Data-driven donation strategies: understanding and predicting blood donor deferral

Vinkenoog, M.

CHAPTER

2

Preliminaries

Preliminaries

With one foot in the world of blood transfusion and the other firmly planted in data science and statistics, the research in this thesis is multidisciplinary and requires an understanding of both fields. In this chapter, background knowledge on blood donation and iron metabolism, as well as explanations of some statistical and data science concepts are presented.

2.1 Blood donation in the Netherlands

Since 1998, when Dutch blood banks were merged into one organisation, Sanquin is the only organisation in the Netherlands authorised to collect human blood for transfusion and manage the blood supply. With 50 fixed collection sites and several mobile collection sites, the country is geographically well-covered.

The two main types of blood donation at Sanquin are whole blood and plasma donation. During a whole blood donation, 500 mL of blood is collected from a donor. The blood is collected as-is, hence the term whole blood. During a plasma donation or plasmapheresis procedure, blood is collected from a donor, but everything except the plasma is returned to the donor after a separation process. In 2022, Sanquin collected 313 386 plasma donations and 430 515 whole blood donations.

Iron is present in red blood cells as a constituent of hemoglobin, and so iron loss is much greater during a whole blood donation than a plasma donation. Our research on iron deficiency in donors is therefore focused on whole blood donors. Donors sign up to become either a whole blood donor or a plasma donor, and while sometimes whole blood donors become plasma donors, donors generally do not alternate between the two types of donations. There is a pool of about 300 000 whole blood donors that is continuously changing as people stop donating and new donors are recruited. Women are allowed to donate whole blood three times a year and men five times a year, provided that all eligibility criteria are met. Whole blood donations are collected by invitation only: donors receive an invitation by postal mail, e-mail, or SMS, with which they can visit the blood bank for a donation within two weeks. More recently, donors can plan their donation by making an online appointment.

2.1.1 Iron metabolism

Blood consists of four main components: plasma, erythrocytes (red blood cells), leukocytes (white blood cells), and thrombocytes (platelets). Erythrocytes make up nearly half of the blood's volume and are responsible for delivering oxygen to the whole body. About one third of the volume of an erythrocyte is filled with hemoglobin, a protein that contains iron, which can bind oxygen in the lungs to release it where needed. [6] Anemia is the condition where hemoglobin or erythrocyte levels are low, which can lead to symptoms as described earlier. [7] When the cause of the anemia is a low iron level, this is called iron deficiency anemia, and this will usually be reflected by a low hemoglobin level. [5]

However, iron is also used in other processes than hemoglobin synthesis, and sufficient hemoglobin levels do not guarantee the absence of iron deficiency. If such a state, called non-anemic iron deficiency, is left untreated, it may evolve into iron deficiency anemia, particularly after blood loss. Aside from oxygen transport, iron is required for many cellular functions, such as the replication and repair of DNA, as well as the synthesis of several enzymes and hormones. [8] Almost all cells use iron as a co-factor for biochemical activities, which is why iron deficiency can cause such a wide range of symptoms.

Approximately 70% of all iron in adults is present in hemoglobin, and 25% is stored in ferritin, a large iron storage protein found mostly in the liver. [8] When hemoglobin levels decrease, iron is released from ferritin to be used in erythropoiesis (the production of red blood cells). A small amount of ferritin is also present in blood. Under steady-state conditions, the concentration of ferritin in the blood is correlated with the size of the total iron stores of the body. However, ferritin levels are often elevated under inflammatory conditions and are therefore not a good indicator of iron stores in those circumstances. [9] This can already occur with low-grade inflammation due to obesity or high levels of air pollution. Under normal (non-inflamed) conditions, ferritin is considered a good indicator of the total body iron store. [10]

2.1.2 Iron monitoring strategies at Sanquin

Sanquin has several eligibility criteria regarding hemoglobin and ferritin levels in donors. Before every donation, a drop of capillary blood is collected from the donor's finger and used to test the concentration of hemoglobin. The European eligibility threshold for hemoglobin levels is 7.8 mmol/L for women and 8.4 mmol/L for men; slightly higher than the WHO cut-offs for anemia (120 g/L for women and 130 g/L for

men, which corresponds to 7.5 and 8.0 mmol/L, respectively). [11] The hemoglobin testing policy at Sanquin is the following: hemoglobin is measured in one drop of blood using a HemoCue device. If the hemoglobin level is below the donation threshold, the same test is repeated once or twice, and the value that gets recorded is the highest value. The policy states that if the difference between the measurements is larger than 0.3 mmol/L, a physician should be consulted. Donors whose hemoglobin level does not meet the threshold are deferred and sent home to be invited again three months later.

Pre-donation hemoglobin testing has always been standard practice at Sanquin, but as described in the previous section, hemoglobin level monitoring does not enable the detection of iron deficiency. Figure 2.1 clearly illustrates how hemoglobin and ferritin levels change after donation in 25 male donors. [12] The concentration of hemoglobin decreases for three days as the blood volume is replenished. [13] During this time, ferritin remains stable at the pre-donation level. The iron that is necessary for the synthesis of hemoglobin for the new erythrocytes is released from ferritin; hence, at three days post-donation we see hemoglobin levels starting to increase while ferritin levels start decreasing. After 56 days, the minimum donation interval has passed, and the donor should in theory be eligible for their next donation. At that point, hemoglobin levels are almost back to the pre-donation level. However, ferritin levels are still low, at only 55% of the pre-donation level. [12] If donors were to donate again at this moment, the same pattern would repeat and there would be an increased risk of iron deficiency (first without, then with anemia). More time between subsequent donations is needed for most of these donors to completely recover their iron stores.

To avoid collecting blood from donors with iron deficiency, Sanquin implemented ferritin-guided donation intervals starting October 2017. [14] A stepped-wedge approach was used to introduce the policy in all blood banks, and since November 2019 all blood collection sites apply the ferritin testing policy. Ferritin levels are now tested in all new donors as part of the new donor intake, which assesses general eligibility for someone to become a blood donor. Hemoglobin is also measured during this visit, so baseline values of both iron markers are available for all donors. Only test tubes of blood are collected during this intake, and no donation takes place. If eligible, the donor is invited for a first donation about three weeks after the intake. Apart from the new donor intake, ferritin is measured in repeat donors every fifth donation. According to WHO guidelines, ferritin levels below 15 µg/L indicate iron deficiency. [11] Donors with a ferritin level below 15 µg/L are therefore deferred from donating for one year. If the ferritin level is between 15 and 30 µg/L, donors are considered ‘at

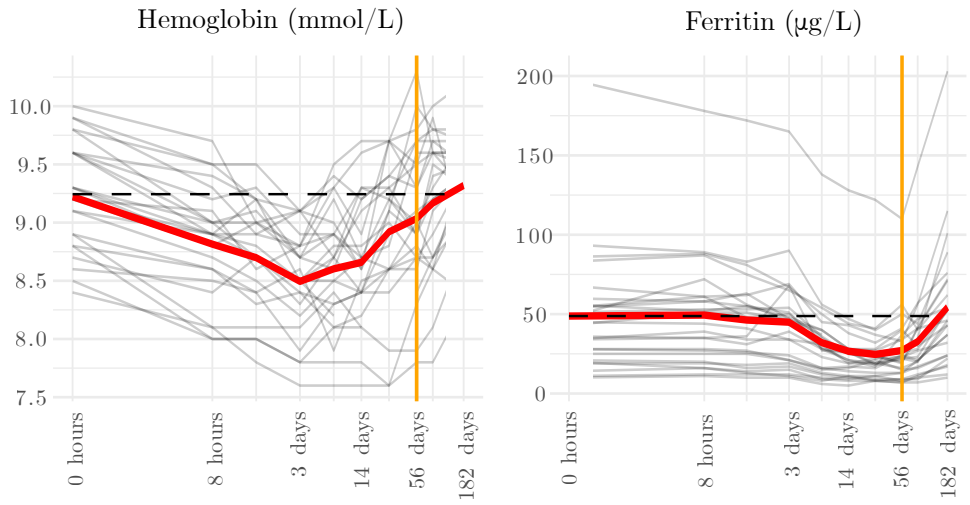

Figure 2.1: Trajectories of hemoglobin and ferritin levels after blood donation for 25 repeat male donors. Solid grey lines are individual trajectories; the red solid line is the average trajectory; the horizontal dashed line indicates the average pre-donation value, and the yellow vertical line shows the minimum donation interval for male donors. Reproduced with permission from Schotten et al., 2016. [12]

risk' of iron deficiency. To prevent the donors from returning to donate with a ferritin level below 15 µg/L, the donor is deferred for six months.

There is an important difference between measuring hemoglobin versus ferritin levels. Hemoglobin testing is done using a point-of-care device and the result is available immediately, so it can be used on-site to decide on the donor's eligibility for donation. However, measuring ferritin levels with such a device is expensive and unreliable, therefore they are measured in the donated blood after the donation has taken place. This means that when a ferritin level is measured, the actual ferritin level of the donor will reduce subsequently as a result of the donation.

After a deferral period due to low ferritin, ferritin is measured again at the next donation. If ferritin is still low after a deferral, the donation frequency is decreased (e.g., from three to two donations a year for women) or the donor can switch donation types (plasma donation might be a better choice) or decide to stop donating altogether.

2.2 Machine learning

Several chapters in this thesis describe models that predict hemoglobin deferral of blood donors. This is a form of binary classification using supervised machine learning algorithms. In those chapters, some statistical concepts are mentioned that are explained in more detail in this section.

2.2.1 Models, explanations and predictions

Machine learning methods can generally be classified as either supervised or unsupervised. In supervised learning, a model learns from examples (training data) where the true outcome is known to the model. After learning from the training data, the resulting model can be used to predict the outcome on unseen test data, for which the model does not know the true outcome. We use supervised learning methods to predict hemoglobin deferral (Chapters 7 through 9), as well as to describe COVID-19 antibody levels (Chapter 6).

In unsupervised learning, there is no outcome to predict. Instead, these models are used to find, for example, groups of either observations (clustering tasks) or variables (dimensionality reduction). We apply unsupervised learning methods to cluster hemoglobin trajectories (Chapter 5), and to group predictor variables into latent constructs for the structural equation model (Chapter 4).

An important distinction in statistics is explanatory versus predictive modelling, both forms of supervised learning. Both serve important roles in scientific research, but a good understanding of the difference is necessary for correct usage and interpretation. Explanatory modelling is concerned with finding and testing causal theories, using data mostly as a tool. The aim here is to find an interpretable statistical model that confirms or rejects the underlying theories. Modelling becomes predictive modelling when the aim is to find the best model for predicting the outcome for unseen observations. The priority here is to generate accurate predictions, rather than being able to understand underlying associations.

Generally, predictive modela use data-driven methods, whereas explanatory models are more hypothesis-driven. There is also a difference in optimisation goal: explanatory models focus on minimizing bias (errors resulting from erroneous assumptions in the model), while minimizing estimation variance (fluctuations resulting from the specific data used) is less important since the models are used for population-level inference. Prediction models are intended to be used for individual-level predictions, and therefore the aim is to minimise estimation variance in addition to bias. In practice,

the biggest difference between explanatory and predictive modelling is the handling of data. In explanatory models, the full dataset can be used to fit the model and calculate coefficients. Validation of the model is done through goodness-of-fit tests and residual analysis to assess how well the model captures the data. In prediction models, however, the models are fit only on the training data, and a separate test dataset is used to validate the accuracy by comparing the predicted to the actual outcomes. It is important to keep the distinction between explanatory and prediction models in mind to avoid conflating the two and drawing incorrect or unsupported conclusions. [14]

In this thesis, both explanatory and prediction models are used. Chapters 3 and 4 focus on explanatory models without any predictive aspect; they relate (changes in) ferritin levels to explanatory variables that are based on biological theory. We do not attempt to minimise estimation variance, as we are interested in the average effect of the variables on ferritin levels in the whole population. The same holds for the analysis of COVID-19 antibodies in Chapter 6. Predicting any individual donor's specific ferritin or antibody level would likely not be very accurate, but we are nonetheless able to explain a considerable amount of population variance with our models. Chapter 5 is neither explanatory nor predictive, as it uses unsupervised learning methods to cluster hemoglobin trajectories and there is no outcome to explain or predict.

Chapters 7 through 9 focus on predictive modelling, which is also expressed by the term 'prediction model' in the titles. In these papers, models are trained using only the training data, and model performance is assessed using the prediction accuracy on a separate set of test data. Since predictions would be used for forecasting, the test set always contains more recent data than the training set. Although these papers focus on prediction, they are not void of explanatory aspects. Especially in Chapter 7, the explanatory aspects are just as important as the predictive aspects. The main difference with 'true' explanatory models is that Chapter 7 focuses on explaining predictions, rather than explaining causal relations with the outcome variable. Where possible, prediction explanations are related to known or hypothesised causal explanations, but this is not required to keep a predictor variable in the model.

2.2.2 Classification for donor deferral

In machine learning, classification refers to the use of a prediction model where class labels are predicted for input data. In binary classification, there are exactly two outcome classes: one is the positive outcome class and the other the negative outcome class, where the positive outcome class is typically the outcome that is of most interest to the researcher. In our context, the two outcome classes for hemoglobin deferral prediction are 'deferral' and 'non-deferral', with deferral being the positive outcome class (even though deferral is not positive in a colloquial sense).

A hemoglobin deferral prediction model could be used by blood banks in the process of inviting donors. In addition to inviting donors based on their eligibility (i.e., has the minimum donation interval been met) and current demand for different blood groups, the prediction of the model could be taken into account. By only inviting donors that are predicted to belong to class non-deferral by an accurate model, the number of donors that are deferred on-site could be reduced.

For each observation, predicting the outcome leads to one of four scenarios:

1. Prediction non-deferral, true outcome non-deferral (true negative, TN)
2. Prediction non-deferral, true outcome deferral (false negative, FN)
3. Prediction deferral, true outcome deferral (true positive, TP)
4. Prediction deferral, true outcome non-deferral (false positive, FP)

If donor invitations are guided by a prediction model, the two types of incorrect predictions (false negatives and false positives) have different effects. For false negatives, the donor is invited to donate and then deferred before donation because the hemoglobin level is below the threshold. For false positives, the donor is not invited even though their hemoglobin level would have been sufficient: a missed donation for the blood bank. A balance must be found between the false positives and the false negatives: we want to reduce the deferral rate, which is determined by the proportion of false positives, but if we predict 'deferral' too liberally, we will find many more false negatives, which means more missed donations, and a risk for the blood bank to not have sufficient blood donations.

Generally, prediction models work best when all outcome classes have the same probability of occurring. In this case, outcome class non-deferral is much more likely than outcome class deferral. Deferral rates in the Netherlands are currently at around 3% for women and 1% for men, meaning that 97% and 99% of donation attempts belong to the outcome class non-deferral, respectively. This has important consequences

for the assessment of model performance. Often, performance is measured as accuracy on a test set (accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$). Another popular choice for model performance is the ROC curve (receiver operating characteristic curve), which plots the true positive rate ($\frac{TP}{TP+FN}$) against the false positive rate ($\frac{FP}{FP+TN}$) at different classification thresholds. A classification threshold is needed to map the output of the model (the deferral probability of an observation) to a binary outcome: if the probability exceeds the threshold, the prediction will be 'deferral'. As the classification threshold decreases (i.e., a lower probability of 'deferral' is needed to classify an observation as such), more observations will be predicted as 'deferral', which increases both the false positive rate and the true positive rate. ROC curves can be summarised in one number by the ROC AUC (Area Under the ROC Curve), which is a number between zero and one, where an ROC AUC of 1 corresponds to a model with 100% correct predictions. The baseline value for the ROC AUC is 0.5: the performance of a random classifier.

However, regular accuracy and the ROC AUC may be misleading when used for datasets with imbalanced outcomes. [15] An extreme illustration of this would be a model that always predicts 'non-deferral'; although its accuracy would be 97% for women and 99% for men, such a model would not have any practical value. Similarly, the ROC curve may be too optimistic in severely imbalanced classification problems. Therefore, throughout this thesis, we use precision and recall instead of accuracy, and AUPR (Area Under the Precision-Recall curve) instead of ROC AUC. Precision is defined as the proportion of correctly predicted observations out of all observations predicted to belong to that class ($\frac{TP}{TP+FP}$ for the positive class, or $\frac{TN}{TN+FN}$ for the negative class). The precision of class deferral is the proportion of true deferrals out of all predicted deferrals. Recall is defined as the proportion of correctly predicted observations in one outcome class ($\frac{TP}{TP+FN}$ for the positive class, $\frac{TN}{TN+FP}$ for the negative class). The recall of class deferral is the proportion of deferrals that are correctly predicted to be deferrals.

There always exists a trade-off between precision and recall: by increasing the classification threshold, the number of false positives decreases, whilst the number of false negatives increases. This causes precision to increase, and recall to decrease. The graph showing precision and recall at different classification thresholds is called the precision-recall curve, and the area underneath this curve is the AUPR. Similar to the ROC AUC, AUPR is a number between zero and one, but its baseline and interpretation are different. For the AUPR, the baseline value is dependent on the proportion of observations belonging to that outcome class. In our case, the baseline AUPR for class deferral would be 0.01 for men and 0.03 for women. The AUPR should

therefore always be interpreted in combination with the baseline value.

An interesting feature is that true negatives are used for the calculation of neither precision nor recall. This is what makes it suitable for imbalanced data: we ignore the large number of non-deferrals that are predicted correctly and focus on the correctness of the prediction of deferrals, and on incorrectly predicted non-deferrals.

The aforementioned metrics provide a fair evaluation of model performance on our imbalanced dataset. All studies involving prediction models within this thesis report these metrics, or a subset thereof, depending on their relevance to the specific research question. Whenever possible, we translate these metrics into statements that clearly show their practical implications. For instance, we calculate the hypothetical deferral rate by taking the complement of the recall of class non-deferral, providing insight into the potential impact of using this model to guide donor invitations. By reporting these metrics and their interpretations, we aim to accurately describe model performance, as well as allow for straightforward interpretation of the effect of using the model for donor management.