1: **Reference table of clinical studies** Clinical study ID used (but remapped) in the database, age information, vaccine type information, and assay data types of clinical studies are in the rest of the columns.

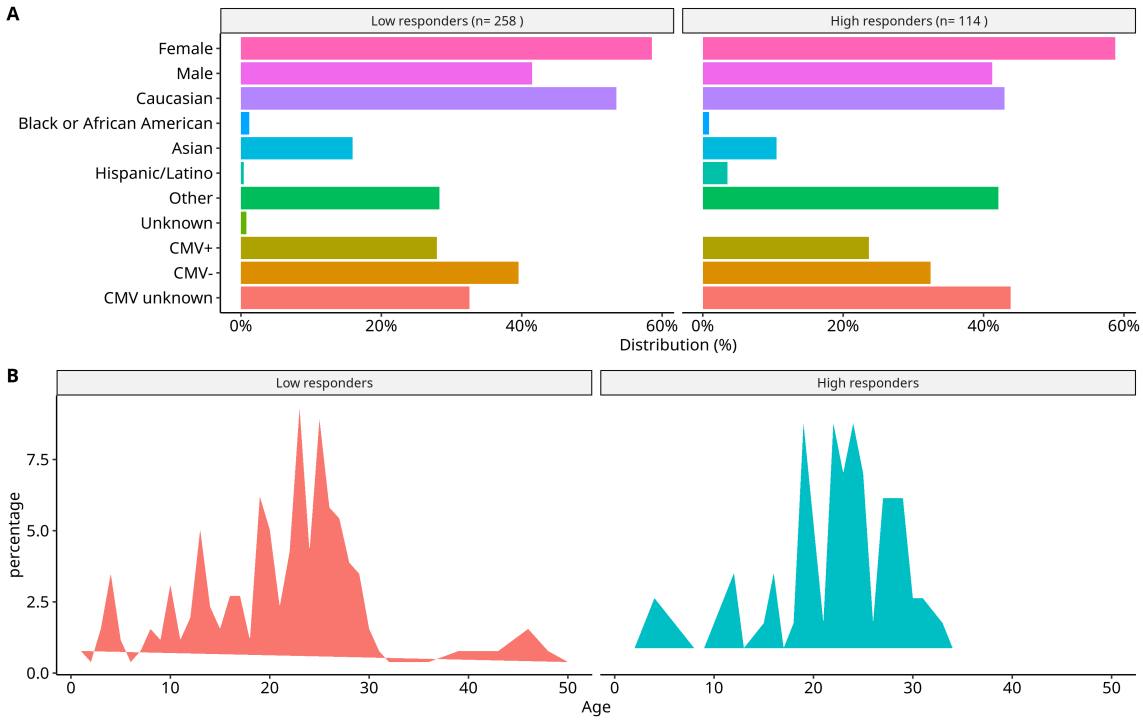

Figure 1: **A.** percentage of donors with factor property within high and low responder groups. Included are sex, race, and CMV status information. **B.** Age distribution of donors with a known response classification.

(Chattopadhyay et al., 2014) The complex heterogeneity of cells, and their interconnectedness with each other, are major challenges to identifying clinically relevant measurements that reflect the state and capability of the immune system. Highly multiplexed, single-cell technologies may be critical for identifying correlates of disease or immunological interventions as well as for elucidating the underlying mechanisms of immunity. Here we review limitations of bulk measurements and explore advances in single-cell technologies that overcome these problems by expanding the depth and breadth of functional and phenotypic analysis in space and time. The geometric increases in complexity of data make formidable hurdles for exploring, analyzing and presenting results. We summarize recent approaches to making such computations tractable and discuss challenges for integrating heterogeneous data obtained using these single-cell technologies.

(Galli et al., 2019) High-dimensional single-cell (HDcyto) technologies, such as mass cytometry (CyTOF) and flow cytometry, are the key techniques that hold a great promise for deciphering complex biological processes. During the last decade, we witnessed an exponential increase of novel HDcyto technologies that are able to deliver an in-depth profiling in different settings, such as various autoimmune diseases and cancer. The concurrent advance of custom data-mining algorithms has provided a rich substrate for the development of novel tools in translational medicine research. HDcyto technologies have been successfully used to investigate cellular cues driving pathophysiolog-

Age (y)	
Mean ± SD	21.02 ± 8.66
Median (min. to max. range)	22.5 (1 - 50)
Gender	
Male (%)	154 (41.4)
Female	218 (58.6)
Ethnicity	
Caucasian (%)	187 (50.3)
African American (Black) (%)	4 (1.1)
Asian (%)	53 (14.2)
Hispanic/Latino (%)	5 (1.3)
Other (%)	121 (32.5)
Unknown (%)	2 (0.5)

Table 2: **Demographic statistics of donors with known vaccine response classification.**

ical conditions, and to identify disease-specific signatures that may serve as diagnostic biomarkers or therapeutic targets. These technologies now also offer the possibility to describe a complete cellular environment, providing unanticipated insights into human biology. In this review, we present an update on the current cutting-edge HDcyto technologies and their applications, which are going to be fundamental in providing further insights into human immunology and pathophysiology of various diseases. Importantly, we further provide an overview of the main algorithms currently available for data mining, together with the conceptual workflow for high-dimensional cytometric data handling and analysis. Overall, this review aims to be a handy overview for immunologists on how to design, develop and read HDcyto data.

(Simoni et al., 2018) Advancement in methodologies for single cell analysis has historically been a major driver of progress in immunology. Currently, high dimensional flow cytometry, mass cytometry and various forms of single cell sequencing-based analysis methods are being widely adopted to expose the staggering heterogeneity of immune cells in many contexts. Here, we focus on mass cytometry, a form of flow cytometry that allows for simultaneous interrogation of more than 40 different marker molecules, including cytokines and transcription factors, without the need for spectral compensation. We argue that mass cytometry occupies an important niche within the landscape of single-cell analysis platforms that enables the efficient and in-depth study of diverse immune cell subsets with an ability to zoom-in on myeloid and lymphoid compartments in various tissues in health and disease. We further discuss the unique features of mass cytometry that are favorable for combining multiplex peptide-MHC multimer technology and phenotypic characterization of antigen specific T cells. By referring to recent studies revealing the complexities of tumor immune infiltrates, we highlight the particular importance of this technology for studying cancer in the context of cancer immunotherapy. Finally, we provide thoughts on current technical limitations and how we imagine these being overcome.

References

Chattopadhyay, Pratip K. et al. (Feb. 2014). "Single-Cell Technologies for Monitoring Immune Systems". English. In: *Nature Immunology* 15.2, pp. 128–135. ISSN: 1529-2916. DOI: 10.1038/ni.2796.

Galli, Edoardo et al. (2019). "The End of Omics? High Dimensional Single Cell Analysis in Precision Medicine". English. In: *European Journal of Immunology* 49.2, pp. 212–220. ISSN: 1521-4141. DOI: 10.1002/eji.201847758.

Simoni, Yannick et al. (Apr. 2018). "Mass Cytometry: A Powerful Tool for Dissecting the Immune Landscape". English. In: *Current Opinion in Immunology* 51, pp. 187–196. ISSN: 0952-7915. DOI: 10.1016/j.coi.2018.03.023.

Tomic, Adriana, Ivan Tomic, Cornelia L. Dekker, et al. (Oct. 2019). "The FluPRINT Dataset, a Multidimensional Analysis of the Influenza Vaccine Imprint on the Immune System". English. In: *Scientific Data* 6.1, p. 214. ISSN: 2052-4463. DOI: 10.1038/s41597-019-0213-4.

Tomic, Adriana, Ivan Tomic, Yael Rosenberg-Hasson, et al. (Feb. 2019). "SIMON, an Automated Machine Learning System Reveals Immune Signatures of Influenza Vaccine Responses". English. In: *bioRxiv*, p. 545186. DOI: 10.1101/545186.