Sanquin

UNIVERSITY OF TWENTE.

Red blood cell transfusion, patient outcomes and anaemia in elective orthopaedic surgery through the lens of supervised learning and mediation analysis

Master Thesis Report

Industrial Engineering and Management (study programme)
Healthcare Technology and Management (specialization)

September 2023

Author:
Lenka Pacnerová

Supervisors:
dr. Cynthia So-Osman
dr. Judith Rijnhart
dr. Julia Mikhal
dr. Engin Topan

University of Twente
Enschede, the Netherlands

Author

L. Pacnerová (Lenka)
S2588293
Industrial Engineering and Management (study programme)
Healthcare Technology and Management (specialization)
Faculty of Behavioural, Management and Social Sciences
University of Twente, Enschede, the Netherlands

Graduation organization

Sanquin Blood Bank, Unit Transfusion Medicine
Plesmanlaan 125, 1066 CX Amsterdam, the Netherlands

University supervisors

Dr. J. Mikhal (Julia)
Faculty of Behavioural, Management and Social Sciences
Department of Technology, Policy and Society
Health Technology and Services Research, TechMed Centre

Dr. E. Topan (Engin)
Faculty of Behavioural, Management and Social Sciences
Department of High-tech Business and Entrepreneurship
Industrial Engineering and Business Information Systems

External supervisors

Dr. C. So-Osman (Cynthia)
Clinical Consultant, Haematology and Transfusion Medicine
Unit Transfusion Medicine
Sanquin Blood Bank

Dr. J.J.M. Rijnhart (Judith)
Assistant Professor of Epidemiology
College of Public Health
University of South Florida

To the patients.

This page is intentionally left blank.

Acknowledgement

First of all, great thanks belong to my young family. Their unconditional love and care made it possible to come to this point – a final version of this Master thesis report and graduation in the growing field of data science and capacity management in healthcare. My gratitude rests on carrying this commitment together. I also thank my family abroad and friends who accompanied me on this journey. Following the exciting Master's programme of Industrial Engineering and Management in the Dutch higher education system helped me fuse into a very-much pragmatic-minded culture on the European soil. I thank all academic project partners and colleagues whom I met and worked with along the way. The learning experience with them contributed to the venture of stepping into transfusion medicine and blood banking already now – the area of my interest for some time.

I very much enjoyed relationship-building on this project. Thank you, Cynthia, Julia, Judith and Engin, for having demonstrated your ongoing support. I thank you again for your patience and curiosity coupled with determination to cross the finish line with me. I feel blessed for the freedom I had while leading this data-driven project and I would love to be a part of such a determined team again some time in the future. These things are close to my heart from sports and I rejoice and celebrate that we found this kind of teamwork in this project. Thank you, Cynthia, for making me feel welcome at Sanquin, particularly, I am enriched by recent, stimulating networking opportunities as I conversed face-to-face with many subject matter experts from Sanquin or from abroad.

I cherish the learning curve tied to the adversity and the intense workload we encountered along the way. This project repeatedly offered room to explore further modelling possibilities. Yet I learned to say "stop" by recognizing the need for model simplification and making assumptions. We were simply constrained by the resources available for this period of time.

Through regular meetings with Cynthia, I was repeatedly reminded about planting a seed. I hope the future of transfusion medicine and patient blood management will yield new funding opportunities to water that seed, transform new project incentives into new insights or scale-up initiatives – for local communities and many nations.

This page is intentionally left blank.

Abstract

Introduction: An urge in the patient blood management and transfusion medicine landscapes to study patient outcomes is prevalent in pursuit of relieving transfusion dependency and enhancing patient-centredness. Clinical researchers prefer simple models and ease in their interpretability. The incentive by the Sanquin organization is to conduct mediation analysis – novel in this field of medicine. Literature offers studies with blood transfusion treated in terms of the product use, not yet through the lens of patient outcomes. A raw, patient-level dataset from the 'TOMaat' study (a double-randomized multi-centre control trial) from the elective orthopaedic surgery featuring 533 variables was available for this research.

Methodology: The research examines the role of red blood cell (RBC) transfusion up to Day 14 relative to post-operative complications up to Day 14. Pre-operative anaemia is the exposure component in the mediation model. A blend of prediction and inference tools were utilized in supervised machine learning model development and mediation analysis on a sample size of 2426 patients. Partial dependence plots, odds ratios and coefficients yielded effect estimates from non-parametric (random forest (RF)) and parametric models (logistic regression (LREG) and lasso). The raw dataset was subject to thorough variable selection to reduce the number of input variables from 533 to 41 and 32. Respectively, this applies to the models with post-operative complication up to Day 14 (Case COM) and RBC transfusion up to Day 14 (Case RBC) being the target dependent variables. Lasso led to a further reduction of input variables to 11-12 and 8 (COM and RBC, respectively). Due to excessive data missingness (34% in COM) and a free text field format (RBC) of event dates, massive data cleaning efforts led to establishing pessimistic and optimistic scenarios (COM_{PES}, COM_{OPT}) to sequence the RBC and COM events in time.

Results: All 12 supervised learning models display moderate performance in terms of the AUC (0.63-0.71) with no significant difference between the RF, tuned RF, LREG and lasso models (built per each Case RBC, COM_{PES}, and COM_{OPT}). Strong confounding variables were consolidated from the inference insights and thoroughly validated with the clinical expert leading to 10 strong confounders for the mediation model. RBC transfusion is a statistically significant predictor for COM_{PES} based on LREG and lasso; however, opposing results are found for COM_{OPT}. Different results may be observed when examining pre-operative anaemia in silo using descriptive statistics ($p<0.001$ for COM_{PES} and COM_{OPT}, chi-squared test) versus in the presence of other covariates resulting in low variable importance based on the supervised learning models. Extreme implications due to data missingness were visible in mediation analysis since opposing findings were observed for these two scenarios. RBC transfusion mediates the relationship between pre-operative anaemia and post-operative complications in the pessimistic scenario (ACME of 0.0445, 95% CI of [0.0268; 0.0700]) whereas there is no significant mediation in the optimistic scenario.

Discussion: RBC transfusion was not previously studied as a mediator between pre-operative anaemia and patient outcomes in elective orthopaedic surgery. Three key takeaway messages are proposed based on this research:

[1] The opposing results of mediation analysis lead to a clear prompt for improving the mediation analysis model to resolve current bias, or digital maturity in hospitals. It is advised to bring the attention towards developing robust data acquisition strategies upstream in the data workflow processes to mitigate risks tied to missingness or unstructured data. This advancement can then offer a stronger platform for analyzing patient outcomes downstream. Otherwise, inference among key variables is hardly deduced, and decision-making tied to transfusion dependency and patient-centredness are very limited.

[2] Transfusion may be used as a dependent variable for modelling. Nevertheless, careful consideration must be given to the choice of input variables that shall not encompass its triggers, such as blood loss or the haemoglobin level. Documenting transfusion triggers and patient consent in databases is a pre-requisite to avoid modelling flaws.

[3] The high complexity of transfusion medicine may lead to the need of developing models more complex than parametric models. Other advanced analytic ('black box') methods may also offer inference insights.

Conclusion: The study offers a roadmap for treating (RBC) transfusion and patient outcomes in supervised learning modelling and mediation analysis. Strong confounding components were extracted from the inference outputs and validated using clinical insight. RBC transfusion mediates the relationship between pre-operative anaemia and post-operative complications up to Day 14 in the pessimistic scenario, yet not in the optimistic scenario. The mediation analysis approach shall be improved to deal with prevailing bias. Next to it, the opposing results due to extensive missingness of the post-operative complication dates prompt for new project incentives and advancements in data acquisition strategies in the PBM landscape.

Keywords: Red blood cell transfusion, post-operative complications, pre-operative anaemia, elective orthopaedic surgery, mediation analysis, supervised learning, patient blood management.

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Term
COM	post-operative complication up to Day 14 (dependent variable)
Hb	haemoglobin
LREG	logistic regression
PBM	Patient Blood Management
PDP	partial dependence plot
RBC	red blood cells, or red blood cell transfusion up to Day 14 (dependent variable)
RF	random forest
SME	subject matter expert

Established Terminology

At the cross-section of data science and epidemiology, the same terms may represent different concepts. Hence, for clarity and consistency across the text, the following terminology is established. The use of synonyms is restricted.

Table 1: Established Terminology.

Term	Explanation	Synonyms
data field	raw data column in a dataset before data exploration some of which can be empty or with a single unique value	column, data column, field, data field, variable
variable	data column in a dataset after data exploration and after data preparation that is considered (and can be eventually included) for modelling or intermediate steps in data preparation; variables can be input or output variables, or intermediate variables necessary to establish independent or dependent variables	
independent variable, input variable	variable that serves as an input variable into a model and that has at least two unique values	predictor, feature, variable
dependent variable	the resulting response (outcome) variable with which independent variables can be associated to various extent; in this work, there are two dependent variables for supervised machine learning model development (RBC transfusion, and post-operative complications)	outcome variable, response, response variable, output variable
variable importance, statistical importance (of a variable)	an association between an independent and a dependent variable that can be measured as statistical significance or effect (given a specific level of statistical significance, α)	association, strong association
statistically significant predictor (variable)	an independent variable of statistical significance of a specific level of statistical significance, α, with respect to the dependent variable	predictor, strongly associated predictor (variable)
effect	variable importance measure, i.e. measured using odds ratios (given a specific level of statistical significance, α)	-
exposure	terms denoting additional model components (variables) relevant for the mediation analysis method	-
mediator		
confounder		

Chapter 1 | Introduction

1.1. The Urgent Need to Tackle the Burden on Healthcare due to Anaemia and Transfusion Dependency

Anaemia and transfusion dependency impose a burden on healthcare systems (WHO, 2021). Alarmingly, statistics indicate that roughly every fourth person on earth has anaemia (Safiri et al., 2021) which is equivalent to more than two billion people. Anaemia is a blood disorder defined as a low level of haemoglobin indicating lower capacity of the blood to transport oxygen to the body's tissues (WHO, 1968). Anaemia is a "serious global public health problem" because it causes negative impacts on people's health, such as fatigue, weakness, dizziness, or shortness of breath, to name a few (WHO, 2023). Thus, anaemia may worsen people's quality of life (WHO, 2023). For anaemic patients in medical or surgical settings, transfusion is typically a common treatment option (WHO, 2021). Hence, in the context of anaemia, healthcare systems carry a load of dependency on transfusions (WHO, 2021) coupled with a resulting burden in terms of logistics, costs, or questionable safety and efficacy. Still, there is a clinical need to bring further clarity on the role of red blood cell (RBC) transfusions in medical or surgical settings given that a patient receives foreign (=allogeneic) blood from a donor. As an example, we may ask this question to trigger patient-centred initiatives in a surgical setting:

Is RBC transfusion harmful in terms of the increased risk for poor post-operative patient outcomes in a specific patient group?

Blood health is a term used more and more commonly to emphasize the approach of treasuring a patient's own blood as a liquid organ. Next to it, Patient Blood Management (PBM) is a concept in medicine that places importance on cherishing a patient's own blood (WHO, 2021; Shander et al., 2022). In this thesis, we adhere to the most recent global definition of PBM, endorsed by 25+ medical societies, from the publication by Shander et al. (2022):

> "Patient blood management is a patient-centered, systematic, evidence-based approach to improve patient outcomes by managing and preserving a patient's own blood, while promoting patient safety and empowerment."

PBM has a tremendous potential for relieving the burden on healthcare caused by anaemia and transfusion dependency, yet there are unmet needs concerning PBM implementation. On one hand, various healthcare establishments across the globe have already implemented and evaluated PBM programmes and strategies to tackle transfusion dependency tied to anaemia (WHO, 2021; So-Osman, 2017). On the other hand, the unmet needs were addressed recently in the World Health Organization's policy brief report titled ***The Urgent Need to Implement Patient Blood Management*** (WHO, 2021). Executing a data-driven project in conjunction with this thesis serves to respond to this urgent need.

Since 1998, in accordance with the Blood Supply Act ('Wet inzake Bloedvoorziening'), Sanquin has been the only organization in the Netherlands authorized to collect, process and distribute blood from donors (Sanquin, 2023a, Sanquin, 2023b). As stated in its 2021 Annual Report, anaemia is currently the #1 medical priority for Sanquin Research (Sanquin Blood Supply Foundation, 2021). Sanquin's staff specialized in PBM and epidemiology in transfusion medicine identifies a clinical need for practising evidence-based medicine to a greater extent across the PBM landscape. The motive is to determine the role of RBC transfusion through executing data-driven projects – by drawing insights from patient medical records in medical or surgical settings. A patient-level dataset shall contain the patient history, transfusion data and patient outcomes. Also, a key criterion is robust data acquisition. For this project, Sanquin provides a patient-level dataset that was collected for the purpose of a double-randomized controlled trial in the elective orthopaedic surgery setting in the Netherlands.

In pursuit of tackling the burden due to anaemia and transfusion dependency, the research serves to investigate and implement selected, suitable methodologies in line with evidence-based medicine – to bring greater granularity to scientific evidence. In this work, the term granularity is equivalent to distinguishing the level of detail when insights are drawn from available data. Eventually, the use of the actual evidence obtained during this research project is left to the careful consideration of healthcare experts, especially, due to the age of the dataset. The data collection occurred during 2004-2009 (So-Osman et al., 2014a; So-Osman et al., 2014b).

1.2. The Four Hypotheses, the Confounding Phenomenon and the Research Gap

Hypothesis 1: The Transfusion Medicine Unit at Sanquin Blood Bank brings forth a hypothesis that RBC transfusion mediates the relationship between pre-operative anaemia and poor post-operative patient outcomes. Figure 1-1 below illustrates the hypothesis in a diagram. RBC transfusion follows pre-operative anaemia (the exposure component) because anaemia is often a trigger (predictor) for RBC transfusion.

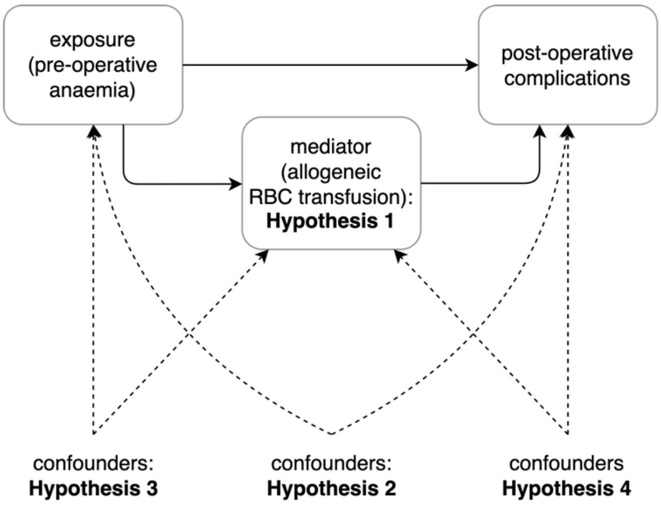

exposure (pre-operative anaemia)
post-operative complications
mediator (allogeneic RBC transfusion): **Hypothesis 1**
confounders: **Hypothesis 3**
confounders: **Hypothesis 2**
confounders **Hypothesis 4**

Figure 1-1: The four hypotheses and the detailed version of the mediation analysis model (allogeneic = a transfusion recipient accepts foreign blood from a donor; RBC = red blood cell).

A method commonly used in health and social sciences to test for mediation is mediation analysis whose implementation Sanquin strongly prefers for this project.

In epidemiology, the phenomenon when a variable significantly affects *both* model components (such as the exposure as well as the patient outcomes) is known as **confounding** (Hernán & Robins, 2020).

On behalf of Sanquin, dr. Cynthia So-Osman is a transfusion specialist dedicated to improving strategies in PBM. She is a contributor to the 2021 WHO report and claims that studying mediation and the confounding phenomenon in transfusion medicine and PBM is currently a research gap. Dr. Judith Rijnhart is an SME in mediation analysis who currently resides with the College of Public Health at the University of South Florida. She advises on the complexity of the problematics through introducing a total of three pairs of confounders. Subsequently, the model in Figure 1-1 is accompanied by three more hypotheses.

Hypothesis 2: There are confounding variables that affect *both*:

- Pre-operative anaemia (the exposure), and
- Post-operative complications (the patient outcomes).

Hypothesis 3: There are confounding variables that affect *both*:

- Pre-operative anaemia (the exposure), and
- Allogeneic RBC transfusion (the mediator).

Hypothesis 4: There are confounding variables that affect *both*:

- Allogeneic RBC transfusion (the mediator), and
- Post-operative complications (the patient outcomes).

Clinical researchers including transfusion professionals desire simple models and ease in their interpretability despite high complexity of transfusion medicine. Therefore, it is wise to start with a simple mediation model if no previous work is available with the dataset at hand particularly involving the desired dependent variables. Next, it is essential to note two considerations to frame the project scope:

[1] From among all available confounders, **strong confounders** (defined further below in this section) will be selected only.

[2] Pre-operative anaemia is not approached as a dependent variable in the model because the data collection of the dataset at hand (further elaborated in Section 1.6.) was not designed to be sufficiently vigorous for this purpose. In the project, this circumstance leads to tackling **Hypotheses 1 and 4**. Hence, **Hypotheses 2 and 3** are out of scope.

In regard to the abundance of variables in the raw dataset (Section 1.6. and Section 3.1.1.), the execution of this data-driven research project revolves around a combination of two applied statistical (learning) methods.

In pursuit of testing Hypothesis 4: The plan is to build interpretable, supervised machine learning models of such choice that insights pertaining to *variable importance* can be retrieved. (i.e. Is a variable a statistically significant predictor?) **Strong confounders** will be identified as the intersection of the strong predictors for the two dependent variables:

- Allogeneic RBC transfusion (the mediator), and
- Post-operative complications (the patient outcomes)

A round of validation through integrating the clinical insight will be required for the final selection of strong confounders.

In pursuit of testing Hypothesis 1: Mediation analysis will be conducted as the project capstone. Per Figure 1-1 shown earlier, the mediator under consideration is allogeneic RBC transfusion in the relationship between pre-operative anaemia and post-operative complications. (Is allogeneic RBC transfusion harmful in the given clinical setting?)

The following sections are dedicated to further introducing the reader to the problem context. In other words, is testing **Hypotheses 1 and 4** the ***core problem***?

1.3. The Anaemia Definition and the Alarming Prevalence

Anaemia is defined as a health condition when a haemoglobin (Hb) level in the blood dropped below a threshold, specifically, below 12 g/dl for non-pregnant women, below 11 g/dl for pregnant women, and below 13 g/dl for men (WHO, 1968). A normal Hb level is essential for the proper transport of oxygen through the cardiovascular system to organs and tissues in the human body (WHO, 2023). In this regard, anaemic patients represent a fragile patient group within a patient population because the restricted oxygen transport leads to limitations during the regenerative process, for instance, after surgery (in a post-operative clinical setting).

Anaemia is *not a disease*. Anaemia is a health condition, a blood disorder, a symptom, a sign, or a signal of *something worse ahead*. Or it can be one of many symptoms that accompany an already existing disease. “Anaemia is often a comorbidity in patients with common noncommunicable diseases”, such as diabetes, cardiovascular disease, chronic kidney disease (WHO, 2021). Also, patients with cancer, surgical, medical, and obstetric patients, or patients with chronic diseases experience anaemia (WHO, 2021; Blood and Beyond, 2021). Upon detecting anaemia, adequate treatment to target the core deficiency is advised by medical specialists. According to WHO (1968): “Anaemia is considered to be a late manifestation of nutritional deficiency, and even mild anaemia is not the earliest sign of such a deficiency… the object of therapy is to correct the underlying deficiency rather than merely its manifestation.”

Anaemia is a global health issue because it imposes a burden on healthcare systems (WHO, 2021; Blood and Beyond, 2021; Safiri et al., 2021). Safiri et al. (2021) estimate the global prevalence of anaemia to be 23% equivalent to roughly 1 in 4 people on earth. WHO (2021) reports an alarming anaemia prevalence of 1.95-2.36 billion people worldwide out of which 1.24-1.46 billion people have iron deficiency anaemia. Estimates indicate that additional 0.98-1.18 billion people (with isolated micronutrient deficiency) may progress towards anaemia (WHO, 2021). WHO (2021) also forecasts that the statistics of only a few countries are heading towards meeting the Global Nutrition Targets 2025 for anaemia set in 2014 one of which is in particular a 50% reduction of anaemia in women of reproductive age (WHO, 2014). Although Safiri et al. (2021) report that Western Europe (which implies the Netherlands) has one of the lowest anaemia prevalence, approx. 5%, anaemia is currently the #1 medical priority for Sanquin Research as stated in the 2021 Annual Report (Sanquin Blood Supply Foundation, 2021).

Incentives to improve the anaemia statistics and to relieve the burden due to anaemia have been an ongoing top priority in healthcare worldwide (WHO, 2021; Blood and Beyond, 2021). WHO (2021) responds with urgency to these alarming statistics.

1.4. Allogeneic RBC Transfusions from the Healthcare Quality Perspective

Administering blood transfusions is generally a safe, widely accessible treatment option for a variety of diseases and health conditions. In this thesis, particularly, a focus is given to the role of RBC transfusions which is an allogeneic intervention (this means that a transfusion recipient accepts foreign blood from a donor, as opposed to autologous blood reinfusion during which the patient’s blood is collected for example, using a blood salvage device during surgery and later reinfused back into the same patient’s body). Besides RBCs, another kind of an allogeneic transfusion intervention is fresh frozen plasma (FFP) transfusions.

Regarding the current transfusion medicine practices in the Netherlands, healthcare establishments including Sanquin follow the Dutch Blood Transfusion Guidelines (de Vries & Haas, 2012). The current version has been in place since 2011. The Dutch Blood Transfusion Guidelines specify strict triggers for allogeneic RBC transfusions to help correct a low Hb level. Specifically, the so-called '4-5-6 rule' sets the thresholds as follows: 4 mmol (6.4 g/dl) for patients aged below 60, 5 mmol (8.1 g/dl) for patients aged above 60, and 6 mmol (9.7 g/dl) for patients of high risk.

Patient groups that receive benefits of RBC transfusions are anaemic patients, patients with other blood disorders, people with cancer or with other various chronic diseases (WHO, 2021; Blood and Beyond, 2021). In the peri-operative clinical setting (around the time of surgery), RBC transfusion can be given prior to, during and/or after surgery (pre-, intra- and/or post-operatively, respectively) but most often during and/or after surgery.

Anaemic patients, when symptomatic, may need treatment, and one of the possible options is to administer allogeneic RBC transfusion. Yet, transfusion is coupled with a burden on the healthcare system in terms of logistics, costs, or questionable safety and efficacy (Blood and Beyond, 2021). In this regard, there are unmet needs accounting for various measures of healthcare quality. Nash et al. (2019) define healthcare quality by the six pillars: safety, efficiency, efficacy, equity, timeliness, and patient-centredness. For example, in terms of safety, blood transfusions may lead to iron overload or immune reactions (Blood and Beyond, 2021) or other adverse outcomes.

1.5. Transfusion Alternatives, Patient Blood Management and the Current Gaps

1.5.1. Transfusion Alternatives

The evidence of adverse outcomes caused by blood transfusions, such as iron overload or immune reactions (Blood and Beyond, 2021), led to a shared perspective across medical professionals that blood transfusion may be harmful to some patients. The Blood and Beyond initiative (2021) on the European level along with WHO (2021) encourages a wide spectrum of stakeholders to further tackle the burden on healthcare systems due to transfusion dependency. In this regard, a lot of focus and efforts by healthcare establishments were put into transfusion alternatives. Some feasible alternative options already exist in current medical practices and some already help reduce the dependency on RBC transfusions. The two alternatives within the scope of the available data in this work will be:

(1) The erythropoietin (EPO) therapy (administered pre-operatively), and
(2) Autologous blood reinfusion (administered intra- and/or post-operatively) during which the patient's own blood is recycled back into the patient's body, after being filtered or washed.

Nevertheless, subject matter experts (Blood and Beyond, 2021; Blood and Beyond, 2020) claim that there have still been unmet needs in terms of the availability of transfusion alternatives.

1.5.2. What is Patient Blood Management?

The concept of Patient Blood Management (PBM) involves but is not limited to the investigation of transfusion alternatives. Considering adverse outcomes of blood transfusions, "our own blood is still the best thing to have in our veins" (Frenzel et al., 2008). "The overarching aim of PBM is to improve patient outcomes, while saving health care resources and reducing costs." (WHO, 2021) PBM is a paradigm shift for transfusion medicine. Preserving patient's own blood reduces the demand for allogeneic blood components (WHO, 2021; Shander et al., 2020). PBM as a concept first emerged in the surgical setting in Australia by Professor James Isbister. In his 2005 publication titled *Updates in Blood Conservation and Transfusion Alternatives*, PBM as a paradigm shift was introduced for the first time (Isbister, 2005). Nevertheless, 25+ medical societies have reached a consensus just recently on a global definition of PBM (Shander et al., 2022). The 15 years needed for reaching a consensus on the PBM definition accentuate the extraordinary level of complexity of the PBM problematics.

> "Patient blood management is a patient-centered, systematic, evidence-based approach to improve patient outcomes by managing and preserving a patient's own blood, while promoting patient safety and empowerment."
>
> (Shander et al., 2022)

Health authorities emphasize key principles of PBM, the so-called three pillars (WHO, 2021; Isbister, 2013; Hofmann et al., 2011):

1. Detection and management of anaemia and iron deficiency,
2. Minimization of blood loss and optimization of coagulation, and
3. Leveraging and optimizing the patient-specific physiological tolerance of anaemia.

Next to PBM, the term **blood health** has been recently used more and more commonly to emphasize treasuring a patient's own blood as a liquid organ (Ozawa, 2023).

1.5.3. Gap in PBM Awareness and Implementation

Under the umbrella of the WHO, subject matter experts from the medical and scientific communities have recently published a report titled ***The Urgent Need to Implement Patient Blood Management*** (WHO, 2021). The list of contributors reflects multi-disciplinarity and a wide spectrum of knowledge that spans transfusion medicine, haematology, hemovigilance, epidemiology, public health administration, health economics, or policy-making, to name a few disciplines. New PBM Implementation Guidelines are currently under development (WHO, 2021) to aim at closing the gap in PBM implementation.

Initiatives to redesign healthcare systems with their infrastructures and to redistribute resources concerning blood transfusion practices have been underway. Various healthcare establishments across the globe have already implemented and reflected on their existing PBM programs and strategies to alleviate blood transfusion overuse and the burden of anaemia as a global health issue (WHO, 2021; So-Osman, 2017). For example, PBM was successfully implemented in Western Australia yielding cost-savings reaching millions of dollars over a six-year period (WHO, 2021).

In the WHO report (2021), subject matter experts bring attention to three drivers for PBM implementation, the so-called three "E's", namely:

(1) Scientific **evidence**,
(2) A strong **economic** argument, and
(3) An **ethical** obligation.

1.5.4. Independent Research involving Blood Product Use together with Patient Outcomes

Patient-centredness is one of the six pillars of healthcare quality (Nash et al., 2019), yet patient-centredness in PBM has been a prevailing gap due to limited studies involving the blood product use together with patient outcomes. This gap is transferred to the limited amount of literature on evidence-based transfusion medicine involving patient outcomes. Thus, new data-driven project incentives involving the analysis of patient outcomes in this context are promising to progressively help close the gap.

Unfortunately, the current commercial ties concerning the initiative on the European level titled Blood and Beyond may hinder independent research. This initiative has recently emerged with a call for multi-disciplinary action to "rethink blood use in Europe to improve outcomes for patients" (Blood and Beyond, 2021). This means, among all, addressing the unmet needs of patients primarily with chronic diseases, solving challenges of transfusion dependency or optimizing blood management (Blood and Beyond, 2021).

1.6. ‘TOMaat’: Transfusion Data Availability from a Randomized Study

The patient-level dataset at hand was originally collected for the purpose of a randomized study, namely, the ‘Transfusie Op Maat’ study (‘TOMaat’ in short, translated as the ‘Customized Transfusion’ study) with data collection performed between 2004 and 2009 (So-Osman et al., 2014a; So-Osman et al., 2014b).

Robust data acquisition led by dr. So-Osman, who served in the role of the study coordinator, was performed on the target patient group that underwent an elective orthopaedic surgery (a total hip- or a total knee-replacement surgery) at four participating hospitals (implying multiple data sources) (So-Osman et al., 2014a; So-Osman et al., 2014b). “This randomized, multicenter, controlled study was registered in the public registry: controlled-trials.com (No. ISRCTN 96327523) and the Dutch Trial Register (No. NTR303).” (So-Osman et al., 2014a)

The raw data has a sample size of 2442 records and contains 533 data fields including patient characteristics, (allogeneic and autologous) transfusion data, data on post-operative patient outcomes (complications) or other intermediate follow-up data (such as lab data).

The research here is approached as a post-hoc, observational study, that is using the RCT dataset as observational data. The further reasoning is that the RCT data were used for another purpose than the RCT itself, that is for the purpose of prediction modelling, inference, and mediation analysis. It is worth noting, nevertheless, that the RCT data collection differs from real-world data (RWD) collection tremendously – especially for the quality of scientific evidence. This RCT particularly was a double-randomized, multi-centre controlled trial (So-Osman et al., 2014a; So-Osman et al., 2014b; So-Osman, 2012). An RCT means that scientifically and statistically, the data exhibits high quality – a very high level of evidence, specifically, level 1b according to the Levels of Evidence by the Centre for Evidence-Based Medicine, CEBM (2009). CEBM adopts the RCT definition from A Dictionary of Epidemiology by Last (2001): “An epidemiological experiment in which subjects in a population are randomly allocated into groups, usually called study and control groups, to receive or not receive a experimental preventive or therapeutic procedure, maneuver, or intervention. The results are assessed by rigorous comparison of rates of disease, death, recovery, or other appropriate outcome in the study and control groups.“

1.7. Challenges in Transfusion Data Availability in the Netherlands

It is problematic to find transfusion data integrated from multiple hospitals that were collected by organizations in the Netherlands other than Sanquin. It seems the data/database infrastructures and/or data acquisition practices have not yet reached maturity due to insufficient funding and/or the lack of incentives.

For example, it was mentioned above in Section 1.4. that cancer patients need transfusions as a part of their treatment. Cancer is ranked with the highest disease burden in the Netherlands (Hilderink et al., 2020), hence, noticeably, cancer research is a key driver in healthcare. In pursuit of longitudinal studies using cancer data, the Netherlands Comprehensive Cancer Organization (‘Integraal Kankercentrum Nederland’, or IKNL) collects cancer patient data from hospitals in the Netherlands and integrates data from the Netherlands Cancer Registry (NCR) (IKNL, 2023). However, after several interviews with IKNL employees, we found that IKNL neither has (direct) access to transfusion data nor uses transfusion data for cancer research.

Dr. So-Osman shared a remark that the discussion topics about transfusion data collection, integration and subsequent opportunities for data analytics remain underrepresented. This state of affairs is due to transfusion being considered a supportive treatment on the hospital floor. Therefore, transfusion data is not usually stored in data registries.

In addition, Sanquin reported difficulties with finding appropriate funding opportunities over the past years for this specific PBM project, which was a reason for postponing this research project until now.

Nonetheless, the above-proposed data from the 'TOMaat' study can be disclosed by Sanquin for this data-driven project. It was shown upon further consultations with dr. So-Osman that data availability is not the core problem for demonstrating the suitability of the proposed methodologies. The reasoning is that this RCT dataset at hand is a "complete" patient-level dataset of sufficient sample size and satisfactory data missingness.

1.8. The Intention of the Study

1.8.1. Problem Formulation

Within patient-level datasets, there could be hidden insights with opportunities to enhance patient-centredness across the PBM and the transfusion medicine landscapes. Plus, there are clinical incentives for evidence-based medicine to relieve the burden due to transfusion dependency (tied to anaemia) in surgical settings. This is the ***action problem*** in this thesis.

Particularly, concerning the patients with pre-operative anaemia, the investigation of the problem context in this chapter has demonstrated that identifying the role of peri-operative RBC transfusion and respective confounding variables is missing granularity (a missing level of detail) in scientific evidence for effective PBM implementation. This means that testing **Hypotheses 1 and 4** (introduced in Section 1.2.) is a research gap and the ***core problem*** for this thesis. The ***knowledge problem*** is to discover whether the hypotheses prove true for the studied patient sample from the elective orthopaedic surgery setting.

Hypothesis 1: RBC transfusion mediates the relationship between pre-operative anaemia and post-operative complications.

Hypothesis 4: There are strong confounding variables associated *both* with:

- Allogeneic RBC transfusion (the mediator), and
- Post-operative complications (the patient outcomes).

1.8.2. The Research Aim and the Research Questions

The aim of the research is to apply suitable applied statistical (learning) methods for testing **Hypotheses 1 and 4**. We will research what supervised learning methods and what mediation analysis model design are appropriate, and if (how) they yield useful results in this context. The execution involves leveraging a patient-level dataset from the 'TOMaat' study (recalling Section 1.6.). The available data restricts the research to the target patient group who underwent elective orthopaedic surgery in the Netherlands (a high-income country).

Definitions of the three dependent variables in the scope of the project are provided in Table 1-1. The respective research question(s) (RQs) are noted in the Context column.

Table 1-1: Boundaries of the dependent variables (target patient outcomes) and corresponding Cases.

Context	Case	Format of the dependent variable	Target patient outcome description with its boundaries
RQ#1, RQ#2,	RBC	binary (1 = yes, 0 = no)	The allogeneic RBC transfusion up to Day 14 (if any) administered first in time to the patient.

Context	Case	Format of the dependent variable	Target patient outcome description with its boundaries
RQ#4			Can be intra- or post-operative.
RQ#1, RQ#2, RQ#3, RQ#4	COM	binary (1 = yes, 0 = no)	Post-operative complication up to Day 14 (if any) after RBC transfusion (if any).
RQ#1	LOS	numeric (days)	Hospital length-of-stay whose timespan may exceed Day 14 after surgery for some patients.

After preliminary data exploration, all project stakeholders agreed that allogeneic RBC transfusion and post-operative complications will be binary dependent variables. Nevertheless, it is known according to consultations with dr. So-Osman (the RCT study coordinator) that some patients may have been administered multiple RBC transfusions and some patients may have experienced more than one post-operative complication.

Dr. So-Osman further advises to focus on allogeneic RBC transfusions up to Day 14 after surgery and post-operative complications up to Day 14 because they are considered more relevant to the clinical need. (The timespan of the care pathway in scope has an upper bound of Day 14 after surgery.) This pertains to the data acquired during the pre-operative phase or in the inpatient clinical setting (during the patient's stay in the hospital) up to Day 14. The reasoning is that during this timeframe, most RBC transfusions are administered to the patients, and most post-operative complications occur in this elective orthopaedic surgery setting. Furthermore, post-operative complications up to Day 14 are more closely tied (correlated) to the length-of-stay (LOS) measure.

Correspondingly, the four RQs are derived based on **Hypotheses 1 and 4** and articulated below.

First, we define the **key patient subgroups** by stratifying the patients based on:

(1) Pre-operative anaemia,
(2) Participating hospitals,
(3) EPO therapy,
(4) Type of surgery (hip or knee, primary or revision),
(5) Blood loss,
(6) Intra-/post-operative autologous reinfusion up to Day 14,
(7) Intra-/post-operative allogeneic RBC transfusion up to Day 14,
(8) Reasonable combinations of (1)-(7), as deemed beneficial.

RQ#1: *How do the target patient outcomes vary for the key patient subgroups?*

Per Table 1-1, the **target patient outcomes** in conjunction with RQ#1 are:

- RBC transfusion up to Day 14 (Case RBC),
- Post-operative complications up to Day 14 (Case COM), and
- Hospital LOS (Case LOS).

Due to its clinical and health-economic relevance, the hospital LOS is also considered as a dependent variable in RQ#1 although the timespan exceeds Day 14 after surgery for some patients.

In response to **Hypothesis 4**:

RQ#2: *What variables are the strong confounding predictors for RBC transfusion up to Day 14, and a post-operative complication up to Day 14?*

RQ#3: *What is the statistical importance of RBC transfusion up to Day 14 acting as the predictor for the occurrence of a post-operative complication up to Day 14?*

Eventually, in response to **Hypothesis 1** and in pursuit of integrating the results of RQ#2:

RQ#4: *What role does RBC transfusion up to Day 14 play in the relationship between pre-operative anaemia and the occurrence of a post-operative complication up to Day 14?*

1.8.3. The Research Scope

The target patient group in this research are elective orthopaedic surgery patients in the Netherlands (a high-income country). We will leverage the patient-level dataset originally collected for the purpose of the 'Transfusie Op Maat' ('TOMaat', 'Customized Transfusion') study. The RCT was registered on controlled-trials.com (No. ISRCTN 96327523) and the Dutch Trial Register (No. NTR303) (So-Osman et al., 2014a; So-Osman et al., 2014b). This data contains 2442 records of patients who underwent a total hip- or a total knee-replacement surgery at one of four participating hospitals.

The timespan of the care pathway has an upper bound of Day 14 after surgery. This pertains to data acquired during the pre-operative phase or in the inpatient clinical setting (during the patient stay in the hospital).

Chapter 2 | Theoretical Background and Literature Review

The Background serves to bridge the terminology of data science with epidemiology. We present the theory relevant to the *core problem* that revolves around utilizing a patient-level dataset with many measured variables that describe a complex surgical setting of transfusion medicine and PBM. A literature review helps investigate and frame an appropriate setup of selected supervised learning models and of mediation analysis to obtain effect estimates for two selected variables (pre-operative anaemia and RBC transfusion relative to post-operative complications). The chapter consists of a presentation of the following concepts: Supervised Statistical Learning and Its Applications in Transfusion Medicine and PBM (Section 2.1.), Considerations for Choosing a Supervised Statistical Learning Model (Section 2.2.), Model Performance and (Cross-)Validation (Section 2.3.), Causal Inference and Confounding (Section 2.4.), and Mediation Analysis (Section 2.5.). Lastly, Section 2.6. is dedicated to the summary and conclusion of this theoretical background.

2.1. Supervised Statistical Learning and Its Applications in Transfusion Medicine and PBM

"*Statistical learning*[1] refers to a vast set of tools for *understanding data.*" (James et al., 2021) In the case of *supervised statistical learning* models (in scope of this work), a dependent variable was already observed and measured. Model development encompasses one or more independent variables (serving as an input for the model, predictors, features, or simply, variables). The independent variables relate to a desired dependent variable(s) (also known as a response, outcome, or output variable).[2] Depending on the format of the dependent variable, two implementations are possible in *supervised statistical learning* model development: *classification* models for binary dependent variables (in scope of this work), and *regression* models for continuous dependent variables.

[1] To clarify the accompanying terminology, the term *machine learning* (ML) substitutable also for artificial intelligence (AI) is considered equivalent to *statistical learning* for some stakeholders. Yet, *machine learning* emerged from the computer science ecosystem whereas *statistical learning* emerged from statistics.

[2] The use of synonyms for these terms is restricted in this work (recalling the Established Terminology).

Particularly for transfusion medicine and/or PBM, the recent literature summarizes promising or already implemented use cases of AI and ML (both supervised and unsupervised) (Meier & Tschoellitsch, 2022; Dhiman et al., 2023; Šuster et al., 2023; Huang et al., 2018). In a scoping review, Meier and Tschoellitsch (2022) list some opportunities for operationalization: prediction of blood loss and transfusion, prediction of the outcome of anaemia and transfusion, decision support, prediction of the efficacy of pre-operative anaemia of a patient with iron and/or EPO, or the development of an expert system for various pathways during PBM.

Many existing prediction models involving blood transfusion as a dependent variable were subject to a systematic review by Dhiman et al. (2023). Specifically, seven scientific publications in this review are related to the orthopaedic (hip and knee) surgery setting. All of the seven articles establish transfusion as a binary RBC transfusion variable (which is consistent with our work) out of which four papers tackle the variable solely as post-operative RBC transfusion, one paper as pre-/intra-/post-operative RBC transfusion, and two papers as intra-/post-operative RBC transfusion. We further pay attention to these two studies (Huang et al., 2018; Rashiq et al., 2004) whose design overlaps with the RBC transfusion as a dependent variable in our study to the greatest extent. Huang et al. (2018) developed logistic regression and random forest. Rashiq et al. (2004) developed a logistic regression model.

2.2. Considerations for Choosing a Supervised Statistical Learning Model

Clinical researchers ask for simple models and ease in their interpretability despite the high complexity of transfusion medicine. A supervised learning model choice arises from the model's primary purpose: *prediction*, *inference*, or both (James et al., 2021). Because no previous modelling using supervised learning was performed on the 'TOMaat' dataset, random forest, logistic regression, and lasso are further examined in the face of *both prediction and inference*. Random forest and logistic regression (Section 2.2.3. and Section 2.2.4.) are one of the most commonly used non-parametric and parametric models, respectively, in academia (Lundberg et al., 2020). Also our results will then be comparable to Huang et al. (2018) and Rashiq et al. (2004) that were introduced in the above section. Lasso (Section 2.2.5., parametric model) is a suitable addition to modelling in pursuit of the ease in model interpretability thanks to a convenient variable selection functionality (James et al., 2021).

2.2.1. Prediction and Inference

In some cases, *prediction* and *inference* approaches could go hand in hand (James et al., 2021).

In *inference*, the modelling focus is targeted to studying the associations between the dependent variable and the independent variables (James et al., 2021). *Inference* is greatly about revealing insights of the variable importance in explaining the dependent variable, and, if possible, estimating the parameters of the model (James et al., 2021; Hastie et al., 2009).

A predictive approach enriches the research with model validation, evaluating the predictive power and the accompanying model performance measures. The function estimate is often known as the *black box* function because one does not necessarily need to know its exact form (James et al., 2021), for example, in case of random forest.

2.2.2. The Trade-off between Model Flexibility and Interpretability

Choosing a suitable model involves a bias-variance trade-off. The *reducible* error element is present due to the squared bias of the model and the *irreducible* error occurs due to natural variability of the data. The *reducible* error can be improved by choosing a more appropriate, more flexible model which often comes at the cost of interpretability. The *irreducible* error caused by the random *error* does not always approach zero either due to unmeasured variables that may still serve well as the model input, or due to leaving out some useful measured variables from the model input (James et al., 2021).

Figure 2-1 depicts this trade-off. Parametric methods feature advantages over non-parametric methods especially because of their ease in interpretability. Yet, choosing a parametric model of a lower flexibility may go at the expense of model performance. And, subsequently, higher flexibility (of random forest) is a promising model characteristic relating to better model performance. Random forests are found as a special case of *Bagging* (bootstrapping + aggregating).

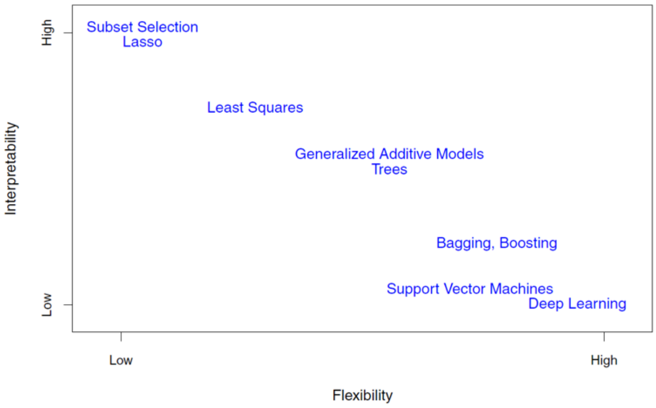


Figure 2-1: The trade-off between model flexibility and interpretability for various statistical learning method; Excerpted from James et al. (2021).

2.2.3. Logistic Regression and the Inference Tools

Logistic regression is a parametric model – a generalized linear model (GLM), and a common choice for inference in classification modelling. It is designed for binary dependent variables to operate with maximum-likelihood parameters, $\beta_0, \beta_1, \dots, \beta_i$ (James et al., 2021; Hastie et al., 2009). The next equation exhibits the expression of *multiple logistic regression* (for multiple independent variables): $\log\left(\frac{p(X)}{1-p(X)}\right) = \log(odds) = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i$. The left-hand side $\log\left(\frac{p(X)}{1-p(X)}\right)$ is known as the *log-odds* or *logit* and has a linear form. $\frac{p(X)}{1-p(X)}$ is known as the *odds* and acquires values between 0 and ∞ (reflecting on very low and very high probabilities of the outcome, respectively).

Variable selection procedures are essential due to issues arising from sparse data or due to the phenomenon of multicollinearity. It is not guaranteed that logistic regression (or lasso) can always be implemented on an arbitrary dataset. Multicollinearity occurs in linear models when two or more input variables are closely related which is undesired because it causes issues later in estimating individual association and effects of variables on the dependent variable (James et al., 2021). Variable selection (such as subset selection) is a desired step in the modelling procedure in order to eventually detect only the significant variables. Filtered in is eventually only a subset of variables that sufficiently explain the joint effect on the dependent variable (Hastie et al., 2009). However, some variables may have non-linear effects within a linear model fit. They shall preferably not be excluded from the model (Hastie et al., 2009).

The inference tools for logistic regression are expressed analytically as standard errors and Z scores (test statistics) for the coefficient estimates. A Z score for each model input variable is tied to a null hypothesis that the coefficient is zero while the others are not – known as the Wald test (Hastie et al., 2009). A given independent variable is significantly important if the Z score is significant. A Wald test built on a chi-squared test can be used to determine statistical significance of categorical variables.

The odds ratio is a powerful, relational inference tool and serves to express the effect of a given independent variable on the dependent variable. An increase or decrease is observed depending on the positive or negative sign of the parameter β, respectively. For example (for binary variables), one can deduce for patients who were administered RBC transfusion compared to the patients who were not that there will be an increase in the odds of the post-operative complications by $100 * \left(e^{\widehat{\beta_{k=1}}} - 1\right)\%$ with the 95% confidence interval of $100 * (e^{\widehat{\beta_{k=1}} \pm 2\widehat{SE_{k=1}}} - 1)\%$.

2.2.4. Random Forest and the Inference Tools

Even some non-parametric models, such as random forest may offer useful variable importance outputs as a means to yield *inference* insights. Random forests are one of the most opted black box off-the-shelf tree-based algorithms especially for their accurate predictions on new datasets with many covariates (Hastie et al., 2009). They can readily accommodate missing data, sparse data, or nonlinear relationships (James et al., 2021; Hastie et al., 2009). McAlexander and Mentch (2020) argue that "the superior predictive performance of random forests can be harnessed to examine the same kinds of relationships in the data that [scientists] typically seek to uncover with conventional parametric models, including making inferences about the marginal effect of independent variables".

Random forest operates on the basis of stratifying or segmenting the variable space into regions (James et al., 2021). The *random* element represents producing multiple trees (as an *ensemble* method) that are combined to nodes and leaves to ultimately yield a single "consensus prediction" (James et al., 2021). The individual building blocks (simple trees) can be called *weak learners* for their tendency to generally perform poorly on their own. Many of these noisy, but quite unbiased models are "bagged" and averaged. Compared to all i independent variables for bagging, random forests operate with a random sample of $\approx \sqrt{i}$ independent variables at each split. This step offers a great advantage called *decorrelating* of the trees. *Decorrelating* decreases the variability and improve the reliability of the averaged tree model. Particularly, on average, $\frac{i-\sqrt{i}}{i}$ of the splits will not contain the strongest variable; hence, other variables will have better opportunities to contribute to the model development and output.

Variable importance for random forest is commonly expressed by these two relative measures (with no thresholds). Generally, the higher the measure is, the more important the variable is in the model input. (The definitions are referenced from James et al. (2021) and Hastie et al. (2009).)

- *Mean Decrease in Accuracy* measures the loss of accuracy if the variable was removed from the model. It is tied to the classification error rate determined as the predictions on the out-of-bag samples. The out-of-bag observations are the data points not used to fit a given bagged tree.
- *Mean Decrease in Gini index:* The Gini index is a measure specific to classification trees and is particularly a measure of node purity (quality of the split). The Gini index acquires values between 0 and ∞. A small value represents that the node contains observed data points from a single category. To obtain the Mean Decrease in Gini index, first, summed up is the total amount of the Gini index decreased by splits over a given independent variable. This sum is then averaged over all individual *weak learner* trees.

Additionally, a *partial dependence plot* (PDP) is useful for making inferences for random forest despite its non-parametric nature, thus, leading to enhancing its interpretability. Molnar (2022) and Friedman (2001) explain that: [1] PDPs exhibit the contribution of the independent variable on the predicted dependent variable through the marginal effect; [2] the PDP progression for numeric variables can be linear, monotonic, or complex; and [3] PDP functions of categorical variables are displayed as bar plots. This effect is calculated by "accounting for the (average) effects of the other variables", not ignoring those effects (Hastie et al., 2009). For classification problems, the PDP function is in the *logit* (*log-odds*) form, on a logarithmic scale (R Documentation, 2023a; Greenwell, 2017; Hastie et al., 2009; Friedman, 2001). This implies that the random forest function acquires the relationship with $\log\left(\frac{p(X)}{1-p(X)}\right)$ as it is the case for logistic regression. Hence, similarly to logistic regression (Section 2.2.2.) the *odds ratios* can be subsequently calculated: $\log\left(\frac{p(X)}{1-p(X)}\right) = \log(odds) = random\ forest\ function\ component.$

2.2.5. Lasso and the Inference Tools

Lasso is a parametric modelling approach and can be applied for the purpose of a variable selection (shrinkage) to improve the (linear) model performance and interpretability (James et al., 2021). Lasso, the Least Absolute Shrinkage and Selection Operator, was first proposed in 1996 and is an alternative fitting procedure as coefficients of some input variables are forced to acquire the value of zero (James et al., 2021; Tibshirani, 1996). Lasso is a powerful regularization method and may lead to a linear (logistic regression) model to be outperformed by its upgraded linear version. (Other common regularization methods include ridge regression, stepwise selection, or principal components regression (James et al., 2021).)

The inference tools for lasso are coefficient estimates and odds ratios coupled with p-values.

2.3. Model Performance and (Cross-)Validation

(Cross-)Validation encompasses procedures of comparing the model fit to the chosen frame of reference. Internal validation (in scope of this project) revolves around splitting the dataset into a train and test set or executing a bootstrap method. (External validation would involve evaluating the model fit using a new, external dataset.)

For *classification* models (with binary dependent variables), model performance is commonly evaluated as:

- The discrimination measure displayed by the Receiver Operating Curve (ROC) with the accompanying Area Under the Curve (AUC) that acquires values between 0 and 1: According to Hosmer and Lemeshow (2000, pp. 160–164), the AUC above 0.7 suggests an acceptable model performance, 0.8 and above is considered excellent, 0.5 means no discrimination;
- The measures deduced from the confusion matrix (Figure 2-2), such as accuracy, sensitivity (recall), or specificity.

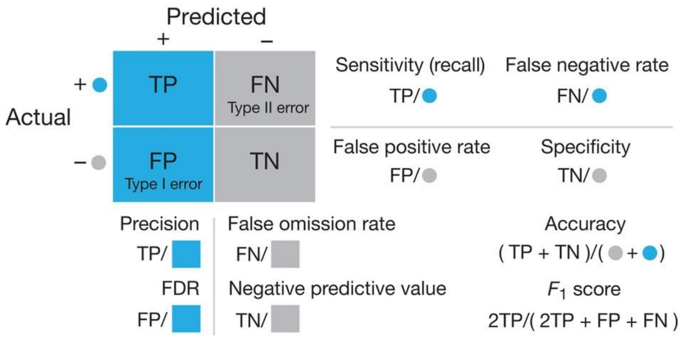


Figure 2-2: Confusion matrix reflecting a binary dependent variable (a binary classifier): Excerpted from Lever et al. (2016). TP = true positives, FP = false positives, FN = false negatives, TN = true negatives, FDR = false discovery rate.

Performance measures of *regression* models are omitted because they are out of scope of this work.

Model performance is an essential accompanying element of uncertainty upon model development in pursuit of interpretation and transferring the modelling insights into practice. Model performance indicates how well the model reflects the reality. In recent scientific literature, there is no clear consensus whether variable importance (such as effect estimates) alone without accompanying model performance measures is transferrable to practical applications. Research solely focused on inference typically lacks validation procedures because validation can be done using external, artificial data.

2.4. Causal Inference and Confounding

Causal inference is an approach different from supervised statistical learning because it places a special attention on the effect of a specific variable under investigation (i.e. pre-operative anaemia, RBC transfusion) relative to the dependent variable (here, RBC transfusion, post-operative complication). In comparison, in supervised learning, variable selection methods, such as lasso, may lead to the elimination of that variable under investigation (i.e. pre-operative anaemia, RBC transfusion) from the model. This elimination would then represent an insignificant (approximately zero) effect of that variable.

Causal inference is founded on the "adage": correlation does not imply causation; in other words, causation does not always occur due to correlation (Hernán & Robins, 2020). The terminology of causal inference distinguishes between association and causation (Hernán & Robins, 2020). To offer transparency and clarity of the terminology, the consistency of the confounding definition in the field of clinical epidemiology and data science is assessed in the paragraphs below.

Accustomed to the field of epidemiology, Hernán and Robins (2020) define confounding as a type of systematic bias resulting from the discrepancy between causation and association. Confounding occurs when the treatment and outcome share a common cause – this means when the association measure generally differs from the effect measure (Hernán & Robins, 2020). This definition describes that the presence of a given independent variable tweaks the resulting effect of another independent variable due to a common causality relation with the dependent variable.

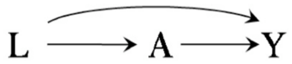


Figure 2-3: An illustration of confounding: The treatment, A, and the outcome, Y, share a common cause, L; Excerpted from Hernán and Robins (2020).

In comparison, the books *An Introduction to Statistical Learning* by James et al. (2021) and *The Elements of Statistical Learning* by Hastie et al. (2009) describe confounding still in terms of correlations among independent variables. Such correlation may cause variability in the resulting coefficients as well as in variable importance depending on the model input choices. Two specific logistic regression examples were adapted from these two literature pieces to illustrate confounding:

- On its own (in a single logistic regression model), a given independent variable X_j results to be significantly important and has a positive coefficient. Yet in a multiple logistic regression, this means in the presence of other independent variables, X_j results to be also significantly important but with a negative coefficient. Reasoning: X_j is correlated with another independent variable in the multiple logistic regression model input (James et al., 2021).
- On its own (in a single logistic regression model), a given independent variable X_k is significant. Yet in the presence of other independent variables (in a multiple logistic regression), X_k is insignificant. Reasoning: This phenomenon is caused due the correlation among other independent variables present in the model input (Hastie et al., 2009).

Causal inference deals with three main types of systematic bias: *confounding*, *selection* and *measurement*, yet, this work is limited to studying *confounding*. In short, a *selection* bias results from selecting limited data for analysis. And a *measurement* error naturally occurs as variables cannot be measured perfectly. Additional relationships and phenomena are described in the terminology of causal inference. For example, in social sciences, Cinelli et al. (2022) reflects on the problem of "bad controls" (as opposed to "good controls" – confounders) when an addition of some variables to the parametric model results in unintended discrepancies of the effect estimates known as the "omitted variable bias".

2.5. Mediation Analysis

Mediation analysis is a technique to study causal inference with a focus on selected exposure and mediator (treatment) variable components (Figure 2-4). Mascha et al. (2013) studied the exposure of anaemic patients on the numeric dependent variable length-of-stay. Saager et al. (2013) supplemented the work by specifying pre-operative anaemia as the exposure in the context of wound contamination (mediator) and infection (outcome). Here, alcohol use was incorporated as a confounder. Later, in donor studies, de Groot et al. (2019) completed mediation analysis to investigate the effect of population density on blood lipid levels.

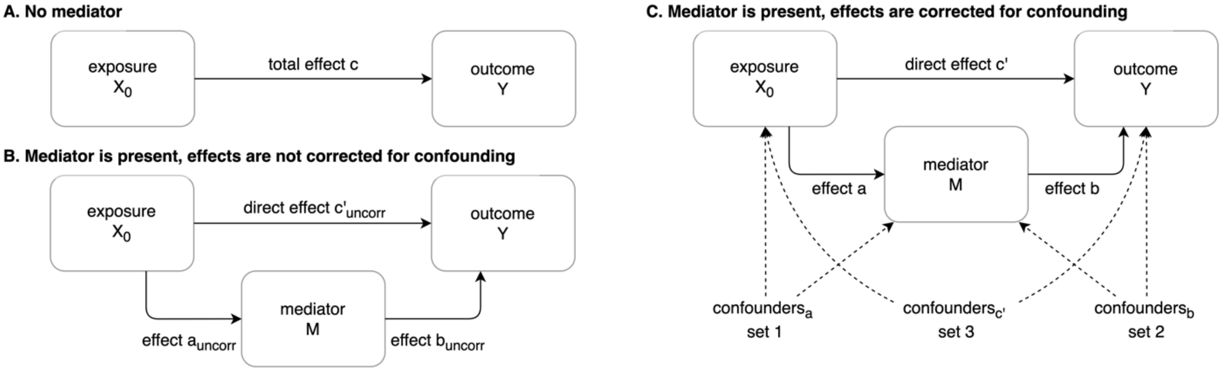

A. No mediator
exposure X_0
total effect c
outcome Y
B. Mediator is present, effects are not corrected for confounding
exposure X_0
direct effect c'_{uncorr}
outcome Y
mediator M
effect a_{uncorr}
effect b_{uncorr}
C. Mediator is present, effects are corrected for confounding
exposure X_0
direct effect c'
outcome Y
mediator M
effect a
effect b
confounders_a set 1
$\text{confounders}_{c'}$ set 3
confounders_b set 2

Figure 2-4: Mediation analysis model uncorrected (B, $uncorr$) and corrected (C) for confounding.

Mediation analysis serves to decompose the *total* effect of the exposure on the outcome into two components:

[1] An *indirect* (mediation) effect explained by the mediator composed of the effect a and effect b: The product-of-coefficients method is a preferred method for estimating an indirect effect in a mediation analysis model with a binary mediator and a binary outcome (Rijnhart et al., 2021).

[2] A *direct* effect not explained by the mediator, c'.

All components of the mediation analysis model can be of a binary, categorical or numeric format. In this work, the outcome, exposure and mediator component formats will be binary, and the effects will be estimated using odds ratios. In this case, mediation analysis involves logistic regression modelling. Next to it, besides odds ratios, other effect measures exist in causal inference. Hernán and Robins (2020) indicate that the portfolio of the causal effect measures also encompasses causal risk difference, risk ratio, or other summaries.

The *total* effect follows an expression PNDE*TNIE (or TNDE*PNIE) considering a binary mediator and a binary outcome (Rijnhart et al., 2021). h represents the XM interaction effect on the outcome. According to Rijnhart et al. (2021): "In the absence of XM interaction, the h coefficient equals zero and drops out of the equations. The CDE, PNDE, and TNDE then all reduce to exp(c'), i.e. the natural direct effect (NDE). The PNIE and TNIE then both equal the PNIE… and is termed the natural indirect effect (NIE)."

2.6. Summary and Conclusion

This theoretical background helped frame a suitable experimental setup of logistic regression and random forest models in pursuit of tackling the core problem. In an interpretable way, clinicians in transfusion medicine and PBM seek to estimate effects of pre-operative anaemia and RBC transfusion based on a patient-level dataset with many covariates. Thus, interpretation of supervised learning models and mediation analysis are the ultimate capstone for this work. Lasso is a possible variable selection method that offers improvement in interpretability of logistic regression by reducing the number of input variables. It is essential to pay careful attention to choices in model input because they are key in determining the desired effect estimates. Neglecting correlations among independent variables and neglecting the confounding phenomenon may lead to misleading, biased conclusions about the effect estimates. Collinearity also shall be treated because it leads to bias. Effect estimates for desired independent variables can be then obtained using odds ratios for both supervised modelling approaches.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \log(odds) = logistic\ regression\ function\ component = \beta_0 + \beta_1 X_1 + \cdots + \beta_i X_i$$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \log(odds) = random\ forest\ function\ component$$

To enrich the research, the purpose of supervised learning modelling can be both inference and prediction. The predictive approach, particularly, offers the information on whether there is a substantial penalty in model performance for logistic regression compared to random forest. It is worth to examine if random forest as a non-parametric model (significantly) outperforms logistic regression (a parametric model) in this setting. Next to it, there is no clear consensus whether choosing a tree-based method, such as random forest, results in a substantial compromise on interpretability. Hence, the implementation of the random forest model will be further considered as an innovative supplement to inference and mediation analysis.

Chapter 3 | Methodology and Experimental Setup

Figure 3-1 on the next page is a bird view on the key methodology steps. This chapter is aligned with it in its composition: Massive data exploration and data cleaning efforts (Step 1) are mapped in Section 3.1. Consolidation of the variable selection process (Step 2) and the preparation of input variables can be found in Section 3.2. Procedures of statistical testing for the Exploratory Data Analysis (Step 3) are written in Section 3.3. The procedure of model fitting and validation (Steps 4-7) is described in Section 3.4. The choices for the selection of strong confounders (Step 8) are reflected in Section 3.5. And the Mediation Model Setup (Step 9) is mapped in Section 3.6.

Data pre-processing, analysis and modelling were executed in the programming environment of the R Statistical Software, version 4.1.2.

3.1. Data Exploration and Data Cleaning

3.1.1. Characteristics of the Raw Dataset

The raw dataset from the 'TOMaat' study was provided by Sanquin and its origin was presented earlier in Section 1.6. The patient flow diagram in Figure 3-2 is a representation of the clinical setting according to So-Osman et al. (2014a; 2014b). Figure 3-3 offers a detailed view. The raw data has a sample size of 2442 records and contains 533 data fields including patient characteristics, allogeneic and autologous transfusion data, data on post-operative patient outcomes (complications) or other intermediate follow-up data (i.e. lab data). The raw data characteristics are summarized in Table 3-2.

Massive data cleaning efforts were performed in order to establish the dependent variables representing RBC transfusion (Section 3.1.3.) and post-operative complication up to Day 14 (Section 3.1.4.). A crucial pre-requisite is the retrieval of dates for these dependent variables according to the definitions in Table 3-2. There is a need to sequence the events in time because of the causality direction in the proposed model for mediation analysis (Section 1.2.). The following content is dedicated to the consolidation of the respective pre-processing tasks.

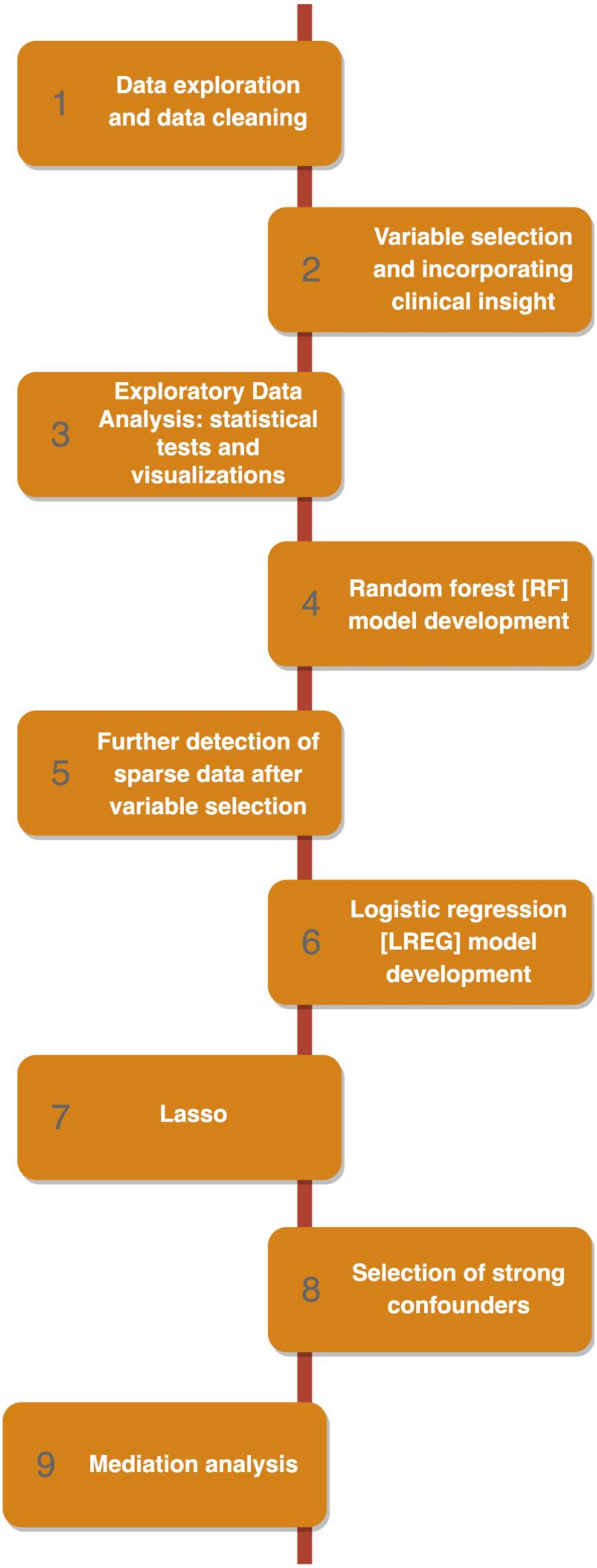

1 Data exploration and data cleaning
2 Variable selection and incorporating clinical insight
3 Exploratory Data Analysis: statistical tests and visualizations
4 Random forest [RF] model development
5 Further detection of sparse data after variable selection
6 Logistic regression [LREG] model development
7 Lasso
8 Selection of strong confounders
9 Mediation analysis

Figure 3-1: Bird view on the key methodology steps.

An overview of the dependent variables for the four research questions (per the Context column) is presented in Table 3-1 below.

Table 3-1: Boundaries of the dependent variables (target patient outcomes) and corresponding Cases (excerpt of Table 1-1).

Context	Case	Format of the dependent variable	Target patient outcome description with its boundaries
RQ#1, RQ#2, RQ#4	RBC	binary (1 = yes, 0 = no)	The allogeneic RBC transfusion up to Day 14 (if any) administered first in time to the patient. Can be intra- or post-operative.
RQ#1, RQ#2, RQ#3, RQ#4	COM	binary (1 = yes, 0 = no)	Post-operative complication up to Day 14 (if any) after RBC transfusion (if any).
RQ#1	LOS	numeric (days)	Hospital length-of-stay whose timespan may exceed Day 14 after surgery for some patients.

Adequate choices and assumptions accompany the experimental setup and the scenario setup (Section 3.1.5. and 3.1.6.).

3.1.2. Reasoning on why RBC transfusion is a variable in the 'TOMaat' study, not a parameter

All patients who entered 'TOMaat' gave their informed consent for study participation (So-Osman, 2014a; So-Osman, 2014b). Patient's agreement to the potential administration of (RBC) transfusion was one of the numerous inclusion criteria.

This patient consent is the key information for treating allogeneic RBC transfusion as a dependent variable in this work (not a pre-defined parameter). Hence, we can study the role of RBC transfusion in this elective orthopaedic surgery setting using the proposed methods.

According to the clinical insight of dr. So-Osman, there are two strict transfusion triggers pertaining to the 'TOMaat' study in accordance with the Dutch Blood Transfusion Guidelines (2011). This means that during the 'TOMaat' study, patients received RBC transfusion as long as at least one of these two triggers was reached.

(1) Specific Hb threshold value according to the '4-5-6 rule' (introduced in Section 1.4.), and
(2) A large amount of blood loss determined per expert-based decisions (by a surgeon or an anaesthesiologist, typically during surgery, thus, in an acute setting).

An important assumption is tied to establishing RBC transfusion as a dependent variable:

- The Dutch Blood Transfusion Guidelines (2011) were strictly followed.

Further elaboration on this key modelling aspect is provided in the chapter Recommendations on Future Work.

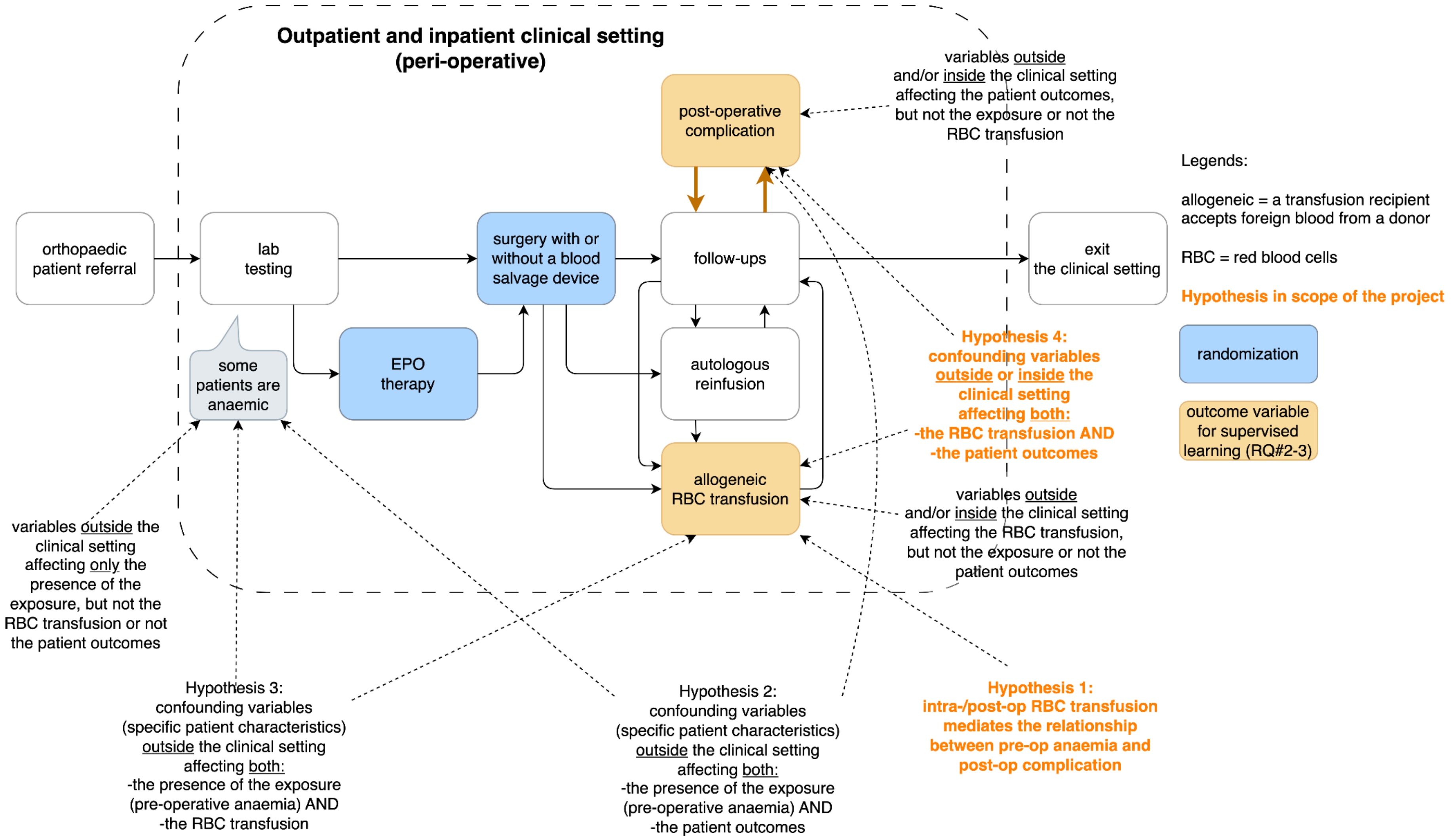

Outpatient and inpatient clinical setting (peri-operative)
orthopaedic patient referral
lab testing
some patients are anaemic
EPO therapy
surgery with or without a blood salvage device
follow-ups
post-operative complication
autologous reinfusion
allogeneic RBC transfusion
exit the clinical setting
variables outside and/or inside the clinical setting affecting the patient outcomes, but not the exposure or not the RBC transfusion
Legends:
allogeneic = a transfusion recipient accepts foreign blood from a donor
RBC = red blood cells
Hypothesis in scope of the project
randomization
outcome variable for supervised learning (RQ#2-3)
Hypothesis 4: confounding variables outside or inside the clinical setting affecting both: -the RBC transfusion AND -the patient outcomes
variables outside and/or inside the clinical setting affecting the RBC transfusion, but not the exposure or not the patient outcomes
variables outside the clinical setting affecting only the presence of the exposure, but not the RBC transfusion or not the patient outcomes
Hypothesis 3: confounding variables (specific patient characteristics) outside the clinical setting affecting both: -the presence of the exposure (pre-operative anaemia) AND -the RBC transfusion
Hypothesis 2: confounding variables (specific patient characteristics) outside the clinical setting affecting both: -the presence of the exposure (pre-operative anaemia) AND -the patient outcomes
Hypothesis 1: intra-/post-op RBC transfusion mediates the relationship between pre-op anaemia and post-op complication

Figure 3-2: Patient flow diagram of the 'TOMaat' study.

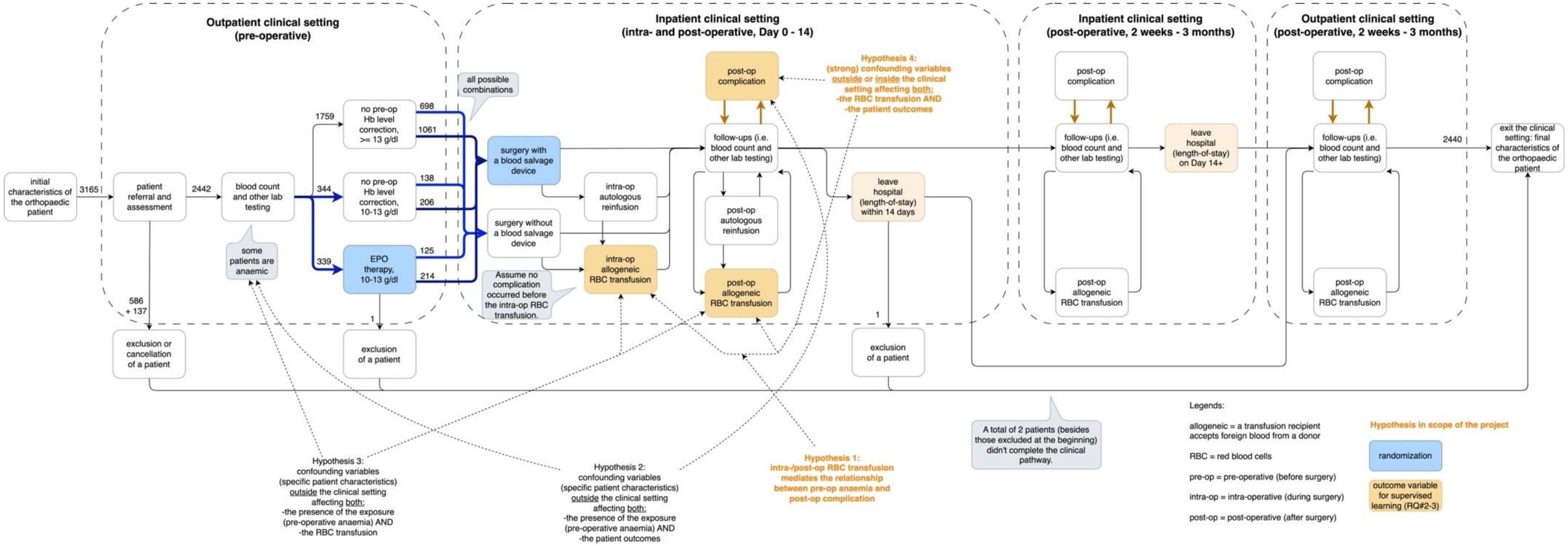

Outpatient clinical setting
(pre-operative)
Inpatient clinical setting
(intra- and post-operative, Day 0 - 14)
Inpatient clinical setting
(post-operative, 2 weeks - 3 months)
Outpatient clinical setting
(post-operative, 2 weeks - 3 months)
initial characteristics of the orthopaedic patient
3165
patient referral and assessment
2442
blood count and other lab testing
some patients are anaemic
1759
344
339
no pre-op Hb level correction, >= 13 g/dl
no pre-op Hb level correction, 10-13 g/dl
EPO therapy, 10-13 g/dl
698
1061
138
206
125
214
all possible combinations
surgery with a blood salvage device
surgery without a blood salvage device
intra-op autologous reinfusion
intra-op allogeneic RBC transfusion
Assume no complication occurred before the intra-op RBC transfusion.
586 + 137
exclusion or cancellation of a patient
1
exclusion of a patient
post-op complication
follow-ups (i.e. blood count and other lab testing)
post-op autologous reinfusion
post-op allogeneic RBC transfusion
Hypothesis 4:
(strong) confounding variables outside or inside the clinical setting affecting both:
-the RBC transfusion AND
-the patient outcomes
leave hospital (length-of-stay) within 14 days
1
exclusion of a patient
post-op complication
follow-ups (i.e. blood count and other lab testing)
leave hospital (length-of-stay) on Day 14+
post-op allogeneic RBC transfusion
post-op complication
follow-ups (i.e. blood count and other lab testing)
post-op allogeneic RBC transfusion
2440
exit the clinical setting: final characteristics of the orthopaedic patient
Hypothesis 3:
confounding variables
(specific patient characteristics)
outside the clinical setting
affecting both:
-the presence of the exposure
(pre-operative anaemia) AND
-the RBC transfusion
Hypothesis 2:
confounding variables
(specific patient characteristics)
outside the clinical setting
affecting both:
-the presence of the exposure
(pre-operative anaemia) AND
-the patient outcomes
Hypothesis 1:
intra-/post-op RBC transfusion
mediates the relationship
between pre-op anaemia and
post-op complication
A total of 2 patients (besides those excluded at the beginning) didn't complete the clinical pathway.
Legends:
allogeneic = a transfusion recipient accepts foreign blood from a donor
RBC = red blood cells
pre-op = pre-operative (before surgery)
intra-op = intra-operative (during surgery)
post-op = post-operative (after surgery)
Hypothesis in scope of the project
randomization
outcome variable for supervised learning (RQ#2-3)

Figure 3-3: Patient flow diagram of the 'TOMaat' study: Detailed version with patient counts to reflect the double randomization.

Table 3-2: Number of data fields in raw dataset per segment of the clinical path.

	Pre-operative data	Intra-operative data	Post-operative data, Day 0-14	Post-operative data, 2 weeks to 3 months	Not identified (i.e. pre-calculated)
Number of data fields	125	59	52 (Day 0-1), 93 (Day 2 up to Day 14)	74	130

3.1.3. Establishing the Dependent Variable: RBC Transfusion up to Day 14

Massive data cleaning efforts were performed to establish the variable representing RBC transfusion (RBC_Transfusion). The date of RBC transfusion was stored as a free text field. A patient may have been administered more than one RBC transfusion. Hence, in this project, the dependent variable is specific to the occurrence of the first RBC transfusion date (Day) per Table 3-2 shown earlier.

The dates were converted to the units of Days where Day 0 is the day of surgery. Two situations occurred during assessing the dates of allogeneic RBC transfusion:

[1] Date of post-operative RBC transfusion was known. Subsequently, for each patient and for each intra-/post-operative RBC transfusion up to Day 14, a calculation was done:

Day of RBC transfusion = date of RBC transfusion – surgery date
where Day of RBC transfusion can acquire the values {0, 1, ..., 14}

Numerous data inconsistencies were treated by visual inspection and resolved by manual correction or exclusion (per Section 3.1.7.).

[2] Date of RBC transfusion was unknown. Subsequently, the patient records were excluded (per Section 3.1.7.).

Figure 3-4 represents the distribution of the occurrence of the first post-operative RBC transfusion up to Day 14, and distinguishes the patients who were also administered intra-operative RBC transfusion. According to the treemap in Figure 3-6 (left), approx. ¼ patients received their first post-operative RBC transfusion on Day 1.

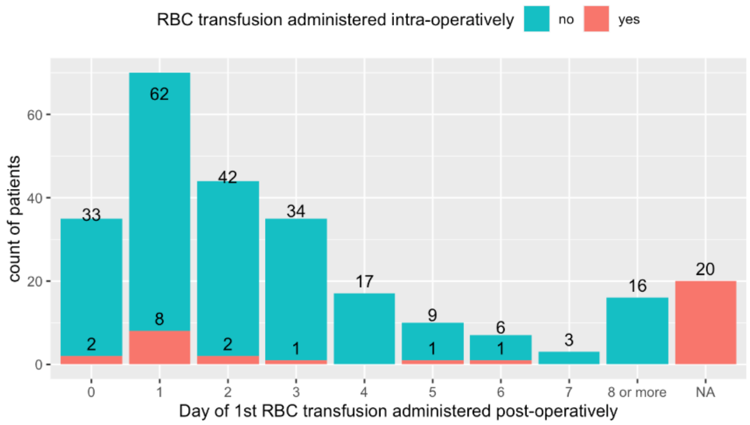


Figure 3-4: Distribution of the occurrence (Day) of the first post-operative RBC transfusion up to Day 14; intra-operative RBC transfusion is treated separately.

[3] Day 0 was assigned to the patients with intra-operative RBC transfusion.

Figure 3-5 represents the distribution of the occurrence of the first RBC transfusion up to Day 14 after incorporating the intra-operative RBC transfusion. According to the treemap in Figure 3-6 (right), approx. ¼ patients received their first RBC transfusion on Day 0 (intra- or post-operative), and another ¼ patients received their first RBC transfusion on Day 1. The highest rates of RBC transfusions are seen on Days 0-3.

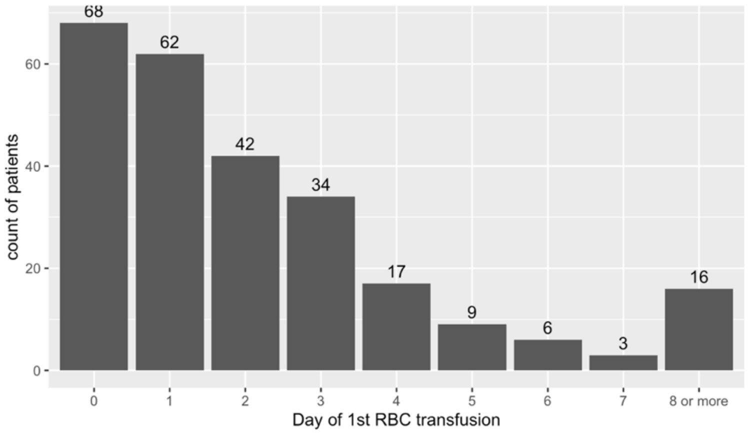


Figure 3-5: Distribution of the occurrence (Day) of the first RBC transfusion up to Day 14 after incorporating intra-operative RBC transfusion.

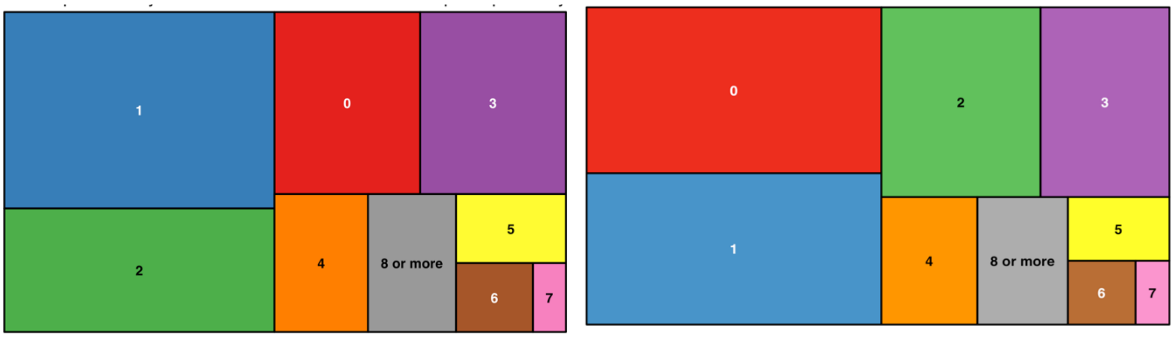


Figure 3-6: Treemaps of the occurrence (Day) of the first RBC transfusion administered solely post-operatively (left) and upon incorporating the intra-operative RBC transfusion (right).

All records with a feasible Day of RBC transfusion acquired 1 for the binary dependent variable, 0 otherwise.

Errors in data entries were found. 5 patients received their first RBC transfusion later than on Day 14 which is out of scope of the timeframe (the data field under investigation was already designated to 0 to 14 Days). These outliers were kept in the dataset, and treated by relabelling the binary dependent variable adequately to acquire the value of 0, no longer 1.

Eventually, a total of 257 (10.6%) patients were administered RBC transfusion up to Day 14 out of the total sample size of 2426 patients.

3.1.4. Retrieving the Dates of Post-operative Complication up to Day 14

Massive data cleaning efforts were performed to establish the variable representing post-operative complication up to Day 14. Again, the dates were converted to the units of Days where Day 0 is the day of surgery. A total of 40 auxiliary variables representing 15 types of complications contributed to the establishment of this dependent variable. There were 15 binary and 15 date variables representing each type of complication, and 10 free text fields denoting details about the complication. A composite variable available in the raw dataset could not have been utilized due to the need to retrieve the complication dates and to sequence the RBC and COM events in time for the purpose of mediation analysis.

An excessive missingness of the complication dates was prevalent. Per the 15 binary fields, 34.2% (count 264) of all 772 recorded complications had no accompanying date (per 15 date fields). Figure

3-7 displays the distribution of the occurrence of the complications up to Day 14 for which the date was known (count 508, 65.8%). The distribution is right-skewed with the highest rates of complications on Days 0-4.

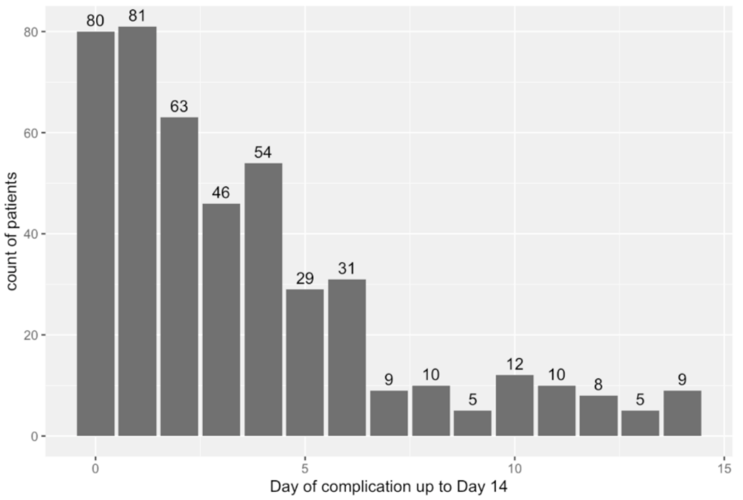


Figure 3-7: Distribution of the occurrence (Day) of the complications up to Day 14; known occurrences only are captured in the plot (count 508, 65.8%).

Some patients experienced more than 1 complication (up to 7 complications per patient). Figure 3-8 shows this observation.

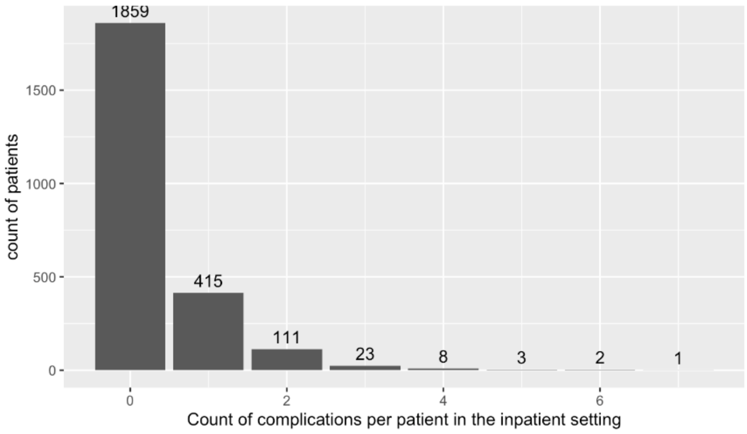


Figure 3-8: The counts of patients in relation to the number of complications up to Day 14 per patient; all complications up to Day 14 are captured.

Insights into the extent of missingness for complication dates can be found in Figures 3-9 and 3-10. Counts and proportions of complications up to Day 14 relative to 15 complication types are displayed. Unfortunately, due to high missingness rates, the complication dates cannot be deduced by (advanced) imputation methods.

A choice was made to set up two scenarios (pessimistic and optimistic) to deal with this extent of date missingness. The process is described further in Section 3.1.5. and 3.1.6.

Errors in data entries were found. The Days of complications up to Day 14 were recorded with values larger than 14 for 53 complications. This is out of scope of the desired timeframe. The respective patient records were kept in the dataset, yet treated by relabelling the respective binary variables adequately to acquire the value of 0, no longer 1, and the dates were erased from the date fields.

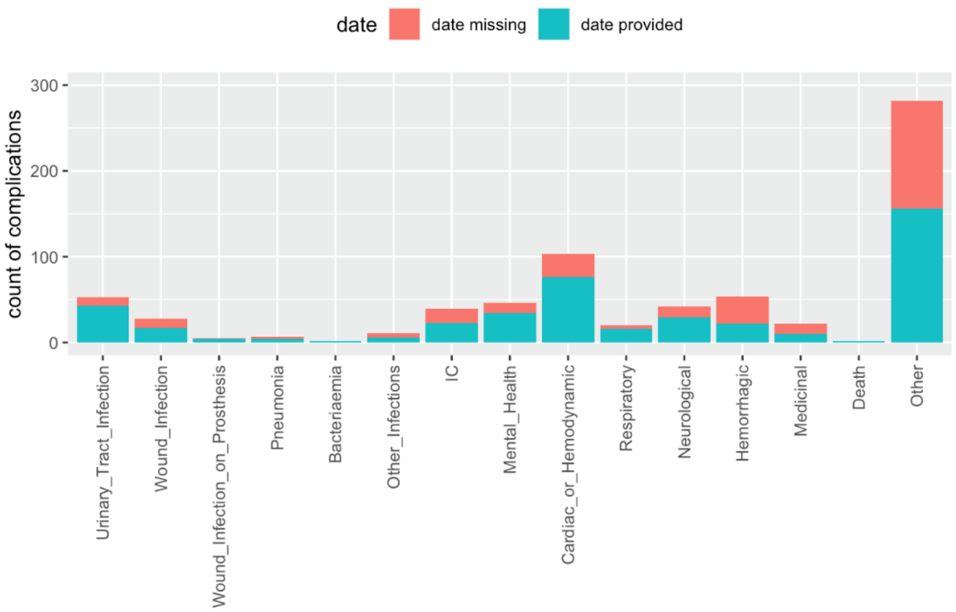


Figure 3-9: Insight into the extent of missingness for complication dates: Counts of complications up to Day 14 relative to 15 complication types.

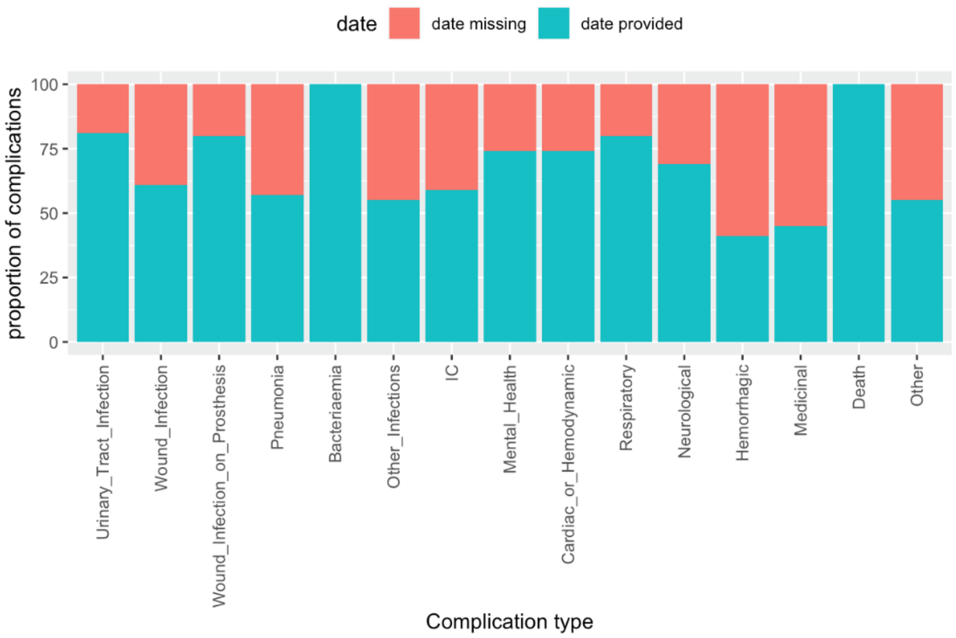


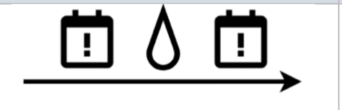
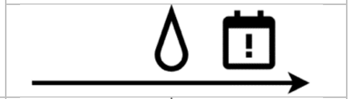
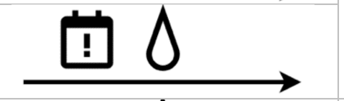
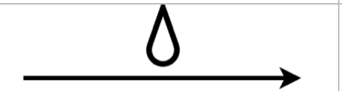
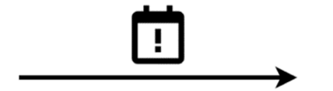
Figure 3-10: Insight into the extent of missingness for complication dates: Proportions of complications up to Day 14 relative to 15 complication types.

3.1.5. Placing the Dependent Variables in Time Sequence and Scenario Setup

By default, four patient groups were distinguished for the purpose of assigning the binary format of the dependent variable. The procedure is illustrated in Table 3-3. The value of the COM variable is assigned per the fifth column if the complication date is known and different from the RBC transfusion date.

Assigning the **COM** variable for Patient Groups 1-0, 0-1 and 0-0 is trivial. Next, a focus is given to Patient Group 1-1 (the patients who had both RBC as well as a post-operative complication) to sequence these events adequately.

Table 3-3: Four patient groups useful for establishing the dependent variables defined up to Day 14.

Patient group	A patient received allogeneic RBC transfusion	Value of RBC transfusion (dependent variable)	A patient had a post-operative complication (not the dependent variable)	Value of post-transfusion complication, COM (dependent variable)	Visual representation*	Patient count
1-1	Yes	1	Yes	1 if at least one complication occurred after RBC transfusion		145
						
				0 if none of the complications occurred after RBC transfusion		
1-0	Yes	1	No	0		112
0-1	No	0	Yes	1		399
0-0	No	0	No	0		1770

* Legends: 💧 = RBC transfusion (dependent variable); 📅 = at least 1 post-operative complication

48 patient records from 145 records in Patient Group 1-1 had at least 1 missing complication date or the same Day of complication as RBC. The procedure of establishing the two scenarios with dependent variables $\mathrm{COM}_{\mathrm{PES}}$ and $\mathrm{COM}_{\mathrm{OPT}}$ is described as follows:

[1] Date of complication is known. Subsequently, for each relevant patient in Patient Group 1-1 and for each post-operative complication up to Day 14:

Day of complication = date of complication – surgery date
where Day of complication can acquire the values {0, 1, ..., 14}

If (Day of complication) > (Day of RBC transfusion), then COM = 1.

If (Day of complication) = (Day of RBC transfusion), then the pessimistic and optimistic scenarios apply (Assumption#1). In the field of epidemiology, this case is called misclassification.

Pessimistic scenario: Assume $\mathrm{COM}_{\mathrm{PES}}$ = 1,
Optimistic scenario: Assume $\mathrm{COM}_{\mathrm{OPT}}$ = 0.

[2] Date of complication is unknown. Subsequently, patient records were examined by integrating the information from an auxiliary file about serious adverse events and clinical insight was consolidated in pursuit of retrieving the sequence of events. Then, two scenarios were considered (Assumption#3). For each relevant patient in Patient Group 1-1:

Pessimistic scenario: Assume for $\mathrm{COM}_{\mathrm{PES}}$ that all remaining complications with missing dates occurred after RBC transfusion.
Optimistic scenario: Assume for $\mathrm{COM}_{\mathrm{OPT}}$ that all remaining complications with missing dates occurred before RBC transfusion.

All complications occurred after intra-operative RBC transfusion, if any (Assumption#2, 15 patients).

The scenario setup according to the three principal assumptions for Patient Group 1-1 and the patient counts for all other patient groups are consolidated in Tables 3-4 and 3-5. The pessimistic and optimistic scenarios, respectively, are distinguished with 112 (4.6%) versus 64 patients (2.6%) who experienced both RBC and a post-operative complication up to Day 14.

Table 3-4: Patient counts for the pessimistic scenario.

Patient group	RBC	$\mathrm{COM}_{\mathrm{PES}}$	Patient count	Total patient count (proportion)
1-1	1	1	64 + 48	112 (4.6%)
		0	33	145 (6.0%)
1-0		0	112	
0-1	0	1	399	399 (16.4%)
0-0		0	1770	1770 (73.0%)

Table 3-5: Patient counts for the optimistic scenario.

Patient group	RBC	$\mathrm{COM}_{\mathrm{OPT}}$	Patient count	Total patient count (proportion)
1-1	1	1	64	64 (2.6%)
		0	33 + 48	193 (8.0%)
1-0		0	112	
0-1	0	1	399	399 (16.4%)
0-0		0	1770	1770 (73.0%)

In Figure 3-11, eventually, two Sankey diagrams represent the two scenarios. The pessimistic scenario led to 511 patients with a post-operative complication (21.1%) compared to the optimistic scenario with 463 patients with a post-operative complication up to Day 14 (19.1%) out of the 2426 patients in total.

3.1.6. Overview on the Exclusion of Patient Records (Listwise Deletion)

Listwise deletion was performed on the patient records due to specific inadequacies leading to the exclusion of:

- 13 records due to the missing value of pre-operative haemoglobin (more in Section 3.2.2.),
- 1 record due to an unknown date of RBC transfusion,
- 1 record because the date of post-operative RBC transfusion up to Day 14 preceded the surgery date,
- 1 record because the date of post-operative complication up to Day 14 preceded the surgery date.

A total of 16 records were excluded from the raw dataset of 2442 records. The cleaned dataset has 2426 patient records.

3.1.7. Establishing the Dependent Variable: Hospital length-of-stay

The dependent variable LOS required minimum cleaning efforts because the raw dataset offers this variable. The extent of missingness is 17 patient records. This raw variable represents the timespan between the admission date at the hospital (1-3 days prior to the surgery date) until the discharge day. The dependent variable LOS has a mean of 8.0 days, median of 7 days, min of 1 day, and max of 98 days.

Furthermore, efforts were made to investigate the calculated timespan between the actual surgery dates and discharge dates as provided in the raw data. Yet, unfortunately, numerous inconsistencies were observed among these calculations. Thus, a choice was made to use the raw LOS variable for the purpose of this project (particularly, RQ#1).

An important assumption is tied to the LOS variable:

- RBC transfusion up to Day 14 was administered during the respective LOS (in the inpatient setting), not in the outpatient setting. If the RBC transfusion was administered in the outpatient setting, then the RBC transfusion is an infeasible independent (input) variable LOS for those patient records. A thorough verification is needed if a follow-up project is considered (i.e. per the mediation analysis segment in Section 6.3.).

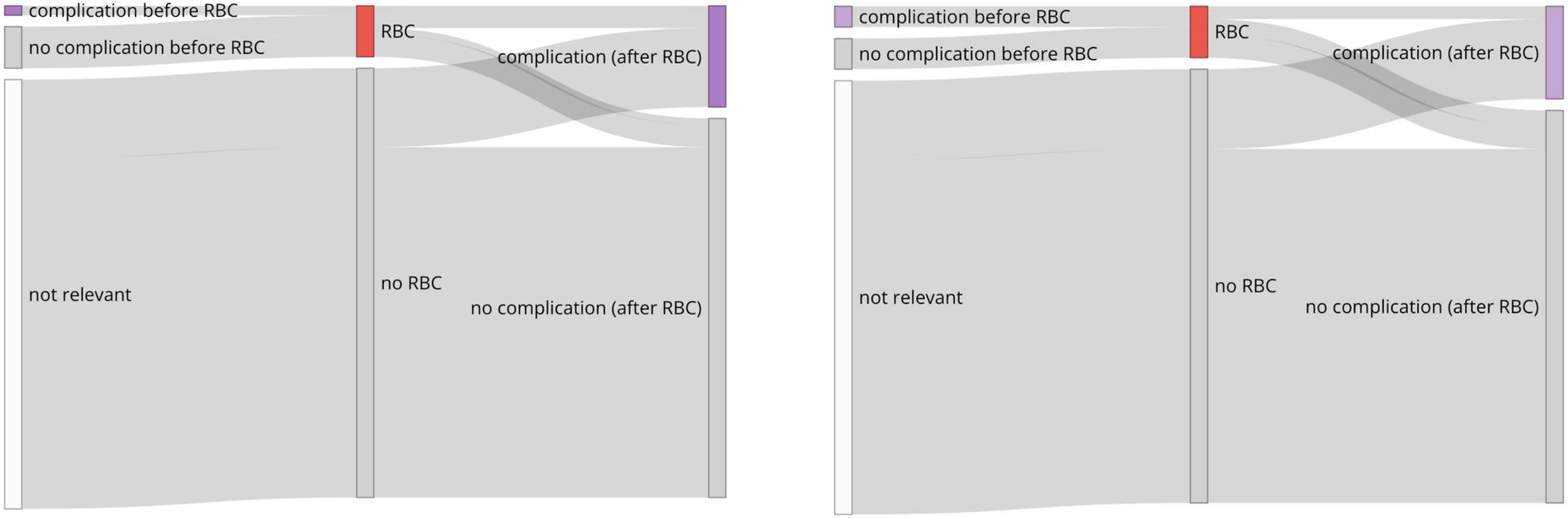


Figure 3-11: Sankey diagrams with proportions of patients who underwent RBC transfusion (if any) and/or post-operative complication up to Day 14 (if any): Pessimistic (left) and optimistic scenario (right).

3.2. Variable Selection

3.2.1. The Variable Selection Process

The variable selection process is sketched in Figure 3-12 below. Variable selection was accompanied with numerous rounds of consultations and verifications with the clinical expert to ensure the understanding of the patient flow and of the information contained in data fields. The variable selection involved subset selection based on variable feasibility consolidated as 12 choices explained in the figure. The threshold for acceptable data missingness was set to 10%.

Eventually, the dataset size for modelling was reduced from 533 data fields to 46 variables out of which 41 are default input variables for modelling, 4 are dependent variables ($\mathrm{RBC\_Transfusion}$, $\mathrm{COM_{PES}}$, and $\mathrm{COM_{OPT}}$), and the remaining one is the variable $\mathrm{Total\_Blood\_Loss\_during\_Surgery}$ not used for modelling (used solely in RQ#1). The COM modelling starts with 41 input variables, and Case RBC comprises of a default of 32 input variables. Table 3-6 provides further details on the 41 input variables. There are 32 binary, 3 categorical, and 6 numeric input variables for the COM Cases. For RBC Case, it is 28 binary, 3 categorical, and 1 numeric variables. All binary and categorical variables were factorized. Later, during model development, 3 additional variables exhibiting sparsity were excluded particularly from LREG and lasso models (as reported in Section 3.4.2.).

3.2.2. Preparation of the Input Variables: Adaptation

Data preparation encompassed adaptation of 4 input variables that were established based on existing data fields. This procedure pertains to participating hospitals, pre-operative anaemia, and the year of surgery. The procedure for the fourth variable $\mathrm{RBC\_Transfusion}$ that was also established through adaptation (grouping) was already presented earlier in Section 3.1.3.

Participating hospitals, $\mathrm{Hospital}$, was subject to pooling of the Hospital2 and Hospital5 because this hospital has two different locations.

Pre-operative anaemia, $\mathrm{Anaemia\_Pre\_Op}$, was determined based on three haemoglobin (Hb) fields measured pre-operatively and before the start of the EPO therapy, if any. A consolidation was done of the Hb fields $\mathrm{VHBINCL}$ (39.4% missingness), $\mathrm{VHB0}$ (0.5% missingness), and $\mathrm{VHB1}$ (0.5% missingness), in this sequence. The remaining 13 patient records (0.5%) for which the Hb value was not retrieved were excluded using listwise deletion. The Hb values were available in the units of mmol/L so the thresholds according to WHO (1968) were converted from g/dL to mmol/L and the binary form was derived depending on gender as follows:

- For men: if Hb falls below 13 g/dL (8.07 mmol/L), pre-operative anaemia = 1, or 0 otherwise;
- For women: if Hb falls below 12 g/dL (7.45 mmol/L), pre-operative anaemia = 1, or 0 otherwise.

Year of surgery, $\mathrm{Surgery\_Year}$, was created by extracting the year from the surgery date field.

3.2.3. Preparation of the Input Variables: Data Imputation Methods

The indicator method deemed feasible for many variables as listed in Table 3-6. The choice of this method was validated with the clinical expert, along with a suitable imputed value.

Multiple imputation using KNN was implemented for $\mathrm{Surgery\_Duration}$. A use was made of the $\mathrm{preProcess()}$ function from the caret library, and eventually the $\mathrm{predict()}$ function of the RANN library. The K argument was set to the square root of the row number (49) of the dataset.

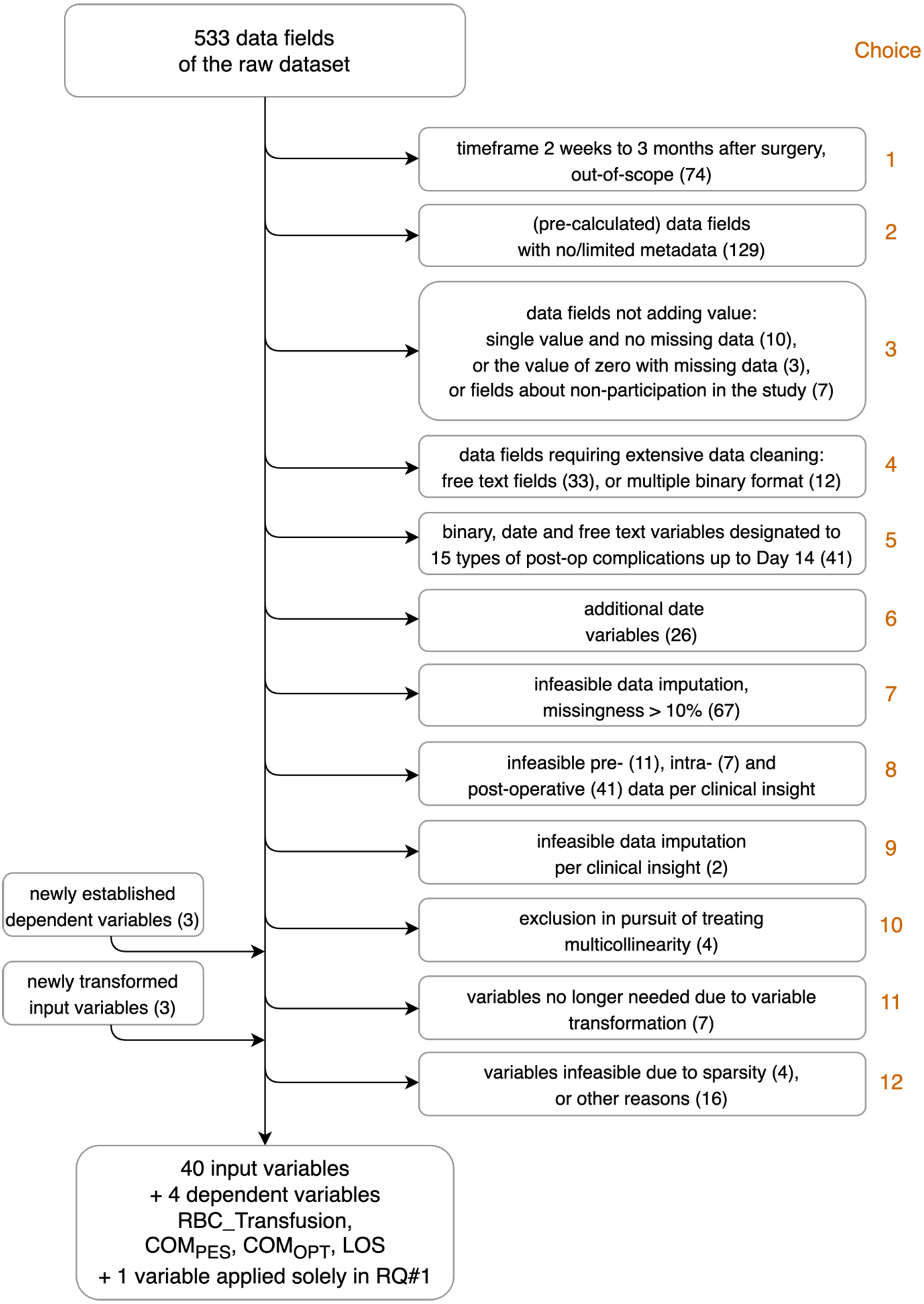

533 data fields
of the raw dataset
Choice
timeframe 2 weeks to 3 months after surgery, out-of-scope (74)
1
(pre-calculated) data fields with no/limited metadata (129)
2
data fields not adding value: single value and no missing data (10), or the value of zero with missing data (3), or fields about non-participation in the study (7)
3
data fields requiring extensive data cleaning: free text fields (33), or multiple binary format (12)
4
binary, date and free text variables designated to 15 types of post-op complications up to Day 14 (41)
5
additional date variables (26)
6
infeasible data imputation, missingness > 10% (67)
7
infeasible pre- (11), intra- (7) and post-operative (41) data per clinical insight
8
infeasible data imputation per clinical insight (2)
9
newly established dependent variables (3)
exclusion in pursuit of treating multicollinearity (4)
10
newly transformed input variables (3)
variables no longer needed due to variable transformation (7)
11
variables infeasible due to sparsity (4), or other reasons (16)
12
40 input variables
+ 4 dependent variables
RBC_Transfusion,
COM$_{PES}$, COM$_{OPT}$, LOS
+ 1 variable applied solely in RQ#1

Figure 3-12: Consolidation of the variable selection process.

Table 3-6: Metadata overview of input variables after subset selection (NA's = missing data).

Variable category	Variable label	Variable name	Case RBC	Case COM	Variable type (#categories excl. NA's, or units)	Listing of unique values (distribution)	% missingness	Data imputation technique (value imputed in indicator method)	Remark
Patient characteristics and pre-operative data	v_{1}	Hospital	✓	✓	categorical: nominal (4)	Hospital1 (401), Hospital2 (956), Hospital3 (602), Hospital4 (467)	0.00%	none	adapted (pooling)
	v_{2}	Age	✓	✓	numeric: integer	min 19, med 71, max 93	0.00%	none	-
	v_{3}	Gender	✓	✓	binary	1 (male, 738), 2 (female, 1688)	0.00%	none	-
	v_{4}	Hip_or_Knee_1	✓	✓	binary	1 (975), 0 (1451)	0.00%	none	-
	v_{5}	Primary_or_Revision_1	✓	✓	binary	1 (181), 0 (2245)	0.00%	none	-
	v_{6}	Hip_or_Knee_2	✓	✓	binary	1 (14), 0 (2412)	99.38%	indicator method (0)	-
	v_{7}	Osteoarthritis	✓	✓	binary	1 (2105), 0 (321)	0.87%	indicator method (0)	-
	v_{8}	Cardiovascular_Disease	✓	✓	binary	1 (1235), 0 (1191)	2.84%	indicator method (0)	-
	v_{9}	CVA	✓	✓	binary	1 (81), 0 (2345)	2.18%	indicator method (0)	-
	v_{10}	COPD	✓	✓	binary	1 (196), 0 (2230)	1.81%	indicator method (0)	-
	v_{11}	Diabetes_Mellitus	✓	✓	binary	1 (286), 0 (2140)	1.94%	indicator method (0)	-
	v_{12}	Rheumatoid_Arthritis	✓	✓	binary	1 (281), 0 (2145)	2.39%	indicator method (0)	-
	v_{13}	Increased_Risk_Group	✓	✓	binary	1 (92), 0 (2334)	0.29%	indicator method (0)	-
	v_{14}	Corticosteroids	✓	✓	binary	1 (132), 0 (2294)	0.16%	indicator method (0)	-
	v_{15}	NSAIDs	✓	✓	binary	1 (768), 0 (1658)	0.29%	indicator method (0)	-
	v_{16}	Anticoagulation	✓	✓	binary	1 (486), 0 (1940)	0.16%	indicator method (0)	-
	v_{17}	Antibiotics	✓	✓	binary	1 (25), 0 (2401)	0.33%	indicator method (0)	-
	v_{18}	Insulin	✓	✓	binary	1 (117), 0 (2309)	0.25%	indicator method (0)	-
	v_{19}	Antihypertensiva	✓	✓	binary	1 (1083), 0 (1343)	0.12%	indicator method (0)	-
	v_{20}	Cardiac_Medication	✓	✓	binary	1 (364), 0 (2062)	0.25%	indicator method (0)	-
	v_{21}	Pulmonary_Medication	✓	✓	binary	1 (205), 0 (2221)	0.29%	indicator method (0)	-
	v_{22}	Smoking	✓	✓	binary	1 (331), 0 (2095)	0.37%	indicator method (0)	-
	v_{23}	EPO	✓	✓	binary	1 (227), 0 (2199)	0.08%	indicator method (0)	-
	v_{24}	Anaemia_Pre_Op	✓	✓	binary	1 (214), 0 (2212)	0.00%	none	adapted

Variable category	Variable label	Variable name	Case RBC	Case COM	Variable type (#categories excl. NA's, or units)	Listing of unique values (distribution)	% missingness	Data imputation technique (value imputed in indicator method)	Remark
	V25	Surgery_Year	✓	✓	categorical: ordinal (6)	2004 (66), 2005 (320), 2006 (454), 2007 (767), 2008 (809), 2009 (10)	0.00%	none	adapted
Intra-operative data	V26	Prosthesis_Type	✓	✓	categorical: nominal (4)	1 (976), 2 (1386), 3 (16), unclassified (48)	1.94%	indicator method (unclassified)	-
	V27	Minimally_Invasive_in_case _of_Total_Hip_Prosthesis	✓	✓	binary	1 (118), 0 (2308)	57.67%	indicator method (0)	-
	V28	Temperature_Drop_Prevention	✓	✓	binary	1 (2104), 0 (322)	2.39%	indicator method (0)	-
	V29	Anticoagulant_Standard	✓	✓	binary	1 (1795), 0 (631)	0.00%	none	-
	V30	Antibiotic_Prophylaxis_Standard	✓	✓	binary	1 (1804), 0 (622)	0.04%	indicator method (0)	-
	V31	Antifibrinolytic_Blood_Loss _Lowering_Medication	✓	✓	binary	1 (7), 0 (2419)	0.00%	none	-
	V32	Antifibrinolytic_Cyclokapron	(✓)	(✓)	binary	1 (5), 0 (2421)	99.79%	indicator method (0)	sparse
	V33	Surgery_Duration	-	✓	numeric: integer (minutes)	min 25, med 90, max 630	2.02%	multiple imputation using KNN	-
	V34	Colloids	-	✓	numeric: integer (mL)	min 0, med 500, max 3500	5.40%	indicator method (0)	-
	V35	Crystalloids	-	✓	numeric: integer (mL)	min 0, med 1500, max 10 000	1.48%	indicator method (0)	-
	V36	Cell_Saver	-	✓	binary	1 (271), 0 (2155)	39.74%	indicator method (0)	-
	V37	Cell_Saver_Collection	-	✓	numeric: integer (mL)	min 0, med 0, max 7097	91.80%	indicator method (0)	-
	V38	Cell_Saver_Reinfusion	-	✓	numeric: integer (mL)	min 0, med 0, max 2117	91.34%	indicator method (0)	-
	V39	Other_Transfusions	-	(✓)	binary	1 (5), 0 (2421)	41.30%	indicator method (0)	sparse
	V40	FFP	-	(✓)	binary	1 (3), 0 (2423)	99.55%	indicator method (0)	sparse
Intra- or post-operative data	V41	RBC_Transfusion	-	✓	binary	1 (257), 0 (2169)	0.00%	none	adapted (grouping)

Figure 3-13 is a representation of the final modelling setup (32 input variables for Case RBC, and 41 input variables for the COM Cases).

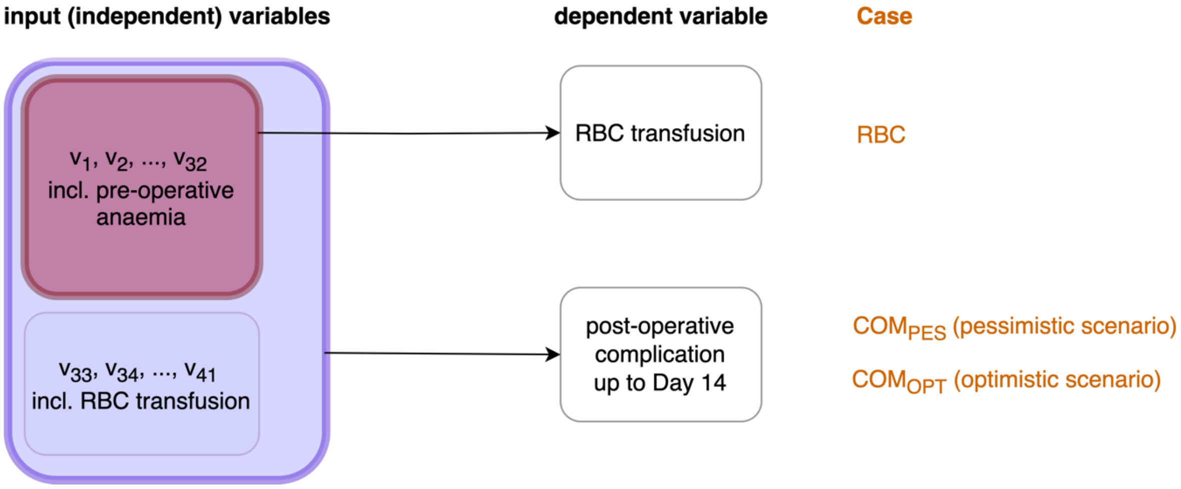

input (independent) variables
dependent variable
Case
$v_1, v_2, \ldots, v_{32}$
incl. pre-operative anaemia
$v_{33}, v_{34}, \ldots, v_{41}$
incl. RBC transfusion
RBC transfusion
post-operative complication up to Day 14
RBC
COM_{PES} (pessimistic scenario)
COM_{OPT} (optimistic scenario)

Figure 3-13: Modelling setup of Case RBC, COM_{PES} and COM_{OPT}.

3.3. Univariate Tests and Visualizations in Exploratory Data Analysis

Exploratory data analysis comprised of univariate tests anvisualizations. Univariate tests were accomplished with chi-squared tests performed using the chisq.test() function of the stats package. The implementation of visualizations in R involved mainly the package ggplot2.

3.4. Choices for Supervised Learning Model Development and (Cross-) Validation

Model development was modularized per Case and model to manage the overall volume of programming tasks and to reduce the risk of human error. Each module represents a separate R code file. The work was structured to a total of 9 modules (R code files): 1 for RF + 1 for LREG + 1 for lasso for each of the three Cases (RBC, COM_{PES}, and COM_{OPT}). For demonstration, Appendix E contains the R code of Case RBC (all three types of models).

Reproducibility was ensured by setting the random seed value (1234, same for all modules).

It was confirmed there is no issue due to an imbalanced dataset. This check constituted of a simple calculation of the percentage of occurrences (cases) of the three binary dependent variables (10.6% for RBC_Transfusion, 21.1% for COM_{PES} and 19.1% for COM_{OPT}). All percentages reach above the generally accepted proportions (there is no unequal distribution).

3.4.1. Random Forest (RF)

The random forest (RF) model development encompasses 7 aspects (substeps):

[1] Model validation using the train/test split: To conduct an internal validation of the model, a dataset was subject to a train/test split with a split ratio of 0.7. This is equivalent to training the model with a trainset size of 1698 patient records, and testing the model on a testset of 728 records.

[2] Model fit: Random forest model fit was achieved using the function randomForest() from the randomForest package. The default parameter mtry (the number of random variables

collected at each split) was set to 6 for the RBC Case and 7 for the COM Cases which corresponds to the square root of the number of variables after executing the model.matrix() operation on the dataset (square root of 42 and 51 variables, respectively). The number of trees to grow (ntree) was set to the default of 500 trees. The model was fit on the trainset. The out-of-bag (OOB) error plot was constructed using the OOB element of the model$err.rate object. Predictions were calculated by the conventional function predict() of the stats package.

[3] Model calibration curve: The graph was plot by executing the val.prob() function of the rms package.

[4] Model performance measures: The auc(roc()) functions from the pROC library served to calculate the AUC. The DeLong method using the function ci.auc() of the same package provided the 95% CI of the AUC. The ROC curve was plot using the geom_roc() function from the plotROC package. Additional model performance measures (accuracy, sensitivity and specificity) per each cutoff value were extracted from a for-loop after running the confusionMatrix() function of the caret library.

[5] Variable importance: The values for the variable importance plot were obtained using the varImpPlot() function of the randomForest package. And the partial dependence plots (PDPs) were constructed using the partial() function of the pdp library coupled with the autoplot() function to achieve ggplot-like aesthetics. For the interpretation of PDPs in RQ#3, it was assumed that the first data point pertains to the positive cases (patients who were administered RBC transfusion) despite the label 0 in the graph. The interpretation of the outputs was tested using different mtry parameter values. The Discussion chapter (Section 5.3.) elaborates more on this observation.

[6] Tuning and cross-validation: The trainControl() function of the caret library was set up to execute 5-fold cross-validation, and a random search was chosen as the suitable mode using the argument search='random'. The mtry parameter was set to default (6 for the RBC Case and 7 for the COM Cases per substep [2] above). The tuning was done using the train() function (also of caret) by passing the trainControl() object to the trControl argument. Accuracy was the target metric for this tuning operation with 3 trees to grow and the tuneLength of 15 to control the computation time.

[7] Variable importance of the tuned RF model: The variable importance tasks as described above (the variable importance plot, the partial dependence plots) were performed in the same manner as in substep [5], this time with the tuned RF model.

3.4.2. Further Detection of Sparse Data

Because the parametric models LREG and lasso are sensitive to sparse data, further treatment was required prior to fitting the models. Then a total of 3 binary variables that caused sparse data issues were removed for the purpose of LREG and lasso model development (reported in Table 3-7, and consistent with Table 3-6 introduced earlier). These variables causing sparse data issues were detected via visual inspection in an increasing manner (from 3 (0.1%) to 5 (0.2%) occurrences out of 2442 patient records). This means that for LREG and lasso, the number of input variables decreased from 41 to 38 for COM Cases, and from 32 to 31 input variables for the RBC Case.

Table 3-7: Variables excluded for LREG and lasso modelling due to sparse data issues.

Variable label	Variable name	RBC[RF]	RBC[tuned RF]	RBC[LREG]	RBC[lasso]	COM[RF]	COM[tuned RF]	COM[LREG]	COM[lasso]	Variable type	# (%) occurrence
v32	Antifibrinolytic_Cyclokapron	✓	✓	-	-	✓	✓	-	-	binary	5 (0.2%)
v39	Other_Transfusions	-	-	-	-	✓	✓	-	-	binary	5 (0.2%)
v40	FFP	-	-	-	-	✓	✓	-	-	binary	3 (0.1%)

3.4.3. Logistic Regression (LREG)

The logistic regression (LREG) model development encompasses 5 aspects:

[1] **Model validation using the train/test split**: [Executed in the same manner as for RF, Section 3.4.1.]

[2] **Model fit:** Using the **glm()** function from the stats package, the LREG model was fit on the testset. Predictions were calculated by the common function **augment()** of the **broom** package.

[3] **Model calibration curve**: [Executed in the same manner as for RF, Section 3.4.1.]

[4] **AUC and ROC:** [Executed in the same manner as for RF, Section 3.4.1.]

[5] **Variable importance**: Coefficients, standard errors, z-values and p-values were available already by running **glm()**, or **summary()**. Odds ratios (as well as p-values) were displayed by executing the function **odds.ratio()** from the **questionr** package. Wald test was performed on the categorical variables (**Hospital**, **Surgery_Year** and **Prosthesis_Type**) using the **wald.test()** function.

3.4.4. Lasso

The lasso model development encompasses 5 aspects:

[1] **Model validation using the train/test split**: [Executed in the same manner as for RF, Section 3.4.1.]

[2] **Model fit and cross-validation:** Lasso using the **glmnet()** function from the **glmnet** library was fit followed by cross-validation using **cv.glmnet()** to proceed with tuning of the lambda parameter. Conventionally, the optimal lambda was chosen to be within one standard deviation of the lowest data point in the graph representing the lambda dependency on GLM deviance. Predictions were calculated by the common function **predict()** of the **stats** package using the best lambda.

[3] **Model calibration curve**: [Executed in the same manner as for RF, Section 3.4.1.]

[4] **AUC and ROC:** [Executed in the same manner as for RF, Section 3.4.1.]

[5] **Variable importance**: The coefficient estimates were obtained using the **coef()** function applied on the cross-validated model. Implementation of calculations of p-values was initiated using the **fixedLassoInf()** function of the **selectiveInference** package, yet not finalized. The Discussion chapter elaborates further on the reasons.

3.5. Choices for the Selection of Strong Confounders

The selection of strong confounders in conjunction with RQ#2 in orchestrated in accordance with this methodology:

[1] First, the variable importance results are consolidated from all 4 models (RF, tuned RF, LREG, and lasso) for all three Cases in this manner:

- For RF models, the variable of high importance is among the top 7 either based on Mean Decrease in Accuracy or Mean Decrease in Gini index.
- For LREG models, high importance pertains to statistical significance accompanied with these codes depending on the level of significance (0.001, 0.01, 0.05, and 0.1). The information about the statistical significance is already contained in the R code output after running the **summary()** function (Section 3.4.3.).
- A Wald test (applicable to LREG models only) is performed for categorical variables (**Hospital**, **Surgery_Year** and **Prosthesis_Type**) to evaluate the importance of the variable, not a single category.
- For lasso, the variable of high importance acquires a non-zero coefficient. (In the case of a categorical variable, at least one category acquires a non-zero coefficient.)

[2] A next step involves a consolidation of resulting overlaps of high variable importance among RF, LREG and lasso models. This summary represents the supervised learning perspective.

[3] Validation of the supervised learning perspective with the clinical insight from a content expert is eventually a crucial determinant for finalizing the selection of the strong confounding variables.

Recalling Figure 3-11, the resulting set of strong confounders is a subset of the input variables for the RBC model (v_1, v_2, ..., v_{32}) because these variables are also input variables for the COM models.

3.6. Mediation Model Setup and Mediation Analysis Execution

In Figure 3-14, the setup of mediation analysis is shown. The effect estimates are denoted appropriately according to the mediation model in Mascha et al. (2013). Table 3-8 provides specifications of the components in this model. All key components (exposure, mediator, and the outcome variables) are binary. The strong confounders will be determined in RQ#2.

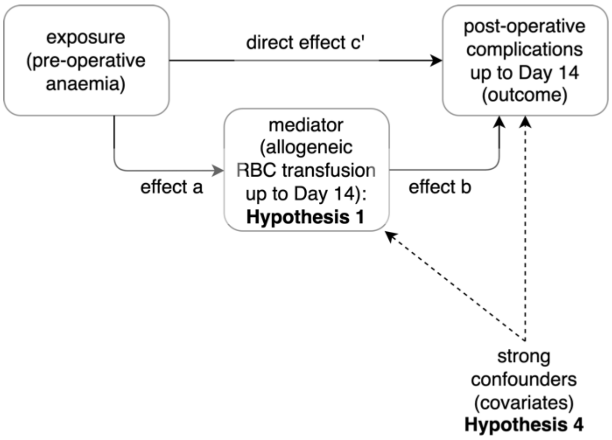

exposure (pre-operative anaemia)
direct effect c'
post-operative complications up to Day 14 (outcome)
mediator (allogeneic RBC transfusion up to Day 14): **Hypothesis 1**
effect a
effect b
strong confounders (covariates) **Hypothesis 4**

Figure 3-14: Setup and components of the mediation model.

Table 3-8: Specifications of the components in the mediation model.

Model component	Variable	Variable name	Format of the variable
Exposure	Pre-operative anaemia	Anaemia_Pre_Op	binary (1 = yes, 0 = no)
Mediator	Allogeneic RBC transfusion up to Day 14	RBC_Transfusion	binary (1 = yes, 0 = no)
Outcome	Post-operative complication up to Day 14	COM$_{PES}$ (pessimistic scenario) COM$_{OPT}$ (optimistic scenario)	binary (1 = yes, 0 = no)
Strong confounders (covariates)	[strong confounders determined in RQ#2, Section 4.2.1.]		

The intended mediation model is coupled with these assumptions:

[1] Unmeasured confounders of the exposure-mediator effect.
[2] Unmeasured confounders of the exposure-outcome effect.
[3] Unmeasured confounders of the mediator-outcome effect.
[4] The absence of the mediator-outcome confounders that are affected by the exposure.

Mediation analysis was conducted for the pessimistic and optimistic scenarios (the PES and OPT subscripts, respectively). Appendix E contains the R code of the implemented mediation analysis.

[1] **Reformatting**: For substep [5], it was necessary to disable factorization of **Anaemia_Pre_Op** and **RBC_Transfusion** by reformatting the variables to the integer class.

[2] **Model fit**: Using the conventional **glm()** function from the **stats** package, three logistic regression models (model.m, model.y$_{PES}$, and model.y$_{OPT}$) were fit on the entire dataset (sample size 2426 records). Input and dependent variables were assigned according to Table 3-9.

Table 3-9: Logistic regression models for the mediation model.

Model	Dependent variable	Input variables
model.m	RBC_Transfusion	Anaemia_Pre_Op (a-path), strong confounders
model.y$_{PES}$	COM$_{PES}$	Anaemia_Pre_Op, (c'-path), RBC_Transfusion (b-path), strong confounders
model.y$_{OPT}$	COM$_{OPT}$	

[3] **Variable importance**: Coefficients, standard errors, z-values and p-values were available already by running **glm()**, or **summary(model.m)** and **summary(model.y)**. Odds ratios (as well as p-values) were displayed by executing the function **odds.ratio()** from the **questionr** package.

[4] **Effect estimates on the a-, b- and c'-paths**: In model.m, the effect estimate for **Anaemia_Pre_Op** is equivalent to the a-path. In model.y, the effect estimates for **Anaemia_Pre_Op** and **RBC_Transfusion** are equivalent to the c'-path and b-path, respectively. (The effects of these components are denoted in this manner in Figure 3-14 above.) Conventionally, corresponding p-values determine statistical significance.

[5] **Evaluation of mediation (Quasi-Bayesian Confidence Intervals):** Mediation was evaluated using the function **mediate()** from the package **mediation**. The arguments were set as follows: **model.m** = model.m, **model.y** = model.y$_{PES}$ or model.y$_{OPT}$ (scenario-specific), **treat** = **Anaemia_Pre_Op**, **mediator** = **RBC_Transfusion**, **covariates** = listing of strong confounders per RQ#2. The argument **boot** was disabled (**boot = FALSE**).

[6] Evaluation of mediation (Nonparametric Bootstrap Confidence Intervals with the Percentile): The function $\mathrm{mediate()}$ was executed per substep [5] while the argument boot was enabled ($\mathrm{boot} = \mathrm{TRUE}$). The number of simulations (the sims argument) was set to the default of 1000.

[7] Desired effect estimates: The $\mathrm{mediate()}$ function in its output provides these effect estimates: average causal mediation effects (ACME, equivalent to the average indirect effect), average proportion mediated, average direct effects (ADE), and total effects. Mediation will be evaluated using the ACME.

Chapter 4 | Results

The Results chapter is structured as follows: Section 4.1. Exploratory Data Analysis (RQ#1), Section 4.2. Prediction and Inference Results composed of Section 4.2.1. The Strong Confounders (RQ#2), and Section 4.2.2. Statistical Importance of RBC Transfusion (RQ#3), Section 4.3. Mediation Analysis Results (RQ#4), and Section 4.4. Summary of Findings. For the reader to be smoothly guided through the series of results, in Table 4-1 we recall the *target patient outcomes* (equivalent to Cases) for all four research questions (RQs). The scenarios established after imputing missing data and after treating misclassification cases were shown earlier in Figure 3-11.

Table 4-1: Target patient outcomes and corresponding Cases (Table 3-1 adjusted upon data preparation).

Context	Case	Format of the dependent variable	Target patient outcome description with its boundaries	
RQ#1, RQ#2, RQ#4	RBC	binary (1 = yes, 0 = no)	The allogeneic RBC transfusion up to Day 14 (if any) administered first in time to the patient. Can be intra- or post-operative.	
RQ#1, RQ#2, RQ#3, RQ#4	COM_{PES}	binary (1 = yes, 0 = no)	Post-operative complication up to Day 14 (if any) after RBC transfusion (if any).	Pessimistic scenario.
	COM_{OPT}			Optimistic scenario.
RQ#1	LOS	numeric (days)	Hospital length-of-stay whose timespan may exceed Day 14 after surgery for some patients.	

4.1. Univariate Testing and Exploratory Data Analysis

Exploratory Data Analysis (EDA) served to respond to RQ#1: *How do the target patient outcomes vary for the key patient subgroups*? The EDA yielded descriptive statistics and insights for the desired ***key patient subgroups*** consolidated in Table 4-2 below. For explanation:

- Statistical significance is accompanied with these codes depending on the **level of significance**: 0 `***` 0.001 `**` 0.01 `*` 0.05 `.` 0.1. (For example, variable importance of the statistical significance level of 0.001 received three stars, `***`.)
- **Weak** correlation represents an overlap between the interquartile (IQR) ranges for numeric input variables, or the LOS variable (of the numeric format).
- Case LOS was analyzed without imputing (17) missing length-of-stay data values; thus, the patient records, for which the hospital LOS value was missing, are not reflected in the EDA.

Table 4-2: Descriptive statistics and EDA results: Insights on associations among selected input variables (key patient subgroups) and target patient outcomes (Cases). (*significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1;* for example, the statistical significance level of 0.001 received three stars, `***`)

Variable category	Key patient subgroup		Variable name	Variable type	Case COM$_{PES}$		Case COM$_{OPT}$		Case RBC		Case LOS	
					Significant association	p-value, or correlation	Significant association	p-value, or correlation	Significant association	p-value, or correlation	Significant association	correlation
Patient characteristics and pre-operative data	Participating hospitals (Hospital1; Hospital2; Hospital3; Hospital4)		Hospital	categorical: nominal	✓***	<0.001	✓***	<0.001	✓***	<0.001		weak
	Type of surgery (total hip; or total knee replacement)		Hip_or_Knee_1	binary		0.107		0.265	✓***	<0.001		weak
	Type of surgery (primary; or revision)		Primary_or_Revision_1	binary		0.162		0.330	✓***	0.002		weak
	Pre-operative anaemia (yes; no)		Anaemia_Pre_Op	binary	✓***	0.002	✓.	0.052	✓***	<0.001		weak
	EPO therapy (yes; no)		EPO	binary	✓.	0.098		0.147		0.726		weak
Intra-operative data	Cell saver device (hip replacement only) as intra-operative autologous reinfusion	Cell saver device (yes; no)	Cell_Saver	binary	✓.	0.100		0.150		0.427		weak
		Collection (mL), subset >0 mL only	Cell_Saver_Collection	numeric: integer		weak		weak		weak		weak positive
		Reinfusion (mL), subset >0 mL only	Cell_Saver_Reinfusion	numeric: integer		weak		weak		weak		weak positive
	Total blood loss during surgery (mL), subset >0 mL only, applies to hip replacement patients only		Total_Blood_Loss_during_Surgery	numeric: integer		weak		weak		weak		weak positive
Intra- or post-operative data	RBC transfusion (yes; no)		RBC_Transfusion	binary	✓***	<0.001	✓***	<0.001	n/a (infeasible)			weak

All aggregate results and visualizations that represent the EDA are available in Appendix A. Selected visualizations reflecting statistically significant differences among stratified patient subgroups based on RBC_Transfusion and Anaemia_Pre_Op are presented on the next pages.

In summary, to respond to RQ#1, the *target patient outcomes* significantly vary only for a few *key patient subgroups* from among all 10 investigated subgroups. In particular:

[1] Differences of the significance level of 0.001 (p-values below 0.001) were observed for categorical input variables among these *key patient subgroups*:
   - For Case COM_{PES}: RBC_Transfusion, Hospital, Anaemia_Pre_Op;
   - For Case COM_{OPT}: RBC_Transfusion, Hospital; and
   - For Case RBC: Hospital, Anaemia_Pre_Op, type of surgery (Hip_or_Knee_1), type of surgery (primary or revision expressed as Primary_or_Revision_1).
   - Interestingly, RBC_Transfusion displays significant associations with post-operative complications up to Day 14 for both pessimistic and optimistic scenarios.
   - Next to it, patient groups stratified based on participating hospitals (Hospital) showed significant differences among all Cases.

[2] Differences of the significance level of 0.1 (p-values below 0.1) were found for categorical input variables of these *key patient subgroups*:
   - For Case COM_{PES}: EPO, Cell_Saver;
   - For Case COM_{OPT}: Anaemia_Pre_Op; and
   - For Case RBC: none.
   - Interestingly, patient groups stratified based on pre-operative anaemia (Anaemia_Pre_Op) eventually showed significant differences among all Cases referring to the finding [1] above.

[3] For all numeric input variables in Cases COM_{PES}, COM_{OPT} and RBC, the IQR ranges overlap among all stratifications. Thus, these correlations were found as weak. All these distributions exhibit a right-skewed characteristic leading to divergences between respective medians and means.

[4] For all numeric input variables in Case LOS, weak positive correlations were detected with respect to the Total_Blood_Loss_during_Surgery (numeric format), and the autologous blood collection and reinfusion in conjunction with the cell saver intervention (Cell_Saver_Collection, and Cell_Saver_Reinfusion, respectively). These trends are reflected in Appendix A in scatter plots where the slopes of the fitted curves clearly fall below 45 degrees.

To conclude RQ#1, it is worth noting three remarks that accompany Table 4-2:

- The blood loss variable (Total_Blood_Loss_during_Surgery) was inquired for EDA. Yet it was evaluated as an infeasible input variable for modelling because it is a trigger of intra-operative RBC transfusion. The chapter Recommendations on Future Work elaborates on the absence of this blood loss variable.
- A drainage device as a second autologous reinfusion device was available for stratification of distinct patient subgroups. The application of this device and corresponding measurements were done solely post-operatively (after some patients already underwent RBC or COM); hence, the drainage device as an input variable was deemed infeasible for modelling.
- RBC_Transfusion is the dependent variable for Case RBC: intra-/post-operative allogeneic RBC transfusion up to Day 14, thus, cannot serve as an input variable for Case RBC.

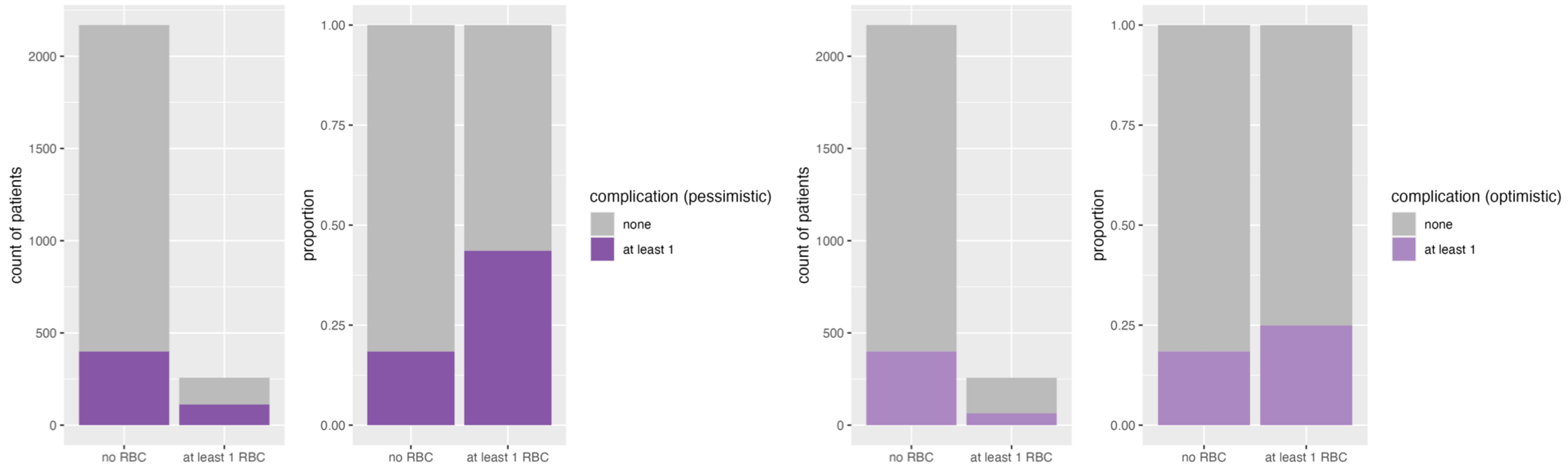


Figure 4-1: Stratification for the RBC transfusion: Case COM$_{\text{PES}}$ (left, p<0.001), COM$_{\text{OPT}}$ (right, p<0.001).

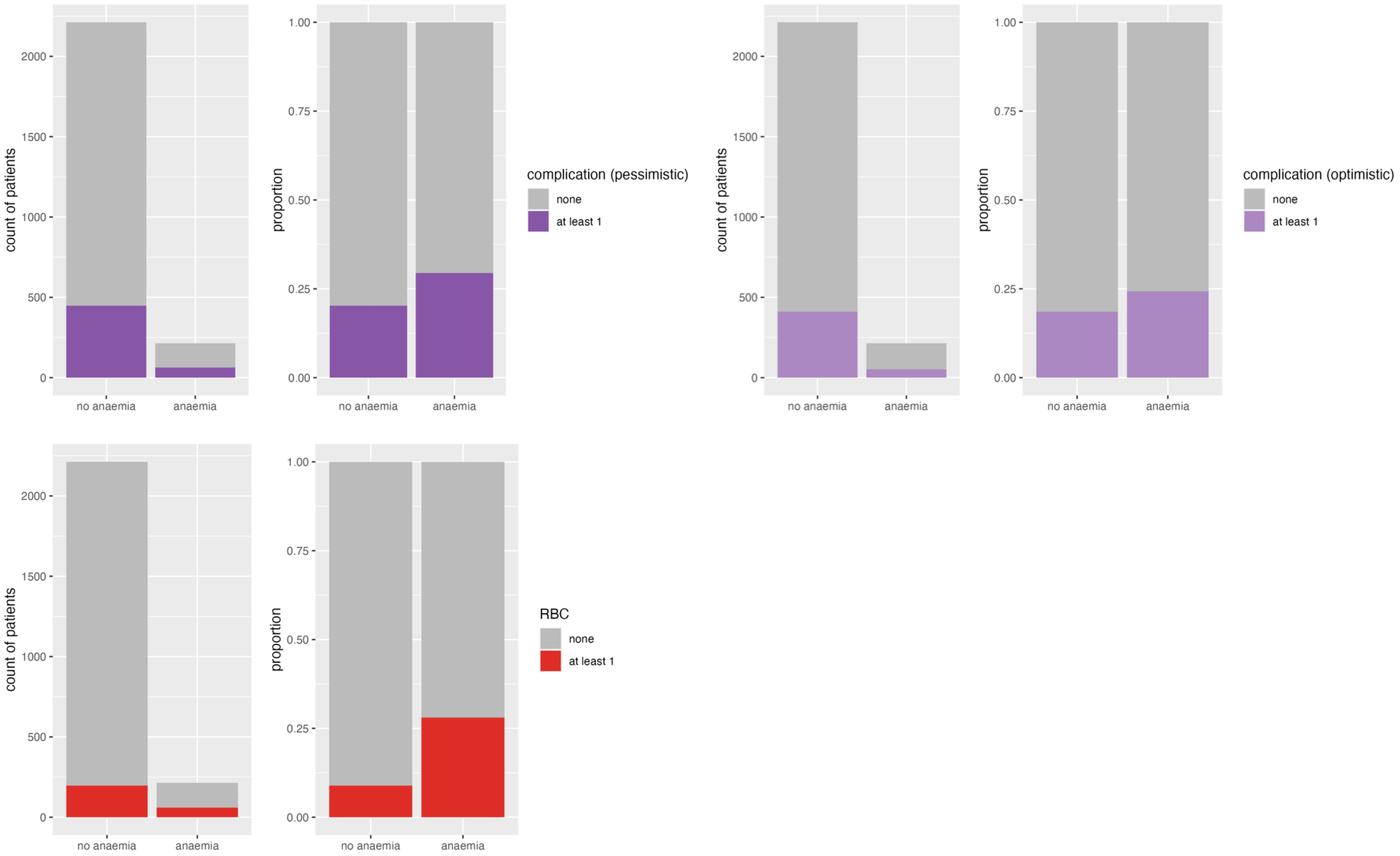


Figure 4-2: Stratification for pre-operative anaemia: Case COMPES (top left, p=0.002), COMOPT (top right, p=0.052), RBC (bottom left, p<0.001).

4.2. Prediction and Inference Results

The behaviour of the setting in elective orthopaedic surgery is described in terms of one non-parametric (RF) and two parametric (LREG and lasso) supervised learning models. As reported in Figure 4-3, all models (RF, LREG, and lasso) exhibit similar predictive abilities because there are no statistically significant differences in the AUC discrimination among all Cases. These findings set the frame of reference for the inference results in the upcoming Section 4.2.1. (RQ#2) and Section 4.2.2. (RQ#3).

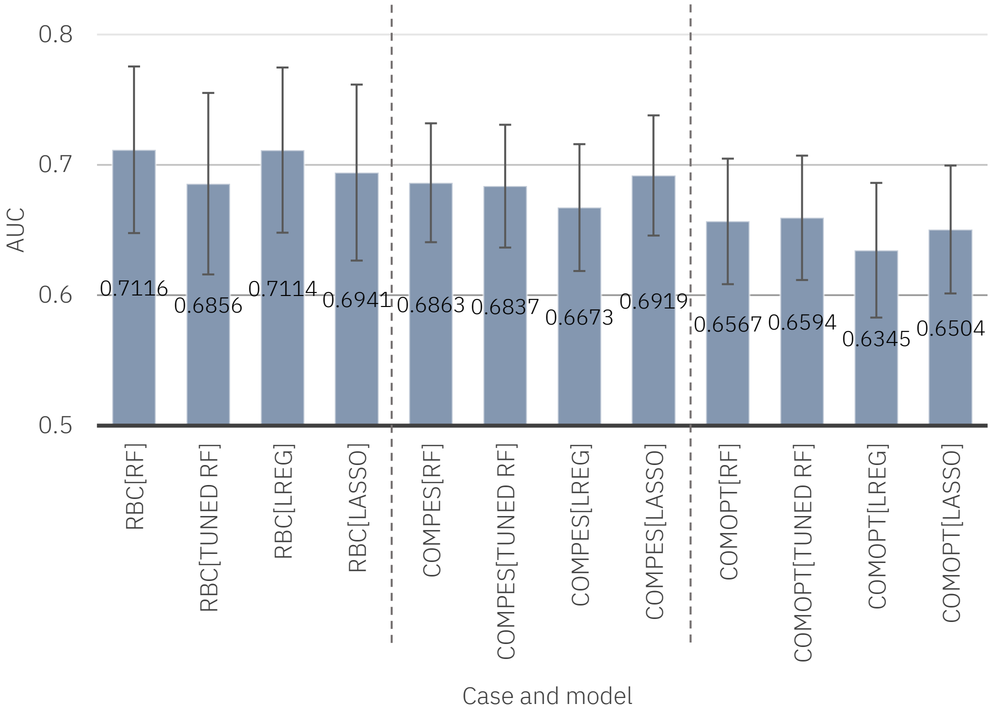


Figure 4-3: Model performance results: 95% CI's of the AUC discrimination measures for each Case and model.

Overall, the 95% CI's are wide for all models and Cases suggesting high natural variability in the dataset. The model performance of all models is evaluated as moderate. The RBC models have slightly (yet not significantly) better predictive power with the AUC's of 0.69-0.71 with 95% CI reaching from 0.62 to 0.78. And the COM models yield AUC's of 0.63-0.69 with 95% CI ranging from 0.60 to 0.74. Interestingly, model performance of lasso models slightly, insignificantly improved for the COM Cases, yet not for the RBC Case.

Table 4-3 reports on hyperparameter tuning results of RF models upon 5-fold cross-validation and random search. The implications of the tuned **mtry** reaching 1 for Case COM$_{\text{PES}}$. are reviewed in the Discussion chapter. Next, Table 4-4 provides regularization results for lasso models.

Table 4-3: Hyperparameter tuning results for RF models upon 5-fold cross-validation and random search.

Case	default parameter **mtry**	parameter **mtry** after tuning
RBC	6	2
COM$_{\text{PES}}$	7	1
COM$_{\text{OPT}}$	7	4

Table 4-4: Regularization results for lasso models.

Case	Regularization parameter lambda	Number of variables with a non-zero coefficient	Number of input variables
RBC	0.0196	8	31
COM$_{\text{PES}}$	0.0263	12	38

Case	Regularization parameter lambda	Number of variables with a non-zero coefficient	Number of input variables
COM$_{OPT}$	0.0259	11	38

To enhance transparency, Appendix B provides additional model validation and model performance results, such as the ROC curves, calibration curves, and for RF specifically, the out-of-bag (OOB) error progression. Plus, the variability of the accompanying performance measures, namely, sensitivity, specificity and accuracy relative to cutoff levels, is also demonstrated graphically in Appendix B. Although many figures do not directly contribute to answering RQ#2, they represent auxiliary modelling outputs to describe the model behaviour and to demonstrate transparency.

4.2.1. The Strong Confounders: Case RBC and Case COM

This segment serves to respond to RQ#2: *What variables are the strong confounding predictors for RBC transfusion up to Day 14, and a post-operative complication up to Day 14 in elective orthopaedic surgery?*

First, we start with Table 4-5 designated to the consolidation of variable importance results for the series of models (random forest (RF), logistic regression (LREG), and lasso) in relation to all Cases (RBC, COM$_{PES}$, and COM$_{OPT}$). Checkmarks with dark green background representing high importance were given according to the criteria defined in Section 3.5. Validation using the clinical insight is marked bright green.

Table 4-5: Variable importance results per Case and model for each input variable, accompanied by validation using the clinical insight. (*significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1;* for example, variable importance with the statistical significance level of 0.001 received three stars, `***`)

Variable category	Variable name	Plausible strong confounder per clinical insight	Variable importance per Case and model											
			RBC[RF]	RBC[tuned RF]	RBC[LREG]	RBC[lasso]	COM$_{PES}$[RF]	COM$_{PES}$[tuned RF]	COM$_{PES}$[LREG]	COM$_{PES}$[lasso]	COM$_{OPT}$[RF]	COM$_{OPT}$[tuned RF]	COM$_{OPT}$[LREG]	COM$_{OPT}$[lasso]
Patient characteristics and pre-operative data	Hospital	✓	✓	✓	✓***	✓	✓	✓	✓**	✓	✓	✓	✓***	✓
	Age	✓	✓	✓	✓**	✓	✓	✓	✓**	✓	✓	✓	✓**	✓
	Gender	✓		✓	✓***	✓			✓*		✓		✓.	
	Hip_or_Knee_1	✓	✓	✓	✓***	✓			✓*				✓**	
	Primary_or_Revision_1	✓												
	Hip_or_Knee_2				✓*									
	Osteoarthritis													
	Cardiovascular_Disease	✓	✓	✓				✓			✓	✓		
	CVA				✓.						✓			
	COPD												✓.	
	Diabetes_Mellitus	✓							✓.					
	Rheumatoid_Arthritis													
	Increased_Risk_Group	✓			✓*	✓			✓**	✓			✓**	✓
	Corticosteroids												✓.	
	NSAIDs		✓						✓**	✓			✓**	✓
	Anticoagulation													
	Antibiotics													
	Insulin													
	Antihypertensiva						✓	✓	✓*		✓	✓	✓*	
	Cardiac_Medication				✓*		✓	✓	✓.	✓	✓		✓*	✓

Variable category	Variable name	Plausible strong confounder per clinical insight	Variable importance per Case and model											
			RBC[RF]	RBC[tuned RF]	RBC[LREG]	RBC[lasso]	COM$_{PES}$[RF]	COM$_{PES}$[tuned RF]	COM$_{PES}$[LREG]	COM$_{PES}$[lasso]	COM$_{OPT}$[RF]	COM$_{OPT}$[tuned RF]	COM$_{OPT}$[LREG]	COM$_{OPT}$[lasso]
	Pulmonary_Medication													
	Smoking													
	EPO	✓	✓	✓	✓**									
	Anaemia_Pre_Op		✓	✓	✓***	✓								
	Surgery_Year	✓	✓	✓		✓	✓	✓	✓.	✓	✓	✓	✓.	✓
Intra-operative data	Surgery_Duration	n/a*					✓	✓		✓	✓	✓	✓.	✓
	Prosthesis_Type		✓	✓	✓**	✓						✓		
	Minimally_Invasive_ in_case_of_Total_Hip_Prosthesis								✓*			✓	✓*	
	Temperature_Drop_Prevention				✓.		✓			✓			✓.	✓
	Anticoagulant_Standard			✓			✓	✓		✓				✓
	Antibiotic_Prophylaxis_Standard			✓				✓						
	Antifibrinolytic_Blood_Loss_ Lowering_Medication													
	Antifibrinolytic_Cyclokapron								n/a†				n/a†	
	Colloids	n/a†					✓				✓	✓		
	Crystalloids						✓	✓			✓	✓		
	Cell_Saver													
	Other_Transfusions								n/a†				n/a†	
	FFP													
	RBC_Transfusion						✓	✓	✓***	✓				

† Not an input variable.

Highlights of the variable importance findings per Table 4-5 are summarized below.

[1] **Hospital** and **Age** are found to be among the most important variables for **all models in all Cases**. In the LREG models, the statistical significance is 0.01 or even higher, 0.001.

[2] **Surgery_Year** is also a highly important variable for all models in all Cases except for RBC[LREG]. In the COM[LREG] models, the statistical significance is 0.1.

[3] **Gender**, **Hip_or_Knee_1**, **Increased_Risk_Group**, and **Cardiac_Medication** are good confounding candidates because they are statistically significant for **all LREG models** (for both scenarios). This corresponds to statistical significance of at least 0.1.

[4] **Cardiovascular_Disease** is deemed highly important for **all RF models**, but not all LREG or lasso models.

[5] **Temperature_Drop_Prevention** shows statistical significance (0.1) only for the optimistic scenario based on the LREG models.

[6] **Anticoagulant_Standard** and **Antibiotic_Prophylaxis_Standard** represent highly important variables and confounding candidates only for the pessimistic scenario.

[7] On the contrary, **Anaemia_Pre_Op**, **EPO**, and **Prosthesis_Type** display high (significant) importance in the RBC model (0.01 or even higher, 0.001 for the LREG model), yet, not in the COM models. Thus, strong confounding is not claimed for the **EPO** and **Prosthesis_Type** variables. And the remaining set of 18 variables are considered rather weak confounding candidates because they either display high importance for the COM model only, or are not highly important for either model.

Table 4-6, a compact format of the previous table, follows with a list of strong confounders for RBC transfusion up to Day 14, and post-operative complication up to Day 14 in elective orthopaedic surgery. In response to RQ#2, a purely supervised learning perspective (dark green in Table 4-6) yields a variety of strong confounders. The findings are reported based on strong importance in either model. The findings per clinical insight are highlighted then in bright green. Several inconsistencies are observed between these approaches with further elaboration below.

Table 4-6: Strong confounders for pessimistic and optimistic scenarios in accordance with the supervised learning (dark green) and clinical perspectives (bright green).

Variable category	Variable name	Plausible strong confounder per clinical insight	Strong confounder (pessimistic)			Strong confounder (optimistic)		
			Per RF models only	Per LREG models only	Per lasso models only	Per RF models only	Per LREG models only	Per lasso models only
Patient characteristics and pre-operative data	Hospital	✓	✓	✓	✓	✓	✓	✓
	Age	✓	✓	✓	✓	✓	✓	✓
	Gender	✓		✓		✓	✓	
	Hip_or_Knee_1	✓		✓			✓	
	Primary_or_Revision_1	✓						
	Cardiovascular_Disease	✓	✓			✓		
	Diabetes_Mellitus	✓						
	Increased_Risk_Group	✓		✓	✓		✓	✓
	Cardiac_Medication			✓			✓	
	EPO	✓						
	Surgery_Year	✓	✓		✓	✓		✓
Intra-operative data	Prosthesis_Type				✓			
	Temperature_Drop_Prevention					✓		
	Anticoagulant_Standard		✓					
	Antibiotic_Prophylaxis_Standard		✓					

Eventually, upon incorporating clinical insight from the content expert to the above findings, 10 plausible strong confounders were selected. They will serve as covariates for the mediation model in RQ#4. These confounders apply both to pessimistic and optimistic scenarios. Three confounders were not anticipated if this final selection of strong confounders is compared to the results from the supervised learning perspective. Overall, the selection of strong confounders was done from a total of 31 input variables both for RBC and COM models (this count excludes Anaemia_Pre_Op).

• Hospital	(anticipated)
• Age	(anticipated)
• Gender	(anticipated)
• Hip_or_Knee_1	(anticipated)
• Primary_or_Revision_1	(not anticipated)
• Cardiovascular_Disease	(anticipated)
• Diabetes_Mellitus	(not anticipated)
• Increased_Risk_Group	(anticipated)
• EPO	(not anticipated)
• Surgery_Year	(anticipated)

Appendix C provides supplemental output with further details about variable importance (in terms of the Mean Decrease in Accuracy or the Mean Decrease in Gini index for RF, and in terms of coefficients and odds ratios for LREG or lasso).

4.2.2. Statistical Importance of RBC Transfusion: Case COM

This segment is dedicated to RQ#3: *What is the statistical importance of RBC transfusion up to Day 14 acting as the predictor for the occurrence of a post-operative complication up to Day 14?* Besides results per Table 4-6, further granularity on the statistical importance results particularly for **RBC_Transfusion** is provided in Table 4-7 (LREG and lasso) and Table 4-8 (RF) where the corresponding odds ratios, coefficients, or variable importance ranking are reported.

LREG reflects **RBC_Transfusion** as a statistically significant predictor in Case COM_{PES} (coefficient 4.979, odds ratio 3.891, p-value <0.001). Yet, Case COM_{OPT} yields no statistical significance of **RBC_Transfusion**. Results of the lasso models report the coefficients of 0.650 and 0.000 for Cases COM_{PES} and COM_{OPT}, respectively.

Per RF models and the pessimistic scenario, the odds ratios of **RBC_Transfusion** are 1.891 and 18.38 for the COM_{PES}[RF] and COM_{PES}[tuned RF], respectively. In terms of the ranking of the Mean Decrease in Accuracy, **RBC_Transfusion** scores 1st to 2nd. In terms of the Mean Decrease in Gini index, RBC transfusion is positioned in the 5th and 7th place. In the optimistic scenario, the odds ratios are 1.225 and 1.271 for the COM_{OPT}[RF] and COM_{OPT}[tuned RF], respectively. The ranking of the Mean Decrease in Accuracy is 26th and 30th, and the ranking of the Mean Decrease in Gini index is 21st and 18th.

Table 4-7: Statistical importance of RBC transfusion in terms of coefficients, odds ratios and p-values: LREG and lasso models.

Scenario	Case	Coefficient	Odds ratio (RBC / no RBC)	p-value
pessimistic	COM_{PES}[LREG]	4.979	3.891	<0.001
	COM_{PES}[lasso]	0.650	[not implemented]	
optimistic	COM_{OPT}[LREG]	-0.254	1.443	0.799
	COM_{OPT}[lasso]	0.000	[not implemented]	

Table 4-8: Statistical importance of RBC transfusion in terms of odds ratios, odds and ranking (from among 41 input variables): RF and tuned RF models.

Scenario	Case	Odds ratio	Odds	log(Odds)	MeanDecrease Accuracy ranking	MeanDecrease Gini ranking
pessimistic	COM_{PES}[RF]	1.891	2.814, 1.488	0.738, 0.598	2	7
	COM_{PES} [tuned RF]	18.38	84.98, 4.624	0.988, 0.822	1	5
optimistic	COM_{OPT}[RF]	1.225	2.927, 2.389	0.745, 0.705	26	21
	COM_{OPT} [tuned RF]	1.271	3.265, 2.569	0.766, 0.720	30	18

4.3. Mediation Analysis Results

As a capstone in this project, we respond to RQ#4: *What role does RBC transfusion up to Day 14 play* ***in the relationship*** *between pre-operative anaemia and the occurrence of a post-operative complication up to Day 14?*

First, the results of the effect estimates in terms of coefficients and the odds ratios for each component of the mediation analysis model are consolidated in Table 4-9. Accounting for the pessimistic and optimistic scenarios, the notation of these results complements the notation established in Figure 3-14, Section 3.6. Statistically significant effects (with a significance level of 0.001) were detected on the a-path and, for the pessimistic scenario, on the b-path. The other effects were deemed statistically insignificant.

Table 4-9: Effect estimates in terms of the coefficients and odds ratios for each path in the mediation model.

Scenario	Variable	Model component	Effect estimate notation	Coefficient	Odds ratio	p-value
-	Pre-operative anaemia	Exposure	a	1.685	5.391	<0.001
pessimistic	Pre-operative anaemia	Exposure	c'_{PES}	-0.016	0.984	0.930
	RBC transfusion	Mediator	b_{PES}	1.078	2.940	<0.001
optimistic	Pre-operative anaemia	Exposure	c'_{OPT}	0.033	1.033	0.864
	RBC transfusion	Mediator	b_{OPT}	0.141	1.152	0.403

The 95% CI's and further details upon the model fit are available in Appendix D.

Next, Table 4-10 reveals the accompanying mediation results including the average causal mediation effects (ACME), average direct effects (ADE), and total effects (TE) for the Quasi-Bayesian Confidence Intervals method. (The results of the second method, Nonparametric Bootstrap Confidence Intervals with the Percentile with 1000 simulations, are presented in Appendix D. This second method yields very similar results as anticipated.)

Table 4-10: Mediation analysis results (the Quasi-Bayesian Confidence Intervals method).

	Average causal mediation effect, ACME (indirect effect)		Average direct effect, ADE		Total effect, TE = ACME + ADE	
Scenario	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
pessimistic	0.0445	[0.0268; 0.0700]	-0.0013	[-0.054; 0.06]	0.0432	[-0.011; 0.11]
optimistic	0.0050	[-0.0072; 0.0200]	0.0046	[-0.046; 0.06]	0.0096	[-0.038; -0.07]

We conclude opposing results about mediation for the two scenarios:

[1] For the pessimistic scenario, the RBC transfusion **mediates** the relationship between pre-operative anaemia and the post-operative complications up to Day 14. This mediated relationship is enumerated with the ACME of 0.0445 with the 95% CI of [0.0268; 0.0700]. The accompanying ADE is -0.0013 (95% CI of [-0.054; 0.06]) and the TE is 0.0432 (95% CI of [-0.011; 0.11]).

[2] For the optimistic scenario, the RBC transfusion **does not mediate** the relationship between pre-operative anaemia and the post-operative complications up to Day 14. The ACME is 0.0050 with the 95% CI of [-0.0072; 0.0200]. The accompanying ADE is 0.0046 (95% CI of [-0.046; 0.06]) and the TE is 0.0096 (95% CI of [-0.038; -0.07]).

Furthermore, it is observed for the pessimistic scenario that the direct effects and indirect effects (ACME) have different signs. Thus, the mediation model is inconsistent (the proportion mediated available in Appendix D does not have a meaningful interpretation).

4.4. Summary of Findings

Clinical researchers including transfusion professionals desire simple models and ease in their interpretability despite the high complexity of transfusion medicine. In pursuit of determining the role of RBC transfusion relative to post-operative complications and in pursuit of describing a multi-centre setting in elective orthopaedic surgery in a simplified way, selected variable selection methods (subset selection, and lasso) were conducted during supervised learning model development. Variable importance was then captured, and mediation analysis was performed.

In the parametric (LREG and lasso) and non-parametric, 'black box' (RF) modelling, a reduction of the number of variables was achieved, particularly, from 32 input variables (default) to 8 variables per lasso for the RBC model. For the COM models, the reduction is counted from 41 input variables to 12 and 11 variables per lasso (pessimistic and optimistic scenarios, respectively). Although lasso models yielded slightly, insignificantly better performance for the COM models, the performance of all implemented models is in fact statistically comparable in terms of their AUC because the AUC's 95% CI's overlap. Overall, the model performance was evaluated as moderate or poor given that all CI's range approximately between 0.58 to 0.78. It was observed that model calibration plots reveal inconsistencies which may be caused due to nonlinearity of the dataset.

In summary, we respond to **Hypothesis 4**: There are strong confounding variables associated *both* with:

- Allogeneic RBC transfusion (the mediator), and
- Post-operative complications (the patient outcomes).

10 strong confounding variables were eventually identified upon incorporating the clinical insight from the content expert in transfusion medicine – namely: hospital, age, gender, type of surgery (hip or knee, and primary vs revision), cardiovascular disease, diabetes mellitus, increased risk group, EPO, and the surgery year. The incorporation was done after a proposal of variables concluded to be of high importance per a supervised learning perspective as follows: Strong confounding among COM[RF] and RBC[RF] models is equivalent to the top 7 ranking in the variable importance for random forest models. The confounding was determined to be strong among COM[LREG] and RBC[LREG] models if at least 0.1 significance level was observed. Strong confounders based on lasso models had non-zero coefficients both for the COM[lasso] and RBC[lasso] models.

The COM[LREG] models already provided answers to the statistical significance of RBC transfusion. RBC transfusion was deemed statistically significant at the significance level of 0.001 in the pessimistic scenario. RBC transfusion was not a significant predictor in the optimistic scenario. These findings of the COM[LREG] models were consistent with the effects on the b-path seen in the mediation model. This auxiliary observation complements the current series of results.

A capstone is to respond to **Hypothesis 1**: RBC transfusion mediates the relationship between pre-operative anaemia and post-operative complications. Mediation analysis revealed that RBC transfusion mediates the relationship between pre-operative anaemia and post-operative complications up to Day 14 in the pessimistic scenario, yet not in the optimistic scenario. Mediation for the pessimistic scenario was enumerated in terms of the average causal mediation effect (ACME) of 0.0445 (with the 95% CI of [0.0268; 0.0700]) that is deemed statistically significant. The accompanying average direct effect (ADE) is -0.0013 (95% CI of [-0.054; 0.06]) and the total effect is 0.0432 (95% CI of [-0.011; 0.11]).

Chapter 5 | Discussion

The Discussion chapter walks the reader through the reflections on each research question. Section 5.1. revolves around the findings to RQ#1 and the comparison of these insights to RQ#2. Strengths and suggestions for improvements of supervised learning that are tied to RQ#2 and RQ#3 are present in Section 5.2. and 5.3., respectively. In Section 5.4., we discuss the opposing results of the mediation model per RQ#4.

5.1. Remarks on Statistical Significance of Input Variables Examined in Univariate Tests versus in a Multivariate Model

Different findings are observed when examining some variables in univariate tests (in silo using the chi-squared test and descriptive statistics per RQ#1 in Section 4.1.) versus in the presence of other covariates (RQ#2 in Section 4.2.1.). This phenomenon is detected for these six key input variables:

- Pre-operative anaemia was deemed significant after a chi-squared test yet low variable importance was observed based on all three types of supervised learning models in both COM Cases (both the pessimistic and optimistic scenarios).
- RBC transfusion was deemed significant for both COM Cases in univariate setting; however, none of the supervised learning models revealed statistical significance and high variable importance for Case COM_{OPT}.
- The type of surgery – hip or knee, was found insignificant in both COM Cases, yet logistic regression models reveal statistical significance for this variable.
- The type of surgery – primary or revision, was found significant in Case RBC, yet none of the supervised learning models say so.
- EPO is significant for Case COM_{PES} and Case RBC in univariate testing, yet the opposite is found per logistic regression.
- The cell saver variable is seen significant for Case COM_{PES} after univariate testing; however, this variable is not significant after exploring it using the supervised learning models.

The inconsistencies in statistical significance of an input variable suggest making careful considerations when analyzing input variables in silo (univariate setting) versus multivariate models. Results of univariate tests are not indicative of the logistic regression model outputs. Hence, the variable selection process for supervised learning models shall not depend entirely on the results of univariate tests.

Interestingly, the hospital variable exhibits statistical significance both in univariate testing and multivariate supervised learning models. This observation represents significant differences among the hospital in terms of the influence on administration of RBC transfusion and the occurrence of post-operative complications. This observation can be somewhat tied to the type of a hospital – academic (Hospital1) or non-academic (Hospital2,3,4). Nevertheless, in future studies it is more useful to stratify data into subsets so that the hospital variable is eventually deemed insignificant. We claim that this modelling choice may lead to extracting valuable evidence and insights about specific aspects of the hospital floor operations (by including adequately measured variables) rather than general information on a hospital level.

5.2. Remarks on Strengths of Supervised Learning Models

We argue that our research exhibits extensive efforts of developing three types of prediction models (random forest, logistic regression, and lasso) including model validation, calibration, a demonstration of variability of performance measures and variable importance. (The results of RQ#2 are found in Section 4.2.1., and Section 6.3. further elaborates on promising follow-up projects.) In Section 2.1., we indicated in a literature review that many existing prediction models involving blood transfusion as a dependent variable were subject to a careful assessment by Dhiman et al. (2023). In this very recent systematic review, Dhiman et al. (2023) depict various flaws and high risk of prediction bias of these predictive modelling studies in transfusion medicine. According to Dhiman et al. (2023), some publications on the prediction models have unreported validation procedures, and most validated models considerably violate handling of predictors, the variable selection process, the sample size considerations, or validation methods.

A variety of these issues in this research project was prevented. We further compare our efforts with Huang et al. (2018) and Rashiq et al. (2004) referenced in the above mentioned systematic review by Dhiman et al. (2023). These two studies can be compared with our research because the dependent variable was established also as intra-/post-operative RBC transfusion. Both studies performed internal validation as it is the case in our research. Huang et al. (2018) performed cross-validation, and Rashiq et al. (2004) used a train/test split. We used both of these methods. Huang et al. (2018) flourished from quite a large sample size of 15 187 patient records out of which 2867 patients (18.9%) received RBC transfusion. Rashiq et al. (2004) had only 884 patient records at hand out of which 239 (27%) patients who received RBC transfusion. Our sample (2426 records out of which 257 patients (10.6%) received RBC) was of a reasonably large sample size with minimal excluded patient records, and lies in between the two sample sizes indicated above.

Furthermore, Huang et al. (2018) and Rashiq et al. (2004) report the AUC of 0.84 (95% CI of [0.81; 0.87]) and 0.76 (unreported 95% CI), respectively, for their logistic regression models. The performance of our logistic regression models for Case RBC are significantly lower, 0.71 (95% CI of [0.65; 0.77]) than that reported by Huang et al. (2018). It is infeasible to claim a comparison with Rashiq et al. (2004) due to the unreported 95% CI. Then, Huang et al. (2018) report the AUC of 0.77 (95% CI of [0.74; 0.79]) for random forest. In our research, the 95% CI of AUC for both random forest models of Case RBC overlap. So we claim a comparable model performance here. Unlike in Huang et al. (2018), it was shown in our research that random forest does *not* significantly outperform logistic regression. The same finding is seen in the publication by Christodoulou et al. (2019) who conducted a systematic review to compare the performance of clinical prediction models, especially, with a focus on logistic regression models.

Clinical researchers (Dhiman et al., 2023) express flaws tied to handling of input variables in predictive modelling involving blood transfusion. It is important to note that our research offers a great extent of transparency into the variable selection process that may help relieve some of this skepticism.

To discuss the variable importance findings of the RBC models, we again account for Huang et al. (2018) and Rashiq et al. (2004). For example, Huang et al. (2018) detected statistical significance at the level of 0.001 in their logistic regression model for gender, pre-operative Hb, the length of surgery, the tranexamic acid use, the drain use, and intra-operative blood loss. Rashiq et al. (2004) identified statistical significance at the level of 0.001 in their logistic regression model for type of surgery (primary vs revision), the categorical Hb variable, and the categorical weight variable. We find an overlap to these studies with gender and type of surgery.

Finally, we argue that the prediction models involving post-operative complications (as the selected patient outcome, here Cases COM) offers an opportunity in patient outcomes research, patient-centred care and in PBM. Unfortunately, we do not report a comparison of our findings to a scientific publication especially due to the specificity of our established dependent variable that encompasses 15 different types of post-operative complications in elective orthopaedic surgery. Still, numerous literature sources involving patient outcomes research and blood transfusion in the surgical setting are available. For illustration, Bramley et al. (2021) present an umbrella review of systematic reviews concerning the risk factors for post-operative mental complication of delirium. From among the 10 publications related to trauma and orthopaedic surgery, they report that intra-operative blood transfusion appears among the risk factors. Future work may be accompanied with a thorough review of these and other publications (systematic reviews) to expand on this discussion topic and to further explore this field. There is room to reflect on a variety of patient outcome variables in many kinds of surgical settings.

5.3. Remarks on Limitations and Potential Improvements of Supervised Learning Models

It is imperative to state that this work is solely the first effort of supervised learning model development using the 'TOMaat' dataset. Hence, there is indeed room for many modelling improvements which can lead to underlining the strengths mentioned in the above section.

Multicollinearity was not entirely treated. This problem was detected by troubleshooting the implementation of inference for lasso leaving it with an unreported measure of uncertainty for variable importance (recalling Table 4-7). There are existing tools to obtain these model outputs, such as the **fixedLassoInf()** function of the **selectiveInference** package (R Documentation, 2023b). Collinearity causes reduction in accuracy of the effect estimates and rapid increase in standard errors (James et al., 2021). This means that the power of the hypothesis test (of detecting a non-zero coefficient) is reduced due to collinearity (James et al., 2021). To determine strong associations among categorical or binary input variables, collinearity can be spotted using the Spearman rank correlation coefficient (for ordinal) or the chi-square test (for nominal variables). Variance inflation factors may serve as an alternative.

Tuning and cross-validation offers room for further investigation. Random search settings led to limited results due to controlling the computation time. For example, recalling Table 4-8, the odds ratio for $\mathrm{COM}_{\mathrm{PES}}$[tuned RF] was quite high, 18.38. The random search did not yield the option to test **mtry** of 2. Instead, this time accuracy was optimized to yield **mtry** of 1. This means high variability was imposed on the model implying a high value of the odds ratio.

Uncertainty measures, such as p-values for RF models may complement future work. It is advised to investigate methods to calculate CI's of odds ratios and validate the odds using alternative functions or other 'black box' models.

This work presents many binary input variables, yet, it is worth noting that random forest tends to be biased towards binary variables in inference outputs (James et al., 2021). The inconsistent findings from the PDPs of the RF models (Appendix C) may pertain to this aspect, the high natural variability, or the nonlinear character of the dataset. Surprisingly, the PDPs favour the presence of **Anaemia_Pre_Op** as well as **RBC_Transfusion** in terms of the lower odds values for the binary category containing ones.

Instead, the binary category containing ones is expected to receive higher odds (as it is the finding of the LREG models). Unfortunately, these results oppose the clinical findings.

Additionally, the variable selection process on the raw data led to leaving unstructured data or many free text fields untapped. Free text fields potentially store valuable information that may enhance the model's predictive and inference ability. However, retrieving data from free text fields is undesirable because it is highly error-prone as well as time-consuming considering common data science tools and practices in workplaces. Future variable selection process on the 'TOMaat' dataset may involve more extensive efforts in grouping of variables based on keywords in free text fields.

Calibration curves (found in Appendix B) need further attention. Due to their offset, the prediction rates neither explicitly represent the probabilities for the occurrence of a post-operative complication (the COM models), nor the probabilities for being administered RBC transfusion (the RBC models). Thus, in this situation, clinicians are advised to continue considering prediction models for clinical use, yet, while choosing a cutoff level other than the intuitive default of 0.5. Appendix B further guides the reader in choosing the suitable cutoff and a possible series of model performance measures.

5.4. Remarks on the Opposing Outputs of the Mediation Model as an Implication of Data Missingness in the Post-hoc Study

RBC transfusion was not previously studied as a mediator between pre-operative anaemia and patient (surgical) outcomes. Literature for comparing the current results is absent (as demonstrated in Appendix E). This argument is supported by the clinical insight from the SME, dr. So-Osman. Numerous recommendations and tips for future work are proposed in the next chapter, Section 6.3.

Mediation analysis revealed that RBC transfusion mediates the relationship between pre-operative anaemia and post-operative complications up to Day 14 in the pessimistic scenario, yet not in the optimistic scenario. The opposing results of mediation analysis lead to a clear prompt for assessing digital maturity of hospitals with the intention to analyze new patient-level datasets. Or alternative modelling approaches may be explored (i.e. per point 1, segment Patient outcomes in Section 6.3.).

In this post-hoc project, massive data preparation efforts took place to establish dependent variables (post-operative complications up to Day 14, and intra-/post-operative RBC transfusion up to Day). To correctly detect the time sequence of the key events – whether RBC transfusion (administered first to the patient) occurred before or after the complication, the RBC transfusion dates were retrieved from a free text field. And not to compromise on the data missingness of the complication dates by making inappropriate assumptions that would alter the problem setting, two COM scenarios for the binary complication variable were established by accounting for extreme cases (0 for optimistic, 1 for pessimistic). Moving forward, instead of post-hoc projects, it would be more favourable to frame project incentives and to set expectations of data quality right from the start.

Monitoring of patient outcomes and drawing insights from patient-level datasets has recently become a priority and urgency for transfusion medicine and PBM. For the future, it is advised to foster the digital maturity of hospitals and other healthcare establishments that are involved in digital transformation and in developing robust data collection strategies. The risks are tied to infeasible missingness or unstructured data. This advancement (monitoring shifts, timestamps, logs) can then offer a stronger platform for analyzing patient outcomes downstream. Otherwise, inference among key variables are hardly deduced, and data-driven decision-making and solutions tied to transfusion dependency and patient-centredness are very limited. Generally, in healthcare settings, the digital maturity may include database design, standardization of data collection practices, and even extensive communication and discussion with the vendors of electronic health record (EHR) systems. A scale-up of data collection strategies and enhancements of digital maturity may potentially also go hand in hand with the new ***PBM Implementation Guidelines*** are currently under development by WHO.

Chapter 6 | Recommendations on Future Work

We are convinced that the methodology of this study suggests quite an innovative roadmap for data analysis involving patient outcomes across the transfusion medicine and PBM landscapes. Section 6.1. discusses generalizability and the degree of innovation. Section 6.2. presents key takeaway messages on how to treat (RBC) transfusion as a dependent variable in modelling. And the content of Section 6.3. serves to touch upon potential future work involving patient outcomes research and transfusion data using the 'TOMaat' dataset or other datasets.

6.1. Generalizability and the Degree of Innovation

The research output has tremendous potential in terms of the degree of innovation. According to dr. So-Osman, the Unit Transfusion Medicine at Sanquin Blood Bank: "The results are going to be very relevant for the transfusion medicine and PBM communities. Transfusion professionals throughout the world will benefit from the results of this project, and Sanquin also (of course)." A greater level of detail (greater granularity) could empower both the patients and the current clinical practice. We argue that the methodologies applied in this study may find applications in other (peri-operative) clinical settings in the future. For example, Blood and Beyond (2021; 2020) reports that most RBCs are rather used in medical indications (67%) than surgical (33%). Hence, there is a great potential to target medical areas, such as solid cancers, gastrointestinal disease, kidney disease, cardiovascular disease, or various blood diseases (Blood and Beyond, 2021). Yet, robust data acquisition on a desired patient group is indeed a key prerequisite for fruitful data analysis and modelling.

New insights could help shed light on the dynamics among many variables from the (anaemic) patient medical record and the role of RBC transfusions in the peri-operative setting. The potential new findings may even help further facilitate ongoing debates about decision-making in healthcare among global multi-disciplinary teams, transfusion medicine experts, and PBM communities. The debates may lead to improving decision-making strategies in PBM (pertaining both to the decisions made by the patient or by the clinician). Subsequently, based on the top-down management framework for healthcare planning and control (Hans, 2015; Hans et al., 2011), further granularity in scientific evidence may foster medical planning and resource capacity planning to better target the right care to the right patient

– to provide more cost-effective, personalized care. For example, a simulation study (such as health economic modelling) may be later conducted using the methodologies in this research on what kind of design is needed in a particular clinical setting for PBM to be cost-effective and worth implementing. Or a data-driven project may be done in the future on estimating the burden on a clinical setting due to anaemia and blood transfusion overuse by varying a patient case mix.

6.2. Recommendations on Treating Transfusion as a Dependent Variable in Model Development

In the future, transfusion may be used as a dependent variable for modelling. Nevertheless, careful consideration must be given to the selection of input variables that shall not encompass transfusion triggers. A reader may have noticed that in this study, the intra-operative blood loss variable is absent among the input variables although it is present and sufficiently measured in the 'TOMaat' dataset. This modelling choice was done because blood loss is one of the two transfusion triggers of our clinical setting (introduced in Section 3.1.2.). The same consideration goes for the transfusion-related Hb level (per the '4-5-6 rule'). This Hb variable (if any) shall be omitted from model inputs. (This Hb variable was in fact not available (not measured) in the 'TOMaat' dataset.)

In essence, transfusion as a clinical intervention involves elements of independence – *choice*. In other words, someone eventually decides if transfusion is administered, ideally by evaluating if a transfusion trigger(s) was met. If transfusion were to be treated as a dependent variable in developing a feasible model, its triggers cannot be explicitly contained among the input variables. Thus, emerging data analysts in the transfusion medicine and PBM landscapes are encouraged in careful discernment, and shall pay close attention to transfusion triggers relevant to the clinical setting in scope of their study. Ideally, to prevent faulty assumptions, careful documentation of transfusion triggers in datasets can be an essential pre-requisite to avoid modelling flaws; otherwise, errors would give rise to infeasible models and misleading planning strategies thereof.

Transfusion triggers often pertain to a specific country. The 'TOMaat' dataset reflects a clinical setting in the Netherlands where strict transfusion triggers are in place per the national guidelines – the Dutch Blood Transfusion Guidelines (de Vries & Haas, 2012). Here, transfusion is administered to patients depending on meeting pre-established thresholds (recalling the '4-5-6 rule' and the blood loss trigger, Section 3.1.2.). In this work, we assumed that the Dutch Blood Transfusion Guidelines with the established transfusion triggers were strictly followed.

However, in practice it is sometimes seen that there is a lot of variability in transfusion triggers (Stanworth, 2023) which may lead to limitations in data analysis or to completely disregarding existing, promising patient-level datasets if traceability is missing. Dr. Stanworth (2023) has also recently appealed to rethink of how we do studies in transfusion medicine and to respond to the question *'Who really needs transfusion?'*. We claim that proper handling of predictors (to predict transfusion, or to predict patient outcomes) is one of the first steps towards patient-centred care in transfusion medicine and PBM. Recalling the articles per discussion points in Section 5.2., the study of Rashiq et al. (2004) was situated in Canada, and the research by Huang et al. (2018) was situated in China. Rashiq et al. (2004) reports that up to 12 hours before the administration of transfusion, the Hb measurements were collected. 80% of transfusion cases were triggered by meeting the Hb threshold (between 71 and 89 g/dL) and surgeon-specific triggers displayed low variability. Nevertheless, it is questionable to see the Hb variable among input variables of the multivariate logistic regression model. Then, we encourage a reader to think critically about Huang et al. (2018) because in the modelling steps of this publications, no clear, explicit elaboration on the specific transfusion triggers in the variable selection process was detected. Another questionable aspect in the two publications is that the information about the patient consent for receiving allogeneic RBC transfusion was not found.

Furthermore, datasets shall also contain the information if an informed consent for the administration of transfusion was given by a patient. Subsequently, if the patient characteristics met a transfusion trigger, transfusion follows. Traceability in databases on patient consent is also a key pre-requisite in similar studies. This allows to only subset records of those patients who gave their consent for the administration of transfusion. For example, if the information about the patient's informed consent to receive transfusion is not available, or if the patient participated in decision-making prior to the administration of transfusion, the application of similar methods may lead to strong misleading conclusions about the behaviour of a specific patient group in a given clinical setting. Otherwise, transfusion would be rather considered a parameter (equivalent to the respective transfusion triggers) if the information, whether a patient agreed beforehand to be administered transfusion, is missing.

6.3. Tips on Alternative Modelling Setup and Follow-up Data-driven Project Incentives involving Transfusion Data and Patient Outcomes

Although clinicians prefer simple models and easy interpretability, the high complexity of transfusion medicine may lead to the need of developing other more complex solutions to deal with uncertainty of the system behaviour. This project completion opens the door for tips on new project incentives involving data on blood products and patient outcomes (in various clinical settings).

Sanquin requests suggestions on follow-up projects involving the 'TOMaat' or other datasets. Future work and improvements may encompass many new, data-driven projects, namely, concerning:

[1] The mediation analysis model:

- Addition of two pairs of confounders (Hypothesis 2 and 3 per Section 1.2., and the assumptions per Section 3.6.). This improvement can lead to obtaining unbiased estimates noting that the current series of results remain biased. Next to it, SMEs in mediation analysis are advised to double-check the candidacy of the age and gender variables (under consideration) that in fact both contribute to the established transfusion triggers. Please refer to Section 6.2. above for details concerning transfusion triggers in modelling.
- Addition of other model components (i.e. moderators) that can improve estimations of variable effects in mediation analysis.
- Comparing outputs of models involving different sets of confounders (i.e. not only strong confounders).
- Dealing with multicollinearity in the current logistic regression models (referring to Section 5.3.). Per publications by Hernán and Robins (2020) and by Rijnhart et al. (2021), no explicit guidance or recommendations on dealing with or preventing collinearity in mediation analysis were found. Collinearity in the context of mediation analysis is known and described by Beasley (2013).
- Reversing the causality (RBC transfusion follows a complication).
- Exploring opportunities to utilize inference outputs from lasso for mediation analysis. This approach would favour a great reduction of the number of input variables.

[2] Patient outcomes:

- Excluding patient records with a missing date of complications (as an additional step to listwise deletion (Section 3.1.6.)). This exclusion is especially important in Patient group 1-1 that applies to the patients who experienced RBC transfusion followed by a complication. Now that we have the knowledge that the pessimistic and optimistic scenarios yield opposing results, this enhancement in data exclusion will reduce the study to a single scenario. (The 'TOMaat' dataset suffers from data missingness in the Patient group 1-1 for 48 patients from among 145 patients which is equivalent to 33% missingness in this Patient group.)
- Systematic literature reviews to further explore the field of patient outcomes research involving blood transfusion with accompanying (country-specific) guidelines (i.e. by

expanding on the discussion topics per Chapter 5). There is room to reflect on a variety of patient outcome variables in many kinds of surgical settings besides elective orthopaedic surgery.

- Exploring complications based on severity or selecting key complication types: From among all types of complications (i.e. infectious, pneumonia, cardiac, respiratory, or mental health to name a few), there is no focus on any particular type of complications in this work. Yet, to note, So-Osman et al. (2014a; 2014b) distinguish between thromboembolic (TE) and non-thromboembolic (non-TE) complications. Additionally, classification into several levels of severity was discussed during project initiation; however, due to the age of the data, data collection practices did not yet revolve around grouping complications based on severity as opposed to today's common practices in hospitals.
- Studying the numeric LOS variable as the patient outcome variable (perhaps, also relative to the health-economic aspect).
- Proceeding to longitudinal studies by analyzing the post-operative complications up to 3 months after surgery.
- Health-economic analysis (i.e. budget impact analysis). The 'TOMaat' dataset offers the Quality of Life, QoL, measures.
- Exploring opportunities using other datasets to study other patient outcomes (i.e. readmission).

[3] The RBC transfusion variable:

- Establishing RBC transfusion as a numeric variable to study RBC as the blood product use.
- Reducing the complexity and variability of the dataset by selecting patient subgroups (into separate models) based on specific transfusion triggers.
- Finding new opportunities for standardization (storing data about transfusion triggers, patient consent) on the national or international level.

[4] Supervised machine learning models:

- Dealing with multicollinearity in the existing lasso models.
- Exploring multi-class classification models (i.e. what is the probability that a patient will have both transfusion and the patient outcome, independent of the time sequence).
- Building generalized additive models to deal with high natural variability of the dataset. Non-linear alternatives of parametric models are available as extensions of linear models, namely, polynomial regression, step functions, regression or smoothing splines, local regression, and generalized additive models (James et al., 2021).
- Improving the model performance with other black-box models (i.e. deep learning) and drawing insights from their PDPs. "Partial dependence functions can be used to help interpret models produced by any "black box" prediction method, such as neural networks... When there is a large number of predictor variables, it is very useful to have a measure of relevance." (Friedman, 2001)

[5] Variable selection:

- Extracting information from free text fields or other variables excluded in this study.
- Studying variable selection methods and comparing the resulting inference results and model uncertainty measures.

[6] Software engineering:

- Mitigating technical/human errors by blending validation activities and other software engineering practices into a workplace.

A series of themes for follow-up projects were proposed. Through their execution (under the umbrella of Sanquin, and perhaps also partnering hospitals), clinical researchers may sequentially get acquainted with the increasing complexity of modelling needs.

It is vital to recall that WHO (2021) being a key, international health authority urges to implement PBM across healthcare systems to tackle the burden tied to transfusion dependency. In this regard, we encourage Sanquin and other healthcare establishments to take action in patient outcomes research and in supporting the design of decision support solutions involving transfusion data for clinical use. We see opportunities for improving and utilizing digital maturity in hospitals by giving appropriate attention to the data quality and data collection efforts which will thoroughly represent transfusion interventions and the accompanying healthcare operations. Although it may take some time to yield useful results based on existing datasets, we advise health authorities to continue paying attention to the field of PBM and making strategic decisions, especially, in the matter of funding and human resources. This may involve building multidisciplinary teams with significant contribution of dataset content experts (i.e. clinical/transfusion specialists) and with professionals committed to patient-centredness. Besides smaller, local initiatives, a great importance while striving for change and scale-up goes to government authorities and patient advocacy groups.

Chapter 7 | Conclusion

"The need for robust methods seems to be intimately mixed up with the need for simple models."
(James Box, 1979)

Our study offers key steps to an innovative roadmap to model a complex surgical setting involving (RBC) transfusion, patient outcomes, pre-operative anaemia and selected (30+) variables from a patient-level dataset ('TOMaat'), here, gathered in the elective orthopaedic surgical setting. The roadmap is represented by the framework of this research if similar methodology was implemented to other patient-level datasets. Clinical researchers favour simple models and ease in their interpretability despite the high complexity of transfusion medicine. Studying RBC transfusion as a mediator in mediation analysis – the method requested by the Sanquin Blood Bank, is novel in this field of medicine.

The central challenge lied in investigating **Hypotheses 1 and 4** (Section 1.2.) which, respectively, pertain to the mediating role of RBC transfusion (RBC) in the relationship between pre-operative anaemia and post-operative complications up to Day 14 (COM), and strong confounding variables in relation to RBC transfusion and these complications. First, to simplify this complex realm, the study employs massive data preparation efforts and variable selection methods during the development of multiple supervised learning models (random forest, logistic regression, and lasso). Two scenarios were established due to extensive data missingness (33%) of the patient outcome dates. The model performance is deemed moderate or poor with overlapping confidence intervals for the AUC metrics indicating similar performance of all models. The RBC models exhibit AUC's between 0.69 and 0.71, and the COM models between 0.63 and 0.69. Their 95% CI's range between 0.58 and 0.78.

The research confirms **Hypothesis 4** – there are strong confounders of RBC transfusion and post-operative complications up to Day 14. 10 strong confounding variables were identified from inference results of supervised learning models, substantiated by clinical insights from a transfusion medicine expert – namely: hospital, age, gender, type of surgery (hip or knee, and primary vs revision), cardiovascular disease, diabetes mellitus, increased risk group, EPO, and the surgery year.

Next to it, supervised learning models reveal the statistical significance of RBC transfusion, particularly, in the pessimistic scenario. This means for this scenario that RBC transfusion was found to be a strong predictor for the occurrence of post-operative complications up to Day 14. In contrary, in the optimistic scenario, there is no statistical significance of RBC transfusion relative to this patient outcome variable.

Interestingly, variable significance differs when analyzed individually (in silo) versus in the presence of other covariates, as seen with pre-operative anaemia and other variables.

In response to **Hypothesis 1**, mediation analysis yielded opposing results for the two scenarios. RBC transfusion mediates the relationship between pre-operative anaemia and post-operative complications up to Day 14 in the pessimistic scenario (with the ACME of 0.0445 and the 95% CI of [0.0268; 0.0700]), yet does not mediate this relationship in the optimistic scenario. Improving the mediation model with additional components is encouraged to reduce the current bias.

Overall, this study sheds light on the dynamics of RBC transfusion, pre-operative anaemia, confounding variables, and their role in post-operative complications up to Day 14. The importance of reporting uncertainty measures of multiple models, and the importance of having incorporated the insights of a transfusion expert are apparent. Numerous suggestions for future work (i.e. through multidisciplinary partnerships between Sanquin and hospitals) were proposed and discussed. Especially the new insights of how to treat RBC transfusion in modelling, standardization of its triggers, or documenting informed consents by patients may contribute to the advancement of patient-centred care, PBM and evidence-based medicine in the future while striving for relieving the burden due to transfusion dependency.

References

Alfons, A., Ateş, N. Y., & Groenen, P. J. F. (2021). A Robust Bootstrap Test for Mediation Analysis. Organizational Research Methods, 109442812199909. https://doi.org/10.1177/1094428121999096

Beasley, T. M. (2013). Tests of Mediation: Paradoxical Decline in Statistical Power as a Function of Mediator Collinearity. The Journal of Experimental Education, 82(3), 283–306. https://doi.org/10.1080/00220973.2013.813360

Blood and Beyond. (2020). Rethinking blood use in Europe to improve outcomes for patients - infographic. In *Blood and Beyond*. Celgene Corporation. https://www.bloodandbeyond.com/wp-content/uploads/Blood_and_Beyond_Infographic_Update_November_2020.pdf

Blood and Beyond. (2021). Rethinking blood use in Europe to improve outcomes for patients. In Blood and Beyond. Celgene Corporation. https://www.bloodandbeyond.com/wp-content/uploads/Blood_and_Beyond_Rethinking_blood_use_in_Europe_Updated_November_2020.pdf

Box, G. E. P. (1979). Robustness in the Strategy of Scientific Model Building. Robustness in Statistics, 201–236. https://doi.org/10.1016/b978-0-12-438150-6.50018-2

Bramley, P., McArthur, K., Blayney, A., & McCullagh, I. (2021). Risk factors for postoperative delirium: An umbrella review of systematic reviews. International Journal of Surgery, 93, 106063. https://doi.org/10.1016/j.ijsu.2021.106063

Centre for Evidence-Based Medicine. (2009, March). Levels of Evidence. University of Oxford. https://www.cebm.ox.ac.uk/resources/levels-of-evidence/oxford-centre-for-evidence-based-medicine-levels-of-evidence-march-2009

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology, 110, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Cinelli, C., Forney, A., & Pearl, J. (2022). A Crash Course in Good and Bad Controls. Sociological Methods & Research, 004912412210995. https://doi.org/10.1177/00491241221099552

de Groot, R., Hoenink, J. C., Mackenbach, J. D., den Braver, N. R., Pinho, M. G. M., Brassinga, D., Prinsze, F. J., Timmer, T. C., de Kort, W. L. A. M., Brug, J., van den Hurk, K., & Lakerveld, J. (2019). The association between population density and blood lipid levels in Dutch blood donors. International Journal of Health Geographics, 18(1). https://doi.org/10.1186/s12942-019-0167-y

de Vries, R., & Haas, F. (2012). English Translation of the Dutch Blood Transfusion Guideline 2011. *Clinical Chemistry*, 58(8), 1266–1267. https://doi.org/10.1373/clinchem.2012.189209

Dhiman, P., Ma, J., Gibbs, V. N., Alexandros Rampotas, Kamal, H., Arshad, S. S., Kirtley, S., Doree, C., Murphy, M., Collins, G. S., & Antony JR. Palmer. (2023). Systematic review highlights high risk of bias of clinical prediction models for blood transfusion in patients undergoing elective surgery. Journal of Clinical Epidemiology, 159, 10–30. https://doi.org/10.1016/j.jclinepi.2023.05.002

Frenzel, T., Van Aken, H., & Westphal, M. (2008). Our own blood is still the best thing to have in our veins. Current Opinion in Anaesthesiology, 21(5), 657–663. https://doi.org/10.1097/aco.0b013e3283103e84

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Greenland, S., & Morgenstern, H. (2001). Confounding in Health Research. Annual Review of Public Health, 22(1), 189–212. https://doi.org/10.1146/annurev.publhealth.22.1.189

Greenwell, B. M. (2017). pdp: An R Package for Constructing Partial Dependence Plots. The R Journal, 9(1), 421. https://doi.org/10.32614/rj-2017-016

Hans, E. W. (2015). Is it better now doctor? Inaugural lecture given upon acceptance of the Chair of Operations Management in Health Care at the Faculty of Behavioural, Management and Social Sciences, University of Twente.

Hans, E. W., Van Houdenhoven, M., & Hulshof, P. J. H. (2011). A framework for health care planning and control. In Handbook of Health Care Systems Scheduling (pp. 303-320). Springer.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction (2nd ed.). Springer.

Heerkens, H., & van Winden, A. (2017). Solving Managerial Problems Systematically. Noordhoff Uitgevers bv.

Hernán M., & Robins, J. M. (2021). Causal inference: What If. CRC Press.

Hilderink, H. B. M., Plasmans, M. H. D., Poos, M. J. J. C., Eysink, P. E. D., & Gijsen, R. (2020). Dutch DALYs, current and future burden of disease in the Netherlands. Archives of Public Health, 78(1). https://doi.org/10.1186/s13690-020-00461-8

Hofmann, A., Farmer, S., & Shander, A. (2011). Five Drivers Shifting the Paradigm from Product-Focused Transfusion Practice to Patient Blood Management. The Oncologist, 16(S3), 3–11. https://doi.org/10.1634/theoncologist.2011-s3-3

Hosmer, D. W., & Lemeshow, S. (2000). Applied Logistic Regression. John Wiley & Sons, Inc. https://doi.org/10.1002/0471722146

Huang, Z., Huang, C., Xie, J., Ma, J., Cao, G., Huang, Q., Shen, B., Byers Kraus, V., & Pei, F. (2018). Analysis of a large data set to identify predictors of blood transfusion in primary total hip and knee arthroplasty. Transfusion, 58(8), 1855–1862. https://doi.org/10.1111/trf.14783

IKNL. (2023). About IKNL. Netherlands Comprehensive Cancer Organisation (IKNL). https://iknl.nl/en/about-iknl

Isbister, J. (2005). Why Should Health Professionals be Concerned about Blood Management and Blood Conservation? Updates in Blood Conservation and Transfusion Alternatives, 2(Dec:3-7).

Isbister, J. P. (2013). The three-pillar matrix of patient blood management – An overview. Best Practice & Research Clinical Anaesthesiology, 27(1), 69–84. https://doi.org/10.1016/j.bpa.2013.02.002

James, G. M., Witten, D., Hastie, T. J., & Tibshirani, R. (2021). An introduction to statistical learning: with applications in R (2nd ed.). Springer.

Last, J. M. (2001). A dictionary of epidemiology (4th ed.). Oxford University Press.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mascha, E. J., Dalton, J. E., Kurz, A., & Saager, L. (2013). Understanding the Mechanism: Mediation Analysis in Randomized and Nonrandomized Studies. Anesthesia & Analgesia, 117(4), 980–994. https://doi.org/10.1213/ane.0b013e3182a44cb9

McAlexander, R. J., & Mentch, L. (2020). Predictive inference with random forests: A new perspective on classical analyses. Research & Politics, 7(1), 205316802090548. https://doi.org/10.1177/2053168020905487

Meier, J. M., & Tschoellitsch, T. (2022). Artificial Intelligence and Machine Learning in Patient Blood Management: A Scoping Review. Anesthesia & Analgesia, 135(3), 524–531. https://doi.org/10.1213/ane.0000000000006047

Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable | Partial Dependence Plot (PDP). In christophm.github.io (2nd ed.). https://christophm.github.io/interpretable-ml-book/pdp.html

Morabia, A. (2010). History of the modern epidemiological concept of confounding. Journal of Epidemiology & Community Health, 65(4), 297–300. https://doi.org/10.1136/jech.2010.112565

Nash, D. B., Joshi, M., Ransom, E. R., & Ransom, S. B. (2019). The healthcare quality book: Vision, strategy, and tools (4th ed.). Health Administration Press, Chicago, Illinois, United States.

Ozawa, S. (2023, January 1). What's Happening Globally in Patient Blood Management? A Sit Down with Sherri Ozawa. Let's Talk Patient Blood Management (podcast series).

R Documentation. (2023a). partial: Partial Dependence Functions. Retrieved May 15, 2023, from https://www.rdocumentation.org/packages/pdp/versions/0.8.1/topics/partial

R Documentation. (2023b). R: Inference for the lasso, with a fixed lambda. Search.r-Project.org. Retrieved May 15, 2023, from https://search.r-project.org/CRAN/refmans/selectiveInference/html/fixedLassoInf.html

Rashiq, S., Shah, M., Chow, A. K., O'Connor, P. J., & Finegan, B. A. (2004). Predicting Allogeneic Blood Transfusion Use in Total Joint Arthroplasty. Anesthesia & Analgesia, 99(4), 1239–1244. https://doi.org/10.1213/01.ane.0000132928.45858.92

Richiardi, L., Bellocco, R., & Zugna, D. (2013). Mediation analysis in epidemiology: methods, interpretation and bias. International Journal of Epidemiology, 42(5), 1511–1519. https://doi.org/10.1093/ije/dyt127

Rijnhart, J. J. M. (2021). Comparison of Methods for Statistical Mediation Analysis within Epidemiological Research [Doctoral thesis]. https://www.globalacademicpress.com/ebooks/judith_rijnhart/index.html#p=1

Rijnhart, J. J. M., Valente, M. J., Smyth, H. L., & MacKinnon, D. P. (2021). Statistical Mediation Analysis for Models with a Binary Mediator and a Binary Outcome: the Differences Between Causal and Traditional Mediation Analysis. Prevention Science. https://doi.org/10.1007/s11121-021-01308-6

Saager, L., Turan, A., Reynolds, L. F., Dalton, J. E., Mascha, E. J., & Kurz, A. (2013). The Association Between Preoperative Anemia and 30-Day Mortality and Morbidity in Noncardiac Surgical Patients. Anesthesia & Analgesia, 117(4), 909–915. https://doi.org/10.1213/ane.0b013e31828b347d

Safiri, S., Kolahi, A.-A., Noori, M., Nejadghaderi, S. A., Karamzad, N., Bragazzi, N. L., Sullman, M. J. M., Abdollahi, M., Collins, G. S., Kaufman, J. S., & Grieger, J. A. (2021). Burden of anemia and its underlying causes in 204 countries and territories, 1990–2019: results from the Global Burden of Disease Study 2019. Journal of Hematology & Oncology, 14(1). https://doi.org/10.1186/s13045-021-01202-2

Sanquin Blood Supply Foundation. (2021). Highlights 2021. Sanquin Annual Reports. https://www.sanquin.nl/binaries/content/assets/sanquinnl/over-sanquin/pers--actueel/jaarverslagen/highlights-2021-stichting-sanquin_uk.pdf

Sanquin. (2023a). The Story of Sanquin. Retrieved January 11, 2023, from https://www.sanquin.nl/en/about-sanquin/the-story-of-sanquin

Sanquin. (2023b). Wordt gedoneerd bloed en plasma commercieel ingezet? Sanquin. Retrieved March 1, 2023, from https://www.sanquin.nl/over-sanquin/dossiers/wordt-gedoneerd-bloed-en-plasma-commercieel-ingezet

Schuster, N. A., Rijnhart, J. J. M., Bosman, L. C., Twisk, J. W. R., Klausch, T., & Heymans, M. W. (2023). Misspecification of confounder-exposure and confounder-outcome associations leads to bias in effect estimates. BMC Medical Research Methodology, 23(1). https://doi.org/10.1186/s12874-022-01817-0

Shander, A., Goobie, S. M., Warner, M. A., Aapro, M., Bisbe, E., Perez-Calatayud, A. A., Callum, J., Cushing, M. M., Dyer, W. B., Erhard, J., Faraoni, D., Farmer, S., Fedorova, T., Frank, S. M., Froessler, B., Gombotz, H., Gross, I., Guinn, N. R., Haas, T., & Hamdorf, J. (2020). The Essential Role of Patient Blood Management in a Pandemic: A Call for Action. Anesthesia and Analgesia. https://doi.org/10.1213/ANE.0000000000004844

Shander, A., Hardy, J.-F., Ozawa, S., Farmer, S. L., Hofmann, A., Frank, S. M., Kor, D. J., Faraoni, D., Freedman, J., & Collaborators. (2022). A Global Definition of Patient Blood Management. Anesthesia and Analgesia. https://doi.org/10.1213/ANE.0000000000005873

So-Osman, C. (2012). Patient Blood Management in Elective Orthopaedic Surgery [Doctoral thesis]. https://scholarlypublications.universiteitleiden.nl/handle/1887/20071

So-Osman, C., Nelissen, R. G. H. H., Koopman-van Gemert, A. W. M. M., Kluyver, E., Pöll, R. G., Onstenk, R., Van Hilten, J. A., Jansen-Werkhoven, T. M., van den Hout, W. B., Brand, R., &

Brand, A. (2014a). Patient Blood Management in Elective Total Hip- and Knee-replacement Surgery (Part 1). Anesthesiology, 120(4), 839–851. https://doi.org/10.1097/aln.0000000000000134

So-Osman, C., Nelissen, R. G. H. H., Koopman-van Gemert, A. W. M. M., Kluyver, E., Pöll, R. G., Onstenk, R., Van Hilten, J. A., Jansen-Werkhoven, T. M., van den Hout, W. B., Brand, R., & Brand, A. (2014b). Patient Blood Management in Elective Total Hip- and Knee-replacement Surgery (Part 2). Anesthesiology, 120(4), 852–860. https://doi.org/10.1097/aln.0000000000000135

So-Osman, C., van der Wal, D. E., & Allard, S. (2017). Patient Blood Management initiatives on a global level: the results of an International Society of Blood Transfusion Survey. ISBT Science Series, 12(3), 327–335. https://doi.org/10.1111/voxs.12356

Stanworth, S. J. (2023). Speech at the NVB-TRIP Symposium, Ede, the Netherlands.

Šuster, S., Baldwin, T., & Verspoor, K. (2023). Analysis of predictive performance and reliability of classifiers for quality assessment of medical evidence revealed important variation by medical area. Journal of Clinical Epidemiology. https://doi.org/10.1016/j.jclinepi.2023.04.006

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Weinberg, C. R. (1993). Toward a Clearer Definition of Confounding. American Journal of Epidemiology, 137(1), 1–8. https://doi.org/10.1093/oxfordjournals.aje.a116591

WHO. (2014). Global nutrition targets 2025: anaemia policy brief (WHO/NMH/NHD/14.4). Geneva: World Health Organization. https://apps.who.int/nutrition/publications/globaltargets2025_policybrief_anaemia/en/

WHO. (2021). The urgent need to implement patient blood management: policy brief. World Health Organization. https://apps.who.int/iris/handle/10665/346655
License: CC BY-NC-SA 3.0 IGO

WHO. (2023). Anaemia. World Health Organisation. Retrieved January 11, 2023, from https://www.who.int/health-topics/anaemia#tab=tab_1

WHO Scientific Group on Nutritional Anaemias & World Health Organization. (1968). Nutritional anaemias : report of a WHO scientific group [meeting held in Geneva from 13 to 17 March 1967]. World Health Organization. https://apps.who.int/iris/handle/10665/40707

Appendix A: Exploratory Data Analysis

The (graphical) content in this appendix serves as a supplement to RQ#1. The sequence of key patient groups in Table A-1 is consistent with Table 4-2:

Table A-1: Descriptive statistics and EDA results: Insights on associations among selected input variables (key patient subgroups) and target patient outcomes (Cases). For binary or categorical variables, the descriptive statistics are counts and proportions (at least 1 COM / all COM, or at least 1 RBC / all RBC). For numeric variables, the descriptive statistics are mean, median and interquartile range (IQR).

Key patient subgroup	no case	Case COM_{PES}		Case COM_{OPT}		Case RBC		Case LOS	
	Descrip-tive statistics	Descriptive statistics (at least 1 COM)	p-value, or corre-lation	Descriptive statistics (at least 1 COM)	p-value, or corre-lation	Descriptive statistics (at least 1 RBC)	p-value, or corre-lation	Descriptive statistics	corre-lation
Participating hospitals (Hospital1; Hospital2; Hospital3; Hospital4)	401; 956; 602; 467	127 (31.7%); 166 (17.4%); 84 (14.0%); 134 (28.7%)	**<0.001**	117 (29.2%); 142 (14.9%); 80 (13.3%); 124 (26.6%)	**<0.001**	54 (13.5%); 122 (12.8%); 29 (4.8%); 52 (11.1%)	**<0.001**	[median 8, mean 9.0, IQR (7;10)]; [median 6, mean 7.8, IQR (5;8)]; [median 6, mean 6.8, IQR (5;8)]; [median 8, mean 9.4, IQR (7;10)]	weak
Type of surgery (total hip; or total knee replacement)	975; 1451	189 (19.4%); 322 (22.2%)	0.107	175 (17.9%); 288 (19.8%)	0.265	57 (5.8%); 200 (13.8%)	**<0.001**	[median 7, mean 7.9, IQR (6;9)]; [median 7, mean 8.1, IQR (6;9)]	weak
Type of surgery (primary; or revision)	2245; 181	465 (20.7%); 46 (25.4%)	0.162	423 (18.8%); 40 (22.1%)	0.330	225 (10.0%); 32 (17.7%)	**0.002**	[median 7, mean 7.9, IQR (6;9)]; [median 8, mean 10.0, IQR (6;10)]	weak
Pre-operative anaemia (yes; no)	214; 2212	63 (29.4%); 448 (20.3%)	**0.002**	52 (24.3%); 411 (18.6%)	0.052	60 (28.0%); 197 (8.9%)	**<0.001**	[median 8, mean 9.5, IQR (6;10)]; [median 7,	weak

Key patient subgroup		no case	Case COM$_{\text{PES}}$		Case COM$_{\text{OPT}}$		Case RBC		Case LOS	
		Descriptive statistics	Descriptive statistics (at least 1 COM)	p-value, or correlation	Descriptive statistics (at least 1 COM)	p-value, or correlation	Descriptive statistics (at least 1 RBC)	p-value, or correlation	Descriptive statistics	correlation
									mean 7.9, IQR (5;9)]	
EPO therapy (yes; no)		227; 2199	58 (25.6%); 453 (20.6%)	0.098	52 (22.9%); 411 (18.7%)	0.147	22 (9.7%); 235 (10.7%)	0.726	[median 8, mean 9, IQR (6;10)]; [median 7, mean 8.0, IQR (6;9)]	weak
Cell saver device (hip replacement only) as intra-operative autologous reinfusion up to Day 14	Cell saver device (yes; no)	271, 2155	68 (25.1%), 443 (20.6%)	0.100	61 (22.5%), 402 (18.7%)	0.150	33 (12.2%), 224 (10.4%)	0.427	[median 8, mean 9.0, IQR (6;10)]; [median 7, mean 7.9, IQR (6;9)]	weak
	Collection (mL), numeric, subset >0 mL only	[median 200, mean 322, IQR (0;478)]	at least 1 COM [median 160, mean 419, IQR (0;476)]; no COM [median 200, mean 290, IQR (0;473)]	weak	at least 1 COM [median 150, mean 411, IQR (0;450)]; no COM [median 210, mean 297, IQR (0;498)]	weak	at least 1 RBC [median 300, mean 595, IQR (0;600)]; no RBC [median 198, mean 285, IQR (0;448)]	weak	slightly positive	
	Reinfusion (mL), numeric, subset >0 mL only	[median 60, mean 121, IQR (0;195)]	at least 1 COM [median 50, mean 171, IQR (0;200)]; no COM [median 70, mean 105, IQR (0;180)]	weak	at least 1 COM [median 50, mean 177, IQR (0;200)]; no COM [median 73, mean 105, IQR (0;180)]	weak	at least 1 RBC [median 75, mean 229, IQR (0;230)]; no RBC [median 60, mean 106, IQR (0;180)]	weak	slightly positive	
Total blood loss during surgery (mL), numeric, subset >0 mL only, applies to hip replacement patients only		[median 350, mean 450, IQR (200;550)]	at least 1 COM [median 400, mean 568, IQR (250;600)]; no COM [median 350, mean 416, IQR (200;500)]	weak	at least 1 COM [median 400, mean 554, IQR (250;600)]; no COM [median 350, mean 423, IQR (200;510)]	weak	at least 1 RBC [median 600, mean 816, IQR (355;948)]; no RBC [median 335, mean 394, IQR (200;500)]†	weak	slightly positive	
RBC transfusion (yes; no)		257; 2169	112 (43.6%); 399 (18.4%)	<0.001	64 (24.9%); 399 (18.4%)	<0.001	n/a (infeasible*)		[median 9, mean 12.2, IQR (7;13.25)]; [median 7, mean 7.6, IQR (5;9)]	weak

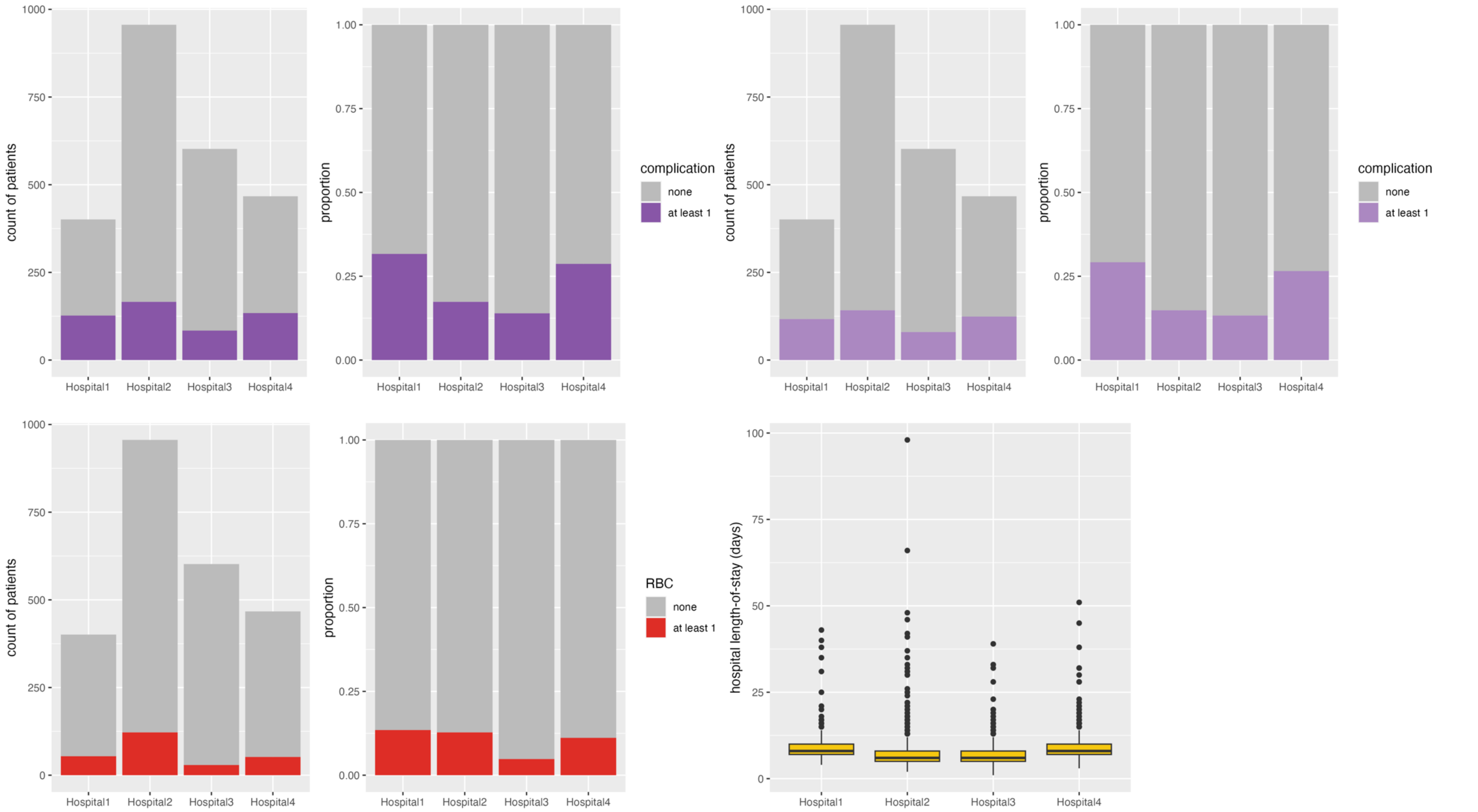


Figure A-1: Stratification for the participating hospitals: Case COM$_{PES}$ (top left, p<0.001), COM$_{OPT}$ (top right, p<0.001), RBC (bottom left, p<0.001), and LOS (bottom right).

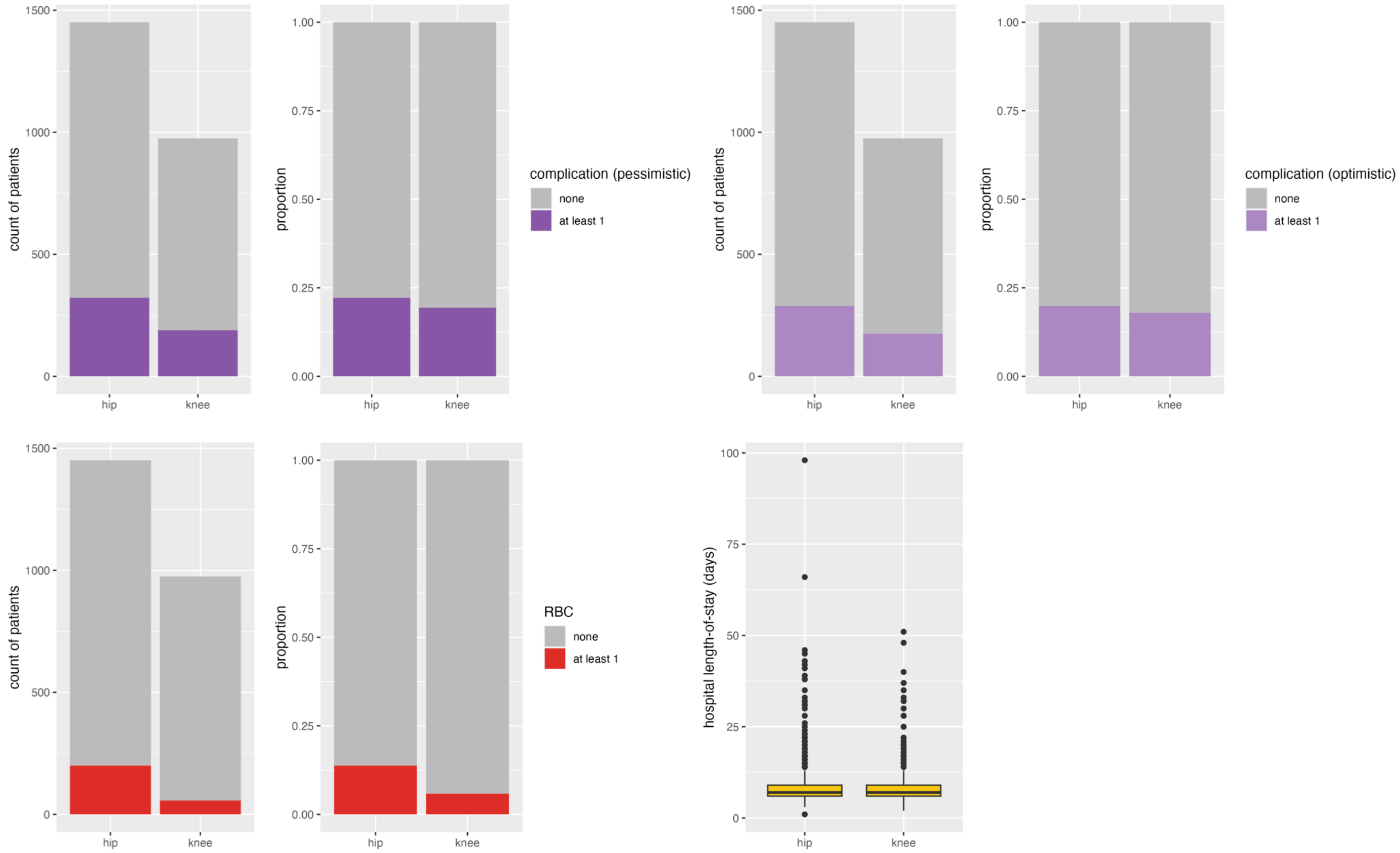


Figure A-2: Stratification for the surgery type (total hip or knee replacement): Case $\mathrm{COM}_{\mathrm{PES}}$ (top left, p=0.107), $\mathrm{COM}_{\mathrm{OPT}}$ (top right, p=0.265), RBC (bottom left, p<0.001), and LOS (bottom right).

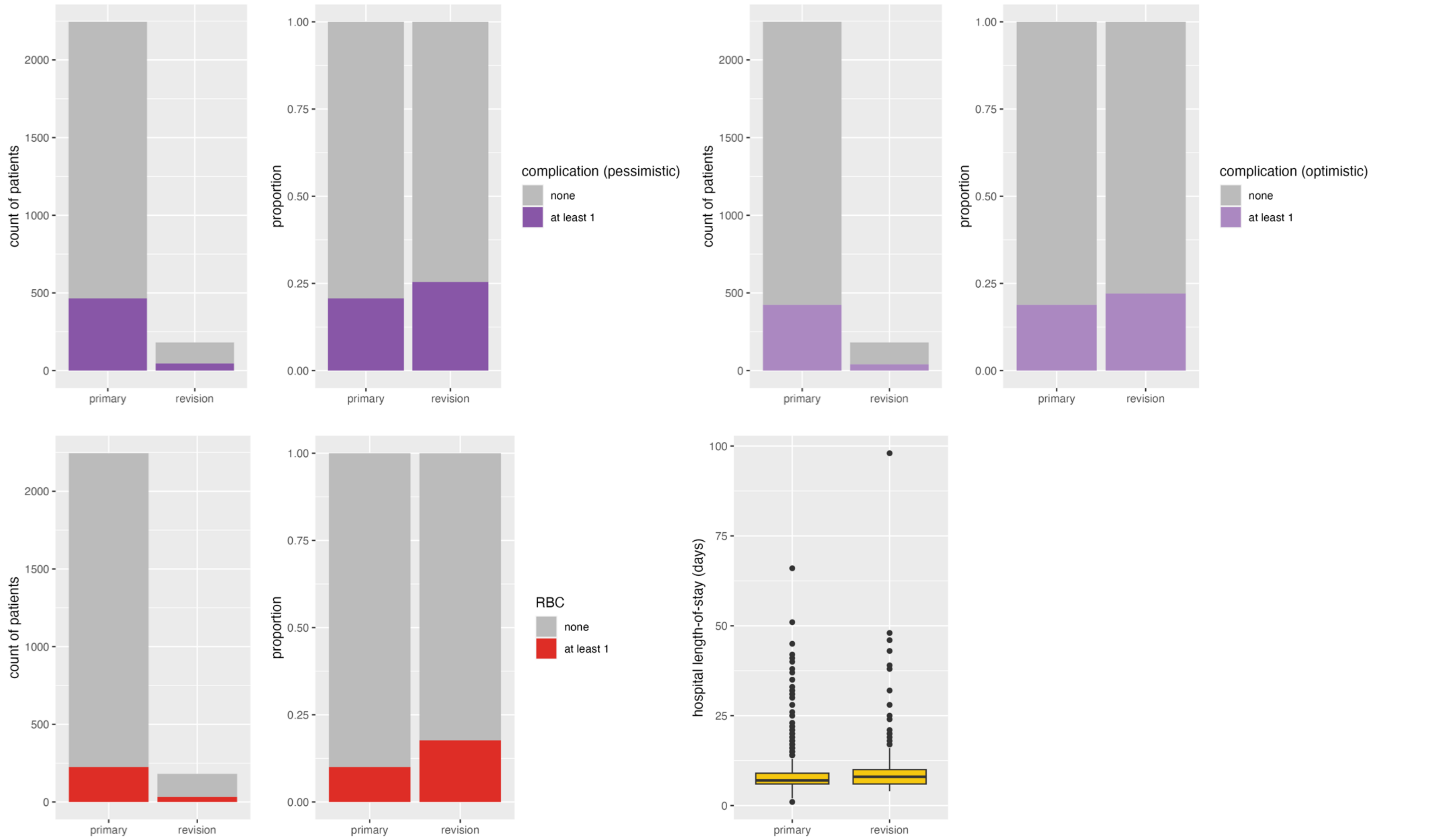


Figure A-3: Stratification for the surgery type (primary or revision): Case COM$_{PES}$ (top left, p=0.162), COM$_{OPT}$ (top right, p=0.330), RBC (bottom left, p=0.002), and LOS (bottom right).

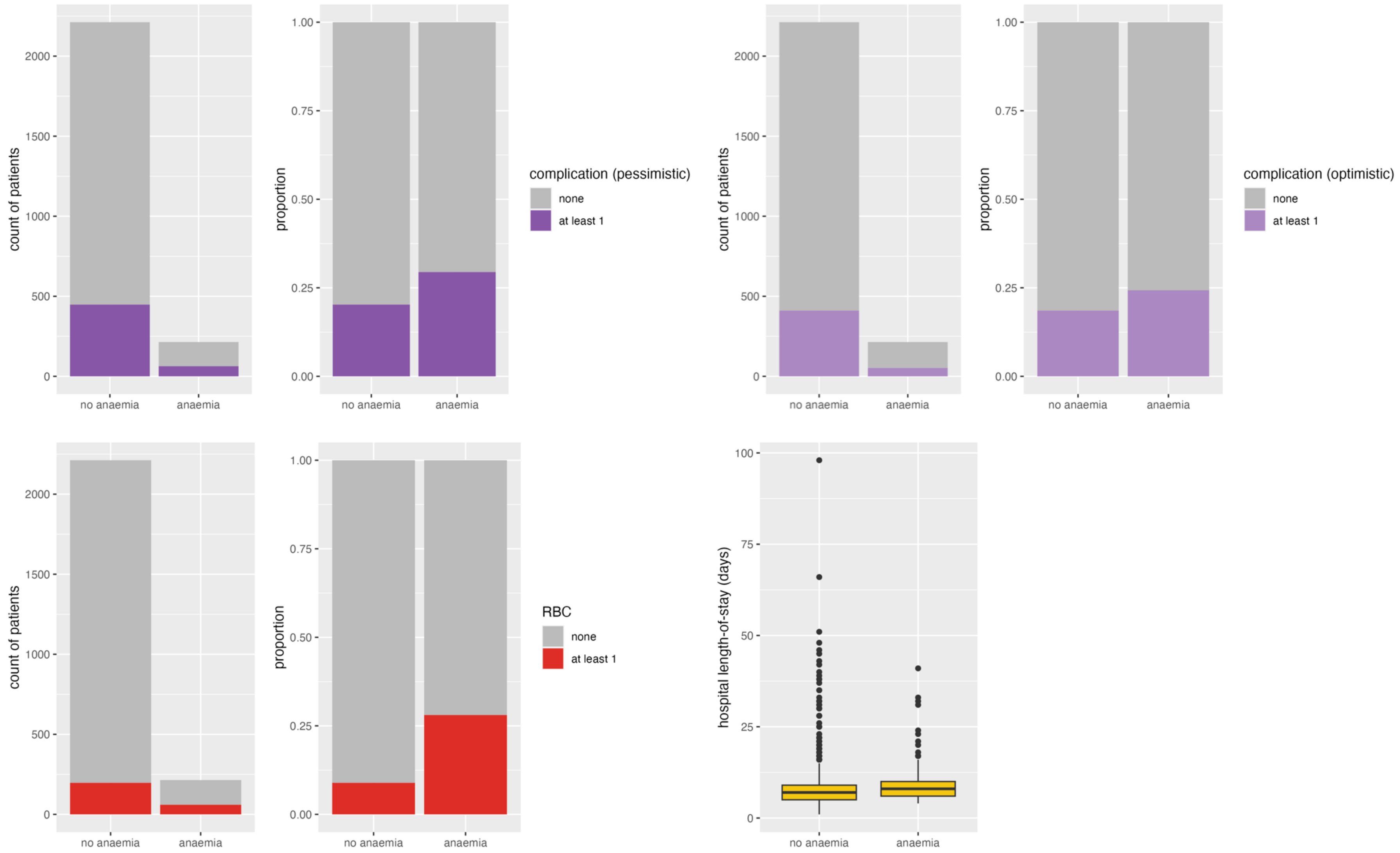


Figure A-4: Stratification for pre-operative anaemia: Case COM$_{PES}$ (top left, p=0.002), COM$_{OPT}$ (top right, p=0.052), RBC (bottom left, p<0.001), and LOS (bottom right).

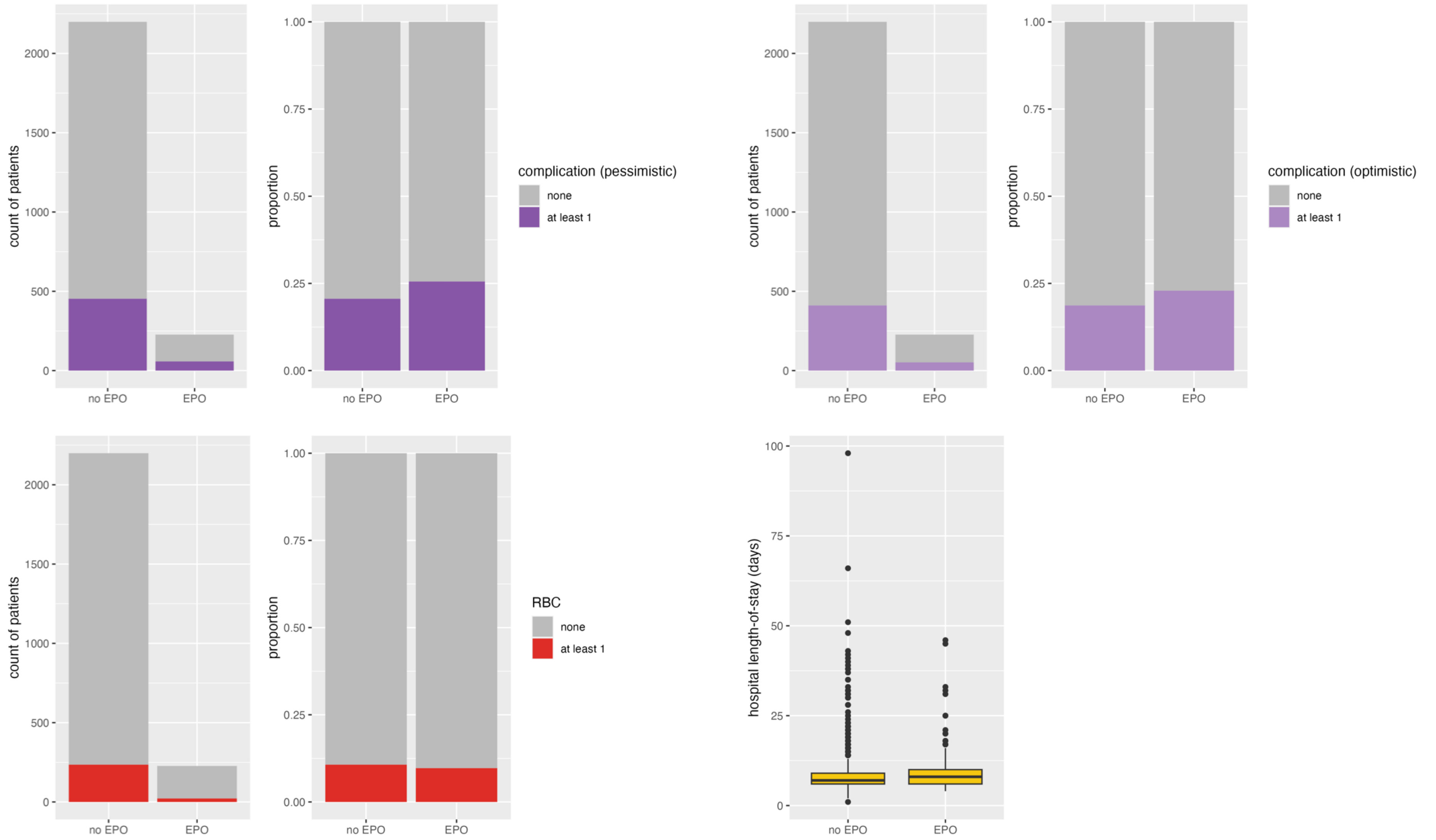


Figure A-5: Stratification for EPO: Case COMPES (top left, p=0.098), COMOPT (top right, p=0.147), RBC (bottom left, p=0.726), and LOS (bottom right).

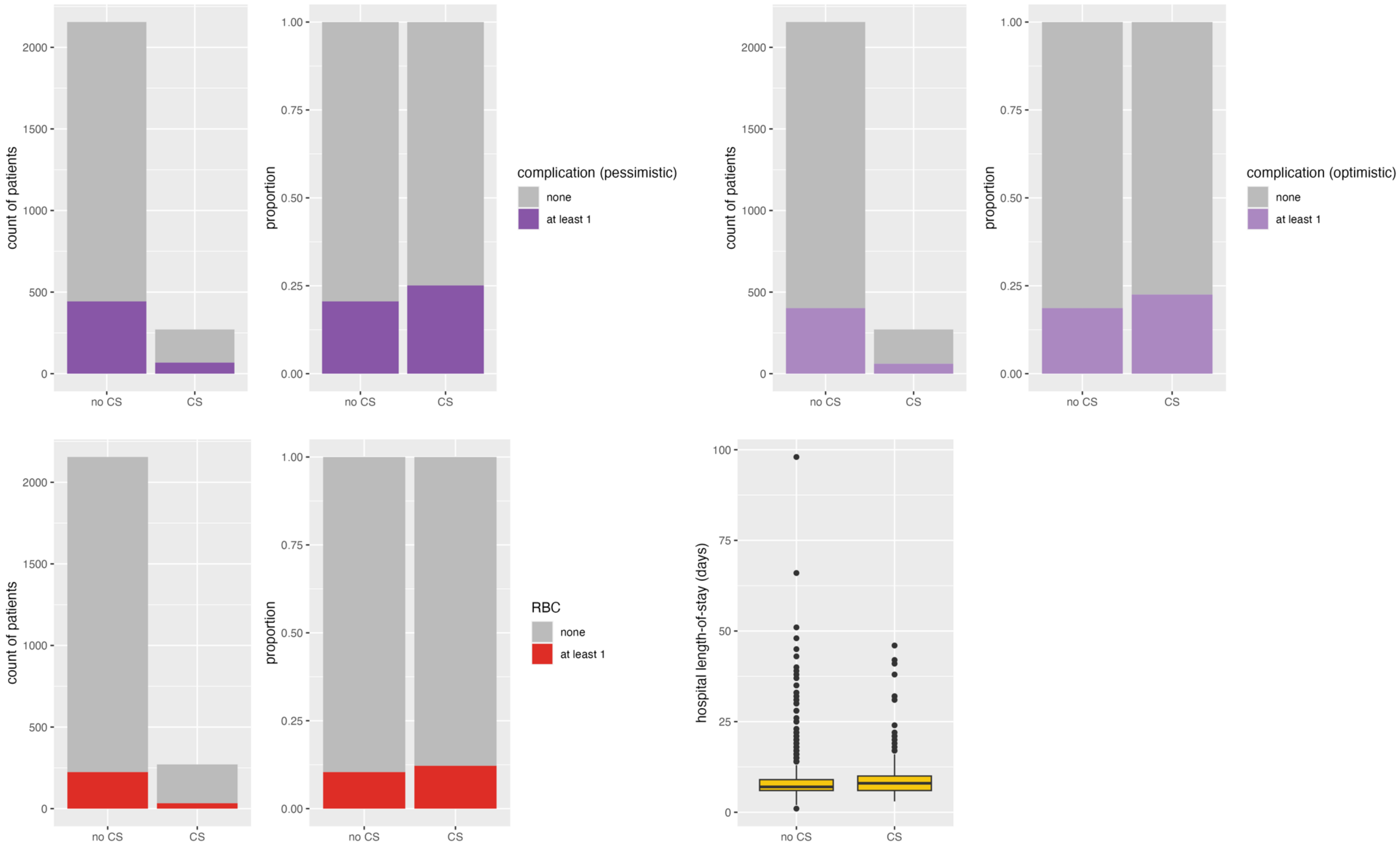


Figure A-6: Stratification for the intra-operative cell saver: Case COM_{PES} (top left, p=0.100), COM_{OPT} (top right, p=0.150), RBC (bottom left, p=0.427), and LOS (bottom right). Only patients with hip replacement could qualify for a cell saver.

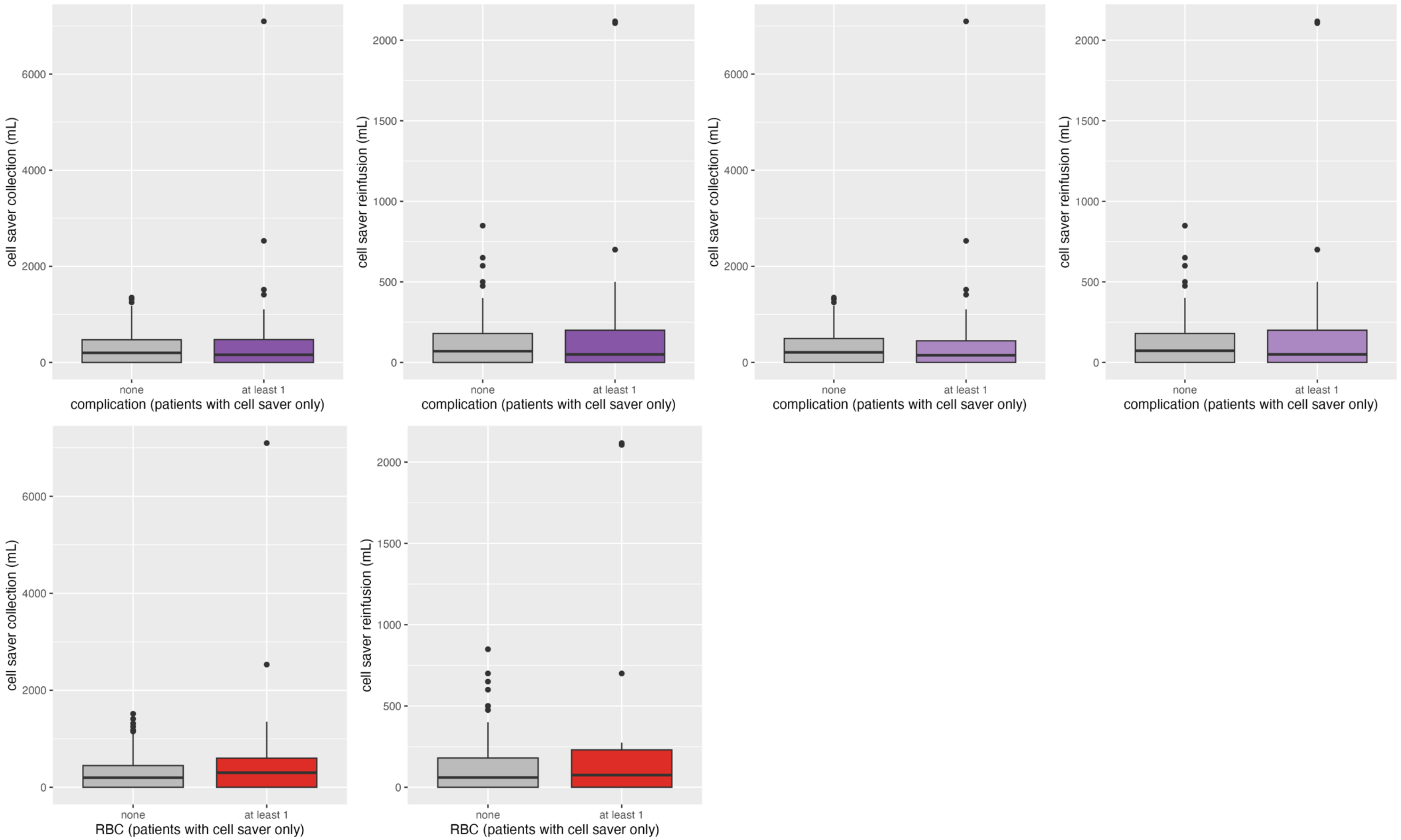


Figure A-7: Stratification for the intra-operative cell saver collection and reinfusion (numeric format, patients with cell saver only): Case COM$_{\text{PES}}$ (top left), COM$_{\text{OPT}}$ (top right), and RBC (bottom left). Only patients with hip replacement could qualify for a cell saver (thus, had non-zero cell saver collection and reinfusion).

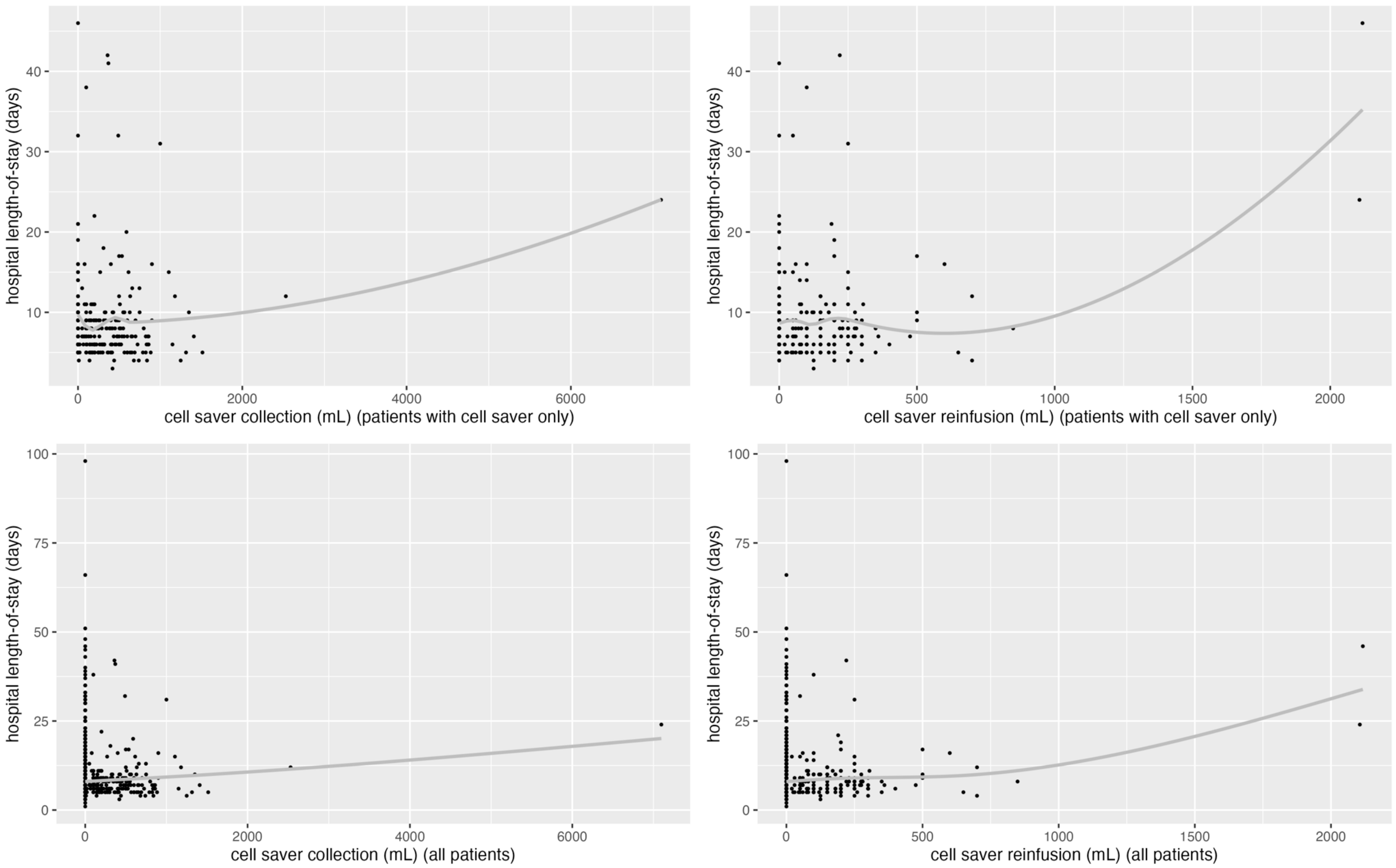


Figure A-8: Stratification for the intra-operative cell saver collection and reinfusion (numeric format): Case LOS: patients with cell saver only (top), and all patients (bottom). Only patients with hip replacement could qualify for a cell saver (thus, had non-zero cell saver collection and reinfusion).

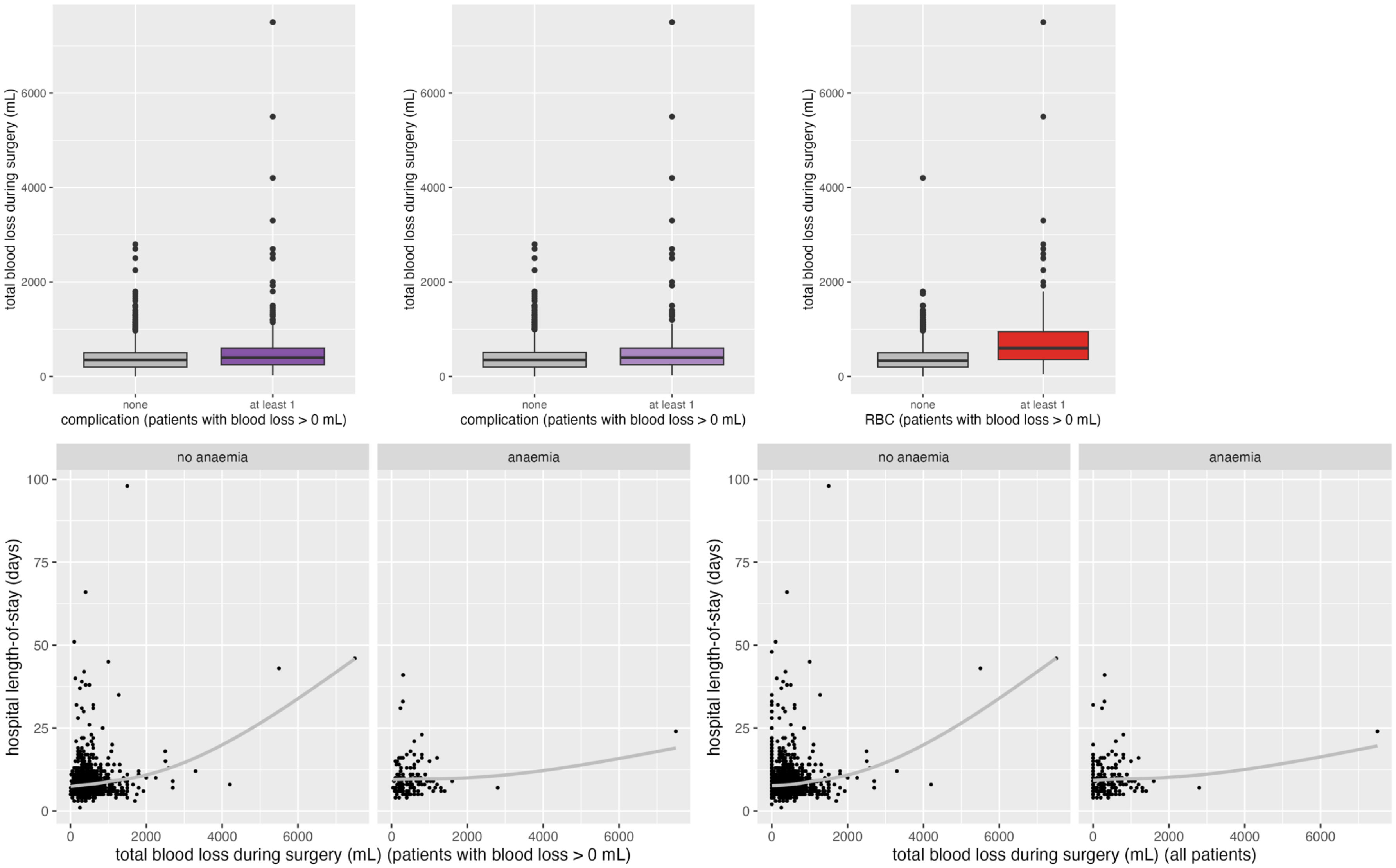


Figure A-9: Stratification for the blood loss (numeric format): Case COM_{PES}, COM_{OPT}, and RBC (top), and LOS (bottom, stratified for pre-operative anaemia).

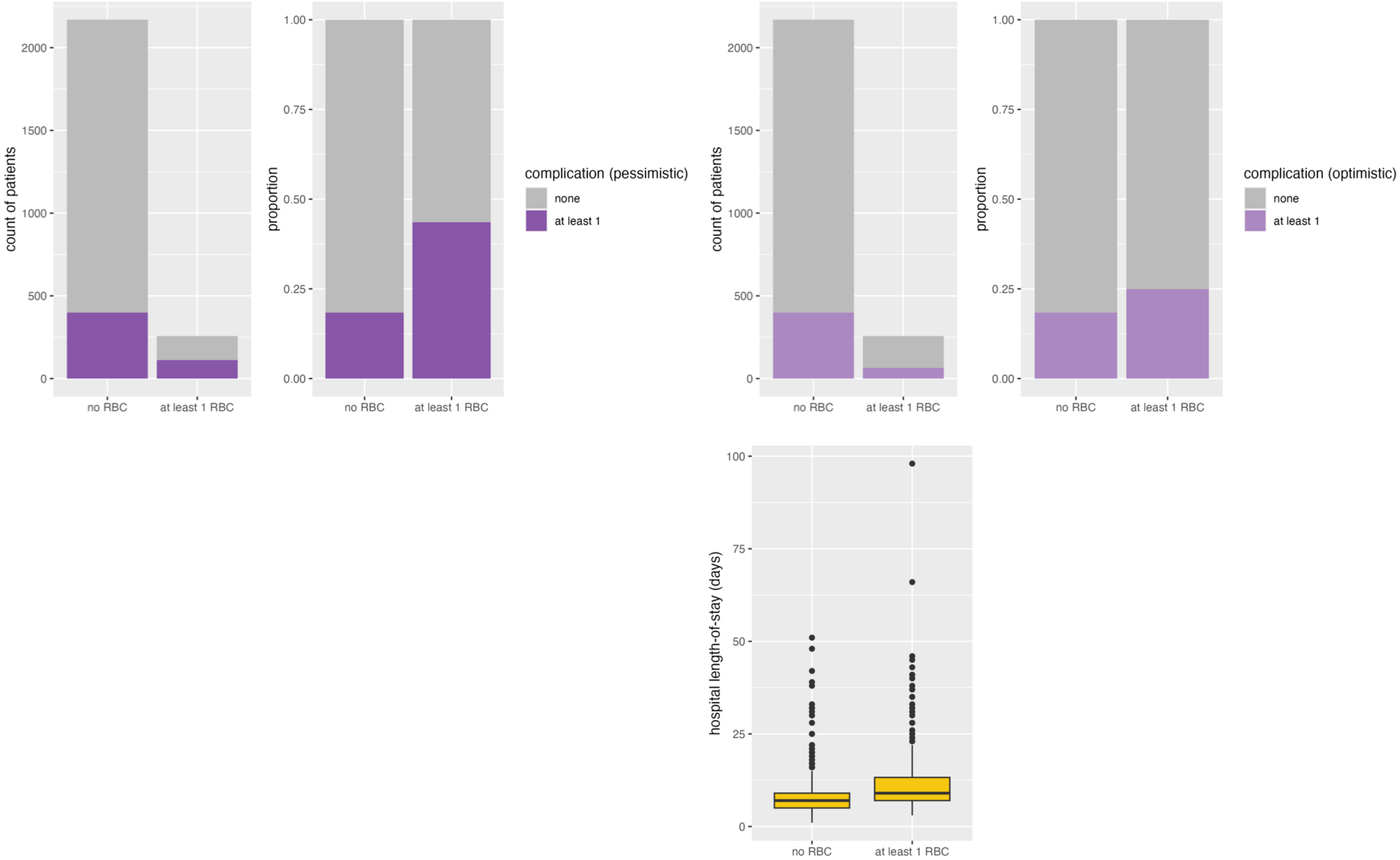


Figure A-10: Stratification for the RBC transfusion: Case COM_{PES} (top left, p<0.001), COM_{OPT} (top right, p<0.001), and LOS (bottom right).

Appendix B: Model Performance Results

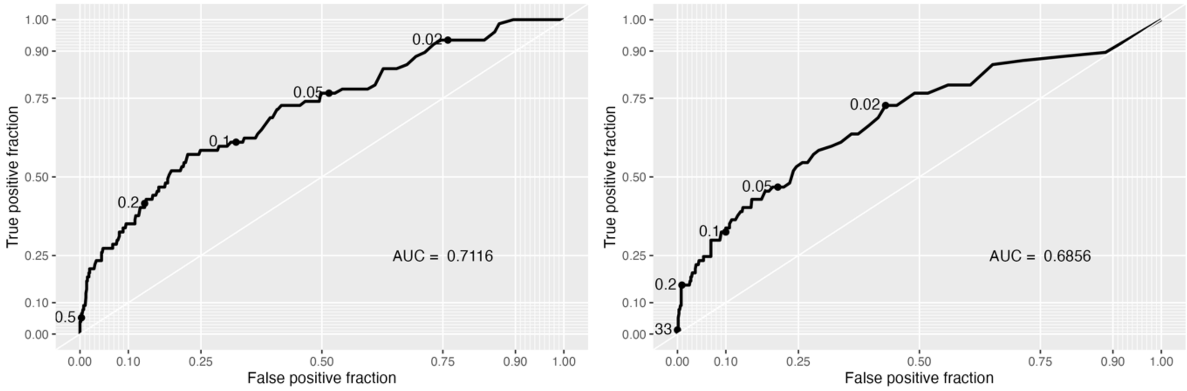


Figure B-1: ROC curves for Case RBC[RF] (left) and RBC[tuned RF] (right).

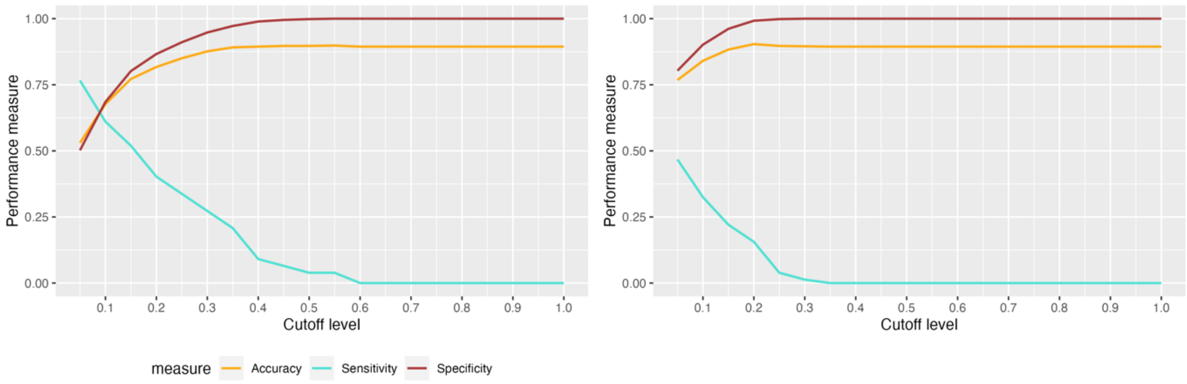


Figure B-2: Performance measures relative to cut-off levels for Case RBC[RF] (left) and RBC[tuned RF] (right).

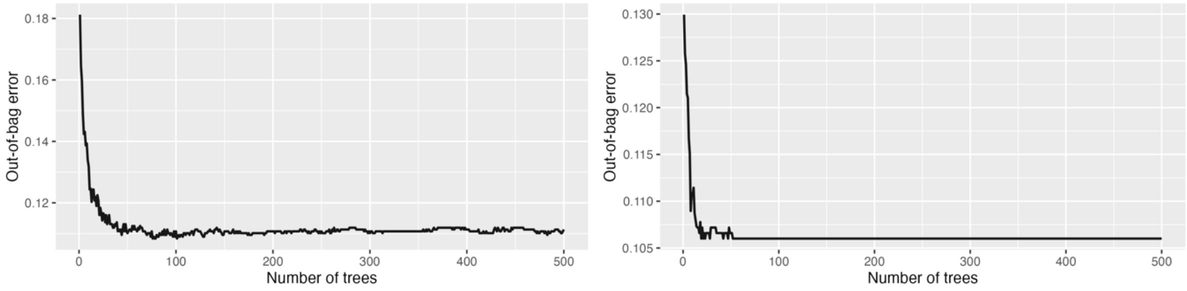


Figure B-3: Out-of-bag error progression for Case RBC[RF] (left) and RBC[tuned RF] (right).

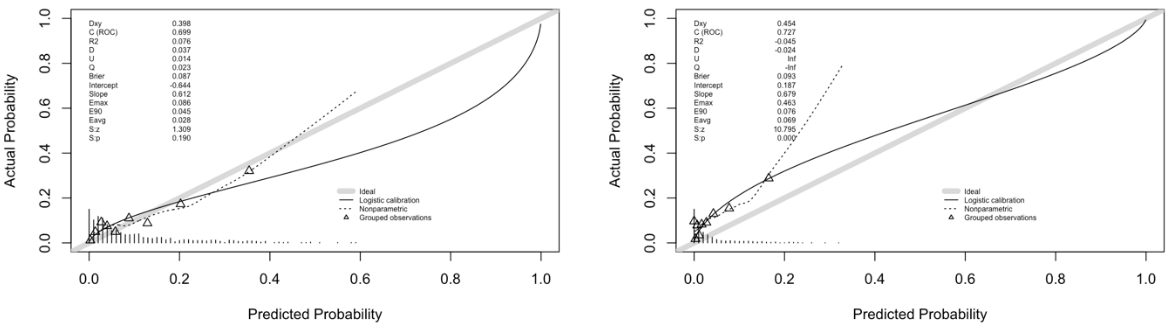


Figure B-4: Calibration plot for Case RBC[RF] (left) and RBC[tuned RF] (right).

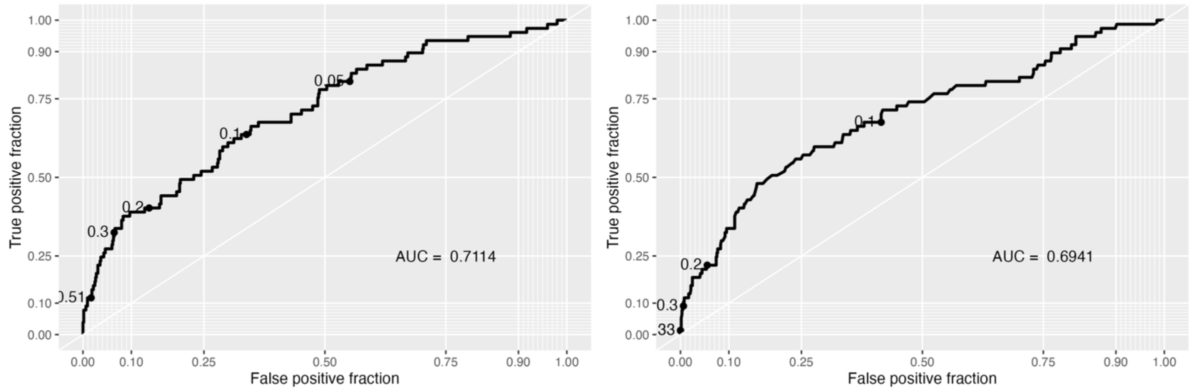


Figure B-5: ROC curves for Case RBC[LREG] (left) and RBC[lasso] (right).

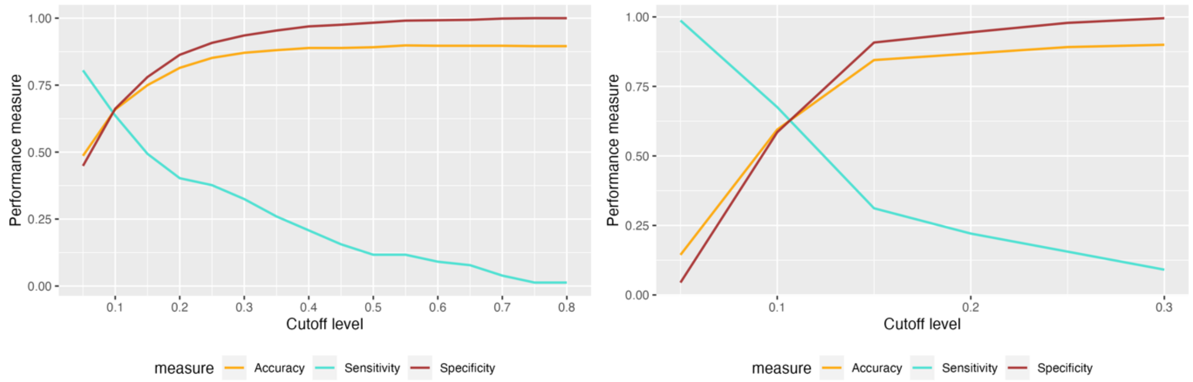


Figure B-6: Performance measures relative to cut-off levels for Case RBC[LREG] (left) and RBC[lasso] (right).

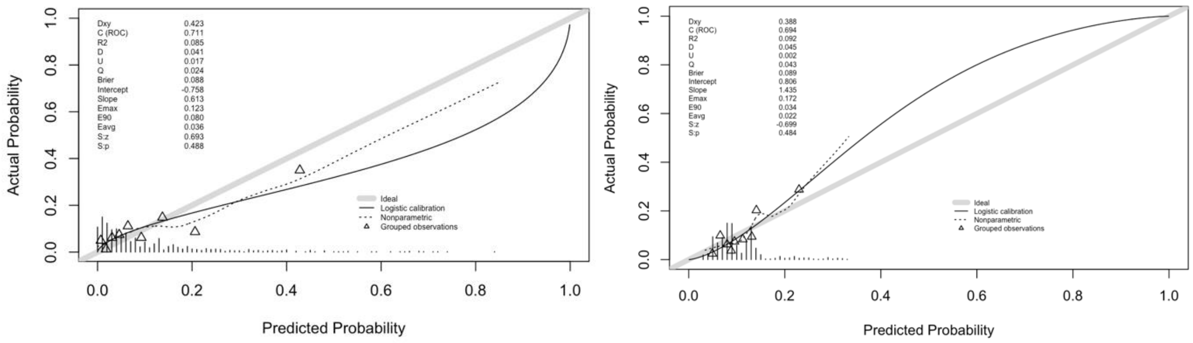


Figure B-7: Calibration plot for Case RBC[LREG] (left) and RBC[lasso] (right).

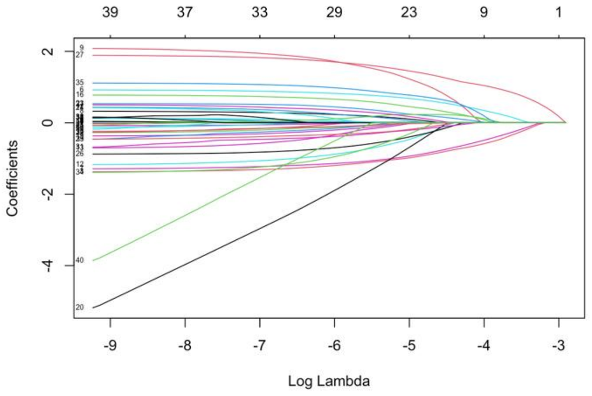


Figure B-8: Convergence of coefficients to zero versus the regularization parameter, log(lambda): RBC[lasso].

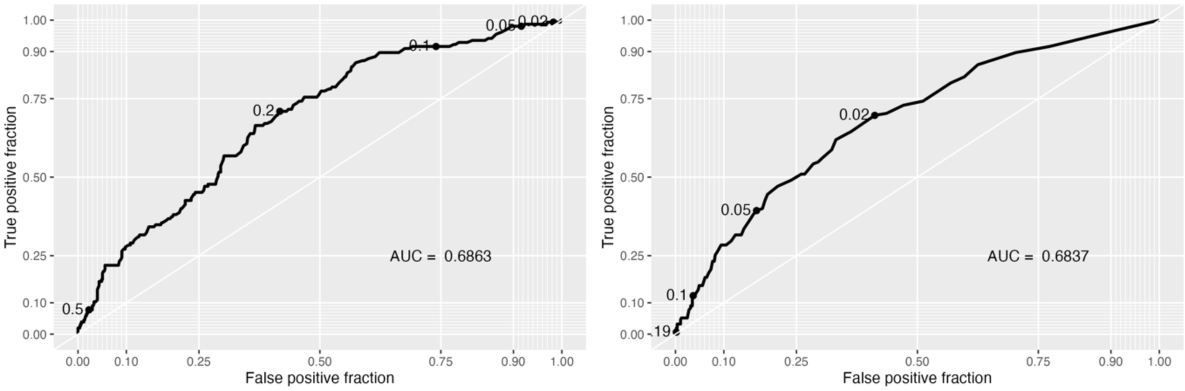


Figure B-9: ROC curves for Case COMPES[RF] (left) and COMPES[tuned RF] (right).

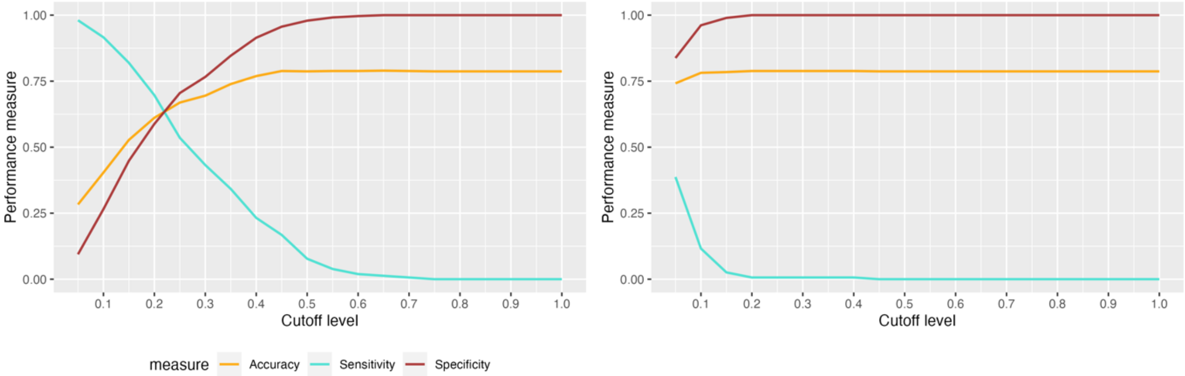


Figure B-10: Performance measures relative to cut-off levels for Case COMPES[RF] (left) and COMPES[tuned RF] (right).

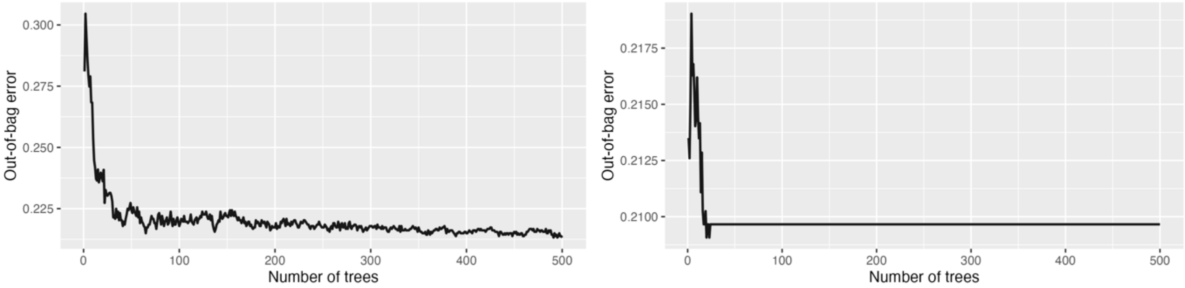


Figure B-11: Out-of-bag error progression for Case COMPES[RF] (left) and COMPES[tuned RF] (right).

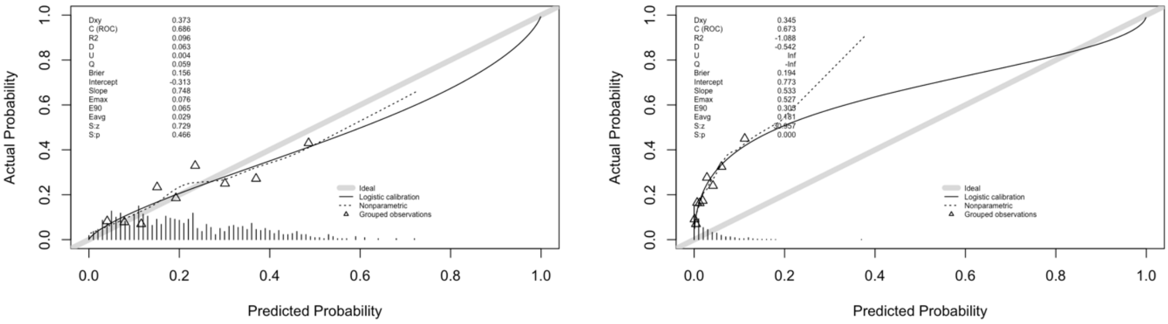


Figure B-12: Calibration plot for Case COMPES[RF] (left) and COMPES[tuned RF] (right).

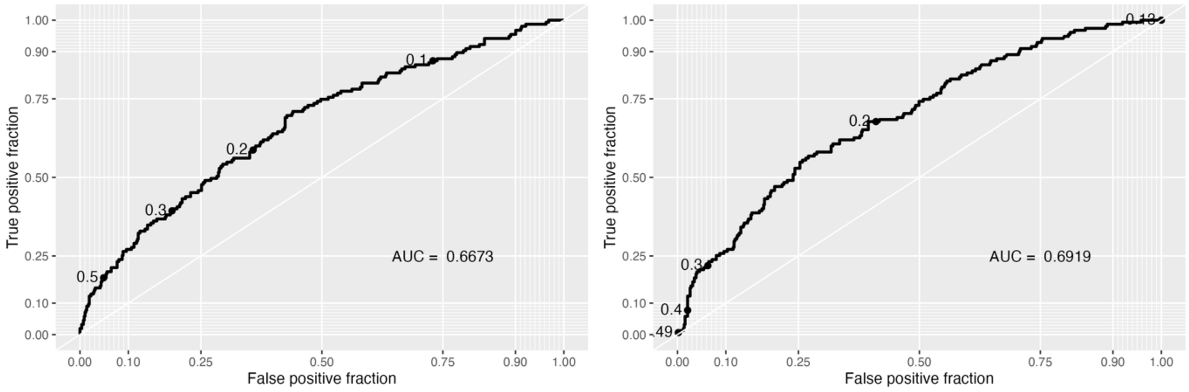


Figure B-13: ROC curves for Case COM$_{PES}$[LREG] (left) and COM$_{PES}$[lasso] (right).

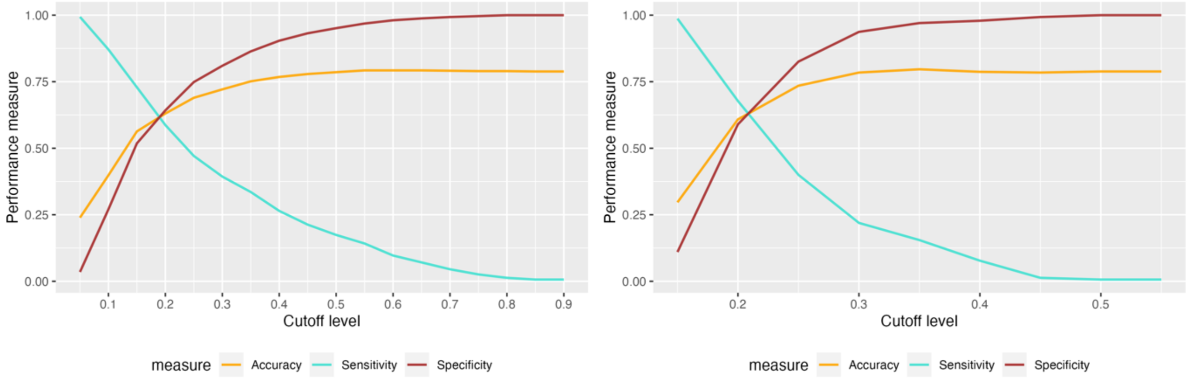


Figure B-14: Performance measures relative to cut-off levels for Case COM$_{PES}$[LREG] (left) and COM$_{PES}$[lasso] (right).

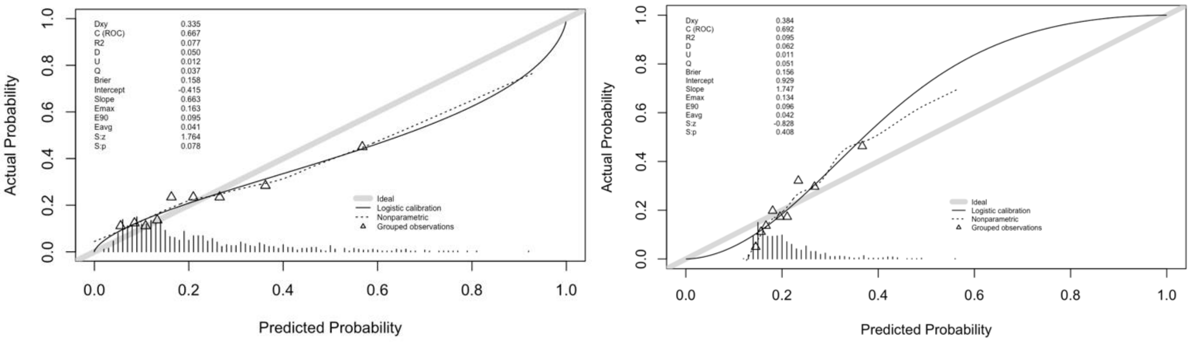


Figure B-15: Calibration plot for Case COM$_{PES}$[LREG] (left) and COM$_{PES}$[lasso] (right).

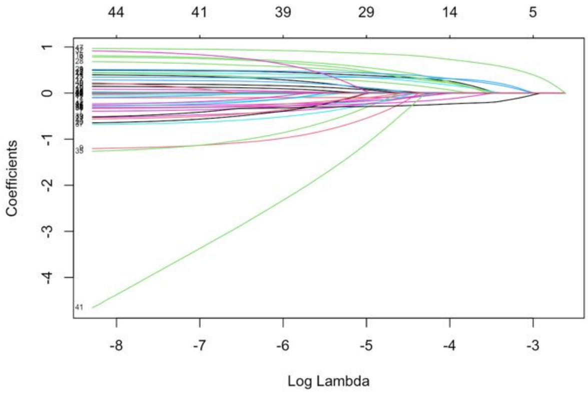


Figure B-16: Convergence of coefficients to zero versus the regularization parameter, log(lambda): COM$_{PES}$[lasso].

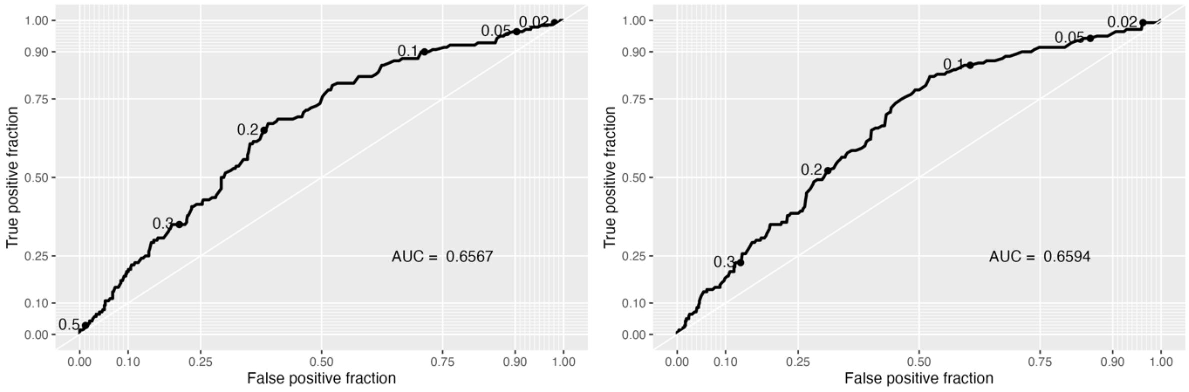


Figure B-17: ROC curves for Case COM$_{\text{OPT}}$[RF] (left) and COM$_{\text{OPT}}$[tuned RF] (right).

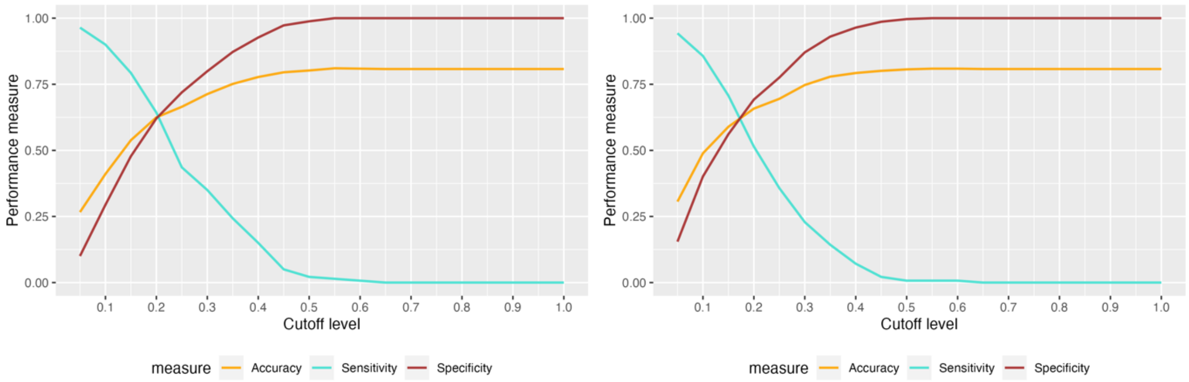


Figure B-18: Performance measures relative to cut-off levels for Case COM$_{\text{OPT}}$[RF] (left) and COM$_{\text{OPT}}$[tuned RF] (right).

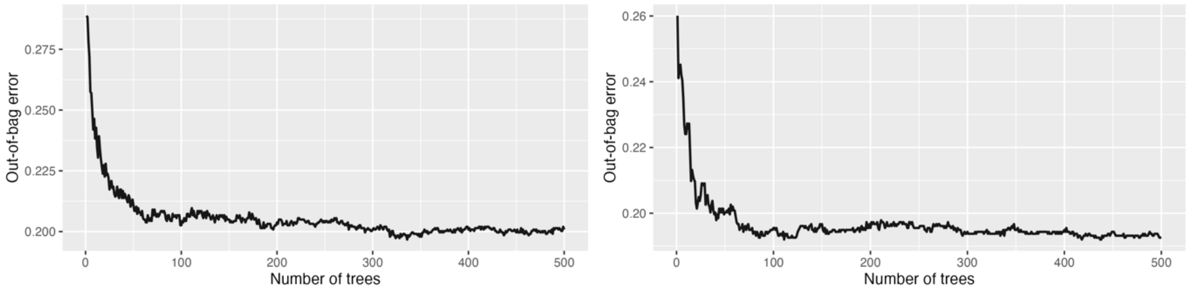


Figure B-19: Out-of-bag error progression for Case COM$_{\text{OPT}}$[RF] (left) and COM$_{\text{OPT}}$[tuned RF] (right).

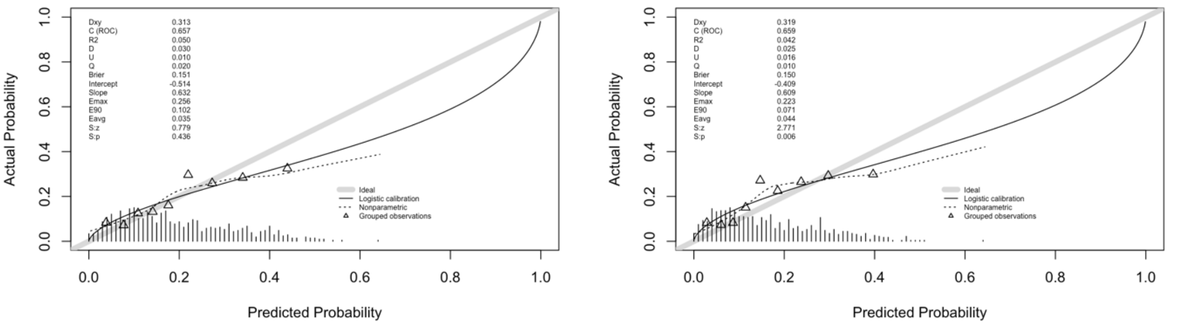


Figure B-20: Calibration plot for Case COM$_{\text{OPT}}$[RF] (left) and COM$_{\text{OPT}}$[tuned RF] (right).

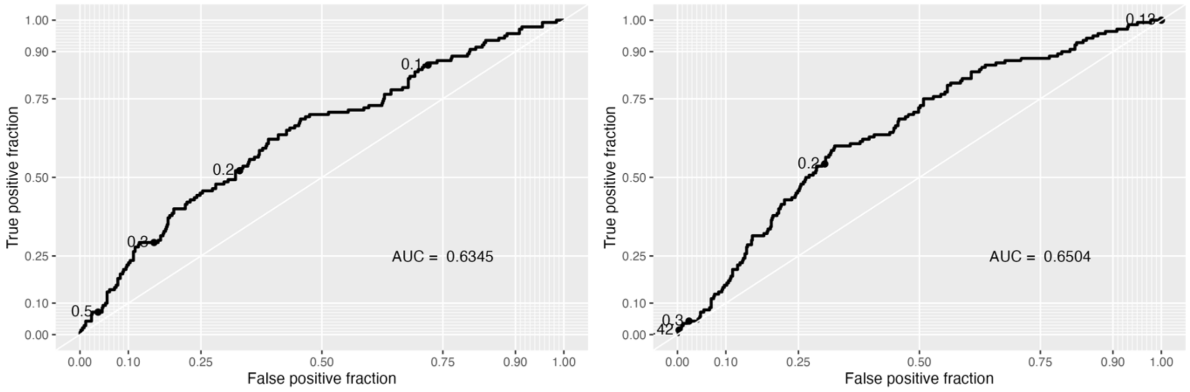


Figure B-21: ROC curves for Case COM$_{OPT}$[LREG] (left) and COM$_{OPT}$[lasso] (right).

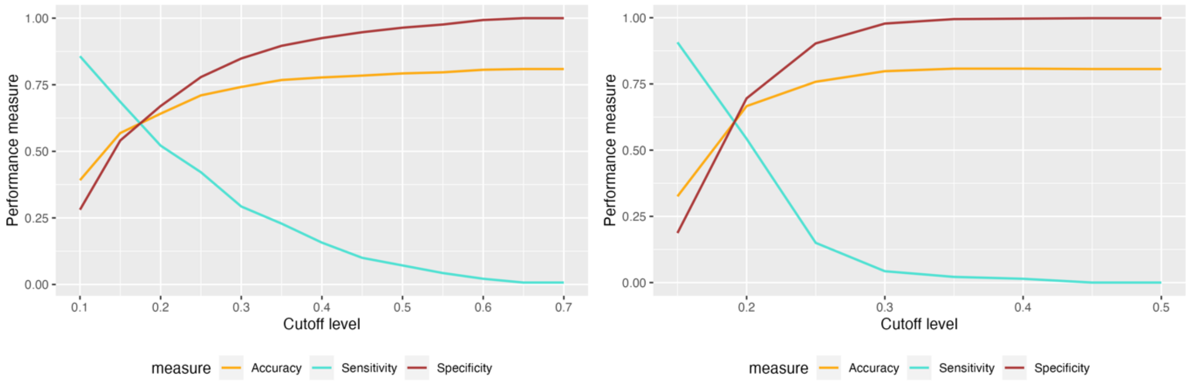


Figure B-22: Performance measures relative to cut-off levels for Case COM$_{OPT}$[LREG] (left) and COM$_{OPT}$[lasso] (right).

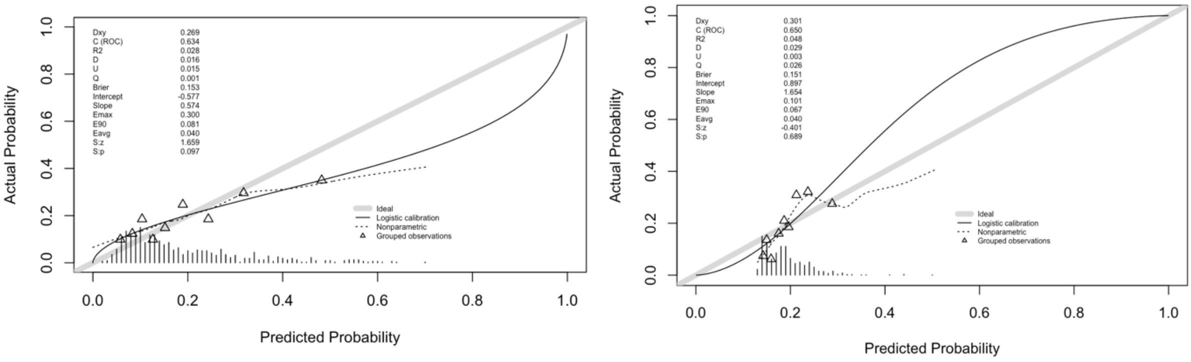


Figure B-23: Calibration plot for Case COM$_{OPT}$[LREG] (left) and COM$_{OPT}$[lasso] (right).

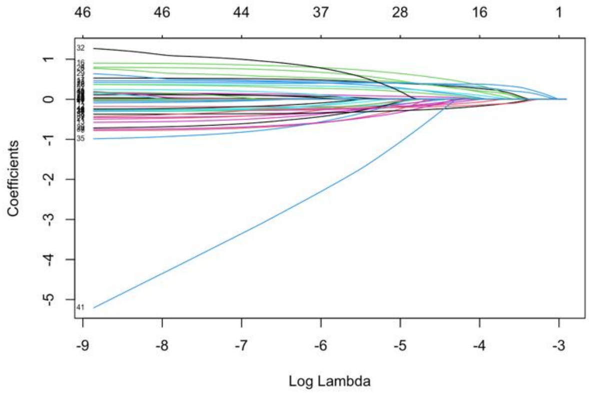


Figure B-24: Convergence of coefficients to zero versus the regularization parameter, log(lambda): COM$_{OPT}$[lasso].

Appendix C: Variable Importance (Supplemental Results)

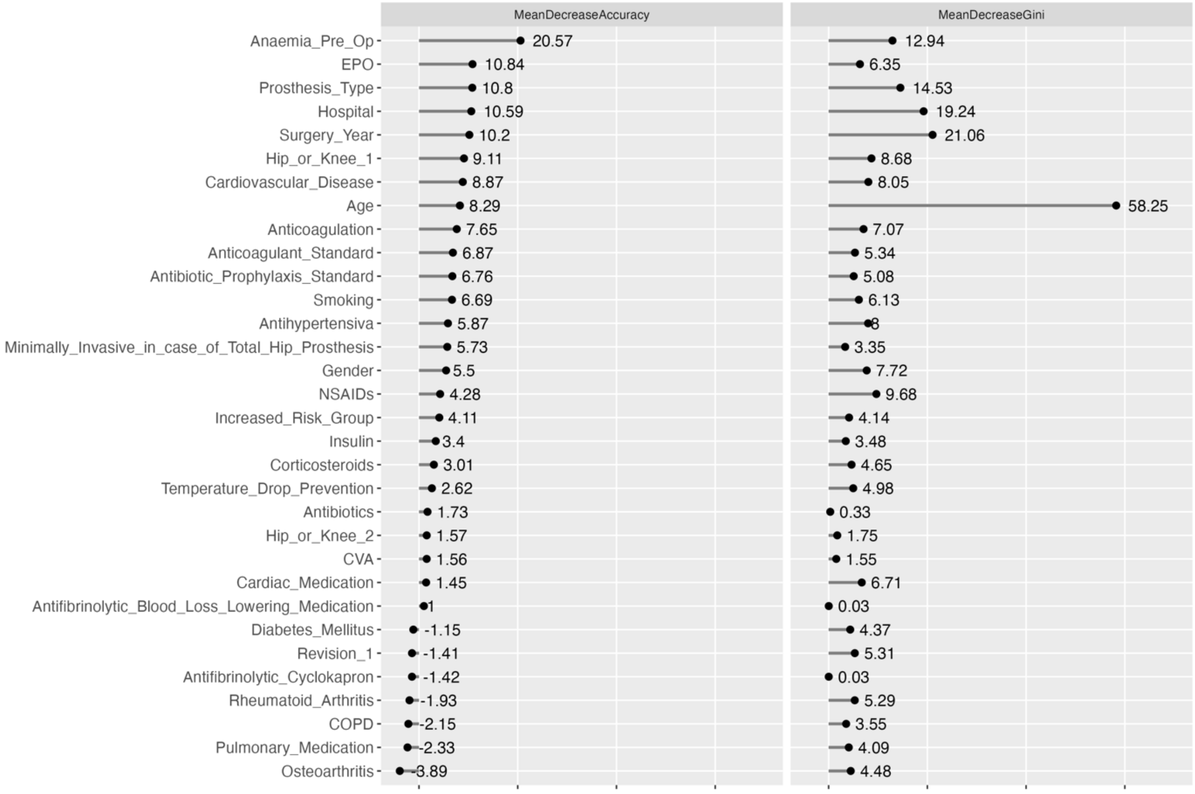


Figure C-1*:* Case RBC[RF]: Variable importance plot for all input variables.

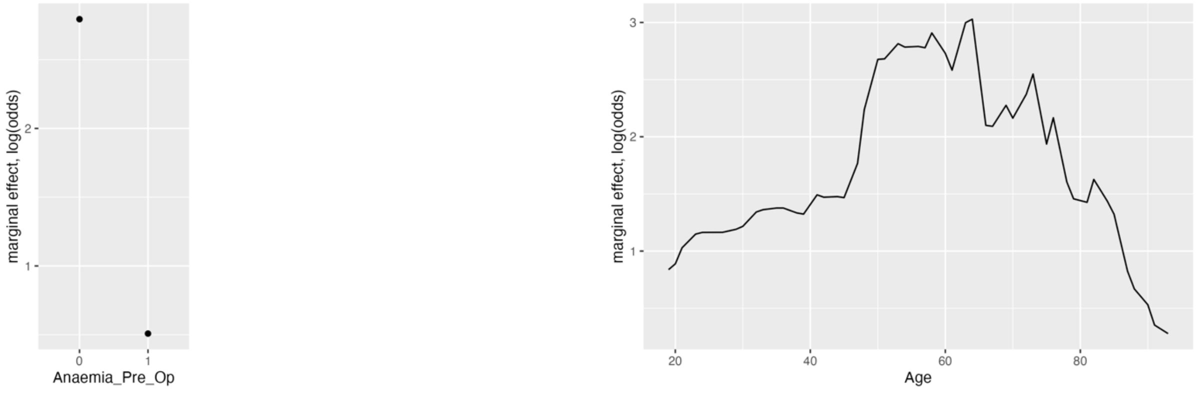


Figure C-2: Case RBC[RF]: Partial dependence plots for pre-operative anaemia (left), and age (right).

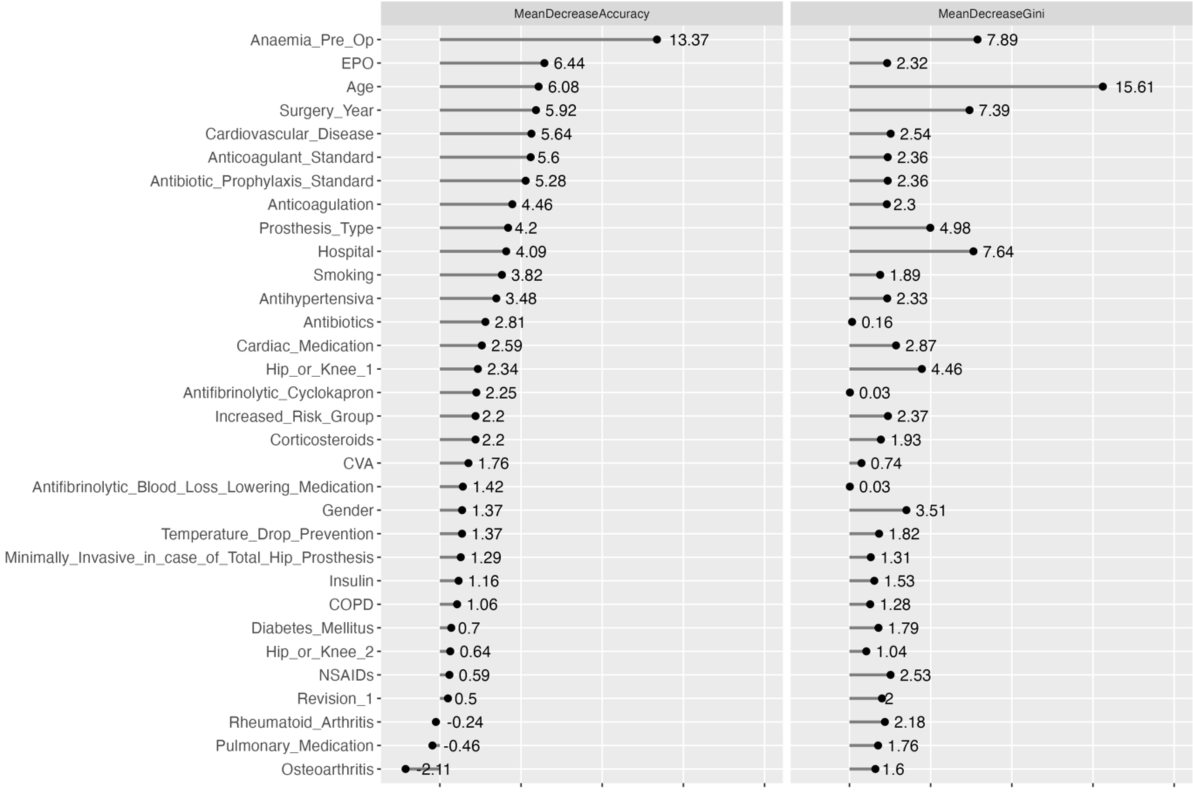


Figure C-3: Case RBC[tuned RF]: Variable importance plot for all input variables.

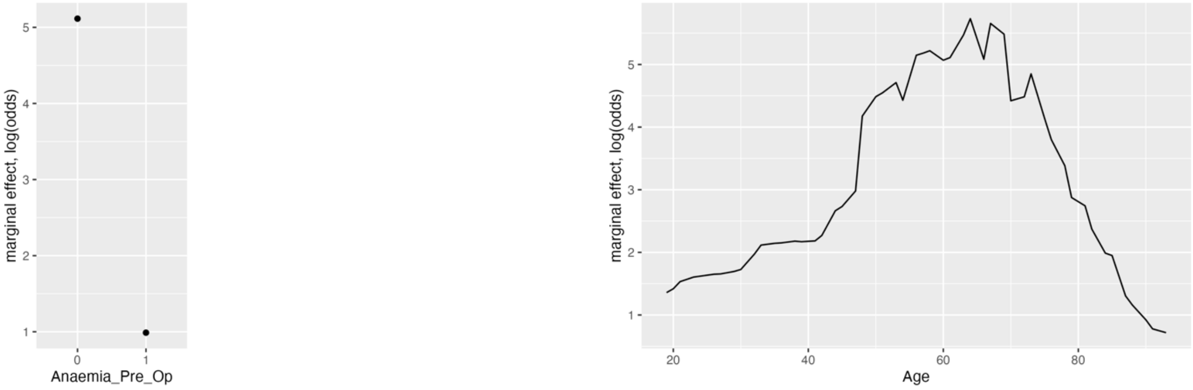


Figure C-4: Case RBC[tuned RF]: Partial dependence plots for pre-operative anaemia (left), and age (right).

Table C-1: LREG[RBC]: p-values and the corresponding statistical significance of all inputs.

Model input component	p-value	Level of statistical significance, α
(Intercept)	<0.001	0.001
HospitalHospital2	0.940	
HospitalHospital3	<0.001	0.001
HospitalHospital4	0.638	
Age	0.008	0.01
Gender2	<0.001	0.001
Hip_or_Knee_11	<0.001	0.001
Revision_11	0.261	
Hip_or_Knee_21	0.015	0.05
Osteoarthritis1	0.851	
Cardiovascular_Disease1	0.576	
CVA1	0.077	0.1
COPD1	0.107	
Diabetes_Mellitus1	0.902	
Rheumatoid_Arthritis1	0.386	
Increased_Risk_Group1	0.038	0.05
Corticosteroids1	0.206	
NSAIDs1	0.499	
Anticoagulation1	0.151	
Antibiotics1	0.980	
Insulin1	0.942	
Antihypertensiva1	0.244	
Cardiac_Medication1	0.041	0.05
Pulmonary_Medication1	0.273	
Smoking1	0.768	
EPO1	0.007	0.01
Anaemia_Pre_Op1	<0.001	0.001
Surgery_Year2005	0.777	
Surgery_Year2006	0.298	
Surgery_Year2007	0.633	
Surgery_Year2008	0.173	
Surgery_Year2009	0.954	
Prosthesis_Type2	0.044	0.05
Prosthesis_Type3	0.198	
Prosthesis_Typeunknown	0.018	0.05
Minimally_Invasive_in_case_of_Total_Hip_Prosthesis1	0.710	
Temperature_Drop_Prevention1	0.093	0.1
Anticoagulant_Standard1	0.657	
Antibiotic_Prophylaxis_Standard1	0.548	
Antifibrinolytic_Blood_Loss_Lowering_Medication1	0.989	

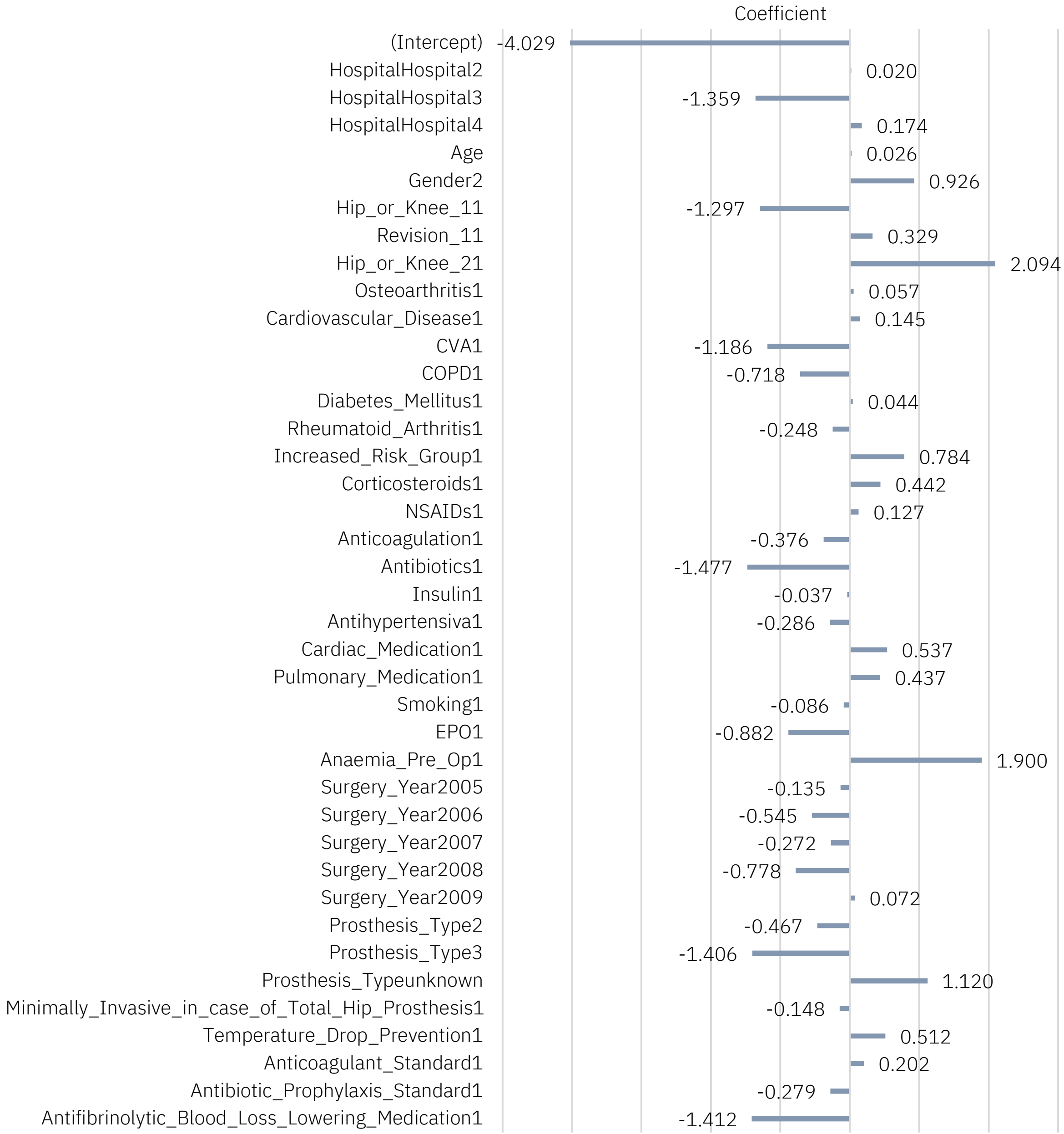


Figure C-5: Case RBC[LREG]: Coefficients of all inputs.

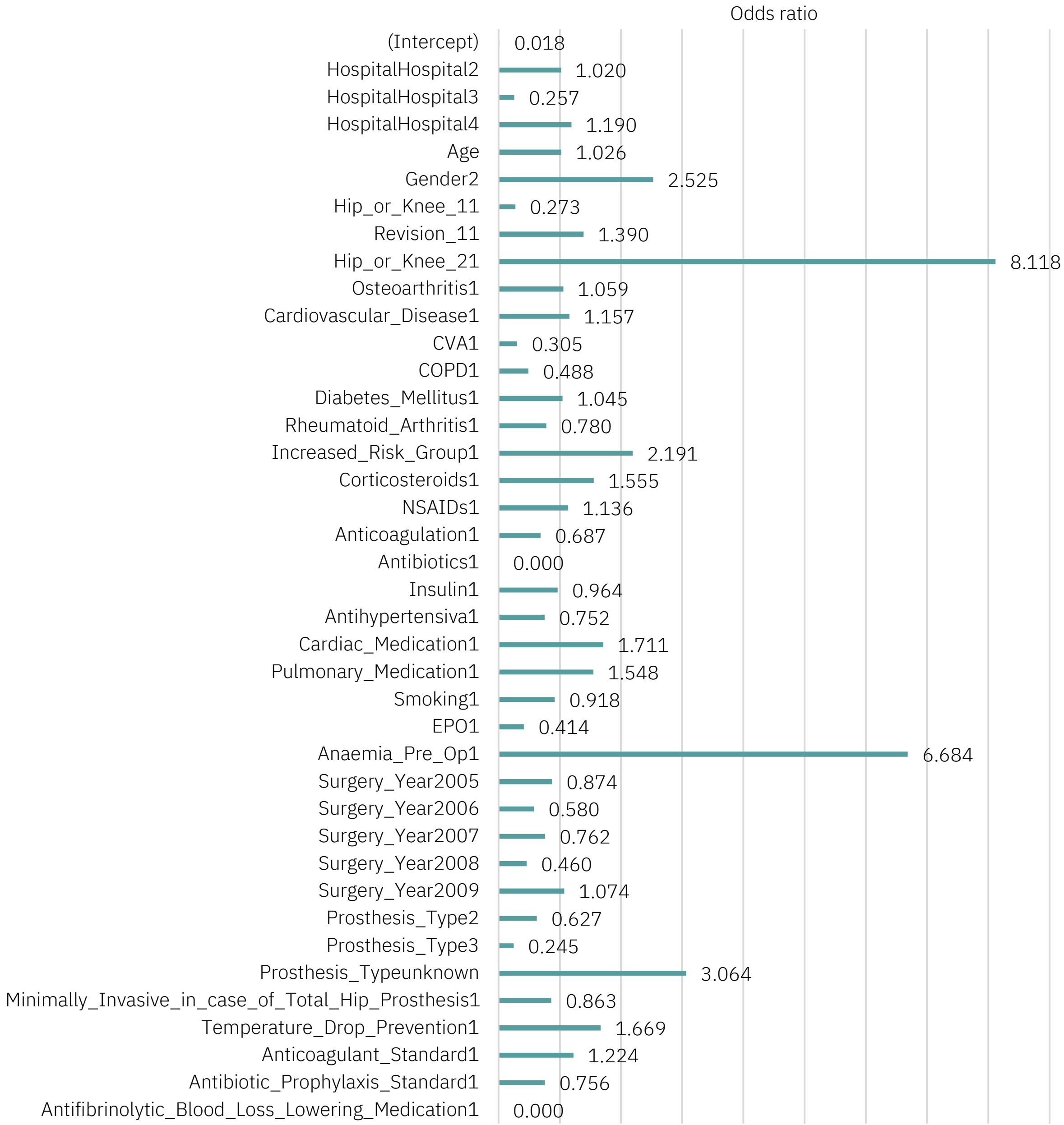


Figure C-6: Case RBC[LREG]: Odds ratios of all inputs.

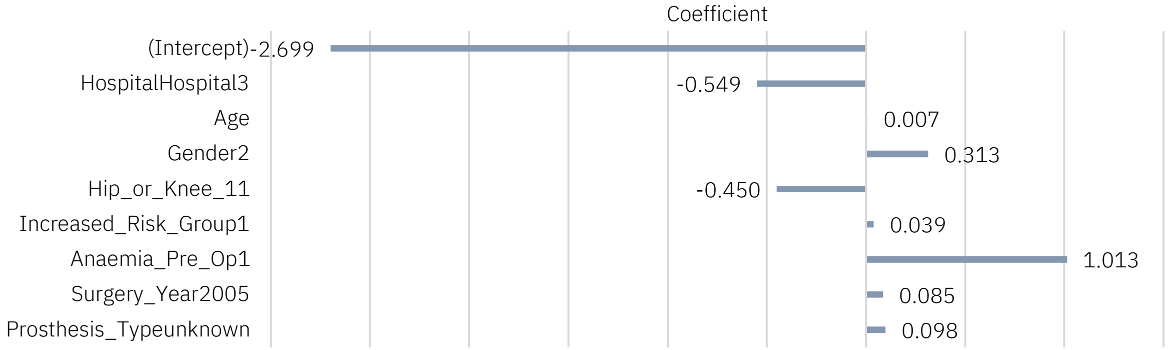


Figure C-7: Case RBC[lasso]: Non-zero coefficients of the inputs.

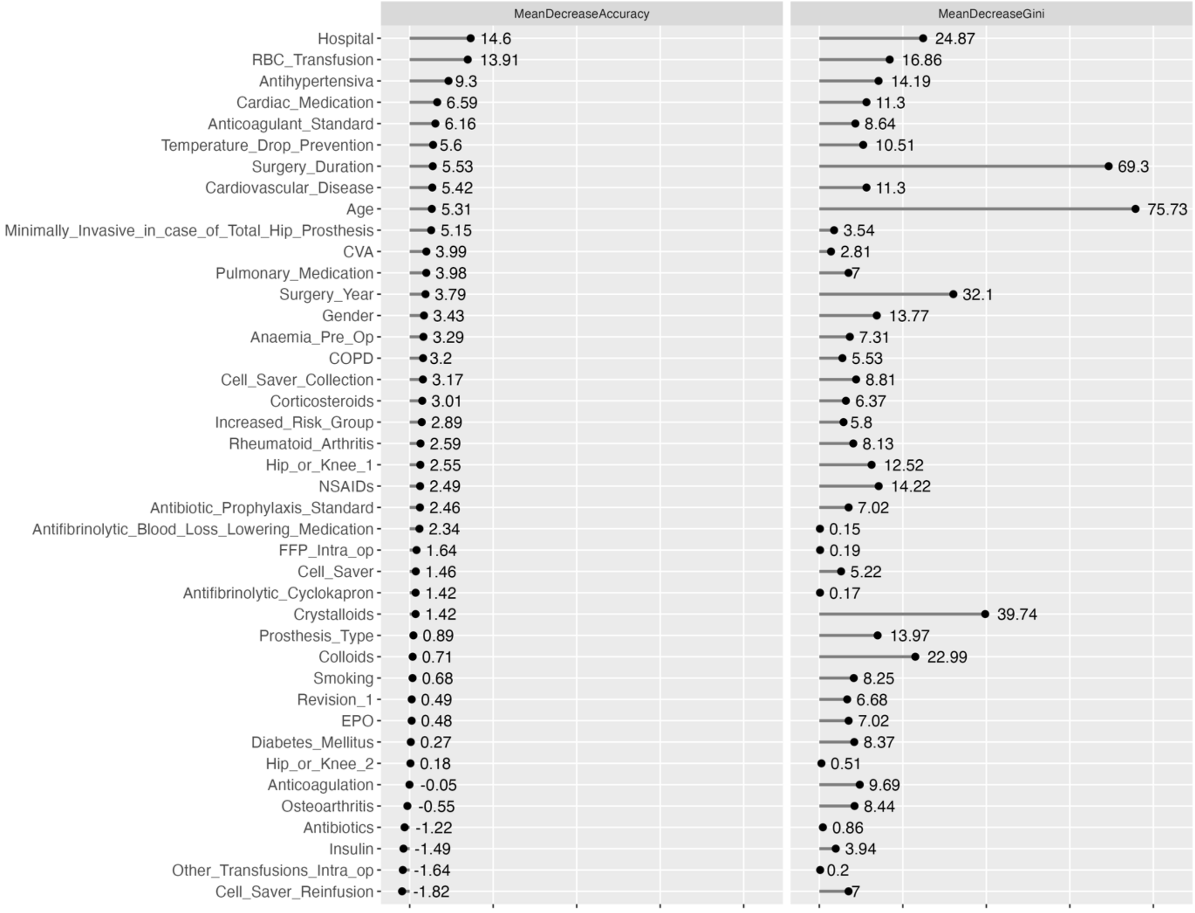


Figure C-8: Case COM$_{PES}$[RF]: Variable importance plot for all input variables.

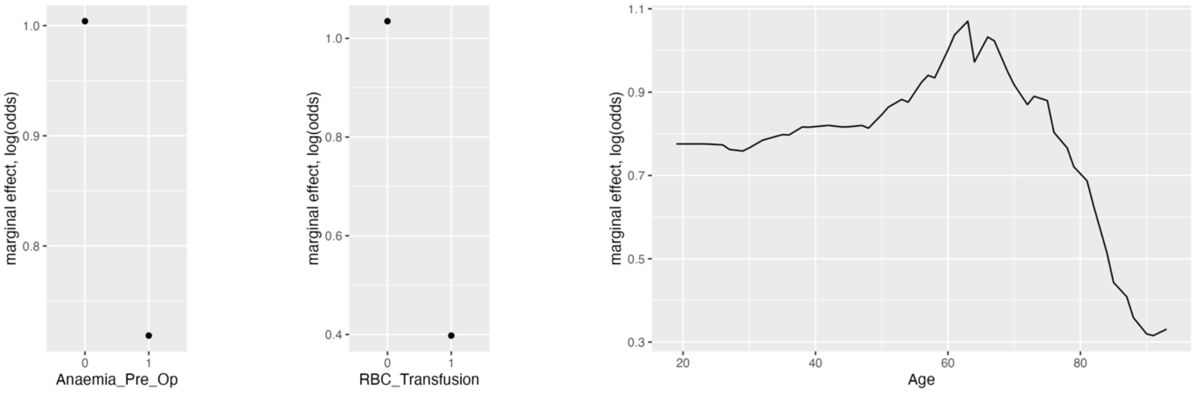


Figure C-9: Case COM$_{PES}$[RF]: Partial dependence plots for pre-operative anaemia (left), RBC transfusion (middle), and age (right).

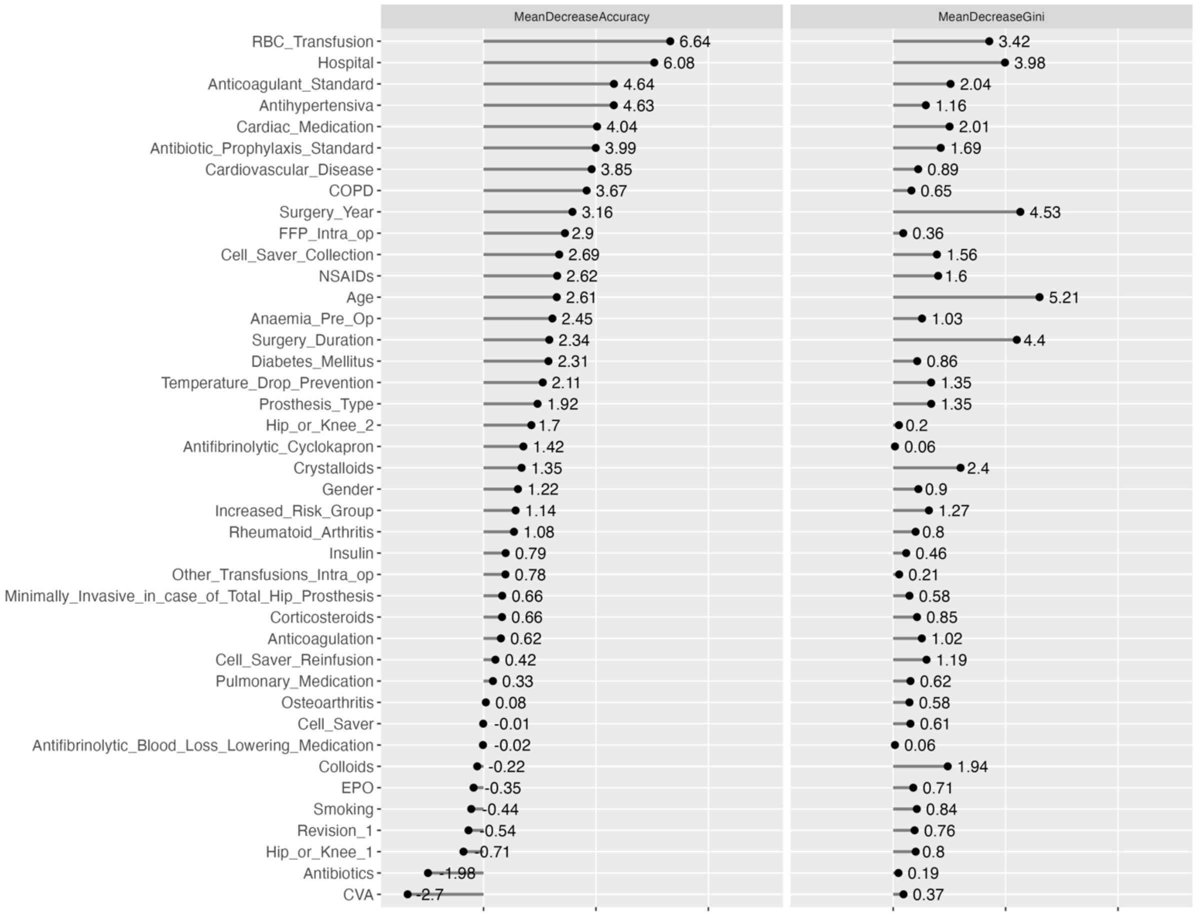


Figure C-10: Case COM_{PES}[tuned RF]: Variable importance plot for all input variables.

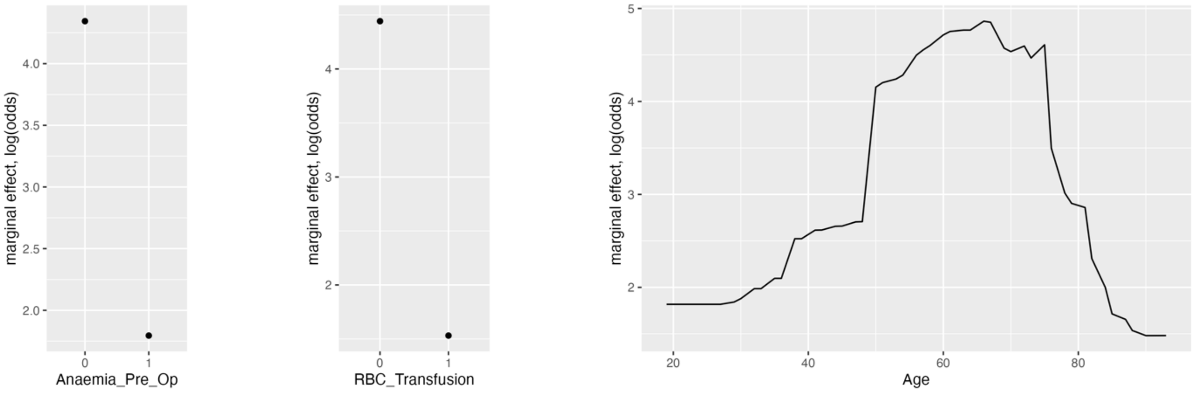


Figure C-11: Case COM$_{PES}$[tuned RF]: Partial dependence plots for pre-operative anaemia (left), RBC transfusion (middle), and age (right).

Table C-2: Case COM$_{\text{PES}}$[LREG]: p-values and the corresponding statistical significance of all inputs.

Model input component	p-value	Level of statistical significance, α
(Intercept)	0.002	0.01
HospitalHospital2	0.009	0.01
HospitalHospital3	0.063	0.1
HospitalHospital4	0.287	
Age	0.002	0.01
Gender2	0.022	0.05
Hip_or_Knee_11	0.012	0.05
Revision_11	0.397	
Hip_or_Knee_21	0.205	
Osteoarthritis1	0.938	
Cardiovascular_Disease1	0.570	
CVA1	0.531	
COPD1	0.103	
Diabetes_Mellitus1	0.092	0.1
Rheumatoid_Arthritis1	0.963	
Increased_Risk_Group1	0.009	0.01
Corticosteroids1	0.279	
NSAIDs1	0.001	0.01
Anticoagulation1	0.639	
Antibiotics1	0.415	
Insulin1	0.127	
Antihypertensiva1	0.014	0.05
Cardiac_Medication1	0.072	0.1
Pulmonary_Medication1	0.880	
Smoking1	0.206	
EPO1	0.507	
Anaemia_Pre_Op1	0.873	
Surgery_Year2005	0.101	
Surgery_Year2006	0.242	
Surgery_Year2007	0.967	
Surgery_Year2008	0.944	
Surgery_Year2009	0.319	
Surgery_Duration	0.142	
Prosthesis_Type2	0.055	0.1
Prosthesis_Type3	0.221	
Prosthesis_Typeunknown	0.509	
Minimally_Invasive_in_case_of_Total_Hip_Prosthesis1	0.038	0.05
Temperature_Drop_Prevention1	0.101	
Anticoagulant_Standard1	0.109	
Antibiotic_Prophylaxis_Standard1	0.459	
Antifibrinolytic_Blood_Loss_Lowering_Medication1	0.971	
Colloids	0.181	
Crystalloids	0.788	
Cell_Saver1	0.345	
Cell_Saver_Collection	0.575	
Cell_Saver_Reinfusion	0.488	
RBC_Transfusion1	<0.001	0.001

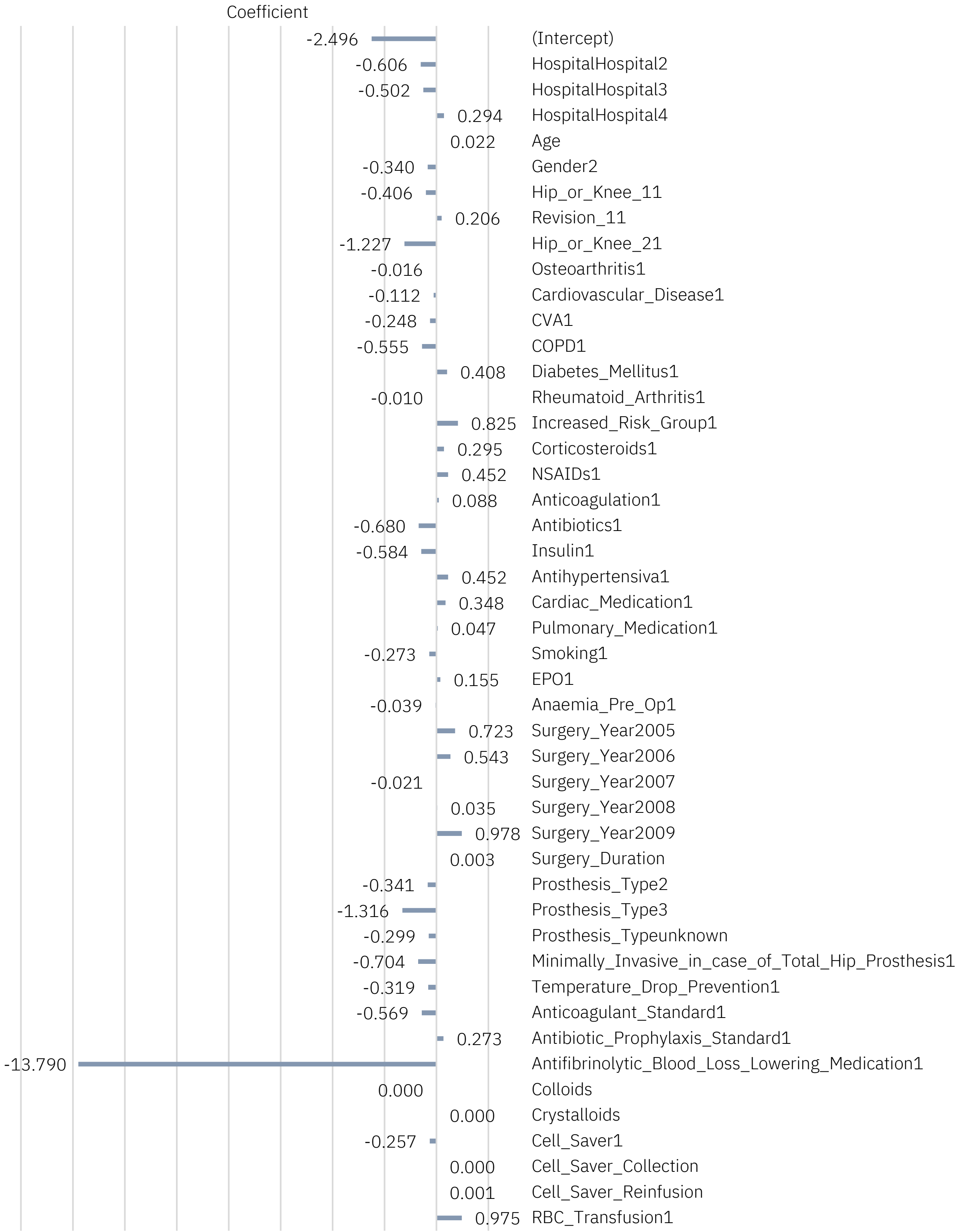


Figure C-12: Case COM$_{PES}$[LREG]: Coefficients of all inputs.

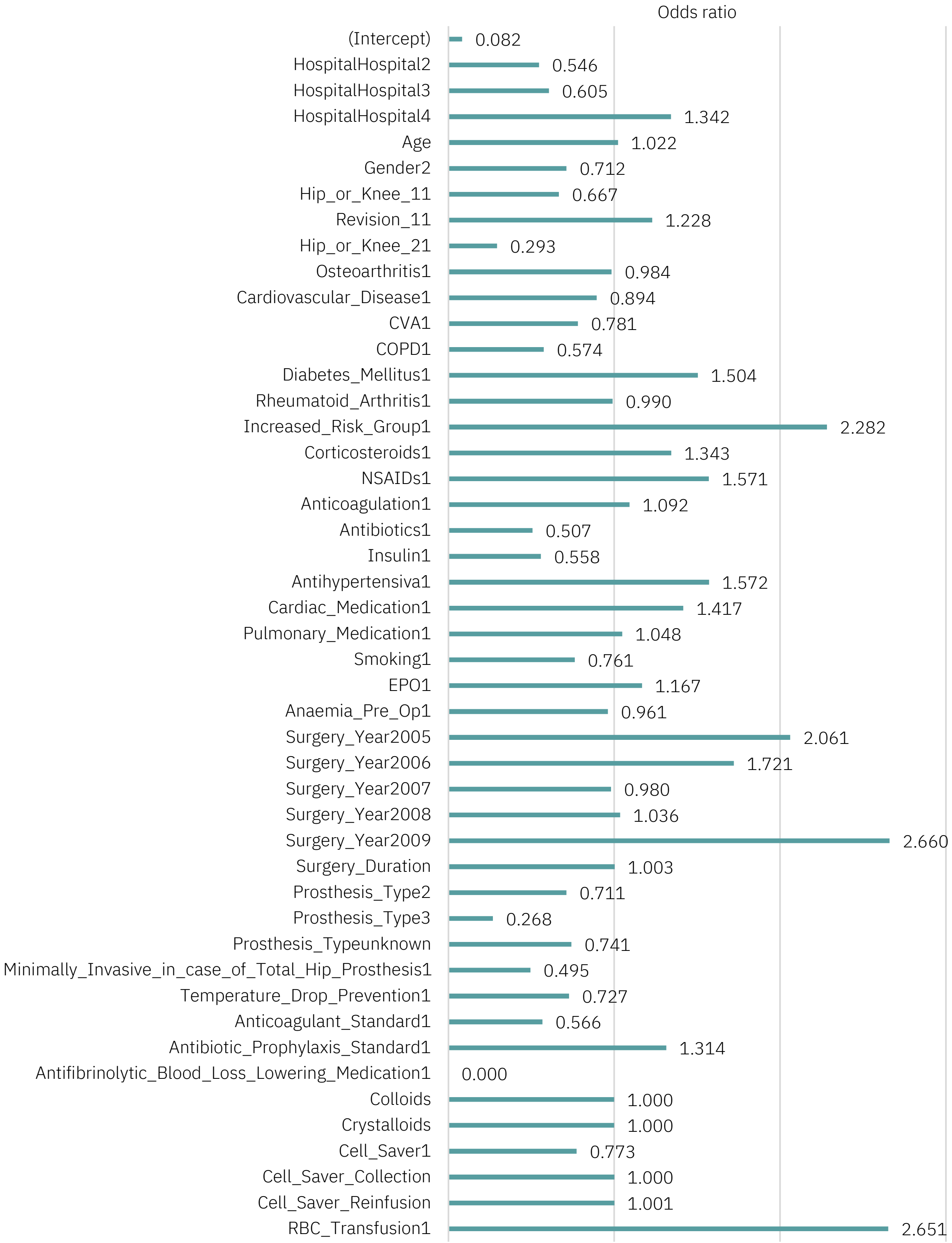


Figure C-13: Case COM$_{PES}$[LREG]: Odds ratios of all inputs.

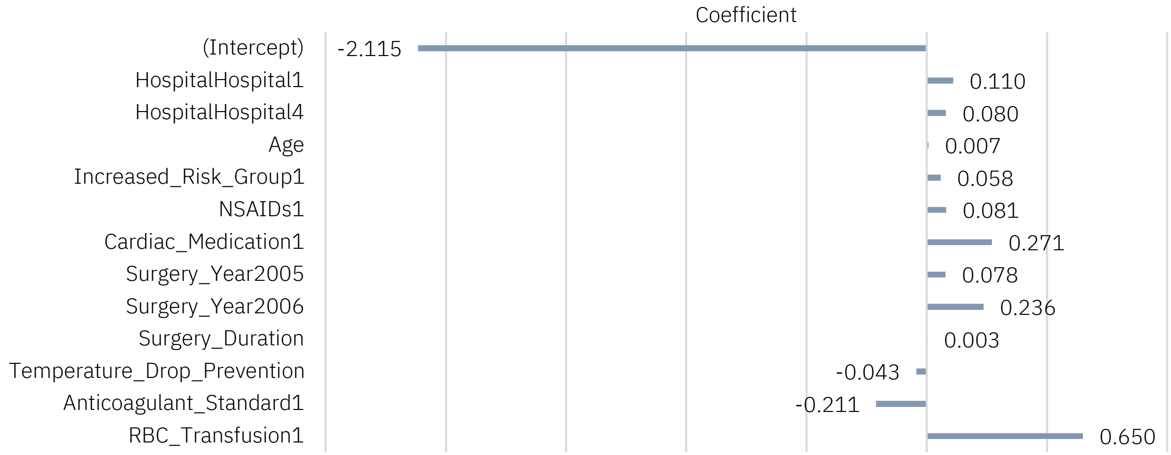


Figure C-14: Case COM$_{PES}$[lasso]: Non-zero coefficients of the inputs.

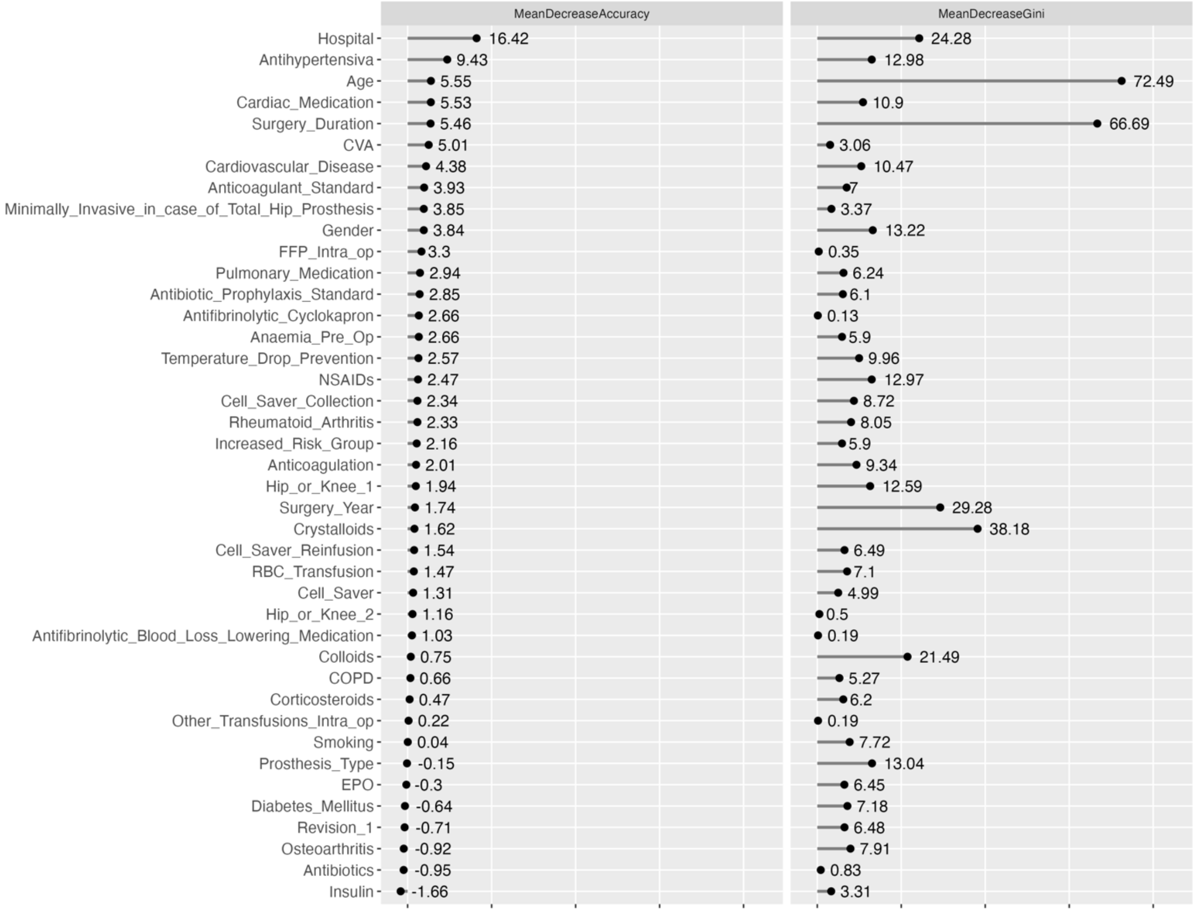


Figure C-15: Case COM$_{OPT}$[RF]: Variable importance plot for all input variables.

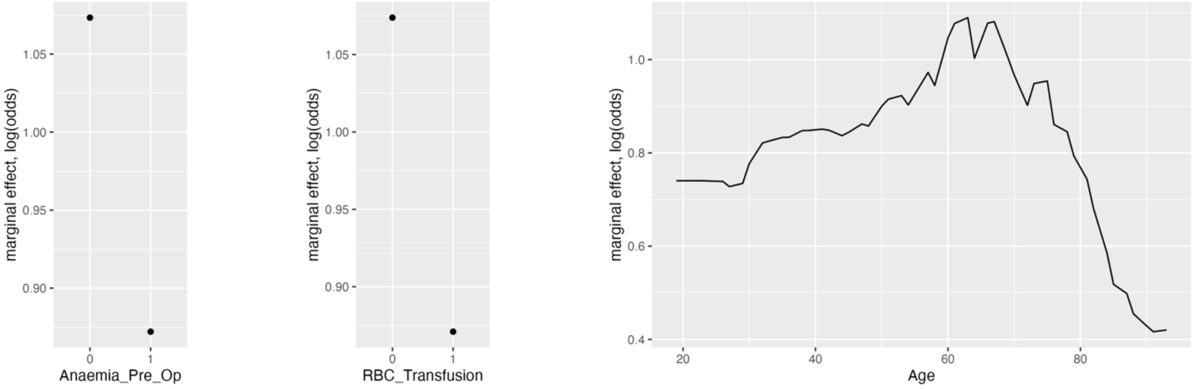


Figure C-16: Case COM$_{OPT}$[RF]: Partial dependence plots for pre-operative anaemia (left), RBC transfusion (middle), and age (right).

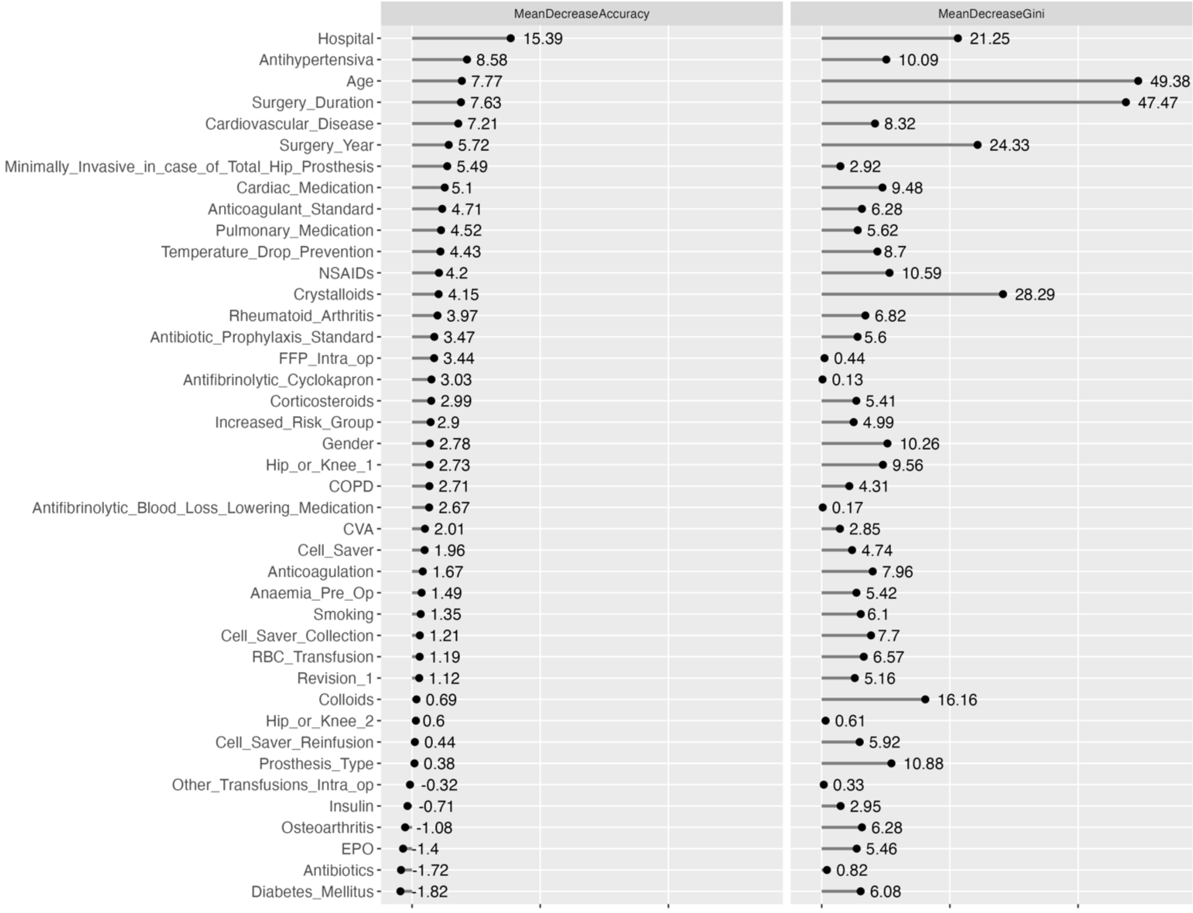


Figure C-17: Case COM$_{OPT}$[tuned RF]: Variable importance plot for all input variables.

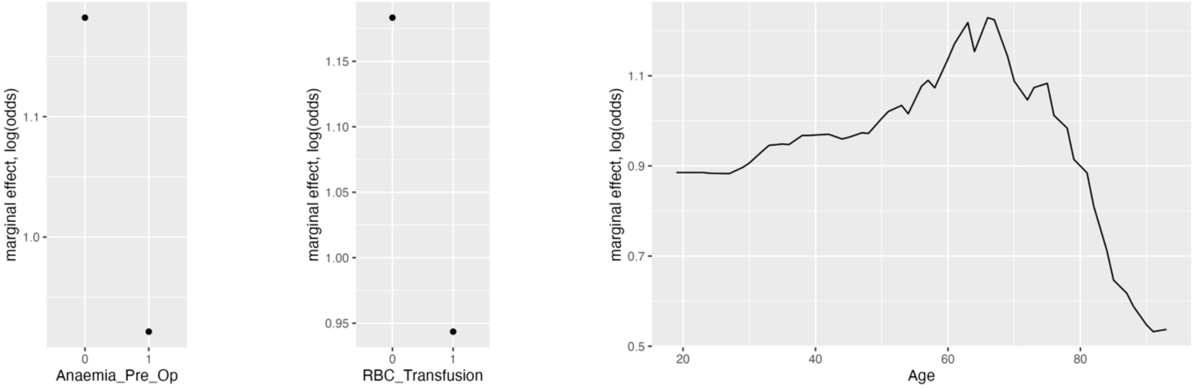


Figure C-18: Case COM$_{OPT}$[tuned RF]: Partial dependence plots for pre-operative anaemia (left), RBC transfusion (middle), and age (right).

Table C-3: Case COM$_{OPT}$[LREG]: p-values and the corresponding statistical significance of all inputs.

Model input component	p-value	Level of statistical significance, α
(Intercept)	0.001	0.01
HospitalHospital2	0.003	0.01
HospitalHospital3	0.050	0.05
HospitalHospital4	0.333	
Age	0.005	0.01
Gender2	0.070	0.1
Hip_or_Knee_11	0.005	0.01
Revision_11	0.664	
Hip_or_Knee_21	0.401	
Osteoarthritis1	0.885	
Cardiovascular_Disease1	0.607	
CVA1	0.488	
COPD1	0.083	0.1
Diabetes_Mellitus1	0.531	
Rheumatoid_Arthritis1	0.897	
Increased_Risk_Group1	0.004	0.01
Corticosteroids1	0.082	0.1
NSAIDs1	0.009	0.01
Anticoagulation1	0.399	
Antibiotics1	0.375	
Insulin1	0.198	
Antihypertensiva1	0.044	0.05
Cardiac_Medication1	0.030	0.05
Pulmonary_Medication1	0.414	
Smoking1	0.187	
EPO1	0.859	
Anaemia_Pre_Op1	0.952	
Surgery_Year2005	0.072	0.1
Surgery_Year2006	0.138	
Surgery_Year2007	0.592	
Surgery_Year2008	0.540	
Surgery_Year2009	0.151	
Surgery_Duration	0.088	0.1
Prosthesis_Type2	0.089	0.1
Prosthesis_Type3	0.336	
Prosthesis_Typeunknown	0.515	
Minimally_Invasive_in_case_of_Total_Hip_Prosthesis1	0.028	0.05
Temperature_Drop_Prevention1	0.054	0.1
Anticoagulant_Standard1	0.192	
Antibiotic_Prophylaxis_Standard1	0.608	
Antifibrinolytic_Blood_Loss_Lowering_Medication1	0.971	
Colloids	0.260	
Crystalloids	0.513	
Cell_Saver1	0.363	
Cell_Saver_Collection	0.864	
Cell_Saver_Reinfusion	0.311	
RBC_Transfusion1	0.799	

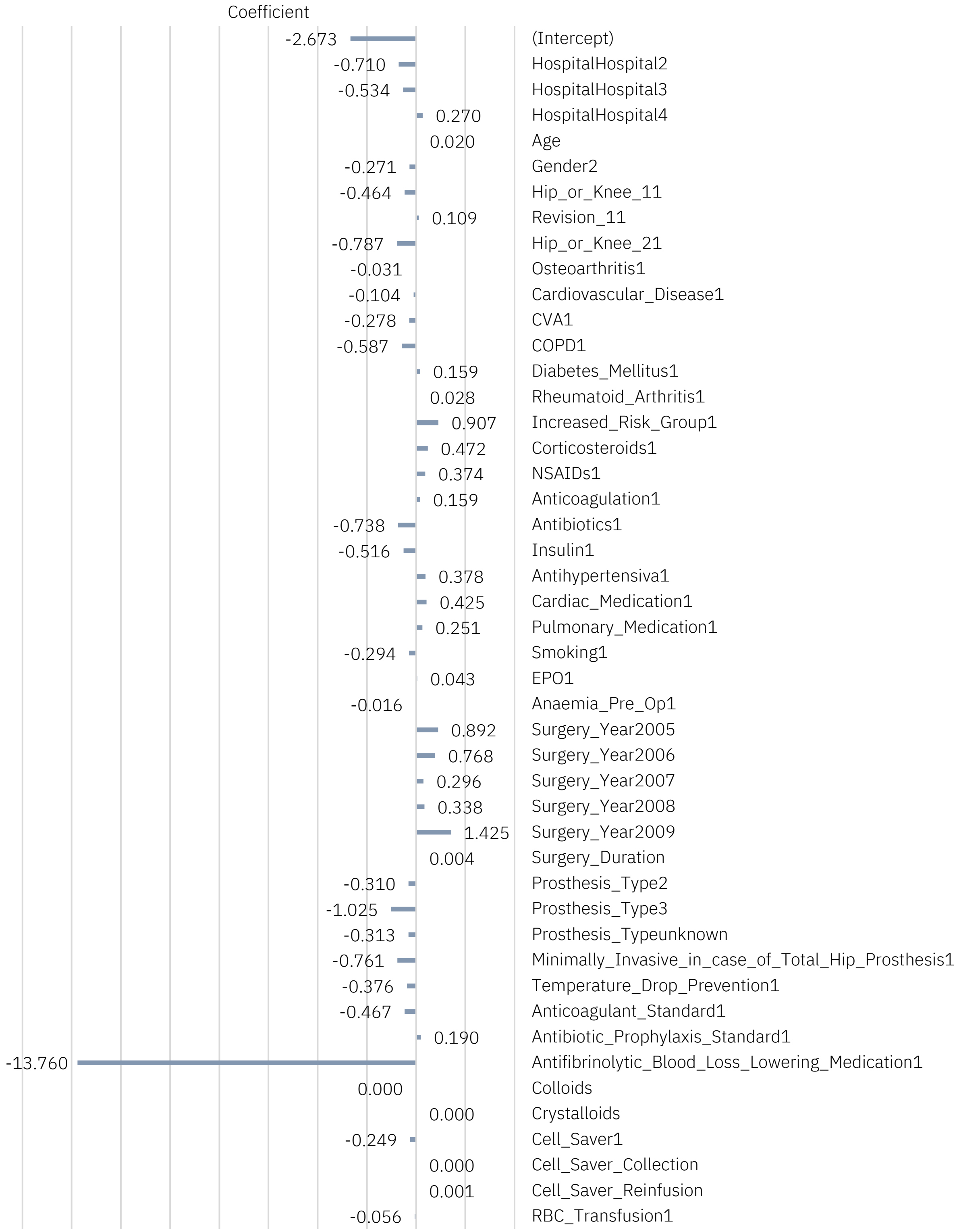


Figure C-19: Case COM$_{OPT}$[LREG]: Coefficients of all inputs.

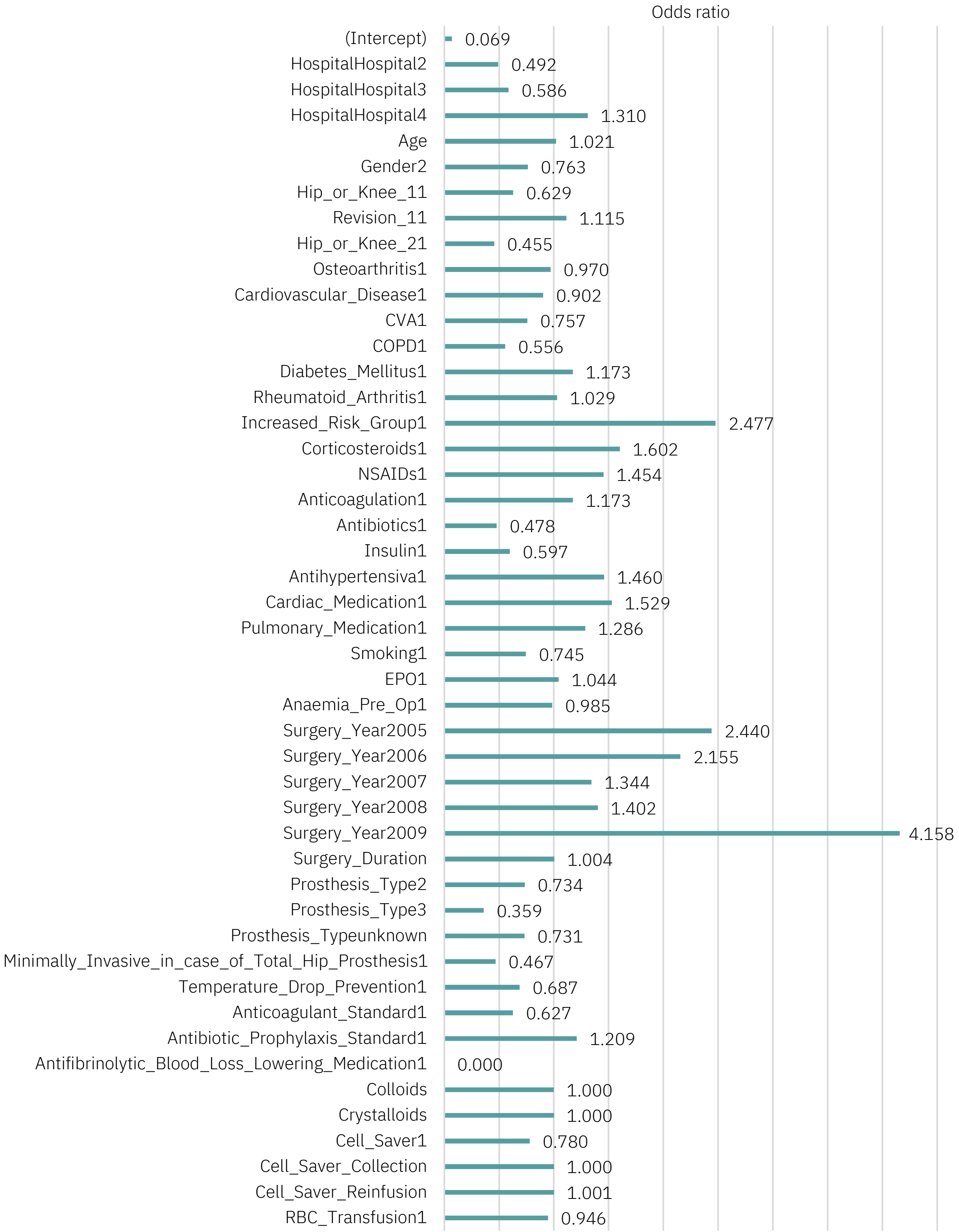


Figure C-20: Case COM$_{OPT}$[LREG]: Odds ratios of all inputs.

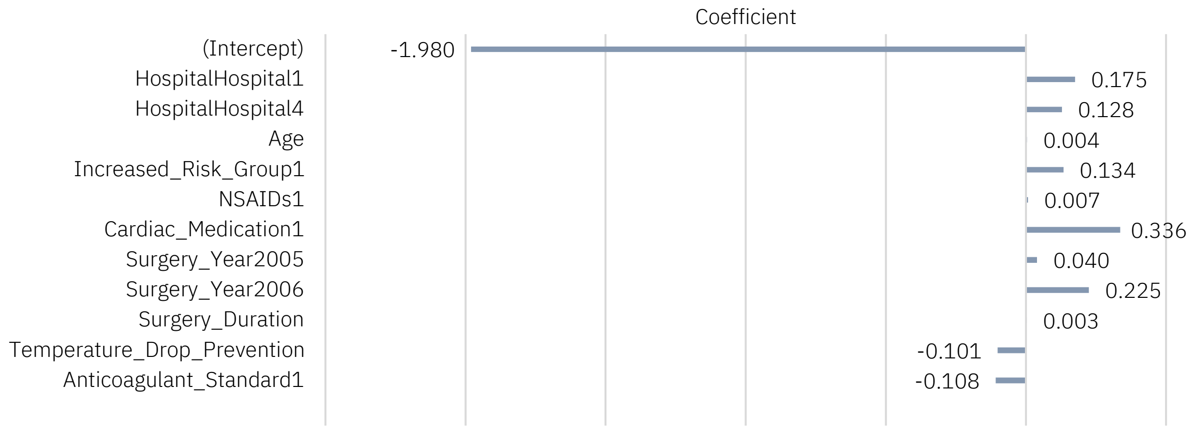


Figure C-21: Case COM$_{\mathrm{OPT}}$[lasso]: Non-zero coefficients of the inputs.

Appendix D: Mediation Analysis (Supplemental Results)

Tables D-1 and D-2 below represent the effect estimates with their 95% CI's and p-values for the ***pessimistic*** scenario. They were exported from the output of the R programming code (specifically, line #238 and #257 of the Programming segment #4 in Appendix E). Notable are the Total Effect, the average ACME, and the average ADE (in green). In Section 4.3., we utilize these three measures for answering RQ#4.

The ACME measures exhibit statistical significance at the level of 0.001.

Table D-1: Causal mediation analysis results (***pessimistic*** scenario): Quasi-Bayesian Confidence Intervals Method (Inference Conditional on the Covariate Values).

	Estimate	95% CI Lower	95% CI Upper	p-value	Significance*
ACME (control)	0.045	0.027	0.07	<0.001	0.001
ACME (treated)	0.044	0.026	0.07	<0.001	0.001
ADE (control)	-0.001	-0.051	0.06	0.93	
ADE (treated)	-0.002	-0.057	0.06	0.93	
Total Effect	0.043	-0.011	0.11	0.15	
Prop. Mediated (control)	0.944	-9.371	11.05	0.15	
Prop. Mediated (treated)	0.949	-8.145	9.88	0.15	
ACME (average)	0.045	0.027	0.07	<0.001	0.001
ADE (average)	-0.001	-0.054	0.06	0.93	
Prop. Mediated (average)	0.947	-8.788	10.55	0.15	

** The last column pertains to the level of statistical significance.*

Table D-2: Causal mediation analysis results (***pessimistic*** scenario): Nonparametric Bootstrap Confidence Intervals with the Percentile Method, 1000 simulations (Inference Conditional on the Covariate Values).

	Estimate	95% CI Lower	95% CI Upper	p-value	Significance*
ACME (control)	0.047	0.027	0.07	<0.001	0.001
ACME (treated)	0.047	0.027	0.07	<0.001	0.001
ADE (control)	-0.002	-0.055	0.05	0.91	
ADE (treated)	-0.003	-0.063	0.06	0.91	
Total Effect	0.044	-0.016	0.10	0.16	
Prop. Mediated (control)	1.060	-7.316	11.06	0.16	
Prop. Mediated (treated)	1.054	-6.424	10.20	0.16	
ACME (average)	0.047	0.027	0.07	<0.001	0.001
ADE (average)	-0.003	-0.059	0.05	0.91	
Prop. Mediated (average)	1.057	-6.870	10.63	0.16	

** The last column pertains to the level of statistical significance.*

Tables D-3 and D-4 contain the coefficients and odds ratios with their p-values of the **model.y** (logistic regression) developed during the mediation analysis procedure (line #196 and #210 of the Programming segment #4 in Appendix E). These values apply to the ***pessimistic*** scenario. The b-path and c'-path are marked in green. Further, more detailed variable selection may be performed in a follow-up study because only six confounding variables exhibit statistical significance at the level of at least 0.1. This observation suggests exclusion of the statistically insignificant confounders that don't significantly contribute to the logistic regression model fit.

RBC_Transfusion (equivalent to the b-path) exhibits statistical significance at the level of 0.001. Pre-operative anaemia (c'-path) exhibits no statistical significance at the level of at least 0.1.

Table D-3: b- and c'-path (model.y, ***pessimistic*** scenario): Coefficients.

	Estimate	Std. Error	z value	p-value	Significance
## (Intercept)	-2.423	0.483	-5.017	**<0.001**	0.001
## Anaemia_Pre_Op (c'-path)	-0.016	0.184	-0.088	0.930	
## RBC_Transfusion (b-path)	1.078	0.152	7.099	**<0.001**	0.001
## HospitalHospital2	-1.008	0.153	-6.597	**<0.001**	0.001
## HospitalHospital3	-0.732	0.186	-3.940	**<0.001**	0.001
## HospitalHospital4	-0.053	0.168	-0.318	0.751	
## Age	0.023	0.005	4.391	**<0.001**	0.001
## Gender2	-0.199	0.117	-1.703	**0.089**	0.1
## Hip_or_Knee_11	-0.119	0.111	-1.075	0.282	
## Primary_or_Revision_11	0.111	0.193	0.577	0.564	
## Cardiovascular_Disease1	0.140	0.113	1.239	0.215	
## Diabetes_Mellitus1	0.307	0.157	1.956	**0.051**	0.1
## Increased_Risk_Group1	0.872	0.248	3.513	**<0.001**	0.001
## EPO1	0.274	0.184	1.491	0.136	
## Surgery_Year2005	0.462	0.349	1.323	0.186	
## Surgery_Year2006	0.186	0.349	0.534	0.593	
## Surgery_Year2007	-0.381	0.348	-1.095	0.273	
## Surgery_Year2008	-0.402	0.348	-1.155	0.248	
## Surgery_Year2009	0.391	0.799	0.490	0.624	

** The last column pertains to the level of statistical significance.*

Table D-4: b- and c'-path (model.y, ***pessimistic*** scenario): Odds ratios.

	OR	2.5 %	97.5 %	p-value	Significance
## (Intercept)	0.089	0.034	0.224	**<0.001**	0.001
## Anaemia_Pre_Op (c'-path)	0.984	0.682	1.406	0.930	
## RBC_Transfusion (b-path)	2.940	2.181	3.959	**<0.001**	0.001
## HospitalHospital2	0.365	0.270	0.492	**<0.001**	0.001
## HospitalHospital3	0.481	0.334	0.692	**<0.001**	0.001
## HospitalHospital4	0.948	0.682	1.319	0.751	
## Age	1.023	1.013	1.034	**<0.001**	0.001
## Gender2	0.820	0.653	1.032	**0.089**	0.1
## Hip_or_Knee_11	0.888	0.714	1.102	0.282	
## Primary_or_Revision_11	1.118	0.759	1.618	0.564	
## Cardiovascular_Disease1	1.150	0.922	1.436	0.215	
## Diabetes_Mellitus1	1.360	0.995	1.843	**0.051**	0.1
## Increased_Risk_Group1	2.392	1.461	3.875	**<0.001**	0.001
## EPO1	1.316	0.912	1.878	0.136	
## Surgery_Year2005	1.588	0.821	3.259	0.186	
## Surgery_Year2006	1.205	0.624	2.471	0.593	
## Surgery_Year2007	0.683	0.354	1.398	0.273	
## Surgery_Year2008	0.669	0.347	1.369	0.248	
## Surgery_Year2009	1.479	0.269	6.701	0.624	

** The last column pertains to the level of statistical significance.*

Tables D-5 and D-6 below represent the effect estimates with their 95% CI's and p-values for the ***optimistic*** scenario. They were exported from the output of the R programming code (specifically, line #326 and #345 of the Programming segment #4 in Appendix E). We utilize three measures, namely, the Total Effect, the average ACME, and the average ADE (marked in green) for answering RQ#4 in Section 4.3.

Compared to the pessimistic scenario, none of the measures exhibit statistical significance at the level of at least 0.1.

Table D-5: Causal mediation analysis results (***optimistic*** scenario): Quasi-Bayesian Confidence Intervals Method (Inference Conditional on the Covariate Values).

	Estimate	95% CI Lower	95% CI Upper	p-value	Significance*
## ACME (control)	0.005	-0.007	0.02	0.39	
## ACME (treated)	0.005	-0.008	0.02	0.39	
## ADE (control)	0.005	-0.045	0.06	0.89	
## ADE (treated)	0.005	-0.047	0.06	0.89	
## Total Effect	0.010	-0.038	0.07	0.75	
## Prop. Mediated (control)	0.071	-3.643	3.63	0.81	
## Prop. Mediated (treated)	0.078	-3.440	3.55	0.81	
## ACME (average)	0.005	-0.007	0.02	0.39	
## ADE (average)	0.008	-0.046	0.06	0.89	
## Prop. Mediated (average)	0.074	-3.519	3.59	0.81	

** The last column pertains to the level of statistical significance.*

Table D-6: Causal mediation analysis results (***optimistic*** scenario): Nonparametric Bootstrap Confidence Intervals with the Percentile Method, 1000 simulations (Inference Conditional on the Covariate Values).

	Estimate	95% CI Lower	95% CI Upper	p-value	Significance*
## ACME (control)	0.005	-0.007	0.02	0.41	
## ACME (treated)	0.005	-0.007	0.02	0.41	
## ADE (control)	0.005	-0.046	0.06	0.86	
## ADE (treated)	0.005	-0.047	0.06	0.86	
## Total Effect	0.010	-0.041	0.07	0.72	
## Prop. Mediated (control)	0.516	-3.599	4.29	0.82	
## Prop. Mediated (treated)	0.524	-3.421	4.18	0.82	
## ACME (average)	0.005	-0.007	0.02	0.41	
## ADE (average)	0.005	-0.046	0.06	0.86	
## Prop. Mediated (average)	0.520	-3.511	4.23	0.82	

** The last column pertains to the level of statistical significance.*

In Tables D-7 and D-8, the coefficients and odds ratios are presented with their p-values of the model.y (logistic regression) developed during the mediation analysis procedure (line #288 and #302 of the Programming segment #4 in Appendix E). These values apply to the *optimistic* scenario. The b-path and c'-path are marked in green.

Compared to the pessimistic scenario, neither RBC_Transfusion (equivalent to the b-path), nor pre-operative anaemia (c'-path) exhibit statistical significance at the level of at least 0.1.

Table D-7: b- and c'-path (model.y, ***optimistic*** scenario): Coefficients.

	Coefficient Estimate	Std. Error	z value	p-value	Significance*
## (Intercept)	-2.470	0.505	-4.896	<0.001	0.001
## Anaemia_Pre_Op (c'-path)	0.033	0.190	0.171	0.864	
## RBC_Transfusion (b-path)	0.141	0.169	0.837	0.403	
## HospitalHospital2	-1.044	0.155	-6.717	<0.001	0.001
## HospitalHospital3	-0.774	0.187	-4.128	<0.001	0.001
## HospitalHospital4	-0.071	0.169	-0.420	0.674	
## Age	0.022	0.005	4.209	<0.001	0.001
## Gender2	-0.161	0.118	-1.365	0.172	
## Hip_or_Knee_11	-0.158	0.112	-1.403	0.161	
## Primary_or_Revision_11	0.119	0.197	0.603	0.547	
## Cardiovascular_Disease1	0.131	0.115	1.143	0.253	
## Diabetes_Mellitus1	0.158	0.163	0.965	0.334	
## Increased_Risk_Group1	0.947	0.249	3.804	<0.001	0.001
## EPO1	0.209	0.188	1.114	0.265	
## Surgery_Year2005	0.531	0.375	1.415	0.157	
## Surgery_Year2006	0.269	0.375	0.717	0.474	
## Surgery_Year2007	-0.197	0.373	-0.527	0.598	
## Surgery_Year2008	-0.267	0.374	-0.714	0.475	
## Surgery_Year2009	0.597	0.797	0.749	0.454	

** The last column pertains to the level of statistical significance.*

Table D-8: b- and c'-path (model.y, ***optimistic*** scenario): Odds ratios.

	OR	2.5 %	97.5 %	p-value	Significance*
## (Intercept)	0.085	0.030	0.221	<0.001	0.001
## Anaemia_Pre_Op (c'-path)	1.033	0.707	1.491	0.864	
## RBC_Transfusion (b-path)	1.152	0.822	1.596	0.403	
## HospitalHospital2	0.352	0.259	0.477	<0.001	0.001
## HospitalHospital3	0.461	0.319	0.666	<0.001	0.001
## HospitalHospital4	0.931	0.669	1.299	0.674	
## Age	1.023	1.012	1.034	<0.001	0.001
## Gender2	0.851	0.676	1.074	0.172	
## Hip_or_Knee_11	0.854	0.685	1.064	0.161	
## Primary_or_Revision_11	1.126	0.758	1.640	0.547	
## Cardiovascular_Disease1	1.140	0.911	1.430	0.253	
## Diabetes_Mellitus1	1.171	0.845	1.604	0.334	
## Increased_Risk_Group1	2.578	1.569	4.174	<0.001	0.001
## EPO1	1.233	0.847	1.772	0.265	
## Surgery_Year2005	1.701	0.846	3.733	0.157	
## Surgery_Year2006	1.308	0.651	2.870	0.474	
## Surgery_Year2007	0.821	0.410	1.797	0.598	
## Surgery_Year2008	0.766	0.382	1.677	0.475	
## Surgery_Year2009	1.817	0.334	8.273	0.454	

** The last column pertains to the level of statistical significance.*

In Tables D-9 and D-10, the coefficients and odds ratios are presented with their p-values of the model.m (logistic regression) developed during the mediation analysis procedure (line #151 and #165 of the Programming segment #4 in Appendix E). The a-path is marked in green.

Pre-operative anaemia (on the a-path) exhibits statistical significance at the level of 0.001.

Table D-9: a-path (model.m): Coefficients.

##	Estimate	Std. Error	z value	p-value	Significance*
## (Intercept)	-4.245	0.636	-6.673	<0.001	0.001
## Anaemia_Pre_Op (a-path)	1.685	0.202	8.357	<0.001	0.001
## HospitalHospital2	-0.174	0.198	-0.876	0.381	
## HospitalHospital3	-1.095	0.275	-3.988	<0.001	0.001
## HospitalHospital4	-0.142	0.237	-0.602	0.547	
## Age	0.032	0.007	4.402	<0.001	0.001
## Gender2	0.825	0.178	4.644	<0.001	0.001
## Hip_or_Knee_11	-1.008	0.164	-6.134	<0.001	0.001
## Primary_or_Revision_11	0.434	0.230	1.887	0.059	0.1
## Cardiovascular_Disease1	0.079	0.150	0.524	0.600	
## Diabetes_Mellitus1	0.133	0.215	0.620	0.535	
## Increased_Risk_Group1	0.832	0.282	2.950	0.003	0.01
## EPO1	-0.942	0.270	-3.490	<0.001	0.001
## Surgery_Year2005	-0.188	0.397	-0.473	0.636	
## Surgery_Year2006	-0.478	0.399	-1.197	0.231	
## Surgery_Year2007	-0.408	0.388	-1.051	0.293	
## Surgery_Year2008	-0.705	0.393	-1.794	0.073	0.1
## Surgery_Year2009	-1.131	1.195	-0.946	0.344	

** The last column pertains to the level of statistical significance.*

Table D-10: a-path (model.m): Odds ratios.

##	OR	2.5 %	97.5 %	p-value	Significance*
## (Intercept)	0.014	0.004	0.048	<0.001	0.001
## Anaemia_Pre_Op (a-path)	5.391	3.625	7.998	<0.001	0.001
## HospitalHospital2	0.840	0.572	1.247	0.381	
## HospitalHospital3	0.335	0.194	0.570	<0.001	0.001
## HospitalHospital4	0.867	0.546	1.381	0.547	
## Age	1.033	1.018	1.048	<0.001	0.001
## Gender2	2.281	1.625	3.264	<0.001	0.001
## Hip_or_Knee_11	0.365	0.263	0.500	<0.001	0.001
## Primary_or_Revision_11	1.544	0.969	2.396	0.059	0.1
## Cardiovascular_Disease1	1.082	0.806	1.453	0.600	
## Diabetes_Mellitus1	1.142	0.740	1.720	0.535	
## Increased_Risk_Group1	2.297	1.301	3.944	0.003	0.01
## EPO1	0.390	0.224	0.648	<0.001	0.001
## Surgery_Year2005	0.829	0.393	1.887	0.636	
## Surgery_Year2006	0.620	0.293	1.417	0.231	
## Surgery_Year2007	0.665	0.322	1.492	0.293	
## Surgery_Year2008	0.494	0.236	1.118	0.073	0.1
## Surgery_Year2009	0.323	0.015	2.441	0.344	

** The last column pertains to the level of statistical significance.*

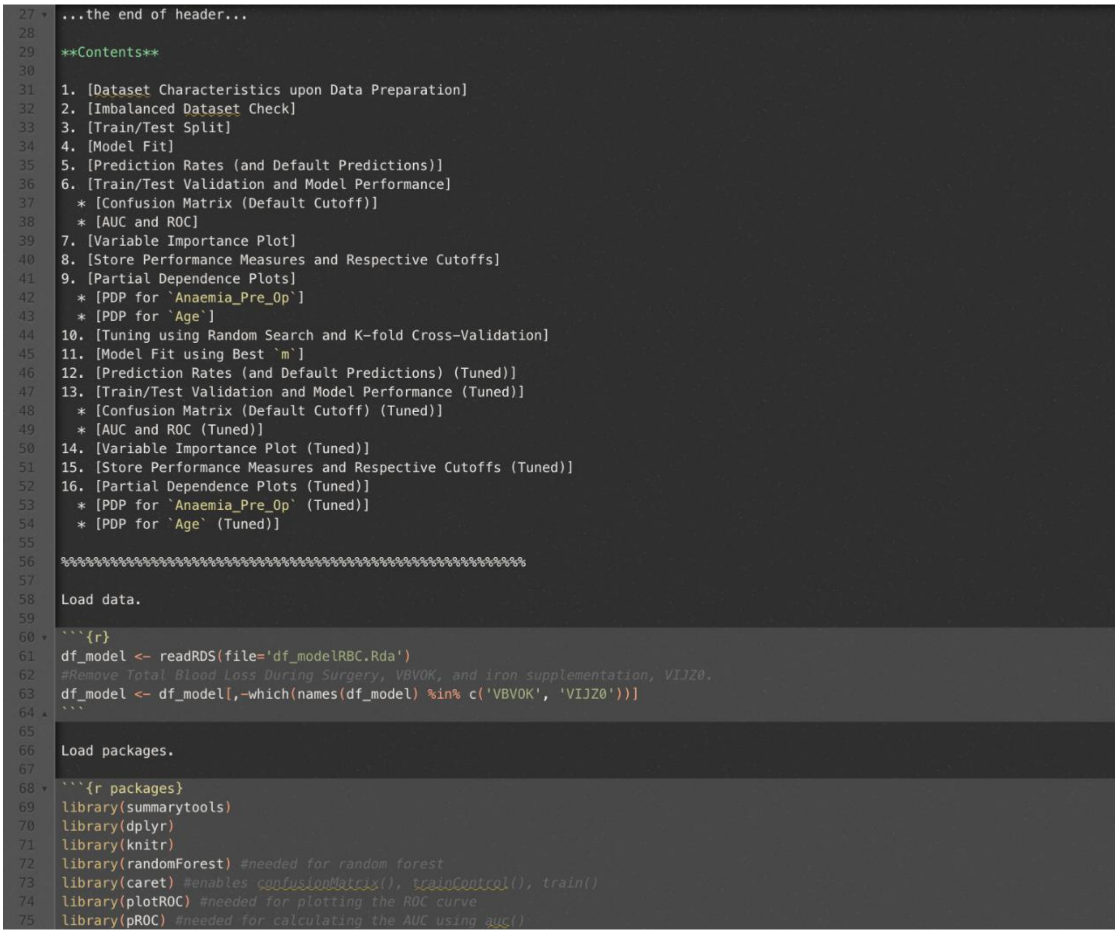
Appendix E: R Programming (Code Disclosure)

Appendix E contains a series of sample R code of modelling implementation for Case RBC. Sequentially, presented is the R code of random forest (segment #1), logistic regression (segment #2) and lasso (segment #3) model development.

Then this appendix is completed by revealing the R code of mediation analysis (segment #4).

The code was executed in the R Statistical Software, version 4.1.2. The code implemented for data preparation purposes and variable selection purposes is not published in this report. Omitted is also the R code implemented for Cases COM_{PES} and COM_{OPT} because it was executed in the same manner as for Case RBC.

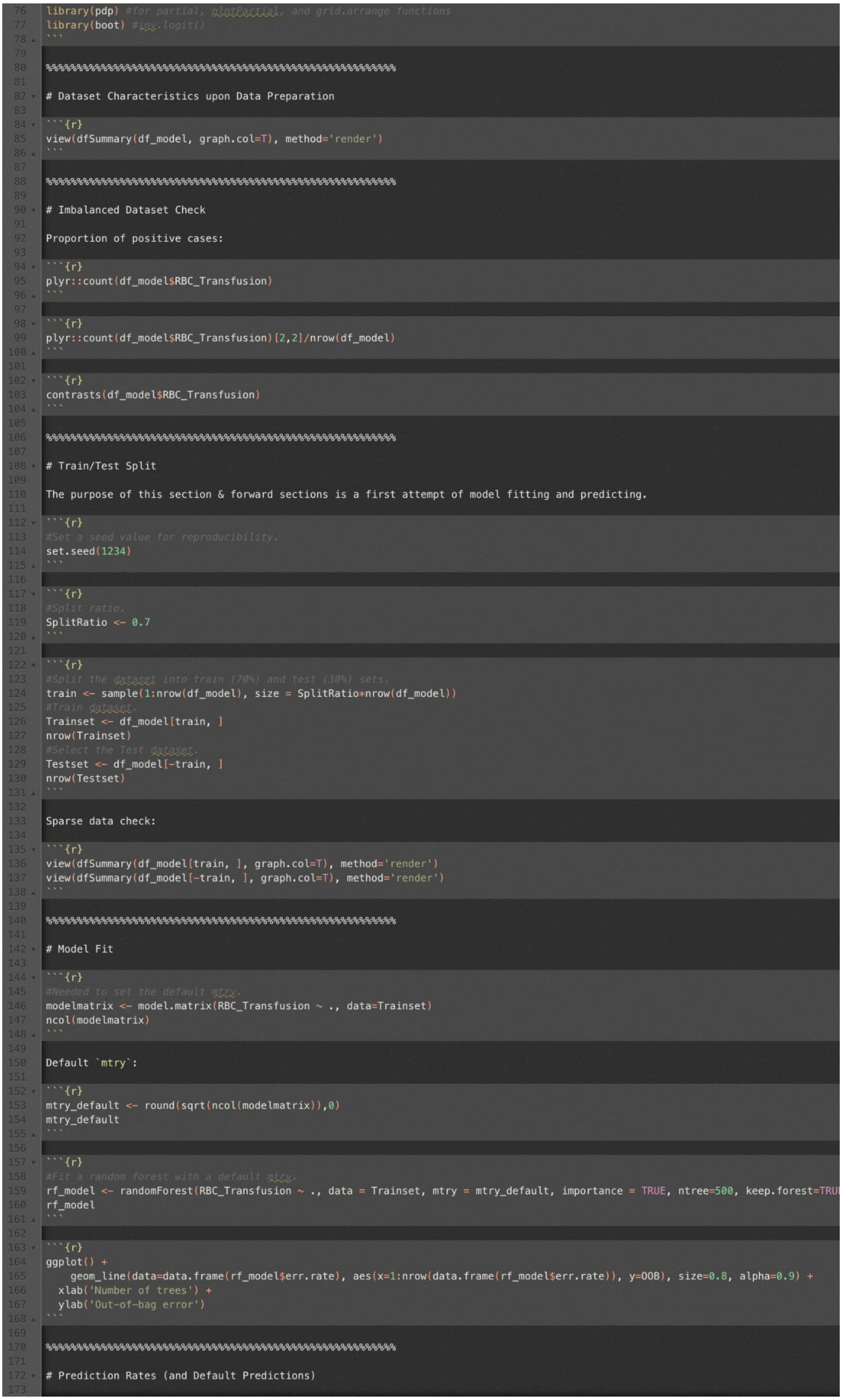
Programming segment #1: The R code of Random Forest, Case RBC[RF], is presented below.



~~~
...the end of header...

**Contents**

1. [Dataset Characteristics upon Data Preparation]
2. [Imbalanced Dataset Check]
3. [Train/Test Split]
4. [Model Fit]
5. [Prediction Rates (and Default Predictions)]
6. [Train/Test Validation and Model Performance]
  * [Confusion Matrix (Default Cutoff)]
  * [AUC and ROC]
7. [Variable Importance Plot]
8. [Store Performance Measures and Respective Cutoffs]
9. [Partial Dependence Plots]
  * [PDP for `Anaemia_Pre_Op`]
  * [PDP for `Age`]
10. [Tuning using Random Search and K-fold Cross-Validation]
11. [Model Fit using Best `m`]
12. [Prediction Rates (and Default Predictions) (Tuned)]
13. [Train/Test Validation and Model Performance (Tuned)]
  * [Confusion Matrix (Default Cutoff) (Tuned)]
  * [AUC and ROC (Tuned)]
14. [Variable Importance Plot (Tuned)]
15. [Store Performance Measures and Respective Cutoffs (Tuned)]
16. [Partial Dependence Plots (Tuned)]
  * [PDP for `Anaemia_Pre_Op` (Tuned)]
  * [PDP for `Age` (Tuned)]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Load data.

```{r}
df_model <- readRDS(file='df_modelRBC.Rda')
#Remove Total Blood Loss During Surgery, VBVOK, and iron supplementation, VIJZ0.
df_model <- df_model[,-which(names(df_model) %in% c('VBVOK', 'VIJZ0'))]
~~~

Load packages.

```{r packages}
library(summarytools)
library(dplyr)
library(knitr)
library(randomForest) #needed for random forest
library(caret) #enables confusionMatrix(), trainControl(), train()
library(plotROC) #needed for plotting the ROC curve
library(pROC) #needed for calculating the AUC using auc()
```



```
library(pdp) #for partial, plotPartial, and grid.arrange functions
library(boot) #inv.logit()
```

%%%

Dataset Characteristics upon Data Preparation

```{r}
view(dfSummary(df_model, graph.col=T), method='render')
```

%%%

Imbalanced Dataset Check

Proportion of positive cases:

```{r}
plyr::count(df_model$RBC_Transfusion)
```

```{r}
plyr::count(df_model$RBC_Transfusion)[2,2]/nrow(df_model)
```

```{r}
contrasts(df_model$RBC_Transfusion)
```

%%%

Train/Test Split

The purpose of this section & forward sections is a first attempt of model fitting and predicting.

```{r}
#Set a seed value for reproducibility.
set.seed(1234)
```

```{r}
#Split ratio.
SplitRatio <- 0.7
```

```{r}
#Split the dataset into train (70%) and test (30%) sets.
train <- sample(1:nrow(df_model), size = SplitRatio*nrow(df_model))
#Train dataset.
Trainset <- df_model[train, ]
nrow(Trainset)
#Select the Test dataset.
Testset <- df_model[-train, ]
nrow(Testset)
```

Sparse data check:

```{r}
view(dfSummary(df_model[train, ], graph.col=T), method='render')
view(dfSummary(df_model[-train, ], graph.col=T), method='render')
```

%%%

Model Fit

```{r}
#Needed to set the default mtry.
modelmatrix <- model.matrix(RBC_Transfusion ~ ., data=Trainset)
ncol(modelmatrix)
```

Default `mtry`:

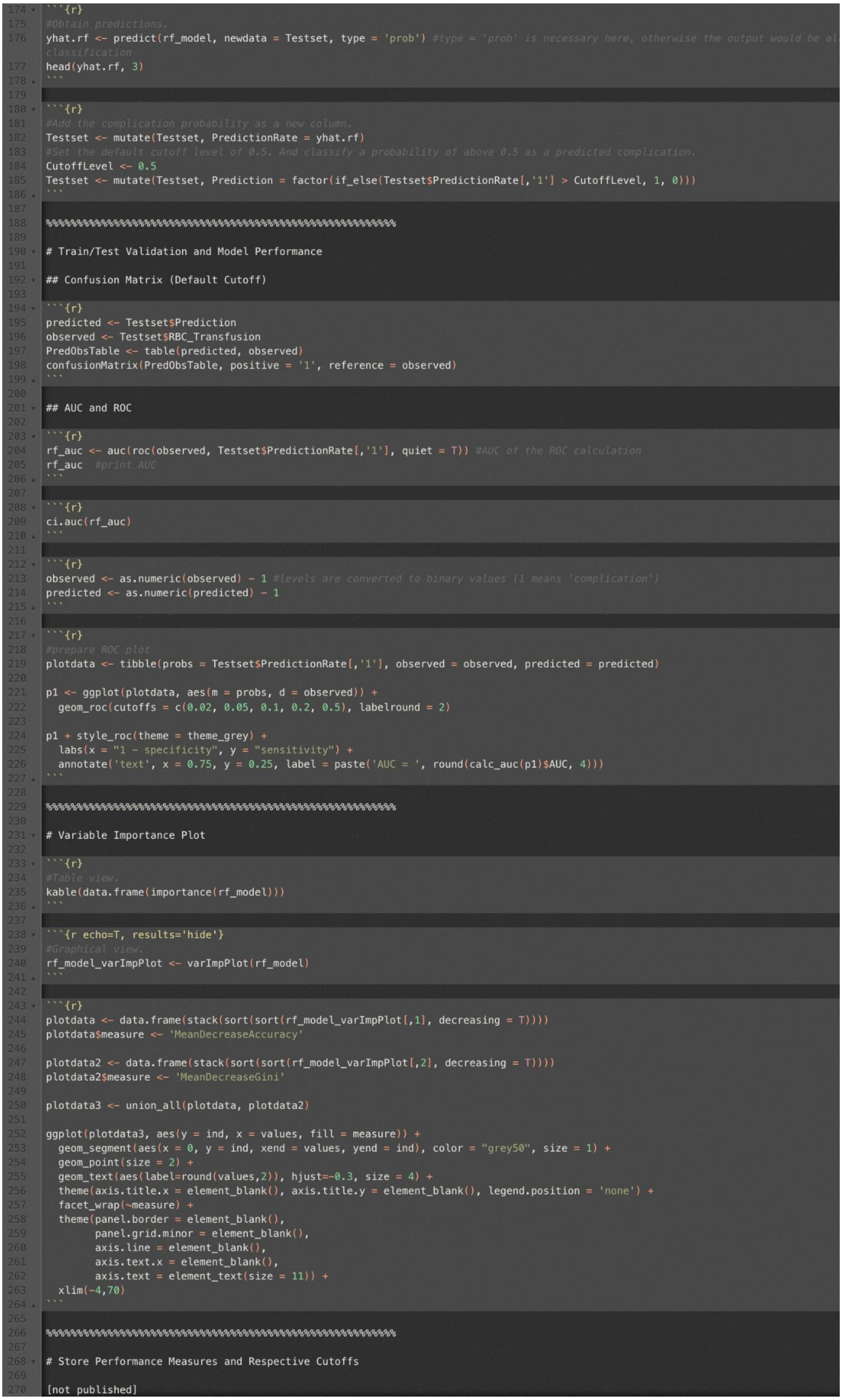
```{r}
mtry_default <- round(sqrt(ncol(modelmatrix)),0)
mtry_default
```

```{r}
#Fit a random forest with a default mtry.
rf_model <- randomForest(RBC_Transfusion ~ ., data = Trainset, mtry = mtry_default, importance = TRUE, ntree=500, keep.forest=TRU
rf_model
```

```{r}
ggplot() +
    geom_line(data=data.frame(rf_model$err.rate), aes(x=1:nrow(data.frame(rf_model$err.rate)), y=OOB), size=0.8, alpha=0.9) +
  xlab('Number of trees') +
  ylab('Out-of-bag error')
```

%%%

Prediction Rates (and Default Predictions)



~~~
```{r}
#Obtain predictions.
yhat.rf <- predict(rf_model, newdata = Testset, type = 'prob') #type = 'prob' is necessary here, otherwise the output would be al
classification
head(yhat.rf, 3)
~~~

```{r}
#Add the complication probability as a new column.
Testset <- mutate(Testset, PredictionRate = yhat.rf)
#Set the default cutoff level of 0.5. And classify a probability of above 0.5 as a predicted complication.
CutoffLevel <- 0.5
Testset <- mutate(Testset, Prediction = factor(if_else(Testset$PredictionRate[,'1'] > CutoffLevel, 1, 0)))
```

%%

Train/Test Validation and Model Performance

Confusion Matrix (Default Cutoff)

```{r}
predicted <- Testset$Prediction
observed <- Testset$RBC_Transfusion
PredObsTable <- table(predicted, observed)
confusionMatrix(PredObsTable, positive = '1', reference = observed)
```

AUC and ROC

```{r}
rf_auc <- auc(roc(observed, Testset$PredictionRate[,'1'], quiet = T)) #AUC of the ROC calculation
rf_auc  #print AUC
```

```{r}
ci.auc(rf_auc)
```

```{r}
observed <- as.numeric(observed) - 1 #levels are converted to binary values (1 means 'complication')
predicted <- as.numeric(predicted) - 1
```

```{r}
#prepare ROC plot
plotdata <- tibble(probs = Testset$PredictionRate[,'1'], observed = observed, predicted = predicted)

p1 <- ggplot(plotdata, aes(m = probs, d = observed)) +
  geom_roc(cutoffs = c(0.02, 0.05, 0.1, 0.2, 0.5), labelround = 2)

p1 + style_roc(theme = theme_grey) +
  labs(x = "1 - specificity", y = "sensitivity") +
  annotate('text', x = 0.75, y = 0.25, label = paste('AUC = ', round(calc_auc(p1)$AUC, 4)))
```

%%

Variable Importance Plot

```{r}
#Table view.
kable(data.frame(importance(rf_model)))
```

```{r echo=T, results='hide'}
#Graphical view.
rf_model_varImpPlot <- varImpPlot(rf_model)
```

```{r}
plotdata <- data.frame(stack(sort(sort(rf_model_varImpPlot[,1], decreasing = T))))
plotdata$measure <- 'MeanDecreaseAccuracy'

plotdata2 <- data.frame(stack(sort(sort(rf_model_varImpPlot[,2], decreasing = T))))
plotdata2$measure <- 'MeanDecreaseGini'

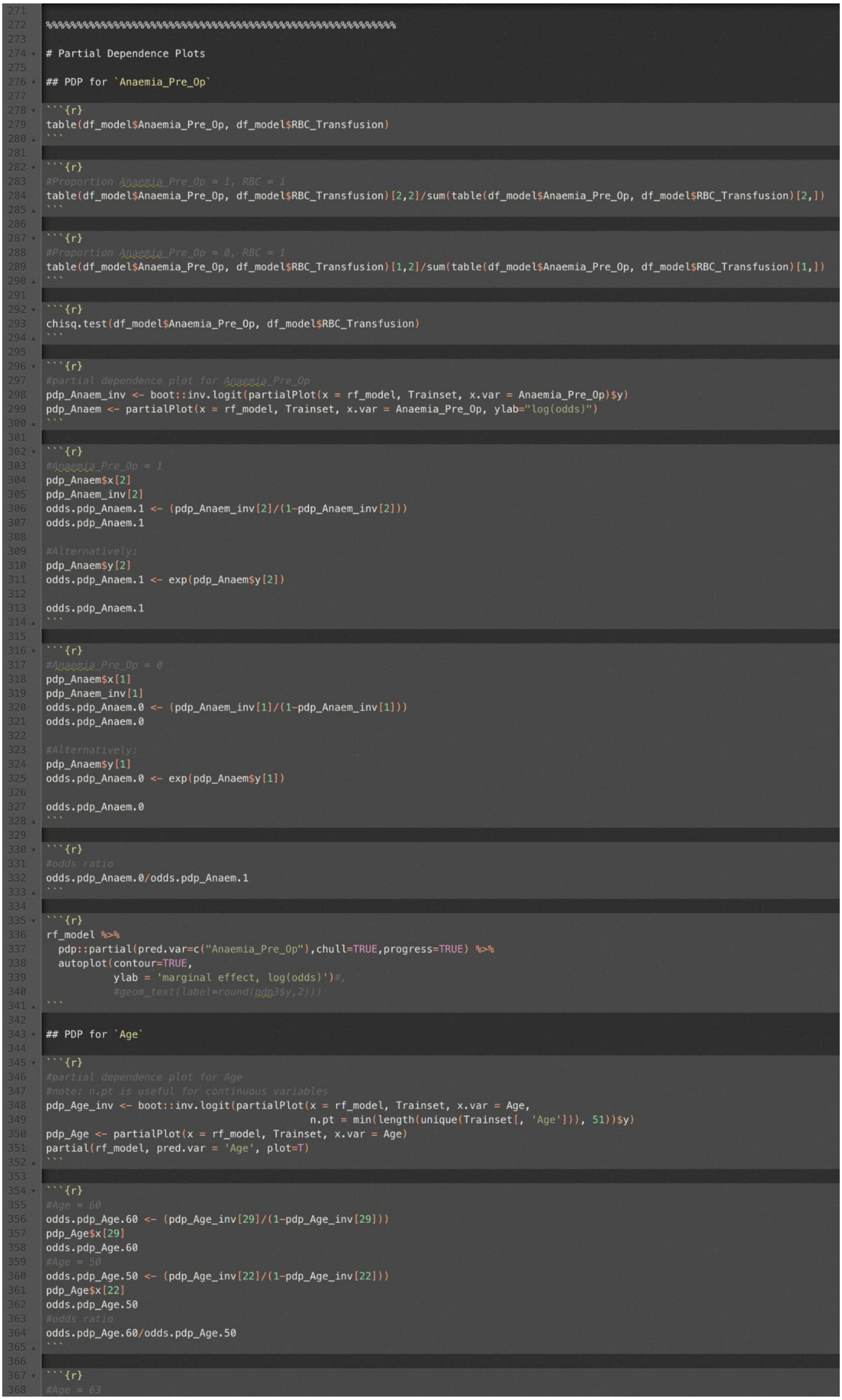
plotdata3 <- union_all(plotdata, plotdata2)

ggplot(plotdata3, aes(y = ind, x = values, fill = measure)) +
  geom_segment(aes(x = 0, y = ind, xend = values, yend = ind), color = "grey50", size = 1) +
  geom_point(size = 2) +
  geom_text(aes(label=round(values,2)), hjust=-0.3, size = 4) +
  theme(axis.title.x = element_blank(), axis.title.y = element_blank(), legend.position = 'none') +
  facet_wrap(~measure) +
  theme(panel.border = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_blank(),
        axis.text.x = element_blank(),
        axis.text = element_text(size = 11)) +
  xlim(-4,70)
```

%%

Store Performance Measures and Respective Cutoffs

[not published]
```
```



~~~
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Partial Dependence Plots

## PDP for `Anaemia_Pre_Op`

```{r}
table(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)
```

```{r}
#Proportion Anaemia_Pre_Op = 1, RBC = 1
table(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)[2,2]/sum(table(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)[2,])
```

```{r}
#Proportion Anaemia_Pre_Op = 0, RBC = 1
table(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)[1,2]/sum(table(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)[1,])
```

```{r}
chisq.test(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)
```

```{r}
#partial dependence plot for Anaemia_Pre_Op
pdp_Anaem_inv <- boot::inv.logit(partialPlot(x = rf_model, Trainset, x.var = Anaemia_Pre_Op)$y)
pdp_Anaem <- partialPlot(x = rf_model, Trainset, x.var = Anaemia_Pre_Op, ylab="log(odds)")
```

```{r}
#Anaemia_Pre_Op = 1
pdp_Anaem$x[2]
pdp_Anaem_inv[2]
odds.pdp_Anaem.1 <- (pdp_Anaem_inv[2]/(1-pdp_Anaem_inv[2]))
odds.pdp_Anaem.1

#Alternatively:
pdp_Anaem$y[2]
odds.pdp_Anaem.1 <- exp(pdp_Anaem$y[2])

odds.pdp_Anaem.1
```

```{r}
#Anaemia_Pre_Op = 0
pdp_Anaem$x[1]
pdp_Anaem_inv[1]
odds.pdp_Anaem.0 <- (pdp_Anaem_inv[1]/(1-pdp_Anaem_inv[1]))
odds.pdp_Anaem.0

#Alternatively:
pdp_Anaem$y[1]
odds.pdp_Anaem.0 <- exp(pdp_Anaem$y[1])

odds.pdp_Anaem.0
```

```{r}
#odds ratio
odds.pdp_Anaem.0/odds.pdp_Anaem.1
```

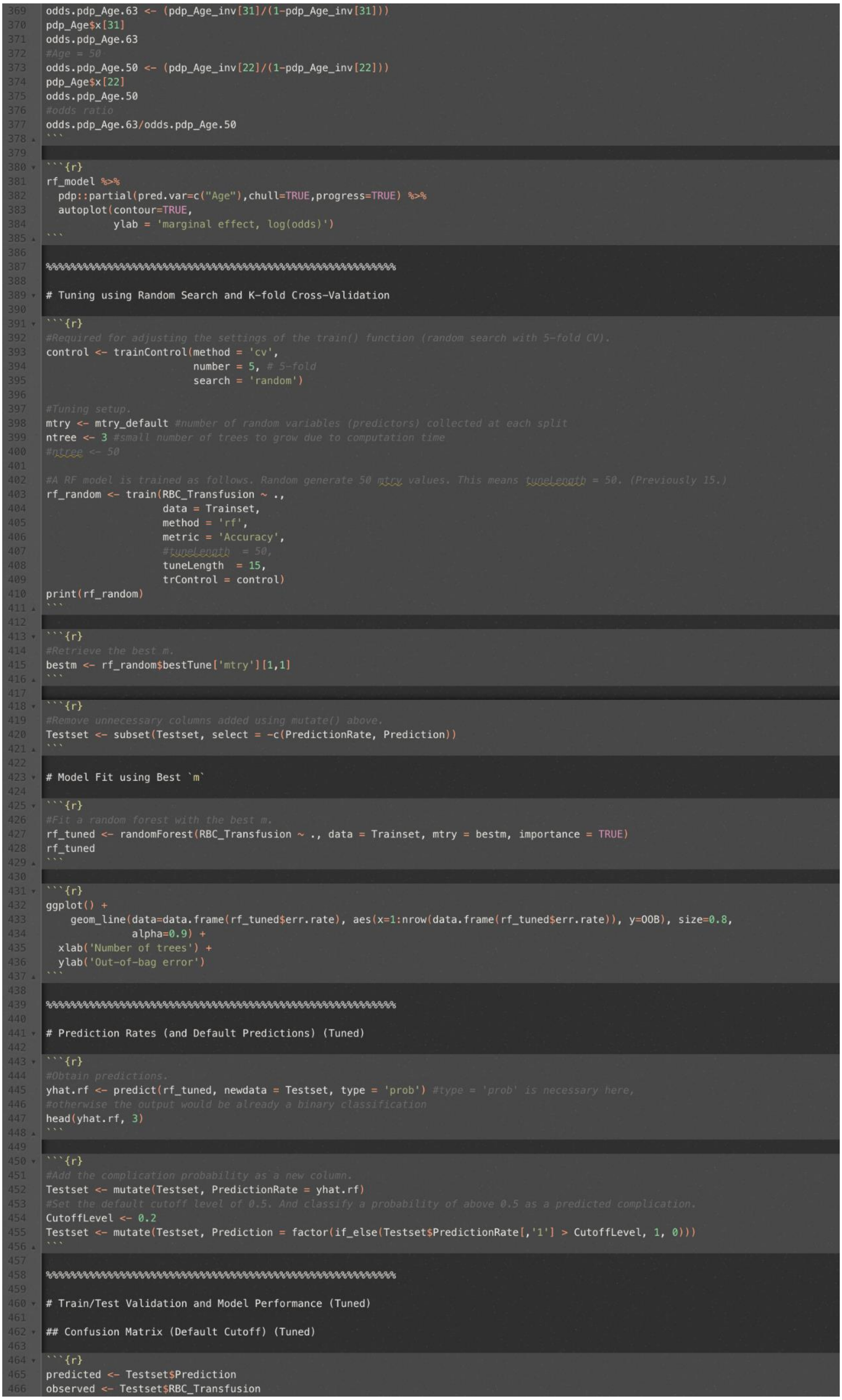
```{r}
rf_model %>%
  pdp::partial(pred.var=c("Anaemia_Pre_Op"),chull=TRUE,progress=TRUE) %>%
  autoplot(contour=TRUE,
           ylab = 'marginal effect, log(odds)')#,
           #geom_text(label=round(pdp3$y,2)))
```

## PDP for `Age`

```{r}
#partial dependence plot for Age
#note: n.pt is useful for continuous variables
pdp_Age_inv <- boot::inv.logit(partialPlot(x = rf_model, Trainset, x.var = Age,
                                           n.pt = min(length(unique(Trainset[, 'Age'])), 51))$y)
pdp_Age <- partialPlot(x = rf_model, Trainset, x.var = Age)
partial(rf_model, pred.var = 'Age', plot=T)
```

```{r}
#Age = 60
odds.pdp_Age.60 <- (pdp_Age_inv[29]/(1-pdp_Age_inv[29]))
pdp_Age$x[29]
odds.pdp_Age.60
#Age = 50
odds.pdp_Age.50 <- (pdp_Age_inv[22]/(1-pdp_Age_inv[22]))
pdp_Age$x[22]
odds.pdp_Age.50
#odds ratio
odds.pdp_Age.60/odds.pdp_Age.50
```

```{r}
#Age = 63
~~~



~~~~
odds.pdp_Age.63 <- (pdp_Age_inv[31]/(1-pdp_Age_inv[31]))
pdp_Age$x[31]
odds.pdp_Age.63
#Age = 50
odds.pdp_Age.50 <- (pdp_Age_inv[22]/(1-pdp_Age_inv[22]))
pdp_Age$x[22]
odds.pdp_Age.50
#odds ratio
odds.pdp_Age.63/odds.pdp_Age.50
```

```{r}
rf_model %>%
  pdp::partial(pred.var=c("Age"),chull=TRUE,progress=TRUE) %>%
  autoplot(contour=TRUE,
           ylab = 'marginal effect, log(odds)')
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Tuning using Random Search and K-fold Cross-Validation

```{r}
#Required for adjusting the settings of the train() function (random search with 5-fold CV).
control <- trainControl(method = 'cv',
                        number = 5, # 5-fold
                        search = 'random')

#Tuning setup.
mtry <- mtry_default #number of random variables (predictors) collected at each split
ntree <- 3 #small number of trees to grow due to computation time
#ntree <- 50

#A RF model is trained as follows. Random generate 50 mtry values. This means tuneLength = 50. (Previously 15.)
rf_random <- train(RBC_Transfusion ~ .,
                   data = Trainset,
                   method = 'rf',
                   metric = 'Accuracy',
                   #tuneLength  = 50,
                   tuneLength  = 15,
                   trControl = control)
print(rf_random)
```

```{r}
#Retrieve the best m.
bestm <- rf_random$bestTune['mtry'][1,1]
```

```{r}
#Remove unnecessary columns added using mutate() above.
Testset <- subset(Testset, select = -c(PredictionRate, Prediction))
```

# Model Fit using Best `m`

```{r}
#Fit a random forest with the best m.
rf_tuned <- randomForest(RBC_Transfusion ~ ., data = Trainset, mtry = bestm, importance = TRUE)
rf_tuned
```

```{r}
ggplot() +
    geom_line(data=data.frame(rf_tuned$err.rate), aes(x=1:nrow(data.frame(rf_tuned$err.rate)), y=OOB), size=0.8,
              alpha=0.9) +
  xlab('Number of trees') +
  ylab('Out-of-bag error')
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Prediction Rates (and Default Predictions) (Tuned)

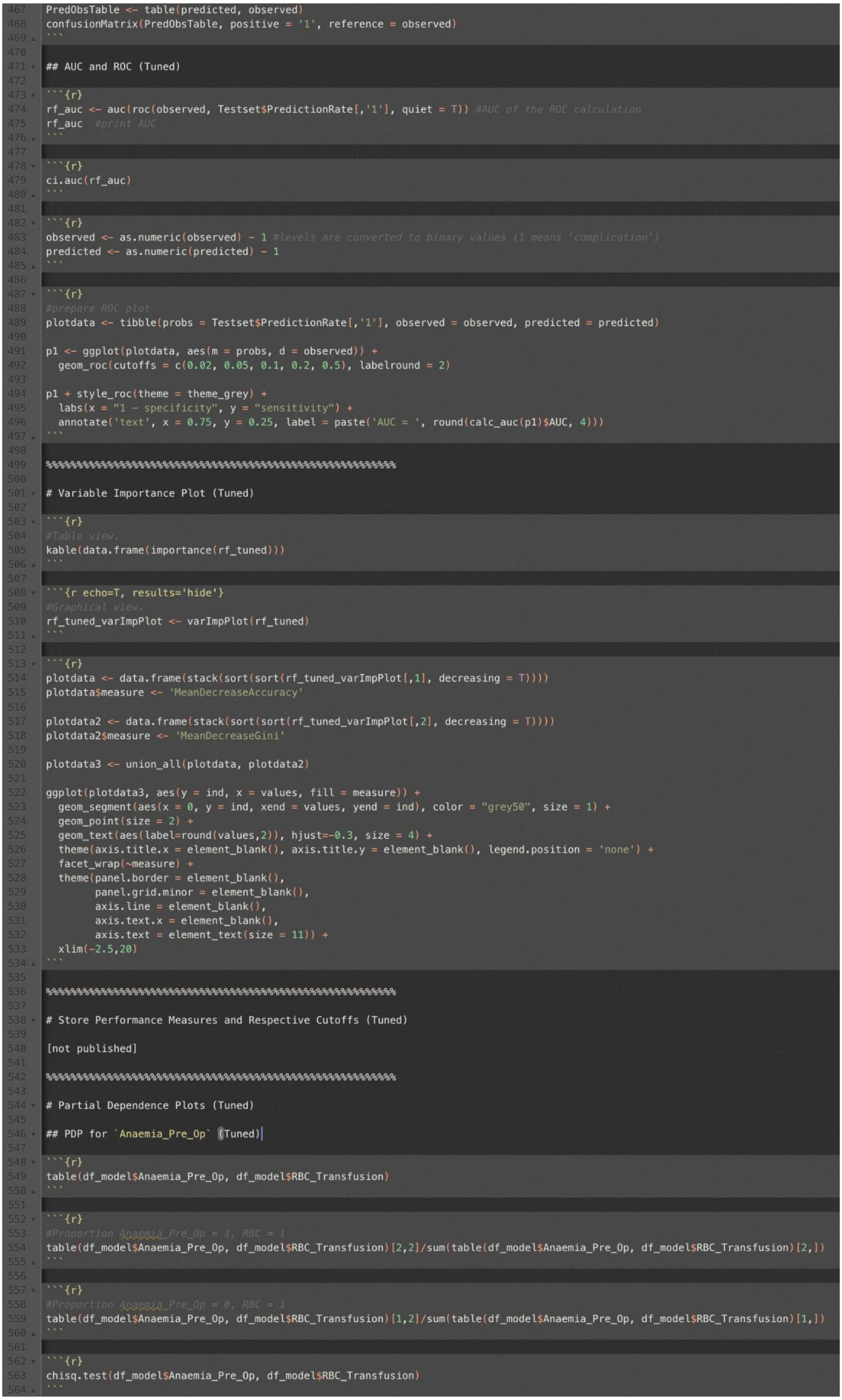
```{r}
#Obtain predictions.
yhat.rf <- predict(rf_tuned, newdata = Testset, type = 'prob') #type = 'prob' is necessary here,
#otherwise the output would be already a binary classification
head(yhat.rf, 3)
```

```{r}
#Add the complication probability as a new column.
Testset <- mutate(Testset, PredictionRate = yhat.rf)
#Set the default cutoff level of 0.5. And classify a probability of above 0.5 as a predicted complication.
CutoffLevel <- 0.2
Testset <- mutate(Testset, Prediction = factor(if_else(Testset$PredictionRate[,'1'] > CutoffLevel, 1, 0)))
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Train/Test Validation and Model Performance (Tuned)

## Confusion Matrix (Default Cutoff) (Tuned)

```{r}
predicted <- Testset$Prediction
observed <- Testset$RBC_Transfusion
~~~~



~~~~
PredObsTable <- table(predicted, observed)
confusionMatrix(PredObsTable, positive = '1', reference = observed)
```

## AUC and ROC (Tuned)

```{r}
rf_auc <- auc(roc(observed, Testset$PredictionRate[,'1'], quiet = T)) #AUC of the ROC calculation
rf_auc  #print AUC
```

```{r}
ci.auc(rf_auc)
```

```{r}
observed <- as.numeric(observed) - 1 #levels are converted to binary values (1 means 'complication')
predicted <- as.numeric(predicted) - 1
```

```{r}
#prepare ROC plot
plotdata <- tibble(probs = Testset$PredictionRate[,'1'], observed = observed, predicted = predicted)

p1 <- ggplot(plotdata, aes(m = probs, d = observed)) +
  geom_roc(cutoffs = c(0.02, 0.05, 0.1, 0.2, 0.5), labelround = 2)

p1 + style_roc(theme = theme_grey) +
  labs(x = "1 - specificity", y = "sensitivity") +
  annotate('text', x = 0.75, y = 0.25, label = paste('AUC = ', round(calc_auc(p1)$AUC, 4)))
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Variable Importance Plot (Tuned)

```{r}
#Table view.
kable(data.frame(importance(rf_tuned)))
```

```{r echo=T, results='hide'}
#Graphical view.
rf_tuned_varImpPlot <- varImpPlot(rf_tuned)
```

```{r}
plotdata <- data.frame(stack(sort(sort(rf_tuned_varImpPlot[,1], decreasing = T))))
plotdata$measure <- 'MeanDecreaseAccuracy'

plotdata2 <- data.frame(stack(sort(sort(rf_tuned_varImpPlot[,2], decreasing = T))))
plotdata2$measure <- 'MeanDecreaseGini'

plotdata3 <- union_all(plotdata, plotdata2)

ggplot(plotdata3, aes(y = ind, x = values, fill = measure)) +
  geom_segment(aes(x = 0, y = ind, xend = values, yend = ind), color = "grey50", size = 1) +
  geom_point(size = 2) +
  geom_text(aes(label=round(values,2)), hjust=-0.3, size = 4) +
  theme(axis.title.x = element_blank(), axis.title.y = element_blank(), legend.position = 'none') +
  facet_wrap(~measure) +
  theme(panel.border = element_blank(),
        panel.grid.minor = element_blank(),
        axis.line = element_blank(),
        axis.text.x = element_blank(),
        axis.text = element_text(size = 11)) +
  xlim(-2.5,20)
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Store Performance Measures and Respective Cutoffs (Tuned)

[not published]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Partial Dependence Plots (Tuned)

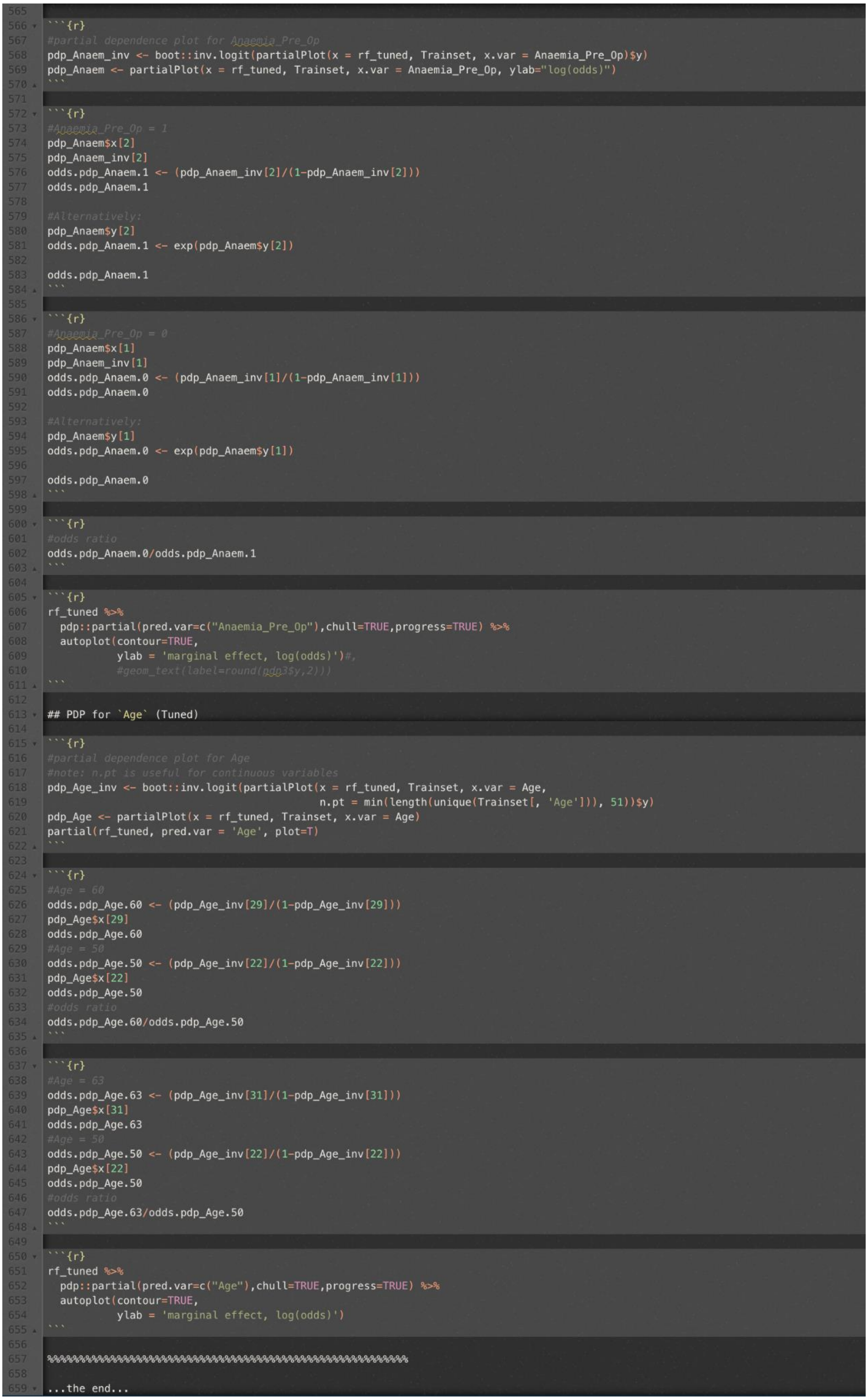
## PDP for `Anaemia_Pre_Op` (Tuned)

```{r}
table(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)
```

```{r}
#Proportion Anaemia_Pre_Op = 1, RBC = 1
table(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)[2,2]/sum(table(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)[2,])
```

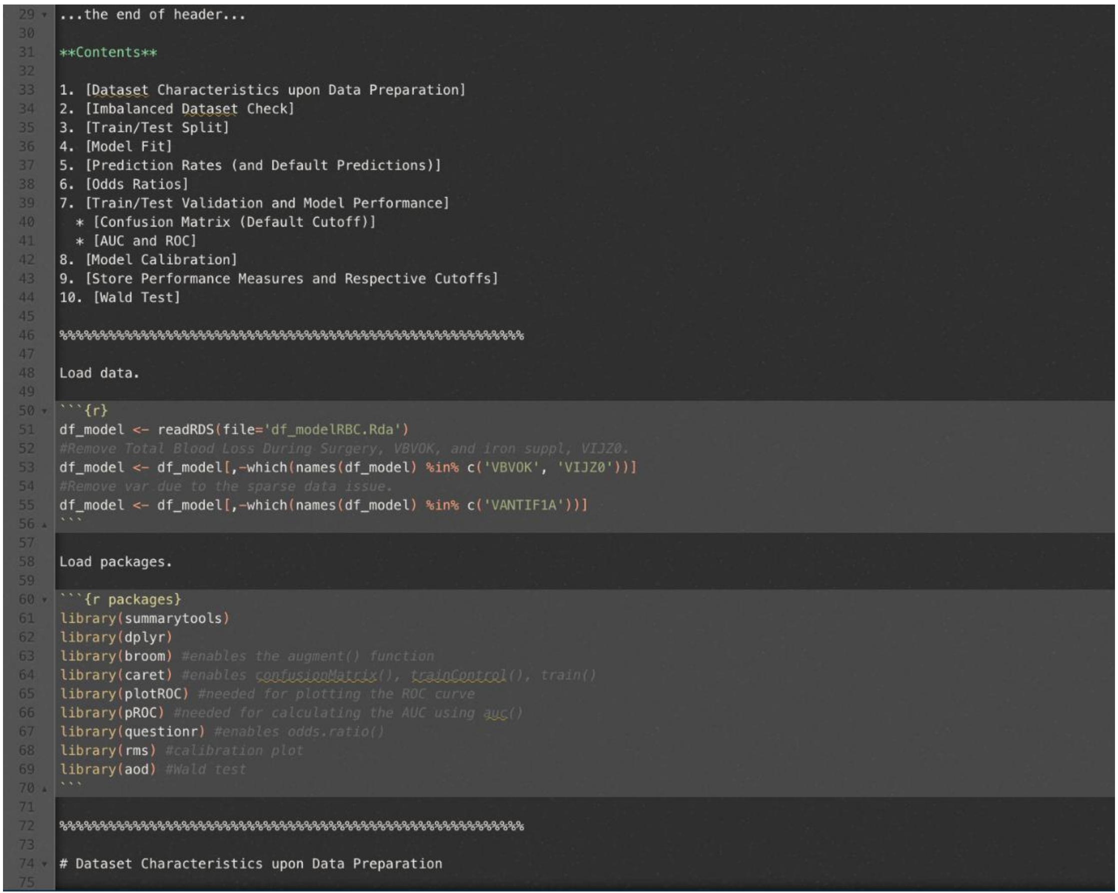
```{r}
#Proportion Anaemia_Pre_Op = 0, RBC = 1
table(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)[1,2]/sum(table(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)[1,])
```

```{r}
chisq.test(df_model$Anaemia_Pre_Op, df_model$RBC_Transfusion)
```
~~~~

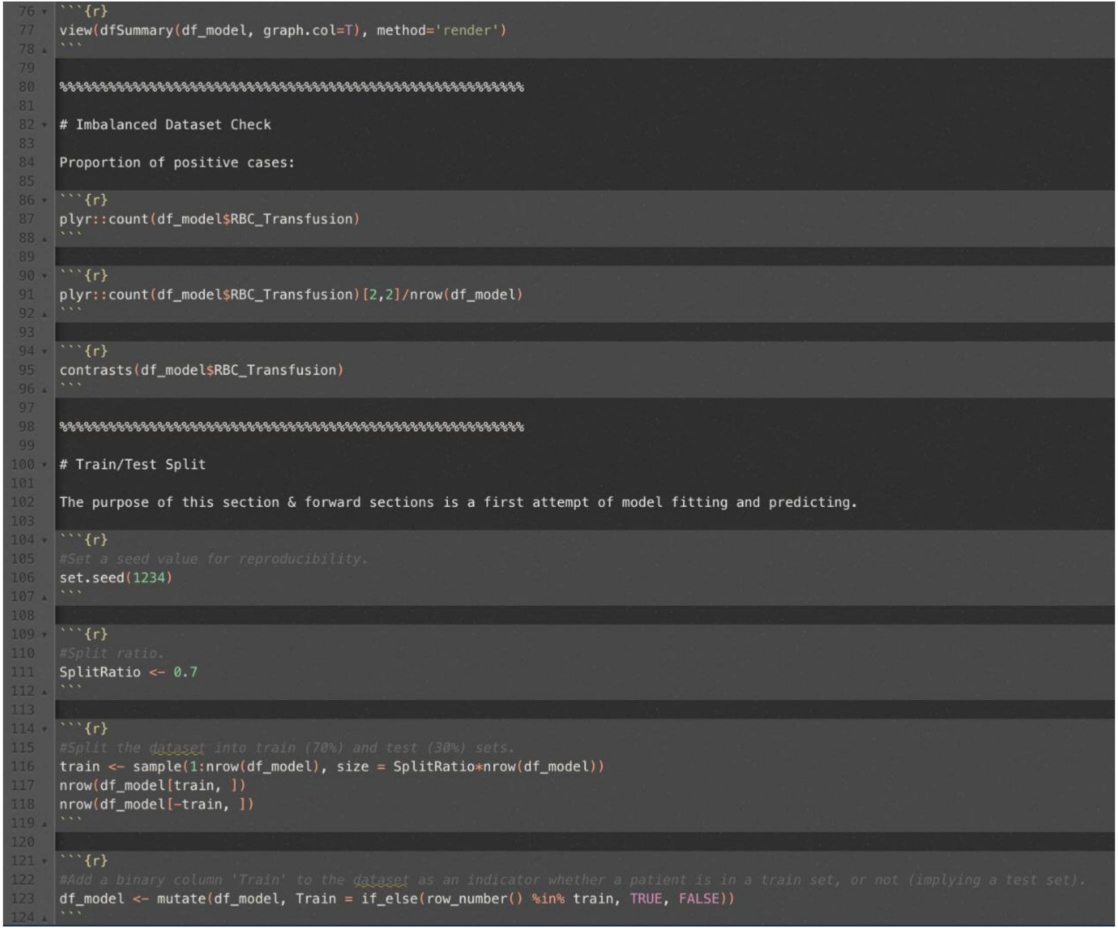


~~~
```{r}
#partial dependence plot for Anaemia_Pre_Op
pdp_Anaem_inv <- boot::inv.logit(partialPlot(x = rf_tuned, Trainset, x.var = Anaemia_Pre_Op)$y)
pdp_Anaem <- partialPlot(x = rf_tuned, Trainset, x.var = Anaemia_Pre_Op, ylab="log(odds)")
```

```{r}
#Anaemia_Pre_Op = 1
pdp_Anaem$x[2]
pdp_Anaem_inv[2]
odds.pdp_Anaem.1 <- (pdp_Anaem_inv[2]/(1-pdp_Anaem_inv[2]))
odds.pdp_Anaem.1

#Alternatively:
pdp_Anaem$y[2]
odds.pdp_Anaem.1 <- exp(pdp_Anaem$y[2])

odds.pdp_Anaem.1
```

```{r}
#Anaemia_Pre_Op = 0
pdp_Anaem$x[1]
pdp_Anaem_inv[1]
odds.pdp_Anaem.0 <- (pdp_Anaem_inv[1]/(1-pdp_Anaem_inv[1]))
odds.pdp_Anaem.0

#Alternatively:
pdp_Anaem$y[1]
odds.pdp_Anaem.0 <- exp(pdp_Anaem$y[1])

odds.pdp_Anaem.0
```

```{r}
#odds ratio
odds.pdp_Anaem.0/odds.pdp_Anaem.1
```

```{r}
rf_tuned %>%
  pdp::partial(pred.var=c("Anaemia_Pre_Op"),chull=TRUE,progress=TRUE) %>%
  autoplot(contour=TRUE,
           ylab = 'marginal effect, log(odds)')#,
           #geom_text(label=round(pdp3$y,2)))
```

## PDP for `Age` (Tuned)

```{r}
#partial dependence plot for Age
#note: n.pt is useful for continuous variables
pdp_Age_inv <- boot::inv.logit(partialPlot(x = rf_tuned, Trainset, x.var = Age,
                                           n.pt = min(length(unique(Trainset[, 'Age'])), 51))$y)
pdp_Age <- partialPlot(x = rf_tuned, Trainset, x.var = Age)
partial(rf_tuned, pred.var = 'Age', plot=T)
```

```{r}
#Age = 60
odds.pdp_Age.60 <- (pdp_Age_inv[29]/(1-pdp_Age_inv[29]))
pdp_Age$x[29]
odds.pdp_Age.60
#Age = 50
odds.pdp_Age.50 <- (pdp_Age_inv[22]/(1-pdp_Age_inv[22]))
pdp_Age$x[22]
odds.pdp_Age.50
#odds ratio
odds.pdp_Age.60/odds.pdp_Age.50
```

```{r}
#Age = 63
odds.pdp_Age.63 <- (pdp_Age_inv[31]/(1-pdp_Age_inv[31]))
pdp_Age$x[31]
odds.pdp_Age.63
#Age = 50
odds.pdp_Age.50 <- (pdp_Age_inv[22]/(1-pdp_Age_inv[22]))
pdp_Age$x[22]
odds.pdp_Age.50
#odds ratio
odds.pdp_Age.63/odds.pdp_Age.50
```

```{r}
rf_tuned %>%
  pdp::partial(pred.var=c("Age"),chull=TRUE,progress=TRUE) %>%
  autoplot(contour=TRUE,
           ylab = 'marginal effect, log(odds)')
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

...the end...
~~~

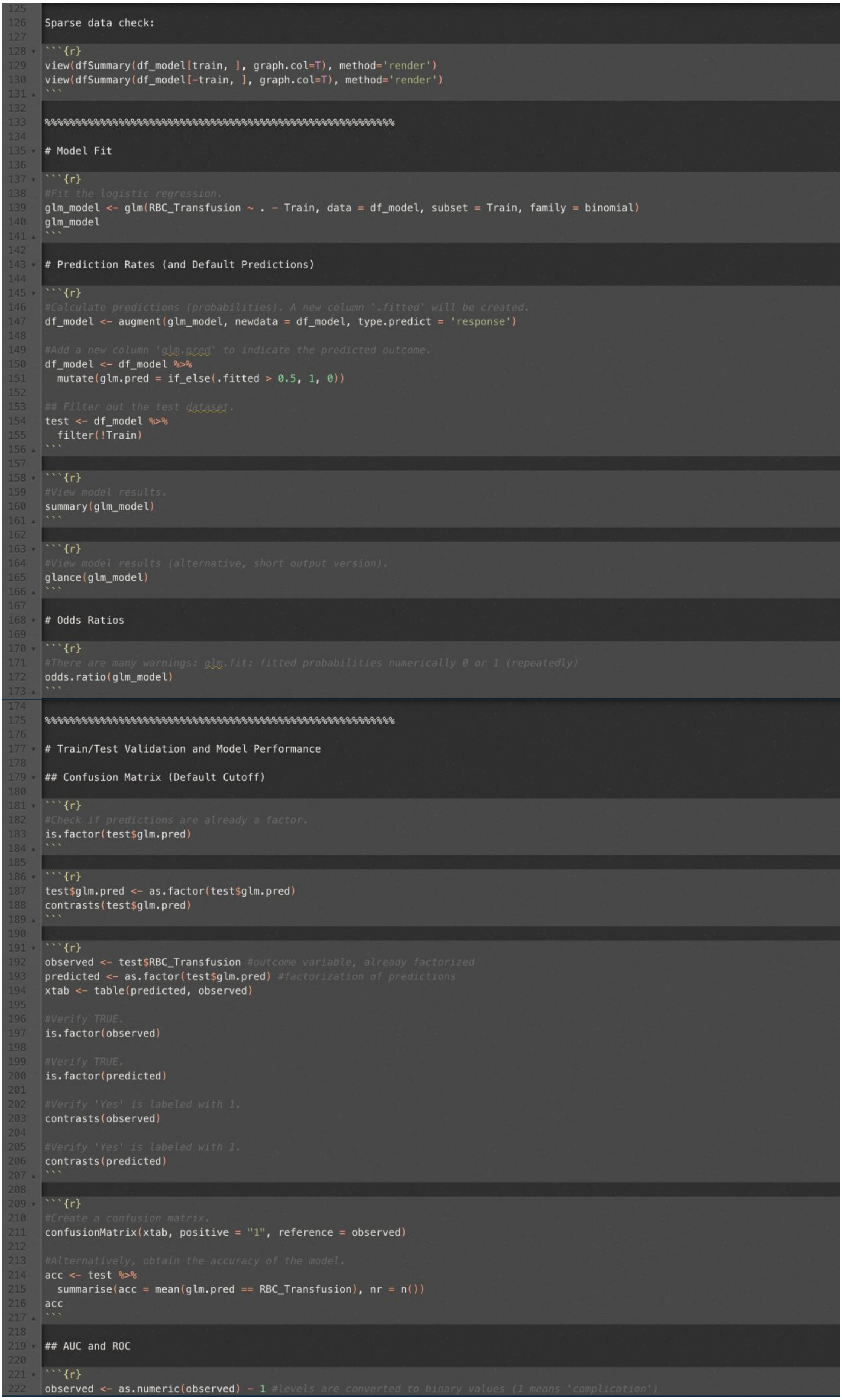
Programming segment #2: The R code of Logistic Regression, Case RBC[LREG], is presented below.



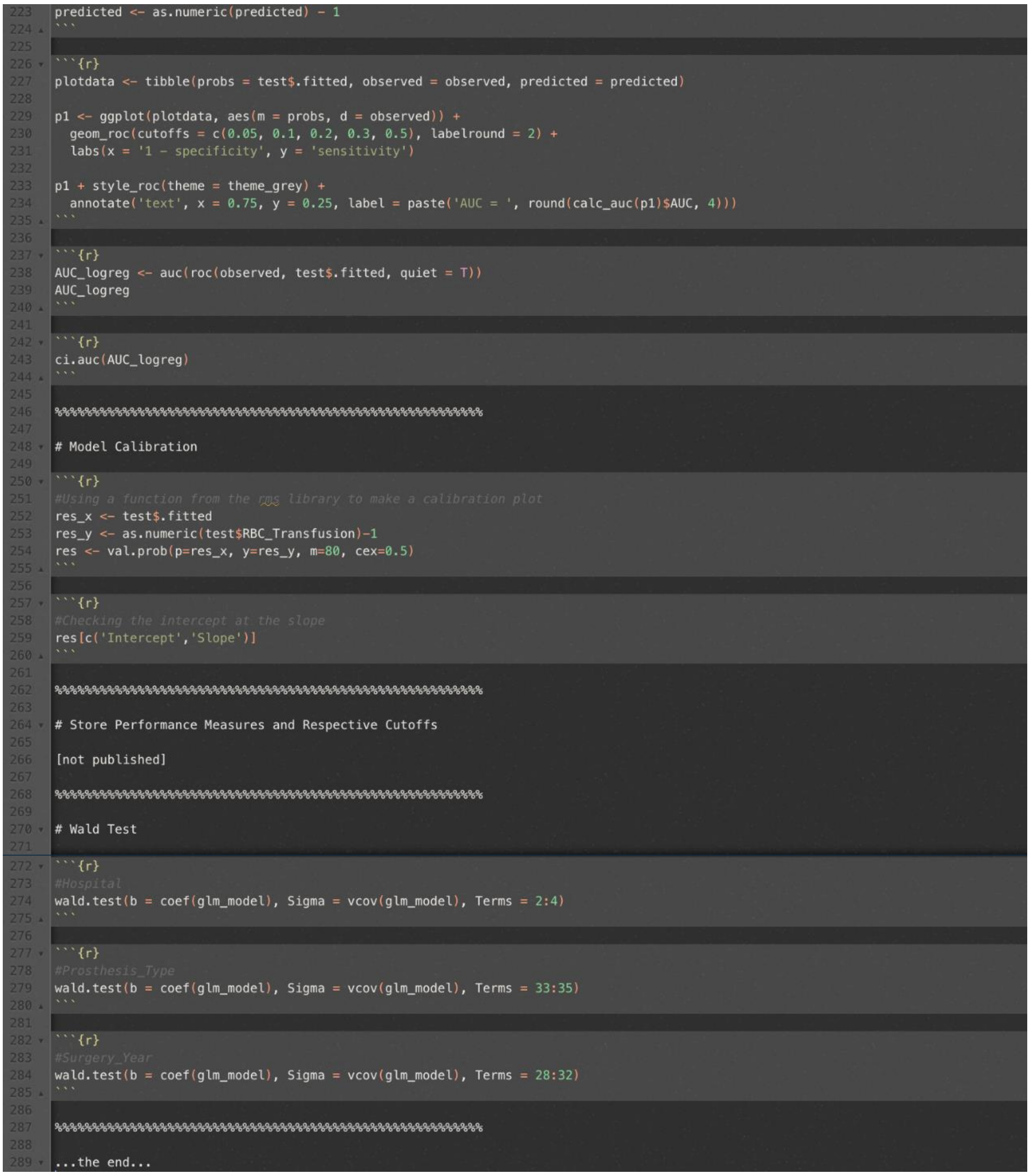
~~~
...the end of header...

**Contents**

1. [Dataset Characteristics upon Data Preparation]
2. [Imbalanced Dataset Check]
3. [Train/Test Split]
4. [Model Fit]
5. [Prediction Rates (and Default Predictions)]
6. [Odds Ratios]
7. [Train/Test Validation and Model Performance]
  * [Confusion Matrix (Default Cutoff)]
  * [AUC and ROC]
8. [Model Calibration]
9. [Store Performance Measures and Respective Cutoffs]
10. [Wald Test]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Load data.

```{r}
df_model <- readRDS(file='df_modelRBC.Rda')
#Remove Total Blood Loss During Surgery, VBVOK, and iron suppl, VIJZ0.
df_model <- df_model[,-which(names(df_model) %in% c('VBVOK', 'VIJZ0'))]
#Remove var due to the sparse data issue.
df_model <- df_model[,-which(names(df_model) %in% c('VANTIF1A'))]
```

Load packages.

```{r packages}
library(summarytools)
library(dplyr)
library(broom) #enables the augment() function
library(caret) #enables confusionMatrix(), trainControl(), train()
library(plotROC) #needed for plotting the ROC curve
library(pROC) #needed for calculating the AUC using auc()
library(questionr) #enables odds.ratio()
library(rms) #calibration plot
library(aod) #Wald test
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Dataset Characteristics upon Data Preparation

~~~



~~~
```{r}
view(dfSummary(df_model, graph.col=T), method='render')
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Imbalanced Dataset Check

Proportion of positive cases:

```{r}
plyr::count(df_model$RBC_Transfusion)
```

```{r}
plyr::count(df_model$RBC_Transfusion)[2,2]/nrow(df_model)
```

```{r}
contrasts(df_model$RBC_Transfusion)
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Train/Test Split

The purpose of this section & forward sections is a first attempt of model fitting and predicting.

```{r}
#Set a seed value for reproducibility.
set.seed(1234)
```

```{r}
#Split ratio.
SplitRatio <- 0.7
```

```{r}
#Split the dataset into train (70%) and test (30%) sets.
train <- sample(1:nrow(df_model), size = SplitRatio*nrow(df_model))
nrow(df_model[train, ])
nrow(df_model[-train, ])
```

```{r}
#Add a binary column 'Train' to the dataset as an indicator whether a patient is in a train set, or not (implying a test set).
df_model <- mutate(df_model, Train = if_else(row_number() %in% train, TRUE, FALSE))
```
~~~

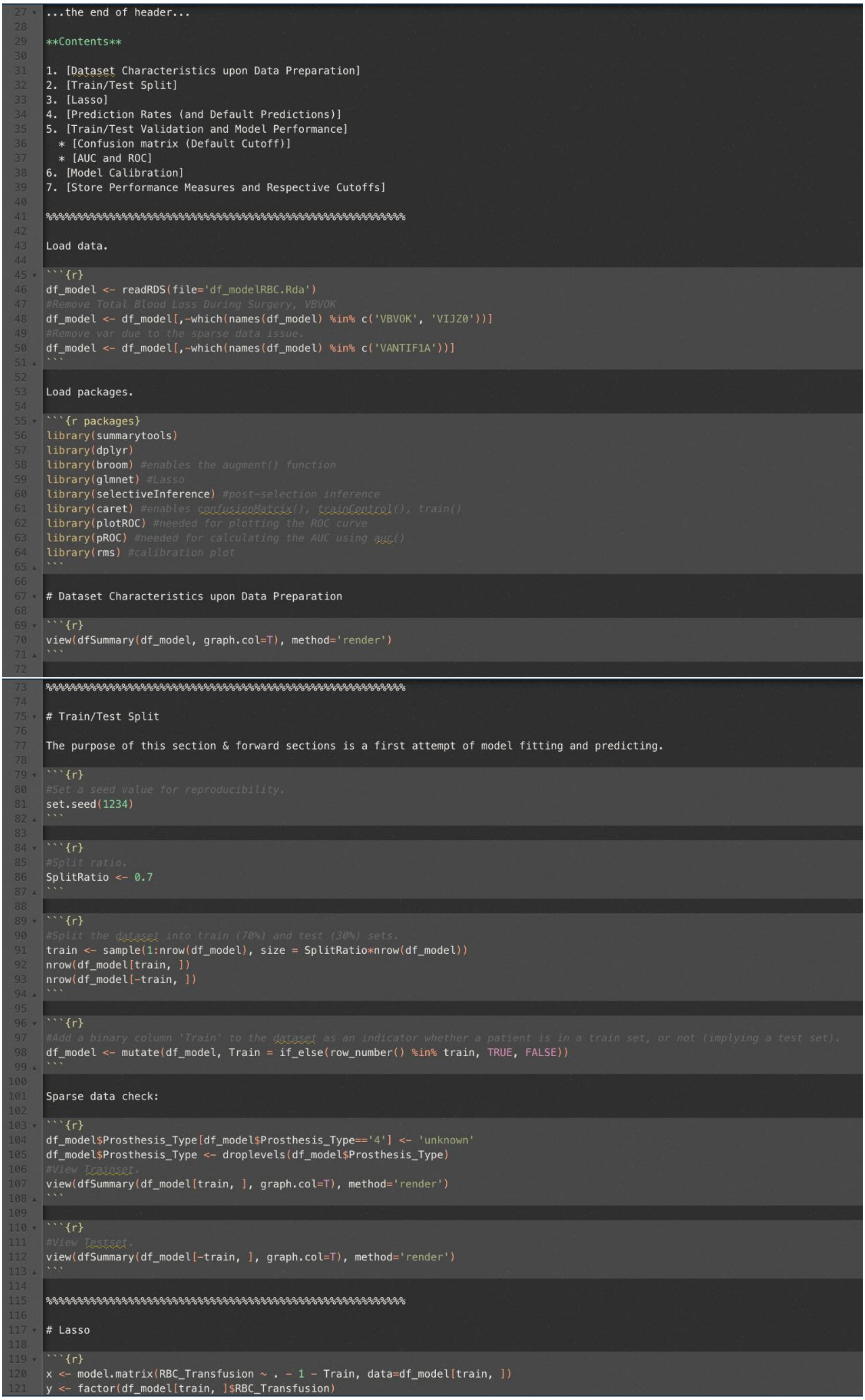


~~~
Sparse data check:

```{r}
view(dfSummary(df_model[train, ], graph.col=T), method='render')
view(dfSummary(df_model[-train, ], graph.col=T), method='render')
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Model Fit

```{r}
#Fit the logistic regression.
glm_model <- glm(RBC_Transfusion ~ . - Train, data = df_model, subset = Train, family = binomial)
glm_model
```

# Prediction Rates (and Default Predictions)

```{r}
#Calculate predictions (probabilities). A new column '.fitted' will be created.
df_model <- augment(glm_model, newdata = df_model, type.predict = 'response')

#Add a new column 'glm.pred' to indicate the predicted outcome.
df_model <- df_model %>%
  mutate(glm.pred = if_else(.fitted > 0.5, 1, 0))

## Filter out the test dataset.
test <- df_model %>%
  filter(!Train)
```

```{r}
#View model results.
summary(glm_model)
```

```{r}
#View model results (alternative, short output version).
glance(glm_model)
```

# Odds Ratios

```{r}
#There are many warnings: glm.fit: fitted probabilities numerically 0 or 1 (repeatedly)
odds.ratio(glm_model)
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Train/Test Validation and Model Performance

## Confusion Matrix (Default Cutoff)

```{r}
#Check if predictions are already a factor.
is.factor(test$glm.pred)
```

```{r}
test$glm.pred <- as.factor(test$glm.pred)
contrasts(test$glm.pred)
```

```{r}
observed <- test$RBC_Transfusion #outcome variable, already factorized
predicted <- as.factor(test$glm.pred) #factorization of predictions
xtab <- table(predicted, observed)

#Verify TRUE.
is.factor(observed)

#Verify TRUE.
is.factor(predicted)

#Verify 'Yes' is labeled with 1.
contrasts(observed)

#Verify 'Yes' is labeled with 1.
contrasts(predicted)
```

```{r}
#Create a confusion matrix.
confusionMatrix(xtab, positive = "1", reference = observed)

#Alternatively, obtain the accuracy of the model.
acc <- test %>%
  summarise(acc = mean(glm.pred == RBC_Transfusion), nr = n())
acc
```

## AUC and ROC

```{r}
observed <- as.numeric(observed) - 1 #levels are converted to binary values (1 means 'complication')
~~~

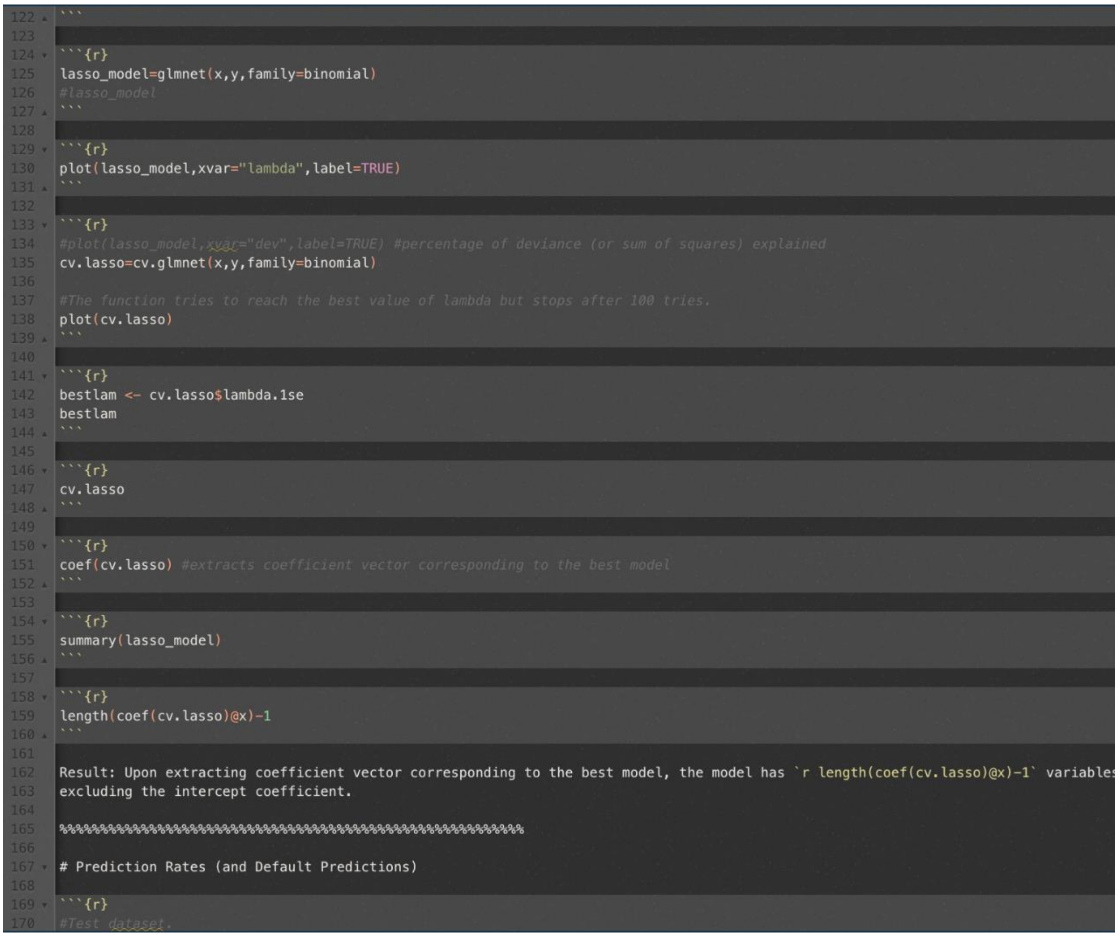


~~~
predicted <- as.numeric(predicted) - 1
```

```{r}
plotdata <- tibble(probs = test$.fitted, observed = observed, predicted = predicted)

p1 <- ggplot(plotdata, aes(m = probs, d = observed)) +
  geom_roc(cutoffs = c(0.05, 0.1, 0.2, 0.3, 0.5), labelround = 2) +
  labs(x = '1 - specificity', y = 'sensitivity')

p1 + style_roc(theme = theme_grey) +
  annotate('text', x = 0.75, y = 0.25, label = paste('AUC = ', round(calc_auc(p1)$AUC, 4)))
```

```{r}
AUC_logreg <- auc(roc(observed, test$.fitted, quiet = T))
AUC_logreg
```

```{r}
ci.auc(AUC_logreg)
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Model Calibration

```{r}
#Using a function from the rms library to make a calibration plot
res_x <- test$.fitted
res_y <- as.numeric(test$RBC_Transfusion)-1
res <- val.prob(p=res_x, y=res_y, m=80, cex=0.5)
```

```{r}
#Checking the intercept at the slope
res[c('Intercept','Slope')]
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Store Performance Measures and Respective Cutoffs

[not published]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Wald Test

```{r}
#Hospital
wald.test(b = coef(glm_model), Sigma = vcov(glm_model), Terms = 2:4)
```

```{r}
#Prosthesis_Type
wald.test(b = coef(glm_model), Sigma = vcov(glm_model), Terms = 33:35)
```

```{r}
#Surgery_Year
wald.test(b = coef(glm_model), Sigma = vcov(glm_model), Terms = 28:32)
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

...the end...
~~~

Programming segment #3: The R code of Lasso, Case RBC[lasso], is presented below.



~~~
...the end of header...

**Contents**

1. [Dataset Characteristics upon Data Preparation]
2. [Train/Test Split]
3. [Lasso]
4. [Prediction Rates (and Default Predictions)]
5. [Train/Test Validation and Model Performance]
  * [Confusion matrix (Default Cutoff)]
  * [AUC and ROC]
6. [Model Calibration]
7. [Store Performance Measures and Respective Cutoffs]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Load data.

```{r}
df_model <- readRDS(file='df_modelRBC.Rda')
#Remove Total Blood Loss During Surgery, VBVOK
df_model <- df_model[,-which(names(df_model) %in% c('VBVOK', 'VIJZ0'))]
#Remove var due to the sparse data issue.
df_model <- df_model[,-which(names(df_model) %in% c('VANTIF1A'))]
```

Load packages.

```{r packages}
library(summarytools)
library(dplyr)
library(broom) #enables the augment() function
library(glmnet) #Lasso
library(selectiveInference) #post-selection inference
library(caret) #enables confusionMatrix(), trainControl(), train()
library(plotROC) #needed for plotting the ROC curve
library(pROC) #needed for calculating the AUC using auc()
library(rms) #calibration plot
```

# Dataset Characteristics upon Data Preparation

```{r}
view(dfSummary(df_model, graph.col=T), method='render')
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Train/Test Split

The purpose of this section & forward sections is a first attempt of model fitting and predicting.

```{r}
#Set a seed value for reproducibility.
set.seed(1234)
```

```{r}
#Split ratio.
SplitRatio <- 0.7
```

```{r}
#Split the dataset into train (70%) and test (30%) sets.
train <- sample(1:nrow(df_model), size = SplitRatio*nrow(df_model))
nrow(df_model[train, ])
nrow(df_model[-train, ])
```

```{r}
#Add a binary column 'Train' to the dataset as an indicator whether a patient is in a train set, or not (implying a test set).
df_model <- mutate(df_model, Train = if_else(row_number() %in% train, TRUE, FALSE))
```

Sparse data check:

```{r}
df_model$Prosthesis_Type[df_model$Prosthesis_Type=='4'] <- 'unknown'
df_model$Prosthesis_Type <- droplevels(df_model$Prosthesis_Type)
#View Trainset.
view(dfSummary(df_model[train, ], graph.col=T), method='render')
```

```{r}
#View Testset.
view(dfSummary(df_model[-train, ], graph.col=T), method='render')
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Lasso

```{r}
x <- model.matrix(RBC_Transfusion ~ . - 1 - Train, data=df_model[train, ])
y <- factor(df_model[train, ]$RBC_Transfusion)
~~~



```
```

```{r}
lasso_model=glmnet(x,y,family=binomial)
#lasso_model
```

```{r}
plot(lasso_model,xvar="lambda",label=TRUE)
```

```{r}
#plot(lasso_model,xvar="dev",label=TRUE) #percentage of deviance (or sum of squares) explained
cv.lasso=cv.glmnet(x,y,family=binomial)

#The function tries to reach the best value of lambda but stops after 100 tries.
plot(cv.lasso)
```

```{r}
bestlam <- cv.lasso$lambda.1se
bestlam
```

```{r}
cv.lasso
```

```{r}
coef(cv.lasso) #extracts coefficient vector corresponding to the best model
```

```{r}
summary(lasso_model)
```

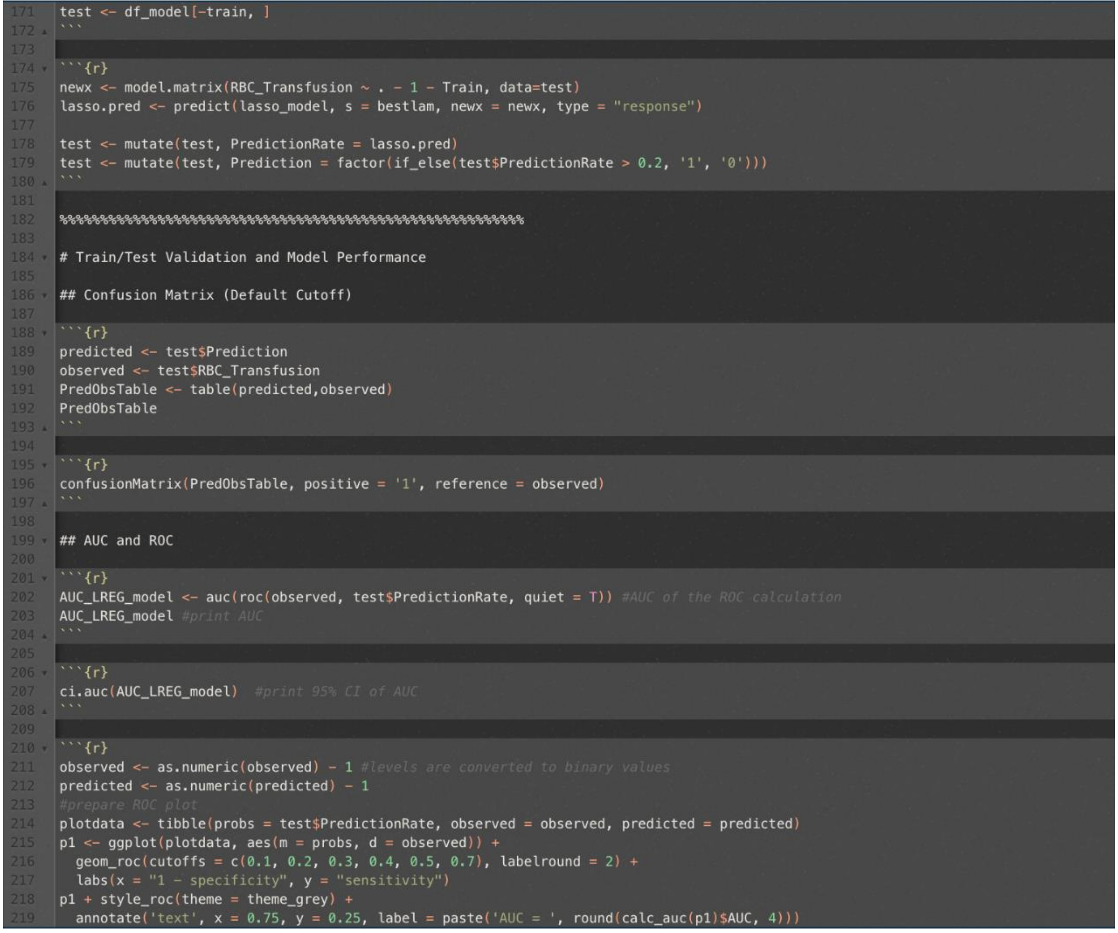
```{r}
length(coef(cv.lasso)@x)-1
```

Result: Upon extracting coefficient vector corresponding to the best model, the model has `r length(coef(cv.lasso)@x)-1` variables
excluding the intercept coefficient.

%%%

Prediction Rates (and Default Predictions)

```{r}
#Test dataset.
```



```
test <- df_model[-train, ]
```

```{r}
newx <- model.matrix(RBC_Transfusion ~ . - 1 - Train, data=test)
lasso.pred <- predict(lasso_model, s = bestlam, newx = newx, type = "response")

test <- mutate(test, PredictionRate = lasso.pred)
test <- mutate(test, Prediction = factor(if_else(test$PredictionRate > 0.2, '1', '0')))
```

%%%

Train/Test Validation and Model Performance

Confusion Matrix (Default Cutoff)

```{r}
predicted <- test$Prediction
observed <- test$RBC_Transfusion
PredObsTable <- table(predicted,observed)
PredObsTable
```

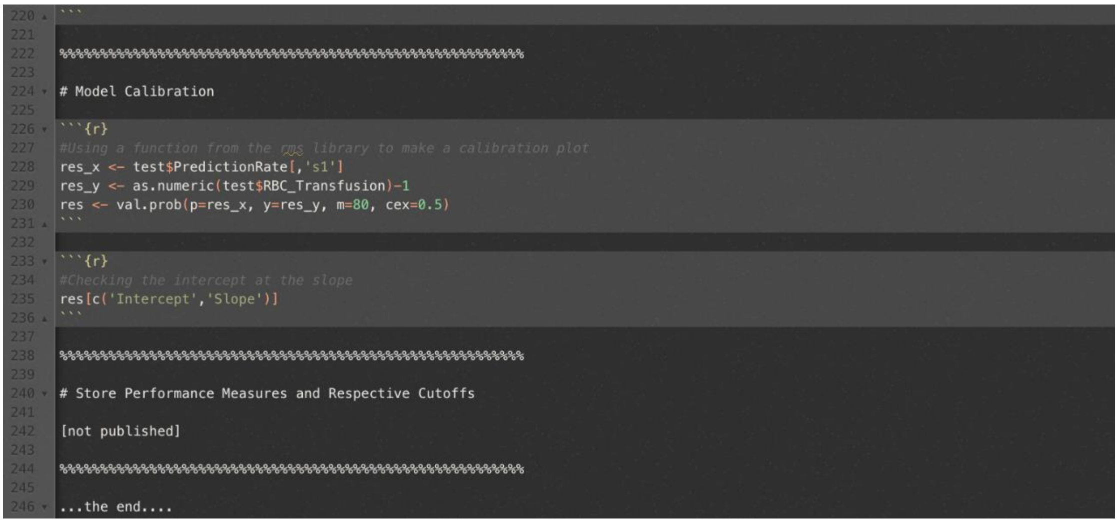
```{r}
confusionMatrix(PredObsTable, positive = '1', reference = observed)
```

AUC and ROC

```{r}
AUC_LREG_model <- auc(roc(observed, test$PredictionRate, quiet = T)) #AUC of the ROC calculation
AUC_LREG_model #print AUC
```

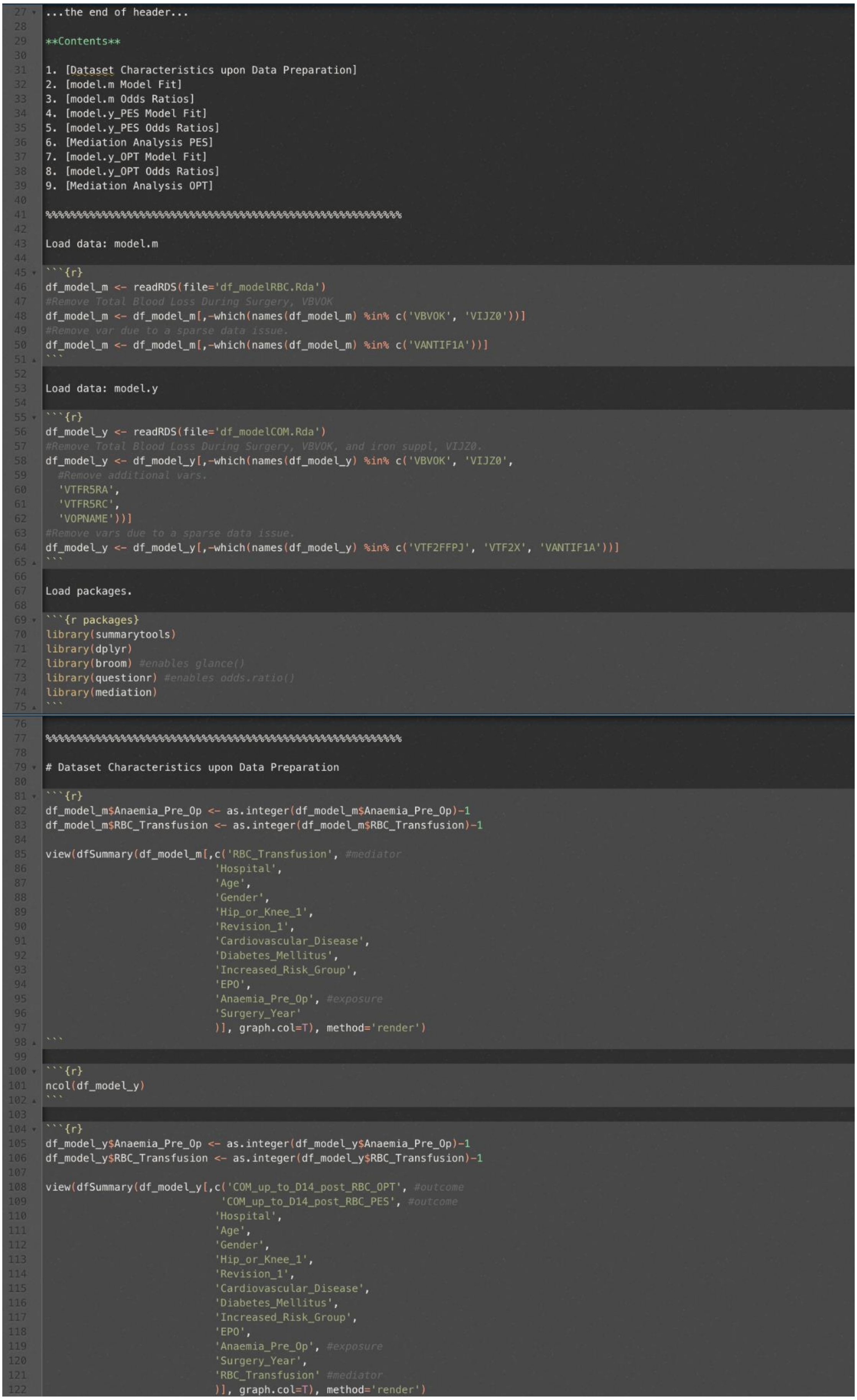
```{r}
ci.auc(AUC_LREG_model)  #print 95% CI of AUC
```

```{r}
observed <- as.numeric(observed) - 1 #levels are converted to binary values
predicted <- as.numeric(predicted) - 1
#prepare ROC plot
plotdata <- tibble(probs = test$PredictionRate, observed = observed, predicted = predicted)
p1 <- ggplot(plotdata, aes(m = probs, d = observed)) +
  geom_roc(cutoffs = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.7), labelround = 2) +
  labs(x = "1 - specificity", y = "sensitivity")
p1 + style_roc(theme = theme_grey) +
  annotate('text', x = 0.75, y = 0.25, label = paste('AUC = ', round(calc_auc(p1)$AUC, 4)))
```

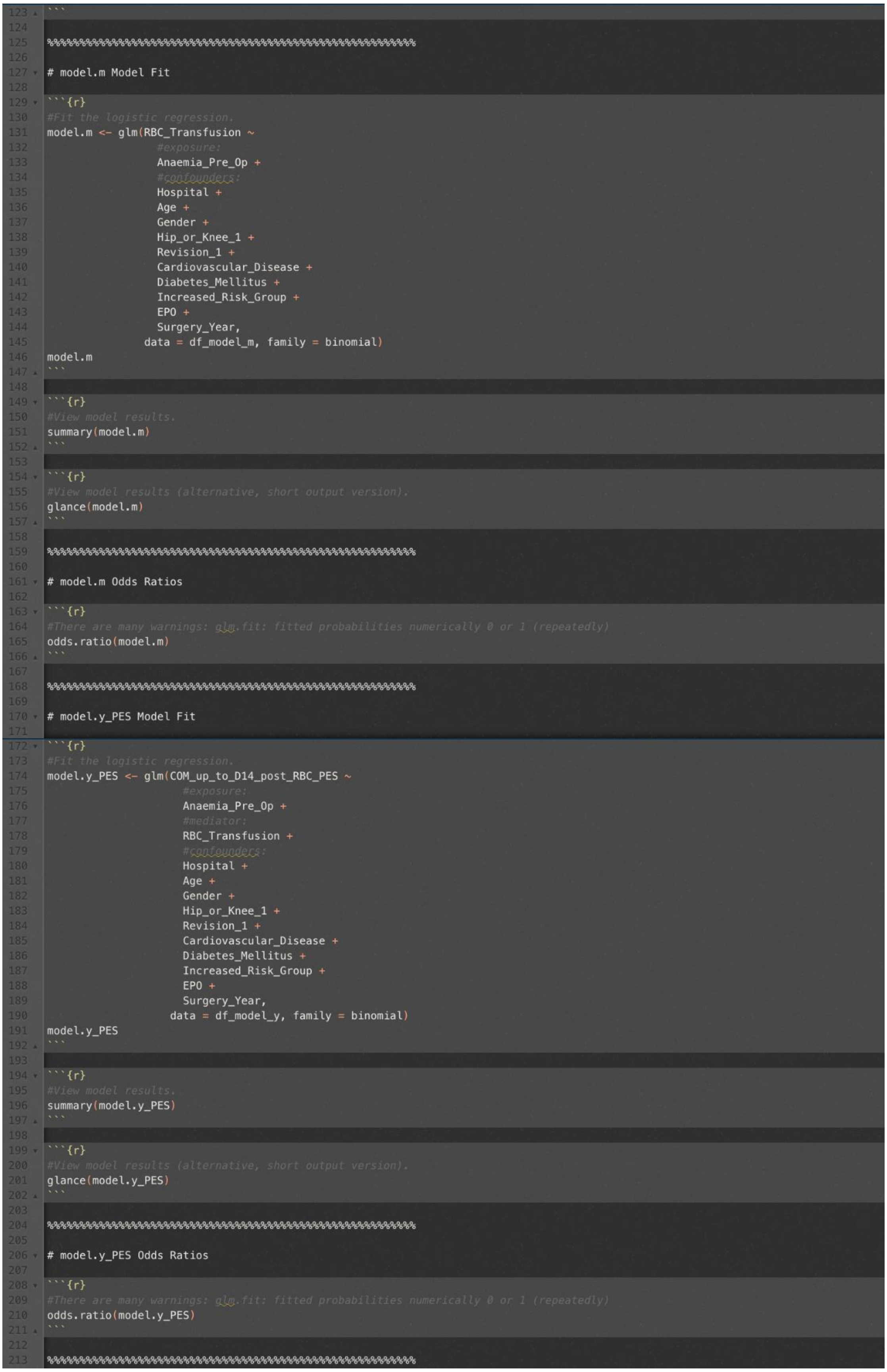


~~~
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Model Calibration

```{r}
#Using a function from the rms library to make a calibration plot
res_x <- test$PredictionRate[,'s1']
res_y <- as.numeric(test$RBC_Transfusion)-1
res <- val.prob(p=res_x, y=res_y, m=80, cex=0.5)
```

```{r}
#Checking the intercept at the slope
res[c('Intercept','Slope')]
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Store Performance Measures and Respective Cutoffs

[not published]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

...the end....
~~~

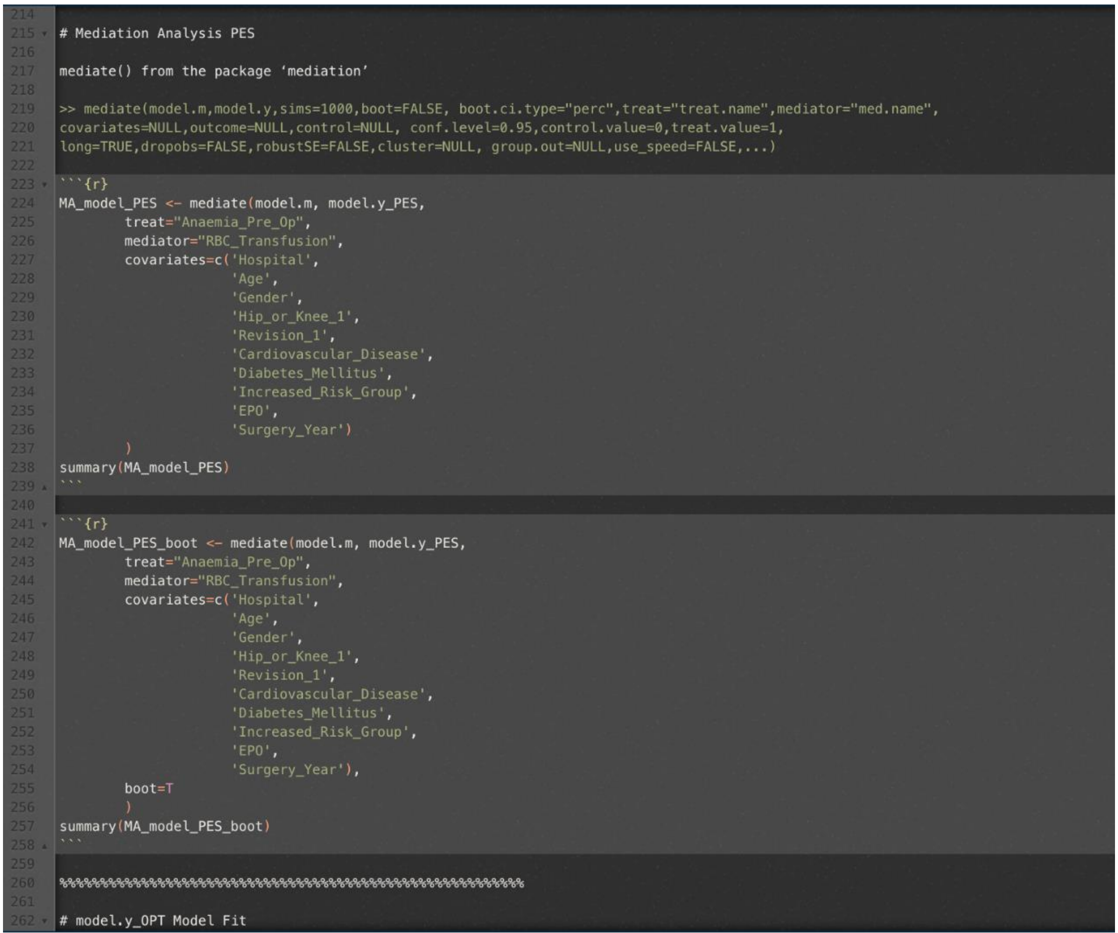
Programming segment #4: The R code of Mediation Analysis is presented below.



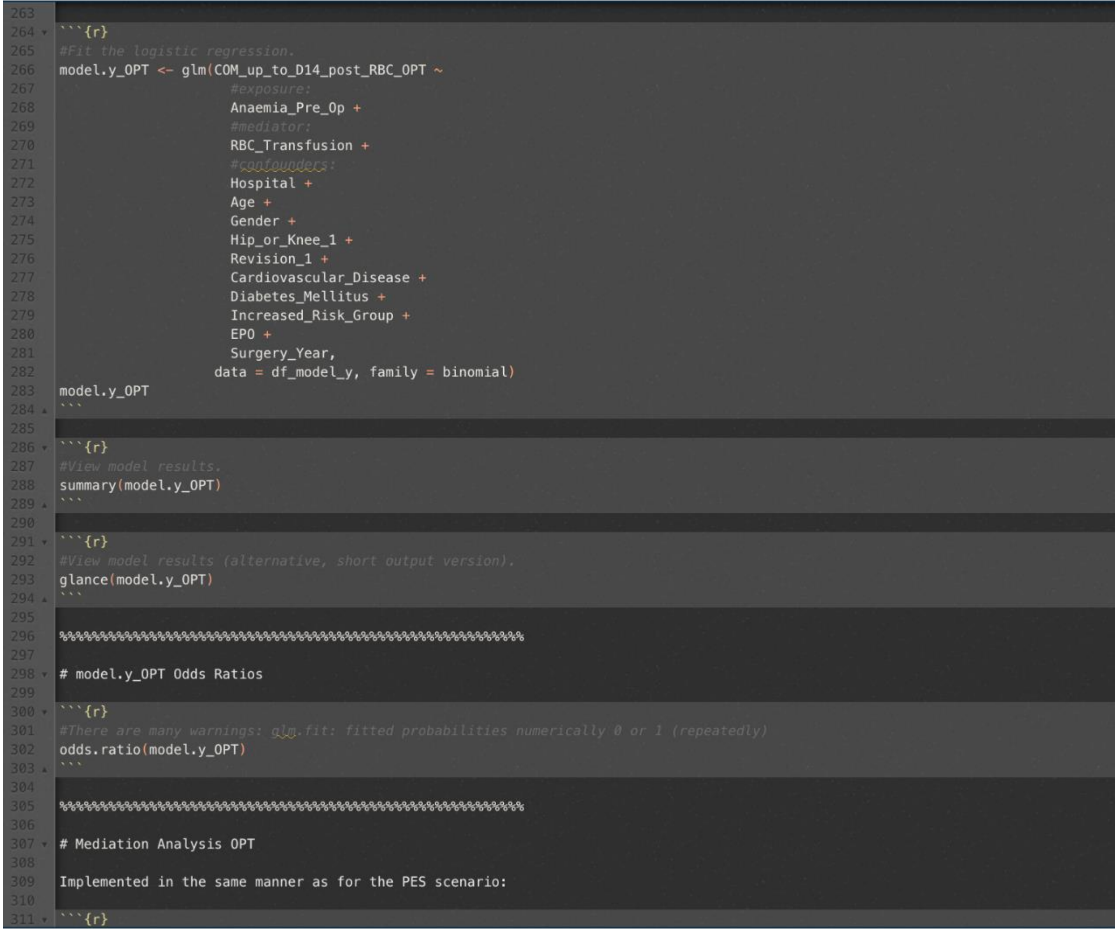
~~~
...the end of header...

**Contents**

1. [Dataset Characteristics upon Data Preparation]
2. [model.m Model Fit]
3. [model.m Odds Ratios]
4. [model.y_PES Model Fit]
5. [model.y_PES Odds Ratios]
6. [Mediation Analysis PES]
7. [model.y_OPT Model Fit]
8. [model.y_OPT Odds Ratios]
9. [Mediation Analysis OPT]

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

Load data: model.m

```{r}
df_model_m <- readRDS(file='df_modelRBC.Rda')
#Remove Total Blood Loss During Surgery, VBVOK
df_model_m <- df_model_m[,-which(names(df_model_m) %in% c('VBVOK', 'VIJZ0'))]
#Remove var due to a sparse data issue.
df_model_m <- df_model_m[,-which(names(df_model_m) %in% c('VANTIF1A'))]
```

Load data: model.y

```{r}
df_model_y <- readRDS(file='df_modelCOM.Rda')
#Remove Total Blood Loss During Surgery, VBVOK, and iron suppl, VIJZ0.
df_model_y <- df_model_y[,-which(names(df_model_y) %in% c('VBVOK', 'VIJZ0',
  #Remove additional vars.
  'VTFR5RA',
  'VTFR5RC',
  'VOPNAME'))]
#Remove vars due to a sparse data issue.
df_model_y <- df_model_y[,-which(names(df_model_y) %in% c('VTF2FFPJ', 'VTF2X', 'VANTIF1A'))]
```

Load packages.

```{r packages}
library(summarytools)
library(dplyr)
library(broom) #enables glance()
library(questionr) #enables odds.ratio()
library(mediation)
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# Dataset Characteristics upon Data Preparation

```{r}
df_model_m$Anaemia_Pre_Op <- as.integer(df_model_m$Anaemia_Pre_Op)-1
df_model_m$RBC_Transfusion <- as.integer(df_model_m$RBC_Transfusion)-1

view(dfSummary(df_model_m[,c('RBC_Transfusion', #mediator
                           'Hospital',
                           'Age',
                           'Gender',
                           'Hip_or_Knee_1',
                           'Revision_1',
                           'Cardiovascular_Disease',
                           'Diabetes_Mellitus',
                           'Increased_Risk_Group',
                           'EPO',
                           'Anaemia_Pre_Op', #exposure
                           'Surgery_Year'
                           )], graph.col=T), method='render')
```

```{r}
ncol(df_model_y)
```

```{r}
df_model_y$Anaemia_Pre_Op <- as.integer(df_model_y$Anaemia_Pre_Op)-1
df_model_y$RBC_Transfusion <- as.integer(df_model_y$RBC_Transfusion)-1

view(dfSummary(df_model_y[,c('COM_up_to_D14_post_RBC_OPT', #outcome
                            'COM_up_to_D14_post_RBC_PES', #outcome
                           'Hospital',
                           'Age',
                           'Gender',
                           'Hip_or_Knee_1',
                           'Revision_1',
                           'Cardiovascular_Disease',
                           'Diabetes_Mellitus',
                           'Increased_Risk_Group',
                           'EPO',
                           'Anaemia_Pre_Op', #exposure
                           'Surgery_Year',
                           'RBC_Transfusion' #mediator
                           )], graph.col=T), method='render')
~~~



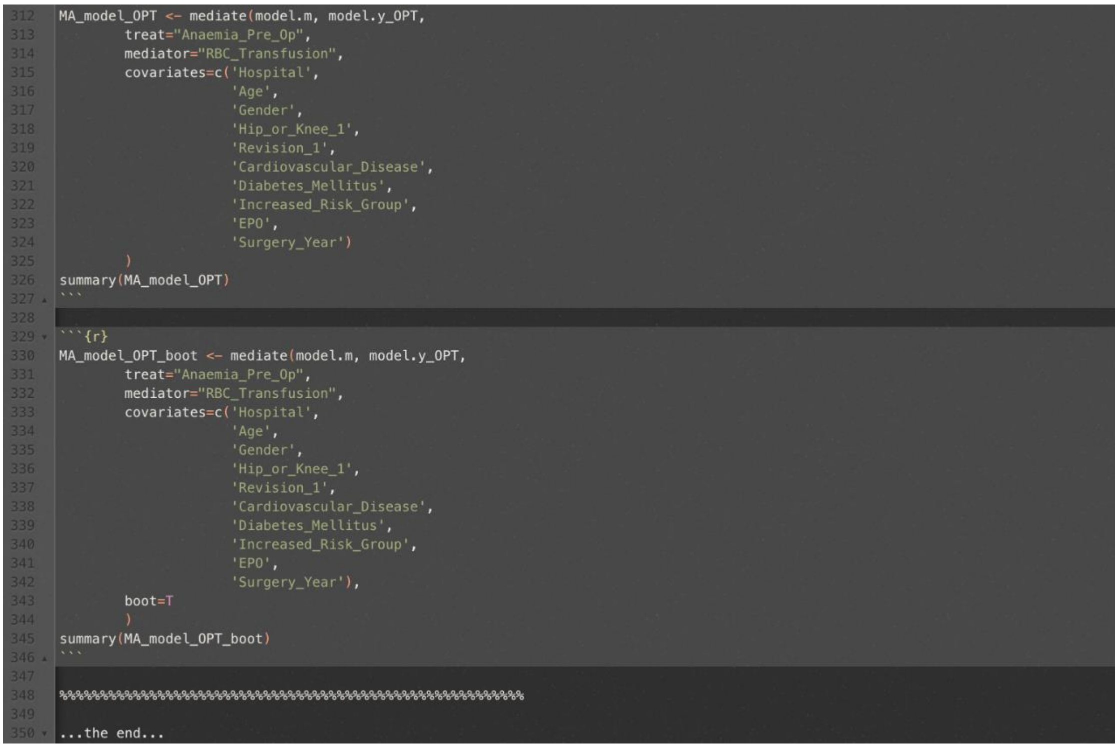
~~~
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# model.m Model Fit

```{r}
#Fit the logistic regression.
model.m <- glm(RBC_Transfusion ~
                 #exposure:
                 Anaemia_Pre_Op +
                 #confounders:
                 Hospital +
                 Age +
                 Gender +
                 Hip_or_Knee_1 +
                 Revision_1 +
                 Cardiovascular_Disease +
                 Diabetes_Mellitus +
                 Increased_Risk_Group +
                 EPO +
                 Surgery_Year,
               data = df_model_m, family = binomial)
model.m
```

```{r}
#View model results.
summary(model.m)
```

```{r}
#View model results (alternative, short output version).
glance(model.m)
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# model.m Odds Ratios

```{r}
#There are many warnings: glm.fit: fitted probabilities numerically 0 or 1 (repeatedly)
odds.ratio(model.m)
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# model.y_PES Model Fit

```{r}
#Fit the logistic regression.
model.y_PES <- glm(COM_up_to_D14_post_RBC_PES ~
                     #exposure:
                     Anaemia_Pre_Op +
                     #mediator:
                     RBC_Transfusion +
                     #confounders:
                     Hospital +
                     Age +
                     Gender +
                     Hip_or_Knee_1 +
                     Revision_1 +
                     Cardiovascular_Disease +
                     Diabetes_Mellitus +
                     Increased_Risk_Group +
                     EPO +
                     Surgery_Year,
                   data = df_model_y, family = binomial)
model.y_PES
```

```{r}
#View model results.
summary(model.y_PES)
```

```{r}
#View model results (alternative, short output version).
glance(model.y_PES)
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

# model.y_PES Odds Ratios

```{r}
#There are many warnings: glm.fit: fitted probabilities numerically 0 or 1 (repeatedly)
odds.ratio(model.y_PES)
```

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
~~~



~~~

Mediation Analysis PES

mediate() from the package ‘mediation’

>> mediate(model.m,model.y,sims=1000,boot=FALSE, boot.ci.type="perc",treat="treat.name",mediator="med.name",
covariates=NULL,outcome=NULL,control=NULL, conf.level=0.95,control.value=0,treat.value=1,
long=TRUE,dropobs=FALSE,robustSE=FALSE,cluster=NULL, group.out=NULL,use_speed=FALSE,...)

```{r}
MA_model_PES <- mediate(model.m, model.y_PES,
        treat="Anaemia_Pre_Op",
        mediator="RBC_Transfusion",
        covariates=c('Hospital',
                     'Age',
                     'Gender',
                     'Hip_or_Knee_1',
                     'Revision_1',
                     'Cardiovascular_Disease',
                     'Diabetes_Mellitus',
                     'Increased_Risk_Group',
                     'EPO',
                     'Surgery_Year')
        )
summary(MA_model_PES)
```

```{r}
MA_model_PES_boot <- mediate(model.m, model.y_PES,
        treat="Anaemia_Pre_Op",
        mediator="RBC_Transfusion",
        covariates=c('Hospital',
                     'Age',
                     'Gender',
                     'Hip_or_Knee_1',
                     'Revision_1',
                     'Cardiovascular_Disease',
                     'Diabetes_Mellitus',
                     'Increased_Risk_Group',
                     'EPO',
                     'Surgery_Year'),
        boot=T
        )
summary(MA_model_PES_boot)
```

%%%

model.y_OPT Model Fit
~~~



~~~

```{r}
#Fit the logistic regression.
model.y_OPT <- glm(COM_up_to_D14_post_RBC_OPT ~
                     #exposure:
                     Anaemia_Pre_Op +
                     #mediator:
                     RBC_Transfusion +
                     #confounders:
                     Hospital +
                     Age +
                     Gender +
                     Hip_or_Knee_1 +
                     Revision_1 +
                     Cardiovascular_Disease +
                     Diabetes_Mellitus +
                     Increased_Risk_Group +
                     EPO +
                     Surgery_Year,
                   data = df_model_y, family = binomial)
model.y_OPT
```

```{r}
#View model results.
summary(model.y_OPT)
```

```{r}
#View model results (alternative, short output version).
glance(model.y_OPT)
```

%%%

model.y_OPT Odds Ratios

```{r}
#There are many warnings: glm.fit: fitted probabilities numerically 0 or 1 (repeatedly)
odds.ratio(model.y_OPT)
```

%%%

Mediation Analysis OPT

Implemented in the same manner as for the PES scenario:

```{r}
~~~



~~~
MA_model_OPT <- mediate(model.m, model.y_OPT,
        treat="Anaemia_Pre_Op",
        mediator="RBC_Transfusion",
        covariates=c('Hospital',
                     'Age',
                     'Gender',
                     'Hip_or_Knee_1',
                     'Revision_1',
                     'Cardiovascular_Disease',
                     'Diabetes_Mellitus',
                     'Increased_Risk_Group',
                     'EPO',
                     'Surgery_Year')
        )
summary(MA_model_OPT)
```

```{r}
MA_model_OPT_boot <- mediate(model.m, model.y_OPT,
        treat="Anaemia_Pre_Op",
        mediator="RBC_Transfusion",
        covariates=c('Hospital',
                     'Age',
                     'Gender',
                     'Hip_or_Knee_1',
                     'Revision_1',
                     'Cardiovascular_Disease',
                     'Diabetes_Mellitus',
                     'Increased_Risk_Group',
                     'EPO',
                     'Surgery_Year'),
        boot=T
        )
summary(MA_model_OPT_boot)
```

%%%

...the end...
~~~

Appendix F: Literature Search Record

The following procedure of a literature search across multiple databases supports the argument that RBC transfusion was not previously studied as a mediator between pre-operative anaemia and patient (surgical) outcomes. Literature for comparing the results of this study (RQ#4) is absent.

The search was conducted across publications in the English language. No filters were applied on the document type.

Table F-1: Search string for literature search.

Prompt	Search string	Database*	Fields searched	Number of articles	Sorted based on
1	("red blood cell transfusion*" OR "RBC transfusion*") AND ("patient outcome*" OR "surgical outcome*") AND ("pre-operative anemia" OR "pre-operative anaemia" OR "preoperative anemia" OR "preoperative anaemia") AND ("mediation" OR “mediator”)	Scopus	Article title, Abstract, Keywords	0	-
2		Web of Science	Abstract	0	-
3		Web of Science	All fields	0	-
4		PubMed	All fields	0	-