Change in immune cell signaling upon repeat vaccination: a data exploration using the FluPrint database

Utrecht University

Mike Vink

May 9, 2021

Contents

Bussiness glossary

antibody Protein used by the immune system to identify and neutralize foreign objects such as pathogenic bacteria and viruses. The antibody recognizes a unique molecule of the pathogen, called an antigen. 3, 7, 8

antigen In immunology, an antigen is a molecule or molecular structure, such as HA and NA, that can be bound by an antigen-specific antibody or immune cell receptor. The presence of antigens in the body normally triggers an immune response . 3, 4, 7, 8

B-cell B-cells produce antibody molecules; however, these antibodies are not secreted. Rather, they are presented on the outside of the cell where they serve as a part of B-cell receptors. When a B-cell is activated by an antigen, it proliferates and differentiates into an antibody-secreting effector cell, known as a plasmablast or plasma cell . 4, 8, 29, 30

CD4+ T-cell The T helper cells, also known as CD4+ cells, "help" the activity of other immune cells by releasing cytokines. These cells help to polarize the immune response into the appropriate kind depending on the nature of the immunological insult (e.g. virus vs. bacterium) . 8, 29, 30

CD8+ T-cell A cytotoxic T cell (also known as CD8+ T-cell) is a T-cell that kills cancer cells, cells that are infected (particularly with viruses), or cells that are damaged in other ways. It does so by recognizing specific part of antigen and then starting a process that kills the targetted cell . 8, 30

CMV Cytomegalovirus (CMV) is a common herpesvirus found in humans. Like other herpesviruses, it is a life-long infection that remains in a latent state inside the human body, until it is 're-activated' by appropriate conditions. Thought to accelerate aging of the immune system and thereby impairing influenza vaccine response (**van˙den˙Berg˙2019**). 13, 17, 21

cytokine Cytokines are a broad and loose category of small proteins important in cell signaling that bind to receptor protein on the outside of (immune) cells to fulfill their signal function . 3, 8, 21

EBV The Epstein–Barr virus (EBV), is one of the nine known human herpesvirus types in the herpes family, and is one of the most common viruses in humans.. 17, 21

glycoprotein Glycoproteins are molecules that comprise protein and carbohydrate chains. Many viruses have external glycoproteins that help them enter bodily cells, but can also serve to be important therapeutic or preventative targets. 7

HAI The hemagglutinin inhibition assay is used to measure the titer of antibody against a strain of influenza virus present in the serum. Antibody levels are measured before vaccination and 28 days after. The antibody levels are used to compute the seroprotection and seroconversion criteria. 7–9, 17, 18, 21, 23

lymphocyte A lymphocyte is a type of white blood cell in the immune system of jawed vertebrates. Lymphocytes include T-cell, and B-cell. These cells work together in the adaptive immune response to generate antibodies against influenza . 4

monocyte Monocytes are a type of white blood cell. Monocytes and their macrophage and dendritic cell progeny serve three main functions in the immune system. These are phagocytosis, antigen presentation, and cytokine production. Phagocytosis is the process of uptake of microbes and particles followed by digestion and destruction of this material . 4, 30

mutation Mutation of genetic material occurs thanks to its chemical instability. The encoded protein molecules can have single amino acid (protein building block) change (minor, but still in many cases significant change leading to disease) or wide-range amino acid changes. 7

PBMC A peripheral blood mononuclear cell is any peripheral blood cell having a round nucleus. These cells consist of lymphocyte and monocytes. 8

ribonucleic acid virus(es) An RNA virus is a virus that has RNA as its genetic material. Inside a host cell this material is used to generate new virusses. Notable human diseases caused by RNA viruses include the common cold and influenza. 7

seroconversion and seroprotection A vaccine is considered succesful if the recipient seroconverted (4-fold or greater rise in antibody against virus after vaccination) and were seroprotected (GMT $\geq$ 40) after vaccination. . 7, 17–19, 23

STAT The signal transducer and activator of transcription (STAT) are transcription factors that work via JAK/STAT pathway regulating the expression of genes involved in cell survival, proliferation, differentiation, development, immune response, and, among other essential biological functions, hematopoiesis.. 8, 29, 30, 32, 33

T-cell A T cell is a type of lymphocyte. T cells are one of the important white blood cells of the immune system and play a central role in the adaptive immune response, for example generating antibodies against influenza. Groups of specific, T cell subtypes have a variety of important functions in controlling and shaping the adaptive immune response . 3, 4, 8

titer Titer is a way of expressing concentration. Titer testing employs serial dilution to obtain approximate quantitative information from an analytical procedure that inherently only evaluates as positive or negative. The titer corresponds to the highest dilution factor that still yields a positive reading . 3, 7, 8, 17, 19

TIV An inactivated trivalent vaccine is a vaccine consisting of antigenic virus particles from viruses that have been grown in culture and then killed to destroy disease producing capacity. In practice vaccines of three main types of influenza were used, hence trivalent. 7, 17

Data mining glossary

FluPrint Database unifying data on donors enrolled in different clinical influenza studies. 2, 5, 8–13, 15–17, 20, 21, 23, 32–34, 43

SIMON Follow up study performed by the creators of the FluPrint database. Applies sequential iterative modeling "overnight" (simon), which is an automatic machine learning pipeline to extract knowledge from clinical datasets. 2, 9–12, 23, 25–28, 32, 33, 37

Acronyms

GMT geometric mean titer. 4, 7, 17–19, 23

HA hemagglutinin. 3, 7

NA neuraminidase. 3, 7

RNA ribonucleic acid. 4, 7

1 Background

Influenza viruses are enveloped ribonucleic acid virus(es) (RNA virus(es)) and are divided into three types on the basis of antigenic differences of internal structural proteins (**fdaGuidanceIndustryClinical2007**). Two influenza virus types, Type A and B, cause yearly epidemic outbreaks in humans and are further classified based on the structure of two major external glycoproteins, hemagglutinin (HA) and neuraminidase (NA) (**fdaGuidanceIndustryClinical2007**). Type B viruses, which are largely restricted to the human host, have a single HA and NA subtype. In contrast, numerous HA and NA Type A influenza subtypes have been identified to date. Type A and B influenza variant strains emerge as a result of frequent antigenic change, principally from mutations in the HA and NA glycoproteins (**fdaGuidanceIndustryClinical2007**).

Since 1977, influenza A virus subtypes H1N1 and H3N2, and influenza B viruses have been in global circulation in humans. The current U.S. licensed inactivated trivalent vaccines (TIV) are formulated to prevent influenza illness caused by these influenza viruses. Because of the frequent emergence of new influenza variant strains, the antigenic composition of influenza vaccines needs to be evaluated yearly, and the TIV are reformulated almost every year. Currently, even with full production, manufacturing capacity would not produce enough seasonal influenza vaccine to vaccinate all those for whom the vaccine is now recommended (**fdaGuidanceIndustryClinical2007**).

1.1 Influenza mortality estimation

Previous works have applied models to estimate seasonal population influenza mortality (**zhouHospitalizationsAssocia greenMortalityAttributableInfluenza2013**; **iulianoEstimatesGlobalSeasonal2018**). Reported were age-specific influenza burdens that varied from one seasonal epidemic to another, that also varied per different region. However, consistently young children an elderly are found to be more susceptible to influenza and are advised to vaccinated annually (**zhouHospitalizationsAssociatedInfluenza2012**). Specifically, within the US based work of **zhouHospitalizationsAssociatedInfluenza2012**, the highest hospitalization rates for influenza were among persons aged $\geq$65 years and those aged $<$1 year.

Nevertheless, overall per age influenza burdens varied per season. Seasonal age variability was shown in **greenMortalityAttributableInfluenza2013**, where an age shift in Wales and England seasonal influenza burden was observed following the 2009 swine flue pandemic. It is also estimated that globally 291.243–645.832 influenza associated seasonal deaths occur annually (**iulianoEstimatesGlobalSeasonal2018**). These varying demographic statistics and the volume of influenza patients can confound decision making on national and international public health policies. Rapid knowledge extraction of vaccine efficacy data from clinical datasets and implementation of that knowledge can be a valuable asset for fighting future seasonal influenza outbreaks.

1.2 Vaccine success criteria

To implement a vaccine clinical efficacy needs to be assessed. However, due to the volume and vulnerability of population groups most at risk for influenza, the young and the elderly, a standard placebo controlled vaccine efficacy study is extremely costly (**zhouHospitalizationsAssociatedInfluenza2012**). Instead, the hemagglutinin inhibition assay (HAI) is used to estimate vaccine efficacy without requiring a placebo controlled study (**dejongHaemagglutinationinhibitingAntibodyInfluenza2003**). The criteria for a successful vaccine is an 4-fold increase in titer of the antibody against a strain of influenza virus and a geometric mean titer (GMT) of $\geq$ 40 28 days after vaccination, these are

called seroconversion and seroprotection. The last is estimated to reduce influenza risk by 50% (**dejongHaemagglutinationinhibitingAntibodyInfluenza2003**).

1.3 Finding immunological factors predicting high vaccine response using machine learning

It is known that pre-existing T-cell populations are correlated with an antibody response after vaccination. But, the role of different T-cell populations in mediating that response is uncertain. In one work it was found that under certain circumstances CD8+ T-cells specific to a conserved part of viral antigens correlated with protection against symptomatic influenza (**sridharCellularImmuneCorrelates2013**). In other work, different populations of CD4+ T-cells that associated with protective antibody responses after seasonal influenza vaccinations were found (**bentebibelInductionICOSCXCR3**). Others, report non-increased CD8+ T-cell populations and an increased CD4+ T-cell population after vaccination (**trieuLongtermMaintenanceInfluenzaSpecific2017**). It was also reported that repeat vaccinations are an important factor in maintaining CD4+ T-cell population (**trieuLongtermMaintenanceInfluenzaSpecific2017**). How exactly these T-cell populations work together to form a protective influenza immunity and vaccination response is not well understood.

Another known factor is that influenza virus infection stimulates various intracellular signaling pathways (**Zhang˙2019**). These pathways are important for viral entry, replication, and propagation, and are involved in host antiviral response, but how these pathways lead to a fully realised vaccine response is not well understood (**Zhang˙2018**). The activation of these pathways is commonly meditated by the phosphorylation and dephosphorylation of several proteins, including signal transducers and activators of transcription (STAT). One example is the JAK-STAT signaling pathway in B-cells, where a large set of B-cell receptors is known to bind cytokines produced by CD4+ T-cells and this results in downstream biological processes that make the immune response to a vaccination (**Papin˙2004**). Further, these pathways are found in all peripheral blood mononuclear cell (PBMC) and control a great amount of biological programs (**Cantrell˙2015**). In general, the phosphorylation pattern of these pathways in PBMCs are used in clinical studies as a measure of cell activation in response to cytokine stimulation (**Toapanta˙2012**; **tomicFluPRINTDatasetMultidimensional2019**).

Machine learning has been applied to clinical datasets to find influenza protection markers, such as the described T-cell populations, titers of PBMCs and related molecules, or cytokine signalling related activity (**furmanApoptosisOtherImmune2013**; **sobolevAdjuvantedInfluenzaH1N1Vaccination2016**; **tsangGlobalAnalysesHuman2014**). However, these studies suffer from multiple issues, such as: inconsistencies between findings depending on the epidemic season, only focussing on one type of biological assay to get data, and a low amount of donors/samples. Furthermore, a successful vaccination is often not well defined within one study and the definition might differ between studies. To reduce these issues and to facilitate data mining of clinical studies the FluPrint database was created. The FluPrint database consists of preprocessed data from multiple clinical studies that span different years and data types, and in FluPrint enrolled donors are classified as high or low responders according to HAI outcomes (**tomicFluPRINTDatasetMultidimensional2019**).

1.4 Bussiness objectives

As described, due to the high volume population that needs vaccines and the rapidly changing nature of the influenza virus, rapid vaccine efficacy knowledge extraction is important. Moreover, identifying patterns between vaccine response and immune correlates furthers the understanding of the underlying immunological mechanism of influenza protection.

This work uses the FluPrint database, which aims to solve data quality issues and low dimensionality of prior studies using clinical datasets comprising different virus, cell and serum sample data types. Specifically, it does so by incorporating eight clinical studies conducted between 2007 to 2015 using in total 740 patients, spanning different types of assays. Further, it preprocesses the data, and provides a binary classification of enrolled donors into high- or low-responder to a vaccine if HAI data is available. This facilitates data mining studies that can identify patterns based on donor vaccine response in the multi-dimensional data collected from multiple clinical studies.

The work done here is structured around providing insight into these questions using the FluPrint database:

- What kind of studies can be done using the FluPrint database?
- What immunological factors correlate to a vaccine responses?
- What is the effect of repeat vaccination?

Since this work is an independent study performed for an assignment, the success criteria for these objective will be loosely defined as providing a statistical description or to provide insight in the questions posed in the objectives.

The rationale behind these questions and success criteria is the limited scope of this project, it is a short 3EC project as part of the Applied data science profile meant to show the ability to use data mining tools. The FluPrint paper on which this work is based provided the directions of these research questions, and a part of this work was to replicate and extend the work done by the authors of FluPrint and their follow-up work in the SIMON paper .

2 Assess situation

2.1 Data and knowledge sources

The sole source of data used in the project was provided by **tomicFluPRINTDatasetMultidimensional2019** (this work is referred to using: "the FluPrint paper" from now on). The FluPrint paper described the MySQL database for which the installation was described in the FluPrint Github Repository. A template query is provided on the github page belonging to an unpublished follow-up study by the same authors of the FluPrint paper (Listing 3). According to the authors, the data belonging to this data is the most interesting for the business objective of finding repeat vaccination effects and will be used in this work too (this unpublished follow-up study is referred to using: "the SIMON paper "). The authors give this brief description of the data:

> *"The influenza datasets were obtained from the Stanford Data Miner maintained by the Human Immune Monitoring Center at Stanford University. This included total of 177 csv files, which were automatically imported to the MySQL database to facilitate further analysis. The database, named FluPrint and its source code, including the installation*

tutorial are freely available here and on project's website. Following database installation, you can obtain data used in the SIMON publication by following MySQL database query (Listing 3)".

2.2 Tools and techniques

Installation of the FluPrint database requires an installation on a unix operating system of MySQL, PHP. More details are at the FluPrint Github Repository.

Database querying was done using a neovim based toolset, personal configuration can be found here.

Since in the FluPrint paper and the SIMON paper R is used, it was also used in this work. Especially crucial was the R package mulset, which was made by the authors of the SIMON paper . This package was used to deal with missing data between different clinical studies and years, and thus was used to generate complete data tables in this paper too. All scripts in this work were written using tidyverse packages and make heavy use of the dplyr package for data wrangling. Additionally the following packages were used:

- ggpubr for making publication quality figures,
- the kable function from knitr to generate latex tables,
- caret and MLeval to streamline model training and evaluation,
- corrplot to visualise correlation between features, and other packages that were used only once.

2.3 Requirements of the project

Requirements of this work are to show ability in using data science methods. And, due to the scope of the project most of the insights gained include replication of the work of the FluPrint paper and the SIMON paper . Rather, all the scripts generated and analysis performed done are original work and are supplied together with the final deliverable (subsection 9.3).

Since the data used here was stored in a MySQL database this makes it more complicated for an examinator to reproduce all code, especially since installing the database requires a unix operating system. This is not considered problematic since the flat data files from the database used in R scripts are included.

Reporting of the project aimed to follow the CRISP-DM methodology, where at each stage of the project a separate report was written during the analysis work. In the end the most important information was kept and incorporated in a final report that was assumed to be the graded work.

2.4 Assumptions of the project

This work assumed that the focus point of the evaluation lies on the methodology used, and the ability to apply the basic data science methods learned in the Applied Data Science profile. The answer to business objectives is assumed to be subjective, and it is assumed that the methods used and clarity of insights into the data gained are more important.

It is also assumed that the FluPrint database and other methods used in the FluPrint paper and the SIMON paper are of high quality, and that this is appropriate for this work. Out of the

scope of this work was investigating whether the preprocessing done for the data in the database is valid, since we are not domain experts. A method for querying, and generating complete data tables was provided by the authors of the FluPrint paper and this was also be used in this work. It was assumed that provided SQL and R methods (in particular the mulset R package) in question are allowed to be used as a starting point in this assignment.

2.5 Constraints of the project

This work is an unsupervised assignment, and only personal hardware were available. This put constraints on dataset size and computational requirements of analyses. The work was done on a Macbook air (2017) with the OSX big-sur operating system. This means that unix tools were available and there were no technical constraints. The relevant filetypes to reproduce this work were csv files generated by the SQL server, and R scripts.

3 Data mining goals

3.1 Translating the problem in data mining terms

All business objectives described involve querying data from the FluPrint database. In this work we first and foremost explore the database, and lastly we apply classification models and feature selection methods on the most interesting dataset.

The business objectives can be translated in data mining terminology like so:

- Explore and describe the database and corresponding tables.
- Apply wrapper feature selection to the most interesting datasets.
- Explore features identified by the models trained in the wrapper feature selection.

In data mining terms, the problem type was a combination of exploratory data analysis and classification. Since this work was for a 2 to 3-weeks/3EC assignment for the Applied Data Science profile, success criteria for all goals are subjective. For the classification type goals we followed the model evaluation procedure used the SIMON paper , models were evaluated using the AUROC metric, other confusion matrix metrics were also reported.

3.2 Project plan

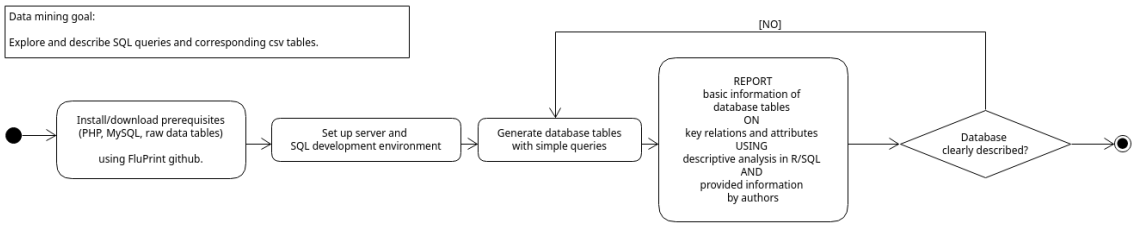

Figure 1: Project plan for the SQL related data description goal.

The first part of the project involved querying the database, and collecting and describing the available data (Figure 1). The first goal was to understand the tables in the SQL database, their key relations, and to describe the attributes within the tables. Valuable info on this part was already provided in the the FluPrint paper , but it was also investigated in this work. The tools that will be used are SQL for querying and R for statistical descriptions.

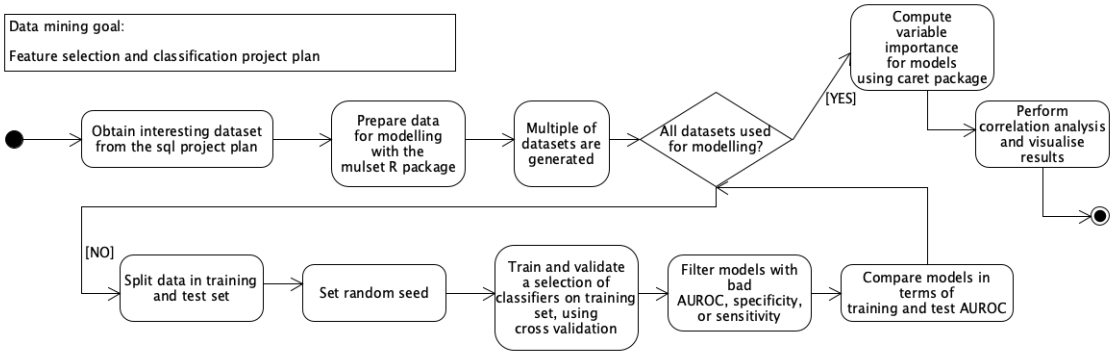

Figure 2: Project plan for the classification and feature selection data mining goals.

For the modeling and feature selection data mining goals the plan was to implement a simplified version of the automatic feature selection pipeline described in the SIMON paper . In the SIMON paper a large set of classifier models was automatically trained and evaluated on a set of small datasets generated from the FluPrint database, the trained models were then used for feature selection. Rather, in this work we manually trained a small selection of models on the same datasets as in the SIMON paper towards providing insight in a specific research question (Figure 2). Further, not all generated datasets were included in the final analysis since we were interested in repeat vaccination data which was not always available.

4 Data description

4.1 Volumetric analysis

Data in the FluPrint database on individuals enrolled in influenza vaccine studies was collected from the Stanford-LPCH Vaccine Program, the data was archived at the Stanford Data Miner archive. The archived data was filtered for a selection of interesting 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 were described in the table accompanying the online publication of the FluPrint dataset, and relevant parts are in the appendix (Table 9).

From those 740 donors a vaccine response classification was only given for 372 donors (Figure 3), these classifications are discussed more in depth later. Further, there was no major difference in demographic statistics when stratifying the data in high or low responder classification (Figure 3).

In the FluPrint paper it was reported that in all studies the donors are only vaccinated once, except in the study SLVP015 (Table 9). However, in the other work of the same authors, the SIMON paper , it was claimed that vaccines were administered in multiple influenza seasons for multiple

studies. Rather, while true that the majority of data on donors that received repeat vaccinations spanning influenza seasons comes from one study there are two more studies containing repeat vaccination data (Figure 8) (SLVP015, SLVP021, and SLVP029 in (Table 9)).

The aggregated donors for which a vaccine respone classification waas available from all clinical studies span a wide age range (Figure 3, B) from 1 - 50 (Table 1), in the original work the demographic statistics include the donors for which no vaccine response classification is given, therefore they report a greater range of 1-90. Considering that data on donors with missing classifications are not included for analyses they are left out of the demographic description.

Demographic attributes that were available include gender, ethnicity, and cytomegalovirus (CMV) status, database specific representation of the values of demographic attributes are described the FluPrint paper . Further, like in the FluPrint paper stratifying the donors on vaccine response did not affect the overall distributions of demographic attributes (Figure 3, A). However, including only the donors for which a vaccine response classification was available made the maximum age lower in the high responders group (Figure 3, B).

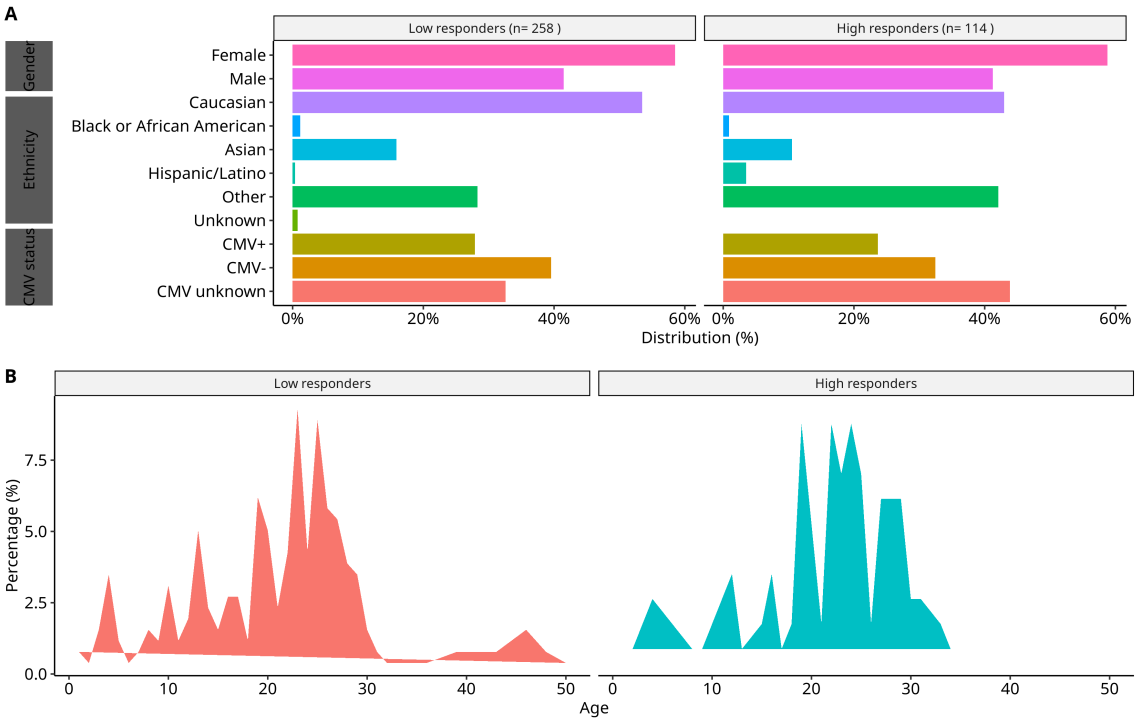

Figure 3: **Demographic attribute distributions and age distribution. A.** percentage of donors/rows having some Gender, Ethnicity, or CMV status within high and low responder groups. **B.** Age distribution of donors with available response classification.

The data from the clinical studies consisted of 121 CSV files that were imported into the FluPrint database. The data was used to build four tables, three of which were described but we omitted the discussion of technical validation details of the database construction. The relation between the tables is best visualised using the schema given in the FluPrint paper , it describes the MySql

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 1: Demographic statistics of donors with known vaccine response classification.

attribute types and columns in the tables (Figure 4) (copied). The volume of the data is also given in the FluPrint paper , per table the number of rows and columns was reported (Table 2).

Table name	**Rows**	**Columns**
donors	740	6
donor_visits	2,937	18
experimental_data	371,260	9
Medical history	740	18

Table 2: Volume of tables in the Fluprint database.

4.2 Attribute types and values

Because of the great number of attributes in the database, we discuss them by table starting with the donors table (Figure 4).

4.2.1 donors table

The *id* attribute is simply an enumeration of unique donors, additionally it is used as a key to get attributes from other tables. The *study_donor_id* attribute is an encrypted identification number. Each donor belongs to the study identified by the *study_id*, these are the last two digit of the name code (those starting with SLVP0 ··) in the reference table (Table 9), the *study_internal_id* is either the digit or a string containing the digit in *study_id*. The *gender* and *race* attribute contain the values used in (Figure 3), a minor note is that in the original paper "American Indian or Alaska

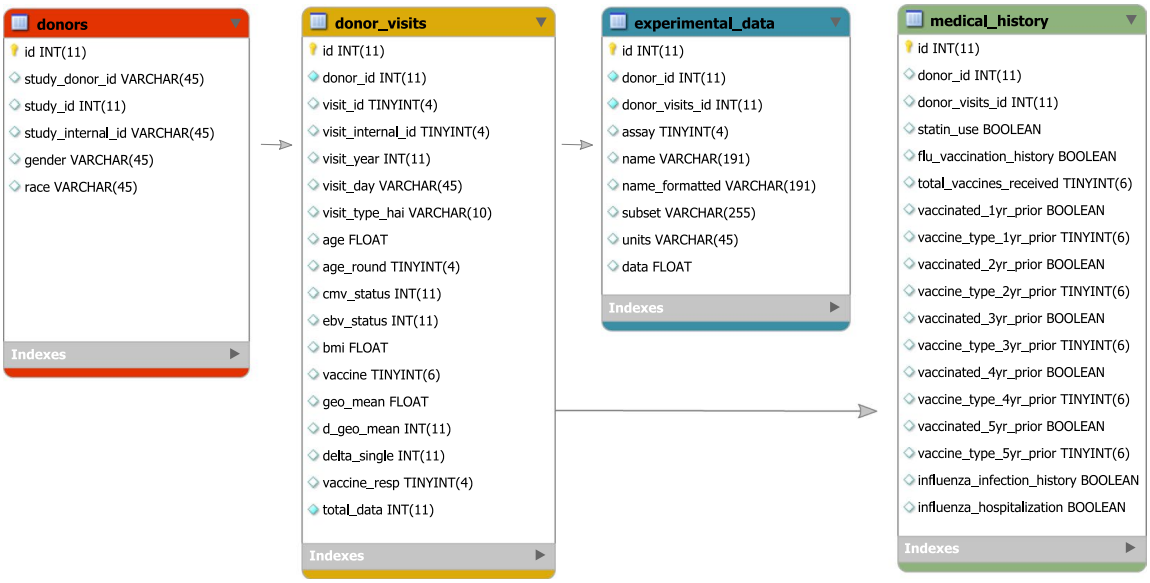

Figure 4: **(taken from original paper)** The FluPrint database model. The diagram shows a schema of the FluPrint database. Core tables, donors (red), donor_visits (yellow), experimental_data (blue) and medical_history (green) are interconnected. Tables experimental_data and medical_history are connected to the core table donor_visits. The data fields for each table are listed, including the name and the type of the data. CHAR and VARCHAR, string data as characters; INT, numeric data as integers; FLOAT, approximate numeric data values; DECIMAL, exact numeric data values; DATETIME, temporal data values; TINYINT, numeric data as integers (range 0–255); BOOLEAN, numeric data with Boolean values (zero/one). Maximal number of characters allowed in the data fields is denoted as number in parenthesis.

Native" is listed as one of the *race* values but is not used in the database, race attribtue processing is described more in the FluPrint paper . There are 5 donors whose race is "NULL", which are mapped to unknown (Figure 3).

id	study_donor_id	study_id	study_internal_id	gender	race
1	e27ad74ff9a5f2f32d8e852533f054c0	30	30	Female	Asian
2	4a89ac4d3f4dc869e5c8e8cf862cffda	30	30	Male	Other
3	a2cde6e54dec92422b0427dd49244350	30	30	Female	Caucasian
4	0f7d8d1c13e876017ea465f99d25581f	30	30	Male	Other
5	1ed2f6409584b7b4e9720b28d794fe91	30	30	Female	Caucasian
6	a575678405e9615bfb87eccfa031f7fc	30	30	Male	Other

Table 3: Head of the donors table.

4.2.2 donor_visits table

The donor visits table is the core table of the database, it contains donor attributes at visit times during enrolment in clinical studies in rows that are uniquely identified by an *id* integer. Each row also includes the *donor_id* identify the donor that visited by the *id* in the donors table.

The database combines datasets from multiple clinical studies spanning multiple years. Within clinical studies the data is often incomplete due to factors that change between influenza seasons, such as changes in the number of features measured in an assay data collected. As a result, the FluPrint database is incomplete and contains heterogeneous data quality (Table 4). More, every attribute in the core table has missing values, which complicates selection of data for further analysis. In summary, the number of visits is inconsistent per season per donor, all columns are incomplete, and classification is sometimes based on single visits or inconsistent with available data (Table 5) (Table 4).

stat	age	cmv_status	ebv_status	bmi	vaccine	geo_mean	d_geo_mean	vaccine_resp	total_data
n	2937.0	1081.0	548.0	516.0	2794.0	984.0	1260.0	1206.0	2937.0
na	0.0	1856.0	2389.0	2421.0	143.0	1953.0	1677.0	1731.0	0.0
mean	47.3	0.4	0.8	24.8	3.7	87.6	8.9	0.3	126.4
sd	27.0	0.5	0.4	5.6	1.0	101.7	30.9	0.4	368.4
se_mean	0.5	0.0	0.0	0.2	0.0	3.2	0.9	0.0	6.8
IQR	50.2	1.0	0.0	6.7	0.0	105.4	4.0	1.0	19.0
skewness	0.2	0.3	-1.4	1.0	-1.7	3.6	9.9	1.1	7.1
kurtosis	-1.5	-1.9	-0.1	2.1	3.0	26.6	114.9	-0.9	49.7

Table 4: Descriptive stats of relevant numeric or binary factor columns in the donor visits table. For geo_mean 0 is considered as missing data.

Per donor all visits are enumerated in chronological order by *visit_id* (Table 5). Further visit info includes: *visit_internal_id* which is a number that indicates the visit order within an influenza season but this differs per clinical study (e.g. some use 1-2-3, orther use 0-7-28), the *vist_year* is the influenza season of the visit, the *visit_day* is the number of days relative to the date of vaccination, *age* and *age_round* indicate the donor's age at time of the visit, and *bmi* gives the donor bmi at visit time, and lastly *visit_type_hai* is the intent of the visit which is either "pre", "post", "other", or "single".

Depending on clinical study, during the "pre" visit a virological assay is performed to determine the CMV and Epstein-Barr virus (EBV) status of the donor, which are indicated by the binary variables *cmv_status* and *ebv_status*.

In most clinical studies, vaccine response was measured using the HAI assay. The procedure measures the influenza antibody titers before vaccination during the *visit_type_hai* "pre" visit of a participant, and 28 days after vaccination during a "post" visit. The GMT at each visit is calculated, and a fold change in GMT is calculated as the ratio of the GMT at day 28 (post) and during the first visit (pre). These values are called *geo_mean* and *d_geo_mean*, respectively. Lastly, there is one more HAI related data attribute which is the *d_single*, this is reported as the antibody titer fold-change per strain of virus used in the vaccine. It is unclear how this value is aggregated over different influenza strains in a TIV and is left out of further analysis. Based on these HAI related attributes a donor is classified as high or low responder, the seasonal vaccine response classifications are given by the *vaccine_resp* attribute.

The type of vaccine used in a study is indicated by the *vaccine* attribute, the meaning of the vaccine id is reported in the appendix (Table 10). The type of experimental assays performed to measure the immunological profile of the donor during the "pre" visit are described later in the section of the experimental data table. All assays are listed in the FluPrint paper and are summarised here (Table 6). This information is relevant to *total_data* attribute of the donor visits table which indicates the number of measurements made during a visit.

visit_id	year	day	type	age	cmv	ebv	bmi	vaccine	geo_mean	d_geo_mean	response	assay_data_rows
1	2011	0	pre	20	1	1	30.31	4	25.20	6	0	343
2	2011	7	other	20	1	1	NULL	4	0.00	6	0	51
3	2011	28	post	20	1	1	NULL	4	160.00	6	0	51
4	2012	0	pre	21	1	1	30.31	4	9.28	4	0	292
6	2013	0	pre	22	1	1	30.31	4	15.91	2	0	2877
7	2013	7	other	22	1	1	NULL	4	0.00	2	0	63
8	2013	28	post	22	1	1	NULL	4	26.75	2	0	82

Table 5: Visit data of donor 166 from study SLVP021 (Table 9). Note that the number of visits and volume of data collected at visit varies per season. Further, the classification is inconsistent with seroconversion and seroprotection criteria in 2011.

The most important data related to the visits of donor 166 is shown in Table 5. As described above, the vaccine response classification is determined per season based on the GMT in the "pre" and "post" visits. However, since the HAI assay requires a "pre" visit and a "post" visit 28 days later to measure the difference in GMT, a classification is inconsistent when there is only one visit

record in a season (Table 5).

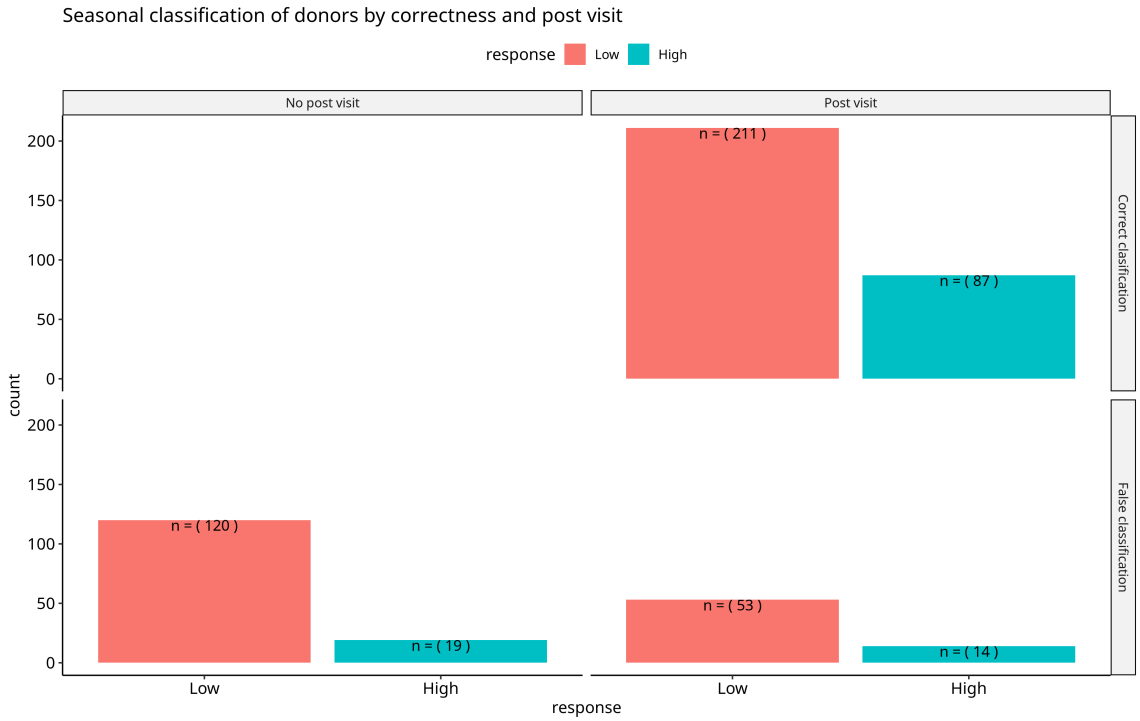

Figure 5: **Classifications inconsistent with seroconversion and seroprotection criteria using given data.** The classifications given in influenza seasons with only a visit with *visit_type_hai* value of "single" or "other", or did not have "pre" and "post" visits were considered inconsistent with the HAI procedure. Additionally, those that did not meet the seroconversion and seroprotection criteria given the GMT data were considered inconsistent.

Furthermore, the example of donor 166 contains another type of inconsistency in the classification, in 2011 the GMT *geo_mean* increases from 25.20 to 160.00, and the *d_geo_mean* is 6, but in this season the donor is classified as a low responder, even though seroconversion and seroprotection criteria are met (Table 5). Because of these apparent inconsistencies the seasonal classification of donors was evaluated using available GMT data and the seroconversion and seroprotection criteria 5, records of incorrectly labelled donors are also saved as a spreadsheet. Given the information in the database classification is inconsistent in a large number of cases. However, the most likely explanation is that antibody titer for one specific strain of virus in the vaccine did not meet the seroconversion and seroprotection criteria. Therefore, in this work it was considered as inconsistent because classification required data not given in the database. Nevertheless, the classification is not necessarily incorrect and the classification data was used in this work without any modifications.

4.2.3 experimental_data table

Name	**Description**	**id** (*experimental_data.assay*)
(Multiplex) cytokine assays	Multiplex ELISA using Luminex polysterene bead or magnetic bead kits. Measures serum cytokine/hormone level in z.log2 units using fluorescent antibodies.	3, 6, 15, 16
Flow and mass cytometry assays	uses labeled antibodies to detect antigens on a cell surface to identify a subset of a cell population, units are in percentage of parent population.	4, 9, 13, 17
Phosphorylation cytometry assays	Uses antibodies to measure phosphorylation of specific proteins stimulated by an immune system event belonging to cell population subsets. Units are a fold change between stimulated and unstimulated cells, for mass cytometry arcsin readout difference, fold-change of 90th percentile readout values otherwise.	7, 10 (mass cytometry) (flow cytometry)
complete blood count (CBCD)	Different cells are counted using flow cytometry Units are usually in Count/μL	11
meso scale discovery assays (MSD)	A setup where serum cytokines or hormones are captured with antibodies, and then detected by using a detection antibody. Units are arbitrary intensity	2, 12, 14

Table 6: Table containing the types of data collected at donor visits. It describes the assay type in the name and description columns. The id column refers to the different specific assays belonging to a data type with the id used in the database. The mapping from id to assay can be found in the appendix (Table 11).

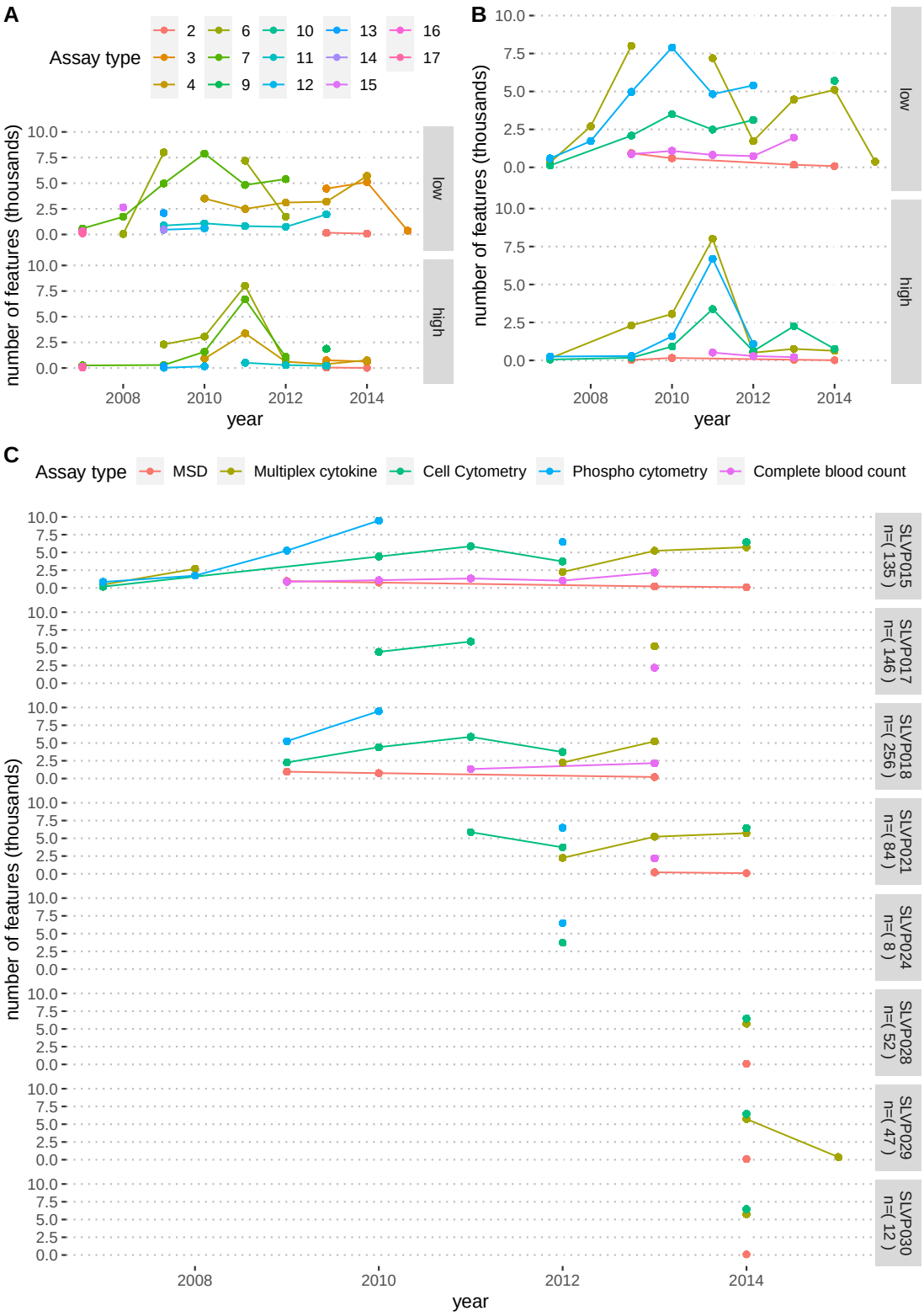

Figure 6: **Description of data volume collected in different years and studies, by data type or experiment id and stratified by classification.** **A.** the aggregated number of data points/features measured per season and by classification. Data is shown per experiment id used in the FluPrint database, indicated with color. **B.** Same as (**A**), but grouping experiments per data type instead of experiment id, the same as in (**C**). **C.** Seasonal data points per data type by study instead of classification.

As reported in the FluPrint paper different assays performed in clinical studies are remapped to and *id* number, but the values in the database do not correspond to the reported remaps (Table 10). The actual data type, units, and assay id contained in the database were described in this work (Table 6).

In total there are 14 different experimental assays used across clinical studies, not counting the virological and HAI assays (Table 6). Further, the virological assays determining the CMV and EBV status are not used in this work, since it is available only in a small subset of the collected data. Those 14 assays have been aggregated in this work to 5 different data types/experiment types (Table 6), in short:

- the multiplex cytokine assays measure levels of molecules such as cytokines and other signaling molecules in human serum/blood,
- flow and mass cell cytometry measure the phenotype of specific immune related cells,
- phosphorylation flow and mass cytometry measures signaling pathway activation after an induced cytokine stimulation or the absence thereof,
- the complete blood count (CBCD) measures the concentration of cells in the serum/blood,
- and meso scale discovery (MSD) measures hormones or cytokines from human serum/blood.

The experimental data table contains all features recorded per donor visit. The number of features collected for each visit is large and varies greatly (mean at 126 , ±368 SD) (Table 4), and in total there are 3285 different features measured across all clinical studies. However, not every assay is done in every clinical study and over the years the data generated by assays has changed, so a table with all features as columns and all donors as rows would be extremely sparse (Figure 6). Describing the 3285 different features in this sparse table would be impossible, but assay value distributions across studies are shown to follow normal or power distributions (Figure 7).

In addition to the sparseness of data, what further complicated selecting relevant data is repeat visits of donors, and missing visits. The problem of repeat visits over a span of multiple influenza seasons is the change in the data type collected per season, and that repeat visits are only a small portion of the database. Furthermore, the potential for studying the effect of repeat vaccination on high versus low vaccine response classification is limited, since the classification in the longitudal study (SLVP015) containing repeat visit data is not available in a majority of data points (Figure 8). As a result, exploring what effect repeat vaccination has on vaccine response was done in this work using the small subset of donors where a classification was available in two influenza seasons.

4.3 Data quality

The database has issues that are inherent to combining multiple studies. Firstly, the vaccine response classification was inconsistent with the given data in some cases (Figure 5). Secondly, the classification was often missing completely because no HAI assay data was available. Thirdly, the classification was set to a null value by the database authors because possibly the antibody titer for a single strain of virus in the vaccine was too low (this data is not in the database) (Figure 8). Lastly, the data is highly sparse when considering data collected on donors in different studies or in different influenza seasons.

The value of the database in terms of knowledge is the great amount of assay data that was collected in different studies across years and was preprocessed. But this information is hard to

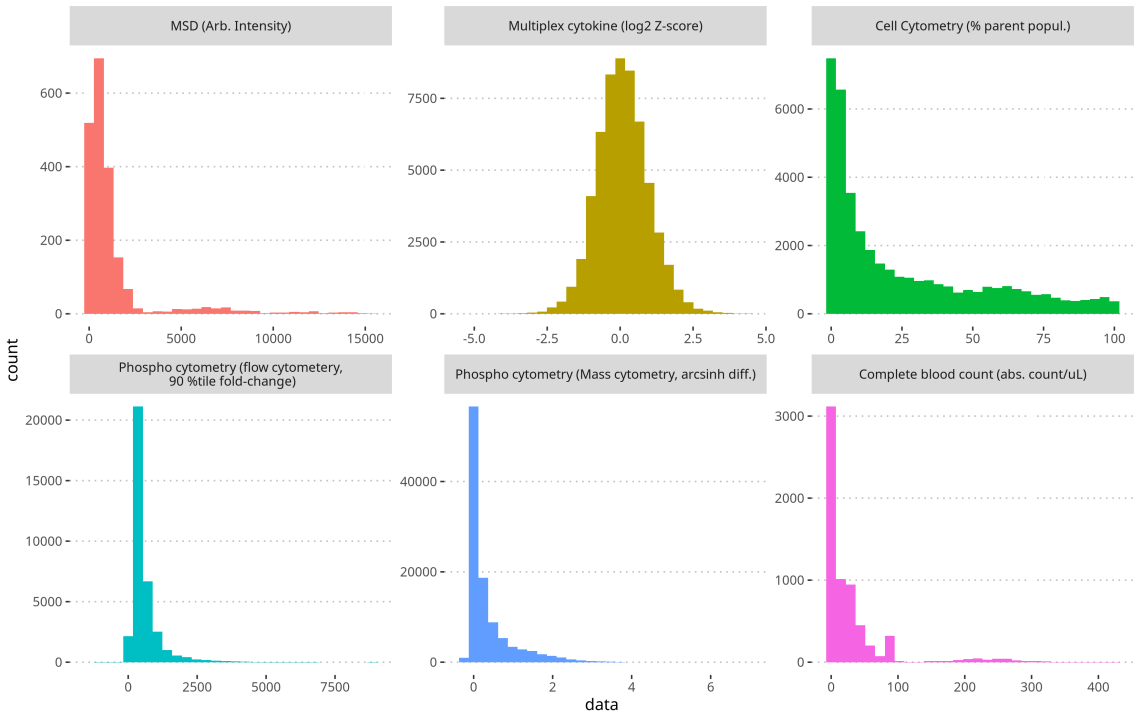

Figure 7: **Distributions of experimental data values.** In this work we grouped experiments into five datatypes, however the phopho(rylation) cytometry data was measured in two different experiments and thus had two different units. The process that is measured is the same between the assays, only the experiment differs. As a result, six different distributions are shown, one for each unit of measurement in the FluPrint database. Importantly, there were outlier values for the phosphorylation flow cytometry 90th percentile values that were removed to show the overall distribution.

Figure 8: The number of donors that visited per number of influenza seasons they visited (years), per study. The color indicates the number of visits for which a classification was available, counted within the groups of donors that visited the same amount of times.

access since all studies do not use different assays (Figure 6), resulting in high sparsity data. Further, the sample size that can be used for further classification studies is limited, since the high versus low vaccine response is only available for a minority of the data points.

5 Data preparation

5.1 Data selection

The data used in this work was based on the data used in the SIMON paper (Listing 3). Using this query template generates a subset of FluPrint comprised data from 5 clinical studies using the same vaccine type (Table 10, Vaccine id 4), most importantly the longitudinal study SLVP015 (Table 9). Presumably the authors of the SIMON paper included only the first visit of donors because the classification is the most complete in this dataset (Figure 9). In this work we firstly use this query to generate initial first visit datasets. Secondly, to explore repeat vaccination, we select a subset of this data that includes only donors with a repeat visit in a second influenza season.

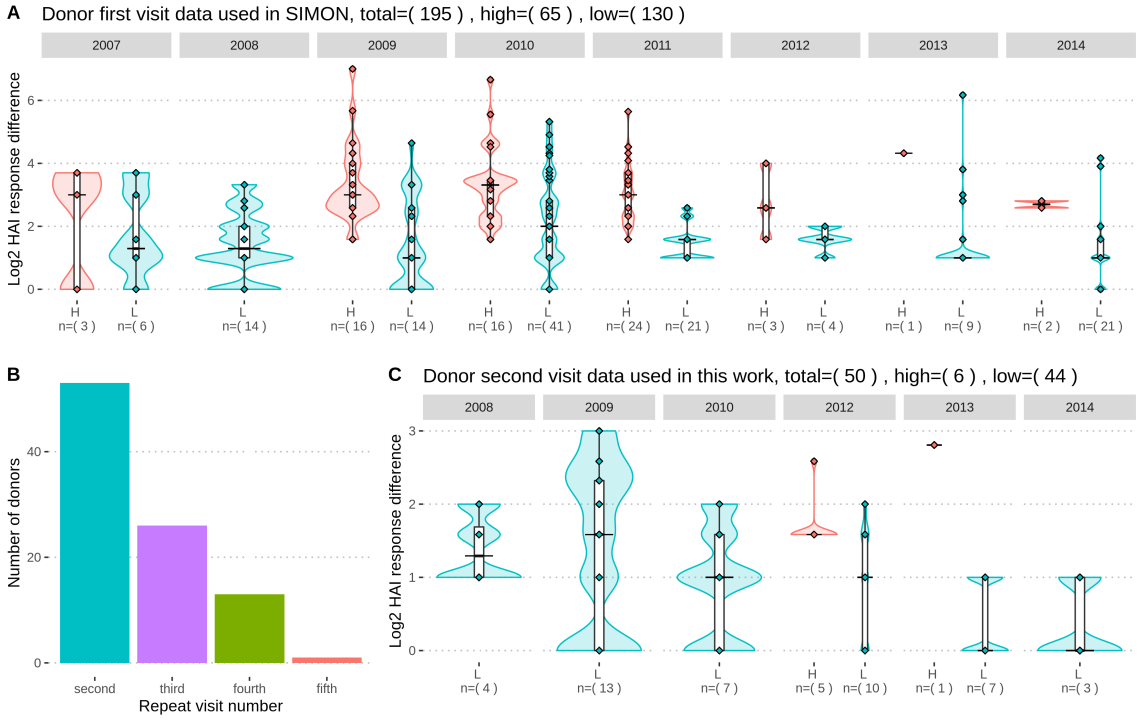

Figure 9: **Description of data used in this work. A.** HAI response distributions of high and low vaccine responders of the data used in this work and in the SIMON paper . Referred to as the first-visit data . **B.** Number of donors for which data was available in multiple influenza seasons. **C.** HAI response distributions of high and low vaccine responders of the donors that got a second vaccination. Referred to as the second-visit data .

The initial query generated a long table for in total 3285 different features recorded at the first

visit of 195 donors in different studies and years (referred to as first-visit data). An observable pattern in this data is that low responding donors are overrepresented and that the classification is mostly consistent with the seroconversion and seroprotection criteria as seen by the log2 GMT change feature (Figure 8, A). Unfortunately, the number of donors in the first-visit data that returned in other influenza seasons decreases quickly, limiting possibilities of comparing models built on first-visit data and subsequent visit data (Figure 8, B). The selected second-visit data lacks the high response class (except for 6 donors) precluding training any models, therefore we only used the second-visit data to explore the knowledge gained by models trained on the first-visit data (Figure 8, C).

Listing 1: Applying the mulset algorithm and preparing the data

```
generate intersection datasets suitable for analysis
for {each donor in data} do:
    Calculate intersection between donor and all other donors using
        ↪ mulset algorithm
    Skip sets that have less than 5 features and less than 15 donors
        ↪  in common
end for;

for {each set in generated datasets} do:
    Partition data in training (75\%) and test (25\%) split
    Skip sets that have less than 10 donors in the test set
end for;

for {each set in prepared datasets} do:
    calculate number of donors that visited a second influenza
        ↪ season
    skip dataset if it is not in the top3 of datasets containing the
        ↪  highest number of second visitors
end for;
```

The first-visit data used for modeling and feature selection had a total of 640575 cells of which 596736 values were missing (sparsity of 93%). In the SIMON paper missing values in this data is not imputed because there is not enough prior knowledge. And, since every donor had a missing feature, dropping all rows/donors was not an option either. A solution used in the SIMON paper was generating complete tables comprising small subsets of donors that had all features in common using the mulset algorithm (Listing 1) (Figure 15).

In this work the procedure in the SIMON paper was replicated and extended to generate usable datasets for feature selection (Listing 1). First, there were duplicate measurements of features in the first-visit data , these were aggregated to unique feature records using the mean. Second, the mulset R package was used to generate 47 complete datasets. These datasets were then reduced to 36 by selecting those that had at least 5 features and 15 donors. Finally, the datasets were split into train (75%) and test (25%) sets, and datasets with less than 10 donors in the test set were discarded reducing the number of datasets to 20 (Table 7).

A significant number of datasets contained more predictors than samples (Table 7). However, we consider this as inevitable and not an absolute obstacle since the purpose of the models is not

dataset	Rows x Cols (Donors x Features)	total (low / high (low %))	train (low / high)	test (low / high)
1	61 x 78	43 / 18 (0.7)	33 / 14	10 / 4
2	105 x 101	62 / 43 (0.59)	47 / 33	15 / 10
3	140 x 50	94 / 46 (0.67)	71 / 35	23 / 11
4	63 x 269	38 / 25 (0.6)	29 / 19	9 / 6
5	62 x 293	38 / 24 (0.61)	29 / 18	9 / 6
6	68 x 237	42 / 26 (0.62)	32 / 20	10 / 6
7	67 x 44	47 / 20 (0.7)	36 / 15	11 / 5
8	111 x 93	66 / 45 (0.59)	50 / 34	16 / 11
9	73 x 54	58 / 15 (0.79)	44 / 12	14 / 3
10	40 x 105	28 / 12 (0.7)	21 / 9	7 / 3
11	46 x 97	32 / 14 (0.7)	24 / 11	8 / 3
12	137 x 53	78 / 59 (0.57)	59 / 45	19 / 14
13	48 x 42	35 / 13 (0.73)	27 / 10	8 / 3
14	**91 x 38**	**62 / 29 (0.68)**	**47 / 22**	**15 / 7**
15	42 x 37	36 / 6 (0.86)	27 / 5	9 / 1
16	**92 x 26**	**62 / 30 (0.67)**	**47 / 23**	**15 / 7**
17	88 x 6	68 / 20 (0.77)	51 / 15	17 / 5
18	82 x 87	56 / 26 (0.68)	42 / 20	14 / 6
19	**151 x 51**	**92 / 59 (0.61)**	**69 / 45**	**23 / 14**
20	83 x 75	56 / 27 (0.67)	42 / 21	14 / 6

Table 7: Datasets generated by applying the mulset algorithm on the first-visit data also used in the SIMON paper , and the balanced train test splits that were performed.

to discriminate vaccine responders with the highest accuracy, but to identify features that correlate with a vaccine response from the great number of features.

In this work we selected only the top 3 datasets that best fit to the business objectives of exploring repeat vaccination, as well as finding features that correlate with vaccine responses. Accordingly, for each dataset we calculated the number of donors that visited a second influenza season, and chose the three with the highest number. The resulting selected datasets for further analysis were 14, 16, and 19 (Table 7, **bold rows**). These datasets had, respectively, 27 out of 91, 27 out of 92, and 21 out of 151 donors that returned for a second vaccination.

Within all three datasets 82 of the donors are shared indicating that using both dataset 14 and 16 might add little additional information, and that all three datasets contain a lot of the same information. Furthermore, 26 of the measured features are shared between dataset 14 and 16, meaning that all features of dataset 16 are in dataset 14 and that dataset 16 is almost a subset of dataset 14. More, all features were from the same phosphorylation flow cytometry data type (Table 6). Nevertheless, in lack of a better alternative and despite these issues, these three datasets were chosen for further analysis of repeat vaccination.

5.2 Data cleaning

In this work features and rows were not changed for the chosen datasets, since this would result in a lower number of rows/donors. As a result, noisy data points were included in the training and evaluation data of the models. Furthermore, since the objective of this work is not obtaining optimal models but exploring repeat vaccination and vaccine responses, this is not considered problematic.

6 Modelling

6.1 Choice of modeling technique

In this work a form of wrapper feature selection is used, since we are training models on different subsets of features and chose those models that discriminate the best between low and high vaccine responders (**hiraReviewFeatureSelection2015**). Although, the aim is to train an at least fair discriminator on any dataset and to then use that model to identify any new knowledge about vaccine response and repeat vaccination.

Four models were chosen for this task: the naive bayes classifier (nb), the random forest model (rf), the regularised logistic regression model (reglog), and regularised linear discriminant analysis (rrlda). In the SIMON paper an automatic machine learning pipeline is used where 2400 models are trained on all 20 datasets, and the best models are then used to explore important features that correlate with a high vaccine response. This approach is out of scope for this work, and instead we change the objective to specifically identifying repeat vaccination effects. Additionally, the datasets chosen for analysis in this work are not discussed in the SIMON paper , so this is a novel analysis using a similar procedure.

6.2 Test design

First, the three selected datasets were split in test and training sets. Secondly, the training set was used for training models using 2 times repeated 10 fold cross-validation where the accuracy was computed on every fold. Thirdly, the models that had the best cross-validated accuracy were

compared using the training and test area under the curve measure (AUC), since we are interested in general discriminative ability. Using these measures the best discriminator was chosen for further exploration repeat vaccination and vaccine response features.

6.3 Model parameters and assessment

dataset	model	SENS	SPEC	MCC	PREC	NPV	FPR	F1	TP	FP	TN	FN	train AUC	test AUC
14	rrlda	0.091	0.915	0.010	0.333	0.683	0.085	0.143	2	4	43	20	0.50	0.62
	nb	0.636	0.702	0.321	0.500	0.805	0.298	0.560	14	14	33	8	0.67	0.59
	rf	0.364	0.851	0.243	0.533	0.741	0.149	0.432	8	7	40	14	0.65	0.61
	reglog	0.227	0.766	-0.007	0.312	0.679	0.234	0.263	5	11	36	17	0.49	0.48
16	rrlda	0.000	1.000	NaN	NaN	0.671	0.000	0.000	0	0	47	23	0.48	0.61
	nb	0.652	0.617	0.253	0.455	0.784	0.383	0.536	15	18	29	8	0.68	0.55
	rf	0.261	0.851	0.135	0.462	0.702	0.149	0.333	6	7	40	17	0.65	0.69
	reglog	0.391	0.723	0.116	0.409	0.708	0.277	0.400	9	13	34	14	0.64	0.47
19	rrlda	0.533	0.391	-0.075	0.364	0.562	0.609	0.432	24	42	27	21	0.47	0.41
	nb	0.489	0.565	0.053	0.423	0.629	0.435	0.454	22	30	39	23	0.54	0.48
	rf	0.244	0.739	-0.018	0.379	0.600	0.261	0.297	11	18	51	34	0.54	0.52
	reglog	0.267	0.754	0.023	0.414	0.612	0.246	0.324	12	17	52	33	0.51	0.32

Table 8: Model evaluation measures on the three chosen datasets. SENS=sensitivity: proportion of true positives, SPEC=specificity: proportion of true negatives, MCC=mathews correlation coefficient: correlation prediction with true labels, PREC=precision: true positive over predicted positive ratio, NPV=negative predictive value: true negative over predicted negative ratio, F1=f1-score: harmonic mean precision and accuracy, TP: true positives, FP: false positives, TN: true negatives, FN: false negatives, AUC: area under the receiver operator curve.

On all three datasets the model with the highest train and test AUC metric was the naive bayes classifier (Table 8). On dataset 14 and 16 the naive bayes model reached a training AUC of 0.67-0.68, which could reflect the fact that these dataset share a large part of donors and features. An AUC value in this range is considered to be a (somewhat) fair discriminator. Although, ideally discriminators would have training and test AUC values in the range 0.7 and up, anything below is considered a weak discriminator (**L˙demann˙2006**). On dataset 19 all models failed to produce fair discriminators, hence we discard this dataset from further analysis.

On dataset 14 and 16 the random forest model had similar performance compared with the naive bayes model, the training AUC score was only slightly lower and the model performed better on unseen test data. This could indicate that the random forest model is overfitting the training data less than the naive bayes model, and would therefore be the preferred choice when choosing a discriminator to be used for new data. Despite this, in this work we consider the naive bayes model the best on dataset 14 and 16. Further, we continue the exploration of vaccine responses and repeat vaccination only using the naive bayes models on dataset 14 and 16. This is motivated by the fact that we are not interested in the best model and the random forest model tends to predict false negatives (sensitivity of 0.364) (Table 8), this last fact is the most problematic since the negative

class is overrepresented in our data.

The final parameters for both naive bayes models were laplace = 0 and usekernel = TRUE and adjust = 1.

7 Exploration of modeling results

Using the models built on dataset 14 and 16 our goal was to identify the features relevant to the generation of antibodies in response to vaccination. The procedure is the same as in the SIMON paper , we calculate the feature importance for the classifier model and rank them based on their contribution to the model from 0 to 100. The top three features with the highest score are explored in more detail. Furthermore, for these features we looked at the values in the second-visit data to explore the effect of repeat vaccination (Figure 12). Lastly, we also calculated the correlation between all features in dataset 14 and 16 to identify feature groups related to the top three most important features (Figure 13) (Figure 14).

7.1 Identifying phosphorylation flow cytometry cell signaling features correlated with vaccine response

Firstly, the top ranked feature in dataset 14 was the phosphorylated STAT transcription factor in unstimulated B-cells (Figure 10, A). However, the difference in the value of this feature between the high and low vaccine responders was not found to be significant (at FDR < 0.01) (Figure 11, B). In contrast, the other two features, IFNg stimulated B-cell phosphorylated STAT and CD4+ T-cell phosphorylated STAT5, were found to be significantly greater in the high responder group (FDR < 0.01).

A correlation analysis of all features showed that the three different STAT protein formed positively correlated clusters (Figure 13) ($p < 0.0001$ after BH adjustment). Further, the most important B-cell STAT5 had negative correlations (pearson's r from -0.2 to -0.5) to STAT1 and STAT3 features ($p < 0.0001$ after BH adjustment). The second most important feature had similar correlations as the first, likely since they are both B-cell STAT features. Lastly, the unstimulated CD4+ T-cell STAT phosphorylation also belonged to the STAT5 positively correlated cluster as the previous B-cell features. These correlations might indicate an interaction pattern between STAT5 and STAT1/3 phosphorylation in different immune cell populations in response to a vaccine.

In dataset 16 there were only four features that had a variable importance score greater than 50 (Figure 11, A). The top two features were phospohorylated STAT1 in unstimulated B-cells and phosphorylated STAT1 in unstimulated CD8+ T-cell. However, only the B-cell feature was found to be significantly greater in the positive class (FDR < 0.01) (Figure 11, B). The B-cell STAT1 feature correlated positively with both unstimulated CD4+ T-cell and CD8+ T-cell STAT1 phosphorylation (pearson's r= 0.7 and 0.4, $p < 0.001$), and there were mild negative correlations with interferon gamma stimulated monocyte STAT3 and STAT5 phosphorylation (pearson's r= 0.3 and 0.2, $p < 0.001$) (Figure 14).

7.2 Repeat vaccination effect on identified features

In the second-visit data of donors in datasets 14 and 16 there were outliers (donors had a value greater than 1000) and nonsensical negative values. These were left out of visualisations, since outliers made the pattern unclear and the negative values were considered as nonsensical values.

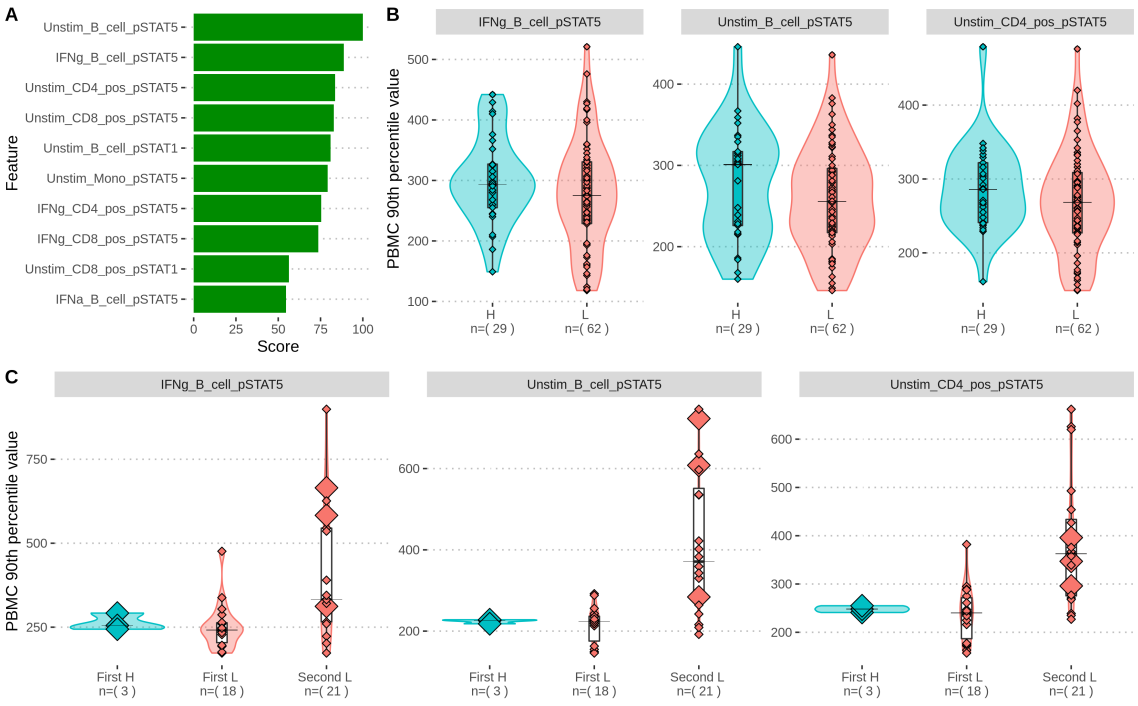

Figure 10: **Exploration of selected features on dataset 14. A.** Features with a variable importance contribution score greater than 50. **B.** Distributions of top 3 most important features grouped by vaccine response classification. Thin horizontal bars show the median value. **C.** Values of the same features as in **B** compared to their value in second-visit data . Donors/rows that changed classification between their first and second visit are indicated as enlarged diamonds.

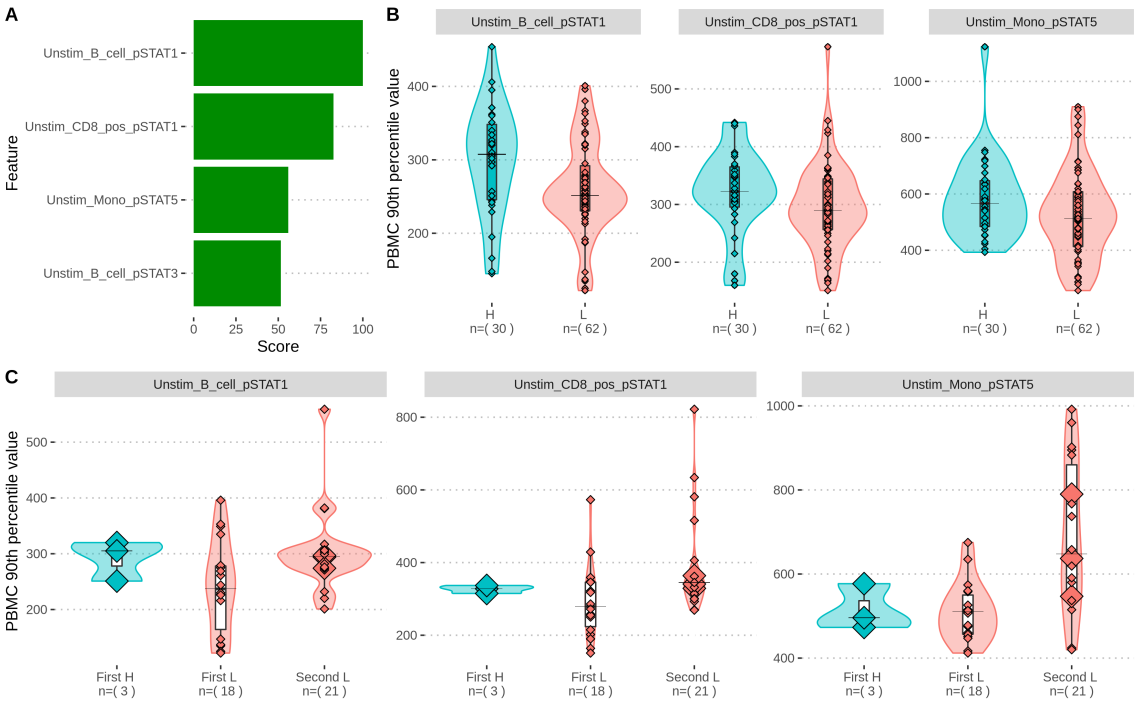

Figure 11: **Exploration of selected features on dataset 16. A.** Features with a variable importance contribution score greater than 50. **B.** Distributions of top 3 most important features grouped by vaccine response classification. Thin horizontal bars show the median value. **C.** Values of the same features as in **B** compared to their value in second-visit data . Donors/rows that changed classification between their first and second visit are indicated as enlarged diamonds.

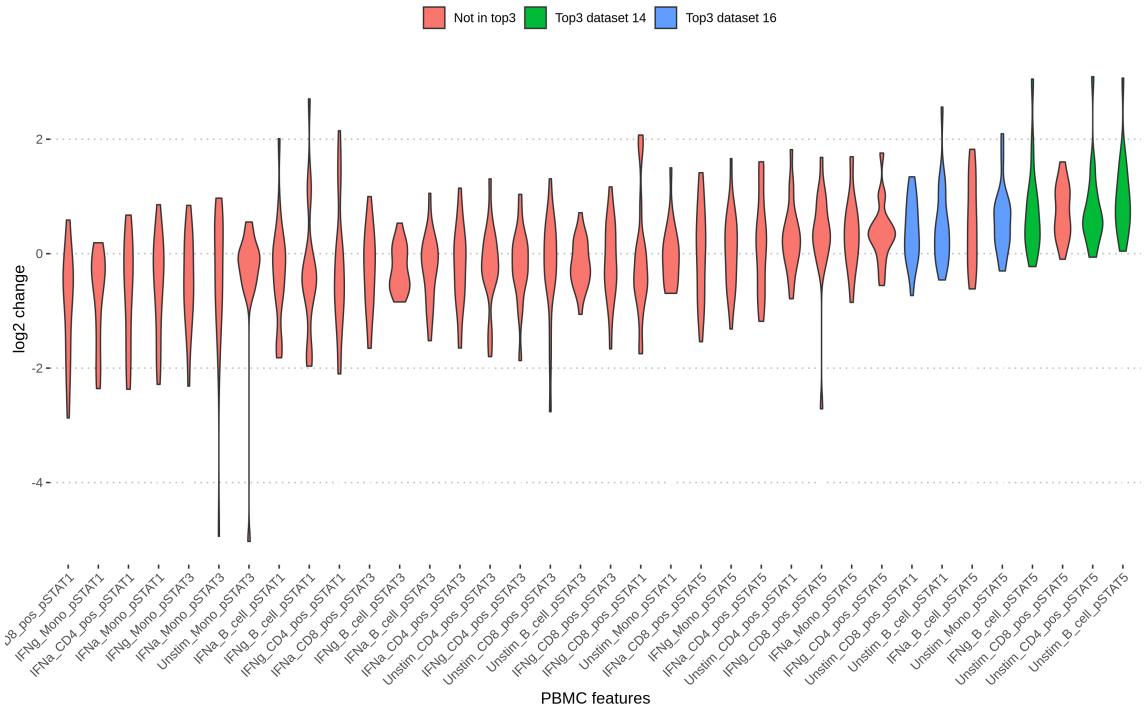

Figure 12: **Log2 change in features of dataset 14 and 16 between subsequent influenza seasons where a vaccine was administered.**

To see how a repeat vaccination affects immune cell signaling, the distribution of the top three features of dataset 14 were compared to their distribution when measured in a subsequent influenza season (Figure 10, C). In the 21 donors that had a second measurement of the features in another influenza season that were not left out (outliers and nonsensical values) there was a consistent pattern that the three high responders were classified as low responders in their second visit (Figure 10, C). Although, overall the values were consistently greater in the second-visit data (Figure 10, C, enlarged diamonds). Thus, vaccination might increase activity in general signaling pathways of PBMC in subsequent influenza seasons, but the classification does not reflect this as increasing influenza antibody response. One possibility is that the donor was classified as low responder due to a lack of response to one specific strain of virus in the vaccine administered in the repeat visit, not necessarily to all strains (Figure 5).

To explore the overall change in the features of dataset 14 between the first and subsequent influenza seasons the distribution of changes for donors were visualised (nonsensical negative readout values were removed) (Figure 12). The overall trend was that the unstimulated PBMCs had higher STAT5 values upon a repeated visit. And, in general STAT5 features increased in value the most in repeat vaccination visits. Furthermore, the values that contributed the most to the model discriminating between high and low responders in the first-visit data also increased the most in a repeat visit. Although, there are outliers that increased a lot in the subsequent influenza season (Figure 12).

On dataset 16 the two top STAT1 features had similar distributions to the first-visit data (Figure 11, C). In contrast, unstimulated monocyte cells had higher STAT5 phoshporylation in the subsequent influenza season (Figure 11, C). Further, the same three donors that were classified as high responders in the first-visit data and as low responders in the second-visit data as in dataset 14 (Figure 10, C) had increased monocyte cell STAT5 phosphorylation (Figure 11, C, enlarged diamonds). Lastly, the top three features of the model trained on dataset 16 also belonged to those that increased the most between the first-visit data and second-visit data (Figure 12).

8 Discussion and conclusion

In this work we gave a brief introduction into influenza vaccination and how vaccine responses are measured, described the FluPrint database, and applied a similar data mining method as in the SIMON paper and additionally explored the available repeat vaccination data. The FluPrint database made it possible to study vaccine responses by providing a classification of donors into high or low responders based on measured antibody level before and after vaccination. Further, it combined and preprocessed data from multiple clinical studies in an accessible database format. This resulted in a wide variety of data on immune cell populations, serum signaling molecules, and cell signaling activity that is suitable for studying immune correlates to vaccine responses using data mining methods. We applied a procedure as described by the authors of the FluPrint database in the SIMON paper , wrapper feature selection using multiple models trained on interesting data subsets of FluPrint . Using this procedure we then explored selected features and how they changed in subsequent influenza seasons. It was found that STAT5 related signaling features correlated with a vaccine response and increased the greatest amount in subsequent influenza seasons. Furthermore, based on correlations between features potential interactions between different immune cell populations could be observed.

Initially, the idea was to focus on building accurate predictors of vaccine response by training models including constructed features based on repeat vaccination. However, during the data

understanding phase of this project it became clear that FluPrint contains only complete classifications in the first-visit data . Instead, the objective was revised to explore the available data on repeat vaccination using models trained on first-visit data data from a selection of clinical studies that received the same vaccine, as done in the SIMON paper .

Firstly, During the data understanding phase it became clear that FluPrint is not suitable for predicting vaccine response with high accuracy, due to data sparsity and small sample size of complete data. Consequently, usage of the FluPrint database requires selecting small datasets without missing values. Further, the available data on repeat vaccinations is limited to one clinical study, and in repeat visits there is often no classification making it impossible to train models using repeat vaccination data.

Secondly, during the data understanding phase we also found that classification is missing in a lot of cases in general. Further, we identified an inconsistency in the classification based on data presented in FluPrint . However, this is likely due to the fact that the before and after antibody titer against individual influenza strains in the vaccine is not completely available in the database and not because the classification is incorrect. Thus, to check the classification quality it is necessary to study the raw data and scripts used to generate the database, which is considered out of the scope of this work.

The data preparation and modeling phases included selecting the data that was most suitable for training models and studying repeat vaccinations. We started with the initial data used in the SIMON paper and also collected repeat vaccination data for the donors in this dataset. To deal with the sparse data the mulset algorithm was applied to generate twenty small but complete datasets, the three datasets that had the highest amount of donors that received a repeat vaccination were then chosen for modeling and further analysis. Four models were built all three datasets, but models with fair discriminative ability were built only on dataset 14 and 16.

The features in dataset 14 and 16 were all from the phosphorylation flow cytometry assay. As a result, the analysis became one dimensional. Using the datasets we models were trained and used to identify features correlated with a high vaccine response. We found that STAT5 phosphorylation in different immune cell populations were associated with a high vaccine response and were increased in subsequent influenza seasons. However, further study of this result is considered out of the scope of this work where the focus lies on the application of data science tools. Instead, we show here that data mining methods described in the SIMON paper can be replicated to answer research questions using complex clinical datasets.

Before beginning this data mining project the following objectives were laid out:

- What kind of studies can be done using the FluPrint database?
- What immunological factors correlate to a vaccine responses?
- What is the effect of repeat vaccination?

In summary, we provided insight into which studies can be done using the FluPrint database by describing the experimental data tables of FluPrint . It became clear that FluPrint is suitable for correlating immunological features with a vaccine response by selecting small complete datasets, but that the possibility of combining large data across years and different studies is limited. Additionally, we found that classifications are not available in a great amount of data points limiting the sample size for classification studies.

We identified that immune cell populations with increased STAT5 phosphorylation activity correlated to vaccine response. Correlation analysis showed that STAT phosphorylation is not

dependent on stimulation or immune cell population as seen by the three main positively correlated clusters. Further, STAT5 features were anti-correlated with the other STAT protein phosphorylation activity.

We identified the features that increased and decreased upon repeat vaccination in a subsequent influenza season, in general STAT1 features decreased the most and STAT5 features increased. Further, this pattern was also seen within the features that contributed the most to models discriminating between high and low vaccine responders.

9 Materials and methods

9.1 Data collection

By following the guide on the FluPrint Github Repository the MySQL server was set up. All file paths mentioned referred to the github repository of this project which can be found below.

In this work the FluPrint github was first added as a submodule. This module provided 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 was 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 2, the password type had to be `mysql_native_password` (how to resolve [SQLSTATEHY000] 2054 the server requested authentication method.).

Listing 2: 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`.

9.2 Statistical methods

9.2.1 Data selection

In this work, immunological features correlating to a vaccine response were identified using wrapper-based feature selection on data from the FluPrint SQL database. Suitable datasets without missing values were generated using the R package mulset, as described in the data preparation section. These datasets were split into training and test splits using the createDataPartition function from the R package caret. As described in the data preparation and selection sections, datasets were not considered if the test set had less than 10 donors. Lastly, from the generated datasets the number of donors in the second-visit data was used to choose datasets for further analysis. The second-visit data was obtained from the database by a query that is avalaible in the github repository of this project.

9.2.2 Model training, evaluation, exploration

Standard procedures were used for model training, models were trained only on the training datasets using 10-fold cross-validation that was repeated two times. The test data was used only as an independent dataset to estimate how much the model overfitted on the training data. Model training itself was done using the caret R package function train. Additionally, parameters were chosen based on the highest cross-validated accuracy automatically train function.

Variable importance of the models generated by the caret package was generated by the function varImp from the same package. This uses model specific feature contribution statistics and ranks them from most important to not important on a scale from 0 to 100, for example for the naive bayes model it uses the class conditional probabilities of features.

Confusion matrix metrics were generated using the MLeval R package which accepts caret objects and computes metrics in a table format as shown in the model evaluation section. Additionally, the test AUC was calculated with another R packages called pROC.

Correlation plots of the features from the selected datasets 14 and 16 were made using the R package corrplot.

9.2.3 Significance tests

To see if features identified by the best classifier trained on datasets 14 and 16 had different distribution in between the two classes the significance analysis of micro arrays (SAM) at a FDR < 0.01 was used in R. P values for all correlations shown in the correlation plots below were calculated using an R package, and only correlations with a p-value less than 0.001 were shown.

9.3 Code and data availability

The code and data belonging to this project can be found in the github repository. The repository contains the directories `bussiness_understand`, `data_understanding` and `data_preparation_modeling` ↪ which contain all the LaTeXsource files for what was written during the project. However, the source files for the final pdf deliverable that is to be graded are in the `deliverable` directory. The directory `csv` contains all the flat data files that were generated in this work, `queries` contains the SQL source files. As mentioned above, the import script for constructing the database is added as a submodule called `fluprint`. Other files and directories are data files used in the latex source files.

Appendices

A Correlation plots

B mulset algorithm

C Query that generates initial SIMON data

Listing 3: Query of initial SIMON data

```
SELECT donors.id                          AS donor_id,
       donor_visits.age                   AS age,
       donor_visits.vaccine_resp          AS outcome,
       experimental_data.name_formatted AS data_name,
       experimental_data.data             AS data
FROM   donors
       LEFT JOIN donor_visits
              ON donors.id = donor_visits.donor_id
                 AND donor_visits.visit_id = 1
       INNER JOIN experimental_data
               ON donor_visits.id = experimental_data.
                  ↪ donor_visits_id
                  AND experimental_data.donor_id = donor_visits.
                     ↪ donor_id
WHERE  donors.gender IS NOT NULL
       AND donor_visits.vaccine_resp IS NOT NULL
       AND donor_visits.vaccine = 4
ORDER  BY donors.study_donor_id DESC
```

D Full description of FluPrint clinical studies

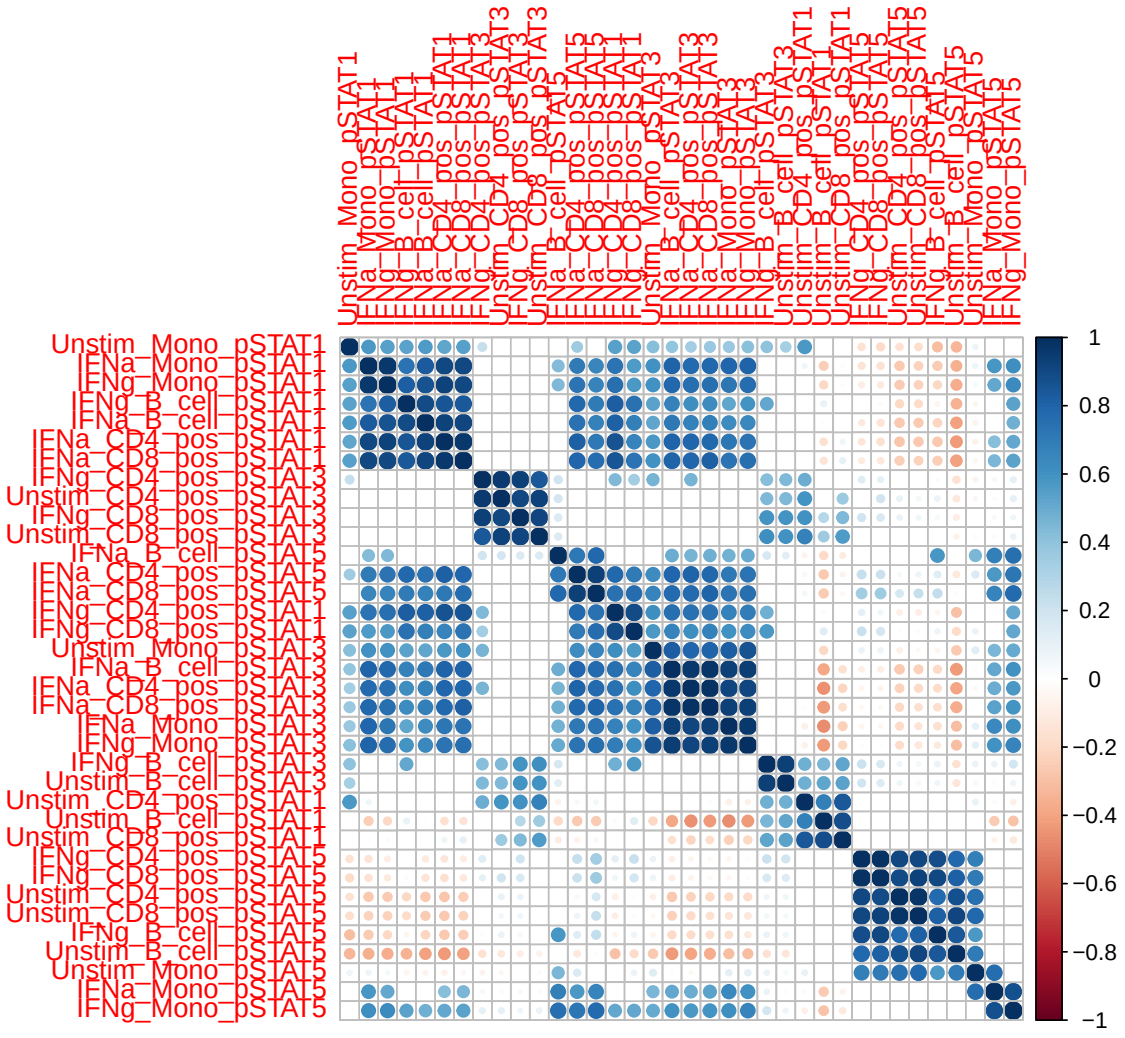

Figure 13: **Correlation heatmap of the features in dataset 14.** Shows the Pearson correlation between the features of dataset 14 (p (0.0001). Insignificant values were not plotted.

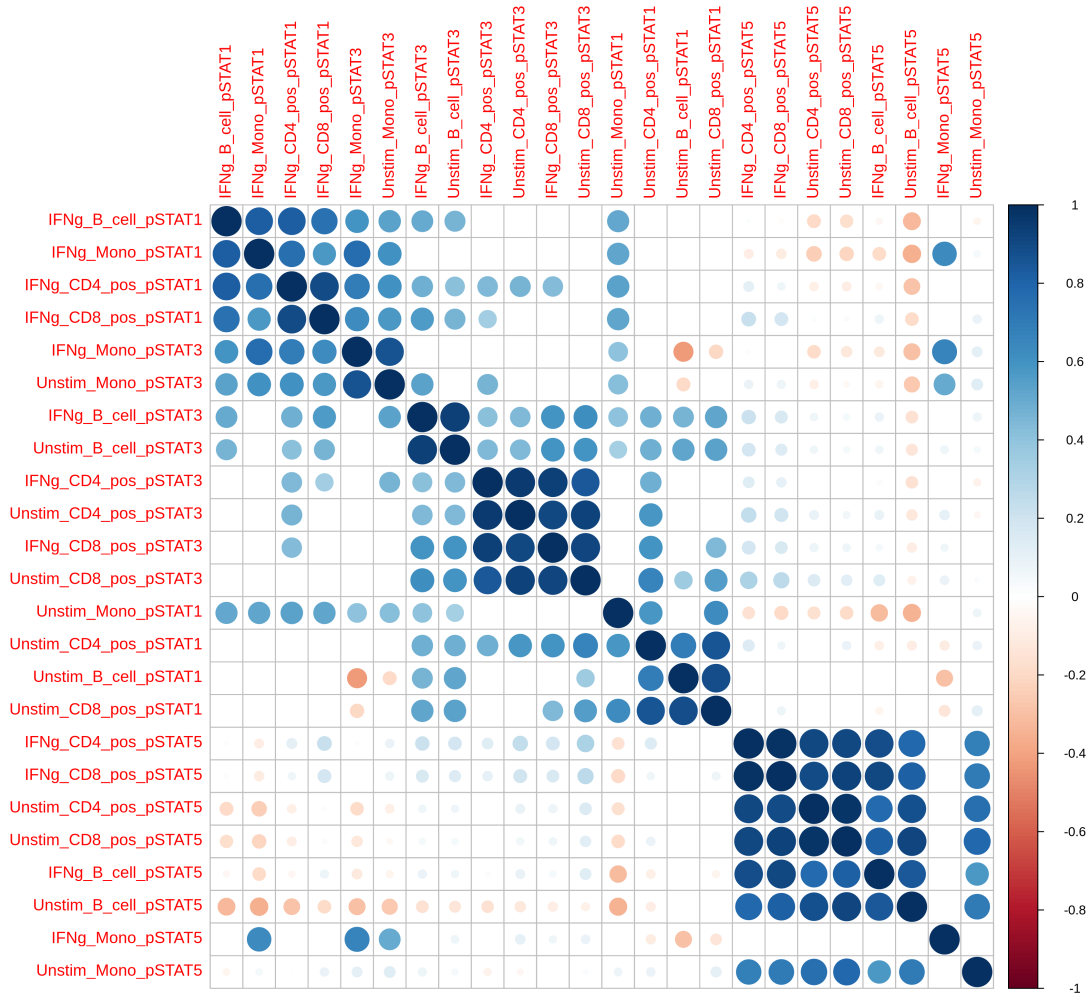

Figure 14: **Correlation heatmap of the features in dataset 16.** Shows the Pearson correlation between the features of dataset 16 (p (0.0001). Insignificant values were not plotted.

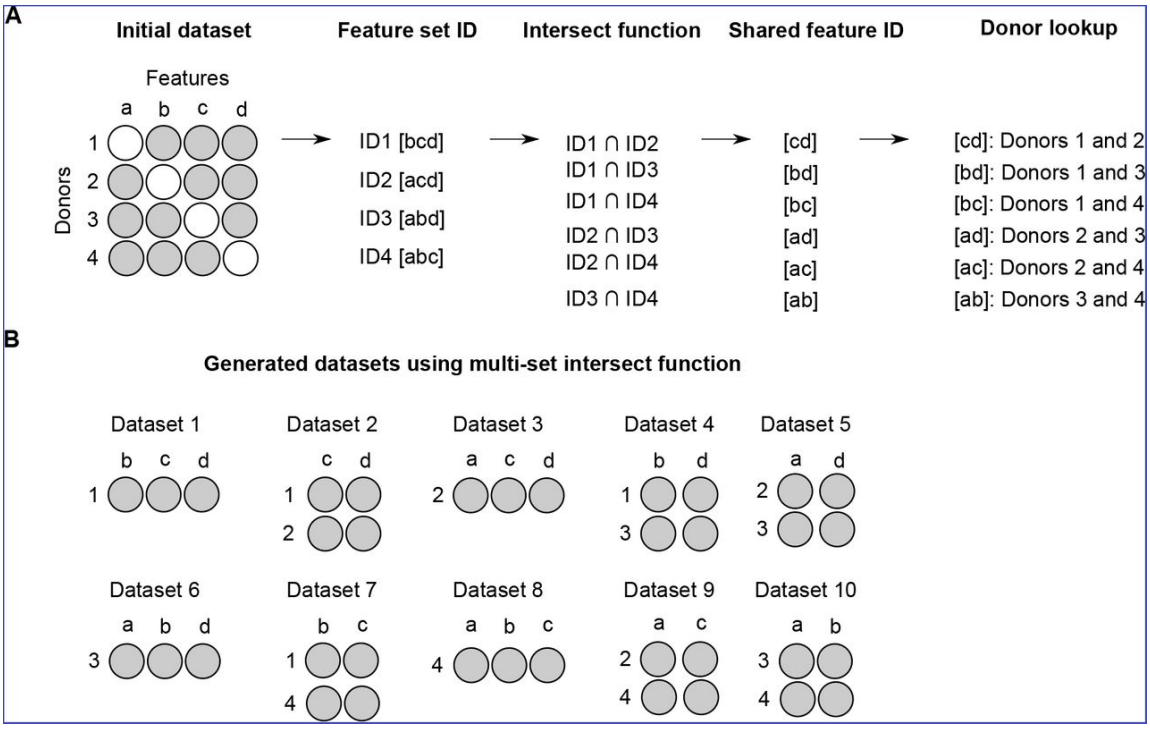

Figure 15: **taken from original work** Graphical explanation of the mulset algorithm. **A.** The intersection function is applied to the powerset of the features. For each intersection a donor look-up is done. If sufficient donors have a value for this feature, then generate a dataset. **B.** A set of small datasets with complete data is generated.

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 9: **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.

E Remaps used in the FluPrint

Vaccine received	Vaccine type ID	Vaccine type name
FluMist IIV4 0.2 mL intranasal spray	1	Flumist
FluMist Intranasal spray	1	Flumist
FluMist Intranasal Spray 2009–2010	1	Flumist
FluMist Intranasal Spray	1	Flumist
Flumist	1	Flumist
Fluzone Intradermal-IIV3	2	Fluzone Intradermal
Fluzone Intradermal	2	Fluzone Intradermal
GSK Fluarix IIV3 single-dose syringe	3	Fluarix
Fluzone 0.5 mL IIV4 SD syringe	4	Fluzone
Fluzone 0.25 mL IIV4 SD syringe	5	Paediatric Fluzone
Fluzone IIV3 multi-dose vial	4	Fluzone
Fluzone single-dose syringe	4	Fluzone
Fluzone multi-dose vial	4	Fluzone
Fluzone single-dose syringe 2009–2010	4	Fluzone
Fluzone high-dose syringe	6	High Dose Fluzone
Fluzone 0.5 mL single-dose syringe	4	Fluzone
Fluzone 0.25 mL single-dose syringe	5	Paediatric Fluzone
Fluzone IIV3 High-Dose SDS	6	High Dose Fluzone
Fluzone IIV4 single-dose syringe	4	Fluzone
Fluzone High-Dose syringe	6	High Dose Fluzone

Table 10: Remaps of vaccine type relevant to to the clinical studies reference table (Table 9), and the section on the donor visits table.

Original	Remapped
CMV EBV	1
Other immunoassay	2
Human Luminex 62–63 plex	3
CyTOF phenotyping	4
HAI	5
Human Luminex 51 plex	6
Phospho-flow cytokine stim (PBMC)	7
pCyTOF (whole blood) pheno	9
pCyTOF (whole blood) phospho	10
CBCD	11
Human MSD 4 plex	12
Lyoplate 1	13
Human MSD 9 plex	14
Human Luminex 50 plex	15
Other Luminex	16

Table 11: Assay to id map used in the FluPrint database. Note that in the actual data 1 is not used, and 17 is used. Refer to (Table 6) for the actual mappings used in the database.