Data Understanding Report

Utrecht University

Mike Vink

April 20, 2021

Contents

1 Initial data collection 1

1.1 Technicalities 1

1.1.1 MySQL database set up and data import 1

1.2 Data Requirements 2

1.2.1 Requirements per data mining goal 2

1 Initial data collection

1.1 Technicalities

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

```
mysql> CREATE USER 'mike'@'localhost' IDENTIFIED BY ';
    ↪ lkj';
mysql> GRANT ALL PRIVILEGES ON * . * TO 'mike'@'
    ↪ localhost';
mysql> ALTER USER 'mike'@'localhost' IDENTIFIED WITH
    ↪ mysql_native_password BY 'mike';
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 **tomicFluPRINTDatasetMultidimensional2019** as well. Using these descriptions we provide insight into which datasets in the database are most interesting, and why in **tomicSIMONAutomatedMachine2019** 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.”

(**chattopadhyaySinglecellTechnologiesMonitoring2014**) 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.

(**galliEndOmicsHigh2019**) 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 pathophysiological 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.

(**simoniMassCytometryPowerful2018**) 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.