Bussiness Understanding Report

Utrecht University

Mike Vink

April 19, 2021

Contents

Acronyms

HA hemagglutinin. 1, 2

NA neuraminidase. 1, 2

RNA ribonucleic acid. 1, 2

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 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 inactivated trivalent vaccines 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 models

Numerous works apply regression models to describe seasonal population influenza mortality (**zhouHospitalizationsAssociatedInfluenza2012**; **greenMortalityAttributableInfluen** **iulianoEstimatesGlobalSeasonal2018**). Reported are varying age-specific influenza burdens during different seasonal epidemics for different regions, but in general young children an elderly are found to be more susceptible to influenza and are adviced to vaccinated annually (**zhouHospitalizationsAssociatedInfluenza2012**).

Specifically, within the US based work of **zhouHospitalizationsAssociatedInfluenza2012**, the highest hospitalization rates for influenza were among persons aged >=65 years and those aged <1 year. And, age-standardized annual rates per 100000 person-years varied substantially for influenza. A similar pattern is in **greenMortalityAttributableInfluenza** where an age shift in Wales and England seasonal influenza burden was observed following the 2009 swine flue pandemic. These patterns can confound decision making on national and international public health policies. The necessity of informed decision making is apperant from estimates of influenza attributed mortality, it is estimated that globally 291.243–645.832 influenza associated seasonal deaths occur annually (**iulianoEstimatesGlobalSeasonal2018**).

1.2 Vaccine success criteria

Due to the volume and vulnerability of population groups most at risk for influenze, the young and the elderly, a placebo controlled vaccine efficacy study is extremely costly (**zhouHospitalizationsAssociatedInfluenza2012**). Instead the haemagglutination-inhibiting (HAI) antibody test for influenza virus antibody is used to assess vaccine protection (**dejongHaemagglutinationinhibitingAntibodyInfluenza200** The policy for a succesful vaccine is an 4-fold increase in HAI antibody titre after vaccination and a geometric mean HAI titer of ≥ 40. The last is predicted to reduce influenza risk by 50% **dejongHaemagglutinationinhibitingAntibodyInfluenza2003**.

1.3 Finding immunological factors predisposing vaccine HAI antibody response using machine learning

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

Machine learning has been applied to clinical datasets to find influenza protection markers, such as the described T cell populations and titers of related molecules (**furmanApoptosisOtherImmune2013**; **sobolevAdjuvantedInfluenzaH1N1Vaccination201** **tsangGlobalAnalysesHuman2014**). These type of studies suffer from data quality 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 patients/samples. A succesful vaccination is also often not well defined.

1.4 Bussiness objectives

Due to the high volume population that needs vaccines, it is important to study immune correlates to vaccine response. For example, repeat vaccination might not be necessary if the response is low, or a different vaccine is desired on a person to person basis depending on immune correlates. 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 some of the data quality issues of prior studies using clinical datasets comprised of blood and serum sample assays. It does so by incorporating eigth clinical studies conducted between 2007 to 2015 using in total 740 patients, including different types of assays and normalizing their values, and by providing a binary classification of high- and low-responder to a vaccine.

The objectives of this work are to answer:

- 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?

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 insigth in the questions posed in the objectives.

The rationale for these questions and succes criteria are based on the scope of the 3EC project as part of the Applied data science profile and the data available. The paper of **tomicFluPRINTDatasetMultidimensional2019** on which this work is mostly based on provides these questions as interesting directions for further analysis, but does not directly provide the data necessary to answer them, only the MySQL database containing a great volume of data.

2 Assess situation

2.1 data sources

The only source of data used in the project is provided by **tomicFluPRINTDatasetMultidimensional2019**. It is a MySQL database for which the installation is described in the FluPrint Github Repository. A template query is also provided by the authors on the github page belonging to an unpublished work by the same authors SIMON Github Repository. According to the authors, this data is the most interesting for the bussiness objective of finding repeat vaccination effects and will be used in this work too **tomicSIMONAutomatedMachine2019**. 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 1: 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
```

2.2 Tools and techniques

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

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

Since the work this paper is based on uses the R toolset, it is also used here (**tomicFluPRINTDatasetMultidimensional2019**; **tomicSIMONAutomatedMachine2019**). Especially crucial is the R package mulset, which was made by the authors. This package is used to deal with missing data between different clinical studies and years, and thus will be used to generate complete data tables in this paper too. All scripts in this work were composed using tidyverse packages in combination with modelling packages.

2.3 Requirements of the project

Requirements of this work are to show ability in using data science methods. As such, most of the insights will inevitably be a replication of the work done by the authors of the FluPrint database **tomicSIMONAutomatedMachine2019**, but all the scripts and analysis done are original work and are supplied together with the final deliverable.

Since the data type used here is a 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 queried tables from the database will be included in the final deliverable.

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

2.4 Assumptions of the project

This work assumes 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 by the authors **tomicFluPRINTDatasetMultidimensional2019**; **tomicSIMONAutomatedMachine2019** are of high quality, and that this is appropriate for this work. Out of the scope of this work is investigating whether the preprocessing done for the data in the database is valid, since we are not domain experts. A method for querying, cleaning, and generating complete data tables has been provided by the authors and will also be used in this work. It is assumed that the 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 filetypes are only csv files generated by the SQL server.

3 Data mining goals

All bussiness objectives described involve querying data from the FluPrint database. The goal of the authors of the FluPrint database was to provide a unqiue opportunity to study immune correlates of high vaccine responders across different years and clinical studies. The authors also provide a binary classification for donors. In this work we first and foremost explore the database, and lastly we apply feature selection methods and classification models on the most interesting dataset.

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

- Explore and describe SQL queries and corresponding csv tables.
- Model and visualise the different clinical study populations.
- Model and visualise the difference between vaccination types.
- Model and visualise repeat vaccination effects.
- Apply standard feature selection methods to the most interesting dataset.
- Fit classification models to the most interesting dataset.

In data mining terms, the problem type is a combination of exploratory data analysis and classification. Since this work is for a 3EC assignment for the Applied Data Science profile and most of the goals are exploratory analyses, success criteria for all goals are subjective. For exploratory and visual type goals the quality is expected to be of the same level as the publications of the authors **tomicFluPRINTDatasetMultidimensional2019**; **tomicSIMONAutomatedMachine2019**. For the classification type goals we follow the model evaluation procedure used by the authors **tomicSIMONAutomatedMachine2019**, models were evaluated by the AUROC metric, and accuracy, specificity and sensitivity were also reported. Insights produced by this work were benchmarked against the work of the original authors.

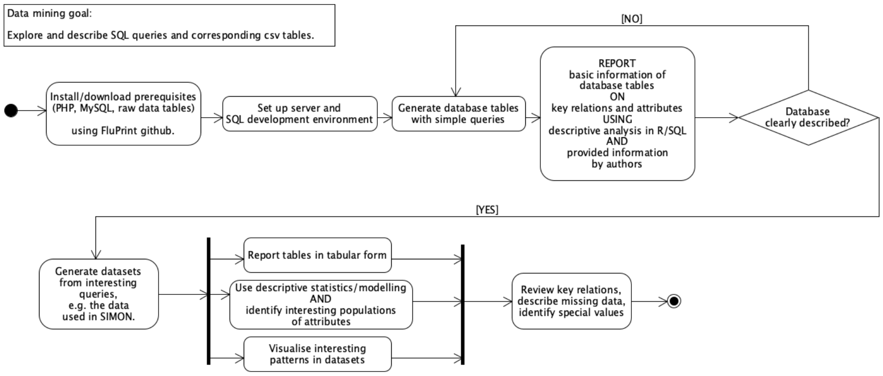
Data mining goal:
Explore and describe SQL queries and corresponding csv tables.
Install/download prerequisites (PHP, MySQL, raw data tables) using FluPrint github.
Set up server and SQL development environment
Generate database tables with simple queries
REPORT basic information of database tables ON key relations and attributes USING descriptive analysis in R/SQL AND provided information by authors
Database clearly described?
[NO]
[YES]
Generate datasets from interesting queries, e.g. the data used in SIMON.
Report tables in tabular form
Use descriptive statistics/modelling AND identify interesting populations of attributes
Visualise interesting patterns in datasets
Review key relations, describe missing data, identify special values

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

4 Project plan

The first part of the project involved querying the database, and collecting and describing the available data (**??**). The first goal is to understand the tables in the SQL database, their key relations, and to describe the attributes within the tables. Valuable info on this part is already provided in the original publication of the database **tomicFluPRINTDatasetMultidimensional2019**, but it was also investigated in this work. The tools that will be used are SQL for querying and R for statistical descriptions.

The second phase of this plan was an iterative process of finding suitable data to answer the modelling and visualisation data mining goals. This is a more involved process since it requires exploration of the database to answer the questions, and therefore was estimated to take time.

Relations between attributes in the generated datasets are visualised and modelled to see if there exist a pattern in the data that is relevant for the business objectives (**??**). A critical point in this plan is deciding whether an objective cannot be answered with the available data. In that case the goal was revised and the second phase of the SQL query plan was reiterated. When deciding if the exploratory analysis was of sufficient quality, the work by the authors of the database used in this work was used as a subjective benchmark **tomicSIMONAutomatedMachine2019**; **tomicFluPRINTDatasetMultidimensional2019**.

For the final two data mining goals the plan was to find the immune correlates of high immune responders using a wrapper based feature selection strategy (**??**)

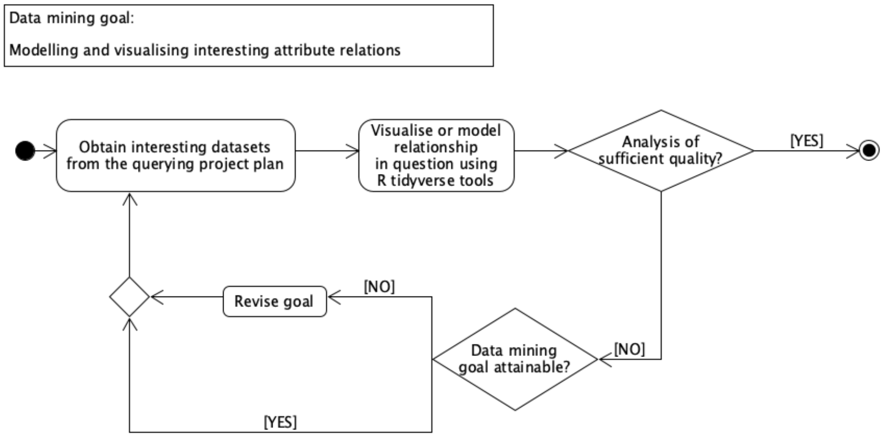
Data mining goal:
Modelling and visualising interesting attribute relations
Obtain interesting datasets from the querying project plan
Visualise or model relationship in question using R tidyverse tools
Analysis of sufficient quality?
[YES]
[NO]
Data mining goal attainable?
[NO]
Revise goal
[YES]

Figure 2: Project plan for the modelling and visualisation data mining goals.

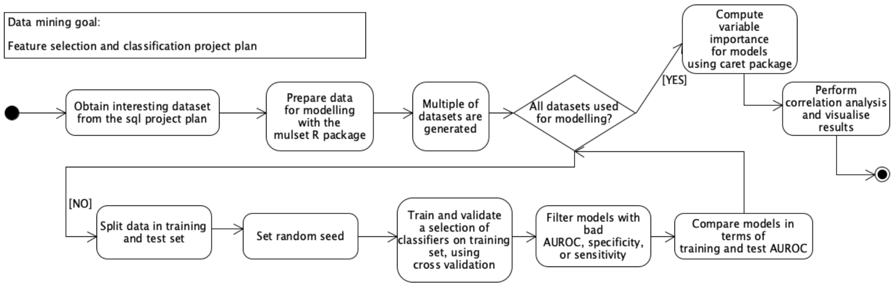
Data mining goal:
Feature selection and classification project plan
Obtain interesting dataset from the sql project plan
Prepare data for modelling with the mulset R package
Multiple of datasets are generated
All datasets used for modelling?
[YES]
Compute variable importance for models using caret package
Perform correlation analysis and visualise results
[NO]
Split data in training and test set
Set random seed
Train and validate a selection of classifiers on training set, using cross validation
Filter models with bad AUROC, specificity, or sensitivity
Compare models in terms of training and test AUROC

Figure 3: Project plan for the classification and feature selection data mining goal.