Taking time into the equation

TAKING TIME INTO THE EQUATION

Potential and pitfalls of using real-world longitudinal data in developing clinical prediction models

Ruben Deneer

Financial support for printing this thesis was kindly provided by the Catharina Hospital, Eindhoven.

Cover	Marleen Deneer, 2022
Thesis layout	Ruben Deneer, 2022
Printed by	ADC Nederland
ISBN	978-90-386-5628-1

A catalogue record is available from the Eindhoven University of Technology Library. This thesis was typeset using LaTeX 2ε starting from the Maastricht Thesis Template (MTT) by Robbert Harms.

TAKING TIME INTO THE EQUATION

Potential and pitfalls of using real-world longitudinal data in developing clinical prediction models

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Technische Universiteit Eindhoven, op gezag van de rector magnificus prof.dr.ir. F.P.T. Baaijens, voor een commissie aangewezen door het College voor Promoties, in het openbaar te verdedigen op woensdag 21 december 2022 om 16.00 uur

door

Ruben Deneer

geboren te Valkenburg aan de Geul

Dit proefschrift is goedgekeurd door de promotoren en de samenstelling van de promotiecommissie is als volgt:

Voorzitter:	prof.dr. M. Merkx
Promotoren:	prof.dr. V. Scharnhorst
	prof.dr.ir. N.A.W. van Riel
Copromotor:	dr. A.-K. Boer (Catharina Ziekenhuis)
Leden:	prof.dr. O. Bekers (Universiteit Maastricht)
	prof.dr. L.R.C. Dekker
	dr. P.M.E. van Gorp
	prof.dr. P.A.J. Hilbers

Het onderzoek of ontwerp dat in dit proefschrift wordt beschreven is uitgevoerd in overeenstemming met de TU/e Gedragscode Wetenschapsbeoefening.

Cleverness as opposed to wisdom.
Wanting and having instead of thinking and making.
We cannot stop it.

DAVID FOSTER WALLACE

Contents

1
Introduction

Background

Historically, medicine was practiced as an art, based on the authority of a master, expert opinion and experience. Nowadays, medicine is considered both an art and a science, founded on results from clinical trials and research. The shift towards evidence based medicine (EBM) has for a great degree been facilitated by the rise of statistics. While the origins of statistical theory lie in the 18th century, medicine before the 20th century was still mostly based on empirical evidence and case studies. This changed drastically in the 20th century and can be linked to several key publications. In 1925 Sir Ronald Fisher published the book *Statistical Methods for Research Workers*, which is considered by some as one of the 20th century's most influential works on statistical methods [1]. Fisher introduced the well known (but poorly understood) concept of significance testing based on the P-value. In 1937 Sir Austin Bradford Hill published a series of papers in *The Lancet* on medical statistics, these were compiled in a book, *Principles of medical statistics*, in the same year [2]. Aided by Philip D'Arcy Hart and Marc Daniels, they conducted the first randomized controlled trial (RCT) on streptomycin (an antibiotic) treatment of pulmonary tuberculosis in 1948 [3]. Finally, in 1972 Archie Cochrane published his influential monograph, *Effectiveness and efficiency: Random reflections on health services*, in which he sets out the vital importance of RCTs in assessing the effectiveness of treatments and discusses the basis for 'evidence based medicine (EBM)' [4]. In response to this publication, the Cochrane Collaboration (now: Cochrane) was founded in 1993, which includes review groups from research intuitions worldwide to conduct systematic reviews to produce credible and accessible health information.

The rise of EBM has gone hand in hand with the development and use of clinical prediction models (CPMs). CPMs can be used either in public health (e.g. prediction of disease prevalence), clinical practice (e.g. for diagnosis or therapeutic decision-making) or research (e.g. selecting high risk patients for inclusion in a RCT or adjusting for covariates and confounding) [5]. In this thesis we focus on the application of CPMs in clinical practice. CPMs combine a set of predictors to predict an outcome, the outcome can either be diagnostic (e.g. does the patient have the disease) or prognostic (e.g. what is the expected time until mortality). In 1976 the first CPM, based on the Framingham Heart Study, was published. The Framingham risk score can be used to estimate the 10-year cardiovascular risk of an individual, based on several characteristics such as cholesterol levels, systolic blood pressure and cigarette use [6].

Before the advent of electronic health records (EHRs) and large registries, CPMs were mostly based on data collected as a result of a study, including RCTs. With the rapid growth in data acquisition, storage, algorithms and computing power, using real-world data (RWD) to develop CPMs has become more popular. This also contributed to an exponential growth in the number of publications on the topic of CPMs. However, actual implementation of CPMs in clinical practice is lagging behind in terms of publications, resulting an unrealized potential (Fig. 1.1).

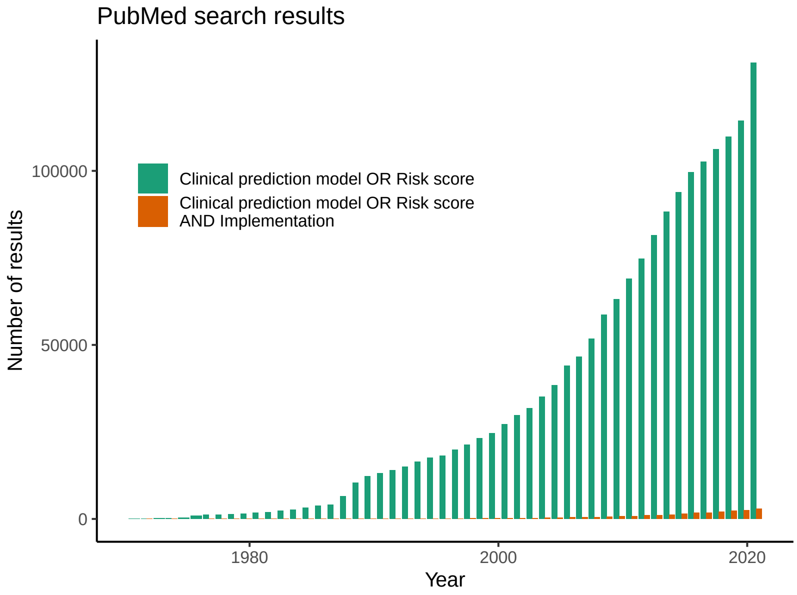


Figure 1.1: Number of PubMed results over time for search terms ”Clinical prediction model OR Risk score” versus terms ”Clinical prediction model OR Risk score AND Implementation”, starting from 1970 until 2021.

This thesis, to which this chapter is the introduction, focuses on the potential and pitfalls in the development, validation and implementation of CPMs based on real-world longitudinal data. In all chapters time plays an essential role, either because populations change over time, the disease changes over time, the data is in the form of repeated measures or the outcome is a time to an event in the future. Many RWD are longitudinal in nature, since patients are followed over time as disease progresses or a treatment is initiated. Exploiting repeated measures data for use in CPMs can make predictions dynamic, more patient-specific and accurate. This chapter is organized as follows: First, we define RWD and describe their potential as a source for CPMs. Secondly, we explain the development, validation and implementation of CPMs. Thirdly, we describe the different forms of repeated measures data and cover the statistical modeling of longitudinal data. Finally, we formulate the aims of this thesis and provide an outline of the chapters.

1.1 Real-world data

Clinical prediction models (CPMs) rely on a dataset from a representative sample of the target population. Historically, CPMs were developed on data from a prospective study, including (randomized controlled) clinical trials or large cohort studies such as the Framingham Heart Study [6]. The benefits of these data is that included patients have a regular follow-up and outcomes and study variables are clearly defined and registered, i.e. the data can be expected to be of high quality with low missingness. However, these data are not without their limitations for CPM development. Aside from costs and duration, clinical studies define inclusion criteria for eligible patients. These criteria can range from strict (e.g. for randomized controlled trials (RCTs)) to more liberal (e.g. for cohort studies). As a consequence, if the data from a study with strict exclusion criteria is used to develop a CPM, the CPM can only be applied to the population that would otherwise be eligible to participate in the study. Even studies with more liberal inclusion criteria cannot be guaranteed to represent all eligible patients, as not all patients can be expected to sign informed consent. With the rapid growth in data acquisition, storage, algorithms and computing power, the use of real-world data (RWD) to develop CPMs is becoming more popular [7]. Although there is no consensus on the definition of RWD [8], it is considered to be observational data relating to patient health status and/or the delivery of health care, generated as part of a healthcare process. This is opposite to data gathered as part of a clinical trial or study. Sources include electronic health records (EHRs), laboratory information systems, claims and billing information, registries, mortality databases, wearables etc. In a recent large-scale review of external validations of cardiovascular CPMs, the EHR was in 54% of the reviewed models the data source for the CPM, rather than a clinical trial (10%), a registry (26%) or any other source (10%) [9]. While EHR data (and RWD) is becoming more and more popular for clinical research due to its accessibility, these data come with many challenges which have been reported in several studies and will be addressed in this thesis [7, 10, 11].

1.2 Clinical prediction models

Clinical prediction models (CPMs) combine a set of predictors (or: independent variables, covariates, etc.) to predict an outcome (or: dependent variable, response, etc.). This outcome can either be diagnostic (e.g. does the patient have the disease) or prognostic (e.g. what is the expected time until mortality). How to choose a model that combines the information from the predictors to obtain a prediction for the outcome is not trivial. The choice depends on the nature of the data, goal and preference of the researcher. In the field of data science, models can be placed on a spectrum ranging from traditional statistical models to more recent machine learning models, see Table 1.1 for a summary. In this thesis we focus on the statistical modeling approach. The main reason for doing so is that statistical models serve multiple goals; study design, hypothesis testing, estimation and prediction. Often, questions such as ”What is the likelihood of outcome y given x?”, or ”What is the added value of measuring x, in predicting y?” or ”Is variable x associated with the outcome y?” arise in daily clinical practice. Statistical models are versatile in the sense that they can be used to answer clinically relevant questions, as well as be used as CPMs. Prediction can be considered a superset of hypothesis testing

Data science	
Statistical modeling	**Machine Learning**
White-box modeling	*Black-box modeling*
probabilistic data generating model,	algorithmic approach,
emphasis on inference and	emphasis on speed and
causal effects,	accuracy of prediction,
finding a ’correct’ model	finding a ’performing’ model

Table 1.1: Key differences between a traditional statistical modeling approach and a machine learning approach. Note that any dichotomization is arbitrary since there are intersections between both approaches and grey areas, e.g. non-parametric statistical models and interpretable machine learning models.

and estimation [12]. Also, when interpretability is required, statistical models are a suitable choice.

1.2.1 Development

Before developing the model itself, it is essential to first carefully formulate the research question. To quote the American mathematician and statistician John Tukey: “An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.” Next, one should get familiar with the available data. This includes an exploratory analysis of the predictor variables which involves checking of the distributions and missingness. In the case of missing values one should uncover the cause for this missingness to estimate the potential for bias. After one is familiar with the data, the model fitting itself can start. This is the most tricky part of model development, as many choices have to be made with respect to variable selection, interactions and non-linear terms. Two books that provide guidance with a focus on clinical data are *Regression modeling strategies* by Frank E. Harrell [12] and *Clinical prediction models* by Ewout W. Steyerberg [5]. After specifying the model, parameters have to be estimated. Again, many choices are available, ranging from a traditional frequentist or Bayesian approach to more recent penalized estimation approaches [13, 14].

1.2.2 Validation

After model fitting is complete, validation is required to asses the quality of the fitted model. The quality of a model is expressed in several performance measures, these can either be of statistical nature such as the akaike information criterion (AIC) [15] or more directly related to clinical practice such as positive predictive value. Ideally, these performance measures are calculated on a dataset that was not used during model development. This can be done by e.g. creating a training and test split of the dataset, or better, by cross validation or bootstrapping [13]. If validation is performed on the original dataset, this is referred to as *internal validation*. Internal validation is essential to quantify a model's tendency for overfitting [16]. A more rigorous test of model performance however, is done by *external validation*. In external validation a dataset from a different center is used to asses the performance of the model. Preferably this center should differ from the development center in patient population and/or treatment guidelines. External validation is the most rigorous test, since this tests ability of the model to produce accurate predictions on patients drawn from a different but plausibly related population [17–19]. Also, after an external validation, model performance has to be evaluated over time as patient populations, diseases and clinical practice change. Reporting guidelines have been established for studies developing, validating, or updating a prediction model, both for diagnostic or prognostic purposes. These are outlined in the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) statement [20].

1.2.3 Implementation

Judging from the vast number of publications of CPMs, one would expect that models are part of daily clinical practice, this is not the case however. Although some models have found acceptance in the clinic, e.g. the EuroSCORE II for assessing the risk of heart surgery in adults [21] and the CHA2DS2-VASc score for atrial fibrillation stroke risk [22], most models have not. This is for a large part a result of a lack of implementation studies, see Fig. 1.1. Implementation studies are even more uncommon than external validation studies. In an implementation study, a model is implemented in routine clinical practice and the benefits are quantified [23, 24]. These studies are also referred to as 'impact studies' and are performed prospectively, ideally, in a cluster-randomized trial [25]. The implementation study also addresses the aspect of model presentation, i.e. how are predictions presented to clinicians? This can either be a directive or assistive approach. In an assistive approach, predicted probabilities are reported without further recommendations, in a directive approach the predicted probabilities are presented as decision recommendations [25]. There is no "one size fits all" approach to an implementation study, each study has to be tailored to each specific setting.

1.3 Repeated measures data

Many real-world data (RWD) are in the form of repeated measures where patients are followed over time, e.g. by electronic health record (EHR) registrations, laboratory test results or wearables. Using repeated measures data for development of clinical prediction models (CPMs) is relatively rare compared to cross-sectional data [26]. Yet, exploiting repeated measures data can result in better prognostic performance of CPMs since one can distinguish within-subject variability from between-subject variability and, under certain assumptions, allow for missing data. Although formal definitions do not exist, we can distinguish two types of repeated measures data: time series data and longitudinal data.

1.3.1 Time series data

Generally, time series are referred to when a single study unit is observed over a long period of time, usually at regular time intervals. In time series analysis one is usually interested in forecasting future time points and assessing properties such as stationarity, trend and seasonality. Often the interest of time series analysis is in the observed unit itself. An example of time series data are the daily number of patients presenting at the emergency department (see Fig. 1.2A). In this case a single unit (the number of patients presenting at the emergency department) is observed over a long period of time (one year) at

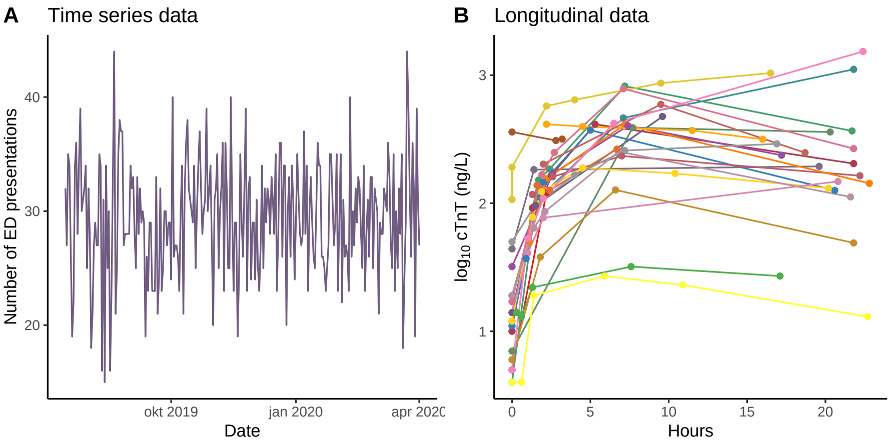


Figure 1.2: **A**: Example of time series data, daily presentations at the emergency department (ED) of the Catharina Hospital in Eindhoven.
B: Example of longitudinal data, 25 patients who underwent surgery had repeated measures of cardiac troponin-T (cTnT).

regular intervals (daily) and the interest may lie in forecasting the daily rates of patients presenting at the emergency department.

1.3.2 Longitudinal data

Longitudinal data is also referred to as panel data. In general, longitudinal data are referred to when multiple study units (e.g. patients) are observed with fewer measurements per unit (although it is also possible to have many, high frequency measurements per unit). In longitudinal data, one often takes a representative sample from a population of interest and performs several measurements on each subject in the sample. The interest is often in the variability of the different profiles and how this is related to an outcome or the population itself. A study collecting longitudinal data is usually designed with a fixed number of measurements. An example is the measurement of a cardiac biomarker at several time points for patients who undergo cardiac surgery (see Fig. 1.2B).

Time series data	Longitudinal data
low n	high n
high m	low m
forecasting future time points	association between profile and outcome
stationarity, trend, seasonality	within-subject, between-subject variability

Table 1.2: Summary of differences between time series and longitudinal data, n = number of subjects, m = number of measurements.

1.3.3 Modeling of longitudinal data

In this thesis we mainly focus on modeling longitudinal data. Longitudinal data can be incorporated in CPMs in various ways, either through statistical modeling or machine learning. As stated previously we follow the statistical modeling path. The linear mixed effects (LME) model is arguably the most widely used method for the analysis of longitudinal data [27]. The LME model was proposed in 1982 by Laird and Ware and can handle irregularly timed and missing measurements in a natural way. If we model a continuous response y_{ij} with a single predictor x_{ij}, for a patient i at occasion j and ignore the grouping structure we obtain a simple linear regression model:

$$\begin{aligned} E(y_{ij}|x_{ij}) &= \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \\ \varepsilon_{ij} &\sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{1.1}$$

where $E(y_{ij}|x_{ij})$ is the expected value, β_0 an intercept term, β_1 the regression coefficient and ε_{ij} the residual error having a Gaussian distribution with mean zero and standard deviation σ, $\mathcal{N}(0, \sigma^2)$. This can also be written in matrix notation:

$$\begin{aligned} E(\mathbf{Y}_i|\mathbf{X}_i) &= \beta \mathbf{X}_i + \varepsilon_i, \\ \varepsilon_i &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}), \end{aligned} \tag{1.2}$$

Note that the grouping structure of the data are ignored and the index i does not have any implications, this is referred to as a *fixed effects (or population average)* model. If we now introduce *random (or subject specific) effects*, we obtain a general formulation of a LME model:

$$E(\mathbf{Y}_i|\mathbf{X}_i) = \beta\mathbf{X}_i + b_i\mathbf{Z}_i + \varepsilon_i, \\ b_i \sim \mathcal{N}(0,\Psi), \varepsilon_i \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}), \tag{1.3}$$

where $\mathbf{Z}_i$ is the random effects design matrix and Ψ is a positive-definite symmetric covariance matrix. Normally $\varepsilon_i \sim \mathcal{N}(\mathbf{0},\sigma^2\mathbf{I})$ but other structures for ε_i can be used to model the residual correlation, this is also referred to as an extended linear mixed effects model. A visual demonstration of the difference between linear regression and mixed effects regression is given in Fig. 1.3. LME models allow for dynamic subject specific predictions which lead to substantial improvements in predictive accuracy [28]. Moreover, LME models can be combined with other outcomes such as a time-to-event in a joint model, which allows for individualized predictions of outcomes such as mortality or readmission [29].

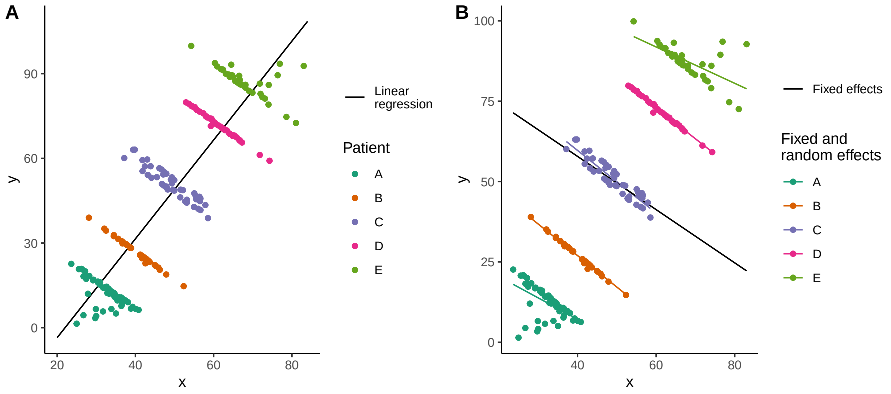


Figure 1.3: Visualization of predictions obtained from a linear regression and LME model based on synthetic data from 5 patients.
A: Shows a linear regression fitted to data from 5 patients. While every patient shows a declining trend, not taking the grouping into account will result in a positive slope for the linear regression fit. **B**: Shows the fit of a mixed effects model with random intercept and random slope, fit to the data from 5 patients. The fixed effects (i.e. population average) show a declining trend and the random effects allow for a subject-specific deviation from the fixed effects based on the measurements from each patient.

1.4 Aims and outline of this thesis

The goal of this thesis is to uncover the potential and pitfalls in the development, validation and implementation of clinical prediction models (CPMs) based on real-world longitudinal data. All aspects from assessing the quality of real-world data (RWD), to the development, validation and implementation of CPMs and the statistical modeling of longitudinal data, are applied to real-world clinical applications. From these applications, the potential and pitfalls of using RWD to develop CPMs are uncovered. This thesis is organized as follows:

Chapter 2 shows how RWD from patients undergoing coronary artery bypass grafting surgery can be used to detect clinically relevant subgroups. This study is a demonstration of unsupervised learning from RWD to improve diagnosis of a complication following surgery when the a gold standard diagnosis is lacking.

Chapter 3 builds upon the previous chapter. The goal of this chapter is to compare several non-parametric statistical modeling approaches to dynamic binary classification. Data is simulated to compare the performance of the different approaches and the data from Chapter 2 is used to show how dynamic longitudinal classification approaches can be used in clinical practice.

Chapter 4 shows the development, validation and implementation of a CPM based on RWD in the context of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). A CPM was developed to screen patients for a possible SARS-CoV-2 infection, based on routine laboratory data from patients presenting at the emergency department. This CPM was externally validated and implemented in multiple hospitals in the Netherlands.

Chapter 5 assesses the validity of the CPM developed in the previous chapter, in a new setting. The validity of the model is assessed for screening symptomatic healthcare workers.

Chapter 6 is an external validation study of a CPM developed to classify the risk for 6-month readmission or mortality for patients admitted for acute decompensated heart failure. The performance of the risk score is assessed for the patient population of the Catharina Hospital, and an association with self-care behavior is examined.

Chapter 7 is a validation study of a wearable device to measure heart rate (HR). The goal of this study is to assess the agreement between the HR extracted from the wearable and the gold standard 5-lead electrocardiogram connected to a patient monitor, during surgery and recovery.

Chapter 8 concludes this thesis with a discussion of the potential and pitfalls of using real-world longitudinal data in developing CPMs, based on the experiences from the previous chapters.

References

1. Fisher, R. A. *Statistical methods for research workers* (Oliver and Boyd, Edinburgh, 1925).
2. Hill, A. B. *Principles of Medical Statistics* (Lancet limited, 1937).
3. A Medical Research Council Investigation. Treatment of pulmonary tuberculosis with streptomycin and para-amino-salicylic acid. *The British Medical Journal,* 1073–1085 (1950).
4. Cochrane, A. L. *Effectiveness and efficiency: Random reflections on health services* (Nuffield Trust, 1972).
5. Steyerberg, E. W. *Clinical Prediction Models* 2nd ed. (Springer Nature, Switzerland, 2019).
6. Kannel, W. B., McGee, D. & Gordon, T. A general cardiovascular risk profile: the Framingham Study. *The American journal of cardiology* **38,** 46–51 (1976).
7. Goldstein, B. A., Navar, A. M., Pencina, M. J. & Ioannidis, J. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* **24,** 198–208 (2017).
8. Makady, A., de Boer, A., Hillege, H., Klungel, O., Goettsch, W., *et al.* What is real-world data? A review of definitions based on literature and stakeholder interviews. *Value in health* **20,** 858–865 (2017).
9. Wessler, B. S. *et al.* External validations of cardiovascular clinical prediction models: a large-scale review of the literature. *Circulation: Cardiovascular Quality and Outcomes* **14,** e007858 (2021).
10. Rusanov, A., Weiskopf, N. G., Wang, S. & Weng, C. Hidden in plain sight: bias towards sick patients when sampling patients with sufficient electronic health record data for research. *BMC medical informatics and decision making* **14,** 1–9 (2014).
11. Hersh, W. R. *et al.* Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical care* **51,** S30 (2013).
12. Harrell, F. E. *Regression modeling strategies* 2nd ed. (Springer International Publishing, Switzerland, 2015).
13. Hastie, T., Tibshirani, R. & Friedman, J. H. *The elements of statistical learning: data mining, inference, and prediction* (Springer New York, NY, 2009).
14. McElreath, R. *Statistical Rethinking* (Chapman and Hall/CRC, Mar. 2020).
15. Akaike, H. in *Selected papers of hirotugu akaike* 199–213 (Springer, 1998).
16. Moons, K. G. M. *et al.* Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. *Heart* **98,** 683–690 (2012).

17. Bleeker, S. E. *et al.* External validation is necessary in prediction research:: A clinical example. *Journal of clinical epidemiology* **56,** 826–832 (2003).

18. Justice, A. C., Covinsky, K. E. & Berlin, J. A. Assessing the generalizability of prognostic information. *Annals of internal medicine* **130,** 515–524 (1999).

19. Collins, G. S. *et al.* External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC medical research methodology* **14,** 1–11 (2014).

20. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Journal of British Surgery* **102,** 148–158 (2015).

21. Nashef, S. A. M. *et al.* Euroscore ii. *European journal of cardio-thoracic surgery* **41,** 734–745 (2012).

22. Lip, G. Y. H., Nieuwlaat, R., Pisters, R., Lane, D. A. & Crijns, H. J. G. M. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest* **137,** 263–272 (2010).

23. Reilly, B. M. & Evans, A. T. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. *Annals of internal medicine* **144,** 201–209 (2006).

24. Moons, K. G. M., Altman, D. G., Vergouwe, Y. & Royston, P. Prognosis and prognostic research: application and impact of prognostic models in clinical practice. *BMJ* **338** (2009).

25. Kappen, T. H. *et al.* Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic and prognostic research* **2,** 1–11 (2018).

26. Plate, J. D. J. *et al.* Incorporating repeated measurements into prediction models in the critical care setting: a framework, systematic review and meta-analysis. *BMC medical research methodology* **19,** 1–11 (2019).

27. Fitzmaurice, G., Davidian, M., Verbeke, G. & Molenberghs, G. *Longitudinal data analysis* (CRC press, 2008).

28. Finkelman, B. S., French, B. & Kimmel, S. E. The prediction accuracy of dynamic mixed-effects models in clustered data. *BioData mining* **9,** 1–21 (2016).

29. Rizopoulos, D. *Joint models for longitudinal and time-to-event data: With applications in R* (CRC press, 2012).

2

Detecting patients with PMI post-CABG based on cardiac troponin-T profiles

Ruben Deneer, Astrid GM van Boxtel, Arjen-Kars Boer, Mohamed A Soliman Hamad, Natal AW van Riel & Volkher Scharnhorst

Adapted from: Detecting patients with PMI post-CABG based on cardiac troponin-T profiles: a latent class mixed modeling approach. *Clinica Chimica Acta* **504,** 23–29 (2020).

Abstract

Background Diagnosis of perioperative myocardial infarction (PMI) after coronary artery bypass grafting (CABG) is fraught with complexity since it is primarily based on a single cut-off value for cardiac troponin (cTn) that is exceeded in over 90% of CABG patients, including non-PMI patients. In this study we applied an unsupervised statistical modeling approach to uncover clinically relevant cTn release profiles post-CABG, including PMI, and used this to improve diagnostic accuracy of PMI.

Methods In 624 patients that underwent CABG, cTnT concentration was serially measured up to 24 h post aortic cross clamping (XC). 2857 cTnT measurements were available to fit latent class mixed models (LCMMs).

Results Four classes were found, described by: normal, high, low and rising cTnT release profiles. With the clinical diagnosis of PMI as golden standard, the rising profile had a diagnostic accuracy of 97%, compared to 83% for an optimally chosen cut-off and 21% for the guideline recommended cut-off value.

Conclusion Clinically relevant subgroups, including patients with PMI, can be uncovered using serially measured cTnT and a LCMM. The LCMM showed superior diagnostic accuracy of PMI. A rising cTnT profile is potentially a better criterion than a single cut-off value in diagnosing PMI post-CABG.

2.1 Introduction

Coronary artery bypass grafting (CABG) surgery is an effective procedure to treat ischemic heart disease. Although the safety of CABG surgery is well-established, the procedure is nevertheless associated with a risk of perioperative and postoperative mortality and morbidity. Elevation of cardiac biomarkers such as creatine kinase and cardiac troponin (cTn) is common following CABG surgery and reflects perioperative myocardial damage [1, 2]. Even small enzyme elevations post-CABG are predictive of long-term prognosis and there is a graded association of elevation with outcome [1]. Perioperative myocardial damage can be ascribed to multiple causes, including direct trauma from surgical handling, inadequate myocardial protection during cardiopulmonary bypass or perioperative myocardial infarction (PMI) [3]. In CABG surgery, PMI is a complication that adversely affects the prognosis of the patient [3]. Incidence of PMI varies depending on the diagnostic criteria and patient population [3, 4]. Some studies report incidence rates up to 30%, though an average incidence of 3.9% established in a large systematic review seems more realistic [4]. The fourth universal definition of myocardial infarction (MI) arbitrarily defines a CABG-related PMI (Type 5) as elevation of cTn values $>$ 10 times the 99th percentile upper reference limit (URL) in patients with normal baseline values during the first 48 h following CABG surgery, combined with other clinical or echocardiographic evidence [5]. However, the current definition has its limitations. The diagnostic cut-off value of cTn $>$ 10 x URL is arbitrarily defined and occurs in over 90% of all patients undergoing CABG surgery [1, 6–8]. As a result, even small degrees of myocardial damage may lead to additional diagnostic procedures and subsequent clinical care pathways [9]. Alternatively, isolated elevations of cardiac biomarkers, which could be prognostically significant, are ignored in the absence of other evidence. Several studies have focused on the release profile (or kinetics) of cTn post-CABG, arguing that insight in the normal postoperative release profile can aid clinicians in recognizing patients with PMI and that timing of the peak is relevant when applying cut-off values [10–12]. Aside from the normal post-operative CABG cTn release profile, studies describe profiles for off-pump coronary artery bypass grafting (OPCAB) surgery [10, 11, 13] and surgeries complicated by PMI [10–12, 14–17]. While these studies demonstrate the variability in cTn release profiles and their value in recognizing PMI, they a priori define subgroups based on clinical characteristics or outcomes and subsequently evaluate cTn profiles. An alternative, assumption-free, approach is to group patients according to cTn release profiles and a posteriori evaluate the clinical characteristics and outcomes of each subgroup. In this study we used an unsupervised statistical learning approach to identify subgroups of patients without using any information other than cTnT release profiles post-CABG. To achieve this we used a statistical modeling technique called latent class mixed models (LCMMs) [18]. Our first aim was to fit a LCMM to data from a cohort of CABG patients where cTn was serially sampled post-operatively. From this model, we investigated the mean cTn release profiles of the uncovered classes and analyzed the subgroups of patients assigned to the classes based on clinical characteristics and outcomes, including PMI. Finally, the added value of the LCMM and serial cTn sampling in diagnosing PMI was determined.

2.2 Materials and Methods

2.2.1 Patient population

This study was a prospective observational cohort study and all patients who underwent coronary artery bypass grafting (CABG) surgery at the Catharina Hospital in Eindhoven between February 2013 and February 2014 were included in this study (N = 1028). Exclusion criteria were patients who underwent CABG with concomitant surgery. If patients underwent a reoperation during the inclusion period, only the first operation was included in the analysis. Blood samples for this study were residual samples obtained during routine withdrawal. Primary endpoints were cardiac troponin (cTn)T profile after uncomplicated cardiac surgery, cTnT profile after cardiac surgery complicated by perioperative myocardial infarction (PMI) and short/long-term mortality. Patients with missing data that had either i) none or only one cTn measurement (N = 123), or ii) where the aortic cross clamping (XC) time was not registered (N = 4), were excluded. Patients were also excluded if there was reasonable doubt whether the labeling of tubes was performed correctly (e.g.

a $>$ 28 ng/L ($>$ 2 x URL) decrease followed by $>$ 28 ng/L increase during the first 5 hours post-CABG (N = 27)).

2.2.2 cTnT measurements

Arterial blood samples were obtained preoperatively ($\leq$ 2 h before surgery) and at 1.5 h, 2 h, 6 h and 12 - 24 h post XC. If the procedure was performed as an off-pump coronary artery bypass grafting (OPCAB), the positioning of the mechanical stabilization device was taken as reference point of time. Samples were collected in BD Vacutainer® heparin tubes and immediately after withdrawal assayed for cTnT concentration using a high-sensitive cTnT Immunoassay from Roche Diagnostics Corporation on a Roche Elecsys® platform. The Roche hs-cTnT assay has a 10% imprecision at 13 ng/L with a 99th percentile URL of 14 ng/L.

2.2.3 PMI diagnosis

At the time this study was performed, diagnosis of PMI in our institution was based on elevation of aspartate aminotransferase (ASAT) activity. PMI was registered as a complication if ASAT activity was $>$ 100 U/L combined with i) new Q waves on an electrocardiogram (ECG) or new left bundle branch block; or ii) angiographic evidence of graft or native coronary artery occlusion; or iii) echocardiographic imaging evidence of new regional wall motion abnormality or new loss of viable myocardium.

2.2.4 Data collection and storage

Patient data was collected prospectively in the database of the department of cardiac surgery of our institution. These data included demographic information, risk factors, and complications. cTnT results were extracted from the laboratory information system. Mortality data was obtained from the municipal personal records database. All study data was merged and stored in a study database, see Section 2.A for the structure of the data.

2.2.5 Model fitting

Linear mixed models [19, 20] provide a flexible method to analyze longitudinal data since they incorporate between-patient variability, can handle irregularly sampled and missing data (under the missing at random assumption). However, linear mixed models assume that the underlying population is homogenous and can be described at the population level by a unique profile. If different subpopulations exist within the total population, these have to be explicitly specified. Latent class mixed models (LCMMs) assume that the population is heterogeneous and composed of latent classes of subjects, characterized by mean profiles of trajectories [18, 21]. LCMMs are also referred to as growth mixture models. After the LCMM is fitted, a posterior classification can be made which calculates the probability that each subject belongs to each of the latent classes. The LCMM consisted of a linear mixed model representing $\log_{10}$ transformed cTnT profiles over time and a multinomial logistic regression model representing (latent) class membership. CTnT was $\log_{10}$ transformed due to the highly skewed distribution. In the linear mixed model the fixed and random effects were modeled with natural cubic splines since cTnT profiles were expected to be highly nonlinear. Splines are preferred to polynomials due to their local nature and better numerical properties [22]. No predictors were included in the multinomial logistic regression model so class membership was not based on any patient characteristics or outcomes. The LCMM was fitted using R version 3.6.1 [23] and the lcmm package version 1.8.1 [18]. A series of LCMMs were fitted, consisting of an increasing number of latent classes. Since it is known that LCMMs can converge to local maxima [18], each LCMM was fitted 100 times with randomly chosen starting values to ensure that each model converged to a global maximum. The number of latent classes was increased until the Bayesian information criteria (BIC) [18, 24] started increasing with the addition of an extra latent class. The selection of the best model was based on i) the BIC and ΔBIC, ii) the posterior classification table and iii) clinical relevance [18, 25]. Mathematical formulation, model selection procedures and R code used to fit the model can be found in Section 2.A.

2.2.6 Post-hoc analysis

After selecting a LCMM, classes were given names based on the evolution of the mean cTnT. Class membership probabilities were calculated and each patient was a posteriori assigned to the class that corresponded to the highest probability. The differences between classes were analyzed with respect to patient characteristics, procedural characteristics and outcomes. If there were statistically significant ($p < 0.05$) differences between classes for a particular variable, post-hoc tests were used to compare which particular pair(s) of classes differed. Tukey's test was used for continuous variables, Dunn's test for non-normal continuous variables and pairwise Fisher tests for categorical variables. The Holm-Bonferroni method was used to adjust p-values for multiple comparisons [26]. Patients that were assigned to a rising profile class (i.e. cTnT still rising at 24 h post XC) were considered positive for PMI by the LCMM. The classification of the LCMM was compared to the clinical diagnosis of PMI in terms of sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) and accuracy. This was also done for the guideline definition [5] and for an optimally chosen cutoff. The optimally chosen cutoff was based on the Youden index [27] of the receiver operating characteristic (ROC) curve of the maximum cTnT value within 24 h post XC and the clinical diagnosis of PMI. Since a LCMM cannot be directly implemented in the clinic, we also assessed the added value of the LCMM in the clinical practice by defining a simple criterion based on visual inspection of the mean profiles.

2.3 Results

2.3.1 Study population

In total, 1028 patients were included in the study, see Fig. 2.1. Patients with concomitant procedures were excluded, leaving 778 patient that underwent coronary artery bypass grafting (CABG) without combined surgery. Of these 778 patients, some patients had missing or incorrect data with respect to the cardiac troponin (cTn) measurements, the aortic cross clamping (XC) time or the labelling of pre- and post-operative samples. After excluding patients with missing or erroneous data, a total of 624 patients remained in the analysis. Patient characteristics and outcomes are summarized in Table 2.1.

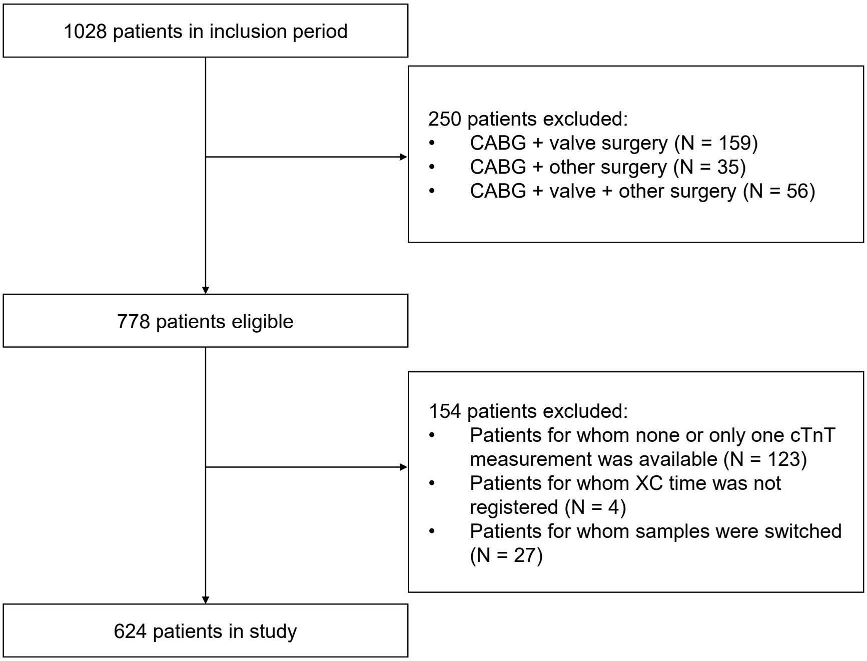

1028 patients in inclusion period
250 patients excluded:
• CABG + valve surgery (N = 159)
• CABG + other surgery (N = 35)
• CABG + valve + other surgery (N = 56)
778 patients eligible
154 patients excluded:
• Patients for whom none or only one cTnT measurement was available (N = 123)
• Patients for whom XC time was not registered (N = 4)
• Patients for whom samples were switched (N = 27)
624 patients in study

Figure 2.1: Inclusion flowchart. Flowchart describing inclusion of patients.

2.3.2 Latent class mixed model

Latent class mixed models (LCMMs) were fitted with up to 6 latent classes. The Bayesian information criteria (BIC) of the LCMMs decreased from 762.80 for a model without latent classes to a minimum of 514.86 for a model with 5 latent classes. The model with 4 latent classes (BIC = 518.86) was chosen as the final model, given the small decrease in BIC (-4.00) when going from 4 to 5 latent classes (reflecting modest evidence for a fifth class [28]) and higher discriminative ability. For more details regarding model selection, see Section 2.A. In Fig. 2.2A the estimated mean profile of each of the four latent classes in the final model is plotted. Patients were assigned to the class with the highest posterior probability. The individual profiles of patients assigned to each of the four classes are shown in Fig. 2.2B. Classes were labelled according to the shape of the profile. The "normal profile" class (N = 523, 83.8 %) contains the majority of patients and shows a typical cTnT profile post-CABG: a sharp increase in cTnT with a peak around 4 – 5 h post-XC, representing periprocedural myocardial damage, followed by a slow steady decline. The "rising profile" class (N = 29, 4.6 %) shows an initial sharp increase similar to the "normal profile" but where the cTnT concentration continues to rise until the end of the measurement period (24 h). The "low profile" class (N = 40, 6.4 %) shows a slower increase and a lower peak cTnT than the "normal profile". The "high profile" class (N = 32, 5.1 %) contains patients with an elevated baseline cTnT that peaks around 10 h post XC and then starts to decline.

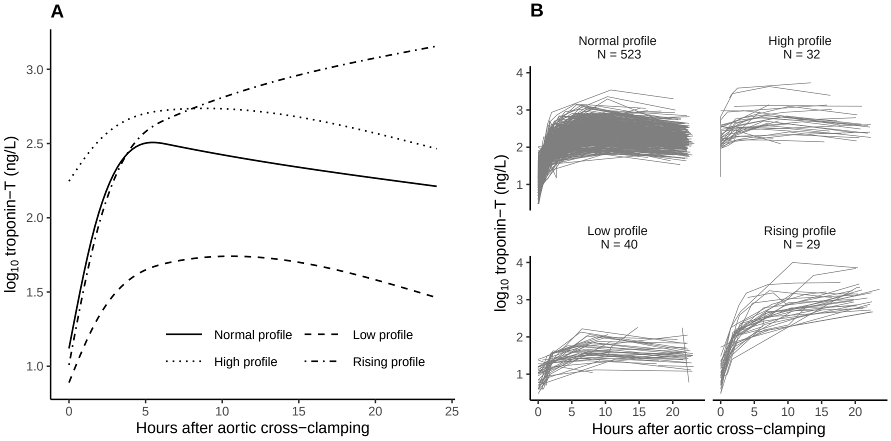


Figure 2.2: A: Mean cTnT profiles of latent classes. Estimated mean profiles of $\log_{10}$ cTnT in ng/L for each latent class from the final four class LCMM.
B: Individual cTnT profiles. Individual log˙10 cTnT profiles of patients that were a posteriori assigned to one of the four latent classes. N is the number of patients a posteriori assigned to that class.

2.3.3 Latent class characteristics and outcomes

Patients were assigned to the class with the highest posterior probability and classes were compared based on patient characteristics, procedural characteristics and outcomes in Table 2.2. Patients in the low profile class almost exclusively underwent off-pump coronary artery bypass grafting (OPCAB) surgery and were on average younger than patients in the high profile class. Patients in the high or rising profile class had higher surgical risk (i.e. higher EuroSCORE) than patients in the normal or low profile class. This was also reflected by the fact that these patients more often underwent emergency procedures than patients in the normal profile class. Patients in the high or rising profile class had a longer length of stay (LoS) in the intensive care unit (ICU) and hospital, than patients in the low or normal profile class. Finally, patients that were assigned to the rising profile class were more often diagnosed with perioperative myocardial infarction (PMI) and showed signs of ischemia on an electrocardiogram (ECG). There were no statistically significant differences between classes in terms of early or late mortality.

2.3.4 Latent classes and the diagnosis of PMI

55 % of patients in the rising profile class were clinically diagnosed with PMI whereas 98.8 % of patients in any of the other classes were not clinically diagnosed with PMI. Patients that were assigned to the rising profile class were considered positive for PMI. Fig. 2.3 shows the agreement between the LCMM rising class PMI classification and the clinical diagnosis of PMI in our clinic. For 16 of the 23 patients that had PMI there is agreement between the LCMM and the clinical diagnosis (true/concordant positives). 5 patients were classified in the normal profile class and 2 patients were classified in the high profile class, in spite of being clinically diagnosed with PMI (false/discordant negatives). 13 patients that were not clinically diagnosed with a PMI were classified in the rising profile class (false/discordant positives), the remaining 599 patients were all not diagnosed with a PMI and did not appear in the rising profile class (true/concordant negatives). From the mean profiles in Fig. 2.2A it can be seen that only the PMI class is still rising between 6 and 24 h post

XC. To assess the added value for the clinic, five different methods to classify patients with PMI are compared based on sensitivity, sensitivity, positive predictive value (PPV), negative predictive value (NPV) and accuracy. The results are shown in Table 2.3. The LCMM approach had the highest accuracy and PPV without sacrificing NPV. The additional criterion that cTnT is rising between 6 and 24 h post XC improved accuracy compared to the current guideline criterion and an optimally chosen cut-off.

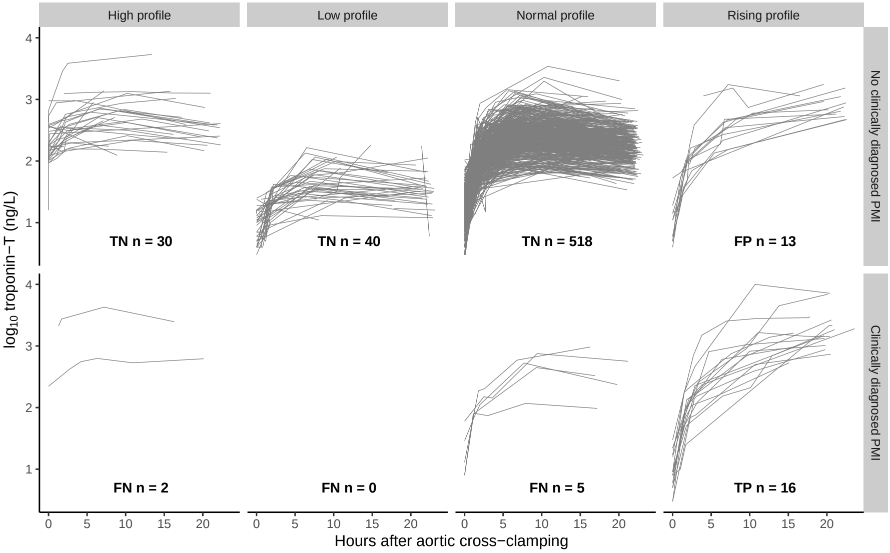


Figure 2.3: Confusion matrix plot. Individual log˙10 cTnT profiles of patients split by clinical diagnosis of PMI (upper row with clinical diagnosis of PMI, bottom row without PMI and posterior class assignment by the LCMM (normal, high, low and rising profile in each column respectively). True positives (TP); false negatives (FN); false positives (FP); true negatives (TN).

		N = 624
Pre-operative		
Age in years (mean (SD))		65.65 (9.68)
Female gender (%)		121 (19.4)
BMI in kg/m^2 (mean (SD))		27.63 (4.01)
Diabetes (%)		134 (21.5)
Hypertension (%)		367 (58.8)
Peripheral vascular disease (%)		65 (10.4)
Previous stroke (%)		34 (5.4)
Left-ventricular function (%)		
	Good	511 (81.9)
	Moderate	90 (14.4)
	Poor	14 (2.2)
	Very poor	1 (0.2)
	Unknown	8 (1.3)
Additive EuroSCORE (median [IQR])		3.00 [2.00, 5.00]
Prior cardiac surgery (%)		16 (2.6)
Emergency (%)		16 (2.6)
Intra-operative		
Intra-aortic balloon pump (%)		
	No	616 (98.7)
	Pre-op	2 (0.3)
	Per-op	3 (0.5)
	Post-op	3 (0.5)
Pre-op hemoglobin (mmol/L) (mean (SD))		9.00 (0.88)
Pre-op creatinine (umol/L) (median [IQR])		88.00 [76.00, 100.00]
Off-pump CABG (%)		133 (21.3)
Aortic cross-clamp time (min) (median [IQR])		45.00 [35.00, 59.00]
Cardiopulmonary bypass time (min) (median [IQR])		73.00 [57.00, 90.00]
Post-operative		
Length of stay on ICU (days) (median [IQR])		1.00 [0.00, 1.00]
Length of stay in hospital (days) (median [IQR])		5.00 [4.00, 6.00]
PMI (%)		23 (3.7)
Required reoperation (%)		38 (6.1)
Early mortality (30 days) (%)		4 (0.6)
Late mortality (5 years) (%)		49 (7.9)
Number of cTnT measurements (%)		
	2	29 (4.6)
	3	45 (7.2)
	4	121 (19.4)
	5	396 (63.5)
	6	31 (5.0)
	7	2 (0.3)

Table 2.1: Characteristics of the study population and outcomes. Body mass index (BMI); coronary artery bypass grafting (CABG); intensive care unit (ICU); cardiac troponin (cTn); perioperative myocardial infarction (PMI); standard deviation (SD); interquartile range (IQR)

	Normal profile	High profile	Low profile	Rising profile	p-value*
N	523	32	40	29	
Age in years (mean (SD))	65.8 (9.4)	68.2 (10.6)	61.8 (12.1)	64.6 (8.9)	0.026
Female gender (%)	97 (18.5)	7 (21.9)	9 (22.5)	8 (27.6)	0.602
EuroSCORE (median [IQR])	3.0 [2.0, 5.0]	6.0 [3.0, 7.0]	3.0 [1.0, 4.0]	3.0 [3.0, 5.0]	<0.001
Emergency (%)	5 (1.0)	6 (18.8)	2 (5.0)	3 (10.3)	<0.001
OPCAB (%)	81 (15.5)	5 (15.6)	38 (95.0)	9 (31.0)	<0.001
XC tim in min. (median [IQR])	45.0 [35.0, 58.5]	49.0 [37.5, 58.5]	61.0 [60.5, 61.5]	59.0 [43.0, 64.2]	0.071
CPB time in min. (median [IQR])	72.0 [55.0, 88.5]	77.5 [64.8, 88.0]	106.0 [102.5, 109.5]	94.0 [70.0, 99.0]	0.028
Days in ICU (median [IQR])	1.0 [0.0, 1.0]	1.0 [0.0, 2.0]	0.0 [0.0, 1.0]	2.0 [0.0, 3.0]	<0.001
Days in hospital (median [IQR])	5.0 [4.0, 6.0]	6.0 [4.0, 9.0]	4.0 [4.0, 6.0]	6.0 [5.0, 7.0]	0.001
PMI (%)	5 (1.0)	2 (6.2)	0 (0.0)	16 (55.2)	<0.001
ECG conclusion (%)					<0.001
No ischemia	440 (84.1)	25 (78.1)	33 (82.5)	12 (41.4)	
Possible	61 (11.7)	6 (18.8)	6 (15.0)	7 (24.1)	
Definite	10 (1.9)	0 (0.0)	0 (0.0)	10 (34.5)	
Inconclusive	12 (2.3)	1 (3.1)	1 (2.5)	0 (0.0)	
30 day mortality (%)	4 (0.8)	0 (0.0)	0 (0.0)	0 (0.0)	0.855
5 year mortality (%)	41 (7.8)	4 (12.5)	3 (7.5)	1 (3.4)	0.628
Pre-op cTnT in ng/L (median [IQR])	12.0 [7.0, 21.0]	165.5 [109.8, 355.2]	10.0 [5.2, 14.0]	9.0 [6.0, 17.0]	<0.001
Peak cTnT in ng/L (median [IQR])	278.0 [180.0, 425.0]	508.0 [368.5, 914.0]	43.5 [32.8, 82.2]	1111.0 [735.0, 1755.0]	<0.001

Table 2.2: Patient characteristics of latent classes. *p-values determine if there are significant differences between classes. χ^2 test for categorical variables, one-way ANOVA for normally distributed variables, Kruskal-Wallis rank sum test for non-normally distributed variables. aortic cross clamping (XC); cardiopulmonary bypass (CPB); intensive care unit (ICU); electrocardiogram (ECG); off-pump coronary artery bypass grafting (OPCAB); perioperative myocardial infarction (PMI); cardiac troponin (cTn); standard deviation (SD); interquartile range (IQR).

cTnT PMI classification method	Sensitivity	Specificity	PPV	NPV	Accuracy
Guideline definition	0.96	0.18	0.05	0.99	0.21
Guideline definition & rising after 6 h	0.91	0.40	0.06	0.99	0.42
Optimal cutoff*	0.91	0.83	0.18	1.00	0.83
Optimal cutoff* & rising after 6 h	0.87	0.87	0.22	0.99	0.87
LCMM rising class	0.70	0.98	0.55	0.99	0.97

Table 2.3: Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) for different approaches to detecting patients with a perioperative myocardial infarction (PMI). * the optimal cutoff for cTnT based on the Youden index was 528.5 ng/L. Latent class mixed model (LCMM); cardiac troponin (cTn); perioperative myocardial infarction (PMI).

2.4 Discussion

In this study we investigated whether subgroups of patients could be identified based on cardiac troponin (cTn)T release profiles post-coronary artery bypass grafting (CABG), in particular patients with perioperative myocardial infarction (PMI). Our results illustrate that by using a latent class mixed model (LCMM), subgroups of patients could be identified that show distinctive cTnT profiles without using any prior information other than serial cTnT measurements taken up to 24h post aortic cross clamping (XC). Using the model's posterior classification of patients to a rising cTnT profile showed substantially greater accuracy and positive predictive value (PPV) (without affecting negative predictive value (NPV)) in diagnosing PMI compared to the guideline criteria. First, a model with latent classes had a substantially lower Bayesian information criteria (BIC) than a model without latent classes (518.86 versus 762.80), indicating that a model with latent classes is a better fit to the data [18, 25, 28]. In addition to the BIC, the diagonal terms in the posterior classification table were close to 1 (0.86, 0.89, 0.96, 0.94) reflecting good discriminative ability [18]. While the five class model had slightly lower BIC, the four class model was chosen due to the better discriminative ability. Also, there is substantive theory from literature which validates the choice for a four class model. The low profile class, which contains almost exclusively off-pump coronary artery bypass grafting (OPCAB) surgeries, is in agreement with literature describing a delayed and lower peak for OPCAB surgery [10, 11, 13]. The rising profile class is also in agreement literature, describing a PMI profile as having a delayed peak or rise following an earlier peak [10, 12, 14–17]. To explain the cTnT release profile of patients with PMI, most studies refer to the work from Katus et. al who suggest that early release represents cytosolic troponin from myocytes that are reversible damaged, whereas late release (after one day) represents structural troponin from irreversibly damaged myocytes [29]. The high profile class is not described in literature. This is explained by the fact that most studies exclude patients with emergency procedures or procedures within seven days of a myocardial infarction (MI), which are most likely patients with elevated baseline cTnT in the high profile class. The post hoc analysis revealed that the rising profile class consisted of more than half of patients clinically diagnosed with PMI. There were thirteen discordant positive patients (of 624 in total), i.e. patients that were assigned to the rising profile class but were not clinically diagnosed with PMI. Nine patients had peak aspartate aminotransferase (ASAT) $\leq$ 100 U/l and did therefore not meet the ASAT-criterion to be diagnosed with PMI. The electronic health records (EHRs) of the remaining four patients were re-examined by a thoracic surgeon. Three patients had no secondary evidence (electrocardiogram (ECG) or echocardiographic) of PMI and were therefore not diagnosed with PMI, one patient started to develop PMI but re-intervention took place before the final diagnosis was made. There were seven discordant negative patients, i.e. patients that were clinically diagnosed with PMI but not assigned to the rising profile class. The EHRs of these seven patients were also re-examined by a thoracic surgeon. Five patients did not have elevated cardiac enzymes but were diagnosed with PMI on the basis of a combination of other diagnostic criteria i.e. ECG abnormalities, echocardiographic evidence or hemodynamic instability. Two patients were incorrectly clinically diagnosed with PMI, one patient had pre-operative MI and one patient was mislabeled as positive. Although only mismatched cases were re-examined, the causes of misclassification were mainly due to the ASAT criterion or (lack of) evidence of other clinical or echocardiographic abnormalities. If the two patients that were incorrectly labelled with PMI were re-classified as negative, the sensitivity of the LCMM increased to 0.76 and the specificity to 0.99. That the LCMM is not in perfect agreement with the clinical diagnosis of PMI, is not merely a shortcoming of the model but also of the variation in the diagnosis of PMI. Confirming or denying a diagnosis of PMI may be of secondary importance to the clinical consequences. Although clinical consequences such as major adverse cardiac and cerebrovascular events were not registered in our study, post hoc tests revealed that patients with a rising cTnT profile had a longer length of stay (LoS) in the hospital (Dunn's test p-value: 0.045), a longer LoS on the ICU (Dunn's test p-value: 0.0036) and more often had signs of ischemia on an ECG (Fisher's exact test p-value: < 0.001) than patients with a normal cTnT profile. This is in agreement with previous studies who reported associations between elevated post-operative cTnT and prolonged stay on the ICU [30, 31]. A limitation of our study is that we did not have samples > 24 h post XC, therefore we could not determine the timing of the cTnT peak for patients with PMI. Previous studies observed a rising pattern even after 48 hours [11, 14]. However, non-PMI patients reach their peak between 4 – 12 h post XC, therefore 24 h is sufficient to distinguish early versus late peak occurrence. This finding also confirms the potential for early ($<$ 12 h post-CABG) cTnT to detect patients at risk for PMI or other adverse events as reported by other studies [31, 32]. Another limitation is that the model in its current form is difficult to implement in a prospective manner in the clinic. However, we have demonstrated that information gathered from the estimated mean cTnT profile (that cTnT in patients in the rising class is still rising between 6 and 24 h post XC) can already improve diagnostic accuracy, both with respect to the guideline and an optimally chosen cutoff. Our approach of visually interpreting estimated mean cTnT profiles does not take any variability into account and is only a proof of concept. Also, given the multifactorial causes for post-operative cTnT elevation, one can expect differences between centers. Consequently the results obtained from this single center analysis may not be generalizable to other centers. Finally, although LCMMs do not prove that the found subpopulations actually exist [33] and skepticism of complex statistical models is appropriate, the fit indices combined with substantive theory and high diagnostic accuracy of PMI patients provide strong evidence to assume that the heterogeneity in cTnT release profiles is the result of actual subpopulations instead of other causes of non-normality. We have demonstrated that a statistical model is capable of recognizing clinically relevant subgroups of patients based on cTnT release profiles post-CABG and that information from this model can be used to improve the guideline for Type 5 MI. Future studies should be done to determine the optimal sampling time-points of cTnT to detect a rising pattern and the associated improvement compared to a single cut-off value in diagnosing PMI.

2.5 Conclusions

This study has shown that characteristic cardiac troponin (cTn)T release profiles exist post-coronary artery bypass grafting (CABG) surgery. These profiles could be uncovered by a latent class mixed model (LCMM) without any prior information other than serial cTnT measurements up to 24 hours post aortic cross clamping (XC). Four classes were discovered that showed a low, high, rising and normal cTnT release profile. Patients were a posteriori assigned to each of one of these classes. The rising profile proved to be predictive for perioperative myocardial infarction (PMI), with higher positive predictive value (PPV) and accuracy than the guideline recommended cutoff or an optimally chosen cutoff. We argue that a rising cTnT release profile is potentially of greater predictive value for PMI than a single value above or below a cutoff.

References

1. Domanski, M. J. *et al.* Association of Myocardial Enzyme Elevation and Survival Following Coronary Artery Bypass Graft Surgery. *JAMA* **305,** 585–591 (2011).

2. Januzzi, J. L. *et al.* A comparison of cardiac troponin T and creatine kinase-MB for patient evaluation after cardiac surgery. *Journal of the American College of Cardiology* **39,** 1518–1523 (2002).

3. Yau, J. M. *et al.* Impact of Perioperative Myocardial Infarction on Angiographic and Clinical Outcomes Following Coronary Artery Bypass Grafting (from PRoject of Ex-vivo Vein graft ENgineering via Transfection [PREVENT] IV). *The American Journal of Cardiology* **102,** 546–551 (2008).

4. Nalysnyk, L., Fahrbach, K., Reynolds, M. W., Zhao, S. Z. & Ross, S. Adverse events in coronary artery bypass graft (CABG) trials: a systematic review and analysis. *Heart* **89,** 767–772 (2003).

5. Thygesen, K. *et al.* Fourth Universal Definition of Myocardial Infarction (2018). *Journal of the American College of Cardiology* **72,** 2231–2264 (2018).

6. Wang, T. K. M. *et al.* Diagnosis of MI after CABG with high-sensitivity troponin T and new ECG or echocardiogram changes: relationship with mortality and validation of the Universal Definition of MI. *European Heart Journal: Acute Cardiovascular Care* **2,** 323–333 (2013).

7. Muehlschlegel, J. D. *et al.* Troponin is superior to electrocardiogram and creatinine kinase MB for predicting clinically significant myocardial injury after coronary artery bypass grafting. *European Heart Journal* **30,** 1574–1583 (2009).

8. Mohammed, A. A. *et al.* Prospective, comprehensive assessment of cardiac troponin t testing after coronary artery bypass graft surgery. *Circulation* **120,** 843–850 (Sept. 2009).

9. Moussa, I. D. *et al.* Consideration of a New Definition of Clinically Relevant Myocardial Infarction After Coronary Revascularization: An Expert Consensus Document From the Society for Cardiovascular Angiography and Interventions (SCAI). *Journal of the American College of Cardiology* **62,** 1563–1570 (2013).

10. Ge, W. *et al.* High-sensitivity troponin T release profile in off-pump coronary artery bypass grafting patients with normal postoperative course. *BMC Cardiovascular Disorders* **18,** 157 (2018).

11. Markman, P. L., Tantiongco, J.-P., Bennetts, J. S. & Baker, R. A. High-Sensitivity Troponin Release Profile After Cardiac Surgery. *Heart, Lung and Circulation* **26,** 833–839 (2017).

12. Tevaearai Stahel, H. T. *et al.* Clinical Relevance of Troponin T Profile Following Cardiac Surgery. *Frontiers in Cardiovascular Medicine* **5,** 182 (Dec. 2018).

13. Wittock, A. *et al.* High-sensitive cardiac troponins in patients undergoing cardiac surgery: friend or foe? *Intensive Care Medicine Experimental* **3,** A952 (2015).

14. Carrier, M., Pellerin, M., Perrault, L. P., Solymoss, B. C. & Pelletier, L. C. Troponin levels in patients with myocardial infarction after coronary artery bypass grafting. *The Annals of Thoracic Surgery* **69,** 435–440 (2000).

15. Lim, C. C. S. *et al.* Early Diagnosis of Perioperative Myocardial Infarction After Coronary Bypass Grafting: A Study Using Biomarkers and Cardiac Magnetic Resonance Imaging. *The Annals of Thoracic Surgery* **92,** 2046–2053 (2011).

16. Onorati, F. *et al.* Determinants and Prognosis of Myocardial Damage After Coronary Artery Bypass Grafting. *The Annals of Thoracic Surgery* **79,** 837–845 (2005).

17. Peivandi, A. A. *et al.* Comparison of Cardiac Troponin I versus T and Creatine Kinase MB after Coronary Artery Bypass Grafting in Patients with and without Perioperative Myocardial Infarction. *Herz* **29,** 658–664 (2004).

18. Proust-Lima, C., Philipps, V. & Liquet, B. Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *Journal of Statistical Software* **78,** 1–56 (2017).

19. Laird, N. M. & Ware, J. H. Random-effects models for longitudinal data. *Biometrics,* 963–974 (1982).

20. Verbeke, G. & Molenberghs, G. *Linear Mixed Models for Longitudinal Data* (Springer Science & Business Media, 2009).

21. Muthén, B. & Shedden, K. Finite Mixture Modeling with Mixture Outcomes Using the EM Algorithm. *Biometrics* **55,** 463–469 (1999).

22. Harrell, F. E. *Regression modeling strategies* 2nd ed. (Springer International Publishing, Switzerland, 2015).

23. R Core Team. *R: A Language and Environment for Statistical Computing* Vienna, Austria, 2020.

24. Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics* **6,** 461–464 (1978).

25. Nylund, K. L., Asparouhov, T. & Muthén, B. O. Deciding on the Number of Classes in Latent Class Analysis and Growth Mixture Modeling: A Monte Carlo Simulation Study. *Structural Equation Modeling: A Multidisciplinary Journal* **14,** 535–569 (2007).

26. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* **6,** 65–70 (1979).

27. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3,** 32–35 (1950).

28. Kass, R. E. & Raftery, A. E. Bayes Factors. *Journal of the American Statistical Association* **90,** 773–795 (1995).

29. Katus, H. A., Remppis, A., Scheffold, T., Diederich, K. W. & Kuebler, W. Intracellular compartmentation of cardiac troponin T and its release kinetics in patients with reperfused and nonreperfused myocardial infarction. *The American Journal of Cardiology* **67,** 1360–1367 (1991).

30. Baggish, A. L. *et al.* Postoperative troponin-T predicts prolonged intensive care unit length of stay following cardiac surgery. *Critical Care Medicine* **32,** 1866–1871 (Sept. 2004).

31. Gahl, B. *et al.* Prognostic value of early postoperative troponin T in patients undergoing coronary artery bypass grafting. *Journal of the American Heart Association* **7** (Mar. 2018).

32. Göber, V. *et al.* Early Troponin T and Prediction of Potentially Correctable In-Hospital Complications after Coronary Artery Bypass Grafting Surgery. *PLOS ONE* **8,** e74241 (Sept. 2013).

33. Bauer, D. J. & Curran, P. J. Distributional Assumptions of Growth Mixture Models: Implications for Overextraction of Latent Trajectory Classes. *Psychological Methods* **8,** 338–363 (2003).

Appendix

2.A Latent class mixed model fitting

In this appendix we describe the development of the latent class mixed model (LCMM). We provide the mathematical formulation of the LCMM, the R code that was used to fit the model, summary statistics, mean trajectories of all the LCMMs that were fitted, estimates and posterior classification table of the final selected model.

2.A.1 Mathematical formulation of the LCMM

The LCMM consisted of a linear mixed model describing cardiac troponin (cTn)T profiles over time and a multinomial logistic regression model representing (latent) class membership.

Linear mixed model

The standard linear mixed model is given by: $Y = \beta\mathbf{X} + b\mathbf{Z} + \varepsilon$. Y is the dependent variable, β are the fixed effects, b are the random effects, $\mathbf{X}$ and $\mathbf{Z}$ are the fixed and random effects design matrices, and ε the residuals. In this study we had one group of patients i, where $i = 1,\ldots,n$, measured at j occasions, where $j = 1,\ldots,m$. For each patient i at each occasion j the log˙10 transformed value of troponin-T, log $cTnT$, was modeled as a function of time t

(in hours post aortic cross-clamping). The nonparametric linear mixed model, which relates the time t_{ij} of patient i at occasion j to the log $cTnT_{ij}$, is given by

$$\log cTnT_{ij} = \sum_{k=1}^{q} \beta_k X_k(t_{ij}) + b_{i,k} Z_k(t_{ij}) + \varepsilon_{ij}, \tag{2.A.1}$$

where X_k and Z_k are bases for spline functions with q fixed knots, t_{ij} is the time in hours post aortic cross-clamping, β the fixed effects coefficients, b_{ik} are the random effects coefficients and ε_{ij} the residual errors. For X_k and Z_k natural cubic splines were used with $q = 4$ knots. Four knots were chosen for identifiability purposes, i.e. the number of observations is greater than the number of random effects in the linear mixed model. The two boundary knots were fixed at the time of the first and last cTnT measurement that appeared in the data (= 0, 23.6 h). The internal knots were chosen based on the 33.3 % and 66.7 % quantiles of the distribution of all the measurement times (= 1.5, 6.6 h). Knots at the quantiles guarantee that each interval, while of varying length, contains an equal amount of measurements.

Multinomial logistic regression model

The linear mixed model assumes that the population of patients is homogenous and can be described at the population level by the unique profile of the fixed effects: $\sum_{k=1}^{q} \beta_k X_k(t_{ij})$. The LCMM assumes that the population is heterogeneous and consists of G latent class, characterized by G mean profiles of trajectories. Latent class membership is defined by a discrete random variable c_i that equals g if subject i belongs to latent class g. We fitted models containing up to 6 latent classes, i.e. $g = 1, \ldots, 6$. The probability that a subject i belongs to a certain class g is calculated from a multinomial logistic regression model without any covariates:

$$\pi_{ig} = \frac{e^{\gamma_{0g}}}{e^{\gamma_{01}} + e^{\gamma_{02}} + e^{\gamma_{03}} + e^{\gamma_{04}} + e^{\gamma_{05}} + e^{\gamma_{06}}}, \tag{2.A.2}$$

where for identifiability purposes the last class is chosen as reference class and is set to 0, i.e. in the case of 6 latent classes: $\gamma_{06} = 0$.

Latent class linear mixed model

The linear mixed model is extended with latent classes by allowing the fixed effects and the distribution of the random effects to be class-specific. The linear mixed model described above, extended with latent classes can be written as:

$$\log cTnT_{ij|c_i} = \sum_{k=1}^{q} \beta_k X_{k_1}\left(t_{ij}\right) + \gamma_{c_i} X_{k_2}\left(t_{ij}\right) + b_{i,k_{c_i}} Z_k\left(t_{ij}\right) + \varepsilon_{ij}, \qquad (2.A.3)$$

where the fixed effects are now split into common fixed effects $\beta_k X_{k_1}\left(t_{ij}\right)$ and class-specific fixed effects $\gamma_{c_i} X_{k_2}\left(t_{ij}\right)$. The random effects are not split but the distributions of the random effects $b_{i,k_{c_i}}$ are now class-specific.

2.A.2 R-code to fit model

The dataset that contained the cTnT measurements for each patient was stored in the so called long format, see Listing 2.A.1.

```
ID  t     logtropT  gender  age ...
13  0.0   1.278754  M       63  ...
13  1.8   2.049218  M       63  ...
13  7.1   2.666518  M       63  ...
13  21.8  3.045714  M       63  ...
18  2.2   2.617000  M       68  ...
18  4.5   2.598791  M       68  ...
18  11.5  2.568202  M       68  ...
18  16.0  2.499687  M       68  ...
17  0.0   1.278754  F       55  ...
17  1.4   1.806180  F       55  ...
17  2.1   1.934498  F       55  ...
17  6.8   2.403121  F       55  ...
17  21.6  2.049218  F       55  ...
26  0.0   0.698970  F       56  ...
26  2.2   2.082785  F       56  ...
27  0.0   0.602060  M       46  ...
27  1.7   2.181844  M       46  ...
27  2.4   2.267172  M       46  ...
```

```
27  7.2   2.914343  M       46  ...
27  21.7  2.564666  M       46  ...
```

Listing 2.A.1: Long format, ID refers to the patient ID, time to the time after aortic crossclaming and logtropT to log10 cTnT. There are more variables (e.g. gender and age) but these are not used in the LCMM.

To fit the model, R version 3.6.1 was used with packages `lcmm` (version 1.8.1) and `splines`. The `set.seed(123)` command was used for reproducibility. The code to fit the (two class) latent class linear mixed model is given in Listing 2.A.2.

```
gridsearch(rep = 100, maxiter = 20, minit = m1lin,
hlme(fixed = logtropT ~ ns(t, knots = c(1.5, 6.6),
    Boundary.knots = c(0, 23.6)), random = ~ ns(t,
    knots = c(1.5, 6.6), Boundary.knots = c(0, 23.6)),
     mixture = ~ ns(t, knots = c(1.5, 6.6), Boundary.
    knots = c(0, 23.6)), data = df.cabg, subject = "ID
    ", ng = 2))
```

Listing 2.A.2: Two class LCMM.

With arguments:

-
  ```
  gridsearch(rep = 100, maxiter = 20, minit =
      m1lin
  ```

  Convergence to a global maximum is not guaranteed for mixture models, because of the existence of local maxima. To ensure convergence to a global maximum, the gridsearch function is used. We used a maximum of 20 iterations from 100 random vectors of initial values which are generated from the linear mixed model (m1lin).

-
  ```
  fixed = logtropT ~ ns(t, knots = c(1.5, 6.6),
      Boundary.knots = c(0, 23.6))
  ```

  The formula for the fixed effects, the log cTnT value is modelled by a natural cubic spline function of t, with interior knots at 1.5 and 6.6 hours and boundary knots at 0 and 23.6 hours.

-
  ```
  random = ~ ns(t, knots = c(1.5, 6.6),
     Boundary.knots = c(0, 23.6))
  ```

  This is the model formula for the random effects, in this case similar to the fixed effects, i.e. the same natural cubic spline function of t.

-
  ```
  mixture = ~ ns(t, knots = c(1.5, 6.6),
     Boundary.knots = c(0, 23.6))
  ```

  The formula for the class specific fixed effects, similar to the random effects.

-
  ```
  data = df.cabg
  ```

  Name of the dataframe that contains the measurements.

-
  ```
  subject = "ID"
  ```

  Name of the covariate in the dataframe representing the patient identifier.

-
  ```
  ng = 2
  ```

  Number of latent classes, in this case 2, but this was varied from 2 to 6.

All other arguments are set to the default values, this implies the use of an unstructured variance-covariance matrix for the random effects which is common over latent classes.

2.A.3 Summary and selection of models

All models converged successfully and were compared using the `summarytable` command in the `lcmm` package. Using natural cubic splines with 4 knots and up to 6 latent classes the results are shown in Table 2.A.1. Models without random effects (this corresponds to latent class growth analysis which assumes that within a specific latent class the repeated measures of the patient are independent) were also tried. However, models without random effects had higher Bayesian information criteria (BIC)

G	loglik	npm	BIC	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
1	-333.13	15	762.80	100.00	NA	NA	NA	NA	NA
2	-265.96	20	660.65	93.43	6.57	NA	NA	NA	NA
3	-197.79	25	556.49	89.10	4.49	6.41	NA	NA	NA
4	-162.89	30	518.86	83.81	5.13	6.41	4.65	NA	NA
5	-144.80	35	514.86	6.57	8.33	77.40	3.21	4.49	NA
6	-131.19	40	519.82	9.46	0.32	75.32	4.33	4.01	6.57

Table 2.A.1: Summarytable using natural cubic splines with 4 knots and up to 6 latent classes. G is number of latent classes, loglik the log likelihood, npm the number of parameters, BIC the Bayesian information criteria and posterior probability of the latent classes. The number represent the percentage of patients assigned to each class.

values (2381.83, 1899.63, 1400.22, 1209.81, 1085.85, 984.14) and were therefore not preferred. To select the number of latent classes from the LCMM we also calculated the ΔBIC when increasing the number of latent classes, see Table 2.A.2. According to the BIC, the 5 class model is preferred since the BIC is lowest. However, the strength of evidence for the 5-class- versus the 4-class-model is positive but not strong, given that the ΔBIC = -4 when going from 4 to 5 latent classes. We plotted the 2, 3, 4 and 5 class models to determine clinical relevance in Fig. 2.A.1. The 5-class-model has an additional class that accommodates patients with a moderately elevated baseline cTnT, as compared to the 4-class-model which only has a single class for highly elevated baseline cTnT. Given that an intermediate class with moderately elevated cTnT has no direct clinical application/relevance and the ΔBIC value is modest, the 4-class-model is preferred. Note that the rising class is always present starting from the 3-class-model, and is unaffected by

Latent classes increased from . . . to . . .	ΔBIC
1 to 2	-102.15
2 to 3	-104.16
3 to 4	-37.63
4 to 5	-4.00
5 to 6	4.96

Table 2.A.2: ΔBIC when increasing the number of latent classes.

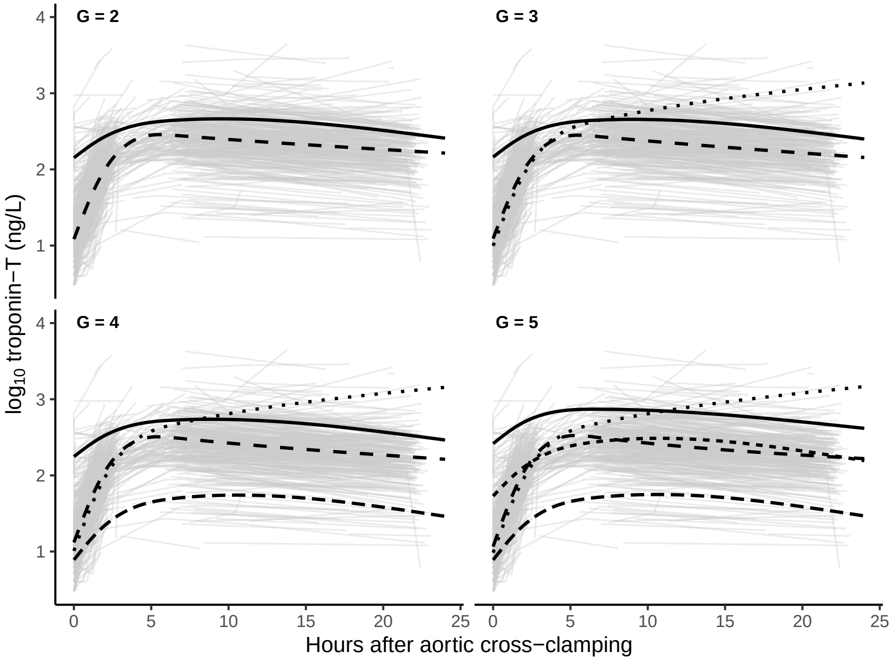


Figure 2.A.1: The 2, 3, 4 and 5 class LCMMs, G is the number of latent classes.

the number of latent classes.

2.A.4 Posterior classification

To compare the 4- and 5-class LCMMs, we assigned names to each of the profiles based on the trajectories: normal, high, intermediate, low and rising. A posteriori classification is made for each patient based on the calculation of the posterior class-membership probabilities. These probabilities are extracted from the model by using the pprob command of the lcmm package, see Listing 2.A.3.

```
ID   class     prob1   prob2   prob3   prob4
50   3         0.492   0.000   0.508   0.000
```

```
54   1         1.000  0.000  0.000  0.000
55   1         0.999  0.000  0.000  0.001
56   1         0.558  0.440  0.000  0.002
58   4         0.017  0.000  0.000  0.982
63   3         0.134  0.000  0.866  0.000
67   1         1.000  0.000  0.000  0.000
68   1         0.969  0.000  0.031  0.000
69   4         0.000  0.000  0.000  1.000
70   1         0.516  0.003  0.000  0.481
```

Listing 2.A.3: Posterior probabilities for individual patients (ID). For example patient 50: the patient was assigned to class 3 (low profile) but only with a probability of 0.508, there is a probability of 0.492 that the patient belongs to class 1 (normal profile).

The mean of all the probabilities can be calculated for the assigned classes and summarized in the posterior classification table as provided by the `lcmm` command `postprob`. These are given in Table 2.A.3 for the 4 class solution and in Table 2.A.4 for the 5 class solution.

		Mean posterior probability in each class			
	N (%)	Normal class	High class	Low class	Rising class
Normal class	523 (83.81)	0.9625	0.0104	0.0150	0.0121
High class	32 (5.13)	0.0590	0.9405	0.0000	0.0004
Low class	40 (6.41)	0.1282	0.0000	0.8709	0.0009
Rising class	29 (4.65)	0.1176	0.0079	0.0001	0.8744

Table 2.A.3: Posterior classification table for 4 class solution.

		Mean posterior probability in each class				
	N (%)	Normal class	High class	Low class	Rising class	Intermediate class
Normal class	483 (77.40)	0.9297	0.0010	0.0149	0.0106	0.0436
High class	20 (3.21)	0.0004	0.9071	0.0000	0.0000	0.0924
Low class	41 (6.57)	0.1097	0.0000	0.8831	0.0008	0.0064
Rising class	28 (4.49)	0.1191	0.0122	0.0001	0.8647	0.0039
Intermediate class	52 (8.33)	0.1295	0.0331	0.0110	0.0111	0.8154

Table 2.A.4: Posterior classification table for 5 class solution.

A LCMM has perfect discriminative ability if all the terms on the diagonal of the posterior classification table are 1. From the posterior classification

Tables 2.A.3 and 2.A.4 it can be concluded that both models show good discriminative ability, given that all the diagonal terms are > 0.8. The 4 class model however, shows slightly better discriminative ability, the minimum of the diagonal terms is 0.87 for the 4 class model versus 0.82 for the 5 class model, the mean is 0.91 versus 0.88 respectively. This, combined with the modest ΔBIC value and a lack of clinical relevance, has led to the choice for the 4 class LCMM.

3

A comparison of non-parametric statistical modeling approaches to dynamic classification of irregularly and sparsely sampled curves

Ruben Deneer, Zhuozhao Zhan, Edwin R van den Heuvel, Astrid GM van Boxtel, Arjen-Kars Boer, Natal AW van Riel & Volkher Scharnhorst

Adapted from: A comparison of non-parametric statistical modeling approaches to dynamic classification of irregularly and sparsely sampled curves. *In preparation for submission*.

Abstract

This paper describes and compares the performance of several popular non-parametric statistical modeling approaches to dynamically classify subjects into two groups, based on an irregularly and sparsely sampled curve. We simulated data and compared the discriminative ability over time for growth charts, conditional growth charts, a tensor product smooth, longitudinal discriminant analysis and a generalized functional linear model. The approaches were subsequently applied to a real world clinical example. Our results show that functional regression approaches that implicitly incorporate historic information through random effects, provide better discriminative ability than approaches that do not take historic information into account or explicitly model historic information through auto-regression terms. The functional regression and tensor product smooth approaches were subsequently applied to a real-world clinical dataset to demonstrate the performance.

3.1 Introduction

Longitudinal data occurs frequently in clinical settings where patients are monitored over time. Combining longitudinal data with a time to event outcome has gained popularity in recent years due to an ongoing increase in the research of joint modeling techniques[1, 2]. Remarkably, modeling approaches that combine longitudinal data with a *binary* outcome are less popular, despite the fact that binary outcomes are frequently encountered in clinical settings when the exact timing of an event cannot be determined or is not clinically relevant. Examples include, diagnosing prostate cancer based on serial prostate-specific antigen measurements, achievement of successful pregnancy based on longitudinal measurements of adhesiveness of certain blood lymphocytes, and predicting the presence of gestational trophoblastic disease based on repeated measurements of human chorionic gonadotropin [3–5]. In literature, prediction and classification based on longitudinal data is also referred to as longitudinal discriminant analysis or longitudinal classification. Compared to various studies that examine the performance of dynamic prediction of joint models with a time to event outcome, research on the performance of different approaches to dynamic classification of longitudinal data is lacking [6–9]. The aim of this study is to compare several recent non-parametric approaches to longitudinal classification. Non-parametric approaches allow for flexible modeling of non-linear profiles and can easily be fit through the use of available software. We assess the potential gain in dynamic classification performance between approaches that incorporate historic information implicitly and explicitly, and more simple approaches do not take the history into account. With this study we provide guidance to researchers developing longitudinal classification models for sparse and irregular data. Results are based on simulated data and illustrated by real-world example.

This study is motivated by a real-world clinical example where a cardiac biomarker is measured sparsely (2 to 5 times) and irregularly in the first 24 hours after coronary artery bypass grafting (CABG) surgery to determine the presence of a complication in the form of a perioperative myocardial infarction (PMI). Early detection of a PMI enables clinicians to intervene and limit injury [10]. Developing a model that can dynamically classify patients as PMI or non-PMI, based on accruing information from biomarker measurements, can assist clinicians in early detection of a PMI.

Longitudinal data introduces an extra challenge compared to cross-sectional data, in the sense that a suitable model for the longitudinal profile has to be selected. The approaches most commonly described in literature, fit a linear mixed effects (LME) model to the longitudinal profile and use the output of the LME model in a discriminant function [3, 11–14]. Extensions to multiple longitudinal profiles which utilize multivariate LME models have also been been developed [15–17], as well as non-parametric approaches in the form of functional discriminant analysis [18, 19]. As an alternative to using a discriminant function to classify individuals into groups, summary measures obtained from a LME model (e.g. subject-specific slopes or intercepts) are also utilized as covariates in a logistic regression model, either by a two-stage approach [20, 21] or a joint modeling approach where parameters of the mixed model and logistic regression model are estimated simultaneously [4, 5, 22, 23]. Some authors have also used non-parametric or non-linear mixed effects models to model the longitudinal trajectory in a two-stage approach [24–27]. Alternatively, one can apply a more simplified approach and ignore the historic information, i.e. the multilevel structure and serial correlation of the longitudinal data. One way to achieve this, is by using a varying-coefficient model. This implies the use of a (logistic) regression model where the interaction of the biomarker with time is used as covariate [28, 29]. Finally, reference growth charts obtained from quantile regression models, are widely used in the clinic as an easy-to-follow tool for differentiating abnormal growth of infants from the population norm[30]. While standard reference growth charts are not developed with screening purposes in mind (due to their inability to account for covariates or past history), conditional growth charts can account for growth history, and are thus recommended as a diagnostic tool to screen for unusual growth [31, 32].

This study is organized as follows. First, we describe the different modeling approaches that are compared. Secondly, we describe the method to simulate data from a mechanistic model and we present the results of the different modeling approaches applied to the simulations. Finally, we apply the approaches to the motivating example as an illustration and show the potential clinical benefit. The study is concluded with a discussion.

3.2 Methods

We consider the situation where, for each patient, a single biomarker is measured sparsely (for some patients down to one or two measurements) and irregularly. The outcome is binary, indicating whether the patient was diagnosed with a complication. The predicted probability of a patient experiencing the outcome should be updated each time a new measurement becomes available, this probability is then used to classify patients as positive or negative. Let Y_i be the binary outcome of experiencing a perioperative myocardial infarction (PMI) for the i-th patient, t_{ij} be the time of the j-th occasion the i-th patient was measured, and $u(t_{ij})$ be the measured biomarker value at t_{ij} where $t_{ij} \in \mathscr{T}$, a bounded interval in $\mathbb{R}$. We model the longitudinal profile of $u(t_{ij})$ as follows:

$$u(t_{ij}) = f(t_{ij}) + \varepsilon_{ij}, \tag{3.2.1}$$

where $f(t_{ij})$ is a smooth function of the time t_{ij} and ε_{ij} is random noise. As stated previously, we assume $u(t_{ij})$ is measured on an irregular and sparse grid. In all approaches we use the generalized additive model (GAM) framework to fit the longitudinal profile $u(t_{ij})$ [29, 33]. To estimate the smooth function $f(t_{ij})$ we choose P-splines with a sufficiently large basis dimension [34]. We can therefore express $u(t_{ij})$ as a set of basis functions:

$$u(t_{ij}) = \beta_0 + \sum_{k=1}^{K} \beta_k b_k(t_{ij}) + \varepsilon_{ij}, \tag{3.2.2}$$

where b_k are a set of K basis functions, β_0 is a parametric intercept term and β_k the associated spline coefficients. The coefficients are estimated by maximizing the penalized log likelihood: $l_p(\beta) = l(\beta) - \lambda J_\beta$, where J_β is a penalty function based on the second-order difference of the coefficients of adjacent splines and λ a smoothing parameter that has to be chosen (for more details see Eilers and Marx) [34]. The solution to maximizing the penalized log-likelihood is obtained by penalized iteratively reweighted least squares and

λ is chosen by minimizing the generalized cross validation score, see Wood [29]. Note that we have not yet taken dependence among observations from the same patient into account, this is deferred to the different approaches. We use the representation in Eq. (3.2.2) for the longitudinal profile and compare the following approaches to classify a new patient with at least one or more measurements: a fixed cut-off, a static growth chart (SGC), a conditional growth chart (CGC), a tensor product smooth (TPS), longitudinal discriminant analysis (LDA) and a generalized functional linear model (GFLM). The fixed cuf-off, SGC and TPS do not take the history into account, while the CGC, LDA and GFLM can be considered historic approaches that incorporate past measurements.

3.2.1 Raw value

Arguably the most straightforward approach is to use the raw value $u(t_{ij})$ itself as an early stopping rule. If a patient rises above a certain threshold then the patient is considered positive for the outcome Y_i. This is the current clinical practice and serves as a reference to compare to the modeling approaches in this study. This approach does not take the time dependent nature of $u(t_{ij})$ into account.

3.2.2 Static growth charts

Static growth charts (SGCs) can be estimated through quantile regression (QR). QR aims at fitting the τ-th conditional quantile ($\tau \in [0,1]$) of u for a given t_{ij}:

$$Q_{u|\tau}(t_{ij}) = t_{ij}\beta_\tau + \varepsilon_{ij,\tau}, \tag{3.2.3}$$

where β_τ is the coefficient belonging to the τ-th quantile and $\varepsilon_{ij,\tau}$ random noise belonging to the τ-th quantile. In Eq. (3.2.3) the assumption is made that the τ-th quantile depends linearly on the covariate t_{ij}. Fasiolo et al. have developed a novel framework that combines QR with GAMs, resulting in a smooth additive quantile regression model (QGAM) [35]. The advantage of this framework is that, aside from incorporating smooth functions, all smoothing and hyperparameters are estimated automatically. Using the QGAM framework, the conditional quantile τ of u is given by:

$$Q_{u|\tau}(t_{ij}) = f_\tau(t_{ij}) + \varepsilon_{ij,\tau}, \tag{3.2.4}$$

where $f_\tau(t_{ij})$ is smoothing function represented by a P-spline and $\varepsilon_{ij,\tau}$ the residual error term. Parameter estimates are obtained by minimizing:

$$\begin{aligned} Q_{u|\tau}(t_{ij}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \rho_\tau(u(t_{ij}) - f_\tau(t_{ij})), \\ \rho_\tau(z) &= (\tau-1)\frac{z}{\sigma} + \lambda \log(1 + e^{\frac{z}{\lambda\sigma}}), \end{aligned} \tag{3.2.5}$$

$\rho_\tau(z)$ is the so called extended log-f loss, σ a scale parameter and λ a penalty factor that determines the smoothness of the loss. By using the fast calibrated Bayesian methods proposed by Fasiolo et al., QGAMs can be fitted with this loss function. Since $Q_{u|\tau}(t_{ij})$ represents the value of the τ-th quantile of $u(t_{ij})$, we fit Eq. (3.2.5) for a vector of quantiles $\tau = (0.01, 0.02, ..., 0.99)$ to each training set containing only patients hat did not experience the outcome (i.e. controls). This way we obtain a reference growth chart for a suitable grid of quantiles for control patients. Subsequently we transform all measurements in the test sets to their estimated quantile, i.e. all measurement pairs $(u(t_{ij}), t_{ij})$ are transformed to τ_{ij} by using the grid of quantiles. We then use the τ_{ij} to classify a patient as positive or negative for a PMI, if τ_{ij} is above a chosen cutoff the patient is classified as positive.

3.2.3 Conditional growth charts

In this study we implement the conditional growth chart (CGC) as described by Wei et al. in the QGAM framework [32]. We expand Eq. (3.2.4) as follows:

$$Q_{cond,u|\tau}(t_{ij}) = f_\tau(t_{ij}) + \sum_{k=1}^{p} (\alpha_{k,\tau} + \theta_{k,\tau} D_{i,j,k}) t_{i,j-k} + \varepsilon_{ij,\tau}, \tag{3.2.6}$$

where $D_{i,j,k} = t_{i,j} - t_{i,j-k}$, the time between the j-th and $(j-k)$-th measurement, and $\alpha_{k,\tau}$ and $\theta_{k,\tau}$ are parametric auto-regressive (AR) coefficients. We choose $p = 1$, i.e. an AR(1) model. Analogous to the SGC approach, all measurements in the test dataset are assigned to their respective conditional quantiles, $\tau_{cond,ij}$, and used to classify patients as positive or negative.

3.2.4 Tensor product smooth

An alternative to the growth chart approach that directly models the probability of the outcome Y_i, rather than estimating quantiles of a healthy population, is the tensor product smooth (TPS) approach. In the case of a TPS model we directly predict the probability of having a PMI, $P(Y_i = 1|u(t_{ij}), t_{ij})$, for a measurement pair $(u(t_{ij}), t_{ij})$. Essentially, this a logistic regression with an interaction between time and the measured value as covariate, also called a varying-coefficient model. For this we use a GAM with a logistic link function and a tensor product smooth interaction (as $u(t_{ij})$ and t_{ij} are on different scales):

$$\begin{aligned} \text{logit}\{Y_i = 1|u(t_{ij}), t_{ij}\} &= f(u(t_{ij}), t_{ij}) \\ &= \sum_{k=1}^{K} \sum_{l=1}^{L} \beta_{kl} b_l(u(t_{ij})) a_k(t_{ij}), \end{aligned} \tag{3.2.7}$$

where a_k and b_l are sets of P-spline basis functions for t_{ij} and $u(t_{ij})$, respectively. The predicted probability $P(Y_i = 1|u(t_{ij}), t_{ij})$ is used to classify patients as positive or negative for a PMI. This is analogous to the growth chart approach, except the TPS approach produces a probability instead of a quantile.

3.2.5 Longitudinal discriminant analysis

The longitudinal discriminant analysis (LDA) approach consists of two steps. In the first step the inherently infinite-dimensional curves are projected onto a low dimensional space and in the second step the low dimensional representation is used to perform discriminant analysis. In this study we implement both a covariance pattern longitudinal discriminant analysis (COV-LDA) as described by Roy et al. [36] and a functional longitudinal discriminant analysis (F-LDA) as described by James and Hastie [18].

Covariance pattern LDA

The COV-LDA model consists of a linear additive model with a parametric intercept term, a factor smooth interaction term and a covariance pattern (i.e. correlation structure) to model dependence among observations within a single patient. The factor smooth interaction term allows for separate smooths for both PMI and non-PMI patients:

$$\begin{aligned} u(t_{ij}) &= \beta_0 + z_i\beta_1 + f_{z_i}(t_{ij}) + \varepsilon_{ij} \\ &= \beta_0 + z_i\beta_1 + \sum_{k=2}^{K} \beta_{z_i,k} b_{z_i,k}(t_{ij}) + \varepsilon_{ij}, \\ \varepsilon_i &= \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{im} \end{bmatrix} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \Lambda_i) \\ z_i &= \begin{cases} 0, & \text{if patient } i \text{ was not diagnosed with a PMI,} \\ 1, & \text{if patient } i \text{ was diagnosed with a PMI,} \end{cases} \end{aligned} \tag{3.2.8}$$

where β_0 and β_1 are class-specific parametric intercepts for non-PMI and PMI patients respectively, $b_{z_i,k}$ are sets of P-spline basis functions with dimension K for non-PMI and PMI patients respectively and Λ_i is a covariance matrix. Λ_i can be decomposed, $\Lambda_i = \mathbf{V}_i\mathbf{C}_i\mathbf{V}_i$, where $\mathbf{V}_i$ is diagonal and $\mathbf{C}_i$ a correlation matrix. Since observations are irregularly sampled, we choose a continuous-time AR(1) correlation structure for $\mathbf{C}_i$ to model the dependence between measurements from the same subject [37].

Functional LDA

The alternative F-LDA approach does not assume a correlation structure for the residual error but utilizes random effects to capture the variability between patients. In the F-LDA approach we model the profile $u(t_{ij})$ as follows:

$$\begin{aligned}
u(t_{ij}) &= \beta_0 + z_i\beta_1 + f_{z_i}(t_{ij}) + U_i(t_{ij}) + \varepsilon_{ij} \\
&= \beta_0 + z_i\beta_1 + \sum_{k=2}^{K} \beta_{z_i,k} b_{z_i,k}(t_{ij}) + U_i(t_{ij}) + \varepsilon_{ij}, \\
U_i(t_{ij}) &\sim \mathcal{N}(0, \Sigma(t_{ij}, t'_{ij})), \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \\
z_i &= \begin{cases} 0, & \text{if patient } i \text{ was not diagnosed with a PMI,} \\ 1, & \text{if patient } i \text{ was diagnosed with a PMI,} \end{cases}
\end{aligned} \tag{3.2.9}$$

where β_0, β_1 and $b_{z_i,k}$ are analogous to Eq. (3.2.8) and $U_i(t_{ij})$ are (functional) random effects representing the subject specific deviation from the overall mean function, modeled as a zero-mean Gaussian process with variance-covariance function $\Sigma(t_{ij}, t'_{ij})$. More specifically, $\Sigma(t_{ij}, t'_{ij}) = \text{cov}(U_i(t_{ij}), U_i(t'_{ij}))$. Since we are dealing with irregular and sparse data, the estimation of the covariance function is not as straightforward as with a suitably dense grid. Therefore we use the fast covariance estimation for sparse functional data (FACEs) approach by Xiao et al. [38] In this approach the covariance function $\Sigma(t_{ij}, t'_{ij})$ is modeled by penalized tensor product splines $\Sigma(t_{ij}, t'_{ij}) = \mathbf{b}(t_{ij})^\intercal \Theta \mathbf{b}(t'_{ij})$, where $\mathbf{b}$ is a spline basis and Θ a symmetric coefficient matrix. The covariance function and error variance are jointly estimated in a two-step procedure and smoothing parameters are selected using leave-one-subject out. For more details see Xiao et al. [38]

After fitting Eq. (3.2.8) and Eq. (3.2.9) to the training set, we can obtain estimates for a new subject given a set measurement times t_{ij}. By plugging t_{ij} in either Eq. (3.2.8) or Eq. (3.2.9) we obtain estimates for the mean if the patient would belong to the PMI group, $\hat{u}(t_{ij}, z_i = 1)$, or the non-PMI group $\hat{u}(t_{ij}, z_i = 0)$ and a covariance matrix (either by imposing a correlation structure in Eq. (3.2.8) or by modeling the covariance matrix with a spline basis in Eq. (3.2.9)). Given the observed set of measurements $u(t_{ij})$ for the new subject, we can then calculate the value of the probability density function in the case the patient belongs to the PMI or to the non-PMI group. These values can then be utilized in a Bayes discriminant rule to obtain the probability of a PMI:

$$P(Y_i = 1|u(t_{ij}), t_{ij}) = \frac{\pi_{\text{PMI}} f_{\text{PMI}}(u(t_{ij}))}{\pi_{\text{no-PMI}} f_{\text{no-PMI}}(u(t_{ij})) + \pi_{\text{PMI}} f_{\text{PMI}}(u(t_{ij}))} \tag{3.2.10}$$

where π_{PMI} is the prior probability of having a PMI and $\pi_{\text{no-PMI}} = 1 - \pi_{\text{PMI}}$, f_{PMI} is the conditional density function if the patient had a PMI and $f_{\text{no-PMI}}$ is the conditional density function if the patient did not have a PMI. The prior probabilities π_{PMI} and $\pi_{\text{no-PMI}}$ are equal to the fraction of non-PMI and PMI patients in the training dataset, respectively. The probability $P(Y_i = 1|u(t_{ij}))$ is then used to classify patients as positive or negative, analogous to the previous approaches.

3.2.6 Generalized functional linear model

The generalized functional linear model (GFLM) is described by Müller et al. for observations observed on dense grids of points [24]. The main idea is to reduce the dimension of the longitudinal data by an orthogonal expansion of the random effects and use the first few components of the expansion as covariates in a generalized linear model. This procedure can also be applied to our study, with some modification as observations are irregularly and sparsely observed. We model $u(t_{ij})$ as follows:

$$\begin{aligned} u(t_{ij}) &= f(t_{ij}) + U_i(t_{ij}) + \varepsilon_{ij} \\ &= \beta_0 + \sum_{k=1}^{K} \beta_k b_k(t_{ij}) + U_i(t_{ij}) + \varepsilon_{ij} \\ U_i(t_{ij}) &\sim \mathcal{N}(0, \Sigma(t_{ij}, t'_{ij})), \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \end{aligned} \tag{3.2.11}$$

analogous to Eq. (3.2.9), except there is no factor smooth interaction for PMI and non-PMI patients and only an overall mean. All variability is captured by the random effects. Analogous to Eq. (3.2.9) we estimate the smoothed covariance matrix $\Sigma(t_{ij}, t'_{ij})$ with the FACEs procedure by Xiao et al. [38] The covariance function $\Sigma(t_{ij}, t'_{ij})$ can be decomposed into functional principal components: $\Sigma(t_{ij}, t'_{ij}) = \sum_{l=1}^{\infty} \lambda_l \phi_l(t_{ij}) \phi_l(t'_{ij})$, where λ_l and $\phi_l(t_{ij})$ are the respective eigenvalues and eigenfunctions. We choose a number of eigenfunctions L that explain 95% of total variance. By the Karhunen-Loève theorem we can project $U_i(t_{ij})$ onto the L-dimensional basis, and Eq. (3.2.11) becomes:

$$\begin{gathered} u(t_{ij}) = \beta_0 + \sum_{k=1}^{K} \beta_k b_k(t_{ij}) + \sum_{l=1}^{L} \xi_{i,l} \phi_l(t_{ij}) + \varepsilon_{ij}, \\ \xi_{i,l} \sim \mathscr{N}(0, \lambda_l), \varepsilon_{ij} \sim \mathscr{N}(0, \sigma^2). \end{gathered} \tag{3.2.12}$$

Since data are sparse and irregular, the scores $\xi_{i,l}$ are estimated by the principal components analysis through conditional expectation approach described by Yao et al. [39] After estimating all scores for the subjects in the training dataset, a logistic regression model is fitted with the scores $\xi_{i,l}$ as covariates and Y_i as outcome. If we want to make a prediction for a new subject, we first estimate the scores $\xi_{i,l}$ by using the conditional expectation and subsequently plug the scores in the logistic regression model to obtain the probability that the patient has a PMI:

$$\text{logit}\{Y_i = 1 | \xi_{i,1}, \xi_{i,2}, \ldots, \xi_{i,L}\} = \sum_{l=1}^{L} \gamma_l \xi_{i,l}, \tag{3.2.13}$$

where γ_l are the coefficients belonging to the L scores. The predicted probabilities are ten used to classify patients as positive or negative for PMI, analogous to the previous approaches.

3.2.7 Performance evaluation

Since the goal of this study is to perform dynamic longitudinal classification, we focus on the performance of the approaches when they are used in a dynamic fashion. To compare the area under the ROC-curve (AUC) over time, we first calculate the cumulative maximum value/quantile/probability over time, within each patient for each approach. Our motivation for doing so, is that in clinical practice, if the probability exceeds a certain threshold, the patient is classified as positive and an intervention will take place. Hence, the cumulative maximum in the time interval $\mathscr{T}$ determines if a patient is classified as positive or negative. The cumulative maxima are either the raw values of $u(t_{ij})$, the predicted quantiles τ_{ij} in case of the growth charts and the predicted probabilities of a PMI $P(Y_i = 1|u(t_{ij}), t_{ij})$ for the other approaches. Next we take the maximum value for each patient and for each approach on the time interval $\mathscr{T}$ and calculate the AUC of this maximum value. Subsequently, the ROC-curves of the maxima are used to determine a threshold for each approach based on the Youden index. This threshold is then used as an early stopping rule, we classify a patient as positive if a value/quantile/probability rises above the threshold and mark the time that this occurs. We then calculate the sensitivity, specificity and average run length (ARL) of each approach using the stopping rule. The average run length represents the mean time until a patient is classified as positive.

3.3 Implementation

All approaches are implemented using `R` version 4.2.1.[40] The P-spline smooths are modeled using the `s` function from `mgcv` package version 1.8.40. The static growth chart (SGC) and conditional growth chart (CGC) are fitted using the `qgam` package version 1.3.4. The tensor product smooth (TPS) model is fitted using the using `gam` and `te` functions from `mgcv`. To fit the covariance pattern longitudinal discriminant analysis (COV-LDA) model, the `gamm` function is used together with the `corCAR1` function as a constructor for the correlation matrix, parameters in the COV-LDA model are estimated using restricted maximum likelihood. To estimate the covariance function in the functional longitudinal discriminant analysis (F-LDA) and generalized functional linear model (GFLM) model we use the `face.sparse` function from the `face` package version 0.1.7. The densities of the normal distributions, required by the Bayes rule, are calculated using the `dmvnorm` function in the `mvtnorm` package version 1.1.3.

3.4 Simulations

We simulate 100 datasets, each containing $N = 500$ patients that are measured on an irregular and sparse grid of measurement times t_j. For each dataset we generate N sequences of measurement times: $t_j = \{0, 2, 4, 6, 8, 12, 16, 20, 24\}$ where j is an index and t_j the measurement time in hours. We introduce irregularity by adding random variation to each $t_{j>1}$ by sampling from a uniform distribution between -0.25 and 0.25 and randomly removing elements $t_{j>1}$ with a probability 0.1, sampled from a binomial distribution. The irregular and sparse sequence t_j is then plugged in a bi-exponential model to obtain simulated biomarker values for a patient i:

$$\begin{gathered}
c_i(t_j) = \phi_{1i} \mathrm{e}^{-\phi_{3i} t_j} + \phi_{2i} \mathrm{e}^{-\phi_{4i} t_j} + \varepsilon_{ij}, \\
\phi_i = \begin{bmatrix} \phi_{1i} \\ \phi_{2i} \\ \phi_{3i} \\ \phi_{4i} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} \gamma_1 z_i \\ \gamma_2 z_i \\ \gamma_3 z_i \\ \gamma_4 z_i \end{bmatrix} + \begin{bmatrix} b_{1i} \\ b_{2i} \\ b_{3i} \\ b_{4i} \end{bmatrix} = \beta + \gamma z_i + b_i, \\
z_i = \begin{cases} 0, & \text{if patient } i \text{ was not diagnosed with a PMI,} \\ 1, & \text{if patient } i \text{ was diagnosed with a PMI,} \end{cases} \\
b_i \sim \mathcal{N}(0, \Psi), \ \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2),
\end{gathered} \tag{3.4.1}$$

$$\begin{gathered}
\phi_i = \begin{bmatrix} -1.59 - 0.233 z_i \\ 2.73 + 0.19 z_i \\ 0.457 - 0.0903 z_i \\ 0.00928 - 0.0118 z_i \end{bmatrix}, \\
\Psi = \begin{bmatrix} 0.1770 & -0.0786 & -0.0218 & 0 \\ -0.0786 & 0.1450 & -0.0233 & 0 \\ -0.0218 & -0.0233 & 0.0357 & 0 \\ 0 & 0 & 0 & 2.36 \times 10^{-6} \end{bmatrix}, \\
\sigma^2 = 0.0159
\end{gathered} \tag{3.4.2}$$

where β and γ are the fixed effects, b_i the random effects with a covariance matrix Ψ. Parameter values for β, γ, Ψ and σ^2 are given in Eq. (3.4.2) and were obtained by fitting the model to clinical trial data from patients that underwent coronary artery bypass grafting (CABG) surgery. The values simulated by the model, taking only the fixed effects into account, are show in Fig. 3.4.1. The bi-exponential model represents a pharmacokinetic model reflecting release of the biomarker after surgery and clearance from the circulation by the kidneys. To simulate data for a new patient i, we draw z_i from a binomial distribution with $P = 0.1$, we plug in the sparse and irregular sequence t_j in Eq. (3.4.1), add between subject variability by sampling from a multivariate normal distribution with covariance matrix Ψ and add Gaussian noise with variance σ^2. If, as a result of random sampling, any $c_{it} < 0$, these values are set to 0.

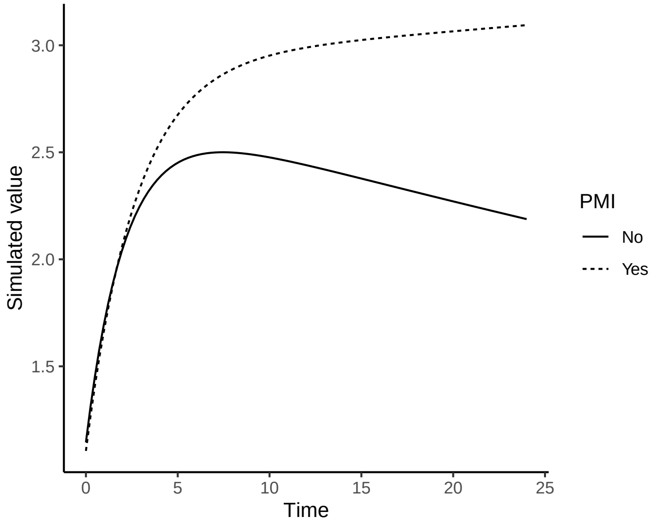


Figure 3.4.1: Simulated values without random effects and residual error (i.e. only fixed effects) from the bi-exponential model using parameter values given in Eq. (3.4.2).

3.5 Results

The area under the ROC-curve (AUC) of the cumulative maximum over time is shown for each approach in Table 3.5.1, with t representing the time. The discriminative ability increases for all approaches as time progresses. Except for the conditional growth chart (CGC) approach, the performance is comparable up to $t = 6$, thereafter the functional longitudinal discriminant analysis (F-LDA) and generalized functional linear model (GFLM) approaches show a clear benefit in terms of AUC. All modeling approaches eventually outperform the raw value. The AUC for each approach, when taking the maximum value for each patient over the time interval $\mathscr{T}$, is given in Table 3.5.2. Again, all classification approaches perform better than using the raw value. The F-LDA and GFLM are clearly superior in terms of discriminative ability. In Table 3.5.3 the sensitivity, specificity and average run length (ARL) of each

t	Raw value	SGC	CGC	TPS	COV-LDA	F-LDA	GFLM
2	0.514 (0.007)	0.514 (0.007)	0.534 (0.007)	0.529 (0.006)	0.528 (0.007)	0.513 (0.008)	0.529 (0.006)
4	0.593 (0.007)	0.567 (0.007)	0.582 (0.007)	0.592 (0.007)	0.586 (0.006)	0.586 (0.009)	0.583 (0.007)
6	0.650 (0.006)	0.616 (0.007)	0.645 (0.006)	0.656 (0.007)	0.635 (0.008)	0.659 (0.008)	0.669 (0.007)
8	0.694 (0.006)	0.664 (0.007)	0.695 (0.006)	0.698 (0.006)	0.676 (0.007)	0.721 (0.008)	0.742 (0.007)
12	0.754 (0.006)	0.735 (0.007)	0.763 (0.006)	0.764 (0.006)	0.736 (0.008)	0.800 (0.008)	0.833 (0.006)
16	0.792 (0.005)	0.794 (0.006)	0.801 (0.005)	0.811 (0.005)	0.783 (0.007)	0.854 (0.006)	0.881 (0.005)
20	0.817 (0.005)	0.839 (0.004)	0.830 (0.004)	0.846 (0.005)	0.818 (0.007)	0.897 (0.005)	0.910 (0.005)
24	0.835 (0.005)	0.874 (0.004)	0.859 (0.004)	0.875 (0.005)	0.847 (0.006)	0.926 (0.004)	0.927 (0.005)

Table 3.5.1: Dynamic classification performance as expressed in area under the ROC-curve (AUC) for each approach at each time t with the standard error in round brackets. Each approach can use all available information up until t to make a prediction for the binary outcome. Static growth chart (SGC); conditional growth chart (CGC); tensor product smooth (TPS); covariance pattern longitudinal discriminant analysis (COV-LDA); functional longitudinal discriminant analysis (F-LDA); generalized functional linear model (GFLM).

	AUC
Raw value$_{max}$	0.834 (0.005)
SGC$_{max}$	0.871 (0.004)
CGC$_{max}$	0.856 (0.004)
TPS$_{max}$	0.872 (0.005)
COV-LDA$_{max}$	0.844 (0.006)
F-LDA$_{max}$	0.923 (0.004)
GFLM$_{max}$	0.925 (0.004)

Table 3.5.2: The AUC of the maximum value in the time interval $\mathscr{T}$ for each approach. Static growth chart (SGC); conditional growth chart (CGC); tensor product smooth (TPS); covariance pattern longitudinal discriminant analysis (COV-LDA); functional longitudinal discriminant analysis (F-LDA); generalized functional linear model (GFLM).

approach are given when using a threshold based on the Youden index. The F-LDA and GFLM approaches show the best results in therms of combining a high sensitivity, specificity, but at the expense of a somewhat longer ARL.

	Sensitivity	Specificity	ARL
Raw value	0.834 (0.010)	0.749 (0.012)	10.768 (0.156)
SGC	0.865 (0.008)	0.794 (0.010)	13.320 (0.244)
CGC	0.832 (0.009)	0.778 (0.009)	12.710 (0.208)
TPS	0.900 (0.008)	0.779 (0.010)	13.954 (0.278)
COV-LDA	0.914 (0.007)	0.720 (0.011)	13.522 (0.276)
F-LDA	0.935 (0.004)	0.844 (0.008)	14.737 (0.210)
GFLM	0.923 (0.007)	0.859 (0.009)	14.466 (0.213)

Table 3.5.3: Sensitivity, specificity and average run length (ARL) for each approach. A threshold based on the Youden index was used as a stopping rule. I.e. if for a patient the predicted value of an approach rises above this threshold, the patient is classified as positive. The mean time until a positive classification is represented by the ARL. Static growth chart (SGC); conditional growth chart (CGC); tensor product smooth (TPS); covariance pattern longitudinal discriminant analysis (COV-LDA); functional longitudinal discriminant analysis (F-LDA); generalized functional linear model (GFLM).

3.6 Illustrative analysis on PMI after CABG surgery

As outlined in the introduction, this study is motivated by the need to detect patients experiencing a perioperative myocardial infarction (PMI) after having undergone coronary artery bypass grafting (CABG) surgery, based on serial measurements of a cardiac biomarker. After surgery, cardiac biomarkers are repeatedly sampled in patients to detect a possible PMI. A PMI is defined as a procedural myocardial infarction whose pathogenesis is multifactorial and can be either graft related or non-graft related [41, 42]. Examples of graft-related PMI include graft failure due to occlusion, kinking or overstretching. Non-graft related PMI can result from procedural difficulties like trauma from surgical manipulation or inadequate myocardial protection. The post-operative rise of cardiac biomarkers, in particular cardiac troponin (cTn), can reflect myocardial damage originating from either (early) graft failure or non-graft related causes. In the former case, minimizing the time to a re-intervention is crucial to save viable myocardium. Yet, all patients experience some unavoidable rise in cardiac biomarkers, simply as a result of the procedure itself. For this study, a dataset of 639 patients who underwent CABG surgery in the Catharina Hospital in Eindhoven, the Netherlands, is available. For more details regarding the study see [43]. For each patient, cTnT was sampled up to 24 hours after surgery and the outcome (PMI yes/no) was recorded. Sampling of cTnT was irregular and more frequent in the first 6 hours after surgery, see Fig. 3.6.1B. Patients with a PMI generally show a sustained release of cTnT from damaged myocardium, instead of a rising-and-falling trend in the first 24 hours after surgery [43–45], see Fig. 3.6.1A. The tensor product smooth (TPS) and generalized functional linear model (GFLM) approaches were fitted to the study dataset, since these approaches showed the best performance for non-historic and historic approaches, respectively. The fitted approaches

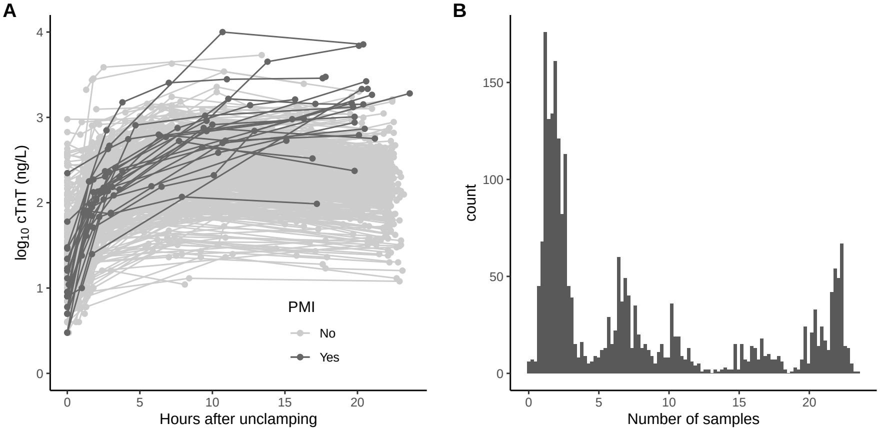


Figure 3.6.1: Overview of studydata.
A: Spaghetti plot of time after aortic unclamping against the $\log_{10}$ transformed value of the measured cardiac troponin (cTn)T concentration in ng/L. A total of 2892 cTnT values were measured for 639 patients undergoing CABG surgery. Profiles of patients diagnosed with a PMI (N = 22) are shown in dark gray, patients without PMI in light gray.
B: Histogram of sampling times (excluding $t = 0$). cTnT was measured at t = 0 (before surgery) and at irregular times after surgery, centering around 1.5, 2, 6, 12 and 24 h after aortic unclamping.

are visualized in Fig. 3.6.2. As the data is too irregularly sampled to calculate the area under the ROC-curve (AUC) as frequent as in the simulated data, the AUC was calculated at $t = 6, 12$ and 24 h after surgery, again using the cumulative maximum as described previously. Table 3.6.1 shows the AUC of the different approaches, using all measurements up to $t = 6, 12$ and 24 h respectively. The AUC of the maximum value of each approach is given in Table 3.6.2. The TPS approach performs best overall. To determine if there is a clinical benefit to using a model based approach instead of the raw biomarker value as a cutoff to initiate further diagnosis for a PMI, we compared the model based approaches to the current clinical guideline in terms of sensitivity, specificity, and average run length (ARL). Since the clinical guide-

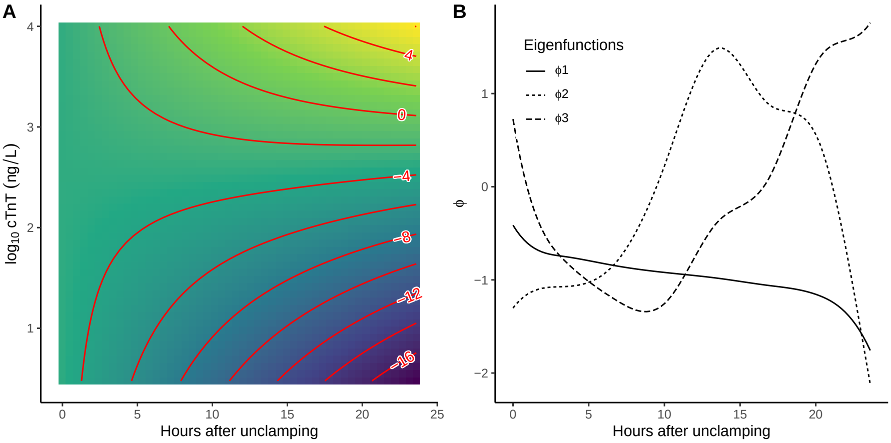


Figure 3.6.2: Predictions from the tensor product smooth (TPS) and generalized functional linear model (GFLM) approaches.
A: Contour plot with contour lines in red, reflecting the predicted probability of a PMI by the TPS approach. Note that the predictions are on the linear predictor scale which can be converted to a probability by applying the logit function. E.g. the ”0” contour line, represents the line with a probability of a PMI of 0.5.
B: This plot shows the three eigenfunctions that explain 95 % of the variance, extracted by the FACEs approach of the GFLM. The first eigenfunction $\phi 1$ is negatively associated with the outcome of a PMI, whereas the second eigenfunction $\phi 2$ is positively associated with a PMI. By obtaining conditional expectations for a new subject, based on these eigenfunctions, the probability of a PMI can be obtained.

	cTnT	TPS	GFLM
$t \leq 6$	0.584 (0.463 - 0.584)	0.517 (0.382 - 0.517)	0.539 (0.416 - 0.539)
$t \leq 12$	0.731 (0.604 - 0.731)	0.743 (0.614 - 0.743)	0.735 (0.599 - 0.735)
$t \leq 24$	0.907 (0.83 - 0.907)	0.931 (0.872 - 0.931)	0.893 (0.793 - 0.893)

Table 3.6.1: The area under the ROC-curve (AUC) and the 95% confidence interval for each approach applied to the study dataset using informtion up to time t. Cardiac troponin (cTn); tensor product smooth (TPS); generalized functional linear model (GFLM).

	AUC
cTnT$_{max}$	0.907 (0.83 - 0.907)
TPS$_{max}$	0.931 (0.87 - 0.931)
GFLM$_{max}$	0.893 (0.793 - 0.893)

Table 3.6.2: The area under the ROC-curve (AUC) for each approach applied to the study dataset using information up to time t. Cardiac troponin (cTn); tensor product smooth (TPS); generalized functional linear model (GFLM).

line recommends a threshold of 140 ng/L for cTnT [42], in the study dataset this results in a sensitivity of 0.955 and a specificity of 0.188. By using ROC curve analysis we calculated a threshold for each approach that corresponded to a 0.955 sensitivity. Subsequently the performance in terms of specificity, true positives, false positives, true negative, false negatives and average run length (ARL) were compared, see Table 3.6.3. We conclude that the modeling based approaches can provide a similar sensitivity as the guideline, whilst offering a higher specificity, at the expense of a longer time until detection.

	Sensitivity	Specificity	TP	FP	TN	FN	ARL
cTnT Guideline	0.955	0.188	21	501	116	1	5.40
TPS	0.955	0.786	21	132	485	1	11.19
GFLM	0.955	0.438	21	347	270	1	8.10

Table 3.6.3: Performance of different modeling approaches when defining a threshold with a sensitivity of 0.955, this sensitivity is similar to the clinical guideline of 140 ng/L for cTnT. True positives (TP); false positives (FP); true negatives (TN); average run length (ARL); cardiac troponin (cTn); tensor product smooth (TPS); generalized functional linear model (GFLM).

3.7 Discussion

In this study we described and compared several popular non-parametric modeling approaches that combine irregularly and sparsely sampled measurements with a binary outcome. Our results show that functional regression models that implicitly incorporate historic information through estimation of a covariance function, outperform models that do not incorporate historic information. The generalized functional linear model (GFLM) performed best of the approaches that incorporate historic information, while the tensor product smooth (TPS) approach performed best of the approaches that do not incorporate historic information.

It would appear that growth charts seem to be less suitable to (dynamic) classification of irregularly and sparsely sampled curves. In part, this is due to the fact that growth charts are not developed with classification in mind [32]. The conditional growth chart (CGC) approach, which explicitly incorporates historical information, seems to offer a benefit in early detection of cases but not for later time points (see Table 3.5.1). In this study, the CGC model as defined in Eq. (3.2.6) is referred to as a ”global model” by Wei et al. [32]. This model is restrictive, in the sense that it assumes that auto-regressive (AR) coefficients are linear functions of measurement time distances. Wei et al. describe several generalizations of the global model, for example allowing the AR coefficients to be functions of measurement time distances. These generalizations could improve CGC model performance. However, both methods are out-preformed by the TPS approach, which can theoretically also be expanded with AR coefficients.

The functional regression approaches (functional longitudinal discriminant analysis (F-LDA) and GFLM) performed best on the simulated data. Although the F-LDA and GFLM approaches perform quite similar in this study, this may not always be the case. In a study by Hughes et al. that compared three different approaches to calculating a patient’s posterior group membership based on random effects [46]. They conclude that the marginal approach (comparable to our F-LDA approach) works best when the mean profile is noticeably different between groups. In the case that the difference between groups is characterized by the variability about the mean profile the GFLM approach may be a better choice. The GFLM approach uses the principal component (PC) scores as covariates in a logistic regression model. However, using PC scores as predictors is not without its downsides. With PC scores there is no guarantee that groups are separated best in the direction of the PC scores with the highest variance [47]. In this study we choose the PC scores based on the percentage variance explained, but an alternative approach could be to apply a variable selection technique to choose PC scores based on their ability to separate groups.

The GFLM approach did not outperform the TPS approach in the illustrative example. This, however, does not invalidate the conclusions from the simulations for two main reasons. First, the data from the illustrative example was more sparse than the simulated data. Therefore, approaches that rely on historic information are affected more in terms of predictive performance. We expect that with more frequent sampling, the GFLM approach is capable of outperforming the TPS approach. In case the data are very sparse or it is evident from theory that past values do not provide any prognostic information, the recommended alternative is a TPS approach. Second, as this is only an illustrative example and there are only a small number of cases (22), the data was not split in a train and test set to objectively evaluate performance.

Future research could be to examine the effect of the degree of sparsity on the performance of functional approaches that incorporate historic information, versus non-functional approaches that do not incorporate historic information. This, to determine the robustness against increasing levels of sparsity that are common in a clinical setting.

References

1. Wulfsohn, M. S. & Tsiatis, A. A. A joint model for survival and longitudinal data measured with error. *Biometrics,* 330–339 (1997).

2. Papageorgiou, G., Mauff, K., Tomer, A. & Rizopoulos, D. An overview of joint modeling of time-to-event and longitudinal outcomes. *Annual review of statistics and its application* **6,** 223–240 (2019).

3. Brant, L. J. *et al.* Screening for prostate cancer by using random-effects models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **166,** 51–62 (2003).

4. Horrocks, J. & van Den Heuvel, M. J. Prediction of pregnancy: a joint model for longitudinal and binary data. *Bayesian Analysis* **4,** 523–538 (2009).

5. Dandis, R. *et al.* A tutorial on dynamic risk prediction of a binary outcome based on a longitudinal biomarker. *Biometrical Journal* **62,** 398–413 (2020).

6. Tanner, K. T., Sharples, L. D., Daniel, R. M. & Keogh, R. H. Dynamic survival prediction combining landmarking with a machine learning ensemble: Methodology and empirical comparison. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **184,** 3–30 (2021).

7. Rizopoulos, D., Molenberghs, G. & Lesaffre, E. M. Dynamic predictions with time-dependent covariates in survival analysis using joint modeling and landmarking. *Biometrical Journal* **59,** 1261–1276 (2017).

8. Ferrer, L., Putter, H. & Proust-Lima, C. Individual dynamic predictions using landmarking and joint modelling: validation of estimators and robustness assessment. *Statistical methods in medical research* **28,** 3649–3666 (2019).

9. Maziarz, M., Heagerty, P., Cai, T. & Zheng, Y. On longitudinal prediction with time-to-event outcome: comparison of modeling options. *Biometrics* **73,** 83–93 (2017).

10. Yau, J. M. *et al.* Impact of Perioperative Myocardial Infarction on Angiographic and Clinical Outcomes Following Coronary Artery Bypass Grafting (from PRoject of Ex-vivo Vein graft ENgineering via Transfection [PREVENT] IV). *The American Journal of Cardiology* **102,** 546–551 (2008).

11. Tomasko, L., Helms, R. W. & Snapinn, S. M. A discriminant analysis extension to mixed models. *Statistics in medicine* **18,** 1249–1260 (1999).

12. Marshall, G. & Barón, A. E. Linear discriminant models for unbalanced longitudinal data. *Statistics in medicine* **19,** 1969–1981 (2000).

13. Wernecke, K.-D., Kalb, G., Schink, T. & Wegner, B. A mixed model approach to discriminant analysis with longitudinal data. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **46,** 246–254 (2004).

14. Kohlmann, M., Held, L. & Grunert, V. P. Classification of therapy resistance based on longitudinal biomarker profiles. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **51,** 610–626 (2009).

15. Fieuws, S., Verbeke, G., Maes, B. & Vanrenterghem, Y. Predicting renal graft failure using multivariate longitudinal profiles. *Biostatistics* **9,** 419–431 (2008).

16. Komárek, A., Hansen, B. E., Kuiper, E. M., van Buuren, H. R. & Lesaffre, E. Discriminant analysis using a multivariate linear mixed model with a normal mixture in the random effects distribution. *Statistics in medicine* **29,** 3267–3283 (2010).

17. Hughes, D. M., Komárek, A., Czanner, G. & Garcia-Finana, M. Dynamic longitudinal discriminant analysis using multiple longitudinal markers of different types. *Statistical methods in medical research* **27,** 2060–2080 (2018).

18. James, G. M. & Hastie, T. J. Functional linear discriminant analysis for irregularly sampled curves. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **63,** 533–550 (2001).

19. Bottomley, C. *et al.* Functional linear discriminant analysis: a new longitudinal approach to the assessment of embryonic growth. *Human Reproduction* **24,** 278–283 (2009).

20. Wang, C., Wang, N. & Wang, S. Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics* **56,** 487–495 (2000).

21. Albert, P. S. A linear mixed model for predicting a binary event from longitudinal data under random effects misspecification. *Statistics in medicine* **31,** 143–154 (2012).

22. Zhang, N., Chen, H. & Zou, Y. A joint model of binary and longitudinal data with non-ignorable missingness, with application to marital stress and late-life major depression in women. *Journal of Applied Statistics* **41,** 1028–1039 (2014).

23. Li, D., Wang, X., Song, S., Zhang, N. & Dey, D. K. Flexible link functions in a joint model of binary and longitudinal data. *Stat* **4,** 320–330 (2015).

24. Müller, H.-g. Functional modelling and classification of longitudinal data. *Scandinavian Journal of Statistics* **32,** 223–240 (2005).

25. Crainiceanu, C. M., Staicu, A.-M. & Di, C.-Z. Generalized multilevel functional regression. *Journal of the American Statistical Association* **104,** 1550–1561 (2009).

26. De la Cruz, R., Marshall, G. & Quintana, F. A. Logistic regression when covariates are random effects from a non-linear mixed model. *Biometrical journal* **53,** 735–749 (2011).

27. De la Cruz, R., Meza, C., Arribas-Gil, A. & Carroll, R. J. Bayesian regression analysis of data with random effects covariates from nonlinear longitudinal measurements. *Journal of multivariate analysis* **143,** 94–106 (2016).

28. Hastie, T. & Tibshirani, R. Varying-coefficient models. *Journal of the Royal Statistical Society: Series B (Methodological)* **55,** 757–779 (1993).

29. Wood, S. N. *Generalized Additive Models* (Chapman and Hall/CRC, Boca Raton, FL, 2017).

30. Cole, T. J. The development of growth references and growth charts. *Annals of human biology* **39,** 382–394 (2012).

31. Wei, Y., Pere, A., Koenker, R. & He, X. Quantile regression methods for reference growth charts. *Statistics in medicine* **25,** 1369–1382 (2006).

32. Wei, Y. & He, X. Conditional growth charts. *The Annals of Statistics* **34,** 2069–2097 (2006).

33. Hastie, T. & Tibshirani, R. Generalized additive models: some applications. *Journal of the American Statistical Association* **82,** 371–386 (1987).

34. Eilers, P. H. & Marx, B. D. Flexible smoothing with B-splines and penalties. *Statistical science* **11,** 89–121 (1996).

35. Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R. & Goude, Y. Fast calibrated additive quantile regression. *Journal of the American Statistical Association* **116,** 1402–1412 (2021).

36. Roy, A. & Khattree, R. Discrimination and classification with repeated measures data under different covariance structures. *Communications in Statistics—Simulation and Computation®* **34,** 167–178 (2005).

37. Pinheiro, J. & Bates, D. *Mixed-effects models in S and S-PLUS* (Springer science & business media, 2006).

38. Xiao, L., Li, C., Checkley, W. & Crainiceanu, C. Fast covariance estimation for sparse functional data. *Statistics and computing* **28,** 511–522 (2018).

39. Yao, F., Müller, H.-G. & Wang, J.-L. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association* **100,** 577–590 (2005).

40. R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing (Vienna, Austria, 2022).

41. Thygesen, K. *et al.* Fourth universal definition of myocardial infarction (2018). *Journal of the American College of Cardiology* **72,** 2231–2264 (2018).

42. Thielmann, M. *et al.* ESC Joint Working Groups on Cardiovascular Surgery and the Cellular Biology of the Heart Position Paper: Peri-operative myocardial injury and infarction in patients undergoing coronary artery bypass graft surgery. *European heart journal* (2017).

43. Deneer, R. *et al.* Detecting patients with PMI post-CABG based on cardiac troponin-T profiles: a latent class mixed modeling approach. *Clinica Chimica Acta* **504,** 23–29 (2020).

44. Ge, W. *et al.* High-sensitivity troponin T release profile in off-pump coronary artery bypass grafting patients with normal postoperative course. *BMC cardiovascular disorders* **18,** 1–7 (2018).

45. Tevaearai Stahel, H. T. *et al.* Clinical relevance of troponin T profile following cardiac surgery. *Frontiers in cardiovascular medicine,* 182 (2018).

46. Hughes, D. M., El Saeiti, R. & Garcia-Fiñana, M. A comparison of group prediction approaches in longitudinal discriminant analysis. *Biometrical Journal* **60,** 307–322 (2018).

47. Jolliffe, I. T. *Principal component analysis for special types of data* (Springer, 2002).

4

Development and validation of an early warning score to identify COVID-19 in the emergency department based on routine laboratory tests: a multicenter case-control study

Arjen-Kars Boer*, **Ruben Deneer***, Maaike Maas, Heidi SM Ammerlaan, Roland HH van Balkom, Wendy AHM Thijssen, Sophie Bennenbroek, Mathie Leers, Remy JH Martens, Madelon M Buijs, Jos J Kerremans, Muriël Messchaert, Jeroen J van Suijlen, Natal AW van Riel & Volkher Scharnhorst

* both authors contributed equally

Adapted from: Development and validation of an early warning score to identify COVID-19 in the emergency department based on routine laboratory tests: a multicentre case–control study. *BMJ Open* **12,** e059111 (2022).

Abstract

Objectives Identifying patients with a possible severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection in the emergency department (ED) is challenging. Symptoms differ, incidence rates vary and test capacity may be limited. As polymerase chain reaction (PCR) testing all ED patients is neither feasible nor effective in most centers, a rapid, objective, low-cost early warning score to triage ED patients for a possible infection is developed.

Design Case-control study

Setting Secondary and tertiary hospitals in the Netherlands.

Participants Patients presenting at the ED with venous blood sampling from July 2019 to July 2020 (N = 10.417, 279 SARS-CoV-2 positive). The temporal validation cohort covered the period from July 2020 to October 2021 (N = 14.080, 1093 SARS-CoV-2 positive). The external validation cohort consisted of patients presenting at the ED of three hospitals in the Netherlands (N = 12.061, 652 SARS-CoV-2 positive).

Primary outcome measures The primary outcome was one or more positive SARS-CoV-2 PCR-test results, within one day prior to, or one week after, ED presentation.

Results The resulting “CoLab-score” consists of 10 routine laboratory measurements, and age. The score showed good discriminative ability (area under the ROC-curve (AUC): 0.930, 95% CI: 0.909 to 0.945). The lowest CoLab-score had a high sensitivity for coronavirus disease 2019 (COVID-19) (0.984, 95% CI: 0.970 to 0.991, specificity: 0.411, 95% CI: 0.285 to 0.520). Conversely, the highest score had high specificity (0.978, 95% CI: 0.973 to 0.983, sensitivity: 0.608, 95% CI: 0.522 to 0.685). Results were confirmed in temporal and external validation.

Conclusions The CoLab-score is based on routine laboratory measurements and is available within one hour after presentation. Depending on the prevalence, COVID-19 may be safely ruled-out in over one third of ED presentations. Highly suspect cases can be identified regardless of

presenting symptoms. The CoLab-score is continuous, in contrast to the binary outcome of lateral flow testing, and can guide PCR testing and triage ED patients.

4.1 Introduction

Coronavirus disease 2019 (COVID-19), caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has evolved into a global pandemic in 2020 [1]. For emergency department (ED) physicians, identifying presenting patients with a possible COVID-19 infection remains challenging since symptoms like fever, shortness of breath or coughing overlap with other illnesses [2, 3] . It is crucial however, to identify a possible COVID-19 infection as early as possible. Early identification prevents further spreading and protects hospital staff by isolating a suspected patient, pending the results of a SARS-CoV-2 RNA polymerase chain reaction (PCR) test and/or chest CT. Conversely, when PCR testing or isolation treatment capacity is limited, ruling-out COVID-19 as soon as possible can save valuable resources. In the era of electronic health records and clinical prediction models, developing an early warning score that can assist ED physicians in identifying patients presenting at the ED with COVID-19 is of great value. Moreover, if only routine ED test results are required as input, the score can be easily adopted by EDs worldwide, potentially reduce diagnostic costs and accelerate patient triage. Many COVID-19 prediction models have already been developed, the living systematic review by Wynants et. al [4] provides an extensive overview and critical appraisal. Unfortunately, only few models have found their way into routine care at the ED [5, 6]. Early models were based on relatively small sample sizes, hampered by selection bias or were over-fitted by selecting too many features [4–6]. Aside from methodological shortcomings of early models, most models are not developed as an early warning score for all ED patients. Firstly, they require features from tests that are not routinely performed or logged for all ED patients (e.g. the COVID-19 Reporting and Data System (CO-RADS) score from a CT-scan [7] or non-lab based clinical variables in the Pandemic Respiratory Infection Emergency System Triage (PRIEST) early warning score [8], and are therefore not straightforward to implement or scale to a large ED patient population. Secondly, the population on which models are commonly based, are PCR-tested patients, i.e. a pre-selection of a possible COVID-19 infection has already been done by physicians. Only two studies were identified that focus on patients presenting at the ED, include unsuspected (and pre-pandemic) patients as controls, and rely solely on routine (laboratory) tests [9, 10]. In this study we report the development and validation of an early warning score that, based on routine ED laboratory tests, estimates the risk of a possible COVID-19 infection in patients who undergo routine laboratory testing at presentation. The score can assist ED physicians in triaging patients and prevent further transmission of COVID-19 by quickly identifying possibly infected patients or ruling out a possible infection when resources are scarce.

4.2 Methods

4.2.1 Study design

This is a retrospective case-control study where routine laboratory test results, combined with age and gender, from all patient presenting at the emergency department (ED) of the Catharina Hospital Eindhoven from July 2019 to July 2020 were combined with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) polymerase chain reaction (PCR) test results in a development dataset. A model that could predict the presence of a coronavirus disease 2019 (COVID-19) infection was fit to this dataset. Performance of the model was assessed by i) internal validation, ii) temporal validation and iii) external validation by using data from the ED of three other centers. The study was reviewed by the Medical research Ethics Committees United (MEC-U) under study number W20.071, which confirmed that the Medical Research Involving Human Subjects Act (In Dutch: WMO) does not apply to this study. The study was thereafter reviewed and approved by the internal hospital review board.

4.2.2 Patient and Public Involvement

Patients were not involved in the design, conduct or reporting of this study.

4.2.3 Development dataset

All ED presentations at the Catharina Hospital Eindhoven from July 2019 to July 2020 were included in the development dataset, provided that routine laboratory testing had been requested by the attending ED physician. The rationale for this inclusion period is to limit the effect of seasonal variation in the ED patient population by including the summer, fall and winter season of 2019 (control patients) and the winter, spring and summer season of 2020 (case and control patients). The routine laboratory panel at the ED consists of 28 laboratory tests. In some cases not all tests in the routine panel were requested or one or more quantitative results were not available due to analytical interference (hemolysis, lipemia or icterus). The routine ED laboratory panel is requested for (adult) patients presenting with abdominal pain, chest pain, shortness of breath, syncope, sepsis or other non-specific complaints, or for patients (including non-adult patients) presenting with specific complaints where a suspected diagnosis has to be ruled-in or ruled-out. Presentations with one or more missing values in any of the 28 laboratory test in the routine ED panel, were excluded. Presentations with one or more extreme lab results, >10 times standard deviation from the median, were also excluded to minimize the effect on the estimation of regression coefficients. The median was chosen as a measure of central tendency due to its resistance for outliers. After the first case of COVID-19 in the Netherlands, all patients with symptoms of COVID-19 (either fever and/or respiratory symptoms) were subjected to nasopharyngeal PCR testing for SARS-CoV-2 RNA. PCR testing was performed by commercial tests that were approved by the Dutch national institute of public health (RIVM). If a patient had a positive PCR result in the past, subsequent presentations were excluded as re-presentations might be clinically different from de novo presentations. The ED lab panel results were matched to SARS-CoV-2 PCR results if the underlying nasopharyngeal swab had been taken $\leq$ 1 day prior, or $\leq$ 1 week after initial blood withdrawal at the ED. If multiple PCR tests were performed in this window, and at least one PCR test was positive, the presentation was labelled “PCR-positive”. If all PCR test results in the time window were negative, the presentation was labelled as “PCR-negative”. If no PCR tests were performed in the time window and the presentation occurred after the first case of COVID-19 in the Netherlands, the presentation was labelled as “Untested”. All presentations before the first case were labelled as “Pre-COVID-19”.

4.2.4 Laboratory tests

The routine laboratory panel consisted of hemocytometric and chemical analyses. The hemocytometric tests were performed on Sysmex XN-10 instruments (Sysmex Corp., Kobe, Japan) and consisted of hemoglobin, hematocrit, erythrocytes, mean corpuscular volume (MCV), mean cellular hemoglobin (MCH), mean cellular hemoglobin concentration (MCHC), thrombocytes, leukocytes, neutrophils, eosinophils, basophils, lymphocytes and monocytes. The chemical analyses were performed on a Cobas 8000 Pro (Roche Dx, Basel, Switzerland) instrument and consisted of glucose, total bilirubin, aspartate aminotransferase (ASAT), alanine aminotransferase (ALAT), lactate dehydrogenase (LD), creatine kinase (CK), alkaline phosphatase (ALP), gamma-glutamyltransferase (gGT), blood urea nitrogen (BUN), creatinine, CKD-EPI estimated glomerular filtration rate (CKD-EPI), potassium, sodium, chloride, albumin (bromocresol green) and C-reactive protein (CRP). These results were combined with age and gender.

4.2.5 Modelling

All data were processed and analyzed in R version 4.1.1 [11]. Laboratory results, combined with age and gender were used as covariates in a regression model. Cases were defined as ED presentations labelled as “PCR-positive”, controls were all other presentations (i.e. “PCR-negative”, “Untested” or “Pre-COVID-19”). To achieve predictive accuracy, limit overfitting and perform feature selection, penalized logistic regression with an adaptive lasso penalty was chosen [12, 13]. To minimize missing data, all non-numeric results at the extremes of the measuring range, were converted to numeric results by removing the “<” and ”>” signs. For CKD-EPI and CRP the raw precursor value was used instead of >90 ml/min/m^2 and <6 mg/L, respectively. Considering that laboratory results of bilirubin, ASAT, ALAT, LD, CK, ALP and gGT can have heavy (right) tailed distributions, which in turn impacts model predictions, these variables were transformed logarithmically. More details regarding model fitting can be found in Section 4.A. Models were fitted using the `glmnet`-package [14].

4.2.6 CoLab-score

Since this is a retrospective case-control study, the sample prevalence may not reflect the true/current COVID-19 prevalence. To obtain well-calibrated probabilities the intercept term in the model should be adjusted according to the current prevalence (details can be found Section 4.A) [15]. However, adjusting the intercept term is not straightforward to implement in clinical practice, therefore the linear predictor of the model was categorized into a score, this score is hereafter referred to as the "CoLab-score". The categorization is based on a number needed to test of 15 (i.e. one is willing to PCR test 15 patients to find one positive) and prevalence cut-points of 1%, 2%, 5%, 10% and 40% using the intercept adjustment formula by King [15]. The intervals obtained through these breaks correspond to CoLab-scores 5 to 0, respectively. Score 0 reflects low-risk for COVID-19 and score 5 reflects high-risk. More details regarding the rationale of the CoLab-score categorization can be found in Section 4.A.

4.2.7 Internal validation

To assess model performance while taking overfitting into account, bootstrapping was performed. 1000 bootstrap samples were generated from the original data. On each bootstrap sample, the full model fitting procedure and CoLab-score conversion were performed. Optimism adjusted performance measures of the CoLab-score were obtained by applying the 0.632 bootstrap rule to the in-sample and out-of-bag-sample performance [16]. Performance measures included, the area under the ROC-curve (AUC), sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of each CoLab-score. The `pROC`-package was used to calculate performance measures [17]. Although the full inclusion period from July 2019 to July 2020 was used for model fitting, the performance was evaluated on the period starting from the first COVID-19 infection (24th of February 2020) to July 2020. This was done to obtain performance measures that would reflect real world performance.

4.2.8 Temporal validation

For temporal validation, results from our center were prospectively analyzed from July 2020 to October 2021. During this period, the Netherlands was struck by a second wave of COVID-19 infections, starting in the fall of 2020 and subsiding in the summer of 2021. In this period there was also more widespread external PCR testing by municipal health services. The results of external conducted PCR tests were not available to our study. To overcome this limitation, the outcome in the temporal validation cohort was chosen as a composite of the hospital registration of a confirmed COVID-19 infection and/or at least one positive PCR test result. This period also covers both the emergence of new SARS-CoV-2 variants as well as vaccine rollout. However, neither vaccination status nor genomic sequencing was available to determine whether a patient was vaccinated or which variant caused the infection. Therefore, data from the Dutch national institute of public health (RIVM) was used, to divide the temporal validation period into three phases: i) from July 2020 until March 2021, no vaccination and no variants of concern identified ii) from March 2021 until June 2021, partial vaccination and B.1.1.7 (Alpha) variant identified as dominant iii) from June 2021 until October 2021, widespread vaccination and B.1.617.2 (Delta) variant identified as dominant. See Section 4.B for more details. The temporal validation consisted of assessing the AUC, sensitivity, specificity, PPV and NPV of each CoLab-score threshold for the entire period, as well as for each phase separately to determine a possible effect of vaccination and new variants on performance. Model calibration was assessed graphically using the `rms`-package [18].

4.2.9 External validation

For the external validation, several centers in the Netherlands were approached and assessed if the required panel of laboratory tests and SARS-CoV-2 PCR test results were available. Seven centers responded and three centers fulfilled the inclusion criteria: Gelre Hospitals (center 1), Atalmedial Diagnostic Centers, location Alrijne Hospital Leiderdorp (center 2) and Zuyderland Medical Center (center 3). The hematological parameters were measured with Sysmex XN10/XN20 (center 1), CELL-DYN-Sapphire

(Abbott Laboratories) (center 2) and Sysmex XN10 instruments (center 3). The clinical chemistry parameters were measured with Architect c14100/c160000 (Abbott Laboratories) (center 1), Architect ci4100 (Abbott Laboratories) (center 2) and Cobas 8000 instruments (Roche Dx) (center 3). The external validation was similar to the temporal validation and consisted of assessing the AUC, sensitivity, specificity, PPV and NPV of each CoLab-score threshold. Calibration was assessed graphically analogous to the temporal validation dataset.

4.3 Results

4.3.1 Development dataset

12.879 emergency department (ED) presentations of 10.327 patients from July 2019 to July 2020 were included. After excluding cases with an incomplete lab panel, patient presentations that occurred after a positive polymerase chain reaction (PCR) test in the past (re-presentations) and presentations with extreme values (>10 times standard deviation) in any of the lab results, 10.417 presentations of 8610 patients remained, see Fig. 4.3.1 for the inclusion flow. Descriptive statistics of ED presentations are shown in Table 4.3.1, dark grey marked figures indicate a clinically relevant difference from the Pre-COVID-19 category (based on the total allowable error [19]). For the PCR positives (N = 279), 91% (95% CI: 88 to 94%) of the cases were tested positive in their first PCR. The remaining 24 patients were positive in their second (N = 18), third (N = 5) or fourth (N = 1) PCR.

4.3.2 CoLab-score

The model obtained through adaptive lasso regression contained eleven variables, which are depicted with their regression coefficients (weights) in Table 4.3.2. A larger β-coefficient does not imply that a variable is more important in predicting the odds of testing positive for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), since variables are on different scales.

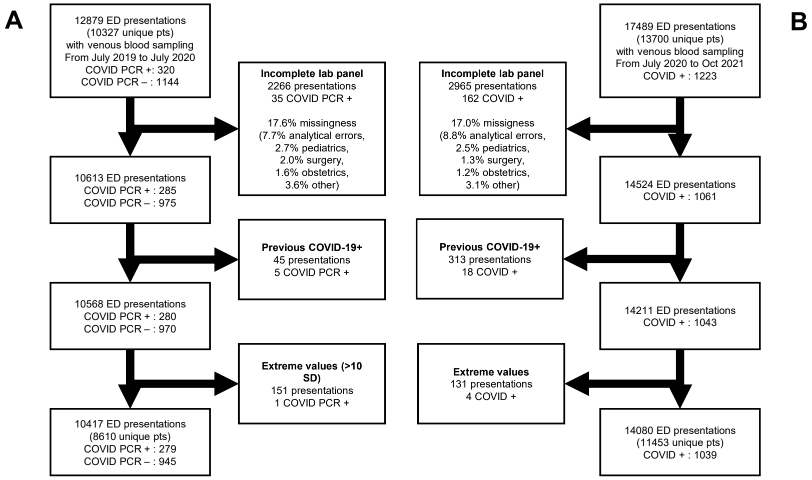


Figure 4.3.1: Inclusion flow of patients in the development (A) and temporal validation (B) dataset.
All patient admissions with routine venous blood sampling at the ED were included. For the development dataset, completeness of the lab panel was assessed for all 28 laboratory tests , for the temporal validation dataset this was only necessary for 10 laboratory tests. The major causes of missingness are described in the text. In the development dataset, presentations with extreme values (>10 SD) were excluded. The same limits were applied to the temporal validation dataset (see Table 4.3.2 for limits).

The most important variables are basophiles, eosinophils and lactate dehydrogenase (LD). As shown in Fig. 4.3.2, the linear predictor clearly discriminates between coronavirus disease 2019 (COVID-19) and non-COVID-19. The linear predictor is converted to CoLab-scores 0 – 5 with the cut-points depicted in Fig. 4.3.2.

4.3.3 Internal validation

The model was validated in the period starting from the first COVID-19 infection to July 2020, in this period the mean prevalence was 7.2%. The area under the ROC-curve (AUC) of the CoLab-score is 0.930 (95% CI: 0.909 to 0.945). Diagnostic performance is shown in Table 4.3.3. A CoLab-score of

	Pre-COVID (N = 5890)	Untested (N = 3303)	PCR negative (N = 945)	PCR positive (N = 279)
Age in years	61.5 (20.8)	60.4 (20.8)	66.0 (17.6)	69.1 (15.1)
Female gender	2909 (49.4)	1659 (50.2)	466 (49.3)	95 (34.1)
Specialism				
Internal medicine	1648 (28.0)	896 (27.1)	244 (25.8)	71 (25.4)
Surgery	1007 (17.1)	679 (20.6)	51 (5.4)	5 (1.8)
Neurology	775 (13.2)	468 (14.2)	64 (6.8)	5 (1.8)
Pulmonary medicine	714 (12.1)	220 (6.7)	326 (34.5)	167 (59.9)
Cardiology	560 (9.5)	322 (9.7)	145 (15.3)	6 (2.2)
Urology	309 (5.2)	148 (4.5)	15 (1.6)	7 (2.5)
Gastroenterology	306 (5.2)	224 (6.8)	27 (2.9)	1 (0.4)
Geriatrics	189 (3.2)	95 (2.9)	52 (5.5)	15 (5.4)
Orthopedics	147 (2.5)	109 (3.3)	11 (1.2)	0 (0.0)
Gynaecology	118 (2.0)	82 (2.5)	2 (0.2)	0 (0.0)
Other	117 (2.0)	60 (1.8)	8 (0.8)	2 (0.7)
Hemoglobin in mmol/L	8.2 (1.3)	8.3 (1.3)	8.2 (1.4)	8.6 (1.1)
Hemoglobin in g/L	13.2 (2.1)	13.3 (2.0)	13.3 (2.2)	13.8 (1.8)
Hematocrit in L/L	0.403 (0.059)	0.405 (0.056)	0.405 (0.062)	0.417 (0.047)
Erythrocytes in /pL	4.41 (0.69)	4.43 (0.66)	4.41 (0.72)	4.61 (0.60)
MCV in fl	91.8 (6.4)	91.9 (6.1)	92.4 (6.7)	90.7 (5.5)
MCH in mmol	1.859 (0.157)	1.876 (0.150)	1.874 (0.172)	1.869 (0.141)
MCHC in mmol/L	20.2 (0.9)	20.4 (0.9)	20.3 (1.0)	20.6 (0.8)
Thrombocytes in /nL	262.5 (98.9)	265.8 (99.7)	269.3 (105.0)	216.8 (122.8)
Leukocytes in /nL	9.30 [7.06, 12.16]	8.92 [7.01, 11.89]	9.66 [7.17, 12.94]	6.33 [4.74, 8.48]
Neutrophils in /nL	6.62 [4.51, 9.53]	6.10 [4.42, 8.94]	7.01 [4.79, 10.02]	4.71 [3.30, 6.94]
Eosinophils in /nL	0.09 [0.03, 0.17]	0.09 [0.03, 0.18]	0.08 [0.02, 0.17]	0.00 [0.00, 0.02]
Basophils in /nL	0.04 [0.02, 0.05]	0.04 [0.02, 0.05]	0.04 [0.02, 0.05]	0.01 [0.01, 0.02]
Lymphocytes in /nL	1.47 [0.93, 2.13]	1.56 [1.05, 2.18]	1.31 [0.80, 2.03]	0.86 [0.59, 1.21]
Monocytes in /nL	0.70 [0.52, 0.93]	0.69 [0.52, 0.91]	0.74 [0.54, 1.01]	0.45 [0.32, 0.64]
Glucose in mmol/L	6.76 [5.83, 8.39]	6.68 [5.76, 8.14]	6.98 [5.95, 8.85]	6.77 [5.98, 8.48]
Bilirubin in umol/L	7.5 [5.0, 11.6]	7.4 [5.1, 10.9]	8.3 [5.6, 12.4]	8.2 [6.3, 11.4]
ASAT in U/L	24.0 [19.1, 32.2]	26.5 [21.6, 35.1]	27.7 [21.7, 39.2]	40.7 [30.2, 57.2]
ALAT in U/L	24.3 [17.8, 35.3]	25.3 [18.4, 36.2]	25.7 [18.4, 40.0]	33.7 [23.3, 50.0]
LD in U/L	201 [173, 240]	198 [170, 236]	215 [178, 263]	300 [238, 403]
CK in U/L	82 [51, 134]	83 [52, 137]	76 [51, 125]	124 [62, 222]
ALP in IU/L	83.0 [68.0, 105.0]	81.0 [65.8, 102.5]	86.9 [67.9, 110.0]	71.0 [58.8, 85.0]
gGT in U/L	27.0 [17.0, 53.0]	28.4 [18.4, 50.5]	37.0 [22.4, 68.9]	42.0 [28.0, 83.5]
BUN in mmol/L	5.7 [4.3, 8.0]	5.8 [4.3, 7.8]	6.2 [4.6, 9.4]	6.1 [4.7, 8.9]
CKD-epi in ml/min/m^2	80.9 [58.0, 99.1]	85.0 [63.5, 103.3]	79.1 [52.1, 96.6]	76.6 [54.9, 91.2]
Creatinine in umol/L	79.0 [64.0, 100.0]	74.1 [60.7, 94.0]	77.7 [62.0, 105.0]	82.0 [67.6, 104.5]
Potassium in mmol/L	4.06 (0.50)	4.03 (0.49)	4.07 (0.55)	3.91 (0.47)
Sodium in mmol/L	139.2 (4.0)	138.5 (3.9)	138.0 (4.3)	136.4 (4.1)
Chloride in mmol/L	104.4 (4.6)	103.8 (4.5)	102.9 (4.8)	101.6 (4.4)
Albumin in g/L	42.4 (4.9)	42.3 (4.5)	40.8 (4.8)	38.4 (3.8)
CRP in mg/L	8.0 [2.0, 41.0]	5.0 [1.4, 30.4]	17.8 [3.5, 68.8]	77.3 [36.5, 135.8]

Table 4.3.1: Descriptive statistics. Normally distributed results are given by the mean and standard deviation (SD) (in round brackets), skewed or heavy tailed distribution by the median and interquartile range (IQR) range [in squared brackets]. Dark grey cells indicate a clinically relevant difference from the Pre-COVID-19 category based on the total allowable error [19].

0 has a negative predictive value (NPV) of 0.997 (95% CI: 0.993 to 0.999) and positive predictive value (PPV) of 0.115 (0.0934 - 0.147), one third (38%, 95% CI: 28 to 514%) of all ED presentations were assigned this score and can

Variable	β	Exclusion limit	Relative importance
Intercept	-6.885		-
Erythrocytes /pL	0.9379	Erythrocytes <2.9 /pL	52 %
Leukocytes /nL	-0.1298		46 %
Eosinophils /nL	-6.834		86 %
Basophils /nL	-47.70	Basophils >0.33 /nL	100 %
$\log_{10}$ of Bilirubin in µmol/L	-1.142	Bilirubin >169 µmol/L	26 %
$\log_{10}$ of LD in U/L	5.369	LD >1564 U/L	58 %
$\log_{10}$ of ALP in IU/L	-3.114	AF >1000 IU/L	45 %
$\log_{10}$ of gGT in U/L	0.3605	gGT >1611 U/L	11 %
Albumin in g/L	-0.1156		45 %
CRP in mg/L	0.002560		15 %
Age in years	0.002275		4 %

Table 4.3.2: Calculation of the CoLab-linear predictor (LP).
The CoLab-LP is calculated by summing the intercept and the products of the 11 variables with their corresponding coefficients (β's). CoLab-LP = – 6.885 + [erythrocytes] × 0.9379 – [leukocytes] × 0.1298 – [eosinophils] × 6.834 – [basophils] × 47.7 – $\log_{10}$([bilirubin]) × 1.142 + $\log_{10}$([LD]) × 5.369 – $\log_{10}$([ALP]) × 3.114 + $\log_{10}$([gGT]) × 0.3605 – [albumin] × 0.1156 + [CRP] × 0.02560 + [age] × 0.002275. The LP can be converted into a CoLab-score (see Fig. 4.3.2) or into a probability if the prevalence is known or estimated (see details in Section 4.A). The CoLab-score is not valid if any of the variables exceed the limits in the third column. The relative importance ranks the importance of variables in predicting the outcome, relative to the most important variable (in this case basophils).

therefore be safely excluded. Conversely, 6% (95% CI: 6 to 8%) of the ED patients had a CoLab-score = 5. Given the PPV of this score (0.683, 95% CI: 0.628 to 0.746, NPV: 0.970, 95% CI: 0.963 - 0.978), subsequent PCR testing is advised.

4.3.4 Temporal validation

As the CoLab-score was developed in our center after the first COVID-19-wave in the Netherlands, the performance was evaluated in our center from July 2020 until October 2021. Lab results from 17.489 ED presentations were collected. After applying the inclusion flow as shown in Fig. 4.3.1B, 14.080 presentations remained, of which 1039 were associated with a COVID-19 infection. The mean prevalence in this period was 7.4%. The AUC of the CoLab-score in the temporal validation set is 0.916 (95% CI: 0.906 to 0.927).

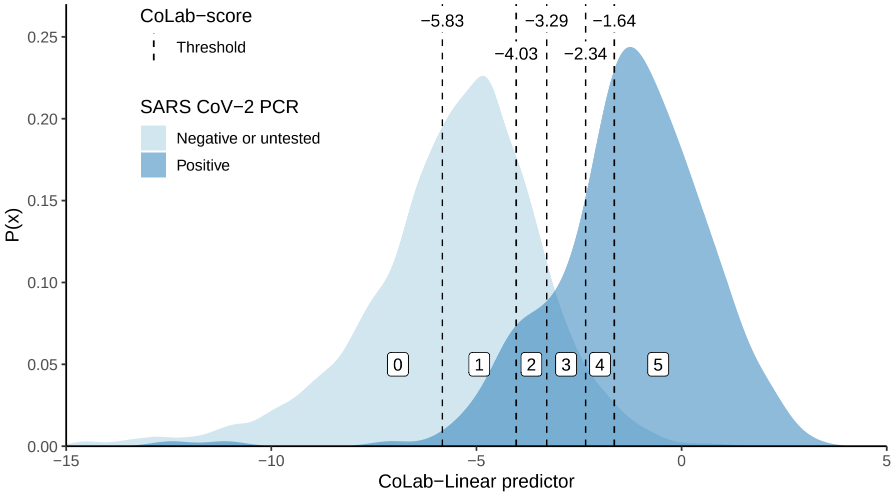


Figure 4.3.2: Probability density plot of the CoLab-linear predictor.
The probability density plots for COVID (dark blue) and non-COVID patients (light blue) are plotted against the linear predictor (for calculation see Table 4.3.2). The CoLab-score cut-offs (–5.83, –4.02, –3.29, –2.34 and –1.64) are depicted with vertical dashed lines. The white-boxed numbers (between the cut-offs) represent the corresponding CoLab-score. Note that although the area under both curves is identical (since these are probability density functions), in absolute numbers the “negative or untested”-group is about 36 times larger than the PCR positive group.

The performance is comparable to the development cohort, although sensitivity is slightly lower and specificity slightly higher (cf. Table 4.3.3 and Table 4.3.4). The temporal validation dataset was also split into three phases according to dominant SARS-CoV-2 variants and vaccine roll-out (see Section 4.B). The discriminative ability was not lower in the second or third phase, compared to the first phase. Diagnostic performance is preserved in terms of sensitivity and specificity, except a moderately reduced sensitivity of scores ≥ 3 in the third phase as compared to the first phase. PPV and NPV are incomparable due to different prevalence/pre-test probabilities in each phase. In terms of the predicted probabilities, model calibration shows that overall predicted probabilities are too low (see Section 4.C for the calibration plot), which is expected since the prevalence differs and the intercept has to be ad-

CoLab-score	Sensitivity	Specificity	PPV	NPV	TP	TN	FP	FN	% of population
0	0.984 (0.969 - 0.991)	0.410 (0.302 - 0.543)	0.115 (0.094 - 0.147)	0.997 (0.993 - 0.999)	273.4 (241.2 - 304.4)	1470.9 (1081.1 - 1950.9)	2119.1 (1633.5 - 2507.6)	4.6 (2.6 - 8.6)	38.0 (28.0 - 51.0)
≤ 1	0.912 (0.892 - 0.952)	0.785 (0.741 - 0.827)	0.248 (0.207 - 0.300)	0.991 (0.989 - 0.995)	253.5 (226.5 - 287.0)	2817.1 (2655.4 - 2961.2)	772.9 (623.2 - 934.5)	24.5 (13.4 - 30.2)	73.3 (69.3 - 77.3)
≤ 2	0.856 (0.816 - 0.895)	0.880 (0.864 - 0.900)	0.357 (0.315 - 0.415)	0.988 (0.984 - 0.991)	238.1 (209.6 - 267.9)	3160.8 (3100.7 - 3233.7)	429.1 (357.3 - 487.1)	39.9 (28.5 - 52.4)	82.9 (80.9 - 83.9)
≤ 3	0.757 (0.706 - 0.809)	0.951 (0.944 - 0.959)	0.546 (0.496 - 0.604)	0.981 (0.976 - 0.985)	210.4 (183.4 - 240.2)	3415.1 (3378.0 - 3456.4)	174.9 (147.0 - 199.3)	67.6 (51.9 - 84.9)	90.0 (89.0 - 91.0)
≤ 4	0.612 (0.530 - 0.706)	0.978 (0.972 - 0.983)	0.683 (0.628 - 0.746)	0.970 (0.963 - 0.978)	170.2 (141.6 - 204.9)	3510.6 (3476.8 - 3547.5)	79.4 (60.3 - 100.4)	107.9 (79.1 - 134.0)	93.7 (91.7 - 93.7)

Table 4.3.3: Bootstrapped diagnostic performance of the CoLab-score in the development dataset.
The development dataset was internally validated for the period March 2020 – July 2020 (N = 3868). The optimism-adjusted bootstrapped sensitivities, specificities, positive predictive value (PPV), negative predictive value (NPV), true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) and fraction of presentations (%) are shown for fixed cut-offs (CoLab-score 0 till ≤4). The numbers in round brackets represent the 95% optimism-adjusted bootstrapped confidence interval (CI). The first column defines the threshold above which CoLab-score a patient is considered positive. Note that "0" lists the sensitivity and NPV of CoLab-score 0 and "≤4" lists the specificity and PPV of CoLab-score 5. Also note that TP, TN, FP and FN are not whole numbers, since these are obtained by applying the 0.632 bootstrap rule.

justed to the prevalence. In this period at least 22 COVID-19 positive patients were identified by the CoLab-score, that initially did not present with COVID-specific symptoms. Most patients had neurological or orthopedic presenting symptoms.

4.3.5 External validation

For external validation, data obtained from three other centers were used, center 1 (N = 1284, 52 COVID-19 positive), center 2 (N = 2899, 99 COVID-19 positive) and center 3 (N = 3545, 336 COVID-19 positive). The inclusion flow is summarized in Fig. 4.3.3.

CoLab-score	Sensitivity	Specificity	PPV	NPV	TP	TN	FP	FN
0	0.967 (0.956 - 0.978)	0.420 (0.411 - 0.428)	0.117 (0.115 - 0.119)	0.994 (0.992 - 0.996)	1005 (993 - 1016)	5476 (5366 - 5587)	7565 (7454 - 7675)	34 (23 - 46)
≤ 1	0.888 (0.870 - 0.908)	0.791 (0.783 - 0.798)	0.253 (0.245 - 0.261)	0.989 (0.987 - 0.991)	923 (904 - 943)	10311 (10215 - 10401)	2730 (2640 - 2826)	116 (96 - 135)
≤ 2	0.820 (0.796 - 0.843)	0.894 (0.889 - 0.899)	0.382 (0.367 - 0.396)	0.984 (0.982 - 0.986)	852 (827 - 876)	11661 (11591 - 11729)	1380 (1312 - 1450)	187 (163 - 212)
≤ 3	0.710 (0.682 - 0.738)	0.962 (0.958 - 0.965)	0.596 (0.573 - 0.618)	0.977 (0.974 - 0.979)	738 (709 - 767)	12540 (12496 - 12582)	501 (459 - 545)	301 (272 - 330)
≤ 4	0.585 (0.556 - 0.615)	0.984 (0.982 - 0.987)	0.750 (0.724 - 0.778)	0.968 (0.965 - 0.970)	608 (578 - 639)	12838 (12811 - 12866)	203 (175 - 230)	431 (400 - 461)

Table 4.3.4: Diagnostic performance of the CoLab-score in the temporal validation dataset. Sensitivities, specificities, positive predictive values (PPV), negative predictive values (NPV), true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are shown for fixed cut-offs (CoLab-score 0 till ≤ 4) with bootstrapped 95% confidence intervals in parentheses.

COVID-19 prevalence differed between the three centers (4.0%, 3.4% and 9.5% respectively) and was lower in centers 1 and 2, and higher in center 3 than in the development dataset. The AUCs of the CoLab-score are 0.904 (95% CI: 0.866 to 0.942), 0.886 (95% CI: 0.851 - 0.922) and 0.891 (95% CI: 0.872 - 0.909), for centers 1, 2, and 3 respectively. Diagnostic performance is shown in Table 4.3.5. The sensitivity of CoLab-score 0 in all centers is $\geq$ 0.96. Therefore, the NPV of CoLab-score 0 was more than 99%. Calibration plots for external centers are shown in Section 4.C, the observed fraction of COVID-19 positives is slightly lower than expected in centers 1 and 2. For center 3, low probabilities appear slightly underestimated and high probabilities slightly overestimated.

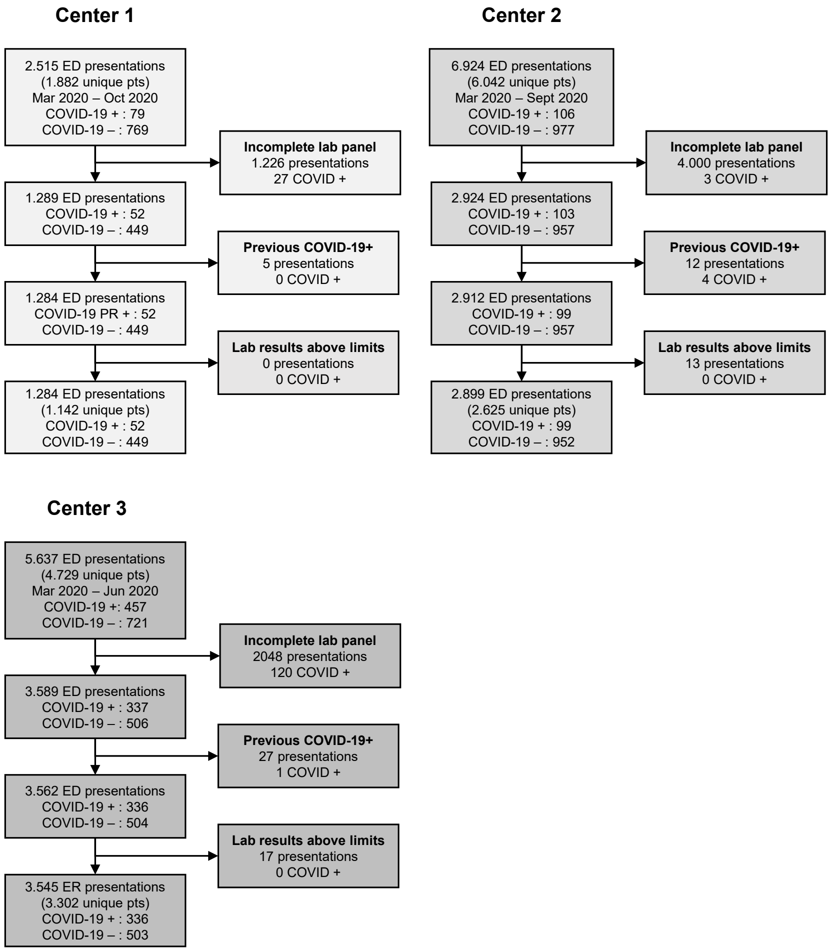


Figure 4.3.3: Inclusion flow of ED patients in three external centers.
All ED presentations with routine venous blood sampling were included. Missingness of lab panels was assessed for the 11 variables in the CoLab-score (see Table 4.3.2). Re-presentations after a positive PCR result or clinical COVID-19 registration were excluded as "previous COVID-19+". Presentations with any laboratory result above the limits of the CoLab-score (see Table 4.3.2) were excluded.

Validation set	CoLab-score	Sensitivity	Specificity	PPV	NPV	TP	TN	FP	FN
Center 1	0	1.000 (1.000 - 1.000)	0.331 (0.307 - 0.358)	0.059 (0.057 - 0.061)	1.000 (1.000 - 1.000)	52 (52 - 52)	410 (380 - 443)	827 (794 - 857)	0 (0 - 0)
Center 2	0	0.961 (0.922 - 0.990)	0.351 (0.333 - 0.369)	0.052 (0.049 - 0.054)	0.996 (0.992 - 0.999)	99 (95 - 102)	985 (935 - 1035)	1823 (1773 - 1873)	4 (1 - 8)
Center 3	0	0.970 (0.950 - 0.988)	0.322 (0.306 - 0.338)	0.130 (0.126 - 0.133)	0.991 (0.984 - 0.996)	327 (320 - 333)	1042 (991 - 1092)	2193 (2143 - 2244)	10 (4 - 17)
Center 1	≤ 1	0.923 (0.846 - 0.981)	0.694 (0.669 - 0.720)	0.113 (0.101 - 0.124)	0.995 (0.991 - 0.999)	48 (44 - 51)	858 (828 - 891)	379 (346 - 409)	4 (1 - 8)
Center 2	≤ 1	0.913 (0.854 - 0.961)	0.678 (0.661 - 0.696)	0.094 (0.087 - 0.101)	0.995 (0.992 - 0.998)	94 (88 - 99)	1905 (1857 - 1953)	903 (855 - 951)	9 (4 - 15)
Center 3	≤ 1	0.914 (0.881 - 0.944)	0.674 (0.657 - 0.691)	0.226 (0.216 - 0.236)	0.987 (0.982 - 0.991)	308 (297 - 318)	2180 (2126 - 2234)	1055 (1001 - 1109)	29 (19 - 40)
Center 1	≤ 2	0.808 (0.692 - 0.904)	0.811 (0.788 - 0.832)	0.152 (0.129 - 0.176)	0.990 (0.984 - 0.995)	42 (36 - 47)	1003 (975 - 1029)	234 (208 - 262)	10 (5 - 16)
Center 2	≤ 2	0.845 (0.777 - 0.913)	0.801 (0.785 - 0.815)	0.135 (0.122 - 0.147)	0.993 (0.990 - 0.996)	87 (80 - 94)	2248 (2205 - 2289)	560 (519 - 603)	16 (9 - 23)
Center 3	≤ 2	0.890 (0.855 - 0.923)	0.794 (0.779 - 0.808)	0.311 (0.294 - 0.328)	0.986 (0.981 - 0.990)	300 (288 - 311)	2569 (2521 - 2615)	666 (620 - 714)	37 (26 - 49)
Center 1	≤ 3	0.750 (0.635 - 0.865)	0.909 (0.892 - 0.925)	0.257 (0.213 - 0.306)	0.989 (0.983 - 0.994)	39 (33 - 45)	1124 (1104 - 1144)	113 (93 - 133)	13 (7 - 19)
Center 2	≤ 3	0.660 (0.563 - 0.748)	0.897 (0.885 - 0.908)	0.190 (0.163 - 0.218)	0.986 (0.983 - 0.990)	68 (58 - 77)	2519 (2486 - 2549)	289 (259 - 322)	35 (26 - 45)
Center 3	≤ 3	0.766 (0.718 - 0.810)	0.887 (0.876 - 0.898)	0.413 (0.386 - 0.442)	0.973 (0.968 - 0.978)	258 (242 - 273)	2869 (2835 - 2905)	366 (330 - 400)	79 (64 - 95)
Center 1	≤ 4	0.654 (0.519 - 0.788)	0.951 (0.939 - 0.962)	0.359 (0.293 - 0.435)	0.985 (0.979 - 0.991)	34 (27 - 41)	1176 (1161 - 1190)	61 (47 - 76)	18 (11 - 25)
Center 2	≤ 4	0.534 (0.437 - 0.621)	0.952 (0.943 - 0.959)	0.287 (0.239 - 0.339)	0.982 (0.979 - 0.986)	55 (45 - 64)	2672 (2649 - 2693)	136 (115 - 159)	48 (39 - 58)
Center 3	≤ 4	0.665 (0.611 - 0.718)	0.930 (0.921 - 0.938)	0.497 (0.462 - 0.534)	0.964 (0.958 - 0.969)	224 (206 - 242)	3008 (2980 - 3036)	227 (199 - 255)	113 (95 - 131)

Table 4.3.5: Diagnostic performance of the CoLab-score in the three external centers. Sensitivities, specificities, positive predictive values (PPV), negative predictive values (NPV), true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) are shown with bootstrapped 95% confidence intervals in parentheses.

4.4 Discussion

Given the impact of coronavirus disease 2019 (COVID-19) on society and healthcare, there is a need for simple and fast detection of patients with a possible COVID-19 infection in the ED. The CoLab-score described in this study, is a fast and accurate risk score to triage patients presenting at the emergency department (ED) based on ten routine blood biomarkers and age. The main strength of this study is that this score can be used as an early-warning or triaging tool for the ED population presenting with abdominal pain, chest pain, shortness of breath, syncope, sepsis or other non-specific complaints where a routine blood panel is requested. This is in contrast to the vast majority of COVID-19 diagnostic models that have been developed on a pre-selected population of polymerase chain reaction (PCR)-tested patients [9, 20–26]. Moreover, the CoLab-score requires only routine blood tests, instead of (features from) imaging such as CT-scans or laboratory tests that are not routinely collected in the ED, e.g. interleukin-6 or 3-hydroxybuteric acid [4]. Compared to rapid lateral flow tests (LFTs), which provide a dichotomous result within 30 minutes and are widely adopted in EDs, the CoLab-score is a continuous score. The lowest CoLab-scores (0 - 1) offer higher sensitivity and are therefore more suitable to rule-out COVID-19 than a LFT, which are only moderately sensitive (albeit more specific) [27, 28]. Two other studies have been published which are similar to this study [9, 10]. Interestingly, the study by Soltan et al., ranked basophils and eosinophils as the two most important features in predicting the outcome, similar to our results [10]. Eosinophils were also seen as one of the most important features by Plante et al. [9]. However, both studies focus on an artificial intelligence/machine learning approach. While their approach likely results in higher predictive performance, due to the ability of machine learning models to capture non-linear and interaction effects, the goal of this study was to develop a simple, fast and robust model that can easily be implemented in current hospital IT systems. Since this is a retrospective case-control study, there is some unavoidable missing data. In our cohort 17.6% of the ED presentations could not be used due to one or more missing laboratory results. This is lower or equal to similar studies; 22% [23], 17% [21] and 11% [26]. Important to note is that 7.7% of missingness is due to analytical errors which can be assumed to be missing completely at random. For the remaining 9.9% of missingness, the full lab panel was most frequently missing for pediatric, obstetric and surgery patients. These patients are presenting with specific complaints for which specific laboratory tests are requested, and hence do not match the inclusion criteria for a routine blood panel. Overall the missingness was significantly lower in the PCR-tested group versus the untested group (χ^2-test p-value < 0.001). It is assumed that all presentations in the untested group are COVID-19 negative. However, some presentations with asymptomatic COVID-19 could be present in the untested control group. The impact of these 'false controls' is most likely small as other studies indicate that there is a very low positivity rate among asymptomatic ED presentations (only a few in over a thousand tested asymptomatic cases) [29, 30]. The vast majority of controls were not tested for COVID-19, because they were either pre-pandemic or untested patients (89% in the development dataset). Clinical data always contains some unavoidable 'noise' in the form of misregistrations, misdiagnoses or patients who were missed. We have tried to mitigate this by including a large pre-pandemic control group and including all PCR tests within 1 week after discharge. In the external centers, there is a high level of missingness as a result of an incomplete laboratory panel. In the case of centers 1 and 2, only internal medicine ED presentations were tested with a laboratory panel containing the 10 tests required for the CoLab-score. The ED lab panel of other disciplines (e.g. urology, surgery or pediatrics) differed and did not contain the required tests. Nevertheless, the majority of COVID-19 patients were internal medicine ED presentations, which is reflected by the few PCR-positive patients excluded. Due to these high levels of missingness, the results of the external centers cannot be used to show that the CoLab-score generalizes to the entire ED population. Rather, the results show that for the majority of COVID-19 positive patients presenting at the ED, a routine laboratory panel is available from which the CoLab-score can be calculated, and that the performance of the CoLab-score in this population is comparable to the development population. Differences in the distribution of CoLab variables between centers are shown in Section 4.C. The performance of the CoLab-score is affected by the time between the onset of symptoms and ED presentations. The score increases with the duration of symptoms and gradually decreases after day 7 (see Section 4.D for a plot of the duration of COVID-19 related symptoms and the

CoLab-linear predictor). As a consequence, some COVID-19 patients with early or late presentation after onset of symptoms can be missed. Optimal performance of the CoLab-score is achieved when the onset of symptoms is > 1 and < 10 days prior to ED presentation. Chemotherapy that causes myeloid suppression, will decrease neutrophilic, basophilic and eosinophilic counts and thereby “falsely” increasing the CoLab-score. Conversely, COVID-19 patients with severe anemia could have “falsely” lowered CoLab-scores. To minimize false negatives, we have therefore advised to report CoLab-scores only when the concentration of erythrocytes is ≥ 2.9 /pL. It was chosen to exclude re-presentations after a previous presentation with COVID-19. Since the median time between initial presentation and re-presentation was 12 days, these patients were most likely not re-infected patients, but patients who deteriorated after initial presentation/treatment. Given that the CoLab-score follows the host-immune response, the score is time sensitive (see Section 4.D). Including these patients would impact the performance of the CoLab-score as patients in a later phase of the disease show different biomarker profiles. The CoLab-score is aimed towards alerting clinicians to patients presenting with a novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection, rather than patients who deteriorate after treatment for COVID-19. Other re-presentations were not excluded, which results in some patients appearing multiple times in a dataset. This was not adjusted for in the regression model since the assumption was made that ED presentations are independent observations. The median time between re-presentations is 38 days, most likely resulting in variations in laboratory results between presentations, and hence, little to no correlation between presentations. A sensitivity analysis was performed whereby only the first presentation was included for each patient (Section 4.D), but no difference was found in performance in terms of sensitivity, specificity and area under the ROC-curve (AUC). The CoLab-score does not serve as a replacement for PCR-testing or LFTs, and can be used to guide PCR-testing when routine blood tests are available. Important to note is that the CoLab-score is only valid for ED presentations where routine blood testing is requested, and as a consequence does not generalize to the ED population who is otherwise well and does not undergo routine blood testing. Using the CoLab-score in a symptomatic/PCR-tested cohort also results in different diagnostic performance characteristics, as compared to using the score on the full ED cohort (see Section 4.D). Finally, the CoLab-score could lead to false positives by other viral infections. However, in an historic patient cohort, the CoLab-score had only limited discriminative ability in separating influenza-PCR-negative from influenza-PCR-positive patients (see Section 4.D) implying specificity for SARS-CoV-2. Since the CoLab-score reflects the host-response to the virus, it is hypothesized that the CoLab-score could also be sensitive to future SARS-CoV-2 variants. This is supported by the fact that the discriminative ability is sustained in periods with different dominant variants, although the sensitivity of scores ≥ 3 is somewhat lower in the third phase (see Section 4.B). Although vaccination status is not registered for all presenting patients, in a small subgroup of 12 patients for whom vaccination status was registered, and were COVID-19 positive, 8 of 12 patients had the highest CoLab-score (= 5) (see Section 4.B). Continuous assessment of the performance of the CoLab-score is required due to the emergence of new variants and changes in the host's immune response. To conclude, the CoLab-score developed and validated in this study, based on 10 routine laboratory results and age, is available within 1 hour for any patient presenting at the ED where routine blood testing is requested. The score can be used by clinicians to guide PCR testing or triage patients and helps to identify COVID-19 in patients presenting at the ED with abdominal pain, chest pain, shortness of breath, syncope, sepsis or other non-specific complaints where a routine blood panel is requested. The lowest CoLab-score can be used to effectively rule-out a possible SARS-CoV-2 infection, the highest score to alert physicians to a possible infection. The CoLab-score is therefore a valuable tool to rule out COVID-19, guide PCR testing and is available to any center with access to routine laboratory tests.

References

1. *Coronavirus Disease (COVID-19) Situation Reports*

2. Guan, W.-j. *et al.* Clinical Characteristics of Coronavirus Disease 2019 in China. *New England Journal of Medicine* **382,** 1708–1720 (Apr. 2020).

3. Vetter, P. *et al.* Clinical features of covid-19. *BMJ* **369** (Apr. 2020).

4. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *The BMJ* **369,** 18 (Apr. 2020).

5. Albahri, A. S. *et al. Role of biological Data Mining and Machine Learning Techniques in Detecting and Diagnosing the Novel Coronavirus (COVID-19): A Systematic Review* July 2020.

6. Hooli, S. & King, C. Generalizability of Coronavirus Disease 2019 (COVID-19) Clinical Prediction Models. *Clinical Infectious Diseases* **71,** 897–897 (July 2020).

7. Prokop, M. *et al.* CO-RADS: A Categorical CT Assessment Scheme for Patients Suspected of Having COVID-19-Definition and Evaluation. *Radiology* **296,** E97–E104 (Aug. 2020).

8. Goodacre, S. *et al.* Derivation and validation of a clinical severity score for acutely ill adults with suspected COVID-19: The PRIEST observational cohort study. *PLOS ONE* **16,** e0245840 (Jan. 2021).

9. Plante, T. B. *et al.* Development and external validation of a machine learning tool to rule out COVID-19 among adults in the emergency department using routine blood tests: A large, multicenter, real-world study. *Journal of Medical Internet Research* **22,** e24048 (Dec. 2020).

10. Soltan, A. A. *et al.* Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test. *The Lancet Digital Health* **3,** e78–e87 (Feb. 2021).

11. R Core Team. *R: A Language and Environment for Statistical Computing* Vienna, Austria, 2020.

12. Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429 (Dec. 2006).

13. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288 (Jan. 1996).

14. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* **33,** 1–22 (Feb. 2010).

15. King, G. & Zeng, L. Logistic Regression in Rare Events Data. *Political Analysis* **9,** 137–163 (2001).

16. Efron, B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association* **78,** 316–331 (1983).

17. Robin, X. *et al.* pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12,** 77 (Mar. 2011).

18. Harrell Jr, F. E. *rms: Regression Modeling Strategies* 2021.

19. Ricós, C. *et al. Current databases on biological variation: Pros, cons and progress* 1999.

20. Brinati, D. *et al.* Detection of COVID-19 Infection from Routine Blood Exams with Machine Learning: A Feasibility Study. *Journal of Medical Systems* **44,** 1–12 (Aug. 2020).

21. Joshi, R. P. *et al.* A predictive tool for identification of SARS-CoV-2 PCR-negative emergency department patients using routine test results. *Journal of Clinical Virology* **129,** 104502 (Aug. 2020).

22. Qin, L. *et al.* A predictive model and scoring system combining clinical and CT characteristics for the diagnosis of COVID-19. *European Radiology* **30,** 6797–6807 (Dec. 2020).

23. Kurstjens, S. *et al.* Rapid identification of SARS-CoV-2-infected patients at the emergency department using routine testing. *Clinical Chemistry and Laboratory Medicine* **58,** 1587–1593 (Aug. 2020).

24. Fink, D. L. *et al.* Development and internal validation of a diagnostic prediction model for COVID-19 at time of admission to hospital. *QJM: An International Journal of Medicine* (Nov. 2020).

25. Giamello, J. D. *et al.* A simple tool to help ruling-out Covid-19 in the emergency department: derivation and validation of the LDH-CRP-Lymphocyte (LCL) score. *Emergency Care Journal* **16** (Dec. 2020).

26. Tordjman, M. *et al.* Pre-test probability for SARS-Cov-2-related infection score: The PARIS score. *PLOS ONE* **15** (ed Moreira, J.) e0243342 (Dec. 2020).

27. Peto, T. *et al.* COVID-19: Rapid antigen detection for SARS-CoV-2 by lateral flow assay: A national systematic evaluation of sensitivity and specificity for mass-testing. *EClinicalMedicine* **36,** 100924 (June 2021).

28. Garciá-Fiñana, M. *et al.* Performance of the Innova SARS-CoV-2 antigen rapid lateral flow test in the Liverpool asymptomatic testing pilot: population based cohort study. *BMJ* **374,** 1637 (July 2021).

29. Ford, J. S. *et al. Testing Asymptomatic Emergency Department Patients for Coronavirus Disease 2019 (COVID-19) in a Low-prevalence Region* Aug. 2020.

30. Ravani, P. *et al.* COVID-19 screening of asymptomatic patients admitted through emergency departments in Alberta: a prospective quality-improvement study. *CMAJ Open* **8,** E887–E894 (Oct. 2020).

Appendix

4.A Model fitting

In adaptive lasso, weights are applied to each of the covariates present in the lasso constraint, the weight vector has to be calculated before the adaptive lasso regression is performed. Prior to model fitting, covariates were scaled to zero mean and unit variance, after model fitting coefficients were unscaled to obtain regression coefficients on the original scale. Due to multicollinearity between laboratory tests in the routine lab panel, weights in the adaptive lasso were based on ridge regression estimates ($\hat{\beta}_{ridge}$) as recommended by Zou. To obtain $\hat{\beta}_{ridge}$ the optimal penalty (λ) for the ridge regression was chosen using 10 fold cross-validation (CV) with the area under the ROC-curve (AUC) as the loss function. The λ corresponding to the maximum AUC was selected to obtain $\hat{\beta}_{ridge}$. The weight vector ($\hat{w}$) was calculated by $\hat{w} = \frac{1}{|\hat{\beta}_{ridge}|^2}$. This weight vector was then used to fit an adaptive lasso regression where λ was chosen by the criterion ±1 SE of the maximum AUC.

4.A.1 Model intercept adjustment

The linear predictor (LP) for a patient i is calculated as follows: $LP_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_n x_{in}$, where n is the number of covariates in the final model, x_{in} are the observed values for the i-th patient of the n covariates and β_n the estimated coefficients. The LP can then be converted to a probability for patient i

(P_i) by the logistic function: $P_i = \frac{1}{1+e^{-LP_i}}$. The intercept term β_0 is sensitive to the fraction of cases versus controls in the dataset/population. Since the model is fitted to a case-control dataset where the number cases is fixed (all patients tested positive for coronavirus disease 2019 (COVID-19)) and the number of controls is randomly chosen (a 6-month period pre-COVID), the intercept term β_0 is a result of this choice and will likely not be generalizable to the real-world setting. In fact, the prevalence of COVID-19 will vary over time. Prior correction is a method to correct the estimate of the intercept based on the true fraction of positives in the population, τ (prevalence of COVID-19 in the emergency department (ED)) and the fraction of cases in the development dataset, $\bar{y}$. The intercept term β_0 can then be corrected to obtain $\beta_{0,corrected}$ using the following formula:

$$\begin{aligned} \beta_{0,corrected} &= \beta_0 + \beta_{adj} \\ \beta_{adj} &= -\ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right] \end{aligned} \tag{4.A.1}$$

For the development dataset $\hat{y} = 0.02675$ therefore:

$$\beta_{adj} = -\ln\left(\frac{1-\tau}{\tau}\right) + 3.594 \tag{4.A.2}$$

Since the true prevalence τ is unknown, we can use an estimate $\hat{\tau}$ and replace β_0 in the LP for $\beta_{0,corrected}$ to obtained predictions that are calibrated according to the estimated prevalence:

$$LP_{i|\hat{\tau}} = \beta_0 - \ln\left(\frac{1-\hat{\tau}}{\hat{\tau}}\right) + 3.594 + \beta_1 x_{i1} + \ldots + \beta_n x_{in} \tag{4.A.3}$$

4.A.2 CoLab-score

An alternative, which is the basis of the CoLab-score, is to choose a threshold for the probability P_i above which one considers a patient eligible for further testing. The probability can be expressed as a number needed to test. If one is willing to test 10 patients to find one positive, all patients with $P_i \geq 0.1$ should be considered positive. In this study a number needed to test of 15 is used, therefore all patients with a $P_i \geq 0.067$ should be considered positive. On the LP scale, this translates to a threshold of $\text{logit}(0.067) = -2.639$. To determine the cutoffs for different prevalence thresholds, one simply solves the following equation:

$$\begin{aligned} \beta_0 + \beta_{adj} + \beta_1 x_{i1} + \ldots + \beta_n x_{in} &\geq -2.639 \\ \beta_0 + \beta_1 x_{i1} + \ldots + \beta_n x_{in} &\geq -2.639 - \beta_{adj} \\ LP_{i|\hat{\tau}} &\geq \ln\left(\frac{1-\hat{\tau}}{\hat{\tau}}\right) - 6.233 \end{aligned} \tag{4.A.4}$$

Choosing threshold values at $\hat{\tau} = 0.4, 0.1, 0.05, 0.02, 0.01$, yields the cutoffs for the CoLab score:

$$\begin{aligned} LP_{i|\hat{\tau}=0.4} &\geq -5.83 \quad \text{CoLab-score = 1} \\ LP_{i|\hat{\tau}=0.1} &\geq -4.03 \quad \text{CoLab-score = 2} \\ LP_{i|\hat{\tau}=0.05} &\geq -3.29 \quad \text{CoLab-score = 3} \\ LP_{i|\hat{\tau}=0.02} &\geq -2.34 \quad \text{CoLab-score = 4} \\ LP_{i|\hat{\tau}=0.01} &\geq -1.64 \quad \text{CoLab-score = 5} \end{aligned} \tag{4.A.5}$$

These thresholds correspond to CoLab-scores 0 to 5. The interpretation of these scores is as follows; if the prevalence is < 1%, only CoLab-score 5 should be classified as positive and CoLab-score 0 till 4 as negative. If the prevalence is 1% – 2%, CoLab-score 4 and 5 should be classified as positive and 1 – 3 negative. Similarly, with a prevalence of 2 – 5% the split is between CoLab-score 2 and 3 and with prevalence of 5 – 10% between CoLab-score 1 – 2. If the prevalence is higher than 10% only CoLab-score 0 is classified as negative. Using the CoLab-score in this fashion, aims to preserve a number need to test of 15.

4.B Temporal validation details

4.B.1 Vaccination status and novel strains

The temporal validation dataset consists of emergency department (ED) presentations from July 2020 until October 2021. As stated in Section 4.2, this period was split into three phases: i) from July 2020 until March 2021, no vaccination and no variants of concern identified ii) from March 2021 until June 2021, partial vaccination and B.1.1.7 (α) variant identified as dominant iii) from June 2021 until October 2021, widespread vaccination and B.1.617.2 (δ) variant identified as dominant. The ED fraction vaccinated is estimated by merging data from the Dutch national institute of public health by the date of the ED presentation and the year of birth of the patient. See Fig. 4.B.1.

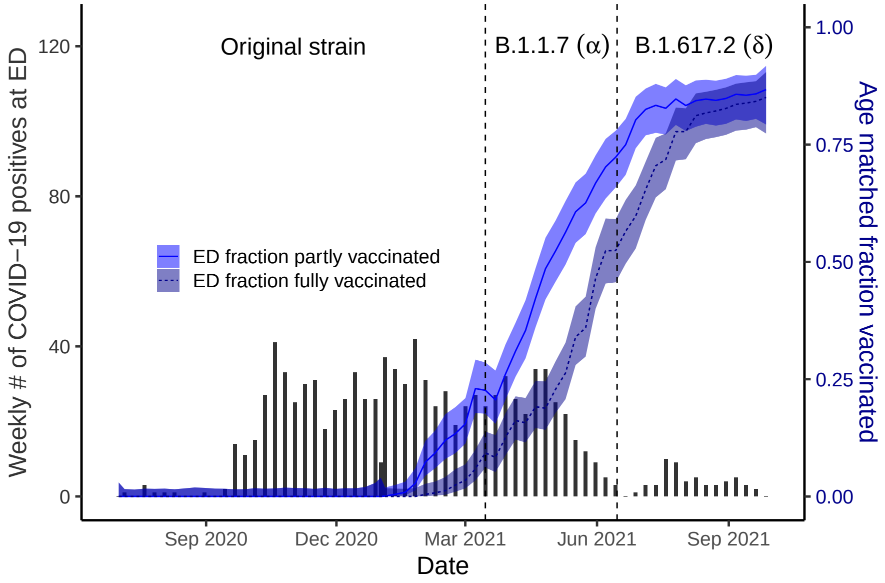


Figure 4.B.1: Temporal validation period split into three phases characterized by weekly number of new COVID-19 cases at the ED and estimated fraction of ED patients vaccinated. The gray bars depict weekly number of new COVID-19 cases at the ED, the blue lines the estimated fraction of ED patients fully or partially vaccinated. The shading depict the 95% Wilson confidence intervals.

4.B.2 CoLab-score performance

In Table 4.B.1 the area under the ROC-curve (AUC) of the CoLab-score is shown for the three different periods, defined in the previous section. In Table 4.B.2 the performance in terms of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) is shown for the three different periods. Finally in Fig. 4.B.2 a boxplot shows the CoLab-linear predictor for the patients with a registered vaccination status.

Phase	Cases/controls (prevalence)	AUC
Original strain & no vaccinations	694/7999 (8.6%)	0.909 (0.896 - 0.923)
α strain & partial vaccination	287/2845 (10.1%)	0.937 (0.921 - 0.953)
δ strain & full vaccination	58/3236 (1.8%)	0.898 (0.857 - 0.939)

Table 4.B.1: AUC with 95% CI over the different time windows defined in Fig. 4.B.1.

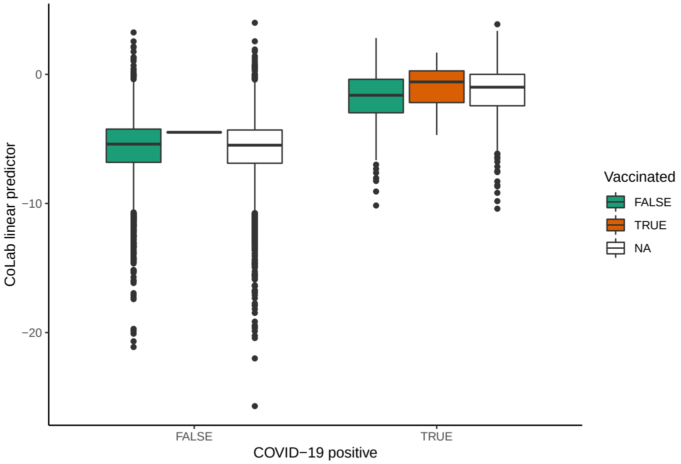


Figure 4.B.2: Boxplots of CoLab linear predictor versus COVID-19 positive, split by registered vaccination status.
The CoLab linear predictor is calculated for all ED presentations in the temporal validation set. Presentations who are registered as vaccinated are labeled TRUE (N = 13). Presentations before vaccine roll-out are labeled FALSE (N = 5855). Presentations during vaccine roll-out but where no status is registered are labeled NA (N = 8212). Of the 13 presentations who were registered as vaccinated, 12 were COVID-19 positive and 1 negative. Note that vaccination status is only registered if a patient is SARS-CoV-2 PCR positive or considered positive until proven otherwise, therefore there is only one COVID-19 negative patient with a registered vaccination status.

CoLab-score	Period	Sensitivity	Specificity	PPV	NPV
0	Original strain no vaccinations	0.960 (0.945 - 0.974)	0.418 (0.407 - 0.429)	0.135 (0.133 - 0.138)	0.991 (0.988 - 0.994)
0	α strain partial vaccination	0.983 (0.965 - 0.997)	0.431 (0.412 - 0.451)	0.162 (0.157 - 0.168)	0.996 (0.991 - 0.999)
0	δ strain full vaccination	0.983 (0.948 - 1.000)	0.415 (0.397 - 0.432)	0.030 (0.028 - 0.031)	0.999 (0.998 - 1.000)
≤ 1	Original strain no vaccinations	0.879 (0.854 - 0.902)	0.789 (0.779 - 0.798)	0.283 (0.272 - 0.294)	0.986 (0.983 - 0.988)
≤ 1	α strain partial vaccination	0.916 (0.885 - 0.948)	0.808 (0.793 - 0.823)	0.349 (0.330 - 0.370)	0.989 (0.984 - 0.993)
≤ 1	δ strain full vaccination	0.862 (0.759 - 0.948)	0.780 (0.766 - 0.793)	0.067 (0.059 - 0.074)	0.997 (0.994 - 0.999)
≤ 2	Original strain no vaccinations	0.813 (0.784 - 0.841)	0.894 (0.887 - 0.901)	0.422 (0.403 - 0.440)	0.981 (0.978 - 0.983)
≤ 2	α strain partial vaccination	0.864 (0.822 - 0.902)	0.896 (0.885 - 0.908)	0.484 (0.455 - 0.517)	0.983 (0.978 - 0.988)
≤ 2	δ strain full vaccination	0.690 (0.569 - 0.810)	0.892 (0.881 - 0.903)	0.104 (0.086 - 0.124)	0.994 (0.991 - 0.996)
≤ 3	Original strain no vaccinations	0.699 (0.664 - 0.732)	0.962 (0.957 - 0.966)	0.634 (0.607 - 0.664)	0.971 (0.968 - 0.974)
≤ 3	α strain partial vaccination	0.760 (0.711 - 0.808)	0.962 (0.955 - 0.970)	0.695 (0.651 - 0.741)	0.973 (0.967 - 0.978)
≤ 3	δ strain full vaccination	0.621 (0.500 - 0.741)	0.960 (0.953 - 0.967)	0.223 (0.177 - 0.271)	0.993 (0.991 - 0.995)
≤ 4	Original strain no vaccinations	0.566 (0.530 - 0.602)	0.984 (0.981 - 0.987)	0.775 (0.739 - 0.808)	0.960 (0.957 - 0.963)
≤ 4	α strain partial vaccination	0.645 (0.589 - 0.700)	0.983 (0.977 - 0.988)	0.808 (0.759 - 0.855)	0.961 (0.955 - 0.967)
≤ 4	δ strain full vaccination	0.517 (0.397 - 0.638)	0.986 (0.982 - 0.990)	0.400 (0.312 - 0.493)	0.991 (0.989 - 0.993)

Table 4.B.2: Diagnostic performance of the CoLab-score in the temporal validation dataset, split by phase.
Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) are shown for fixed cut-offs (CoLab-score 0 till ≤ 4) with bootstrapped 95% confidence intervals in parentheses. The temporal validation dataset is split into three phases according to dominant SARS-CoV-2 strains in the Netherlands and estimated fraction of ED patients vaccinated (see Fig. 4.B.1). Note that "0" lists the sensitivity and NPV of CoLab-score 0 and "≤ 4" lists the specificity and PPV of CoLab-score 5.

4.C Model calibration

In the calibration plots in Fig. 4.C.1, the proportion of observed coronavirus disease 2019 (COVID-19) positives versus expected probabilities are plotted. Observations are grouped with an average of 150 observations per group. The expected probabilities follow from applying the inverse logit function to the CoLab-linear predictor calculated from Table 4.3.2. If the observed proportion in an external dataset is lower than the expected proportion, this means risks are over-estimated, if the observed fraction is higher, risks are under-estimated. Ideally, observed proportions are equal to expected proportions, this ideal-calibration-line is shown as a straight line through the origin with a slope of 1. The logistic calibration line is a logistic regression fit of the predicted probabilities. [Intercept, slope]: Temporal [1.34, 1.08], Center 1 [-0.39, 0.92], Center 2 [-0.76, 0.77], Center 3 [0.08, 0.79]. Although no validation datasets show perfect calibration, this is the result of differences in COVID-19 prevalence in the temporal validation dataset (7.4% versus 2.2%) and differences in calibration of laboratory equipment in the three external centers. Probability density plots are shown for all control patients of the development dataset and the three external centers in Fig. 4.C.2. Ideally all distributions should overlap since this implies that control patient populations are most likely similar in the development dataset to the external datasets. When comparing the distribution of the CoLab variables for all control-patients across different external validation datasets, albumin and lactate dehydrogenase (LD) show the largest deviations.

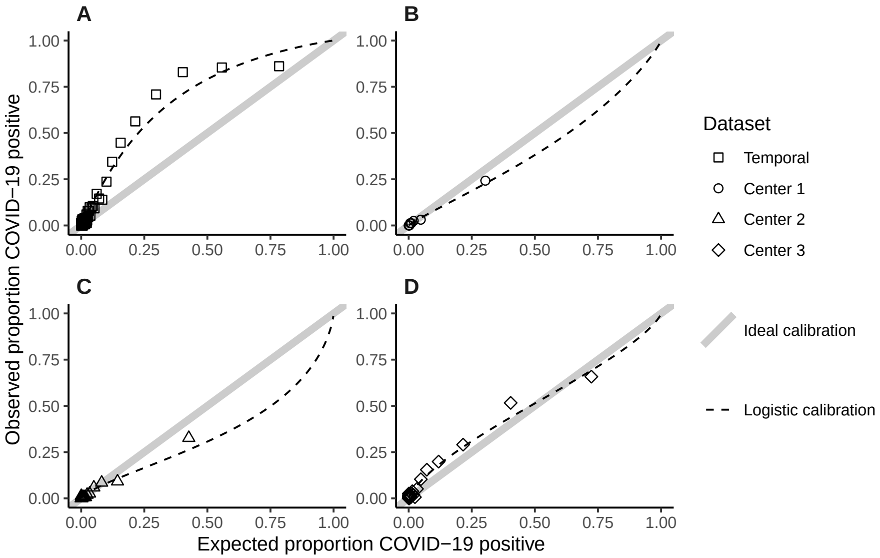


Figure 4.C.1: CoLab-score calibration plots of the temporal validation (A), external validation center 1 (B), external validation center 2 (C) and external validation center 3 (D).

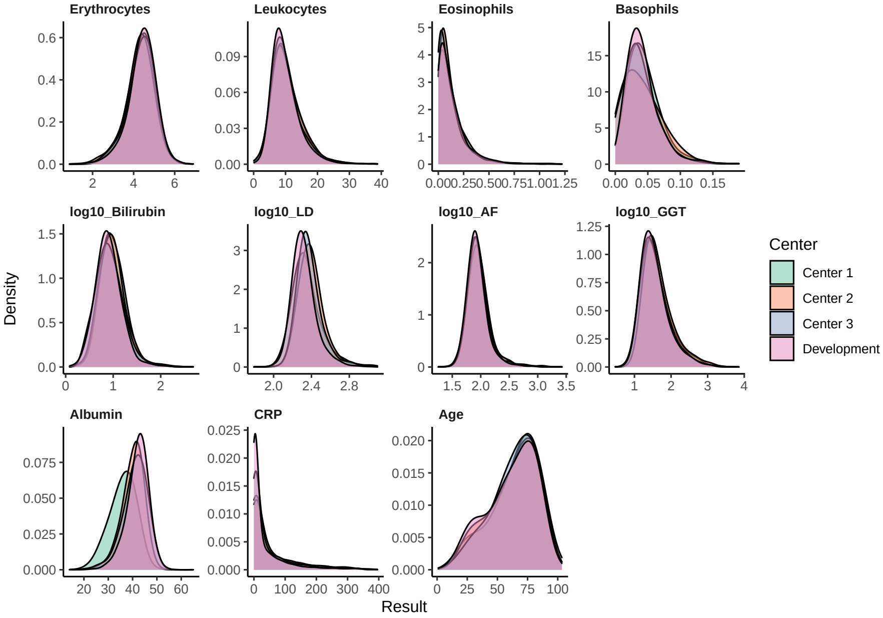


Figure 4.C.2: Probability density plots of laboratory parameters.

4.D Further validation

4.D.1 CoLab-score versus duration of symptoms

For all polymerase chain reaction (PCR)-positive emergency department (ED) presentations in the development and temporal validation dataset, the CoLab-linear predict is plotted in Fig. 4.D.1 against the duration of COVID-related symptoms as registered in the electronic patient records. Patients with unknown duration are not plotted. Patients without symptoms were plotted at 0 days.

4.D.2 CoLab-score versus RS-, Rhino- and Influenza-virus

For 183 ED presentations that were PCR tested for either RS-, Rhino- and Influenza-virus the CoLab-score was calculated. 91 presentations were PCR positive, 92 were PCR negative. The CoLab-score is only marginally elevated for PCR positive patients, the area under the ROC-curve (AUC) in separating both groups is 0.573 (95% CI: 0.4896-0.6563). See Fig. 4.D.2.

4.D.3 Sensitivity analysis

As a sensitivity analysis, the temporal validation dataset is used to compare the performance of the CoLab-score with inclusion criteria that differ from the development dataset. First, we examine the performance of the temporal validation dataset with the original inclusion criteria as specified in Section 4.2. Second, we examine the performance of the CoLab-score when all re-presentations are excluded (i.e. no repeated presentations). Thirdly, we examine the performance of the CoLab-score in the subgroup of patients that underwent PCR-testing. The AUC for the three different scenarios are given in Table 4.D.1, the performance in terms of sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) are given in Table 4.D.2.

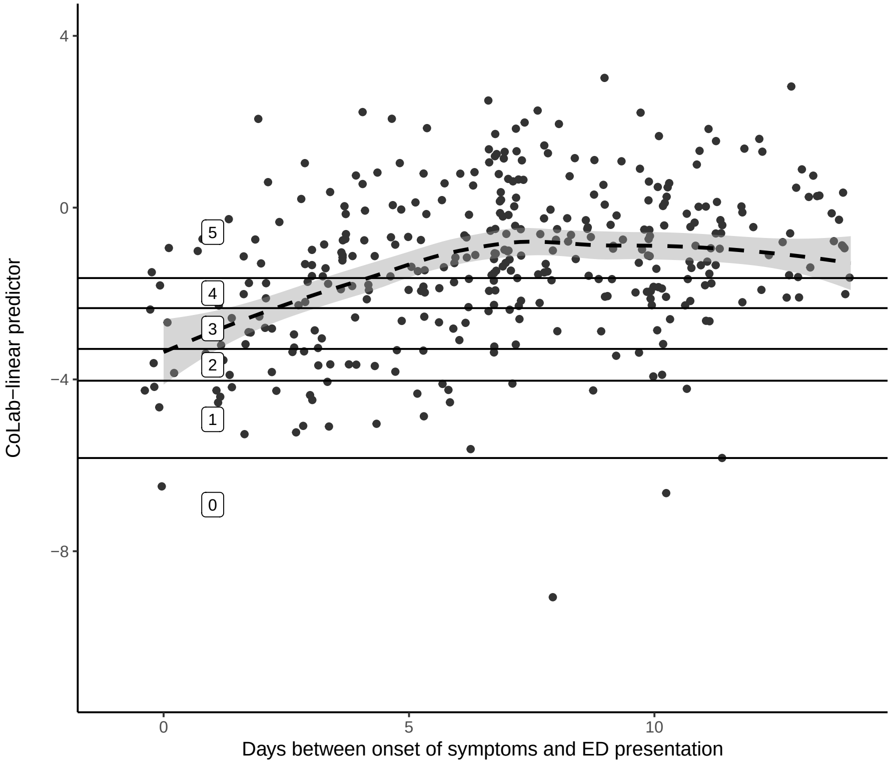


Figure 4.D.1: Association between the CoLab-linear predictor and the duration of COVID-19-related symptoms.
The solid horizontal lines represent the CoLab-score thresholds, the dashed line is a LOESS regression curve with 95% CI. As the duration of symptoms is an integer, some random jitter was added to the days, for visualization purposes. Note that only the first 14 days are shown in this graph.

Inclusion criterion	Cases/controls (prevalence)	AUC
Temporal validation (reference)	1039/14080 (7.4%)	0.916 (0.906 - 0.927)
Only first presentations, re-presentations are excluded	937/11166 (8.4%)	0.919 (0.909 - 0.930)
Only PCR-tested presentations	372/4062 (9.2%)	0.840 (0.817 - 0.862)

Table 4.D.1: AUC with 95% CI over using different inclusion criteria for the temporal validation set.

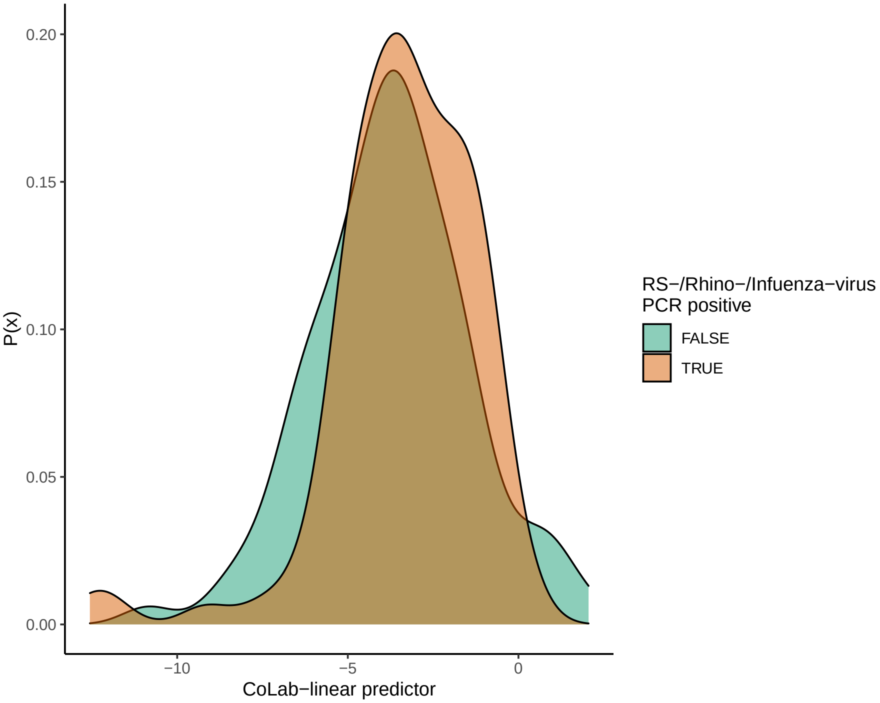


Figure 4.D.2: Probability density plot of CoLab-score for RS-, Rhino- and Influenza-virus PCR tested ED patients.

For 183 ED presentations that were PCR tested for either RS-, Rhino- and Influenza-virus the CoLab-score was calculated. 91 presentations were PCR positive, 92 were PCR negative.

CoLab-score	Validation set	Sensitivity	Specificity	PPV	NPV	TP	TN	FP	FN
0	Reference	0.967 (0.956 - 0.978)	0.420 (0.411 - 0.428)	0.117 (0.115 - 0.119)	0.994 (0.992 - 0.996)	1005 (993 - 1016)	5476 (5366 - 5587)	7565 (7454 - 7675)	34 (23 - 46)
0	Re-presentations excluded	0.968 (0.956 - 0.979)	0.416 (0.406 - 0.426)	0.132 (0.130 - 0.134)	0.993 (0.990 - 0.995)	907 (896 - 917)	4259 (4156 - 4353)	5970 (5876 - 6073)	30 (20 - 41)
0	Only PCR-tested presentations	0.946 (0.922 - 0.968)	0.353 (0.338 - 0.368)	0.129 (0.125 - 0.132)	0.985 (0.979 - 0.991)	352 (343 - 360)	1303 (1246 - 1359)	2387 (2331 - 2444)	20 (12 - 29)
≤ 1	Reference	0.888 (0.870 - 0.908)	0.791 (0.783 - 0.798)	0.253 (0.245 - 0.261)	0.989 (0.987 - 0.991)	923 (904 - 943)	10311 (10215 - 10401)	2730 (2640 - 2826)	116 (96 - 135)
≤ 1	Re-presentations excluded	0.890 (0.870 - 0.908)	0.793 (0.785 - 0.801)	0.282 (0.273 - 0.292)	0.987 (0.985 - 0.990)	834 (815 - 851)	8112 (8030 - 8194)	2117 (2035 - 2199)	103 (86 - 122)
≤ 1	Only PCR-tested presentations	0.852 (0.817 - 0.887)	0.671 (0.656 - 0.686)	0.207 (0.197 - 0.217)	0.978 (0.973 - 0.983)	317 (304 - 330)	2477 (2421 - 2533)	1213 (1157 - 1269)	55 (42 - 68)
≤ 2	Reference	0.820 (0.796 - 0.843)	0.894 (0.889 - 0.899)	0.382 (0.367 - 0.396)	0.984 (0.982 - 0.986)	852 (827 - 876)	11661 (11591 - 11729)	1380 (1312 - 1450)	187 (163 - 212)
≤ 2	Re-presentations excluded	0.824 (0.798 - 0.845)	0.898 (0.892 - 0.904)	0.426 (0.410 - 0.441)	0.982 (0.980 - 0.985)	772 (748 - 792)	9187 (9127 - 9249)	1042 (980 - 1102)	165 (145 - 189)
≤ 2	Only PCR-tested presentations	0.734 (0.688 - 0.777)	0.800 (0.786 - 0.812)	0.270 (0.252 - 0.287)	0.968 (0.962 - 0.973)	273 (256 - 289)	2951 (2902 - 2997)	739 (693 - 788)	99 (83 - 116)
≤ 3	Reference	0.710 (0.682 - 0.738)	0.962 (0.958 - 0.965)	0.596 (0.573 - 0.618)	0.977 (0.974 - 0.979)	738 (709 - 767)	12540 (12496 - 12582)	501 (459 - 545)	301 (272 - 330)
≤ 3	Re-presentations excluded	0.716 (0.687 - 0.744)	0.966 (0.962 - 0.969)	0.658 (0.633 - 0.682)	0.974 (0.971 - 0.976)	671 (644 - 697)	9880 (9844 - 9915)	349 (314 - 385)	266 (240 - 293)
≤ 3	Only PCR-tested presentations	0.591 (0.540 - 0.640)	0.911 (0.902 - 0.921)	0.403 (0.370 - 0.433)	0.957 (0.952 - 0.962)	220 (201 - 238)	3363 (3328 - 3397)	327 (293 - 362)	152 (134 - 171)
≤ 4	Reference	0.585 (0.556 - 0.615)	0.984 (0.982 - 0.987)	0.750 (0.724 - 0.778)	0.968 (0.965 - 0.970)	608 (578 - 639)	12838 (12811 - 12866)	203 (175 - 230)	431 (400 - 461)
≤ 4	Re-presentations excluded	0.590 (0.558 - 0.621)	0.987 (0.985 - 0.989)	0.805 (0.776 - 0.832)	0.963 (0.961 - 0.966)	553 (523 - 582)	10095 (10071 - 10117)	134 (112 - 158)	384 (355 - 414)
≤ 4	Only PCR-tested presentations	0.452 (0.401 - 0.503)	0.959 (0.953 - 0.965)	0.526 (0.480 - 0.575)	0.945 (0.941 - 0.950)	168 (149 - 187)	3539 (3516 - 3562)	151 (128 - 174)	204 (185 - 223)

Table 4.D.2: Sensitivity analysis of the CoLab-score in the temporal validation dataset using different inclusion criteria. ”Reference” represents the temporal validation dataset with the original inclusion criteria. ”Re-presentations excluded” refers to the performance of the CoLab-score when all re-presentations are excluded (i.e. no repeated presentations). ”Only PCR-tested presentations” refers to the subgroup of patients that underwent PCR-testing.

5

Use of an algorithm based on routine blood laboratory tests to exclude COVID-19 in a screening-setting of healthcare workers

Mathie PG Leers*, **Ruben Deneer***, Guy JM Mostard, Remy LM Mostard, Arjen-Kars Boer, Volkher Scharnhorst, Frans Stals, Henne A Kleinveld & Dirk W van Dam

* both authors contributed equally

Adapted from: Use of an algorithm based on routine blood laboratory tests to exclude COVID-19 in a screening-setting of healthcare workers. *PLoS One* **17,** e0270548 (2022).

Abstract

Background Coronavirus disease 2019 (COVID-19) is an ongoing pandemic leading to exhaustion of the hospital care system. Our health care system has to deal with a high level of absenteeism of healthcare workers (HCWs) with COVID-19 related complaints, in whom an infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has to be ruled out before they can return back to work. The aim of the present study is to investigate if the recently described CoLab-algorithm can be used to exclude COVID-19 in a screening setting of HCWs.

Methods In the period from January 2021 till March 2021, HCWs with COVID-19-related complaints were prospectively enrolled in this study. Next to the routinely performed SARS-CoV-2 polymerase chain reaction (PCR), using a set of naso- and oropharyngeal swab samples, two blood tubes (one EDTA- and one heparin-tube) were drawn for analysing the 10 laboratory parameters required for the CoLab-score.

Results In total, 726 HCWs with a complete CoLab-laboratory panel were included in this study. In this group, 684 HCWs were tested SARS-CoV-2 PCR negative and 42 cases PCR positive. ROC curve analysis showed an area under the ROC-curve (AUC) of 0.853 (95% CI: 0.801-0.904). At a safe cut-off value for excluding COVID-19 of -6.525, the sensitivity was 100% with a specificity of 34% (95% CI: 21 to 49%). No SARS-CoV-2 PCR cases were missed with this cut-off and COVID-19 could be safely ruled out in more than one third of HCWs.

Conclusions The CoLab-score is an easy and reliable algorithm that can be used for screening HCWs with COVID-19 related complaints. A major advantage of this approach is that the results of the score are available within 1 hour after collecting the samples. This results in a faster return to labour process of a large part of the COVID-19 negative HCWs (34%), next to a reduction in PCR tests (reagents and labour costs) that can be saved.

5.1 Introduction

Coronavirus disease 2019 (COVID-19) is an ongoing pandemic with at present over 150 million of cases and over three million deaths worldwide [1]. The initial clinical symptoms for COVID-19 are nonspecific and similar to other seasonal viral diseases, which encompass fever, dyspnoea, dry cough and fatigue. Many countries, including the Netherlands, are struggling to control COVID-19 outbreaks, especially in the detection of silent infections in the pre- or asymptomatic patient that can contribute to transmission [2]. Empirical studies have indicated that individuals may be highly infectious during the presymptomatic phase [3]. Healthcare workers (HCWs) potentially experience greater risks for emerging infectious diseases [4, 5] due to occupational exposure to sick patients and virus-contaminated surfaces [6]. Contagious HCWs may infect patients, co-workers and family members. However, the absenteeism of ill HCWs from duty can threaten essential healthcare staffing during an epidemic [7]. Therefore, infection prevention and quick, accurate diagnosis of potential COVID-19 in HCWs are crucial to maintain hospital operations [8]. Consequently, understanding the prevalence of, and factors associated with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection among frontline HCWs who care for COVID-19 patients are important to protect both HCWs and their patients. Next to this, modelling analyses show that rapid case identification of infected persons is critical to interrupt transmission, especially for infectious cases without clinical symptoms [2]. At the moment, the health care system has to deal with a high level of absenteeism of HCWs with COVID-19 related complaints, and in whom an infection with SARS-CoV-2 has to be ruled out before they can return back to work. Polymerase chain reaction (PCR) based methodologies are the gold standard in confirming that the individual presenting with COVID-19 has active viral shedding of SARS-CoV-2 [9]. However, there are some important limitations to PCR. First, current techniques take up to 6-8 hours in order to obtain first results. Next to this, laboratories often cannot handle the overload of tests. A third important limitation is that PCR on a nasopharyngeal swab, may be false negative in the initial phase of the disease, in spite of the presence of typical symptoms [10–12]. In addition, the standard test used has an 80% accuracy (compared to chest CT scan results) [12], which may depend on the specific level of viral shedding by any individual at the time of sample test. Fourth, the PCR technique carries a certain cost, which could mean a considerable financial burden [13]. Recently, a scoring algorithm was developed by Boer et al., called the 'CoLab'-score [14]. The score is calculated using 10 numeric values of routine-laboratory parameters next to the age of the patient. The linear predictor of the CoLab-score is continuous, therefore a cut-off can be chosen such that a high sensitivity and high negative predictive value can be achieved. This algorithm was developed and validated to exclude COVID-19 in patients presenting at the emergency department (ED). The aim of the present study was to investigate if the CoLab-score could be used to exclude within one hour COVID-19 in a screening setting of healthcare workers, who requested a SARS-CoV-2 PCR test because of COVID-19 related complaints, or because they were in close proximity to a SARS-CoV-2 infected person.

5.2 Materials and Methods

5.2.1 Study design and inclusion of healthcare workers

We conducted a prospective screening study to assess the comparability between naso-/oropharyngeal swabs and the CoLab-score (based on routine blood tests). Healthcare workers (HCWs) were included during the period from January 2021 till March 2021 either:

- because of coronavirus disease 2019 (COVID-19) related complaints or
- because they were in close proximity to a person with COVID-19.

HCWs were required to have complete data on clinical chemistry and hematologic parameters, needed to calculate the CoLab-score. From the validation study [14] it is known that there are some external factors influencing the predictive value of the score. For this reason, the following HCWs were excluded from this study. HCWs with:

- more than 10 days complaints at the time of screening

- a known positive severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) RT PCR in the past 4 weeks

- invalid polymerase chain reaction (PCR) test results due to contamination

Next to this, the data of the following HCWs were also excluded because of known interference with the algorithm (see Table 4.3.2):

- a deep anemia (hemoglobin $<$ 5.5 mmol/L)

- extreme laboratory values ($>$10 times standard deviation (SD)) in one or more of the Colab-values, see Table 4.3.2)

Using a standard protocol, paired naso-/oropharyngeal swabs from HCWs were collected using sterile flocked E-swabs and placed both in one sterile tube containing viral transport medium. Next to this, blood was collected in heparin- and EDTA-anticoagulated blood containers. The samples were transported to the central laboratory and immediately prepared for analysis.

5.2.2 Laboratory measurements

For clinical chemistry and hemocytometric analyses, heparin- and EDTA-anticoagulated venous blood samples respectively were collected. All analyses were performed at presentation. Clinical chemistry parameters C-reactive protein (CRP), albumin, total bilirubin, alkaline phosphatase (ALP), gamma-glutamyltransferase (gGT) and lactate dehydrogenase (LD) were obtained on routine chemistry analysers from Roche (Cobas; Roche Dx, Basel, Switzerland). The hemocytometric parameters (leukocytes, erythrocytes, eosinophilic and basophilic granulocytes) were derived from a complete blood count measured on a XN-1000 (Sysmex, Kobe, Japan). The nasopharyngeal and oropharyngeal swab samples were obtained for SARS-CoV-2 detection using multiplex Real-Time Polymerase Chain Reaction with QIAsymphony DSP Virus/-Pathogen Mini Detection Kit (Qiagen Inc). Both the binary outcome of the PCR (positive or negative), as well as the cycle threshold (Ct) value (in case of a positive PCR) were registered. In case of a negative PCR result in a HCW with persistent high suspicion for COVID-19 (e.g. suggestive symptoms without apparent alternative cause) the PCR test could be repeated after 48 hr of the initial PCR. The exclusion criterium of no more than 10 days complaints at the time of screening, nevertheless applies.

5.2.3 CoLab-score calculation

The Colab-score is described in detail in Chapter 4. In short, it is calculated by plugging the ten obtained laboratory measurands, next to the age of the HCW, into a formula: – 6.885 + [erythrocytes] × 0.9379 – [leukocytes] × 0.1298 – [eosinophils] × 6.834 – [basophils] × 47.7 – $\log_{10}$([bilirubin]) × 1.142 + $\log_{10}$([LD]) × 5.369 – $\log_{10}$([ALP]) × 3.114 + $\log_{10}$([gGT]) × 0.3605 – [albumin] × 0.1156 + [CRP] × 0.02560 + [age] × 0.002275. This results in a numeric value, called the CoLab-linear predictor (LP). This linear predictor can be converted to a score using the cut-offs depicted in Fig. 4.3.2.

5.2.4 Statistical analysis

Since the CoLab-score was developed to screen patients presenting at the emergency department (ED) for a possible COVID-19 infection, rather than exclude a SARS-CoV-2 infection in HCWs, the suitability for screening HCWs was investigated in this study. First, the discriminative ability of the CoLab-linear predictor was assessed by calculating the area under the ROC-curve (AUC). Secondly, model calibration was visually assessed with a calibration plot where the CoLab-linear predictor was converted to the predicted probability (through the inverse logit function) and the proportion of observed outcomes was plotted versus expected probabilities [15]. A logistic regression model was fitted to the CoLab predicted probabilities to assess model calibration in terms of intercept and slope [15]. This was done by plotting the proportion of observed COVID-19 positives versus expected probabilities. Ideally, observed proportions are equal to expected proportions, and this ideal-calibration line is shown as a straight line through the origin with a slope of 1. The logistic calibration line will be a logistic regression fit of the predicted probabilities. Using the intercept and/or slope from the logistic regression model, recalibrated probabilities were obtained and also plotted in a calibration plot. Thirdly, a cut-off for the CoLab-linear predictor was calculated to safely rule-out a COVID-19 infection in HCWs with an estimated 95% sensitivity. This was done by fitting a Gaussian to the distribution of the CoLab-linear predictor for all HCWs tested positive for SARS-CoV-2. The cut-off to safely rule out COVID-19 was chosen as the 5th percentile of the fitted Gaussian distribution. The number needed to screen (defined as the number of HCWs needed to PCR test to find one positive) is calculated by dividing the total number of HCWs below the cut-off by the number of HCWs above the cut-off and tested PCR positive. The fraction of HCWs falling below the cut-off was calculated to determine the potential reduction in PCR tests. Confidence intervals for the Gaussian fit, 5th percentile and potential reduction in PCR tests were obtained by bootstrapping and calculating the bias-corrected and accelerated bootstrap (BCa) confidence intervals (CIs). Finally, the relation between the PCR Ct and CoLab-linear predictor was plotted to determine if higher Ct-values corresponded to lower CoLab-linear predictor values. All statistical analyses were performed in R version 4.0.5 [16], calibration plots were made using the `rms`-package [17], bootstrapping was done using the boot-package [18].

5.2.5 Ethical considerations

The medical ethics committee of the Zuyderland Medical Center (METC Z) approved this study (registration nr. METCZ2021002). Data were acquired after informed consent and obtained in accordance with the Declaration of Helsinki, version 2013. Participation in this study was voluntary, and each participating HCW obtained a hard copy of the ”Test subject information sheet”, in which the study is explained and were the participant has to give written consent. These signed consent forms were also signed by the study personnel member responsible for the venipuncture. Each participant was also aware that they could opt out at any time. Because the study was restricted to HCWs, no minors (<18 years) were included.

5.3 Results

In total, 775 healthcare workers (HCWs) were included in this study. Forty-nine out of the 775 HCWs were excluded Fig. 5.3.1. A total of 42 HCWs (5.8 %) were severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) polymerase chain reaction (PCR) positive.

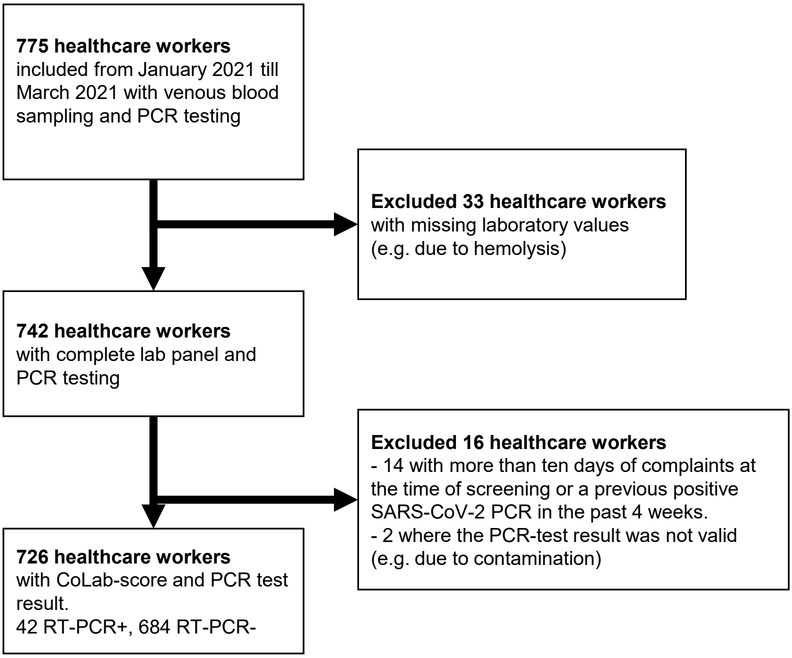

775 healthcare workers
included from January 2021 till March 2021 with venous blood sampling and PCR testing
Excluded 33 healthcare workers
with missing laboratory values (e.g. due to hemolysis)
742 healthcare workers
with complete lab panel and PCR testing
Excluded 16 healthcare workers
- 14 with more than ten days of complaints at the time of screening or a previous positive SARS-CoV-2 PCR in the past 4 weeks.
- 2 where the PCR-test result was not valid (e.g. due to contamination)
726 healthcare workers
with CoLab-score and PCR test result.
42 RT-PCR+, 684 RT-PCR-

Figure 5.3.1: Inclusion flow of patients.

Descriptive statistics for the 726 included HCWs, grouped by PCR test result, are shown in Table 5.3.1. Age, erythrocytes, lactate dehydrogenase (LD), alkaline phosphatase (ALP) and gamma-glutamyltransferase (gGT) do not show significant differences between PCR positive and PCR negative groups. All other variables included in the CoLab-score differ significantly between the 2 groups.

	PCR negative	PCR positive	p-value
N	684	42	
Age in years (mean (SD))	43.3 (12.9)	46.7 (12.3)	0.094
Male gender (%)	105 (15.4)	7 (16.7)	0.993
Erythrocytes in /pL (mean (SD))	4.76 (0.39)	4.76 (0.40)	0.897
Leukocytes in /nL (median [IQR])	6.92 [5.68, 8.43]	4.69 [3.90, 5.87]	<0.001
Eosinophils in /nL (median [IQR])	0.14 [0.09, 0.21]	0.08 [0.05, 0.12]	<0.001
Basophils in /nL (median [IQR])	0.04 [0.03, 0.06]	0.02 [0.01, 0.03]	<0.001
Bilirubin in umol/L (median [IQR])	7.0 [5.0, 9.2]	5.0 [4.0, 6.0]	<0.001
LD in U/L (mean (SD))	186.7 (34.5)	197.2 (41.8)	0.059
ALP in IU/L (median [IQR])	74.0 [61.0, 89.0]	75.0 [64.5, 96.8]	0.561
gGT in U/L (median [IQR])	17.0 [13.0, 25.0]	21.0 [14.0, 26.0]	0.166
Albumin in g/L (mean (SD))	45.7 (3.0)	44.3 (2.7)	0.005
CRP in mg/L (median [IQR])	1.7 [0.7, 4.0]	2.8 [1.4, 5.6]	0.012

Table 5.3.1: Descriptive statistics.
Shown are the laboratory tests required for the CoLab-score and their mean/median results split by PCR test result. For results with normal distributions, the mean value and SD (in round brackets) are shown. For results that have skewed or heavy tailed distributions, the median value and the interquartile range (IQR) is shown [in squared brackets]. The p-value corresponds to a t-test in case of a normal distribution (summarized by mean and standard deviation (SD)), a Mann-Whitney U-test in case of non-normally distributed variables (summarized by median and IQR) and a Fisher exact test for categorical variables. Lactate dehydrogenase (LD); alkaline phosphatase (ALP); gamma-glutamyltransferase (gGT); C-reactive protein (CRP).

ROC-curve analysis of the CoLab-linear predictor is shown in Fig. 5.3.2. The area under the ROC-curve (AUC) of the CoLab-linear predictor in discriminating between PCR positive and negative HCWs was 0.853 (95% CI: 0.801 – 0.904). The calibration plot corresponding to the predicted probabilities and observed proportion of PCR positives is plotted in Fig. 5.3.3A. The logistic regression calibration slope is equal to 1.056 (standard error (SE): 0.1438) and the intercept 2.322 (SE: 0.6197). This implies that predicted probabilities are systematically too low but re-calibration is straightforward, as there is no evidence that the slope is $\neq 1$, hence only the intercept term needs to be added to the original CoLab-linear predictor to obtain a re-calibrated linear predictor suitable for screening HCWs. The re-calibrated calibration plot is show in Fig. 5.3.3B. This also illustrates that the discriminative ability of the CoLab-linear predictor is preserved but that thresholds for screening HCWs should be lower than emergency department (ED) patients. To define a safe

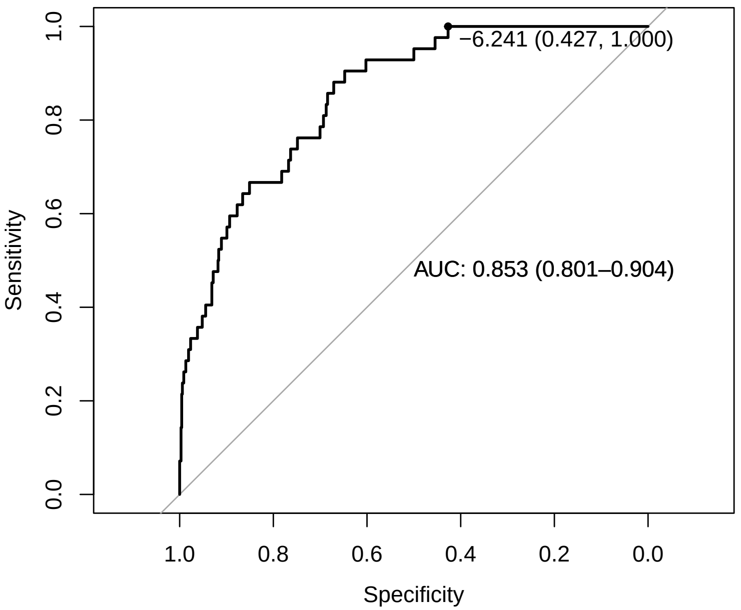


Figure 5.3.2: ROC curve of the CoLab-linear predictor.
The area under the ROC curve is shown with the 95% DeLong confidence interval in round brackets. The displayed threshold of -6.241 corresponds to a sensitivity of 100%, i.e. no HCWs below this linear predictor were PCR positive.

cut-off for excluding COVID-19 in HCWs, a Gaussian is fitted to the distribution of CoLab-linear predictor of HCWs that were tested PCR positive (Fig. 5.3.4). The Shapiro-Wilk test showed no evidence of non-normality (P-value = 0.621). The 5th percentile of the Gaussian fit of the CoLab-linear predictor is equal to -6.525 (95% CI: -7.147 to -5.999), which is recommended as the cut-off below which COVID-19 can be safely ruled-out in HCWs. Using the -6.525 cut-off, the percentage of HCWs that can be safely excluded is 34% (95% CI: 21 to 49%), with a specificity of 34%, a sensitivity of 100%, a positive predictive value (PPV) of 9% and a negative predictive value (NPV) value of 100%. The number need to test is 12 (95% CI: 10 to 14). In Fig. 5.3.5 the

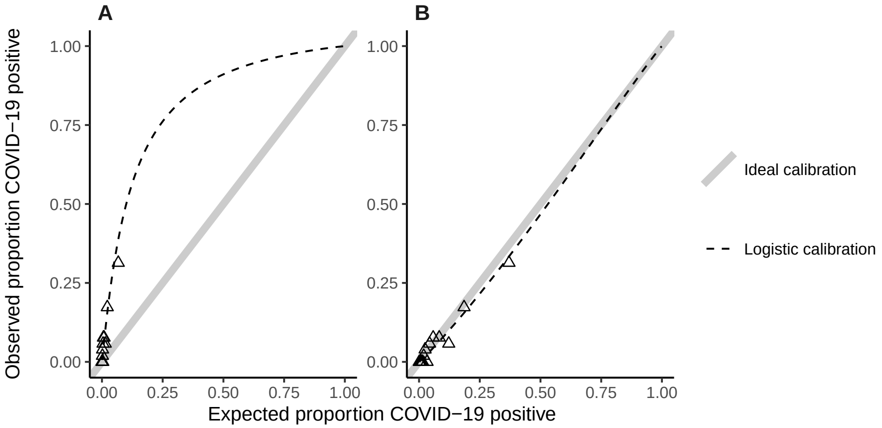


Figure 5.3.3: Calibration plots.
A: In the calibration plot the proportion of observed COVID-19 positives versus expected proportion of positives are plotted. Observations are grouped with an average of 50 observations per group. The expected probabilities follow from applying the inverse logit function to the CoLab-linear predictor. If the observed proportion in an external dataset is lower than the expected proportion, this means risks are over-estimated, if the observed fraction is higher, risks are under-estimated. Ideally, observed proportions are equal to expected proportions, this ideal-calibration-line is shown as a straight line through the origin with a slope of 1. The logistic calibration line is a logistic regression fit of the predicted probabilities.
B: Using the intercept and/or slope from the logistic regression model, recalibrated probabilities were obtained and plotted in a second calibration plot.

relationship between the CoLab-linear predictor and the PCR cycle threshold (Ct) value is plotted. The fitted smooth in Fig. 5.3.5 shows a rising Ct value (implying a decreasing amount of template) near the lower end of the CoLab-linear predictor.

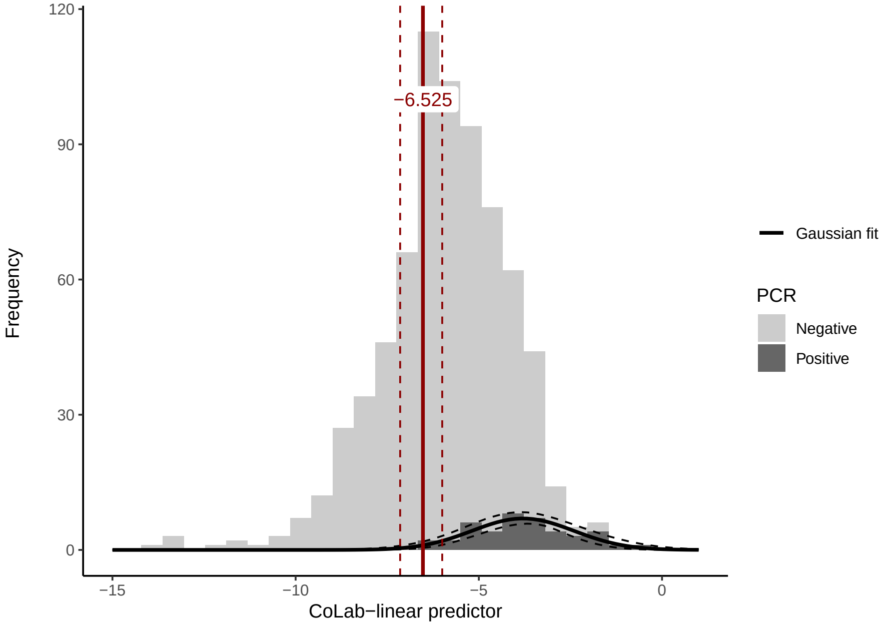


Figure 5.3.4: Histograms and fitted Gaussian distribution of the CoLab-linear predictor split by PCR result.

A normal distribution was fitted to the PCR negative group (mean: -6.04, SD: 1.73), the dashed lines represent the 95% confidence interval (CI). The 5th percentile of the Gaussian distribution is shown in red and dashed lines represent the 95% CI. Linear predictor values below this 5th percentile are regarded as non-COVID-19.

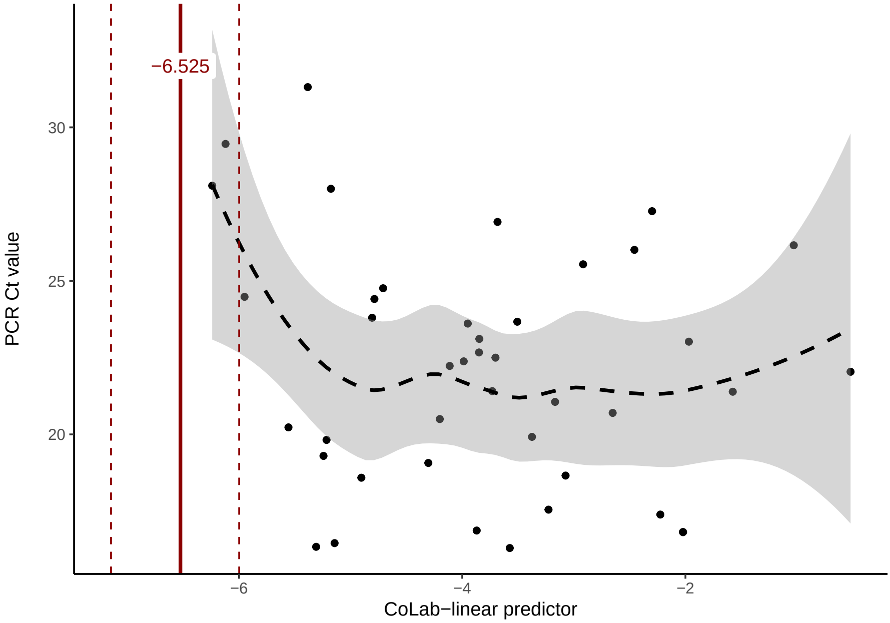


Figure 5.3.5: CoLab-linear predictor versus PCR Ct value.
The CoLab-linear predictor is plotted versus the PCR Ct value. The red line is the CoLab-linear predictor cut-off below which HCWs are regarded as non-COVID-19, the dashed red lines represent the 95% confidence interval (CI) of the cut-off. The dashed line is a LOESS smooth where the 95% CI is shown in gray.

5.4 Discussion

In this prospective study among healthcare workers (HCWs), it was shown that the model behind the CoLab-score could be used to safely exclude coronavirus disease 2019 (COVID-19) in HCWs. The original cut-off for the CoLab-linear predictor was adapted for excluding COVID-19 in HCWs. Using this adapted cut-off, a negative predictive value (NPV) of 100% was found with a specificity of 34% (95% CI: 21 to 49%). The number needed to screen by using the CoLab guided severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) polymerase chain reaction (PCR) testing was 12 (95% CI: 10 to 14). The overall prevalence of COVID-19 in this group of HCWs was 5.8%. In 33 cases (4.3%) the CoLab score could not be calculated. This was due to haemolysis of the blood sample, caused by an improper venipuncture.

PCR-based methodologies are the gold standard for confirming COVID-19. There are several factors that can contribute to false-negative results, including the adequacy of the specimen collection technique, time from exposure and specimen source. Furthermore, current and future viral changes could affect viral based diagnostics [19–21]. In addition, several studies have already shown that COVID-19 is characterized by biochemical as well as haematological changes in peripheral blood [12, 22–24]. Next to focusing on the viral response (PCR), investigating the host immune response by analysing biochemical and haematological changes in peripheral blood is an attractive alternative method [25]. As shown in the living systematic review from Wynants et al. [26], a considerable number of prediction models for COVID-19 have been published until recently, and biochemical and haematological parameters are often an important part of these prediction models. Recently the CoLab-score was developed and externally validated by Boer et al. [14] to exclude COVID-19 in patients presenting at the Emergency Department, using an adaptive lasso-regression technique [27]. This score is based on 6 biochemical and 4 haematological parameters, next to the age of the patient. It appears that the strength of the high NPV derived from this algorithm is driven by the absence of specific COVID-19 related biochemical and haematological changes in peripheral blood. As the CoLab-score is based on a categorization of the underlying continuous linear predictor, the cut-offs define the diagnostic properties of the individual scores. Our results show that the discriminative ability of the CoLab-linear predictor is preserved when screening HCWs instead of patients presenting at the emergency department, as indicated by the area under the ROC-curve (AUC) and calibration slope. The AUC is lower than the development group of the original CoLab study [14], but similar to the AUC reported for external datasets in the original CoLab study. Therefore, the discriminative ability seems to be preserved when classifying HCWs. The cut-offs defined in the original CoLab publication are however not suitable for excluding COVID-19 in HCWs. This is confirmed by the calibration intercept which shows that probabilities predicted by the original CoLab-linear predictor are systematically too low for HCWs. As reliable exclusion of COVID-19 is of utmost importance in screening HCWs, a logical choice would be to select the highest cut-off with 100% sensitivity. However, one might speculate that when sampling more HCWs, the sensitivity could drop, and PCR positive HCWs could occur even below this threshold. Therefore, the Gaussian distribution was fitted to the data and the 5th percentile chosen as safe cut-off. Doing so, the reliability increases at the expense of the number of “negative” results. We recommend that the optimal cut-off value is -6.525, where COVID-19 could be excluded in about 34% of the HCWs. Furthermore, Fig. 5.3.5 suggests that potentially “missed” COVID-19 HCWs might have relative high cycle threshold (Ct) values, potentially resulting in lower disease burden and contagiousness. It must be kept in mind that this study was performed in a time that the prevalence of COVID-19 was high ($>$ 10%) and avoiding false-negatives had the highest priority. For this reason, the optimal cut-off is now set a very low threshold. It can be hypothesized that when the COVID-19 prevalence drops, the cut-off value can be adapted by investigating which NPV can be allowed in this new setting. It turned out that the cut-off to be used in the CoLab-algorithm in the HCW screening setting is different from that of patients presenting at the emergency department (ED). This could be explained by the fact that in the screening setting, the duration of the infective period is shorter, the complaints are milder and consequently the host-immune response is also less pronounced [25]. In addition, in our study group there were 2 HCWs in which the SARS-CoV-2 PCR was initially negative, while the CoLab-score was not negative. A week later both HCWs had a re-test because of persistent COVID-19 related complaints. At that time the CoLab-linear predictor had worsened, and at that time the repeated PCR test turned out to be positive. Because these 2 HCWs had more than 10 days COVID-19-related complaints, they were excluded from this study. It would appear that the outcome of the CoLab-score is dynamic and follows the host immune response. At this time, also so-called rapid lateral flow tests (LFTs) are available which detects the presence of the SARS-CoV-2 antigen. They are widely adopted In EDs because of their ease of use and the rapid result (< 30 minutes). However, compared to these LFTs which provide a dichotomous results, the CoLab-score provides a continuous score. Using the above-mentioned cut-off value, the CoLab-score offers a higher sensitivity and are therefore more suitable to rule-out COVID-19 than a LFT, which are only moderately sensitive [28, 29]. A limitation of this study is that this study has been performed in a period where only the original SARS-CoV-2 strain, next the alpha-variant, were dominant. Since the CoLab-score reflects the host-response to the virus, it is expected that the accuracy of the score will not be changed by emerging SARS-CoV-2 variants. This assertion is supported by Boer et al., who found sustained diagnostic performance of the CoLab-score in periods with different dominant variants (especially Alpha- and Delta-variant) [14]. A control group of HCWs who did not have complaints and were not in close proximity of a COVID-19 patient, was ideally a good control group to test for false-positive results. Unfortunately, this was practically not feasible because the study was designed and performed in a period with high absenteeism of HCWs due to COVID-19. The medical board as well as the ethical committee of our hospital found it unethically to test HCWs with no complaints, especially with the knowledge that a positive SARS-CoV-2 PCR test result does not always mean that the person can spread the virus actively. Because the focus of our study was directed to develop a screening method with a high NPV, the lack of this control group can be seen as a limitation of our study. Next to this, we don not have any information about the kind of work the HCW performs (e.g. nurse, physician, laboratory personnel, cleaning staff, transport) and for that reason it is not possible to assess the group of HCWs who work in close proximity of a COVID-19 patient, separately from the group of HCWs who don’t work in close proximity.

In conclusion, the CoLab-score is an easy and reliable algorithm that, using an adapted cut-off, can be used in screening HCWs with COVID-19 related complaints. Major advantages of this approach are that the results of the score are available within 1 hour after collecting the samples, it can be implemented in almost every hospital, even in a 24/7 setting, and the costs are minimal compared to PCR testing. This results in a faster return to labour process of a significant part of the COVID-19 negative HCWs (34%), next to a reduction in PCR tests (reagents and labour duties).

5.5 Acknowledgments

The authors would like to thank Christel Jacquot, Anke Linssen and Audrey Merry for their help with this study. Additionally, authors would like to thank the laboratory technician René Nieuwenhuysen for extracting the data from the laboratory information systems. Finally, we appreciate the valuable assistance of the phlebotomists of the Department of Clinical Chemistry and Hematology of Zuyderland MC, Heerlen/Sittard-Geleen.

References

1. World Health Organization. *COVID-19 weekly epidemiological update, 23 February 2021* tech. rep. (2021).

2. Moghadas, S. M. *et al.* The implications of silent transmission for the control of COVID-19 outbreaks. *Proceedings of the National Academy of Sciences* **117,** 17513–17515 (2020).

3. He, X. *et al.* Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature medicine* **26,** 672–675 (2020).

4. Lan, F.-Y., Wei, C.-F., Hsu, Y.-T., Christiani, D. C. & Kales, S. N. Work-related COVID-19 transmission in six Asian countries/areas: A follow-up study. *PloS one* **15,** e0233588 (2020).

5. Offeddu, V., Yung, C. F., Low, M. S. F. & Tam, C. C. Effectiveness of masks and respirators against respiratory infections in healthcare workers: a systematic review and meta-analysis. *Clinical Infectious Diseases* **65,** 1934–1942 (2017).

6. Reusken, C. B. *et al.* Rapid assessment of regional SARS-CoV-2 community transmission through a convenience sample of healthcare workers, the Netherlands, March 2020. *Eurosurveillance* **25,** 2000334 (2020).

7. Fraher, E. P. *et al.* Ensuring and sustaining a pandemic workforce. *New England Journal of Medicine* **382,** 2181–2183 (2020).

8. Rueda-Garrido, J. C. *et al. Return to work guidelines for the COVID-19 pandemic* 2020.

9. Van Kampen, J. J. *et al.* Duration and key determinants of infectious virus shedding in hospitalized patients with coronavirus disease-2019 (COVID-19). *Nature communications* **12,** 1–6 (2021).

10. Guo, L. *et al.* Profiling early humoral response to diagnose novel coronavirus disease (COVID-19). *Clinical infectious diseases* **71,** 778–785 (2020).

11. Li, Q. *et al.* Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England journal of medicine* (2020).

12. Lippi, G., Simundic, A.-M. & Plebani, M. Potential preanalytical and analytical vulnerabilities in the laboratory diagnosis of coronavirus disease 2019 (COVID-19). *Clinical Chemistry and Laboratory Medicine (CCLM)* **58,** 1070–1076 (2020).

13. Langer, T. *et al.* Development of machine learning models to predict RT-PCR results for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in patients with influenza-like symptoms using only basic clinical data. *Scandinavian journal of trauma, resuscitation and emergency medicine* **28,** 1–14 (2020).

14. Boer, A.-K. *et al.* Development and validation of an early warning score to identify COVID-19 in the emergency department based on routine laboratory tests: a multicentre case–control study. *BMJ Open* **12,** e059111 (2022).

15. Steyerberg, E. W. *et al.* Assessing the performance of prediction models: a framework for some traditional and novel measures. *Epidemiology (Cambridge, Mass.)* **21,** 128 (2010).

16. R Core Team. *R: A Language and Environment for Statistical Computing* Vienna, Austria, 2020.

17. Harrell Jr, F. E. *rms: Regression Modeling Strategies* 2021.

18. Canty, A., Ripley, B., *et al. boot: Bootstrap R (S-Plus) functions* 2017.

19. Artesi, M. *et al.* A recurrent mutation at position 26340 of SARS-CoV-2 is associated with failure of the E gene quantitative reverse transcription-PCR utilized in a commercial dual-target diagnostic assay. *Journal of clinical microbiology* **58,** e01598–20 (2020).

20. Zhou, Y. *et al.* Sensitivity evaluation of 2019 novel coronavirus (SARS-CoV-2) RT-PCR detection kits and strategy to reduce false negative. *PLoS One* **15,** e0241469 (2020).

21. Ziegler, K. *et al.* SARS-CoV-2 samples may escape detection because of a single point mutation in the N gene. *Eurosurveillance* **25,** 2001650 (2020).

22. Kurstjens, S. *et al.* Rapid identification of SARS-CoV-2-infected patients at the emergency department using routine testing. *Clinical Chemistry and Laboratory Medicine* **58,** 1587–1593 (Aug. 2020).

23. Linssen, J. *et al.* A novel haemocytometric covid-19 prognostic score developed and validated in an observational multicentre european hospital-based study. *eLife* **9,** 1–37 (Oct. 2020).

24. Martens, R. J. *et al.* Hemocytometric characteristics of COVID-19 patients with and without cytokine storm syndrome on the sysmex XN-10 hematology analyzer. *Clinical Chemistry and Laboratory Medicine (CCLM)* **59,** 783–793 (2021).

25. Wiersinga, W. J., Rhodes, A., Cheng, A. C., Peacock, S. J. & Prescott, H. C. Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review. *Jama* **324,** 782–793 (2020).

26. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: Systematic review and critical appraisal. *The BMJ* **369,** 18 (Apr. 2020).

27. Zou, H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101,** 1418–1429 (Dec. 2006).

28. Garciá-Fiñana, M. *et al.* Performance of the Innova SARS-CoV-2 antigen rapid lateral flow test in the Liverpool asymptomatic testing pilot: population based cohort study. *BMJ* **374,** 1637 (July 2021).

29. Peto, T. *et al.* COVID-19: Rapid antigen detection for SARS-CoV-2 by lateral flow assay: A national systematic evaluation of sensitivity and specificity for mass-testing. *EClinicalMedicine* **36,** 100924 (June 2021).

6

Validation of the ELAN-HF Score and self-care behaviour on the nurse-led heart failure clinic after admission for heart failure

Tineke AM Vinck, **Ruben Deneer**, Cindy CAG Verstappen, Wouter E Kok, Khibar Salah, Volkher Scharnhorst & Luuk C Otterspoor

Adapted from: Validation of the ELAN-HF score and self-care behaviour on the nurse-led heart failure clinic after admission for heart failure. *BMC Nursing* **21,** 1–10 (2022).

Abstract

Aim To validate the predictive value of the European coLlaboration on Acute decompeNsated Heart Failure (ELAN-HF) score, and to assess the effect of self-care behaviour on readmission and mortality in patients after admission with acute decompensated heart failure (ADHF).

Design Quantitative, prospective, single centre, cohort study.

Methods N-Terminal pro–B-type natriuretic peptide (NT-proBNP) levels were measured on admission and discharge, and were used together with clinical and laboratory parameters to calculate the ELAN-HF score. Patients were stratified into four risk groups (low, intermediate, high, very high) according to their ELAN-HF score. The performance of the ELAN-HF score was evaluated and compared to the original study. Self-care behaviour was assessed by the European Heart Failure Self-care Behaviour Scale (EHFScBS-9). Survival analysis was used to estimate the association between both scores and re-admission for heart failure (HF) and/or all-cause mortality within 180 days.

Results 88 patients were included. The median age of the study population was 75 years (interquartile range (IQR) 69–83), 43% was female. New York Heart Association (NYHA) III/IV functional class was present at discharge in 68 patients (85%) and 27 patients (34%) had a left ventricular ejection fraction < 40%. Complete data and 180 day follow up was available for 80 patients. 55% reached the endpoint of readmission and/or all-cause mortality. There was a significant association between the ELAN-HF score and re-admission and/or mortality < 180 days (HR = 1.25, 95% CI 1.08—1.45, p = 0.003). The median EHFScBS-9 score was 68.1 (IQR 58.3 – 77.8). There was no significant association between the EHFScBS-9 score and readmission and/or mortality < 180 days (HR = 1.01, 95% CI 0.99—1.03, p = 0.174).

Conclusion This study confirms the validity and therefore the potential of the ELAN-HF score to triage patients with ADHF before discharge. Using this score may optimize the follow-up treatment on the nurse-led heart failure clinic in order to decrease readmission and mortality.

Self-care behaviour was non-significantly associated with readmission and/or mortality in our study population.

Trial Registration This study has been registered with the ethics committee MEC-U (Nieuwegein, The Netherlands), registration nr: V.160999/W18.208/HG/mk.

6.1 Introduction

Heart failure (HF) is defined by the European Society of Cardiology as a clinical syndrome and an inadequacy of the pumping function of the heart, characterized by symptoms such as shortness of breath, persistent coughing or wheezing, ankle edema and fatigue, that may be accompanied by the following signs: elevated jugular venous pressure, pulmonary crackles, increased heart rate and peripheral edema [1]. HF is a major health problem, with a prevalence of 238.700 patients and a mortality rate of 7.264 patients in 2019 in the Netherlands [2]. At present, approximately 26 million people worldwide are living with HF. The prognosis of patients suffering from HF is poor, with survival rates worse than those for bowel, breast or prostate cancer [3]. HF causes severe economic, social and personal costs. Globally, the increasing burden of HF is taking its toll on society, in particular on patients, caregivers and healthcare systems [3]. After hospitalization for acute decompensation, approximately 20% of patients with HF are readmitted within 30 days and over 50% within 6 months, with a 60-day mortality rate after admission of 15.2% [4]. Nurses on the nurse-led HF clinic play a crucial role in the prevention of readmission and mortality in patients with HF. In order to improve this care, accurate prediction scores on these events may be of great value. A common measure in the nurse-led HF outpatient clinic, in order to reduce readmission and mortality, is patient education in self-care behaviour [5]. The present study addresses both the prediction of readmission and mortality and self-care behaviour after an admission for HF.

6.1.1 Background

The concept of HF nurses working in an outpatient clinic was for the first time described in 1983 [6]. This was followed by the first nurse-led HF clinic in Sweden in 1990 after which they spread out to many Swedish hospitals. Nurse-led HF clinics reduce the need for hospital care since titration of drugs can be rapidly achieved. Furthermore, studies indicate that early follow-up after hospitalization may prevent readmissions [6]. Similar to these achievements on HF clinics, several studies on other nurse-led clinics also indicated positive effects [7]. Rich et al. investigated the effect of a multidisciplinary, nurse-directed intervention and found that the intervention improved the patients' compliance, quality of life and decreased the rate of readmission and the healthcare costs [8]. Nurses independently perform anamnesis and physical examination, and are responsible for the diagnostic processes. Nurse-led HF clinics provide education on self-care and psychosocial support to patients and their family. Programs employing multidisciplinary teams and in-person communication led to fewer HF hospital readmissions [9]. High-risk HF patients (advanced stage, low self-care skills, elderly, and those with frequent readmissions) may be expected to benefit the most from improvements in HF self-care knowledge and home care behaviour skills [9].

6.1.2 Risk of readmission; ELAN-HF score

Prognostication of patients is useful in triaging patients during and after hospitalization [10, 11]]. For this purpose, specific predictors for readmission of patients with HF have emerged [12]. By combining different clinical and laboratory parameters in a clinical prediction model, patients can be triaged just before discharge. In patients with HF, plasma biomarkers brain natriuretic peptide (BNP) and N-Terminal pro–B-type natriuretic peptide (NT-proBNP) are commonly used. They indicate the severity of congestion and cardiac dysfunction and predict morbidity and mortality [13]]. Various risk models for readmissions and mortality in HF have already been developed [11, 14–17]. They incorporate the natriuretic peptide levels, measured either at admission or discharge, while some models also use their change during hospitalization. The PRIMA II trial, which investigated the influence of changing of NT-proBNP with guided therapy and intensified HF care pre-discharge, did not demonstrate an improvement in prognosis in these patients [18]. The European coLlaboration on Acute decompeNsated Heart Failure (ELAN-HF) score is a model which is different from other risk models because it incorporates absolute discharge NT-proBNP levels, but also the percentage change in NT-proBNP, along with clinical risk markers [14]. Although it has been already validated retrospectively in a cohort of 325 patients, multiple external validations are needed to generalize the ELAN-HF score as a prediction model before it can be implemented in the clinical practice of the nurse-led HF clinic

[19, 20].

6.1.3 Self-care behaviour

Self-care behaviour is defined as the behaviour that consists of the decisions and strategies that a person undertakes for the sake of livelihood, healthy functioning and well-being [21, 22]. Previous studies indicated that optimal self-care behaviour can lead to fewer hospital admissions for HF [23]. Furthermore, is has been prospectively demonstrated that the use of information to improve self-care in HF led to a 30% decrease in readmission and outpatient visits within 30 days of discharge [23]. Self-care behaviour can be scored by using the European Heart Failure Self-care Behaviour Scale (EHFScBS-9), containing 9 items grouped around consulting behaviours and adherence with the regimen. Each of the items is graded with a 5-point Likert scale, see Table 6.1.1. This questionnaire was validated in several countries and was revised in 2014 [5].

		Totally agree	Agree	Neutral	Disagree	Totally disagree
1	I weigh myself every day.	1	2	3	4	5
2	If SOB (shortness of breath) increases, I contact my doctor or nurse.	1	2	3	4	5
3	If my legs/feet are more swollen, I contact my doctor or nurse.	1	2	3	4	5
4	If I gain weight more than 2 kg in 7 days, I contact my doctor or nurse.	1	2	3	4	5
5	I limit the amount of fluids.	1	2	3	4	5
6	If I experience fatigue, I contact my doctor or nurse.	1	2	3	4	5
7	I eat a low-salt diet.	1	2	3	4	5
8	I take my medication as prescribed.	1	2	3	4	5
9	I exercise regularly.	1	2	3	4	5

Table 6.1.1: The European Heart Failure Self-care Behaviour Scale (EHFScBS-9).

6.2 Materials and methods

6.2.1 Aims

The first aim of this study was to validate the predictive value of the European coLlaboration on Acute decompeNsated Heart Failure (ELAN-HF) score on readmission and/or mortality in a prospective study of a hospitalized heart failure (HF) population. The second aim was to assess the effect of self-care behaviour on readmission and mortality in these patients.

6.2.2 Design

We conducted a quantitative, prospective, single centre cohort study. The primary endpoint is a composite endpoint of re-admission and /or all-cause mortality at 180 days. The secondary endpoint is all-cause mortality at 180 days. Patients hospitalized for acute decompensated heart failure (ADHF) were included for three months (October - December 2017). A readmission was defined as an urgent clinical admission with a duration of at least 24 hours after a previous discharge from the hospital. Two sources were used for data collection. The data for both the outcome variables (readmission and / or mortality within 180 days), as well as the baseline characteristics and biomarkers (independent variables), were extracted from the electronic health record (EHR), whereas missing data on mortality were retrieved from the general physician. The ELAN-HF score was calculated from these data.

6.2.3 Participants

Patients with ADHF were included at admission. Excluded were patients with cognitive limitations, inability to speak the Dutch language and patients who could not be followed-up after discharge.

6.2.4 Ethical considerations

The local ethics committee of the Catharina Hospital, Eindhoven approved the study. All investigators adhered to the principles of the Declaration of

Helsinki. The measurements performed during this study were part of routine care. In addition, all included patients were informed orally and in writing by the investigator and gave written consent.

6.2.5 ELAN-HF score

The ELAN-HF score, can be calculated using the scoring rule or regression coefficients as shown in Table 6.2.1. The score can be categorized into four risk categories as defined by Salah et al., low (≤ 2 points), intermediate (3-4 points), high (5-7 points) and very high (≥ 8 points) which correspond to increasing 6-month mortality rates (3.6%, 9.2%, 23.5% and 51.1%) [14].

Predictor	Score	Regression coefficient
NT-proBNP reduction <30%	1	0.511
NT-proBNP discharge value,		
1500-5000 pg/ml	1	0.713
5001-15000 pg/ml	3	1.426
>15000 pg/ml	4	1.776
Age at admission ≥75 years	1	0.345
Peripheral edema at admission	1	0.517
SBP at admission ≤115 mmHg	1	0.431
Hyponatremia at admission, sodium <135 mmol/L	1	0.374
Serum urea at discharge ≥ 15 mmol/L	1	0.486
NYHA class III/IV at discharge	1	0.403

Table 6.2.1: Calculation of ELAN-HF score. The NT-proBNP reduction is the percentage change between NT-proBNP on admission and NT-proBNP on discharge. Maximum “penalty points” in the risk score is 11. N-Terminal pro–B-type natriuretic peptide (NT-proBNP); systolic blood pressure (SBP); New York Heart Association (NYHA).

6.2.6 Self-care behaviour

To assess self-care behaviour, the Dutch version of the European Heart Failure Self-care Behaviour Scale (EHFScBS-9), was used, see Table 6.1.1 [24]. The questionnaire was handed out by the investigator during admission and was completed by the patient without the presence of the investigator. The

EHFScBS-9 score ranges from 9 – 45, but has been standardised to a scale of 0 – 100 with higher scores indicating better self-care [5]. The details to calculate the standardised score are given in the study of Vellone et al.[5]. The standardised score has an easier interpretation, where a score < 50 means suboptimal score of self-care, and a score 50-100 can be seen as an optimal/good score of self-care [5].

6.2.7 Statistical analysis

The statistical analyses consisted of two parts: i) external validation of the ELAN-HF score in our study cohort and ii) survival analysis of the ELAN-HF score, EHFScBS-9 score and other clinically relevant variables.

External validation

External validation is the process of evaluating model performance in a sample independent of that used to develop the model. The outcome used for external validation was 6-month all-cause mortality, analogous to Salah et al. [14]. The 6-month mortality rates for the four risk groups as reported by Salah et al. were compared to those of our study cohort. The external validation steps performed in this study are described in more detail by Royston et. al [25]. First, the ELAN-HF linear predictor was calculated by using the regression coefficients in Table 6.2.1. The ELAN-HF linear predictor was then used as a covariate in a Cox proportional hazards (PH) model. A likelihood ratio (LR) test was performed to test whether the slope of the ELAN-HF linear predictor was equal to 1. Secondly, the model misspecification was tested formally by running a Cox PH model on all the ELAN-HF covariates and constraining the coefficient of the ELAN-HF linear predictor to 1. Thirdly, the discriminative ability of the ELAN-HF score was evaluated using Harrell's c-index. Finally, calibration was evaluated for predicting all-cause mortality at 6-months. The baseline-hazard at 6 months was obtained through personal correspondence with the authors of the ELAN-HF paper [14]. Patients were grouped based on expected/predicted probabilities and observed probabilities were calculated. Plotting expected versus observed probabilities yielded a calibration plot.

Survival analysis

Survival curves were analysed for the ELAN-HF and EHFScBS-9 scores and compared with log-rank tests. For the ELAN-HF the score categories described by Salah et.al were used as reference. In case of the EHFScBS-9 score, patients with an EHFScBS-9 normalized score lower than or equal to the median EHFScBS-9 normalized score were categorized as "low", and patients above the median as "high". Kaplan-Meier (KM) curves were analysed using the log rank test to assess if there were significant differences between groups in cumulative incidence of events. An event was defined as time to first readmission or time till death from any cause. Secondly, survival was analysed by Cox PH models by using time to readmission or all-cause mortality within 6 months as the outcome. Univariate Cox PH models were fit to a subset of clinically relevant variables that were not in the ELAN-HF score. The variables that were tested significant in univariate analysis were then included in a multivariate Cox PH model to assess whether the ELAN-HF score could be improved by the EHFScBS-9 score or other variables. Statistical analyses were performed in R version 4.0.3 [26] and a p-value < 0.05 was considered statistically significant.

6.2.8 Validity, reliability and rigour

The ELAN-HF score consists of clinical variables (age, peripheral edema on admission, systolic blood pressure and New York Heart Association (NYHA) class) and biomarkers (NT-proBNP, sodium and urea). All variables were collected from the EHR. Biomarkers were measured in the clinical laboratory on a Cobas 8000 Pro (Roche Dx, Basel, Switzerland) instrument. The reliability and validity of the EHFScBS-9 to measure self-care has been extensively researched in multiple studies [16, 27, 28]. These studies show that the psychometric properties of the EHFScBS-9 are satisfactory.

6.3 Results

Eighty-eight patients fulfilled inclusion criteria of whom 8 patients were excluded due to lack of follow-up after discharge. Baseline characteristics are demonstrated in Table 6.3.1. The median age was 75 years (interquartile range (IQR) 69-83), 38 patients (47.5%) were diagnosed with atrial fibrillation at admission and 41 patients (51.2%) had a history of ischemic heart disease. Twenty-six patients (32.5 %) had been previously hospitalized for acute decompensated heart failure (ADHF) in the penultimate year. After 180 days, more than half (n = 44, 55%) of the patients had an event. Thirty-five patients were readmitted and twenty-one patients died, within 180 days after discharge

6.3.1 External validation

Table 6.3.2 presents 6-month all-cause mortality according to subdivisions of the European coLlaboration on Acute decompeNsated Heart Failure (ELAN-HF) score risk groups, comparing actual and predicted mortality. Fig. 6.3.1 shows the calibration plot for predicting 6-month all-cause mortality. Although there is a relatively small of number of patients in each group (20), there are no signs of miscalibration. The slope of the ELAN-HF linear predictor in the validation cohort was 0.80 (standard error (SE) = 0.22), the slope is not significantly different from 1 (likelihood ratio (LR) test $\chi^2_{df=1} = 0.81$, p = 0.367), so the discrimination of the ELAN-HF score seems to be preserved in our cohort. There was also no evidence of model misspecification, a joint test of all the predictors was non-significant ($\chi^2_{df=10} = 14.71$, p = 0.143), meaning that the regression coefficients of the ELAN-HF score do not appear biased. The discriminative ability expressed in Harrell's c-index was 0.719 (SE = 0.056) in our cohort, this is similar to the reported index by Salah et al. of 0.71 [14].

6.3.2 Survival analysis

Fig. 6.3.2 show the relationship between the risk groups derived from both European Heart Failure Self-care Behaviour Scale (EHFScBS-9) (panel A) and

ELAN-HF scores (panel B) and the composite endpoint of readmission and/or all-cause mortality. There was no significant difference in composite endpoint among patients with low EHFScBS-9 score (i.e., below or equal to the median of the normalized EHFScBS-9 score) in comparison to patients with a high score (24%, versus 31% respectively Kaplan-Meier (KM)-log rank test p = 0.15). Readmission and/or mortality rate was significantly higher in patients with higher ELAN-HF scores in comparison to those with low scores (KM log-rank test p = 0.0071). Due to the smaller sample size, there is an overlap between survival curves of the low and intermediate, and high and very high-risk groups. Univariate Cox regression analysis for the composite endpoint results is shown in Table 6.3.3. Univariate analysis did not show that the normalized EHFScBS-9 score was associated with 6-month readmission and/or mortality. Other than the ELAN-HF score, two additional variables showed a significant association; whether the patient was admitted with ADHF in the previous year, and whether the patient is an outpatient clinic patient. Both factors increased the risk of 6-month readmission and/or mortality. This association remained significant in multivariate analysis. A LR-test revealed that adding these variables to the ELAN-HF score improved the model fit ($\chi^2_{df=2} = 10.61$, p = 0.005) in predicting risk of 6-month readmission and/or mortality.

	ELAN-HF score low or intermediate	ELAN-HF score high or very high	Overall
n	29	51	80
Female gender (%)	10 (34.5)	24 (47.1)	34 (42.5)
Age in years (mean (SD))	72.2 (10.7)	76.5 (8.5)	74.9 (9.5)
BMI in kg/m^2 (mean (SD))	26.6 (5.5)	27.1 (6.2)	26.9 (5.9)
History of DM (%)	6 (20.7)	13 (25.5)	19 (23.8)
History of COPD (%)	5 (17.2)	8 (15.7)	13 (16.2)
Atrial fibrillation at admission (%)	13 (44.8)	25 (49.0)	38 (47.5)
Admitted with ADHF in past year (%)	7 (24.1)	19 (37.3)	26 (32.5)
History of valvular disease (%)	20 (69.0)	34 (66.7)	54 (67.5)
Ischaemic aetiology (%)	14 (48.3)	27 (52.9)	41 (51.2)
Outpatient clinic patient (%)	1 (3.4)	15 (29.4)	16 (20.0)
NYHA class at discharge (%)			
II	5 (17.2)	7 (13.7)	12 (15.0)
III	19 (65.5)	31 (60.8)	50 (62.5)
III-IV	5 (17.2)	13 (25.5)	18 (22.5)
Left Ventricular Ejection Fraction (%)			
Preserved	11 (37.9)	20 (39.2)	31 (38.8)
Moderately reduced	11 (37.9)	11 (21.6)	22 (27.5)
Reduced	7 (24.1)	20 (39.2)	27 (33.8)
NT-proBNP at admission pg/ml (median [IQR])	3440.0 [2617.0, 5241.0]	6781.0 [3884.5, 14211.5]	5604.0 [3038.5, 10005.2]
NT-proBNP at discharge pg/ml (median [IQR])	1892.0 [728.0, 2376.0]	5942.0 [3056.5, 10968.0]	3505.0 [1911.5, 7860.8]
NT-proBNP change % (mean (SD))	-57.9 (24.3)	-6.1 (54.2)	-24.8 (51.9)
ELAN-HF score (median [IQR])	3.0 [2.0, 4.0]	6.0 [5.0, 7.0]	5.0 [3.8, 6.0]
ELAN-HF score risk category (%)			
Low	10 (34.5)	0 (0.0)	10 (12.5)
Intermediate	19 (65.5)	0 (0.0)	19 (23.8)
High	0 (0.0)	41 (80.4)	41 (51.2)
Very high	0 (0.0)	10 (19.6)	10 (12.5)
Normalized EHFScBS-9 score (median [IQR])	61.1 [50.0, 75.0]	69.4 [61.1, 77.8]	68.1 [58.3, 77.8]
Outcome (%)			
Event-free	20 (69.0)	16 (31.4)	36 (45.0)
Readmission	6 (20.7)	17 (33.3)	23 (28.7)
Mortality	3 (10.3)	6 (11.8)	9 (11.2)
Readmission and mortality	0 (0.0)	12 (23.5)	12 (15.0)

Table 6.3.1: Baseline characteristics.
BMI (Body Mass Index) based on clinical measurements of weight en length. Standard deviation (SD); interquartile range (IQR); Chronical Obstructive Pulmonary Disease (COPD); Diabetes Mellitus (DM); acute decompensated heart failure (ADHF); New York Heart Association (NYHA); N-Terminal pro–B-type natriuretic peptide (NT-proBNP); European coLlaboration on Acute decompeNsated Heart Failure (ELAN-HF); European Heart Failure Self-care Behaviour Scale (EHFScBS-9).

ELAN-HF score risk group	ELAN-HF cohort	Study cohort (95% CI)
Low (≤ 2)	3.6 %	10.0 % (0 – 28.8 %)
Intermediate (3 - 4)	9.2 %	10.8 % (0 – 23.3 %)
High (5-7)	23.5 %	29.3 % (13.8 – 41.9 %)
Very high (≥ 8)	51.1 %	60.0 % (14.5 – 81.3 %)

Table 6.3.2: 6-month mortality rates.
Comparison between 6-month mortality rates in the ELAN-HF development cohort and in this study cohort. If calibration is good, mortality rates should agree.

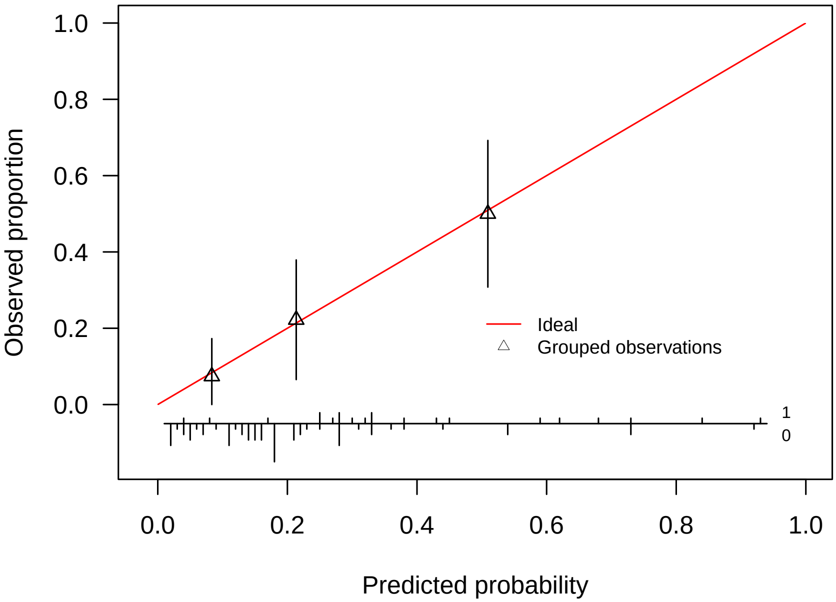
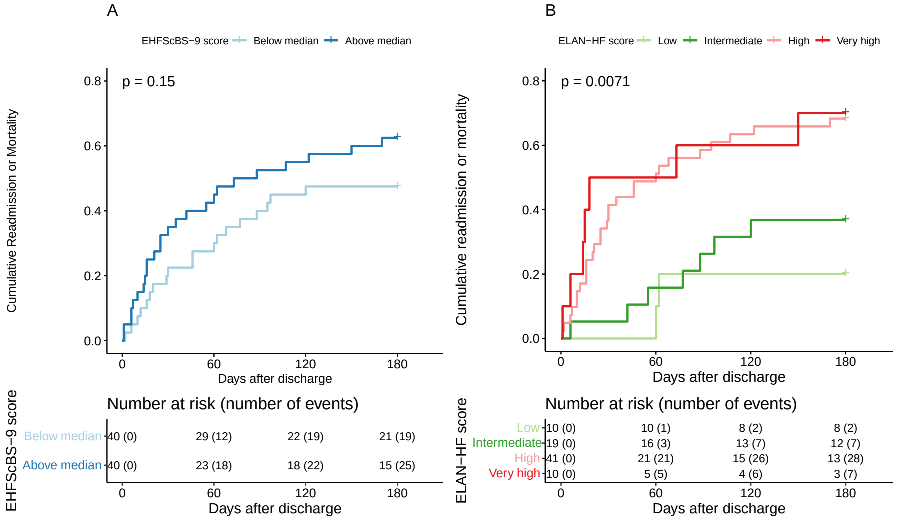


Figure 6.3.1: Calibration plot for predicting 6-month all-cause mortality.
Observations are grouped into groups of 20 patients, the ideal line represents the diagonal along which there is perfect calibration. The histogram on the bottom shows the distribution of patients with (= 1) and without (= 0) an event.



Number at risk (number of events)

EHFScBS−9 score	0	60	120	180
Below median	40 (0)	29 (12)	22 (19)	21 (19)
Above median	40 (0)	23 (18)	18 (22)	15 (25)

Days after discharge

Number at risk (number of events)

ELAN−HF score	0	60	120	180
Low	10 (0)	10 (1)	8 (2)	8 (2)
Intermediate	19 (0)	16 (3)	13 (7)	12 (7)
High	41 (0)	21 (21)	15 (26)	13 (28)
Very high	10 (0)	5 (5)	4 (6)	3 (7)

Days after discharge

Figure 6.3.2: Kaplan-Meier (KM) curves.
A: KM curve for composite endpoint of readmission and/or mortality within 180 days in relation to the self-care behaviour EHFScBS-9 score. On the X-axis the time in days until the first HF readmission or all-cause mortality within 180 days. On the Y-axis the event rate in percentages.
B: KM curve for composite endpoint of readmission and/ or mortality within 180 days in relation to the ELAN-HF risk score categories. On the X-axis the time in days until the first HF readmission or all-cause mortality within 180 days. On the Y-axis the event rate in percentages.

	Univariate HR	Univariate p-value	Multivariate HR	Multivariate p-value
Female gender	1.45 (0.8 to 2.61)	0.223		
History of DM	1.05 (0.53 to 2.08)	0.888		
History of COPD	1.87 (0.92 to 3.8)	0.084		
Atrial fibrillation at admission	1.51 (0.83 to 2.73)	0.175		
Admitted with ADHF in past year	2.42 (1.33 to 4.4)	0.004	1.90 (1.02 - 3.54)	0.044
Outpatient clinic patient	2.78 (1.45 to 5.35)	0.002	2.16 (1.10 - 4.24)	0.025
Left Ventricular Ejection Fraction,				
Preserved	Reference			
Moderately reduced	1.02 (0.63 to 1.66)	0.927		
Reduced	1.28 (0.73 to 2.23)	0.387		
ELAN-HF score	1.27 (1.11 to 1.46)	<0.001	1.24 (1.085 - 1.44)	0.003
EHFScBS-9 score normalized	1.01 (0.99 to 1.03)	0.174		

Table 6.3.3: Cox regression analysis (univariate and multivariate) for readmission and/or mortality $\leq$ 180 days.
Diabeters mellitus (DM); chronical obstructive pulmonary disease (COPD); acute decompensated heart failure (ADHF); European coLlaboration on Acute decompeNsated Heart Failure (ELAN-HF); European Heart Failure Self-care Behaviour Scale (EHFScBS-9).

6.4 Discussion

6.4.1 Validation of the ELAN-HF score

The European coLlaboration on Acute decompeNsated Heart Failure (ELAN-HF) study retrospectively defined a risk score model for all-cause mortality for patients discharged after acute decompensated heart failure (ADHF) [14], which was validated in an independent cohort [20]. In our study, we validated the ELAN-HF score with prospectively collected data and demonstrated that 180-day mortality can be robustly predicted. To assess the added value of the European Heart Failure Self-care Behaviour Scale (EHFScBS-9) score, we used the composite endpoint of mortality and/or readmission. Note that the composite endpoint event rate in our study was 55%, compared to 43% in the ELAN-HF study. This can be explained by the fact that patients in our study had a relatively higher New York Heart Association (NYHA)-class compared to the ELAN-HF study. While the ELAN-HF score demonstrated significant association with the composite endpoint, the EHFScBS-9 self-care score did not. Therefore we argue that implementing the ELAN-HF risk score on the nurse-led heart failure (HF) clinic can offer sufficient guidance to follow-up high risk patients and we strongly suggest to add this score to the discharge checklist as standard care. The high-risk population could benefit from more aggressive treatment and also from a closer follow-up by intensive (tele-)monitoring throughout the entire HF care network.

6.4.2 Self-care behaviour and prognosis

While earlier studies demonstrated a relationship between better self-care and a reduced readmission rate [23], self-care behaviour was non-significantly associated with readmission and/or mortality in our study population. This is most likely caused by an on-average high normalized self-care score (median of 68) within our cohort. These patients already received self-care education and were experienced with adjusting their lifestyle, knowledge of their disease and alarming symptoms. However, an optimal self-care score can always benefit from improvement [21, 22]. Therefore, it remains important to invest in improving self-care behaviour and to optimize patient education in HF and self-care activities by nurses, during the discharge and outpatient phases.

6.4.3 Limitations

Several limitations of our analyses should be acknowledged. First, is the number of participants. While the sample size was sufficient to demonstrate the prognostic value of the ELAN-HF model, this limits the power to detect additional prognostic factors. A second limitation is that using self-reports as in the EHFScBS-9 may be affected by memory and social desirability biases.

6.5 Conclusion

Patients admitted with acute decompensated heart failure (ADHF) have a high risk of post-discharge readmission and death. In this study, we validated the European coLlaboration on Acute decompeNsated Heart Failure (ELAN-HF) model that can be used to triage these patients into different risk groups. Based on this knowledge, follow-up treatment in the nurse-led heart failure (HF) clinic can be adjusted in order to improve prognosis. Self-care behaviour was non-significantly associated with readmission and/or mortality in our study population, most likely due to the fact that most patients already score optimal in terms of self-care. However, in our opinion, to achieve optimal outcomes, combining risk stratification and applying self-care behaviour is of great importance on the nurse-led HF clinic.

References

1. Ponikowski, P. *et al.* 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution. *European Journal of Heart Failure* **18,** 891–975 (Aug. 2016).

2. De Boer, A. R., van Dis, I., Wimmers, R., Vaartjes, I. & Bots, M. *Hart- en vaatziekten in Nederland* (Nederlandse Hartstichting, Den Haag, 2020).

3. Ponikowski, P. *et al.* Heart failure: preventing disease and death worldwide. *ESC Heart Failure* **1,** 4–25 (Sept. 2014).

4. Desai, A. S. & Stevenson, L. W. Rehospitalization for heart failure: predict or prevent? *Circulation* **126,** 501–506 (July 2012).

5. Vellone, E. *et al.* The European heart failure self-care behaviour scale: New insights into factorial structure, reliability, precision and scoring procedure. *Patient Education and Counseling* **94,** 97–102 (Jan. 2014).

6. Strömberg, A., Mårtensson, J., Fridlund, B. & Dahlström, U. Nurse-led heart failure clinics in Sweden. *European Journal of Heart Failure* **3,** 139–144 (2001).

7. Savarese, G., Lund, L. H., Dahlström, U. & Strömberg, A. Nurse-Led Heart Failure Clinics Are Associated With Reduced Mortality but Not Heart Failure Hospitalization. *Journal of the American Heart Association* **8** (May 2019).

8. Rich, M. W. *et al.* A multidisciplinary intervention to prevent the readmission of elderly patients with congestive heart failure. *The New England journal of medicine* **333,** 1190–1195 (Nov. 1995).

9. Smith, C. E. *et al.* Nurse-led multidisciplinary heart failure group clinic appointments: Methods, materials, and outcomes used in the clinical trial. *Journal of Cardiovascular Nursing* **30,** S25–S34 (2015).

10. Bhardwaj, A. & Januzzi, J. L. Natriuretic peptide-guided management of acutely destabilized heart failure: Rationale and treatment algorithm. *Critical Pathways in Cardiology* **8,** 146–150 (Dec. 2009).

11. O'Connor, C. M. *et al.* Triage After Hospitalization With Advanced Heart Failure. The ESCAPE (Evaluation Study of Congestive Heart Failure and Pulmonary Artery Catheterization Effectiveness) Risk Model and Discharge Score. *Journal of the American College of Cardiology* **55,** 872–878 (Mar. 2010).

12. Roger, V. L. Epidemiology of heart failure. *Circulation Research* **113,** 646–659 (2013).

13. Januzzi, J. L. *et al.* NT-proBNP testing for diagnosis and short-term prognosis in acute destabilized heart failure: An international pooled analysis of 1256 patients: The international collaborative of NT-proBNP study. *European Heart Journal* **27,** 330–337 (Feb. 2006).

14. Salah, K. *et al.* A novel discharge risk model for patients hospitalised for acute decompensated heart failure incorporating N-terminal pro-B-type natriuretic peptide levels: A European coLlaboration on Acute decompeNsated Heart Failure: ÉLAN-HF Score. *Heart* **100,** 115–125 (Jan. 2014).

15. Fonarow, G. C., Adams, K. F., Abraham, W. T., Yancy, C. W. & Boscardin, W. J. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. *JAMA* **293,** 572–580 (Feb. 2005).

16. Lee, C. S. *et al.* Validity and reliability of the European Heart Failure Self-care Behavior Scale among adults from the United States with symptomatic heart failure. *European Journal of Cardiovascular Nursing* **12,** 214–218 (Apr. 2013).

17. Kociol, R. D. *et al.* Admission, discharge, or change in B-type natriuretic peptide and long-term outcomes: Data from Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF) linked to Medicare claims. *Circulation: Heart Failure* **4,** 628–636 (Sept. 2011).

18. Stienen, S. *et al.* Rationale and design of PRIMA II: A multicenter, randomized clinical trial to study the impact of in-hospital guidance for acute decompensated heart failure treatment by a predefined NT-PRoBNP target on the reduction of readmIssion and Mortality rAtes. *American Heart Journal* **168,** 30–36 (2014).

19. Debray, T. P. *et al.* A new framework to enhance the interpretation of external validation studies of clinical prediction models. *Journal of Clinical Epidemiology* **68,** 279–289 (Mar. 2015).

20. Salah, K. *et al.* External Validation of the ELAN-HF Score, Predicting 6-Month All-Cause Mortality in Patients Hospitalized for Acute Decompensated Heart Failure. *Journal of the American Heart Association* **8** (2019).

21. Jaarsma, T., Strömberg, A., Mårtensson, J. & Dracup, K. Development and testing of the European Heart Failure Self-Care Behaviour Scale. *European Journal of Heart Failure* **5,** 363–370 (2003).

22. Carlson, B., Riegel, B. & Moser, D. K. Self-care abilities of patients with heart failure. *Heart and Lung: Journal of Acute and Critical Care* **30,** 351–359 (Sept. 2001).

23. Stamp, K. D., Machado, M. A. & Allen, N. A. Transitional care programs improve outcomes for heart failure patients: An integrative review. *Journal of Cardiovascular Nursing* **29,** 140–154 (2014).

24. Jaarsma, T., Årestedt, K. F., Mårtensson, J., Dracup, K. & Strömberg, A. The European Heart Failure Self-care Behaviour scale revised into a nine-item scale (EHFScB-9): A reliable and valid international instrument. *European Journal of Heart Failure* **11,** 99–105 (Jan. 2009).

25. Royston, P. & Altman, D. G. External validation of a Cox prognostic model: Principles and methods. *BMC Medical Research Methodology* **13** (2013).

26. R Core Team. *R: A Language and Environment for Statistical Computing* Vienna, Austria, 2020.

27. Sedlar, N. *et al.* Measuring self-care in patients with heart failure: A review of the psychometric properties of the European Heart Failure Self-Care Behaviour Scale (EHF-ScBS). *Patient Education and Counseling* **100,** 1304–1313 (July 2017).

28. Lin, C. Y. *et al.* Psychometric Properties of the 9-item European Heart Failure Self-care Behavior Scale Using Confirmatory Factor Analysis and Rasch Analysis Among Iranian Patients. *The Journal of cardiovascular nursing* **33,** 281–288 (May 2018).

7

Validation of heart rate extracted from wrist-based photoplethysmography in the perioperative setting: a prospective observational study

Eveline Mestrom, **Ruben Deneer**, Alberto G Bonomi, Jenny Margarito, Jos Gelissen, Reinder Haakma, Hendrikus HM Korsten, Volkher Scharnhorst & R Arthur Bouwman

Adapted from: Validation of heart rate extracted from wrist-based photoplethysmography in the perioperative setting: prospective observational study. *JMIR Cardio* **5,** e27765 (2021).

Abstract

Background Measurement of heart rate (HR) through an unobtrusive, wrist-worn optical heart rate monitor (OHRM) could enable earlier recognition of patient deterioration in low acuity settings and enable timely intervention.

Objective The goal of this study is to assess the agreement between the HR extracted from the OHRM and the gold standard 5-lead electrocardiogram (ECG) connected to a patient monitor during surgery and in the recovery period.

Methods In patients undergoing surgery requiring anesthesia, the HR reported by the patient monitor's ECG module, was recorded and stored simultaneously with the photopletysmography (PPG) signal from the OHRM attached to the patient's wrist. The agreement between the HR reported by the patient monitor and the HR extracted from the OHRM's PPG signal was assessed using Bland-Altman analysis during the surgical and recovery phase.

Results A total of 271.8 hours of data in 99 patients was recorded simultaneously by the OHRM and patient monitor. Median coverage was 86% (interquartile range (IQR) range: 65% to 95%) and did not differ significantly between surgery and recovery (Wilcoxon paired difference test p-value: 0.17). Agreement analysis showed the limits of agreement (LoA) of the difference between the OHRM and the ECG HR were within the range of ± 5 beats per minute (bpm). The mean bias was -0.14 bpm (LoA between -3.08 and 2.79 bpm) and -0.19% (LoA between -4.79 and 4.39%) for the PPG measured HR compared to the ECG measured HR during surgery and -0.11 bpm (LoA between -2.79 and 2.59 bpm) and -0.15% (LoA between -3.92 and 3.64%) during recovery.

Conclusions This study shows that an OHRM equipped with a PPG sensor can measure HR within the ECG reference standard of ± 5 bpm or ± 10% in the perioperative setting when the PPG signal is of sufficient quality. This implies that an OHRM can be considered clinically acceptable for heart rate monitoring in low acuity hospitalized patients.

7.1 Introduction

Timely recognition of deterioration in hospitalized patients is important because early intervention improves clinical outcomes such as mortality, unplanned intensive care unit (ICU) admissions and reduce length of stay [1]. Especially in perioperative care, complications related to surgery limit effectiveness of the surgery and are associated with increased mortality and costs [2, 3]. From previous studies, it is known that vital signs such as heart rate (HR) and respiratory rate are important indicators of critical illness and are often altered long before a deterioration is clinically apparent [4–6]. In general, patients' vital signs are assessed multiple times a day on general wards. However, patients may deteriorate between the scheduled measurements [1]. Therefore, both remote as well as continuous monitoring of heart rate and respiratory rate is considered a promising tool for early detection of patient deterioration in the low acuity or home setting.

The gold standard for measurement of HR in the perioperative setting is the multiple lead electrocardiogram (ECG). However, there are practical limitations to continuous measurements of vital signs using ECG due to the obtrusiveness and limited mobility of patients. Novel solutions to monitor vital signs have been proposed in literature [7]. One of these novel solutions is the wrist-based optical heart rate monitor (OHRM). The OHRM has the advantage of offering unobtrusive, remote and continuous monitoring. The photopletysmography (PPG) sensor in the OHRM has shown potential to provide robust peak detection from which heart rate may be calculated [8, 9]. Validation studies have been presented on the accuracy of these devices in healthy subjects [10–17]. However, it remains unclear whether these tools are also reliable for monitoring vital signs in patients during hospital stay. The robustness of an OHRM should be studied in hospitalized patients before it can be reliably adopted in a clinical setting. Few studies were performed in hospitalized patients, and these included mainly stable ward patients [13, 18, 19]. To check the accuracy of the OHRM in acute phases of disease, the study population should ideally experience some deterioration in heart rate during the study period. Hospitalized patients are a heterogeneous population where HR can be influenced by all kinds of pathologies. Particularly during surgery, which induces hemodynamic, metabolic, endocrine and immunological alterations [20, 21]. The objective of this study was to assess the agreement between the HR extracted from a PPG-sensor based OHRM and the gold standard patient 5-lead ECG connected to the patient monitor during surgery and recovery.

7.2 Methods

7.2.1 Study design

The study is a prospective, non-randomized observational single center study covering the perioperative period. The study was performed in the Catharina Hospital in Eindhoven, the Netherlands, a tertiary hospital performing an average of 20000 surgical procedures annually. The study was reviewed and approved by the medical ethical committee MEC-U (study number NL65134.100.18).

7.2.2 Study population

All adult patients scheduled for non-cardiac surgery were screened by anesthesiologists for inclusion in the study. Patients were selected by the anesthesiologist on a weekly basis and informed of the study prior to the surgical procedure. In total, 203 patients were eligible for inclusion and 100 patients signed informed consent. Cardiac surgeries were excluded since the required extracorporeal circulation and scheduled intensive care unit (ICU) admission would complicate analysis. To obtain a representative case-mix of patients undergoing surgery, patients were categorized and stratified based on the American Society of Anesthesiologists physical status (ASA-PS) score [22] and risk of the surgery [23]. Patients were divided into two groups: i) low risk (ASA-PS score I or II and low or intermediate risk surgery) and ii) high risk (ASA-PS score III or IV and intermediate or high risk surgery). If the ASA-PS score and risk were discordant (e.g. ASA-PS score IV and low risk surgery) the ASA-PS score took precedence over the surgical risk.

7.2.3 Study procedure

The measurements on the optical heart rate monitor (OHRM) started as soon as the device was placed on the patient's wrist in the holding area. The vital sign measurement started upon arrival in the operating room (OR) when sensor modules were connected to the patient monitor. The choice of wrist depended on the placement of the blood pressure cuff. Unless not otherwise possible, the OHRM was placed on the wrist of the arm opposite to the blood pressure cuff to prevent disturbance in the optical measurements of the cardiac pulse. Measurements continued during surgery (surgical phase). After completion of the surgery, the patient was disconnected from the patient monitor located in the OR and transferred to the recovery room. Upon arrival in the recovery room, the patient monitor was reconnected to the patient monitor located in the recovery room and measurements continued (recovery phase) until the patient was transferred to the general ward. Upon transfer, the patient monitor was disconnected and the OHRM removed from the patient's wrist.

7.2.4 Data collection

The wrist-worn OHRM was developed by Philips and equipped with a Philips Cardio and Motion Monitoring Module, which integrates a photopletysmography (PPG) and an accelerometer sensor, see Fig. 7.2.1. Photoplethysmography is an optical technique used to detect volumetric changes in blood in peripheral circulation. It continuously measures the reflectivity of the skin in the green part of the light spectrum in combination with the 3-axial acceleration of the body part where it is located. Accelerometry is a technique used to quantify movement patterns through the detection of rotational and translational acceleration. The sampling frequency of both the PPG and accelerometer sensors was 32 Hz [24]. The patient monitor in both the operating and recovery room was a GE Carescape B850 connected to a 5-lead electrocardiogram (ECG), pulse oximeter, body temperature sensor and oscillometric cuff for noninvasive blood pressure measurements or an arterial line for invasive blood pressure measurements. All patient monitors were linked to a patient data collection system which logged data for every patient. The application used for logging data was AnStat (CarePoint). AnStat logs trends and waveforms with a sampling frequency of 100 Hz, and events like administration of drugs.

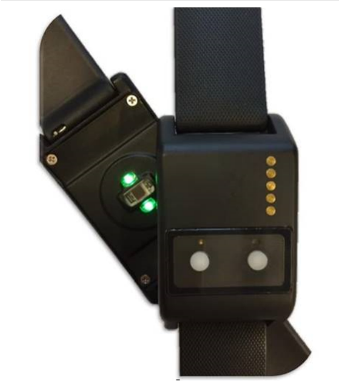


Figure 7.2.1: The wrist-worn optical heartrate monitor.

7.2.5 Data processing

The heart rate (HR) from the 5-lead ECG was derived by the GE Carescape B850 patient monitor's software. The HR from the OHRM was extracted from the logged PPG signal using an algorithm that was previously validated in healthy volunteers in various conditions of rest and physical activity [25]. In brief, the algorithm processed simultaneously the PPG and motion signal to derive HR and a quality index (QI) for the HR measurements on a 1-sec interval. Both HR and QI are assessed real-time. The algorithm provides an output every 1-second, the data however, is processed using a sliding window of 5 seconds. HR from ECG and PPG were synchronized using a cross-correlation function and visual inspection of the resulting overlapped time-series. The QI characterizes the confidence in the provided metric estimated by the algorithm itself. It is represented on a 5-point scale [0, . . . , 4], where 0 denotes "lowest confidence/output unavailable" and 4 denotes "highest confidence". The QI is determined by proprietary methods and aims to provide a monotonically increasing relation between availability and reliability. The QI of the HR would typically be influenced by the signal-to-noise ratio of the PPG signal, the ability of the algorithm to cope with motion artefacts, and the periodicity of the detected pulse signal. A PPG-based arrhythmia detection algorithm [26] was also used to identify periods in which the PPG signal was not in accordance with a normal sinus rhythm. In brief, the arrhythmia detection algorithm would first identify inter-pulse-intervals (IPIs) in a 30-seconds interval form the PPG signal and then reject the IPIs in presence of motion during the IPI period. The final set of IPIs in the 30 sec period are then processed by a Markov Model to define the probability of atrial fibrillation (AF). If $>50\%$ of the detected IPIs in the 30-sec interval were rejected by the algorithm the output of the algorithm was an "undefined rhythm" label. For measurement intervals during which events of arrhythmias were detected by this algorithm, the QI was set to 0. To summarize the PPG signal coverage, each HR measurement was assigned to one of three categories: i) good quality (QI = 4), ii) low quality (QI $\leq$ 3), iii) arrhythmia. Only HR data associated with QI = 4 were used in the agreement analysis. Coverage was measured as the ratio between the measurements with good quality and the entire measurement duration for a patient. If patients had less than 5 minutes of coverage during surgery or recovery, the session was excluded from analysis. The hospital health records were screened to find potential causes for patients that were excluded, since this would indicate that the OHRM was not usable for these patients. Bland-Altman plots were made to visualize the agreement between ECG and PPG HR [27]. Limits of agreement (LoA) and confidence intervals of the LoA were calculated by taking into account both within and between patient variability [28]. The modified method of Bland and Altman to estimate LoA with repeated measurements where the true value varies, as described by Zou was used [29]. The confidence intervals of the LoA were constructed using the method of variance estimates recovery (MOVER). In short, a one-way random effects model was used to model the difference d_{ij} of the j-th measurement for the i-th patient as:

$$d_{ij} = d + a_i + e_{ij}, \tag{7.2.1}$$

Where d is the unknown true difference between ECG and PPG HR. The difference d is either the difference between the PPG and ECG HR, i.e. $d = HR_{PPG} - HR_{ECG}$, or the percentage difference calculated by $d = d_{\%} = \frac{HR_{PPG}-HR_{ECG}}{HR_{mean}}$. a_i and e_{ij} are zero-mean normally distributed with variance σ_b^2 and σ_{dw}^2 corresponding to the true between and within subject variances respectively. The bias is estimated by $\hat{d}_{..}$ where $\hat{d}_{..} = \frac{\sum_i \hat{d}_i}{m_i}$ and $\hat{d}_i = \frac{\sum d_{ij}}{m_i}$ and m_i is the number of pairs per patient i. The between and within subject variances are estimated by $s_b^2 = \frac{\left(\sum_i \hat{d}_i - \hat{d}_{..}\right)^2}{n-1}$ and $s_{dw}^2 = \sum_i \frac{m_i-1}{N-n} s_i^2$ where $s_i^2 = \frac{\left(\sum_j (d_{ij}-\hat{d}_i)\right)^2}{m_i-1}$. s_b^2 and s_{dw}^2 are summed to obtain an estimate of the total variance s_{tot}^2. The 95% LoA are then calculated by: $\hat{d}_{..} \pm 1.96\sqrt{s_{tot}^2}$. Confidence intervals around the LoA are estimated by the MOVER [29]. The Bland-Altman analysis was done for both the absolute difference, as well as the percentage difference in HR between PPG and ECG. The HR evaluation was compared to the reference standard ANSI/AAMI EC13:2002, which requires an accuracy of ±5 beats per minute (bpm) or ±10% (whichever is largest)[30].

7.3 Results

A total of 100 patients were included. One patient was excluded because the patient monitor data was missing due to technical difficulties. Recovery data of one patient was missing because this patient was transferred to the intensive care unit (ICU) immediately after surgery. Three patients had too few (<5 minutes) good quality photopletysmography (PPG) measurements during both the surgery and recovery phase and were therefore omitted from the agreement analysis. Twelve patients had <5 minutes of good quality measurements during either the surgery or recovery phase, only the respective phase was omitted from the agreement analysis. Patient demographics are shown in Table 7.3.1.

An example of the data that was captured for each patient in the study is shown in Fig. 7.3.1. In total 159.08 hours of data was captured during surgery, of which 76.5% was of good quality (quality index (QI) = 4) and 112.59 hours of data was captured during recovery, of which 74.4% was of good quality.

Coverage varied between patients, see Fig. 7.3.2. Median coverage was 86% (interquartile range (IQR): 65% to 95%) and did not differ significantly between surgery and recovery (Wilcoxon paired difference test p = 0.17). Coverage statistics are shown in Table 7.3.2.

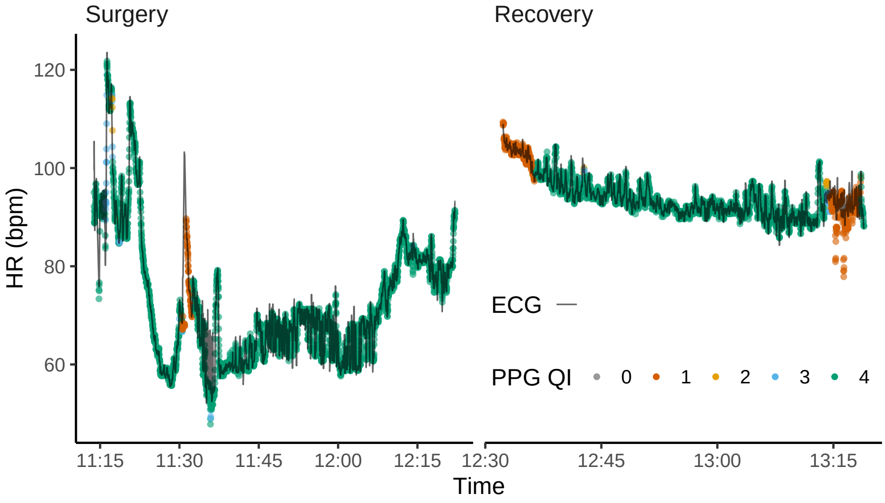


Figure 7.3.1: Example of data captured for a representative patient in the study.
The ECG signal is represented by the gray line and the individual PPG measurements by the colored points. The QI of the PPG signal is represented by a different color which ranges from 0 (lowest quality) to 4 (highest quality). Beats per minute (bpm); electrocardiogram (ECG): heart rate (HR); photopletysmography (PPG); quality index (QI).

7.3.1 Blant Altman analysis during surgery

The mean bias was -0.15 (± 0.05) beats per minute (bpm) and -0.20 (± 0.06) % for the PPG measured heart rate (HR) compared to the ECG measured HR, where the limits of agreement (LoA) (including the standard errors) fall within the reference standard of ±5 bpm and ±10%, see Table 7.3.3.

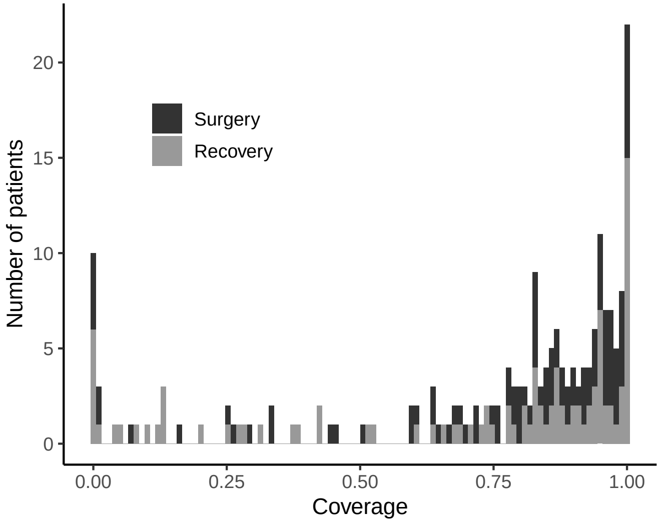


Figure 7.3.2: Histogram with distribution of coverage fraction (ie, proportion of recorded data that corresponds to a photopletysmography signal with good quality).

7.3.2 Bland Altman analysis during recovery

The mean bias was -0.10 (± 0.04) bpm and -0.14 (± 0.04) % for the PPG measured HR compared to the ECG measured HR, where the LoA (including the standard errors) fall within the reference standard of ±5 bpm and ±10 %, see Table 7.3.4.

	Overall
N	99
Age in years (median [IQR])	58.0 [44.5, 68.0]
Male gender	36
BMI kg/m^2 (median [IQR])	28.7 [24.8, 37.1]
ASA-PS score	
I	10
II	39
III	45
IV	5
Surgical risk	
High	9
Intermediate	63
Low	27
Diabetes	7
Hypertension	37
Hypercholesterolemia	21
Previous stroke or TIA	13
Structural heart disease	8
Atrial fibrillation	8
Blood pressure sensor	88
Wrist device location	
Left	45
Right	53
Unknown	1
Surgery type	
Bariatric surgery	22
Gastroenterological surgery	8
Neurosurgery	3
Orthopedic surgery	31
Plastic surgery	7
Thyroid surgery	1
Urogenital surgery	17
Vascular surgery	10
Surgery duration (min.) (median [IQR])	87.0 [48.0, 115.0]
Recovery duration (min.) (median [IQR])	58.0 [41.2, 78.0]

Table 7.3.1: Patient demographics.
Continuous variables are summarized by median and interquartile range (IQR). Body Mass Index (BMI), American Society of Anesthesiologists physical status (ASA-PS), transient ischemic attack (TIA).

	Surgery	Recovery	Surgery and Recovery
Total hours	159.6	112.2	271.8
Good quality PPG	124.1 h (78.0 %)	83.8 h (74.4 %)	207.9 h (76.5 %)
Low quality PPG	33.3 h (21.0 %)	28.7 h (25.5 %)	62.0 h (22.8 %)
Arrhythmia	1.7 h (1.1%)	0.2 h (0.2 %)	1.9 h (0.7 %)

Table 7.3.2: Coverage statistics of total hours for analyses including all patients.

	Difference in bpm		Difference in %	
Bias, mean of differences (± SE)	-0.15 bpm	(± 0.05)	-0.20 %	(± 0.06)
SD of differences	1.50 bpm		2.34 %	
Lower LoA (95% CI)	-3.08 bpm	(-2.99, -3.19)	-4.79 %	(-4.92, -4.66)
Upper LoA (95% CI)	2.79 bpm	(2.69, 2.89)	4.39 %	(4.26, 4.53)
Within-subject variance	2.04 bpm		5.12 %	
Between-subject variance	0.20 bpm		0.37 %	
Intraclass correlation coefficient	0.09		0.07	

Table 7.3.3: Bland-Altman analysis results during surgery. Beats per minute (bpm); standard error (SE); standard deviation (SD); limits of agreement (LoA); confidence interval (CI).

	Difference in bpm		Difference in %	
Bias, mean of differences (± SE)	-0.10 bpm	(± 0.04)	-0.14 %	(± 0.04)
SD of differences	1.38 bpm		1.93 %	
Lower LoA (95% CI)	-2.80 bpm	(-2.72, -2.87)	-3.92 %	(-3.83, -4.01)
Upper LoA (95% CI)	2.59 bpm	(2.52, 2.67)	3.64 %	(3.56, 3.74)
Within-subject variance	1.78 bpm		3.56 %	
Between-subject variance	0.11 bpm		0.16 %	
Intraclass correlation coefficient	0.06		0.04	

Table 7.3.4: Bland-Altman analysis results during recovery. Beats per minute (bpm); standard error (SE); standard deviation (SD); limits of agreement (LoA); confidence interval (CI).

7.4 Discussion

A wrist-worn optical heart rate monitor (OHRM) may be able to provide continuous unobtrusive heart rate (HR) monitoring in the low acuity care or home settings. To determine this, the validity of OHRM-derived HR must first be assessed in a representative target population and compared to the gold standard 5-lead electrocardiogram (ECG). In this study, the agreement between the HR derived from an OHRM and a 5-lead ECG connected to a patient monitor was assessed for a representative patient population during the perioperative period. The OHRM could provide an accurate HR (–5 beats per minute (bpm) to 5 bpm and –10% to 10% compared to the ECG-derived HR) during both the surgical and recovery phase when the photopletysmography (PPG) signal was of good quality. A vast majority (76.5%) of the PPG signal was good quality.

Given the hemodynamic changes during the perioperative period and the diversity in surgical procedures, a technical validation, as performed in this study, is essential before the OHRM can be introduced into clinical practice. Very few studies were found in the literature that validated wrist-worn OHRMs in hospitalized patients. One study, with a goal of early warning detection using an OHRM, was performed in patients during and after discharge from the intensive care unit (ICU) [13]. The OHRM was a personal fitness tracker, and 24 hours of monitoring started in the ICU while patients were still being monitored by means of a continuous ECG. The authors concluded that personal fitness tracker–derived HRs were slightly lower than those derived from continuous ECG monitoring and not as accurate as pulse oximetry-derived HRs. A feasibility study was performed by the same research group regarding bradycardia and tachycardia detection in the same population [18]. The authors stressed in both studies the importance of subgroup analysis of patients not in sinus rhythm since this negatively impacted measurement accuracy. This corresponds to the findings in our study where measurements during arrhythmia were of low quality.

Another study was designed for atrial fibrillation (AF) detection, but also showed good results in sinus rhythm in patients undergoing elective cardioversion for AF [31]. There were fewer patients (N = 20) included than in our study, and the agreement analysis was based on QRS intervals as the reference, with a mean difference of 1.3 ms being found between ECG and PPG. Other studies were performed in healthy participants and focused on assessing accuracy during physical activity [11, 12, 14, 16, 17, 32–34]. However, the results obtained in these studies cannot be translated to our results since surgery was the underlying cause for changes in HR in our study and not physical activity. Factors influencing HR during surgery are hemodynamic changes induced by anesthesia, intraoperative factors such as blood loss and hypothermia, or involvement of vital organs in the area of surgery. Results of previous studies did conclude that motion artifacts remain a challenge in OHRMs. In this study, motion artifacts were less likely to occur since patients were mostly immobilized. Nevertheless, motion artifacts are relevant to consider if the OHRM is to be used in the future for remote monitoring of patients.

The agreement between the ECG- and PPG-derived HR was within the limits of agreement (LoA) of –5 bpm to 5 bpm and –10% to 10% (whichever was largest) both during surgery and recovery. However, this only applied when the quality of the PPG signal was labeled as “good”. Nevertheless, a vast majority (during surgery 78.0%; during recovery 74.4%) of the PPG signal was good quality. Ideally, the coverage should be 100%, but this may not be realistic since a poor signal-to-noise ratio in the PPG measurements can perturb the detection of a sinus rhythm. Arrhythmias such as ectopic beats, AF, premature ventricular or atrial complex, and paced beats also contributed to a reduction of measurement coverage of the OHRM. This is confirmed by the fact that patients with a medical history of AF had lower overall coverage compared to patients without previous diagnosis of AF, resulting in 25% versus 85% overall coverage, respectively. This was also true for those patients with severe congenital heart disease where median coverage was 47% versus 85% for patients without structural heart disease. Finally, a very small group of patients had an extremely low coverage, but a consequently large influence on the mean coverage. Median coverage was higher, with 85% being good quality data. Furthermore, 3 patients were excluded since <5 minutes of data were captured in total, which could be explained in one case by serious congenital heart disease which involved aberrant anatomy. Another 12 patients with <5 minutes of good quality data during surgery or recovery were also excluded. The gold standard ECG, is considered capable of providing 100% coverage. However, in clinical practice, this is most likely not the case since ECG HR detection can also fail in the presence of the aforementioned abnormalities.

The limitations of this study are the following. Despite a heterogeneous group of elective procedures and hospital setting, no general ward patients were included. Nevertheless, translation of our findings to patients in the general ward is reasonable as patients are transitioning from immobile to a more mobile state during stay in the recovery room. By using a 1-way random-effects model, the between- and within-patient variance was quantified to explore the effect of heterogeneity of the study group. As indicated by Hamilton and Lewis [35], not accounting for repeated measures can lead to a falsely narrow LoA, mainly with a small number of patients and a large number of measurements per patient. Both the mean bias and between-patient variance are weighted according to the number of observations, available for each patient. Hence, patients with more observations will contribute more to the final results. As the distribution of observation times was skewed, some patients contributed substantially more than others. Therefore, results could have been biased to these patients. It is also worth mentioning the assumptions underlying the 1-way random-effects model. Specifically, the model assumes that repeated differences on a single patient are independent and that the within-patient variance of these differences is constant and the same for all patients. First, the independence assumption could have been too strong since hemodynamic changes occurred during surgery or recovery which could have led to autocorrelation in the HR and subsequent differences arising between the PPG- and ECG-derived HR. The effect of autocorrelation on the within-patient variance is unknown, and further studies are needed to take autocorrelation into account [28]. Second, the assumption of homoscedasticity was not formally tested, and it could have been the case that the variance of the differences increased with higher HR. Finally, the possible influence of surgery-specific factors, such as electrosurgical instruments causing interference was not investigated. Furthermore, we observed that the oscillometric blood pressure cuff can interfere with the measurements of the OHRM by compromising the blood flow.

In summary, the current study found that the OHRM is clinically acceptable when good quality data are captured and in settings when high-intensity monitoring, such as in the ICU or operating room, is not mandatory. The OHRM seems less suitable for patients with congenital anatomical changes of the heart or patients with arrhythmias. When the OHRM captures a significant amount of low-quality data in a patient, the suggestion would be to use another monitoring type to ensure reliable monitoring. Since the OHRM can report the quality of the PPG signal instantaneously, the decision to switch to ECG monitoring can be made on the spot. The reliability of an OHRM to measure HR in patients known to suffer from arrhythmias or structural heart disease requires further research.

7.5 Acknowledgments

We would like to express our gratitude for the valuable help of Marleen van Aartrijk in the design and collection of data for this study. This research was funded by Rijksdienst voor Ondernemend Nederland, Ewatch (Grant no. ITEA 161006).

References

1. Cardona-Morrell, M., Prgomet, M., Turner, R., Nicholson, M. & Hillman, K. Effectiveness of continuous or intermittent vital signs monitoring in preventing adverse events on general wards: a systematic review and meta-analysis. *International journal of clinical practice* **70,** 806–824 (2016).

2. Ahmad, T. *et al.* Use of failure-to-rescue to identify international variation in postoperative care in low-, middle-and high-income countries: a 7-day cohort study of elective surgery. *BJA: British Journal of Anaesthesia* **119,** 258–266 (2017).

3. Noordzij, P. G. *et al.* Postoperative mortality in The Netherlands: a population-based analysis of surgery-specific risk in adults. *The Journal of the American Society of Anesthesiologists* **112,** 1105–1115 (2010).

4. Walston, J. M. *et al.* Vital signs predict rapid-response team activation within twelve hours of emergency department admission. *Western Journal of Emergency Medicine* **17,** 324 (2016).

5. Fieselmann, J. F., Hendryx, M. S., Helms, C. M. & Wakefield, D. S. Respiratory rate predicts cardiopulmonary arrest for internal medicine inpatients. *Journal of general internal medicine* **8,** 354–360 (1993).

6. Taenzer, A. H., Pyke, J. B., McGrath, S. P. & Blike, G. T. Impact of pulse oximetry surveillance on rescue events and intensive care unit transfers: a before-and-after concurrence study. *The Journal of the American Society of Anesthesiologists* **112,** 282–287 (2010).

7. Suaste-Gómez, E., Hernández-Rivera, D., Sánchez-Sánchez, A. S. & Villarreal-Calva, E. Electrically insulated sensing of respiratory rate and heartbeat using optical fibers. *Sensors* **14,** 21523–21534 (2014).

8. Vadrevu, S. & Manikandan, M. S. A robust pulse onset and peak detection method for automated PPG signal analysis system. *IEEE Transactions on Instrumentation and Measurement* **68,** 807–817 (2018).

9. Chang, K.-M., Chang, K.-M., *et al.* Pulse rate derivation and its correlation with heart rate. *J Med Biol Eng* **29,** 132–7 (2009).

10. Sartor, F., Papini, G., Cox, L. G. E., Cleland, J., *et al.* Methodological shortcomings of wrist-worn heart rate monitors validations. *Journal of medical Internet research* **20,** e10108 (2018).

11. Valenti, G. & Westerterp, K. R. *Optical heart rate monitoring module validation study* in *2013 IEEE international conference on consumer electronics (ICCE)* (2013), 195–196.

12. Gillinov, S. *et al.* Variable accuracy of wearable heart rate monitors during aerobic exercise. *Med Sci Sports Exerc* **49,** 1697–1703 (2017).

13. Kroll, R. R., Boyd, J. G. & Maslove, D. M. Accuracy of a wrist-worn wearable device for monitoring heart rates in hospital inpatients: a prospective observational study. *Journal of medical Internet research* **18,** e6025 (2016).

14. Delgado-Gonzalo, R. *et al. Evaluation of accuracy and reliability of PulseOn optical heart rate monitoring device* in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2015), 430–433.

15. Parak, J. & Korhonen, I. *Evaluation of wearable consumer heart rate monitors based on photopletysmography* in *2014 36th annual international conference of the IEEE engineering in medicine and biology society* (2014), 3670–3673.

16. Shcherbina, A. *et al.* Accuracy in wrist-worn, sensor-based measurements of heart rate and energy expenditure in a diverse cohort. *Journal of personalized medicine* **7,** 3 (2017).

17. Spierer, D. K., Rosen, Z., Litman, L. L. & Fujii, K. Validation of photoplethysmography as a method to detect heart rate during rest and exercise. *Journal of medical engineering & technology* **39,** 264–271 (2015).

18. Kroll, R. R. *et al.* Use of wearable devices for post-discharge monitoring of ICU patients: a feasibility study. *Journal of intensive care* **5,** 1–8 (2017).

19. Hochstadt, A. *et al.* Continuous heart rhythm monitoring using mobile photoplethysmography in ambulatory patients. *Journal of Electrocardiology* **60,** 138–141 (2020).

20. Lord, J. M. *et al.* The systemic immune response to trauma: an overview of pathophysiology and treatment. *The Lancet* **384,** 1455–1465 (2014).

21. Manou-Stathopoulou, V., Korbonits, M. & Ackland, G. L. Redefining the perioperative stress response: a narrative review. *British journal of anaesthesia* **123,** 570–583 (2019).

22. Doyle, D. J., Goyal, A. & Garmon, E. H. *American Society of Anesthesiologists Classification* (StatPearls Publishing, Treasure Island (FL), 2017).

23. Kristensen, S. D. *et al.* 2014 ESC/ESA Guidelines on non-cardiac surgery: cardiovascular assessment and management: The Joint Task Force on non-cardiac surgery: cardiovascular assessment and management of the European Society of Cardiology (ESC) and the European Society of Anaesthesiology (ESA). *European heart journal* **35,** 2383–2431 (2014).

24. Papini, G. B. *et al.* Wearable monitoring of sleep-disordered breathing: Estimation of the apnea–hypopnea index using wrist-worn reflective photoplethysmography. *Scientific reports* **10,** 1–15 (2020).

25. Kathirvel, P., Sabarimalai Manikandan, M., Prasanna, S. & Soman, K. An efficient R-peak detection based on new nonlinear transformation and first-order Gaussian differentiator. *Cardiovascular Engineering and Technology* **2,** 408–425 (2011).

26. Bonomi, A. G. *et al.* Atrial fibrillation detection using a novel cardiac ambulatory monitor based on photo-plethysmography at the wrist. *Journal of the American Heart Association* **7,** e009351 (2018).

27. Bland, J. M. & Altman, D. G. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* **327,** 307–310 (1986).

28. Bland, J. M. & Altman, D. G. Agreement between methods of measurement with multiple observations per individual. *Journal of biopharmaceutical statistics* **17,** 571–582 (2007).

29. Zou, G. Confidence interval estimation for the Bland–Altman limits of agreement with multiple observations per individual. *Statistical methods in medical research* **22,** 630–642 (2013).

30. Association for the Advancement of Medical Instrumentation. *ANSI/AAMI EC13:2002 Cardiac monitors, heart rate meters, and alarms* tech. rep. (American National Standards Institute, Inc., 2002).

31. Hochstadt, A. *et al.* Continuous heart rate monitoring for automatic detection of atrial fibrillation with novel bio-sensing technology. *Journal of electrocardiology* **52,** 23–27 (2019).

32. Dobbs, W. C. *et al.* The accuracy of acquiring heart rate variability from portable devices: a systematic review and meta-analysis. *Sports Medicine* **49,** 417–435 (2019).

33. Wallen, M. P., Gomersall, S. R., Keating, S. E., Wisløff, U. & Coombes, J. S. Accuracy of heart rate watches: implications for weight management. *PloS one* **11,** e0154420 (2016).

34. Wang, R. *et al.* Accuracy of wrist-worn heart rate monitors. *Jama cardiology* **2,** 104–106 (2017).

35. Hamilton, C. & Lewis, S. The importance of using the correct bounds on the Bland–Altman limits of agreement when multiple measurements are recorded per patient. *Journal of clinical monitoring and computing* **24,** 173–175 (2010).

8
Discussion

The goal of this thesis is to examine the potential and pitfalls in the development, validation and implementation of clinical prediction models (CPMs) based on real-world longitudinal data. Below these are discussed for real-world data (RWD) and CPMs, respectively. The discussion is followed by the conclusion of this thesis and future perspectives.

8.1 Real-world data

As defined in the introduction, real-world data (RWD) is considered observational data relating to patient health status and/or the delivery of health care, generated as part of a healthcare process. RWD can be regarded as all data that are generated outside of clinical trials (both interventional and observational). A common source of RWD for clinical research are electronic health record (EHR)s.

8.1.1 Accessibility, volume and speed

The first and most obvious potential of using RWD is that it comes without the high costs, resources and often long time-spans that clinical trials require, there is instant access to a large and historic patient population. EHRs exceed many existing registries and repositories in volume, range of measurements and outcomes. Clinical trials can be limited in sample size and follow-up as they are costly and time-consuming to conduct [1, 2]. An example of this potential was demonstrated during the outbreak of the coronavirus pandemic, when a fast response was required. It would have taken time for the first clinical trials to be approved and finished, whilst RWD obtained from EHRs and the laboratory information system (LIS) was readily available. In this scenario RWD enabled a quick response to an immediate problem.

8.1.2 Consent bias versus selection bias

Before patients can take part in a clinical trial, inclusion criteria have to be fulfilled and informed consent must be signed. Inclusion criteria are generally more of a limitation in a randomized controlled trial (RCT) than a cohort study, the generalizability of RCT patient samples to the full population is a long standing issue [3, 4]. Even so, in cohort studies, patients have to sign informed consent and are subjected to extra tests such as blood sampling or filling out questionnaires. If consent is required, there is a potential for consent bias (also known as authorisation bias or volunteer bias), where those who consent differ in measured or unmeasured baseline characteristics from those who do or cannot consent [5–7]. In a large systematic review by Kho et. al significant differences between participants and non-participants were found, but no clear direction or magnitude of the effect [5]. However, what can be gained by using RWD to prevent consent bias, can more easily be lost by introducing selection bias. Selection bias, is bias as a result of a systematic difference in the population selected by the researcher for inclusion in the analysis, and the target population. Time can introduce selection bias as patient populations, treatments and diseases change over time. An example of this phenomenon was observed during the coronavirus pandemic. Different variants and the roll-out of vaccines caused changes in the patient population presenting at the emergency department (ED) over time. The CoLab-score developed in Chapter 4 was developed during the first wave of the pandemic and subsequently implemented in multiple hospitals in the Netherlands. In the temporal validation the performance of the score was preserved, even during the emergence of new variants and the roll-out of vaccines. Nevertheless, it should be taken into account that the time-frame in which patients are included can lead to selection bias and performance should be monitored continuously after implementation.

8.1.3 Informative missingness

Directly related to selection bias, is informative missingness. Missing data is common in EHRs. Studies have shown that missingness in EHRs is often informative, this implies that the presence or absence of data carries information in itself about the health status of a patient [8–10]. This is not a bug but a feature, EHRs are observational in nature and data is entered for a different purpose than research. This is especially true for laboratory tests, since these are the result of a clinicians judgement. It is known that laboratory tests are ordered more frequently for seriously ill patients [11]. In a study by Agniel et al. it was demonstrated that the hour of the day a laboratory test was ordered, the day of the week and the amount of time between tests was more predictive of survival than the actual value of the test result itself [8]. Also, depending on the expertise of the center, more seriously ill patients can be transferred to specialized centers, which is also a case of informative missingness. In Chapters 2 and 3 cardiac troponin (cTn)T was sampled repeatedly after cardiac surgery to detect the presence of a perioperative myocardial infarction (PMI). In clinical practice however, clinicians will stop sampling cTnT if the patient is stable and values are falling. In this case, the presence or absence of a cTnT measurement at a certain time, can be regarded as informative missingess. For our study (Chapters 2 and 3), it was ensured that cTnT was sampled at prescribed points in time, even if values were falling, to prevent informative missingness. The phenomenon of longitudinal data 'thinning' over time can result from informative missingness.

8.1.4 Data quality

The quality of RWD, and EHRs in particular, is not guaranteed. As stated previously, this is a result of the fact that EHR data is collected during clinical practice by clinicians who are otherwise busy, rather than purposefully collected research data. Several data quality issues can arise.

Inaccurate or incorrect data

While it is difficult to draw specific conclusions with respect to EHR data quality, in a large review it was shown that mainly medication lists are prone to significant errors, whereas laboratory test results appear to exhibit greater accuracy than other types of data [12]. When developing the CoLab-score (Chapter 4), the hospital registration of a confirmed coronavirus disease 2019 (COVID-19) case was used (together with polymerase chain reaction (PCR)-testing results) to label patients as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) positive or negative. However, clinicians sometimes erroneously registered a confirmed COVID-19 infection in a patient. This registration would then quickly be removed when the error was discovered. Simply using a registration without setting a minimal time between start and end of a registration would result in including these mis-registrations.

Lack of gold standard

Outcomes registered in RWD or EHRs may differ from the clinical gold standard. In Chapter 2, at the time the study was performed, clinicians used aspartate aminotransferase (ASAT) activity instead of cTnT levels to determine the presence of a PMI. Therefore, the clinical gold standard differed from the guideline gold standard for diagnosing a PMI [13]. The lack of a gold standard was addressed in Chapter 2 by using a probabilistic model-based clustering approach, this approach did not make any *a priori* assumptions regarding patient diagnoses. Therefore, the characteristic cTnT release profiles that were found, were not affected by a subjective center-specific way of diagnosing a PMI, but rather by the variation in release profiles themselves. This demonstrates that clustering based approaches are a good alternative when a clinical gold standard is lacking or absent.

Loss of data

Multiple sources of data feed into EHRs, the source data can however differ from the data that is stored in the EHR. Again this is intended behaviour as EHRs should display information relevant to the clinician. The source data can e.g. be measured with several decimals precision but if these are not relevant for clinical decision making, the values are rounded when reported to the EHR. This can be illustrated by an example from Chapter 4. The CoLab-score is developed using laboratory test results obtained from the LIS. The LIS contains the raw measured values of all laboratory tests, whereas the values reported to the EHR are truncated above or below certain clinically relevant boundaries. However, the predictive value in separating two patients groups (with/without SARS-CoV-2 infection) is mostly concentrated in the range below the reported (truncated) value, see Fig. 8.1.1. Not taking the source data into account can lead to a loss of predictive performance.

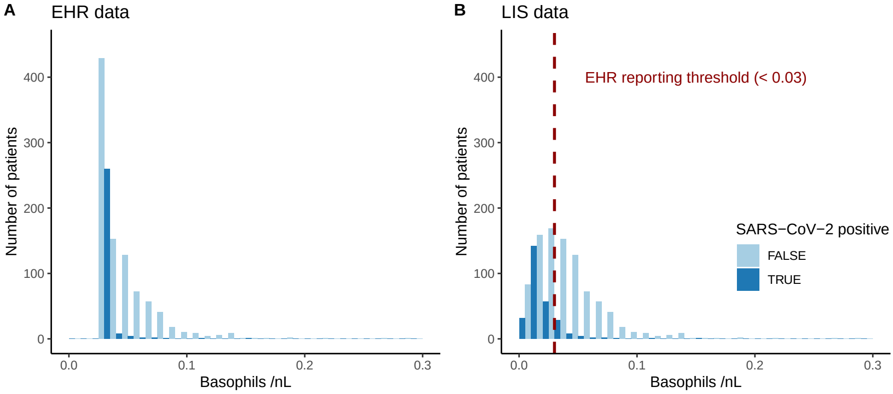


Figure 8.1.1: Histogram of absolute basophil count for patients tested for SARS-CoV-2 at the Catharina Hospital.
A: truncated values as stored in the EHR.
B: raw values as stored in the LIS.
As can be seen from the distributions, the predictive value of basophils lies in the lower range < 0.03, whereas the EHR stores only values ≥ 0.03. The discriminative ability as expressed in the AUC is 0.843 for the basophil count from the LIS and 0.741 from the EHR.

Source data can change over time

RWD data can be structured or unstructured. Structured data is organized by a pre-defined data model and usually stored in a database. Unstructured data can be considered qualitative data and does not have a pre-defined data model. Although a great part of data in EHRs is in the form of structured data, there is no guarantee that the structure is the same over time. Multiple columns or tables can cover the same variable and they have to be placed in the context of the current state of the healthcare process. For example in developing the CoLab-score Chapter 4, shortly after the first positive cases were detected, the PCR testing machine was replaced. The PCR-test results produced by the new machine had a different structure and were stored in a different column. Not taking changes in structured data over time into account can lead to a loss of information. Moreover, in the absence of a gold standard, as described previously, the diagnostic criteria can (and most likely will) change over time. As a consequence, a negative diagnosis made in the past, can be a positive diagnosis in the present.

8.1.5 Mobile health

In clinical practice, measurements are commonly performed during admission and at brief intermittent visits to the clinic. However, in the case of a chronic condition, the use of mobile health (mHealth) devices can provide a more comprehensive collection of data to inform clinical decision making. An example is the wrist-worn optical heart rate monitor (OHRM) validated in Chapter 7. Although many studies have been performed with wearables, these are usually performed on healthy subjects or ward patients [14, 15]. In our study we included a heterogeneous patient population undergoing surgery and showed that the accuracy of the OHRM fell within the reference standard of the patient monitor. This shows the potential for wearables to provide accurate data for monitoring patients over time.

8.2 Clinical prediction models

Clinical prediction models (CPMs) can be used either in public health (e.g. prediction of disease prevalence), clinical practice (e.g. for diagnosis or therapeutic decision-making) or research (e.g. selecting high risk patients for inclusion in a randomized controlled trial (RCT) or adjusting for covariates and confounding) [16]. We focus on the potential of CPMs in clinical practice. CPMs combine a set of predictors to predict an outcome, the outcome can either be diagnostic (e.g. does the patient have the disease) or prognostic (e.g. what is the expected time until mortality).

8.2.1 Sample size

Developing a CPM with a relatively small sample size has proven to be difficult. The sample size (n) should always be considered in relation to the number of predictors or dimension (p). If n is small in relation to p overfitting can be a serious problem [16, 17]. To prevent this pitfall, a sample size calculation should be preformed before starting data collection. If the sample size is fixed, penalization and shrinkage methods can be tried to minimize the problem of overfitting [17, 18]. Note that penalization and shrinkage methods are not a free pass to develop CPMs without regard for the sample size [19]. Note also that the sample size does not cover the whole story, but rather the number of events, otherwise expressed as the number of events per variable [20]. For example in the case of a binary outcome, this is the minimum of the number of patients who experienced the outcome and those who did not experience the outcome. Using repeated measures can be of benefit to studies with small sample sizes. This can be illustrated by Chapter 3, where only 23 patients experience the outcome of a perioperative myocardial infarction (PMI). Although there are only 23 patients with the outcome, mixed effects models can 'borrow strength' over patients to improve individual patient estimates. Therefore, using the appropriate modeling techniques, time can used to improve individual estimates when the sample size is small.

8.2.2 Choosing predictors

For a CPM to be successful it must be implementable in other centers with relative ease. Should the sample size allow it, developing a CPM that requires e.g. 30 predictors is possible. Nevertheless, this implies that other centers have to measure all 30 predictors to be able to use the model. Cost effectiveness of a CPM will also have to be taken into account, a CPM should not lead to increased healthcare costs. Therefore, the costs of measuring the variables required for the CPM to make a prediction, have to be compared to the current clinical practice. After development, the costs of false negatives versus false positives also need to be taken into account in determining a suitable threshold. Limiting the number of predictors can be achieved by applying feature selection techniques. This also holds for including ”exotic” predictors. An example from Chapter 4 is that many CPMs for diagnosing severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) were developed requiring e.g. features from chest CT-scans/X-rays [21]. This will limit the usability of such models as it requires a chest scan in order for the model to make a prediction. Finally, the time required to measure a predictor can also play an important role. If predictions have to be made within a certain amount of time, then the ”slowest” predictor is the limiting factor. In the case of the CoLab-score (Chapters 4 and 5) all blood tests required to calculate the score, are available at the same time (less than 1 hour after presentation). This also proved the motivation for Chapter 5, since the CoLab-score was quickly available, healthcare workers with low probability of a SARS-CoV-2 infection could rapidly return to work.

8.2.3 Dichotomization

For binary classification models such as logistic regression, the outcome is dichotomous. Yet, outcomes that are registered as binary in clinical practice, are sometimes strictly speaking, continuous. An example can be found in Chapters 2 and 3, where the outcome of a PMI is either present or absent. However, the reality is more complex, it involves an area of heart muscle that is affected and a sufficiently large area is considered a PMI [22]. In practice, the area of necrosis is never quantified and only cases fulfilling the diagnostic criteria are labeled as positive, hence, a clearly defined gold standard is essential.

8.2.4 Calibration

In CPMs, patients are assigned probabilities of having an outcome. Since CPMs are used as decision support, it is important that these probabilities are accurate. If a model predicts a 30% probability of an outcome, in the long run 30% of the population should experience the outcome. Models can have good discriminative ability but if they have poor calibration they are potentially harmful for clinical decision-making [23].

Calibration drift

Commonly, CPMs are developed only once when they are fitted to the development dataset, after validation they are implemented until they are either revised or become obsolete. Before CPMs become revised or obsolete, there is usually a degradation of performance over time due to changes in the disease, treatment and/or underlying patient population. This is referred to as calibration drift [24, 25]. Calibration drift is a major issue for diagnostic CPMs for coronavirus disease 2019 (COVID-19), given that the prevalence of COVID-19 can change rapidly, resulting in miscalibrated probabilities during peaks of high or low incidence. Little to no attention has been given to this limitation by published CPMs. In developing the CoLab-score (Chapter 4) this limitation was recognized and addressed by introducing the concept of a prevalence-dependent threshold above which a score should be considered positive. Until more refined approaches are developed and implemented, the approach of a prevalence-dependent threshold preserves the performance of the CoLab-score over time. In this respect, there is potential for models that signal when calibration drift occurs or dynamically adjust regression coefficients [24, 25].

Analytical variation

Next to changes in incidence or prevalence, analytical variation is also a cause of miscalibration when using laboratory test results as predictors. In the external validation of the CoLab-score (Chapter 4), systematic differences in the reported albumin values were observed between centers, see Fig. 8.2.1. Albumin is measured by a dye-binding assay using bromocresol (BC) green or BC purple. In the development center and centers 2 and 3 the BC green method is used, whereas center 1 uses the BC purple method. From literature it is known that BC green results in higher values than BC purple [26]. For the CoLab-score an albumin conversion factor was calculated for center 1 to avoid miscalibration as a result of analytical variation. This demonstrates that analytical variation must be taken into account when assessing calibration of

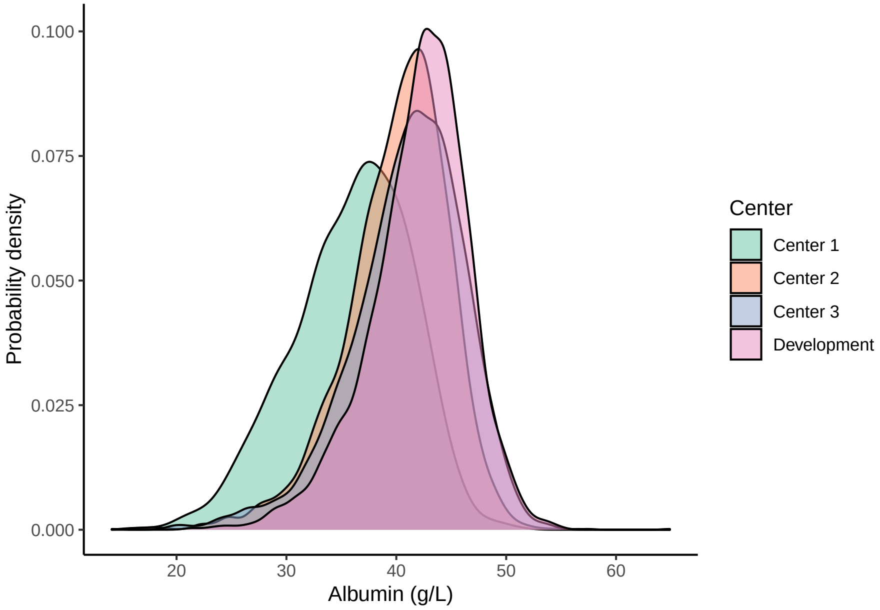


Figure 8.2.1: Smoothed density estimates of albumin levels of patients presenting at the emergency department for the development dataset of the CoLab-score, and three external centers.

CPMs. Analytical variation is also be a concern for miscalibration over time, as laboratory equipment has a finite life-span and is replaced over time.

8.2.5 Validation and implementation

To take overfitting into account, reported performance measures should be calculated using internal validation by e.g. cross-validation or bootstrapping [16, 17]. Next to internal validation, external validation is required. In the absence of validation, reported performance measures are most likely overly optimistic [27]. An example of the external validation of a time-to-event model is given in Chapter 6.

After internal and external validation, a CPM can be implemented in daily clinical practice. The predictions from the model can either be used assistive, in which case the probabilities are reported to the clinician without further recommendations. An alternative approach is presenting the prediction as a decision recommendation, which is referred to as an directive approach. Although evidence is scarce, studies suggest that a directive approach has a greater impact on decision making [28]. For a directive approach one has to consider which decisions are recommended for which probability thresholds. CPMs can be combined with decision theory to weigh the risks, benefits and cost-effectiveness of interventions (e.g. treatment or surgery) and make a decision if the benefits of the intervention outweigh the risks.

The European coLlaboration on Acute decompeNsated Heart Failure (ELAN-HF) score, externally validated in Chapter 6, is used in an assistive manner. The ELAN-HF score is reported in the electronic health record (EHR) of a patient upon visiting the heart failure clinic. High-risk patients are an indication for the heart failure nurse that these patients may require more intensive monitoring and guidance to prevent a readmission or mortality.

The CoLab-score developed in Chapter 4 is presented to clinicians in a directive approach. This was achieved by choosing a *number willing to test* to determine a cutoff above which a clinician should be advised to request a polymerase chain reaction (PCR)-test. If this number was too high then this would result in increased costs since more PCR-tests would be requested, a number too low would result in too many positive patients being missed. Based on data from routine clinical practice, it was inferred that clinicians on average request 15 PCR-tests to find one positive patient. The threshold for the CoLab-score was therefore based on routine clinical practice and taking cost-effectiveness into account. Also in this case, time plays an important role since associated costs often change over time, shifting the decision threshold required to maintain cost-effectiveness.

8.2.6 Longitudinal data

Finally, we discuss the incorporation of time into CPMs itself. Longitudinal data has the potential to improve predictive performance and make CPMs more personalized. However, several studies have shown that incorporation of longitudinal data in CPMs is lagging [29, 30]. In a systematic review by Goldstein et al. it was shown that in 60% of the cases where EHR data was used to develop a CPM, longitudinal data was not incorporated [29]. Moreover, for those studies that did incorporate longitudinal data, most studies resorted to using summary measures such as the peak value, mean/median or slope, leading to a loss of information [29, 30]. In Chapter 3 we demonstrate how non-parametric modeling techniques can incorporate the full historic information that is available for a patient without resorting to summary measures and taking into account irregularity and sparsity that often come with real-world data (RWD). We conclude that the statistical modeling approaches that have been developed in literature, allow for far greater flexibility to model longitudinal data than what is currently being utilized by most published CPMs. This highlights that there is unrealized potential in incorporating longitudinal data in CPMs and the relevance of clinical data science to uncover this potential.

8.3 Conclusions

Real-world data (RWD), and electronic health records (EHRs) in particular, are both a goldmine and minefield. Developing clinical prediction models (CPMs) based on RWD is surrounded by pitfalls. Not taking time into account may be the biggest pitfall of them all. Time has the ability to make a CPM obsolete, faster than it might have taken to develop the model. Healthcare is always evolving, the focus of CPMs should therefore not lie on trying to achieve the best performance today, but on achieving a consistent performance in the future and providing a proven benefit to clinicians and patients. Only by understanding the healthcare process that generated the data, recognizing its limitations and ever changing nature, can a CPM based on RWD be developed successfully. Time will tell if the endeavor was successful.

8.4 Future perspectives

People today are living longer than ever before. This achievement, however, comes at the price of a growing strain on the healthcare system. Keeping healthcare affordable and maintaining quality of care, are challenges in the near future. This is where clinical prediction models (CPMs) can play a role by assisting clinicians in decision making, resulting in faster and more accurate diagnoses. Currently, the role of CPMs is limited and only a tiny fraction of all published CPMs are used in daily clinical practice. While it seems inevitable that CPMs will play a bigger role in clinical decision making in the future, the speed of this transition depends on several bottlenecks that have been uncovered in this thesis. First, the clinical need must be critically assessed, not only at the current moment, but also in the future. Many CPMs are being developed without there ever being a clinical need. In developing the CoLab score (Chapter 4), we focused on an accessible screening tool for the full emergency department (ED) population (instead of a symptomatic population), therefore the CoLab-score did not become obsolete when rapid polymerase chain reaction (PCR) testing was available. Second, with the rise of data-hungry artificial intelligence (AI) and machine learning (ML) techniques, accessibility to unbiased high quality healthcare data will become vital for the success or failure of these approaches. We have seen that considerable loss of predictive power can occur when using data that is derived from secondary sources where some conversion took place. Aside from the availability of high quality data as input for a CPM, equal attention must be paid to the outcome. If the clinical standard is not in agreement with the gold standard of diagnosis (see e.g. Chapter 2), the predictions from the CPM will be affected. Standardized outcome reporting will therefore be equally important as high quality input data. Third, the current IT infrastructure in hospitals does not allow for straightforward implementation of CPMs that are more complex than regression equations. For more advanced modeling techniques such as those described in Chapter 3 or AI and ML techniques, investment in hospital IT infrastructure is required to facilitate the use of these techniques. Fourth, the performance of CPMs will need to be monitored over time to detect calibration drift or loss of discriminative ability. A first step was made with the CoLab-score by defining prevalence dependent thresholds, further development and implementation of monitoring tools to detect loss of performance or calibration drift of CPMs over time is an interesting direction for future studies. Finally, there seems to be great promise for wearable devices, the challenge will be to present these data in a meaningful way to clinicians and integrating these data in the hospital infrastructure. To conclude, the question is not if CPMs will change healthcare in the future, but when and at what scale.

References

1. Martin, L., Hutchens, M., Hawkins, C. & Radnov, A. How much do clinical trials cost. *Nat Rev Drug Discov* **16,** 381–382 (2017).

2. Hopewell, S., Clarke, M. J., Stewart, L. & Tierney, J. Time to publication for results of clinical trials. *Cochrane database of systematic reviews* (2005).

3. Rothwell, P. M. External validity of randomised controlled trials:"to whom do the results of this trial apply?" *The Lancet* **365,** 82–93 (2005).

4. Stuart, E. A., Bradshaw, C. P. & Leaf, P. J. Assessing the generalizability of randomized trial results to target populations. *Prevention Science* **16,** 475–485 (2015).

5. Kho, M. E., Duffett, M., Willison, D. J., Cook, D. J. & Brouwers, M. C. Written informed consent and selection bias in observational studies using medical records: systematic review. *BMJ* **338** (2009).

6. Hewison, J. & Haines, A. Overcoming barriers to recruitment in health research. *BMJ* **333,** 300–302 (2006).

7. Junghans, C. & Jones, M. Consent bias in research: how to avoid it. *Heart* **93,** 1024 (2007).

8. Agniel, D., Kohane, I. S. & Weber, G. M. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. *Bmj* **361** (2018).

9. Groenwold, R. H. Informative missingness in electronic health record systems: the curse of knowing. *Diagnostic and prognostic research* **4,** 1–6 (2020).

10. Bots, S. H., Groenwold, R. H. & Dekkers, O. M. Using electronic health record data for clinical research: a quick guide. *European Journal of Endocrinology* **186,** E1–E6 (2022).

11. Hripcsak, G., Albers, D. J. & Perotte, A. Parameterizing time in electronic health record studies. *Journal of the American Medical Informatics Association* **22,** 794–804 (2015).

12. Chan, K. S., Fowles, J. B. & Weiner, J. P. Electronic health records and the reliability and validity of quality measures: a review of the literature. *Medical Care Research and Review* **67,** 503–527 (2010).

13. Thygesen, K. *et al.* Fourth universal definition of myocardial infarction (2018). *Journal of the American College of Cardiology* **72,** 2231–2264 (2018).

14. Gillinov, S. *et al.* Variable accuracy of wearable heart rate monitors during aerobic exercise. *Med Sci Sports Exerc* **49,** 1697–1703 (2017).

15. Kroll, R. R., Boyd, J. G. & Maslove, D. M. Accuracy of a wrist-worn wearable device for monitoring heart rates in hospital inpatients: a prospective observational study. *Journal of medical Internet research* **18,** e6025 (2016).

16. Steyerberg, E. W. *Clinical Prediction Models* 2nd ed. (Springer Nature, Switzerland, 2019).

17. Harrell, F. E. *Regression modeling strategies* 2nd ed. (Springer International Publishing, Switzerland, 2015).

18. Steyerberg, E. W., Eijkemans, M. J., Harrell Jr, F. E. & Habbema, J. D. F. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Medical Decision Making* **21,** 45–56 (2001).

19. Riley, R. D. *et al.* Penalization and shrinkage methods produced unreliable clinical prediction models especially when sample size was small. *Journal of clinical epidemiology* **132,** 88–96 (2021).

20. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R. & Feinstein, A. R. A simulation study of the number of events per variable in logistic regression analysis. *Journal of clinical epidemiology* **49,** 1373–1379 (1996).

21. Wynants, L. *et al.* Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. *bmj* **369** (2020).

22. Lim, C. C. *et al.* Early diagnosis of perioperative myocardial infarction after coronary bypass grafting: a study using biomarkers and cardiac magnetic resonance imaging. *The Annals of thoracic surgery* **92,** 2046–2053 (2011).

23. Van Calster, B., McLernon, D. J., Van Smeden, M., Wynants, L. & Steyerberg, E. W. Calibration: the Achilles heel of predictive analytics. *BMC medicine* **17,** 1–7 (2019).

24. Jenkins, D. A. *et al.* Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagnostic and Prognostic Research* **5,** 1–7 (2021).

25. Davis, S. E., Greevy Jr, R. A., Lasko, T. A., Walsh, C. G. & Matheny, M. E. Detection of calibration drift in clinical prediction models to inform model updating. *Journal of biomedical informatics* **112,** 103611 (2020).

26. Garcia Moreira, V. *et al.* Overestimation of albumin measured by bromocresol green vs bromocresol purple method: influence of acute-phase globulins. *Laboratory medicine* **49,** 355–361 (2018).

27. Siontis, G. C., Tzoulaki, I., Castaldi, P. J. & Ioannidis, J. P. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of clinical epidemiology* **68,** 25–34 (2015).

28. Kappen, T. H. *et al.* Evaluating the impact of prediction models: lessons learned, challenges, and recommendations. *Diagnostic and prognostic research* **2,** 1–11 (2018).

29. Goldstein, B. A., Navar, A. M., Pencina, M. J. & Ioannidis, J. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association* **24,** 198–208 (2017).

30. Plate, J. D. J. *et al.* Incorporating repeated measurements into prediction models in the critical care setting: a framework, systematic review and meta-analysis. *BMC medical research methodology* **19,** 1–11 (2019).

Summary

Historically, medicine was practiced as an art, based on the authority of a master, expert opinion and experience. Nowadays, medicine is considered both an art and a science, founded on results from clinical trials and research. The shift from authority-based medicine towards evidence-based medicine has been facilitated by the rise of statistics. An essential role of statistics in medicine is to model the generated data and translate results into clinical decision-making. One way of achieving this, is by developing clinical prediction models (CPMs) that can predict the likelihood of a certain outcome. With the rapid growth in data acquisition, storage, algorithms and computing power, the use of real-world data (RWD) to develop CPMs is becoming more and more popular. RWD are observational data relating to patient health status and/or the delivery of health care, generated as part of a healthcare process. This is opposite to data gathered as part of a clinical trial or study. Given that RWD is collected from the beginning until the end of the healthcare process, longitudinal information is a rule rather than an exception. To uncover the potential and pitfalls of using real-world longitudinal data in the development and validation of CPMs, several clinical challenges were addressed in this thesis.

First, we focused on the diagnosis of a perioperative myocardial infarction (PMI) after coronary artery bypass grafting (CABG), based on serial measurements of the biomarker cardiac troponin (cTn). The current clinical guideline is based on a single threshold for cTn to screen patients for a possible PMI. However, the release of cTn after CABG is characterized by a highly non-linear pattern, which varies between patients. The guideline does not take this characteristic cTn release profile into account, and moreover, results in many false positives. Also, the final diagnosis of a PMI requires additional criteria that vary between centers. Therefore, comparing cTn release profiles based on a center-specific diagnosis of a PMI, will lead to subjective findings. We have shown that a latent class mixed model, that does not make any subjective *a priori* assumptions about patient subgroups, can uncover clini-

cally relevant subgroups based solely on cTnT release profiles post-CABG. The class with a rising cTnT profile showed superior diagnostic accuracy over the clinical guideline. While this study demonstrated that the characteristic cTnT release profile is a better diagnostic criterion for screening patients for a PMI than a fixed cutoff, the next step was to develop predictive models that could be used in the clinic to dynamically classify patients based on accruing information from cTn measurements. We implemented several state of the art non-parametric statistical modeling approaches to dynamic classification of irregularly and sparsely sampled curves, and compared their performance in a simulation study. Results demonstrated that the generalized functional linear model (GFLM) was superior to the other approaches when historic information was taken into account, and the tensor product smooth (TPS) when historic information was not taken into account. The GFLM and TPS approaches were also applied to the cTnT data post-CABG, and showed better performance in diagnosing a PMI than the current clinical guideline.

Secondly, we focused on the screening of patients presenting at the emergency department (ED) for a possible coronavirus disease 2019 (COVID-19) infection. Routine laboratory test from a pre-pandemic cohort were combined with a cohort of patients presenting at the ED during the COVID-19 pandemic. These data were then combined with polymerase chain reaction (PCR) test results and used in a penalized regression model to obtain the CoLab-score. The CoLab-score is available within one hour after presentation and could safely rule-out COVID-19 in over one third of ED presentations, depending on the prevalence. Highly suspect cases could be identified regardless of presenting symptoms. The score was temporally validated in the development center and externally validated in three other centers. The novelty of the CoLab-score, compared to other CPMs developed during the COVID-19 pandemic, is that it is valid for the entire ED patient population where routine blood withdrawal is required, requires only ten routine laboratory tests and is adaptable to both high and low prevalence situations. Given the need for fast screening during outbreaks, a prospective study was done to assess if the CoLab-score could also be used to rule-out in infection in symptomatic healthcare workers (HCWs). From this study, it was concluded that the CoLab-score could also be used to safely rule-out a possible COVID-19 infection in symptomatic

HCWs. Using a safe threshold for the CoLab-score, COVID-19 could be ruled out in over one third of symptomatic HCWs.

Thirdly, two validation studies were performed. First, the European coLlaboration on Acute decompeNsated Heart Failure (ELAN-HF) risk score, developed by Salah et al. to predict readmission and/or mortality in patients admitted with acute decompensated heart failure (ADHF), was externally validated. The Cox proportional hazards model showed no signs of miscalibration, bias or reduced discriminative ability. Two factors that provided additional prognostic information, were i) if the patient was admitted with ADHF in the past year and ii) if the patient was managed by the outpatient heart failure clinic. Additionally, the association between self-care behavior and 6-month readmission and/or mortality was investigated. No evidence was found that self-care behavior was associated with 6-month readmission and/or mortality. However, self-care behavior was already adequate in the study population, making it difficult to quantify the effect of low self-care on the outcome. Secondly, a validation study was performed to assess the agreement between the heart rate (HR) extracted from a wrist worn optical heart rate monitor (OHRM) (integrating a photopletysmography (PPG) and accelerometer sensor) and the gold standard 5-lead electrocardiogram (ECG) connected to the patient monitor, in patients undergoing non-cardiac surgery. While many OHRM validation studies have been published, very few studies are performed in hospitalized patients and most include only stable ward patients. This study included a heterogeneous sample of patients, wearing the OHRM during surgery and recovery. It was concluded that, if the PPG signal is of sufficient quality, the HR measured by the OHRM falls within the ECG reference standard during both surgery and recovery. This study demonstrated the potential of an unobtrusive OHRM, to provide accurate HR measurements in hospitalized patients.

Finally, from the clinical challenges addressed in this thesis, some general conclusions were made regarding the potential and pitfalls of using real-world longitudinal data in the development and validation of CPMs. The biggest potential of RWD lies in its accessibility, volume and speed. While RWD does not suffer from consent bias, its major pitfalls are selection bias, informative missingness and data quality issues. Not taking these pitfalls into account can lead to propagation of bias and errors in CPMs. CPMs have the potential to make accurate predictions even if the sample size is small, either by penalization or through the use of mixed effects models in case of longitudinal data. Penalization also enables feature selection, reducing the number of predictors that need to be collected and subsequently simplifying implementation in other centers. Uncovered pitfalls in CPM development are, choosing ’exotic’ predictors, not taking concurrency of predictors into account, calibration drift and analytical variation. In conclusion, RWD are both a goldmine and minefield in developing CPMs. Healthcare is always evolving, the focus of CPMs should not lie on trying to achieve the best performance today, but on achieving a consistent performance in the future and providing a proven benefit to clinicians and patients.

Samenvatting

Historisch gezien werd geneeskunde beoefend als een kunst, gebaseerd op het gezag van een meester, de mening van experts en ervaring. Tegenwoordig wordt geneeskunde beschouwd als zowel een kunst als een wetenschap, gebaseerd op resultaten van klinische studies en onderzoek. De verschuiving van op autoriteit gebaseerde geneeskunde naar ”evidence-based” geneeskunde is mede mogelijk gemaakt door de opkomst van statistiek. Een essentiële rol van statistiek in de geneeskunde is het modelleren van de gegenereerde gegevens en het vertalen van resultaten naar klinische besluitvorming. Een manier om de besluitvorming te ondersteunen, is door het ontwikkelen van klinische predictie modellen (clinical prediction models (CPMs)). CPMs kunnen de waarschijnlijkheid van een bepaalde klinische uitkomst voorspellen, op basis van een set variabelen. Met de snelle groei van data-acquisitie, opslag, algoritmen en rekenkracht, wordt het gebruik van data uit de ’echte’ wereld (real-world data (RWD)) om CPMs te ontwikkelen steeds populairder. RWD zijn observationele gegevens met betrekking tot de gezondheidsstatus van de patiënt en/of de verstrekking van gezondheidszorg, gegenereerd als onderdeel van een zorgproces. Dit in tegenstelling tot gegevens die zijn verzameld als onderdeel van een klinische trial of studie. Aangezien RWD van begin tot einde van het zorgproces wordt verzameld, is longitudinale data eerder regel dan uitzondering. Om de potentie en valkuilen van het gebruik van real-world longitudinale data bij het ontwikkelen en valideren van CPMs te achterhalen, werden in dit proefschrift verschillende klinische problemen uitgewerkt.

Ten eerste, hebben we ons gericht op de diagnose van een perioperatief myocardinfarct (PMI) na een coronaire-bypassoperatie (CABG), gebaseerd op seriële metingen van de biomarker cardiaal troponine T (cTn). De huidige klinische richtlijn is gebaseerd op een vaste afkapwaarde voor cTn om patiënten te screenen op een mogelijke PMI. Het vrijkomen van cTn na een CABG wordt echter gekenmerkt door een sterk niet-lineair profiel, dat per patiënt varieert. De richtlijn houdt geen rekening met dit karakteristieke cardiac troponin (cTn) profiel en levert bovendien veel vals positieven op.

Ook vereist de definitieve diagnose van een PMI aanvullende criteria die per centrum verschillen. Daarom zal het vaststellen van een PMI-specifiek cTn profiel, op basis van een centrum-specifieke diagnose van een PMI, leiden tot subjectieve bevindingen. We hebben aangetoond dat een "latent class mixed model", dat afziet van *a priori* veronderstellingen over de diagnose van patiënten, klinisch relevante subgroepen kan ontdekken, uitsluitend gebaseerd op cTnT profielen post-CABG . De klasse met een stijgend cTnT-profiel had een superieure diagnostische nauwkeurigheid ten opzichte van de klinische richtlijn. Hoewel deze studie aantoonde dat het cTnT-profiel een beter diagnostisch criterium is voor het screenen van patiënten op een PMI dan een vaste afkapwaarde, was de volgende stap het ontwikkelen van voorspellende modellen die in de kliniek zouden kunnen worden gebruikt om patiënten dynamisch te classificeren op basis van herhaalde cTnT-metingen. Hiertoe hebben we verschillende state-of-the-art niet-parametrische statistische modelleringsbenaderingen geïmplementeerd voor dynamische classificatie van onregelmatig en schaars gemeten curves, en hun prestaties vergeleken in een simulatiestudie. De resultaten toonden aan dat het generalized functional linear model (GFLM) superieur was aan de andere benaderingen wanneer er rekening werd gehouden met historische informatie, en het tensor product smooth (TPS) wanneer er geen rekening werd gehouden met historische informatie. De GFLM- en TPS-benaderingen werden vervolgens toegepast op de data uit klinische praktijk, en presteerden beter in het diagnosticeren van een perioperative myocardial infarction (PMI) dan de huidige klinische richtlijn.

Ten tweede, hebben we ons gericht op het screenen van patiënten die zich op de spoedeisende hulp (SEH) presenteren met een mogelijke coronavirus disease 2019 (COVID-19)-infectie. Routine laboratoriumtesten van een pre-pandemisch cohort werden gecombineerd met een cohort van patiënten die zich presenteerden op de SEH tijdens de COVID-19-pandemie. Deze gegevens werden vervolgens gecombineerd met polymerase chain reaction (PCR)-testresultaten en gebruikt in een "penalized" regressiemodel om de CoLab-score te verkrijgen. De CoLab-score is binnen een uur na presentatie beschikbaar en kan COVID-19 veilig uitsluiten in meer dan een derde van de SEH-presentaties, afhankelijk van de prevalentie. Hoog verdachte gevallen konden worden geïdentificeerd, ongeacht de symptomen bij presentatie. De score werd temporeel gevalideerd in het ontwikkelingscentrum en extern gevalideerd in drie andere centra in Nederland. Het innovatieve aan de CoLab-score is dat deze score geschikt is voor het screenen van de gehele SEH-patiëntenpopulatie waarbij bloedafname plaatsvindt, slechts tien routine laboratoriumtesten vereist zijn om de score te berekenen en deze kan adaptief worden gebruikt tijdens zowel hoge als lage prevalentie. Gezien de noodzaak van snelle screening tijdens uitbraken, werd een prospectieve studie gedaan om te beoordelen of de CoLab-score ook gebruikt kon worden om infectie uit te sluiten bij zorgmedewerkers met symptomen van COVID-19. Uit een prospectieve studie werd geconcludeerd dat de CoLab-score ook gebruikt kan worden om een mogelijke COVID-19-infectie veilig uit te sluiten bij symptomatische zorgmedewerkers. Met een veilige afkanwaarde voor de CoLab-score kon een infectie worden uitgesloten bij meer dan een derde van de symptomatische zorgmedewerkers.

Ten derde, zijn er twee validatiestudies uitgevoerd. Allereerst werd de European coLlaboration on Acute decompeNsated Heart Failure (ELAN-HF) risicoscore extern gevalideerd. De ELAN-HF score is ontwikkeld door Salah et al. om heropname en/of mortaliteit te voorspellen bij patiënten opgenomen voor acuut gedecompenseerd hartfalen (ADHF). Het Cox-proportioneel risicomodel vertoonde geen tekenen van mis-kalibratie, bias of verminderd onderscheidend vermogen. Twee factoren die aanvullende prognostische informatie opleverden, waren i) of de patiënt het afgelopen jaar was opgenomen voor ADHF en ii) of de patiënt werd behandeld door de polikliniek hartfalen. Daarnaast werd het verband tussen zelfzorggedrag en heropname en/of mortaliteit na 6 maanden, onderzocht. Er werd geen bewijs gevonden dat zelfzorggedrag geassocieerd was met heropname en/of mortaliteit. Het zelfzorggedrag was echter al adequaat in de onderzoekspopulatie, waardoor het effect van lage zelfzorg op de uitkomst moeilijk te kwantificeren was. Ten tweede werd een validatiestudie uitgevoerd om de overeenkomst tussen de hartslag (HR) gemeten door een om de pols gedragen optische hartslag meter (OHRM) en de gouden standaard 5-lead electrocardiogram (ECG) aangesloten op de patiëntmonitor, vast te stellen. Dit gebeurde in een prospectieve studie bij patiënten die een niet-cardiale operatie ondergingen. Hoewel er veel OHRM validatiestudies zijn gepubliceerd, zijn er zeer weinig studies uitgevoerd bij gehospitaliseerde patiënten en de meeste richten zich op stabiele patiënten op een verpleegafdeling. Deze studie includeerde echter een heterogene groep van patiënten die de OHRM droegen tijdens operatie en herstel. Geconcludeerd kon worden dat, als het gemeten signaal van de OHRM van voldoende kwaliteit is, de HR van de OHRM binnen de ECG-referentiestandaard valt, tijdens zowel operatie als herstel. Deze studie toonde aan dat een OHRM potentie heeft om op een laagdrempelige manier, nauwkeurig HR te meten in een laag-risico setting.

Tenslotte, zijn er op basis van de klinische toepassingen die in dit proefschrift aan bod zijn gekomen, enkele algemene conclusies getrokken met betrekking tot de potentie en de valkuilen van het gebruik van real-world longitudinale data bij het ontwikkelen en valideren van CPMs. De grootste potentie van RWD ligt in de toegankelijkheid, het volume en de snelheid. Hoewel RWD geen last heeft van toestemmingsbias, zijn de grootste valkuilen selectiebias, ontbrekende informatie en problemen met de gegevenskwaliteit. Het niet in acht nemen van deze valkuilen kan leiden tot propagatie van bias en fouten in CPMs. CPMs hebben de potentie om nauwkeurige voorspellingen te doen, zelfs als de steekproefomvang klein is, hetzij door “penalization”, hetzij door het gebruik van hiërarchische modellen in het geval van longitudinale data. Penalization maakt ook selectie van variabelen mogelijk waardoor het aantal variabelen dat moet worden verzameld wordt verminderd, hetgeen de implementatie in andere centra vereenvoudigt. Belangrijke valkuilen bij CPM ontwikkeling zijn het gebruik van ’exotische’ variabelen, geen rekening houden met gelijktijdigheid van variabelen, kalibratiedrift en analytische variatie. Concluderend, is RWD zowel een goudmijn als een mijnenveld voor het ontwikkelen van CPMs. De gezondheidszorg is voortdurend in beweging, de focus van CPMs zou daarom niet moeten liggen op het bereiken van de beste prestaties vandaag, maar op het waarborgen van betrouwbare prestaties in de toekomst en het bieden van een bewezen voordeel voor clinici en patiënten.

List of abbreviations

ADHF acute decompensated heart failure.

AF atrial fibrillation.

AI artificial intelligence.

ALAT alanine aminotransferase.

ALP alkaline phosphatase.

AR auto-regressive.

ARL average run length.

ASA-PS American Society of Anesthesiologists physical status.

ASAT aspartate aminotransferase.

AUC area under the ROC-curve.

BC bromocresol.

BIC Bayesian information criteria.

BNP brain natriuretic peptide.

bpm beats per minute.

BUN blood urea nitrogen.

CABG coronary artery bypass grafting.

CGC conditional growth chart.

CI confidence interval.

CK creatine kinase.

CKD-EPI CKD-EPI estimated glomerular filtration rate.

CO-RADS COVID-19 Reporting and Data System.

COV-LDA covariance pattern longitudinal discriminant analysis.

COVID-19 coronavirus disease 2019.

CPM clinical prediction model.

CRP C-reactive protein.

Ct cycle threshold.

cTn cardiac troponin.

EBM evidence based medicine.

ECG electrocardiogram.

ED emergency department.

EHFScBS-9 European Heart Failure Self-care Behaviour Scale.

EHR electronic health record.

ELAN-HF European coLlaboration on Acute decompeNsated Heart Failure.

F-LDA functional longitudinal discriminant analysis.

FACEs fast covariance estimation for sparse functional data.

GAM generalized additive model.

GFLM generalized functional linear model.

gGT gamma-glutamyltransferase.

HCW healthcare worker.

HF heart failure.

HR heart rate.

ICU intensive care unit.

IPI inter-pulse-interval.

IQR interquartile range.

KM Kaplan-Meier.

LCMM latent class mixed model.

LD lactate dehydrogenase.

LDA longitudinal discriminant analysis.

LFT rapid lateral flow test.

LIS laboratory information system.

LME linear mixed effects.

LoA limits of agreement.

LoS length of stay.

LP linear predictor.

LR likelihood ratio.

MCH mean cellular hemoglobin.

MCHC mean cellular hemoglobin concentration.

MCV mean corpuscular volume.

mHealth mobile health.

MI myocardial infarction.

ML machine learning.

MOVER method of variance estimates recovery.

NPV negative predictive value.

NT-proBNP N-Terminal pro–B-type natriuretic peptide.

NYHA New York Heart Association.

OHRM optical heart rate monitor.

OPCAB off-pump coronary artery bypass grafting.

OR operating room.

PC principal component.

PCR polymerase chain reaction.

PH proportional hazards.

PMI perioperative myocardial infarction.

PPG photopletysmography.

PPV positive predictive value.

QGAM smooth additive quantile regression model.

QI quality index.

QR quantile regression.

RCT randomized controlled trial.

ROC receiver operating characteristic.

RWD real-world data.

SARS-CoV-2 severe acute respiratory syndrome coronavirus 2.

SD standard deviation.

SE standard error.

SGC static growth chart.

TPS tensor product smooth.

TU/e Eindhoven University of Technology.

URL upper reference limit.

XC aortic cross clamping.

List of publications

Deneer, R., van Boxtel, A. G., Boer, A.-K., *et al.* Detecting patients with PMI post-CABG based on cardiac troponin-T profiles: a latent class mixed modeling approach. *Clinica Chimica Acta* **504,** 23–29 (2020)

Deneer, R., Zhan, Z., van den Heuvel, E. R., *et al.* A comparison of non-parametric statistical modeling approaches to dynamic classification of irregularly and sparsely sampled curves. *In preparation for submission*

Boer, A.-K.[*], **Deneer, R.**[*], Maas, M., *et al.* Development and validation of an early warning score to identify COVID-19 in the emergency department based on routine laboratory tests: a multicentre case–control study. *BMJ Open* **12,** e059111 (2022)

Leers, M. P.[*], **Deneer, R.**[*], Mostard, G. J., *et al.* Use of an algorithm based on routine blood laboratory tests to exclude COVID-19 in a screening-setting of healthcare workers. *PLoS One* **17,** e0270548 (2022)

Vinck, T. A., **Deneer, R.**, Verstappen, C. C., *et al.* Validation of the ELAN-HF score and self-care behaviour on the nurse-led heart failure clinic after admission for heart failure. *BMC Nursing* **21,** 1–10 (2022)

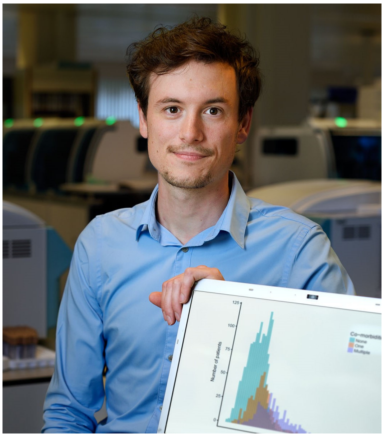
Mestrom, E., **Deneer, R.**, Bonomi, A. G., *et al.* Validation of heart rate extracted from wrist-based photoplethysmography in the perioperative setting: prospective observational study. *JMIR Cardio* **5,** e27765 (2021)

[*] both authors contributed equally.

Other publications not related to this thesis

van Loon, S. L., **Deneer, R.**, Nienhuijs, S. W., *et al.* Metabolic Health Index (MHI): assessment of comorbidity in bariatric patients based on biomarkers. *Obesity Surgery* **30,** 714–724 (2020)

About the author



Ruben Deneer was born on the 21st of April 1986 in Valkenburg aan de Geul. After finishing his VWO in 2005, he studied Biomedical Engineering at the Eindhoven University of Technology (TU/e) in the Netherlands. In 2009 he pursued his childhood ambition to become a public transit bus driver. After working fulltime in public transit, he returned to the TU/e to graduate in 2017 within the research group of Computational Biology. His master thesis was performed in collaboration with the Department of Clinical Chemistry at the Catharina Hospital in Eindhoven. After his graduation in 2017 he started working as a PhD student, part of the ITEA eWatch project, at the Department of Clinical Chemistry in the Catharina Hospital. His PhD research focused on the development, validation, and implementation of clinical prediction models, specializing in longitudinal data. He remained part of the Computational Biology research group at the TU/e, supervising students. During the coronavirus pandemic he performed the statistical modeling for an early warning score to detect patients infected with COVID-19 in the emergency department. This score was implemented in multiple hospitals in the Netherlands and formed the basis for a grant funded by the EU's Horizon 2020 programme for further development. Since 2021 he is employed at the Catharina Hospital as a researcher, whilst working parttime as a public transit bus driver. The results of studies performed during his PhD period are presented in this dissertation.

Dankwoord

Hoewel promoveren vaak (terecht) als een solitaire aangelegenheid wordt beschouwd, prijs ik mij gelukkig met de hulp, inzet en steun die ik heb mogen ontvangen tijdens mijn promotietraject. Zonder deze personen zou dit werk nooit het huidige niveau hebben bereikt, laat staan dat het überhaupt zou zijn afgerond.

Volkher, bedankt dat je mij de kans hebt gegeven om te promoveren in het AKL. Ik bewonder je scherpe inzichten en kritische blik op de klinische datawetenschap. Mijn neiging om helemaal op te gaan in de statistiek en daarbij andere zaken (lees: includeren) te laten voor wat ze zijn, moet je soms tot wanhoop hebben gedreven. Bedankt dat je altijd bleef zoeken naar oplossingen en mij met de feiten bleef confronteren. Natal, vanuit jouw vakgroep ben ik ooit begonnen als afstudeerder. Voor mij was Computational Biology een voor de hand liggende keuze, de perfecte combinatie van programmeren en modelleren. Bedankt voor je ondersteuning vanuit de TU. Arjen-Kars, tijdens mijn afstuderen had ik al een voorproefje gehad van jouw enorme drive om het beste uit klinische data naar boven te halen. Je zou dus kunnen stellen dat ik wist waar ik aan begon met jou als copromotor. Dat is ook gebleken, discussies met jou waren nooit saai en konden zo een middag in beslag nemen. Jouw afkeer van "data-cowboys" en het blind gebruiken van klinische data, deel ik ten zeerste. Tenslotte wil ik je bedanken voor de vele uren die je hebt gestopt in het COVID-X project. Daarnaast wil ik graag alle commissieleden bedanken voor het kritisch lezen van mijn dissertatie en hun bereidheid deel te nemen aan de promotiezitting.

Mathie, allereerst bedankt voor je inzet bij het aanleveren van de externe validatie data uit het Zuyderland. Daar bleef het niet bij. Je hebt daarna in razend tempo een prospectieve studie opgezet én een ZonMw aanvraag ingediend (en gehonoreerd gekregen!). Ik kijk uit naar onze toekomstige samenwerking. Madelon en Muriël, jullie ook bedankt voor het snel aanleveren van de externe validatie datasets voor de CoLab-score.

Tineke en Cindy, bedankt voor jullie ondersteuning vanuit de hartfalen poli. Jullie namen altijd de tijd voor me als ik vragen had over hartfalen, hebben me wegwijs gemaakt op de afdeling, en hielpen me met het lezen van patiëntendossiers. Luuk, jij bedankt voor je input als cardioloog en intensivist. Je reageerde altijd met veel enthousiasme op onze ideeën en bleef meedenken als we ergens niet uitkwamen. Graag wil ik Yosha, Carmen, Jip, Anna, Senna en Sophie bedanken voor hun hulp bij het includeren van patiënten, het invullen van de CRFs en het verzamelen en meten van samples. Zonder jullie hulp was de inclusie nooit zover gekomen. Anton en Anjolie, jullie bedankt voor de coördinatie van de HFOW studie samples op het lab. Jullie hadden altijd tijd voor een gesprek en waren bereid 'out of the box' te denken zodat samples voor mijn studie via routine afnames konden worden aangevraagd. Geniet straks van jullie pensioen. Astrid en Mo, bedankt voor jullie kennis op het gebied van de cardiochirurgie en het met mij delen van de TROPACS studie.

(Zhuozhao) Zhan, thank you for helping me understand the theory behind mixed effects models and for your contributions to the longitudinal classification paper. I admire your extensive knowledge about statistics and thank you for always making time for a meeting in your busy schedule.

Remco, toen we begonnen als promovendi in de "herenkamer" was de sfeer al goed, dat is altijd zo gebleven. Je was eerder klaar dan ik (gelukkig) en sindsdien heb ik de goede gesprekken en grappen wel gemist. Saskia, bedankt voor je begeleiding tijdens mijn afstuderen. Hierdoor is mijn interesse in de klinische datawetenschap dusdanig gegroeid dat ik van mijn standpunt dat ik nooit zou gaan promoveren, ben afgestapt. Jonna, je bent altijd behulpzaam en hebt mij ondersteund in de taken waar ik zelf niet aan toe kwam, zonder dat ik daarom hoefde te vragen, bedankt! Sophie, Esther, Sylvia en Sebas, jullie ook bedankt voor de samenwerking. Ik kijk al uit naar jullie boekjes en mochten jullie statistische vragen hebben, weten jullie me te vinden. Eveline, bedankt voor de fijne samenwerking tijdens de perioperatieve monitoring studie, mooi om te zien hoe technisch een arts kan zijn! Reinder en andere collega's van Philips, bedankt voor het aanleveren van de ELAN horloges en jullie ondersteuning bij het analyseren van de data. To all the people from the CBIO group, thank you for all the interesting talks and the interest you have shown for my research.

Robert, je was mijn mentor toen ik lang geleden mijn carrière als buschauffeur begon. Die rol vervul je nog steeds, in de breedste zin van het woord, en daarbij ben je ook nog eens een goede vriend. Bedankt dat je me de wereld buiten de universiteit en het ziekenhuis laat zien. Dorien, jij ook bedankt, tijdens mijn promotie is mijn bewondering voor verpleegkundigen zeker gegroeid. Tim, al van kinds zijn we vrienden, ondanks de fysiek steeds groter geworden afstand. Jij en Carola (en Ziva en Cody) bedankt voor de bourgondische etentjes en weekendjes weg. Dennis en Laura (en Milou), bedankt voor de afleiding tijdens alle weekendjes in Banholt. Dirk en Soraya (en Max, Ben en Mart) bedankt voor jullie gezelligheid en leuke gesprekken. Marleen, altijd handig om talent in de familie te hebben, heel erg bedankt voor het maken van de kaft. Erik, Tessa, Eric, Linda, Kim, Tim, Pim, Cindy, Sander, Carmen, Sander en Jin, bedankt voor de feestjes en gezelligheid, ik heb veel bewondering voor wat jullie allemaal bereikt hebben.

Ans en Lei, bedankt voor de fijne jeugd die mij heeft gevormd tot wie ik nu ben. Met enige nostalgie denk ik terug aan vroeger. Bedankt dat jullie mijn interesse in techniek hebben gestimuleerd en me vrij hebben gelaten in het maken van keuzes. Imre, bedankt voor je waardevolle leesadviezen en samen met Djur, Géza en Alek, bedankt voor jullie gezelligheid en afleiding. Anna en Jona, als promovendi weten jullie natuurlijk precies wat promoveren inhoudt, ik heb veel bewondering voor jullie onderzoek en motivatie. Kirian en Diede (en Muna), jullie hebben de meeste up en downs van dichtbij meegemaakt. Bedankt voor jullie interesse en dat ik altijd welkom was in Blauwgraspad. Ad en Martine, bedankt voor jullie steun en het zorgen voor afleiding in de vorm van diverse vakantiehuisjes.

Liefste Rachna, jij kent me als geen ander, en ondanks al mijn eigenaardigheden, waardeer je me zoals ik ben. Je laat me vrij om mijn eigen weg te volgen maar bent er altijd voor me als het tegenzit. Bedankt voor al je steun, vooral tijdens de laatste loodjes. Nu het meeste werk er op zit, kijk ik met veel plezier uit naar wat de toekomst ons te bieden heeft.