ERASMUS UNIVERSITY ROTTERDAM AND ERASMUS MC

ERASMUS SCHOOL OF ECONOMICS

MASTER THESIS ECONOMETRICS AND MANAGEMENT SCIENCE

Simulation of haemoglobin concentrations in MISCAN-Colon using a mixed-effect machine learning model.

Danica VAN DEN BERG (470755)

Supervisor:
dr. A. Alfons

Second Assessor:
dr. L.E. Westerink - Duijzer

Supervisors Erasmus MC:
L. de Jonge
dr. I. Lansdorp-Vogelaar

Date: October 8, 2021

Abstract

The MISCAN-Colon model is a microsimulation model, developed by the Erasmus University Medical Center of Rotterdam, for the evaluation of colorectal cancer screening. To evaluate the benefits of personalised screening strategies, based on the previous haemoglobin (Hb) concentration found in a person's stool, MISCAN-Colon needs an accurate simulation model for the Hb values. The aim of this paper is to extend the MISCAN-Colon model with a simulation model for the Hb concentration found in a person's stool.

This thesis presents a mixed-effect machine learning (MEml) model to simulate the Hb concentrations. Tree-based machine learning algorithms, *K*-nearest neighbours and artificial neural network were explored for the machine learning component in the MEml model. We compared the performance of the MEml models to the performance of a linear mixed-effect model, namely a mixed-effect zero-inflated negative binomial (ZINB) model. We concluded that the MEml models outperform the mixed-effect ZINB model significantly. The MEml model with the best predictions uses a *K*-nearest neighbours algorithm, however the differences with the other MEml models is small. Since a regression tree is more interpretable and better in dealing with large data sets than *K*-nearest neighbours, we developed a simulation model in MISCAN-Colon using a mixed-effect regression tree.

Contents

1 Introduction

Colorectal cancer (CRC) is the second leading cause of cancer death in Europe and other developed countries (Ferlay et al., 2018). Fortunately, early detection of CRC by means of screening prevents CRC-related dead and detection of precursor lesions prevents the development of CRC (Van Hees et al., 2015). In response, many countries have decided to offer CRC screening to their population to reduce morbidity and mortality from the disease. However, screening also poses serious harms, such as false-positive test results and over-diagnosis (Van Hees et al., 2015). It is therefore important to strive to improve the benefits of screening and reduce its harms.

In the Netherlands, a national CRC screening programme was implemented in 2014, with faecal immunochemical test (FIT) screening for persons between 55 and 75 years of age (Toes-Zoutendijk et al., 2017). The FIT identifies hidden blood in the stool, which can be an early sign of CRC or a precursor lesion of CRC. Participants with a positive FIT, meaning that the FIT result exceeds the cut-off level of 47 μg haemoglobin (Hb) per gram faeces, are referred for colonoscopy. Participants with a negative FIT result are reinvited for a new screening two years after the previous invitation date.

The MISCAN-Colon (MIcrosimulaten SCreening ANalysis) model has been very useful in the decision, planning and implementation phase of the Dutch CRC screening programme (Van Hees et al., 2015). MISCAN-Colon is a microsimulation model that simulates a large population from birth to death with similar life expectancy and risk of CRC as the Dutch population. The model is used to predict the costs and benefits of different screening strategies and determine the optimal screening strategy for the Dutch population. This can be done, for example, by varying the cut-off for referral to colonoscopy, as well as the age range and screening interval.

Currently, research is being done to explore the benefits of personalised FIT screening strategies (Kuipers & Grobbee, 2020). For example, the FIT screening interval could depend on an individual's Hb concentration in previous FITs. MISCAN-Colon can be used to evaluate the benefits of such a personalised FIT screening. However, since the personalised screening policy is based on previous Hb concentrations, MISCAN-Colon needs an accurate simulation model for the Hb values.

So far, the proposed method to simulate the Hb values is a mixed-effect zero-inflated negative binomial model (ZINB). One important characteristic of mixed-effect models is that they are better able to capture correlations between different observations in comparison to a "standard" linear

regression model (Ngufor, Van Houten, Caffo, Shah, & McCoy, 2019). This is useful for modeling Hb concentrations since there is a high degree of correlation between multiple screening rounds of the same individual. The model is called zero-inflated because it contains relatively many zeroes compared to a regular negative binomial distribution. Zero-inflation is chosen because the majority of the people has no blood in their stool and consequently approximately 85% of the FITs detect zero Hb.

One drawback of the ZINB model is that it assumes a parametric distribution and imposes restrictive linear relationships between the response variable and the features. Alternatively, non-linear machine learning techniques can be applied to extract informative patterns from the data without making this strong assumption. However, one problem with most machine learning algorithms is that they assume that the training data is independent and identically distributed (Ngufor et al., 2019). This assumption is violated in our setting since multiple FIT results of the same individual exhibit a high degree of correlation.

Therefore, the purpose of this research was to investigate the application of machine learning algorithms to longitudinal and clustered data sets. Specifically, we developed a simulation model for the Hb concentrations in MISCAN-Colon which combines the structure of mixed-effect models with a machine learning algorithm. For this purpose, we used the mixed-effect machine learning (MEml) model proposed by Ngufor et al. (2019). With this approach, random-effects can be incorporated in any supervised machine learning algorithm. In this paper, we explore tree-based machine learning algorithms, *k*-nearest neighbours and artificial neural network for the machine learning component in the MEml model. Since Ngufor et al. (2019) focus only on the performance of MEml models with different tree-based machine learning algorithms, we contribute to the existing literature by investigating the performance of MEml models with a larger variety of machine learning algorithms.

The remainder of this paper is structured as follows. In Section 2 we provide some background information about the development of CRC and the MISCAN-Colon model, followed by a description of the problem in Section 3. Next, in Section 4 we discuss relevant existing literature. Then, in Section 5 we describe the data set used in this research. This is followed in Section 6 by an explanation of the different econometric methods we apply. The results of our models are discussed in Section 7. Finally, the conclusion and discussion of our research follow in Section 8.

2 Background knowledge

In this section, we first describe how colorectal cancer (CRC) develops and what methods are used to screen for CRC. Then, we give a description of the MISCAN-Colon model.

2.1 Colorectal cancer development and screening

Generally, CRC develops slowly and does not produce symptoms until the cancer reaches a considerable size of several centimeters (Simon, 2016). Cancer in the colon mainly develops from adenomas arising from the lining of the intestine (Cooper et al., 2010). Adenomas can be classified as non-advanced or advanced based on their diameter and histopathological features. Non-advanced adenomas are defined as adenomas with a diameter smaller than 10 mm, while advanced adenomas are larger than 10 mm (Cottet et al., 2012). The development of adenomas is quite common, but only a small percentage, 3.3%, eventually becomes malignant (Winawer, 1999). Therefore we distinguish between progressive and non-progressive adenomas. Progressive adenomas will develop into a cancer while non-progressive adenomas will not. Progressive adenomas first develop into preclinical cancer stage I. Preclinical indicates that the cancer is not yet diagnosed because of symptoms. Preclinical cancer can then progress from stage I to stage IV (Compton & Greene, 2004). In each stage symptoms may develop and preclinical cancer can change to clinical cancer in which the disease is diagnosed. The disease progression from adenoma to cancer is shown in Figure 1. In this figure, the term lesion is also introduced. A lesion is a general term used to refer to an abnormal region of tissue, in this case referring to both adenomas and cancer.

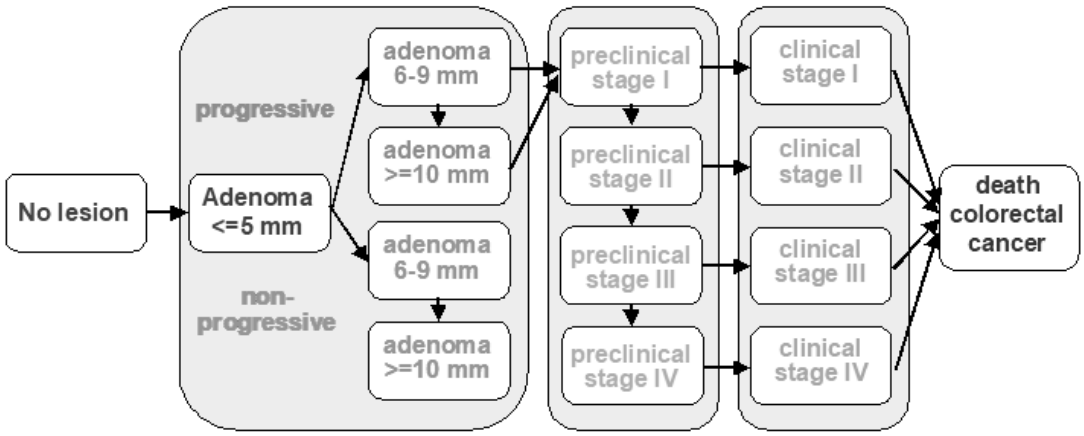


Figure 1: Disease progression from adenoma to colorectal cancer. Source: Gini (2020)

CRC screening has the potential to detect adenomas even before they become symptomatic. In this way, further development into cancer can be prevented. The Dutch national CRC screening programme includes two types of screening procedure: stool-based tests and endoscopic tests (Toes-Zoutendijk et al., 2017).

Each invitee of the Dutch national CRC screening programme is asked to perform a faecal immunochemical test (FIT). This is a stool-based test designed to measure the presence of blood in the stool. A high haemoglobin (Hb) concentration indicates a higher risk of having an adenoma or cancer. The FIT is preferred to other stool-based test, due to the high participation and a high diagnostic performance (Van Rossum et al., 2008). In addition, the FIT is quick, inexpensive and noninvasive in comparison to endoscopic tests. The Dutch national CRC screening programme uses biennial FIT screening between ages 55 and 75 (Van Hees et al., 2015). If the Hb concentration is above the cut-off of 47 μg Hb per gram faeces, the FIT result is considered positive.

If the result of the FIT is positive, the participant is referred to a colonoscopy center nearby for a follow-up test. The endoscopic test, called a colonoscopy, is used as the standard follow-up test. With a colonoscopy, the colon and rectum are visualised and evaluated by an endoscopist directly. If any lesions are detected, they are removed and in case of advanced adenoma or CRC, the person is referred for further treatment and surveillance. The person is then removed from the regular FIT screening programme. If, on the other hand, no lesions are found, then the individual is invited again for a FIT screening, but only after ten years.

2.2 MISCAN-Colon model

Despite the benefits of CRC screening, it also has potential harms. The main harms of screening are the chance of having a false-positive FIT followed by overtreatment and complications during endoscopic tests. In addition, many participants suffer from anxiety or discomfort of endoscopic examination (Trevisani, Zelante, & Sartori, 2014), as well as the inconvenience of the FIT (Worthley et al., 2006). Moreover, both stool-based tests and endoscopic procedures bring financial costs.

To analyse the benefits, harms and costs of CRC screening the MISCAN-Colon model can be used. MISCAN-Colon is a microsimulation model that can evaluate different CRC screening policies by comparing their costs and effectiveness (Loeve, Boer, van Oortmarssen, van Ballegooijen, & Habbema, 1999). A general MISCAN model was first described by Habbema, Van Oortmarssen, Lubbe, & Van der Maas (1985), followed by the description of MISCAN-Colon (Loeve et al., 1999), a MISCAN model specifically adapted for colorectal cancer. The model consists of three parts: demography, natural history and screening.

The demography part simulates a population of fictitious individuals without CRC. For each individual, a time of birth and a time of death of other causes than CRC is simulated. The input for this simulation consists of birth and life tables.

Next, in the natural history part, adenomas are simulated for some individuals. Most individuals do not develop adenomas, while others develop multiple adenomas. The simulated adenomas can progress into preclinical cancer and subsequently into clinical cancer. This simulated disease progression corresponds to the disease progression described in Section 2.1. This part of MISCAN-Colon requires the input of disease-related characteristics, e.g. the risk of developing the disease and the duration of each stage of the disease.

In the screening part of the model, screening for CRC is simulated. A screening policy consists of the following characteristics: the age at which individuals are invited for screening, the time interval between repeated screenings, the type of screening tests used at each examination, and the diagnostic follow-up appointments after a positive test. These properties are used as input for the screening part of the MISCAN-Colon model. Due to screening, some simulated life histories will change. This is due to the fact that for some individuals CRC may be prevented by the early detection and removal of adenomas.

Figure 2 gives an illustration of the simulation done in each part of MISCAN-Colon. First, in the demography part, a person is simulated without CRC who dies at the age of 80 due to an-

other cause than CRC. Then, in the natural history part, an adenoma that becomes malignant is simulated. Without screening, this person dies of CRC at the age of 70. After the model has simulated a life-history without screening, screening is overlayed. In this individual, at the moment of screening, the adenoma that is present is detected and removed. Since the adenoma is removed, the person does not develop cancer and therefore dies at the the age of 80 from other causes. Thus, in this case, the individual gains 10 life years with the screening intervention.

By comparing the simulated life histories with and without screening, MISCAN-Colon can evaluate the costs and benefits of a specific screening strategy.

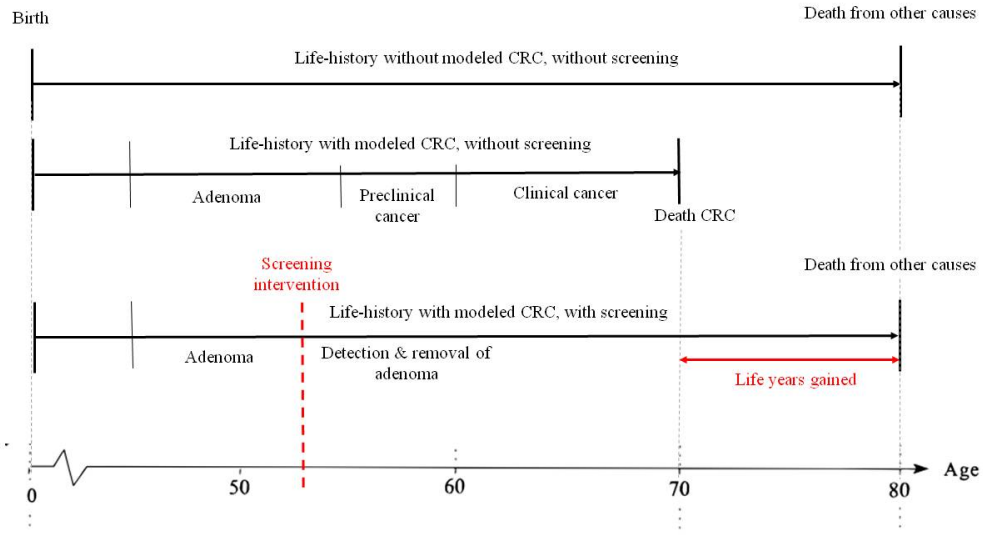


Figure 2: An example of a life history simulated in MISCAN-Colon for a person who develops colorectal cancer during his life. The upper bar shows only the demography part, the middle bar adds the natural history and the lower bar adds both the natural history and screening.

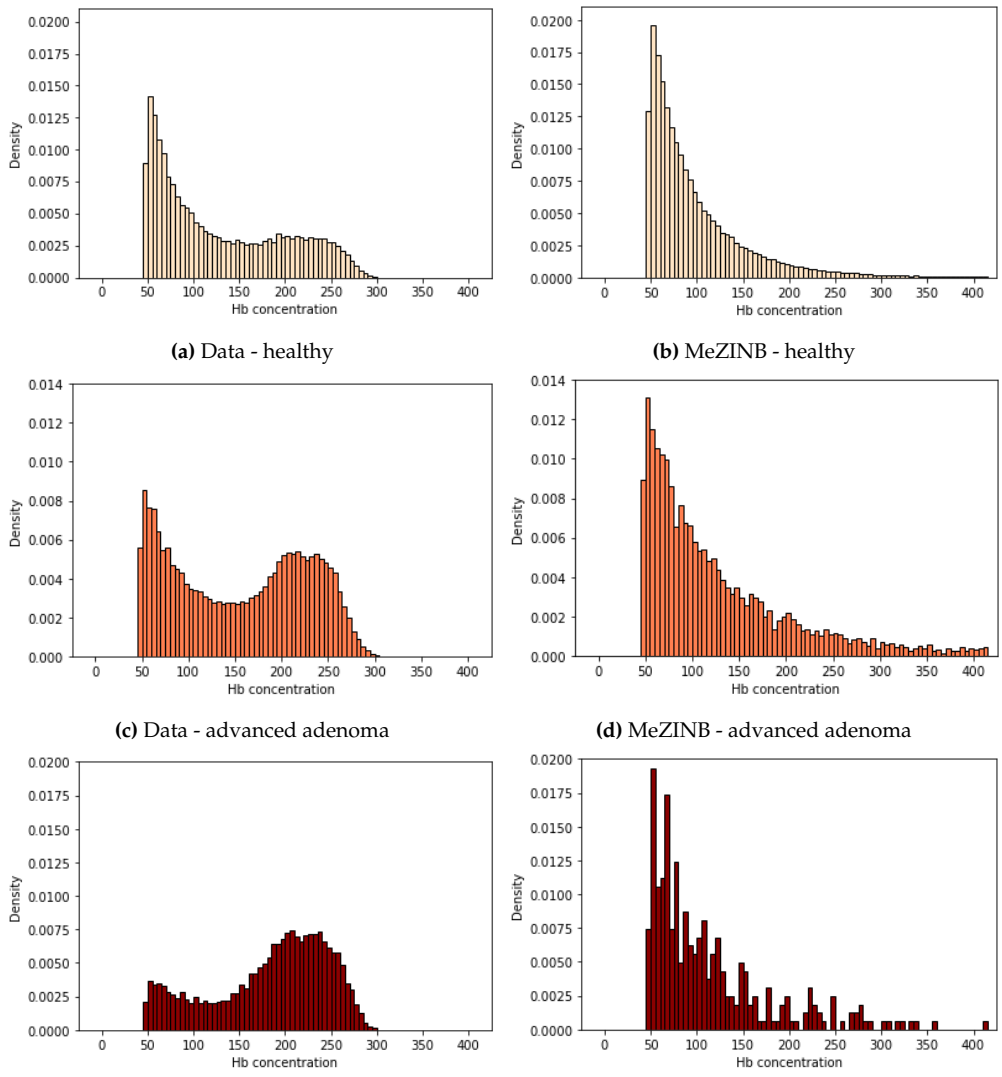
3 Problem description

To evaluate personalised CRC screening policies that base their screening intervals and/or cut-off levels on the individual's haemoglobin (Hb) concentration, MISCAN-Colon needs an accurate simulation model for the Hb values. Every time a simulated individual in MISCAN-Colon is screened, a Hb concentration needs to be generated to determine the outcome of the faecal immunochemical test (FIT). If the simulated Hb value is above a predefined cut-off, the outcome of the FIT is considered positive, while a simulated Hb value below the predefined cut-off is considered negative. The main goal is to obtain a Hb simulation model for which the simulated Hb concentrations resemble

the observed Hb concentrations of real-life Dutch population screening data.

The MISCAN-Colon model that is currently used for the simulation of the Dutch population screening does not simulate a Hb value for every FIT. Instead it uses the absence or presence of simulated lesions and a pre-defined specificity and sensitivity to determine whether the FIT has a positive or negative result. The specificity indicates the probability of a negative result in a person without lesions, whereas the sensitivity indicates the probability of a positive result due to a lesion in a particular stage.

Recently, the Public Health department of Erasmus Medical Center explored the extension of the MISCAN-Colon model with a simulation model for the Hb values. The model that they use to simulate Hb concentrations is a mixed-effect zero-inflated negative binomial (ZINB) model. This model can take into account that the majority of the population has no blood in their stool and that there is a high degree of correlation between multiple screening rounds of the same individual. However, they concluded that the simulation model needs further calibration since the simulated distribution of Hb values is not yet similar to the observed distribution. The difference between the theoretical Hb distribution and the observed Hb distribution is shown in Figures 3.



(a) Data - healthy

(b) MeZINB - healthy

(c) Data - advanced adenoma

(d) MeZINB - advanced adenoma

(e) Data - CRC

(f) MeZINB - CRC

Figure 3: Distributions of the haemoglobin concentrations (Hb) simulated by MISCAN-Colon with the mixed-effect ZINB (MeZINB) model and distributions of the Hb concentrations obtained from the Dutch CRC screening programme for the stages healthy, advanced adenoma and CRC. 1e6 individuals are simulated with MISCAN-Colon. The figures show only the Hb values above the cut-off of 47 μg/g.

We used the mixed-effect ZINB model as a starting point for the search for an accurate Hb simulation model. Our research can be described in two phases. First, different linear and non-linear mixed-effect models were explored for the prediction of Hb concentrations. These models were trained and tested using a real-life data set of the Dutch national CRC screening programme. We compared the performance of all proposed models using multiple performance measures.

In the second phase we incorporated the most promising model of phase one in MISCAN-Colon. We calibrated the parameters of this model in MISCAN-Colon in order for the simulated Hb distribution to match the observed Hb distribution as closely as possible. Since calibration is a

time-consuming task, it is not preferable to perform the calibration multiple times. Therefore, only the best performing method of phase one was calibrated. With this two-phase approach, we tried to answer the following research questions:

1. How can we combine the structure of mixed-effect models with machine learning algorithms?
2. Are non-linear MEml models better in predicting the Hb concentration than the linear mixed-effect ZINB model?
3. Which MEml model is best suited for predicting the Hb concentration in CRC screening?
4. Using a MEml model, can we develop a Hb simulation model in MISCAN-Colon for which the simulated Hb distribution resembles the observed Hb distribution of the Dutch CRC screening population?

4 Literature review

In this section we review the existing literature regarding models for the simulation of haemoglobin (Hb) concentrations. We start by reviewing existing parametric models, in particular generalized linear mixed-effect models. Next, we introduce the mixed-effect machine learning model.

4.1 Parametric models for Hb simulation

To accurately simulate Hb concentrations in MISCAN-Colon, we need a model that can capture all characteristics of the observed Hb distribution. One important characteristic is that the Hb concentrations can only take non-negative values. In addition, the distribution is characterised by a long right-tail and a mass-point at zero. Ridout, Demétrio, & Hinde (1998) offer some solutions for the problem of modelling non-negative data with excess zeros: a mixed Poisson distribution, hurdle models, birth process models and threshold models. An alternative method to analyse such data is a zero-inflated Poisson (ZIP) regression, first proposed by Lambert (1992). However, Yau, Wang, & Lee (2003) argue that a zero-inflated negative binomial (ZINB) model might be more appropriate than the ZIP, because the ZIP model assumes that the nonzero counts follow a zero-truncated Poisson distribution. This assumption is often violated in practice.

4.2 Generalized linear mixed-effect models

Yau et al. (2003) extend the ZINB model to account for situations where the data is longitudinal, such as repeated measures data. They propose a mixed-effect ZINB regression model to predict the length of hospital stay. In addition to the fixed clinical- and patient-related effects, random effects are incorporated into the model to take into account the correlation between observations within the same hospital.

The mixed-effect ZINB model belongs to the class of generalized linear mixed-effect models (GLMMs). GLMMs are an extension of generalized linear models in which random effects are added to the linear predictor (Stroup, 2012). The GLMM framework has been frequently used in different disciplines, especially in biology, to analyse longitudinal data, and in physical sciences and engineering, to analyse correlated observations (Myers, Montgomery, Vining, & Robinson, 2012). In this paper, the mixed-effect ZINB model is used as a benchmark.

4.3 Combining random effects with machine learning

A drawback of GLMMs is that it assumes a parametric distribution and imposes restrictive linear relationships between the response variable and the features. In addition, the ZINB model is used to model discrete data, while the Hb concentration is a continuous variable. An alternative is the use of non-linear machine learning algorithms. For example, Zhao et al. (2019) use random forest, gradient boosting trees, convolutional neural networks and recurrent neural networks for disease prediction with longitudinal data. Kinreich et al. (2019) predict risk for alcohol use disorder with support vector machines using longitudinal data and Perveen et al. (2020) use machine learning techniques for prognostic modeling of diabetes.

Nonetheless, using machine learning algorithms for the analysis of longitudinal data without adjusting for the inherent correlation structure in the data often leads to mediocre performance and potential for misleading inference (Ngufor et al., 2019). As a solution, Hajjem, Bellavance, & Larocque (2010) propose the generalized mixed-effect regression trees (GMERT) model, which combines the structure of mixed-effect models with the flexibility of tree-based machine learning algorithms. They conclude that this combination is able to improve predictive performance over standard trees and allows the modeling of target variables without assuming that linearity holds. Moreover, Sela & Simonoff (2012) propose an estimation method, called RE-EM tree, for mixed-effect regression tree models. The estimation method they propose alternates between estimating

the regression tree, assuming that the estimates of the random effects are correct, and estimating the random effects, assuming that the regression tree is correct.

Ngufor et al. (2019) propose a mixed-effect machine learning (MEml) model that incorporates random-effects into machine learning algorithms for efficient analysis of longitudinal data. The MEml model uses the structure of the GMERT model and estimates the fixed and random effects with the RE-EM tree estimation method. They find the performance of the MEml model to be better than the performance of classical machine learning methods, especially when the number of repeated observations increased. This indicates that MEml is able to take advantage of the correlation between observations to obtain more accurate models. Any supervised machine learning algorithm can be used in the MEml model, however Ngufor et al. (2019) focus only on tree-based algorithms for interpretability. In our research, the MEml model is explored for the simulation of Hb concentrations.

4.4 Machine learning algorithms for Hb simulation

For the application of the MEml model to the simulation of Hb concentrations, one property that should be taken into consideration is that Hb concentrations cannot be negative. Therefore, the machine learning methods that we used in the MEml model should only predict non-negative Hb concentrations. Wah, Nasaruddin, Voon, & Lazim (2012) and Beshah & Hill (2010) show that tree-based prediction algorithms and K-nearest neighbours (KNN) are appropriate machine learning algorithms for the prediction of non-negative variables. For both methods it holds that the predictions are in between the minimum and maximum value of the response variable in the training set. Since the Hb concentrations in the training set are always non-negative, all predictions will be non-negative as well.

Haghani, Sedehi, & Kheiri (2017) show that an artificial neural network (ANN) is another suitable machine learning algorithm for the prediction of non-negative variables. They compare statistical models like the ZIP and ZINB to ANN for the prediction of a zero-inflated count variable, which is always non-negative. Chang (2005) compares the performance of a negative binomial regression model and ANN for the prediction of vehicle accident frequency. Both studies show that that the ANN model is a good alternative for the classic, statistical models.

Nonetheless, we are still left with the problem of having excess zeros for the Hb concentrations in our data set. Due to this imbalanced distribution of the training sample, a high accuracy can be obtained even when there are large prediction errors for non-zero Hb concentrations. A common

solution for imbalanced data is to use resampling techniques. Cochran (1977) present a solutions for sampling imbalanced data sets, called stratified sampling. In stratified sampling the entire data set is divided into clusters or groups. Then, a random sample is drawn from each cluster. The percentage of observations sampled from each cluster does not need to be equal to their population representation.

In this paper, we used stratified sampling to increase the number of samples from the individuals with a lesion and decrease the number of samples from healthy individuals. After resampling, we trained the MEml model of Ngufor et al. (2019) using tree-based methods, KNN and ANN. To the best of our knowlegde, a comparison of different mixed-effect machine learning models, beyond tree-based algorithms, has not been done before.

5 Data

All data for this research is obtained from the Dutch national CRC screening programme during the period 2014-2019. Between 2014 and 2019 a population of 3,597,486 persons participated at least once in the CRC screening programme. Up to three screening rounds are available in the data set, since individuals did not participate in every screening round, or individuals were invited just for one or two screening rounds. A total of 3,474,328 persons participated in the first round, followed by 1,681,626 and 299,606 persons in the second and third screening round, respectively. For every faecal immunochemical test (FIT) a participant has taken, the data set includes the Hb concentration detected with the test (y). In addition, for every participant, the data set includes the age at the time of the screening (x_{AGE}) and the sex (x_{SEX}).

For participants that obtained a positive FIT and subsequently underwent a follow-up colonoscopy, the findings of follow-up examination are available in the data set. Hence, if a colonoscopy was performed, we know whether the participant is healthy or has adenomas/cancer (x_{STAGE}). In the data set a distinction is made between non-advanced and advanced adenomas. If CRC is found, the data set includes whether the person has CRC stage I, II, III or IV. Persons with a positive FIT for whom the results of any follow-up examination was missing, are excluded. For participants with a negative FIT, there is also no follow-up examination and therefore the health condition of the colon is unknown. A solution for this problem of missing data is presented in Section 5.1.

Since previous FIT results can contain valuable information to predict the current FIT result, we also constructed three new variables using the provided data set. The first variable (x_{PREV}) is

a binary variable which is 1 if the participant had an undetectable Hb concentration (between 0 and 2.6 μg/g faeces) in the previous screening round, thus two years ago. An undetectable Hb concentration two years ago decreases the current risk of having CRC (Grobbee et al., 2017), and therefore we expect a negative correlation between x_{PREV} and the current Hb concentration. For FITs in the first screening round this variable is always equal to 0.

The second variable (x_{SEQNR}) is the sequence number of the FIT. A higher sequence number indicates that the individual has participated in more screening tests before. With more frequent screening, there is a higher chance of detecting cancer in a earlier stage. Generally, the more progressed the cancer the higher the Hb concentration found in the stool and, therefore, we expect individuals that are more frequently screened to obtain less extreme Hb concentrations. In addition, individuals with a positive FIT are referred to the hospital for further examination and not invited for the regular FIT screening again after two years. Consequently, high sequence numbers are more likely to belong to healthy individuals with low Hb values.

The third variable (x_{MAX}) is the maximum Hb concentration measured in the previous FITs of a participant. A participant that had a negative FIT before but with a Hb concentration just below the cut-off level, has a higher chance of having an adenoma or cancer compared to a patient with no blood in their stool in previous screening rounds (Grobbee et al., 2017). Therefore, we expect the current FIT result to be positively correlated with the maximum Hb concentration found in earlier tests. For FITs in the first screening round this variable is always equal to 0. An overview of all variables is shown in Table 1.

Table 1: Description of the variables included in the data set.

Variable	Description	Details/ Range
y	The Hb concentration detected with the FIT.	0-437.14
x_{AGE}	The age of the participant.	55-78
x_{SEX}	The sex of the participant.	M = male, F = female
x_{STAGE}	The heath condition of the colon.	1 = healthy, 2 = non-advanced adenoma, 3 = advanced adenoma, 4 = cancer stage I, 5 = cancer stage II, 6 = cancer stage III, 7 = cancer stage IV
x_{PREV}	Indicates if the individual had an undetectable Hb concentration (between 0 and 2.6 μg/g faeces) in the screening round two years ago.	0 = No, 1 = Yes
x_{SEQNR}	Sequence number of the FIT (e.g. 2 indicates that it is the second FIT taken by the individual).	1-3
x_{MAX}	The maximum Hb concentration measured in previous FITs.	0-46.98

5.1 Imputing missing values

For participants with a negative FIT, there is no follow-up examination and therefore the health condition of the colon is unknown. However, since a greater share of the FITs is negative, omitting the FITs with an unknown stage would imply losing a lot of observations. In addition, we want our model to accurately predict the Hb concentration both above and below the cut-off level. This is not possible if the model is trained only with data above the cut-off level of 47 μg/g. Therefore, we replaced the missing data with new values by applying Multivariate Imputation by Chained Equations (MICE) (Van Buuren, 2007).

The MICE algorithm uses a series of models to impute the missing data. Each variable with missing values is modeled conditional upon the other variables in the data set. In the observed data set we only have information about the stage variable if the FIT is positive, therefore we aimed to add information about the stage variable for negative FITs using the MISCAN-Colon model. With MISCAN-Colon we can simulate a population that resembles the Dutch population in the available data set. In the simulated data set the health condition of the colon at the time of the FIT is known as well as the final conclusion of the FIT (positive/negative). However, for the simulated data set we do not know the exact Hb concentration. Even though we have no knowledge about the exact Hb concentrations, including more information on the stage variable could add to the accuracy of the imputation of the missing values. Figure 4 illustrates how we combined the observed data with the simulated data. Next to the variables stage and Hb value, we added the variable age. This variable is known for all observations.

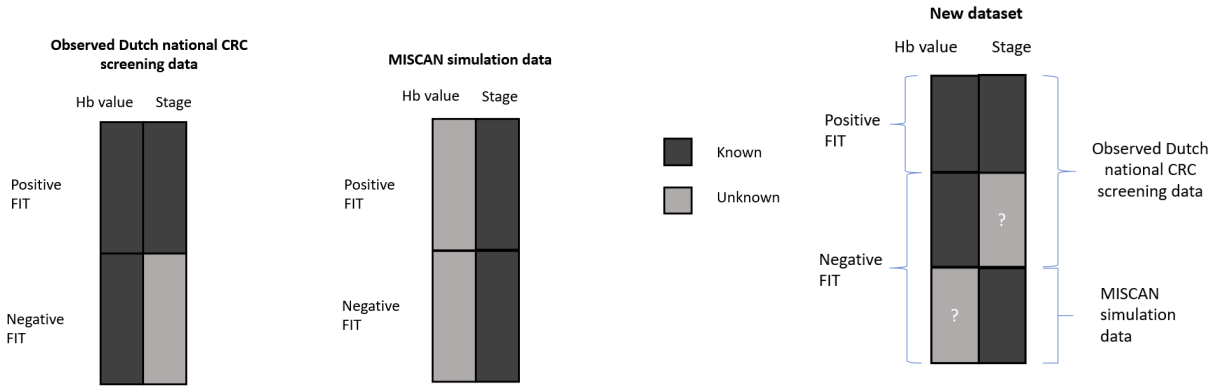


Figure 4: Combining the observed Dutch national CRC screening data with data simulated in MISCAN-Colon in order to obtain the marginal distributions of both the Hb values and the stages.

The chained equation approach in the MICE process can be described in five steps:

1. For each variable, the missing values are randomly drawn with replacement from the known values of that variable in the data set.
2. The imputed values of the Hb concentrations are set back to missing.
3. A linear regression of Hb concentration on age and stage is run using all cases where the Hb concentration is observed. All missing Hb values are imputed using the estimated coefficients of the linear regression.
4. The originally missing values of the stage variable are set back to missing.
5. Predictive Mean Matching is used to replace the missing values of the stage variable.

Steps 2-5 are repeated until convergence, resulting in a data set without missing values. Since we want as little as possible imputed values in our final data set, we remove the MISCAN simulation data once the imputation is done. Thus, we are left with the original observed Dutch national CRC screening data, only now with an imputed stage for every observation for which the stage was unknown.

5.2 Splitting longitudinal data in a train and test set

Since our data set contains repeated measures from the same individual, it is not a good idea to randomly split the data set in a train and test set. This could lead to one observation of an individual being in the training set and another observation of that same individual in the test set, which may bias the test set performance. Therefore, instead of randomly splitting the FIT results in a train and test set, we ensured that the same individual is not represented in both the test and training set. This means that if a person is selected for the training data then all observations belonging to that person are included in the training set. Similarly, if a person is selected for the test set of the data then all observations for that person are included in the test set.

The fact that we use longitudinal data should also be taken into account when we use a validation set to find the optimal hyperparameters for the different machine learning algorithms. In the same way that we created the train and test set, we ensured that if an individual is in the validation set, is in not in the training or test set.

For computational reasons the training set and validation set contain 200,000 observations each, leaving the remaining observations to test the performance of the models. The test set contains a total of 2,546,197 observations.

6 Methodology

In this section, we describe the different econometric methods that are used in this paper. We start with a description of our benchmark model, the mixed-effect ZINB model. Next, we explain the MEml model, followed by a description of the performance measures. Finally, we describe how we implement the MEml model into MISCAN-Colon.

6.1 Notation

The response variable y_{in} is a continuous variable that represents the Hb concentration of participant i found at his n_{th} FIT, where $i \in 1, ..., I$ and $n \in 1, ..., N$, with I the total number of participants and N the total number of FITs for each participant.

For each test of participant i, we have a vector X_{in} of fixed-effect (i.e. population-level) features. The benchmark ZINB model uses the following features to simulate the Hb concentrations:

$$X_{in} = \{1, x_{AGE}, x_{SEX}, x_{STAGE}\},$$

where the first element of the vector X_{in} represents the intercept. The variables x_{AGE}, x_{SEX} and x_{STAGE} are explained in more detail in Table 1 in Section 5.

In the models that we propose, we want to include more information on previous FITs of an individual. Therefore, we added x_{PREV}, x_{SEQNR} and x_{MAX} to the feature set (see Section 5):

$$X_{in}^{NEW} = \{1, x_{AGE}, x_{SEX}, x_{STAGE}, x_{PREV}, x_{SEQNR}, x_{MAX}\}$$

In addition, each participant i has its own risk factor γ_i. This is a random variable that has the same value for every FIT of participant i. It is responsible for the correlation between multiple FITs of the same participant. γ_i is assumed to follow a normal distribution: $\gamma_i \sim N(0, \sigma^2)$.

Phase one

6.2 Mixed-effect zero-inflated negative binomial model

Our benchmark model for the simulation of Hb concentrations is a mixed-effect zero-inflated negative binomial (ZINB) model, which is a specific type of the generalized linear mixed models (GLMM). The mixed-effect ZINB model is a discrete simulation model, meaning that the simulated Hb values are always non-negative integers. The ZINB model assumes that there are two

distinct data generation processes. With probability p_{in}, the Hb concentration is below the detection limit and with probability $(1 - p_{in})$ the Hb concentration is drawn from a mixed-effect negative binomial (NB) distribution.

A mixed-effect model consists of two parts: the fixed-effects part and the random-effects part. Fixed-effects represent relations between the features and the Hb concentration irrespective to which individual the FIT belongs. They are also called population-level effects. Random-effects, on the other hand, are used to account for the correlation between multiple observations of the same individual. The mixed-effect ZINB model is written as:

$$y_{in} \sim \begin{cases} \text{below detection limit} & \text{with probability } p_{in} \\ NB(y_{in}|X_{in}, \gamma_i; \mu_{in}, \theta) & \text{with probability } 1 - p_{in}, \end{cases} \tag{1}$$

where μ_{in} and θ are the mean and dispersion parameter of the NB distribution, respectively.

The probability p_{in} depends on the stage the participant is in. For participants in a healthy stage, the probability of having zero Hb in their stool is larger than for participants with cancer. The probability is assumed to be related to the stage through a logit link function (see Equation 2). Although including the other features to determine p_{in} could improve the accuracy of the mixed-effect ZINB model, it would also complicate the implementation and calibration in MISCAN-Colon considerably since the number of parameters would increase substantially.

$$\text{logit}(p_{in}) = \log\left(\frac{p_{in}}{1 - p_{in}}\right) = \alpha^T x_{STAGE}, \tag{2}$$

where α is a vector of coefficients.

The mean μ_{in} of the NB distribution depends both on the fixed and random effects, and is defined as:

$$\mu_{in} = \mu(X_{in}, \gamma_i) = \exp\{\beta^T X_{in} + \gamma_i\}, \tag{3}$$

where β represents the coefficients of the fixed effects (i.e. population-level effects).

This model is used as a benchmark model with the new set of variables (X_{in}^{NEW}). In addition, we created a second model in which we extend the mixed-effect ZINB model to allow for non-linearities through the input features. In this model polynomials of the numeric features and interactions with categorical variables are added to the feature set X_{in}^{NEW}. All possible features that we consider to include in this model are displayed in Section A of the Appendix.

Since we can create a large number of polynomial and interaction features, the selection of features is important. There are a number of ways to select features, including statistical techniques or a priori feature selection based on expert knowledge. We used both strategies to choose our features. First, we used the opinion of experts to make a subset of possible features that are plausible and relevant for our application. Subsequently, we applied a statistical method called backward elimination, to the subset of features based on expert knowledge. Chatterjee & Hadi (2015) recommend using backward elimination over the alternative forward selection if the number of candidate features is not larger than your sample size, because backward elimination is better able to handle multicollinearity than forward selection.

Backward elimination Backward elimination is an iterative procedure that begins with all candidate features in the model. It can be described in five steps:

1. Choose a significance level, for example 5% (0.05).
2. Fit the model with all features.
3. Consider the feature with the highest p-value. If the p-value is larger than the significance level, go to step 4. Otherwise, your feature set is ready.
4. Remove this feature from the feature set.
5. Fit the model with the new feature set, and go to step 3.

The final feature set we obtained for the extented ZINB model after backward elimination consists of the following variables:

$$X_{in}^{EXT} = \{1, x_{AGE}, x_{SEX}, x_{STAGE}, x_{PREV}, x_{SEQNR}, x_{MAX}, x_{MAX}^2, x_{PREV} * x_{MAX}, x_{AGE} * x_{SEX}, x_{AGE} * x_{MAX}\}$$

6.3 Mixed-effect machine learning model

Ngufor et al. (2019) propose a mixed-effect machine learning (MEml) model as alternative for the mixed-effect ZINB model. As opposed to the mixed-effect ZINB model, the fixed-effects of the MEml model are estimated using machine learning algorithms. The MEml model is therefore able to simulate continuous variables. The expected Hb concentration is determined as follows:

$$E(y_{in}) = f(X_{in}) + \gamma_i, \tag{4}$$

where the function f is estimated using different machine learning algorithms.

For estimation, the MEml model follows the Expectation Maximisation (EM) algorithm (Dempster, Laird, & Rubin, 1977). This algorithm alternatively estimates the fixed and random effects until convergence. If the random effects, γ_i, are known, we can estimate $f(X_{in})$ using a machine learning algorithm with adjusted response variable: $y_{in} - \gamma_i$. Since the MEml framework makes the assumption that all correlation between observations of the same individual is captured in γ_i, the adjusted response variable is assumed to be independent and identically distributed. Therefore, we do not need to make adjustments for having multiple observations of the same person when training the machine learning algorithm.

Next, if the fixed effects $f(X_{in})$ are estimated, we can estimate the random effect using the traditional estimation method of GLMM with the fixed effects corresponding to $\hat{f}(X_{in})$. Algorithm 1 provides the pseudo-code for the EM algorithm. More details on the estimation procedure of the MEml model can be found in the paper of Ngufor et al. (2019).

Algorithm 1 Expectation Maximisation algorithm for the mixed-effect machine learning model.

Initialize the random effects $\hat{\gamma}_i = 0$.
for $k = 1, 2, ..., max\_iterations$ **do**
    E-step:
    Compute $y^*_{in} = y_{in} - \hat{\gamma}_i$.
    Train a machine learning algorithm to estimate $f(X_{in})$ using the adjusted response variable y^*_{in}.
    M-step:
    Estimate γ_i using the estimation methods of GLMM with fixed effects $\hat{f}(X_{in})$.
end for

Stratified sampling As mentioned in Section 4, the excess number of zeros in our data set requires resampling before we can train the MEml models. The reason for the large proportion of zeros in our data set is that the majority of the individuals does not have any lesions in the colon and only a very small part of the data set consists of individuals with colorectal cancer (CRC).

We used stratified sampling (Cochran, 1977) to reduce this imbalance in the training set. Stratified sampling is a type of sampling method in which we, first, split the data set into clusters and then randomly select elements from those clusters.

In more detail, we start by categorizing the individuals in our trainingset based on their worst lesion. CRC is defined as the worst lesion, followed by advanced adenoma, then non-advanced adenoma and finally no lesions. Next, we randomly over-sample individuals in the categories advanced adenoma and CRC. This means that we randomly choose an individual with replacement

from the category advanced adenoma or CRC. For this individual we duplicate the faecal immunochemical test (FIT) result with the worst lesion. We continue this process until we have 35,000 FIT results belonging to people with an advanced adenoma and 100,000 FIT results for people with CRC. In this way the advanced adenoma category is almost the same size as the non-advanced category. We draw more samples for the CRC stage to make sure that there are enough observations to separate CRC in CRC stage I, II, III and IV.

Then, we randomly under-sample those individuals that never had any lesion in the colon. When a person is selected, we remove all his FIT results from the data set. This process is continued until there are only 30,000 FIT results left that belong to people who do not have any lesions. Consequently, the number of FITs belonging to an individual with a healthy colon constitutes only 15% of the data set instead of 67%. Table 2 shows the results after resampling.

Table 2: Percentage of FIT results in each stage before and after resampling.

	Before resampling	**After resample**
No lesions	67.2%	14.8%
Non-advanced adenoma	20.4%	18.9%
Advanced adenoma	12.0%	17.2%
CRC	0.4%	49.2%
stage I	0.21%	23.3%
stage II	0.09%	10.0%
stage III	0.12%	13.2%
stage IV	0.03%	2.7%

We now elaborate on the five different machine learning algorithms that were used in the MEml models in this paper. Two of them are very simple models: K-nearest neighbours and regression tree. The other three methods are more complex 'black-box' algorithms: random forest, gradient boosting machine and artificial neural network. Together, these five algorithms cover a large variety of different methods, where each algorithm has his own advantages and disadvantages.

6.3.1 Decision trees

Decision tree is a powerful method for both classification and regression. It is a non-parametric method that is capable of detecting complex relations between the response variable and the features (Brezigar-Masten & Masten, 2012). A decision tree consists of branches and nodes, where the topmost node is called the root and the terminal nodes are called leaves. Each node that is not a leaf represents a question on a feature in X_{in} that will result in the subdivision of the data into two

or more mutually exclusive subsets. To decide on which feature the split is based, a cost function is minimized. After the feature is chosen, the node is split into child nodes. If there are only two child nodes the tree is called binary, whereas a tree with more than two child nodes is called non-binary. This splitting procedure continues until pre-determined stopping criteria are met. There exist many different tree-based algorithms ranging from very interpretable to more complex algorithms. To have a diverse selection of methods we explore three different types of tree-based algorithms: classification and regression trees (CART), random forest (RF) and gradient boosting machine (GBM). In the following paragraphs we first describe the CART algorithm, followed by an explanation of RF and GBM.

Classification And Regression Tree One of the most famous decision tree algorithms is the Classification And Regression Tree (CART) algorithm (Breiman et al., 1984). Since we want to predict a continuous variable, namely the Hb value, we use the CART algorithm to construct a regression tree. The constructed regression tree is a binary tree and takes the mean squared error as a cost function for splitting. In our regression tree, each leaf node holds a possible Hb value prediction. The Hb value prediction belonging to leaf node j is calculated as the mean Hb value of all data points from the training set belonging to leaf node j. Algorithm 2 shows how the regression tree is constructed.

Algorithm 2 Pseudocode for tree construction using the CART algorithm.

1: Define the maximum depth of the tree d.
2: Start at the root node and set $depth\_tree = 0$.
3: **for** $depth\_tree < d$ **do**
4: For each feature, calculate the cost function C for every possible split.

$$C = \sum_{i \in I_{tr}} \sum_{n=1}^{N} (y_{in} - \hat{y}_{in})^2,$$

where y_{in} and $\hat{y}_{in}$ are the observed and predicted Hb value of the n_{th} FIT of person i, respectively. I_{tr} is the subset of individuals in the training set.
5: Select and execute the split that results in the smallest cost function C.
6: Update the depth of the tree.
7: **end for**

A critical choice for the construction of a regression tree is when to stop splitting. This is important, because extremely deep trees can lead to overfitting your data. In order to prevent this, we applied cross-validation to get the optimal maximum depth d of the tree.

After the regression tree is constructed, it can be used to make predictions. We follow the decision path of the new data point until we reach the leaf node it belongs to. The prediction of the new data point is then defined as the Hb value belonging that particular leaf node.

The biggest advantage of the CART algorithm is that it is simple to understand, interpret and visualize. In addition, the algorithm is fast, leading to a manageable computation time. A disadvantage of the CART algorithm is that it is prone to overfitting and very unstable (Ngufor et al., 2019). A small change in the input data can have a large impact on the final prediction. An alternative is to build many trees and combine the predictions, which is called ensembling. Two of the most popular tree-based ensemble methods are Random Forest (RF) (Breiman, 2001) and Gradient Boosting Machine (GBM) (Friedman, 2001). RF uses the idea of bagging, while GBM is based on the idea of boosting. Ensemble methods like RF and GBM have been proven to achieve greater accuracy than CART, however due to increasing complexity they loose interpretability (Seni & Elder, 2010).

Random Forest Random Forest (RF) (Breiman, 2001) is an ensemble machine learning algorithm that builds a predefined number of independent decision trees and combines their predictions. The RF algorithm uses the idea of bagging. This means that every single tree in a RF is trained with a different bootstrap sample from the original training data. Bootstrapping is a specific way of sampling with replacement in which the new sample has the same number of observations as the original data. This means that the training set for a single tree can contain the same data point multiple times. After training the individual regression trees with the different bootstrap samples, they are combined through averaging.

In addition, RF only uses a random subset of the features $s \subseteq S$ at each splitting node of the tree, with S representing the total feature set. The 'best' split is then obtained from the features in s instead of S. The randomness in the selection of the training sample and features ensures variety in the trees, which helps to reduce the amount of overfitting. Algorithm 3 shows a pseudocode for the construction of a RF.

Algorithm 3 Pseudocode for the construction of a Random Forest

1: Define the maximum depth of the decision trees d, the number of trees T in the forest and the number of features l to consider when looking for the best split.
2: **for** $t = 1, 2, ..., T$ **do**
3: Draw a bootstrap sample B_t from the training set.
4: Start at the root node and set $depth\_tree = 0$.
5: **for** $depth\_tree < d$ **do**
6: Randomly select l features from the feature set
7: For each selected feature, calculate the cost function C for every possible split.

$$C = \sum_{i \in B_t} \sum_{n=1}^{N} (y_{in} - \hat{y}_{in})^2,$$

where y_{in} and $\hat{y}_{in}$ are the observed and predicted Hb value of the n_{th} FIT of person i, respectively.
8: Select and execute the split that results in the smallest cost function C.
9: Update the depth of the tree.
10: **end for**
11: **end for**
12: Combine the trees through averaging:

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} \hat{f}_t(x),$$

RF improves the prediction accuracy by reducing the variance. According to Friedman, Hastie, Tibshirani, et al. (2001), the variance of a RF can be calculated as follows:

$$\sigma^2_{RF} = \rho\sigma^2 + \frac{1-\rho}{T}\sigma^2, \tag{5}$$

where σ^2 represents the variance of a single tree, ρ the correlation between the trees and T the number of trees in the forest.

From Equation 5 we can deduce that there are three ways in which we can decrease the variance of a RF:

1. Decrease the correlation between the trees. This can be done by selecting a smaller subsample of features at each splitting node.
2. Decrease the variance of each individual tree. One way to achieve this is to decrease the maximum depth of each tree. Regression trees with a larger depth have a higher probability of overfitting the training data, leading to a higher variance.

3. Increase the number of trees in the forest. If the number of trees in the forest is large enough, the second term of Equation 5 goes to zero.

This shows us that the choice of the hyperparameters of the RF is important. Therefore, we used cross-validation to get the optimal maximum depth d of the trees and the optimal number of trees T in the forest. We choose l to be equal to five features.

Gradient Boosting Machine Gradient Boosting Machine (GBM) (Friedman, 2001) is an ensemble machine learning algorithm that builds consecutive decision trees in which each tree tries to correct the loss function from the previous tree. In regression trees, this loss function is the mean squared error. At each iteration a new decision tree is fitted to the current residual and added to the previous model to update the residual. This process, called boosting, continues until the predefined maximum number of iterations is reached.

Instead of using the full training sample, a random subsample is selected at each iteration to fit the regression tree. The randomization improves the accuracy and reduces the computational cost by a factor equivalent to the factor of the subsampling (Touzani, Granderson, & Fernandes, 2018).

How much each tree contributes to the final prediction depends on the learning rate η $(0 < \eta \leq 1)$. The higher the value for η, the more the added decision tree contributes to the final prediction. There is a trade-off between the learning rate η and the total number of decision trees used in the model. A smaller value for η implies that a larger number of decision trees is needed to achieve convergence.

The following pseudo-code provides an illustration of the GBM algorithm:

Algorithm 4 Gradient Boosting Machine algorithm

1: Define the number of iterations K, the depth of the regression trees d, the factor of subsampling λ and the learning rate η.
2: Initialization: set the mean value of y as an initial guess of $\hat{f}$ and the residual $r_0 = y_{in} - \hat{f}$.
3: **for** $k = 1, 2, ..., K$ **do**
4: Randomly select a subsample S_k from the full training sample with the number of observations corresponding to the factor λ.
5: Use S_k to fit regression tree $\hat{f}_k$ of depth d to the residuals r_{k-1} (see Algorithm 2).
6: Update $\hat{f}$ by adding the regression tree: $\hat{f}(x) \longleftarrow \hat{f}(x) + \eta \hat{f}_k(x)$.
7: Update the residual: $r_k \longleftarrow r_{k-1} - \eta \hat{f}_k(x)$.
8: **end for**

We used cross-validation to obtain the optimal number of iterations K, the depth of the regression trees d and the learning rate η. The factor of subsampling, λ, in our GBM is equal to 0.7,

meaning that the subsample contains 0.7 times the number of observations of our full training set.

6.3.2 K-Nearest Neighbours

K-Nearest Neighbours (KNN) is one of the simplest algorithms when there is little or no prior knowledge about the distribution of the data. KNN uses the K closest observations to predict the Hb concentration of a new data point. The distance between the training data and a new data point is measured by the Euclidean distance. A lower Euclidean distance indicates that there are more similarities in the features.

Distance measures like the Euclidean distance are affected by the scale of the features. Since we do not want our distance measure to be biased towards features with a higher magnitude, it is important that we standardize the features before calculating the distance.

After determining which K observations are closest, we use these observations to make the prediction. The prediction of the new data point is defined as the weighted average of the Hb concentrations of the K nearest neighbours. In this way, closer neighbours have a greater influence than neighbours which are further away. Algorithm 5 shows the pseudocode for KNN.

Algorithm 5 K-nearest neighbours algorithm

Input: X: features of training data, y: response variable belonging to X, x: features of the new data point.

1: Define the number of neighbours K.
2: Standardize X and x.
3: **for** $i \in I_{tr}$, $n \in \{1, 2, ..., N\}$ **do**
4: Calculate the Euclidean distance between X_{in} and x:

$$d(X_{in}, x) = \sqrt{\sum_{j=1}^{p} (X_{in,j} - x_j)^2},$$

where p is the total number of features.
5: **end for**
6: Sort the distances from low to high and select the K observations from X with the smallest distance to x.
7: For the prediction, take the weighted average of the K closest neighbours:

$$\hat{f}(x) = \frac{\sum_{X_{in} \in C} d(X_{in}, x)^{-1} \, y_{in}}{\sum_{X_{in} \in C} d(X_{in}, x)^{-1}},$$

where C is the subset of the K closest neighbours.

The choice of k is important. A small value for K leads to a low bias, but high variance. A large

value for K, on the other hand, leads to a high bias and low variance. This is known as the bias-variance trade-off. The proper choice of K depends on the data, therefore we used cross-validation to find the optimal hyperparameter value.

6.3.3 Artificial neural network

The last algorithm is an Artificial Neural Network (ANN). ANNs have a layered structure with different artificial neurons, also called nodes. This structure consists of an input layer, one or more hidden layers and an output layer, which are all connected. For regression purposes, an ANN only has one node in the output layer. In our setting, the output node represents the predicted Hb concentration. An example of an ANN with one hidden layer is shown in Figure 5.

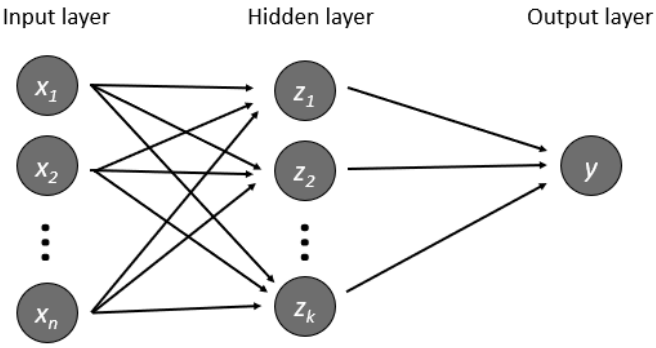


Figure 5: An illustration of an artificial neural network with n nodes in the input layer, k nodes in the hidden layer and one node in the output layer.

The connections between different nodes in the ANN have weights. The higher the weight, the greater influence one node has on another. During the learning process of the ANN these weights adjust.

The input layer of our ANN consists of p nodes, where p is equal to the number of features in X_{in}. Thus, our input layer consists of six nodes, where each node is given the value of a feature. The nodes in the hidden layer(s) and output layer receive their values from the nodes in the preceding layer in two steps. An illustration is shown in Figure 6. First, it adds up each input from the preceding layer multiplied by its corresponding weight. Then, the value is passed on to an activation function. These functions are necessary to learn the complex patterns in the data and introduce non-linearity into the network. The activation function calculates the output value for the node. Both for the hidden layers and the output layer, we use the ReLU activation function, which is defined as $g(x) = \max(0, x)$. Hence, the values for hidden layers of the neural network

are calculated as follows:

$$z_j^1 = g\Big(\sum_{h=1}^{p} w_{hj}^1 x_h\Big) \qquad \forall j \in \{1, ..., J\} \tag{6}$$

$$z_j^l = g\Big(\sum_{s} w_{sj}^l z_s^{l-1}\Big) \qquad \forall j \in \{1, ..., J\}, \forall l \in \{2, ..., L\} \tag{7}$$

where z_j^l is the j_{th} node of hidden layer l, w_{hj}^l is the weight of node h in layer $l-1$ on node j in hidden layer l, p is the total number of features, J is the total number of nodes in each hidden layer and L is the total number of hidden layers.

Using the values of the hidden layers, we can calculate the value of the output layer:

$$\hat{f}(x) = g\Big(\sum_{t} w_t^{L+1} z_t^L\Big) \tag{8}$$

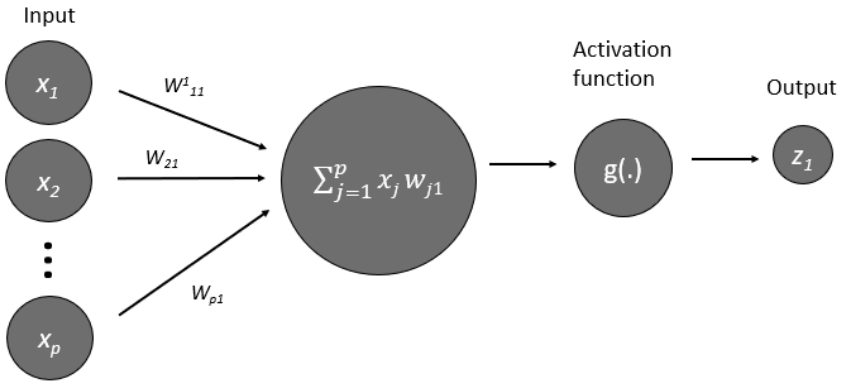


Figure 6: An illustration of the calculation of the value for the first node of the first hidden layer. The values for the other nodes in the hidden layer(s) and the value for the node in the output layer are obtained in the same way.

Two decisions must be made regarding the hidden layers of the ANN: the number of hidden layers, L, and the number of nodes, J, in the hidden layers. To determine L and J we used cross-validation.

To estimate our weights, we used the back-propagation algorithm, which is one of the most famous estimation algorithms to train ANNs (Rumelhart, Hinton, & Williams, 1986). This algorithm uses stochastic gradient descent to look for the weights that minimize the error function, which is the squared loss in this case.

6.4 Performance measures

We evaluated the performance of the different models in two ways. First, we looked at the accuracy of the individual predictions. Second, we evaluated the distribution of the predictions as a whole.

6.4.1 Individual prediction level

We evaluated the individual predictions by looking at the root mean squared error (RMSE), mean absolute error (MAE) and median absolute error (MedAE). In addition, we used the Diebold-Mariano (DM) test as a statistical measure of the relative forecasting performance between two models in order to determine if one model generates significantly better predictions than another.

Root mean squared error The RMSE is the square root of the average of the squared errors across all observations in the test set. By taking the square root, the RMSE provides a performance measure in the same unit as the response variable. This makes the results more interpretable. A lower RMSE value indicates better predictions. The RMSE is calculated as:

$$RMSE = \sqrt{\frac{1}{O_t} \sum_{i \in I_t} \sum_{n=1}^{N} (y_{in} - \hat{y}_{in})^2}, \tag{9}$$

where $\hat{y}_{in}$ and y_{in} are the predicted and observed Hb concentration for the n_{th} FIT of individual i, respectively. O_t is the total number of observations in the test set and I_t is the subset of individuals in the test set.

Mean absolute error The MAE is the average of the absolute value of the errors. Similar to the RMSE, this performance measure has the same unit as the response variable. Differently from the RMSE, the MAE does not penalize large prediction errors more than smaller ones, because the MAE does not square the errors. Therefore, the MAE is less sensitive to large prediction errors than the RMSE. A lower MAE value indicates better predictions. The MAE is calculated as:

$$MAE = \frac{1}{O_t} \sum_{i \in I_t} \sum_{n=1}^{N} |y_{in} - \hat{y}_{in}|. \tag{10}$$

Median absolute error Opposed to the RMSE and MAE, the MedAE is completely insensitive to outliers. This is because it is the median of the absolute value of the errors, and the median is not affected by values in the tails. In combination with the MAE we can determine if there are large

outliers among the errors. A large difference between the MAE and MedAE suggests that there are large prediction errors present. Similar to the RMSE and MAE, this performance measure has the same unit as the response variable. The MedAE is calculated as:

$$MedAE = \text{median}\Big\{ |y_{in} - \hat{y}_{in}| \;\Big|\; i \in I_t,\, n \in \{1, \ldots, N\} \Big\} \tag{11}$$

Diebold-Mariano test The Diebold-Mariano (DM) test (Diebold & Mariano, 2002) enables statistical comparison of the accuracy of two competing forecasts of the same data. The null hypothesis of the DM test is that the forecasting performance of the two compared models is equal. Stated otherwise, the difference in loss function between two models is equal to zero.

The DM test assumes that the forecast errors are normally distributed. However, since we predict the Hb values using different MEml models, it is questionable whether the forecast errors are indeed normally distributed. Therefore, we applied the adjusted DM test, as proposed by Harvey, Leybourne, & Newbold (1997). Two modifications are made to the test statistic: (i) an unbiased estimator of the variance of $\overline{d}$, the sample mean of the differences in loss function, is used. This results in a bias correction of $\sqrt{\frac{n-1}{n}}$, where n is the number of predicted Hb values. (ii) the DM statistic is compared to critical values from a Student-t distribution with $n-1$ degrees of freedom, rather than the standard normal. The latter corrects for forecast error distributions with heavier tails than the normal distribution. The modified DM test statistic is defined as:

$$DM = \sqrt{\frac{n-1}{n}} \frac{\overline{d}}{\sqrt{\text{Var}(\overline{d})}} \sim t(n-1), \tag{12}$$

where d is the vector of differences in loss function between two different sets of predictions and $\overline{d}$ is its average over all observations. we define d as $d = |e_1| - |e_2|$, with e_j is the difference between the predicted Hb value of model j and the observed Hb value.

As the number of hypotheses increases, so does the probability of finding at least one statistically significant test statistic just by chance. Since we pairwise compare the predictions of seven different models, we need to correct the significance level for this problem of multiple comparison. We use the Bonferroni method (Dunn, 1961) to do this, which defines the new significance level as:

$$\alpha_{new} = \frac{\alpha_{original}}{c},$$

where c is total number of comparisons.

6.4.2 Distribution level

To evaluate how much the predicted Hb distribution resembles the observed Hb distribution, we looked at the following measures: Kullback-Leibler divergence and the percentages of positive FIT results. The latter can easily be calculated as the proportion of the total (predicted) FIT results that is above 47 μg/g.

Kullback-Leibler divergence The KL divergence (Cover & Thomas, 1991) is a non-symmetric measure that determines how much information is lost when the distribution $q(x)$ is used to approximate the distribution $p(x)$. In our case $q(x)$ represents the predicted Hb distribution and $p(x)$ the observed Hb distribution. The KL divergence measure is defined as follows:

$$D_{KL}(p(x), q(x)) = \sum_{x \in X} p(x) \ln \left(\frac{p(x)}{q(x)} \right). \tag{13}$$

The lower the KL divergence, the better we have matched the true distribution with our approximation.

Phase two

6.5 Calibration in MISCAN-Colon

After evaluating all the mixed-effect models, we used the most promising one as a starting point for the implementation in MISCAN-Colon. In Section 3 we saw that the distribution of Hb values in the data set has a two-peaked shape. To capture this two-peaked shape in our simulation model we made an adjustment to the random-effects. Instead of drawing the random effects from a normal distribution, we drew the random effects from a bimodal distribution. The bimodal distribution we used is a mixture of two gamma distributions. The density of the bimodal distribution is defined as follows:

$$g(x; \xi_1, \theta_1, \rho_1, \xi_2, \theta_2, \rho_2) = p f(x; \xi_1, \theta_1, \rho_1) + (1 - p) f(x; \xi_2, \theta_2, \rho_2), \tag{14}$$

where $0 < p < 1$ and $f(x; \xi_i, \theta_i, \rho_i)$ is the pdf of the gamma distribution with location ξ_i, scale θ_i and shape ρ_i.

This bimodal distribution is stage dependent, meaning that the parameters $\theta_1, \rho_1, \theta_2$ and ρ_2 are different for the stages healthy, non-advanced adenoma, advanced adenoma and CRC. We used grid search to find the parameters for which the shape of the simulated distributions resembles the

shape of the observed distributions the most.

In addition to the adjustment of the random effects, we added a threshold component to our Hb simulation model. This threshold is used to regulate the number of zeros that is simulated with our model and works in the following way. We draw a random variable u from the standard uniform distribution. If u is above a fixed threshold T, the MEml model is used to obtain a Hb value. If u is below T, a Hb value of 0 is assigned. The threshold T depends on the stage, so we have four different thresholds: T_H for the healthy stage, T_{NAA} for non-advanced adenoma, T_{AA} for advanced adenoma and T_{CRC} for CRC. Since we do not know the optimal values for T_H, T_{NAA}, T_{AA} and T_{CRC}, they are calibrated in MISCAN-Colon. This means that we searched for the thresholds for which the output of MISCAN-Colon is as close as possible to a known calibration target. The calibration targets are calculated as the number of positive FITs, negative colonoscopies, detected adenomas and detected CRC divided by the total number of people screened. These values were obtained from the data set described in Section 5 and are displayed in Section B of the Appendix.

7 Results

In this section, we first evaluate the performance of the different mixed-effect models by looking at the individual predictions. Second, we evaluate which model obtains a Hb distribution that resembles the observed Hb distribution the most. Finally, we show the results after implementing the MEml model in MISCAN-Colon.

Results phase one

7.1 Individual prediction level

Table 3 shows the final hyperparameters obtained using cross-validation. Using the parameters in this table, we obtained the predicted Hb concentration for our test set. Table 4 shows the RMSE, MAE and MedAE for every model. We observe that the mixed-effect ZINB (MeZINB) model has very high values for both the RMSE and MAE. This is likely caused by the fact that this model predicts a number of unrealistically high Hb values. The maximum predicted Hb value with the MeZINB model has a value of 43,044,931, while the maximum Hb value in the observed data set is equal to 437. Adding interaction and polynomial terms to the feature set (MeZINB-ext) does not seem to improve the predictions when looking at the RMSE and MAE.

The MedAE of 0.00 for the MeZINB models indicates that at least 50% of the predicted Hb

values are exactly equal to their true values. This is likely due to the fact that the MeZINB models predict a lot of zeros. Since the test set contains a high number of healthy people with zero blood in their stool, a prediction of zero is correct in most cases. The large difference between the MedAE and MAE confirms that there are large prediction errors present.

For the mixed-effect machine learning (MEml) models the MedAE is in the range 6-9. This means that that at least 50% of the errors are smaller than 9 μg for every MEml model. The difference between the MedAE and MAE is smaller in comparison to the MeZINB models which indicates that there are less outliers among the errors.

The RMSE of the MEml models is approximately 700 times smaller than the RMSE of the MeZINB model and the MAE approximately 4 times smaller. The lowest RMSE and MAE belong to the MeKNN model, however the differences in RMSE and MAE among the machine learning models are small. Therefore, we look at the Diebold-Mariano test statistic to investigate the accuracy of the competing forecasts (Table 5).

The test statistic is positive and significant for every entry in the first two column, except for the test statistic between the extended MeZINB model and the 'standard' MeZINB model. This confirms that the MeZINB model is significantly outperformed by every MEml model. In addition, it confirms that adding interaction and polynomial terms does not improve the accuracy of the MeZINB predictions.

In the third column, we observe a negative and significant test statistic for the MeRF and MeGBM model. This means that the the MeCART model is the tree-based model with the most accurate predictions. Between the MeRF model and MeGBM model we obtain a test statistic of -0.707 with a p-value of 0.479, indicating that there is no significant difference in prediction accuracy between these two models.

The last two rows consist of only positive and significant test statistics, except for the second-last column. This indicates that all tree-based models are significantly outperformed by MeKNN and MeANN. Of these two, MeKNN has the most accurate predictions.

Table 3: Optimal hyperparameters found using grid-search with cross-validation.

MeCART	max depth: 6
MeRF	max depth: 6, number of trees: 200
MeGBM	max depth: 4, number of trees: 200, learning rate: 0.1
MeKNN	number of neighbors: 11
MeANN	number of hidden layers: 1, number of nodes in hidden layer: 10

Table 4: Accuracy of the haemoglobin predictions measured with the RSME, MAE and MedAE. The test set consists of 2,546,197 observations from Dutch national CRC screening data.

	MeZINB	**MeZINB-ext**	**MeCART**	**MeRF**	**MeGBM**	**MeKNN**	**MeANN**
RMSE	28627.35	33709.02	39.59	39.79	39.66	38.54	39.74
MAE	92.47	137.93	22.11	22.27	22.29	21.03	21.69
MedAE	0.00	0.00	8.25	8.50	8.78	7.09	6.68

Table 5: Diebold-Mariano test statistic using the absolute error as loss function. The test is run on 2,546,197 observations. A positive statistic indicates that the model in the row outperforms the model in the column. A negative statistic indicates that the model in the column outperforms the model in the row. The value between parentheses beneath the test statistic denotes the *p*-value of the statistic. * indicates that the difference in loss function is significant with a 0.24% significance level. The Bonferroni correction is used to adjust for multiple comparison testing (5%/21 = 0.24%) .

	MeZINB	**MeZINB-ext**	**MeCART**	**MeRF**	**MeGBM**	**MeKNN**	**MeANN**
MeZINB	x						
MeZINB-ext	-3.518	x					
	(0.000*)						
MeCART	3.922	5.483	x				
	(0.000*)	(0.000*)					
MeRF	3.913	5.475	-7.454	x			
	(0.000*)	(0.000*)	(0.000*)				
MeGBM	3.912	5.474	-8.158	-0.707	x		
	(0.000*)	(0.000*)	(0.000*)	(0.479)			
MeKNN	3.982	5.534	51.055	58.273	59.181	x	
	(0.000*)	(0.000*)	(0.000*)	(0.000*)	(0.000*)		
MeANN	3.945	5.503	19.812	27.230	27.944	-30.649	x
	(0.000*)	(0.000*)	(0.000*)	(0.000*)	(0.000*)	(0.000*)	

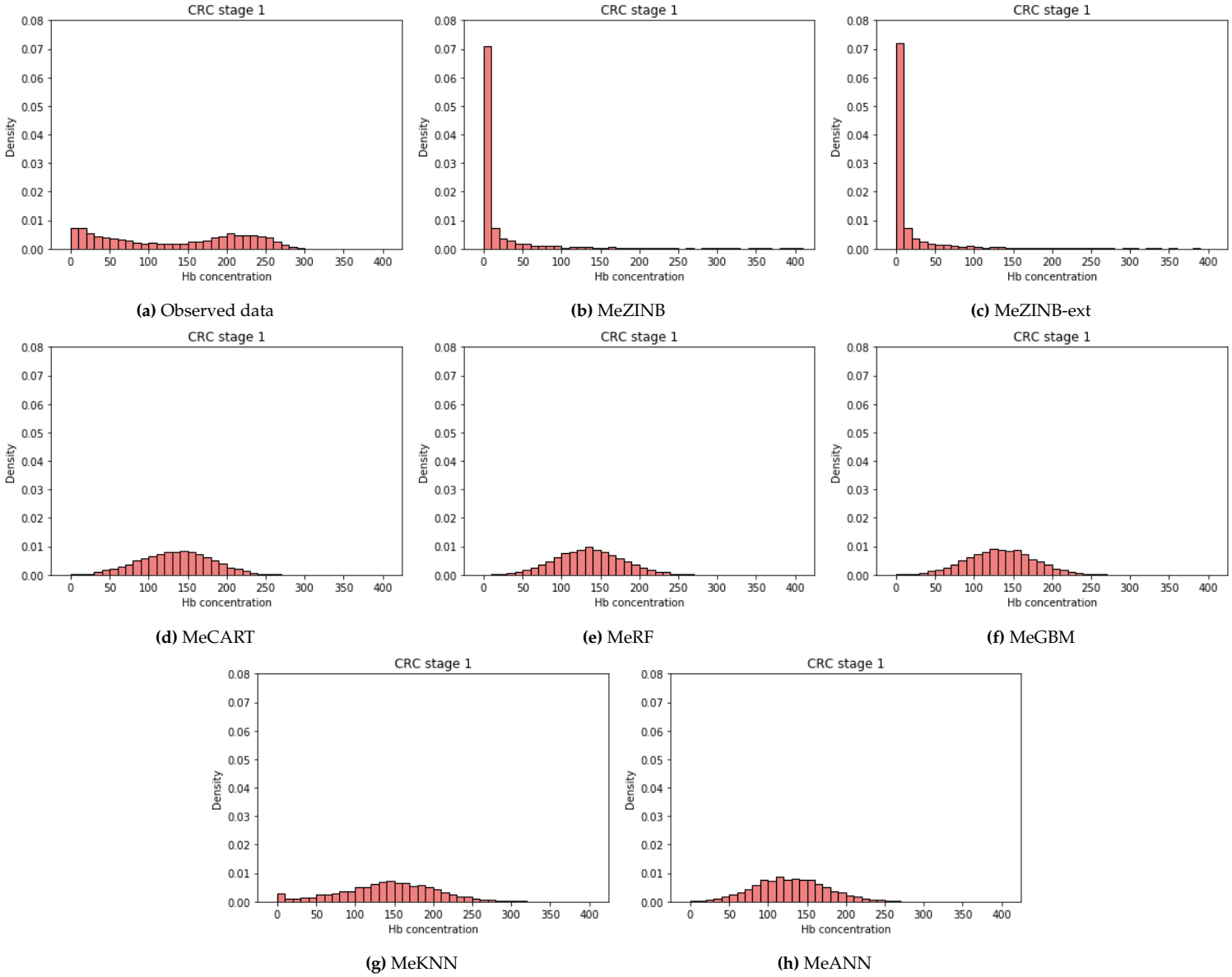
7.2 Distribution level

In addition to having an individual Hb prediction that is as accurate as possible, we also want the full distribution of Hb values to resemble the observed distribution as closely as possible. The results in terms of the KL divergence are shown in Table 6. For the stages healthy and adenoma, the KL divergence is below 1 for every model. This indicates a relatively good resemblance with the observed Hb distribution. For these stages, the ZINB distribution comes closest to the observed distribution.

Table 6: Kullback-Leibler divergence score between the actual haemoglobin probability distribution and the predicted haemoglobin probability distributions. The Kullback-Leibler divergence score is calculated for the stages healthy (H), non-advanced adenoma (NAA), advanced adenoma (AA), colorectal cancer stage I (CRC1), colorectal cancer stage II (CRC2), colorectal cancer stage III (CRC3) and colorectal cancer stage IV (CRC4).

	H	**NAA**	**AA**	**CRC1**	**CRC2**	**CRC3**	**CRC4**
MeZINB	0.03	0.07	0.25	3.57	7.25	5.08	12.56
MeZINB-ext	0.04	0.06	0.24	3.40	7.23	5.56	13.01
MeCART	0.54	0.47	0.67	2.00	4.10	2.68	5.03
MeRF	0.54	0.50	0.73	2.54	4.65	3.29	4.88
MeGBM	0.56	0.50	0.75	2.48	3.32	2.51	3.25
MeKNN	0.52	0.41	0.42	0.65	1.08	0.46	3.02
MeANN	0.46	0.51	0.80	1.97	3.06	3.37	3.53

For the cancer stages, all seven models have more difficulty to match the observed Hb distribution. The difference in distribution becomes very clear in Figure 7 and Section C of the Appendix. Figure 7 shows that the MeZINB models predicts way too many Hb values in the range 0-10 and not enough predictions above 50 for individuals with CRC stage I. This explains the high KL divergence for the ZINB models (see Table 6). The distributions of the MEml models are more similar to the observed distribution, however all of them have too little predictions below 50, too many between 100 and 200, and too little above 200. Table 6 shows that, for the cancer stages, the distributions obtained with the MeKNN model are most similar too the observed Hb distribution. This is also visible in Figure 7, where we can see that the MeKNN model has the most predictions below 50 and above 200 of all MEml models.



(a) Observed data **(b)** MeZINB **(c)** MeZINB-ext

(d) MeCART **(e)** MeRF **(f)** MeGBM

(g) MeKNN **(h)** MeANN

Figure 7: Distributions of the haemoglobin concentrations (Hb) in the observed data set and predicted by the different mixed-effect models. The Hb distributions are shown for the stage colorectal cancer stage I.

The problems with the predicted distributions are confirmed if we compare the percentage of positive predicted FITs with the actual percentage of positive FITs (see Table 7). We find that the ZINB models predict too little positive FITs for all stages, except for the healthy stage. The difference is especially large (above 50%) for the cancer stages. The MEml models, on the other hand, predict too many Hb values above 47 for every stage.

Table 7: The percentage of positive (predicted) FITs, calculated for the stages healthy (H), non-advanced adenoma (NAA), advanced adenoma (AA), colorectal cancer stage I (CRC1), colorectal cancer stage II (CRC2), colorectal cancer stage III (CRC3) and colorectal cancer stage IV (CRC4). Each percentage is based on 2,546,196 observations.

	H	NAA	AA	CRC1	CRC2	CRC3	CRC4
Observed	1.2	5.2	15.9	72.8	82.6	77.6	96.6
MeZINB	1.3	3.6	7.6	22.1	22.7	20.7	25.9
MeZINB-ext	1.6	4.0	7.7	21.1	21.9	22.1	24.8
MeCART	12.2	16.6	30.6	96.6	99.9	99.0	99.3
MeRF	12.3	16.5	30.4	97.8	99.6	99.5	99.6
MeGBM	12.5	16.7	29.7	97.9	99.4	98.6	99.8
MeKNN	11.1	16.4	28.0	92.7	92.8	91.3	98.0
MeANN	10.5	17.6	35.5	96.6	99.3	99.1	100

Results phase two

7.3 Calibration in MISCAN-Colon

In Section 7.1 and 7.2, we found that the MeKNN model obtained the best results. However, since the differences are small, we prefer to use the MeCART model for the implementation in MISCAN-Colon. There are three reasons why we prefer the MeCART model to the MeKNN model. Firstly, making predictions with a KNN algorithm is a lot more computationally intensive than with a regression tree. Therefore, the MeCART model is more suitable for large population runs with a lot of data in MISCAN-Colon. Secondly, the obtained regression tree can be expressed in a readable form. This representation is easy to understand and provides a lot of insights into the most discriminatory features. Thirdly, the MeCART model is easier to adjust in comparison to the MeKNN model. A regression tree can be simplified for example by decreasing the maximum depth of the tree.

In Section 7, we found that the optimal depth for the MeCART model was 6. However, a tree with 64 (2^6) leaves is less ideal for visual representation. Therefore, we simplified the model by reducing the maximum depth to 4, thus obtaining a regression tree with 16 leaves. The final regression tree that we used in MISCAN-Colon is displayed in Figure 8.

The tree shows us which features are most influential for the prediction of Hb values. We see that the stage variable is a very important feature. The worse the lesion, the higher the predicted Hb value. Age is also an influential variable. We observe that a younger person with a lesion always gets a higher Hb prediction than an older person with the same lesion. This is both the case for adenomas and CRC.

Another interesting insight that the tree provides is that individuals with a higher maximum

Hb value in previous FITs, get a higher prediction for the current FIT. This confirms our expectation that previous FIT results can contain valuable information for the prediction of current Hb values. The other two variables, x_{PREV} and x_{SEQNR}, are not in the tree and therefore do not influence the predictions made with our model.

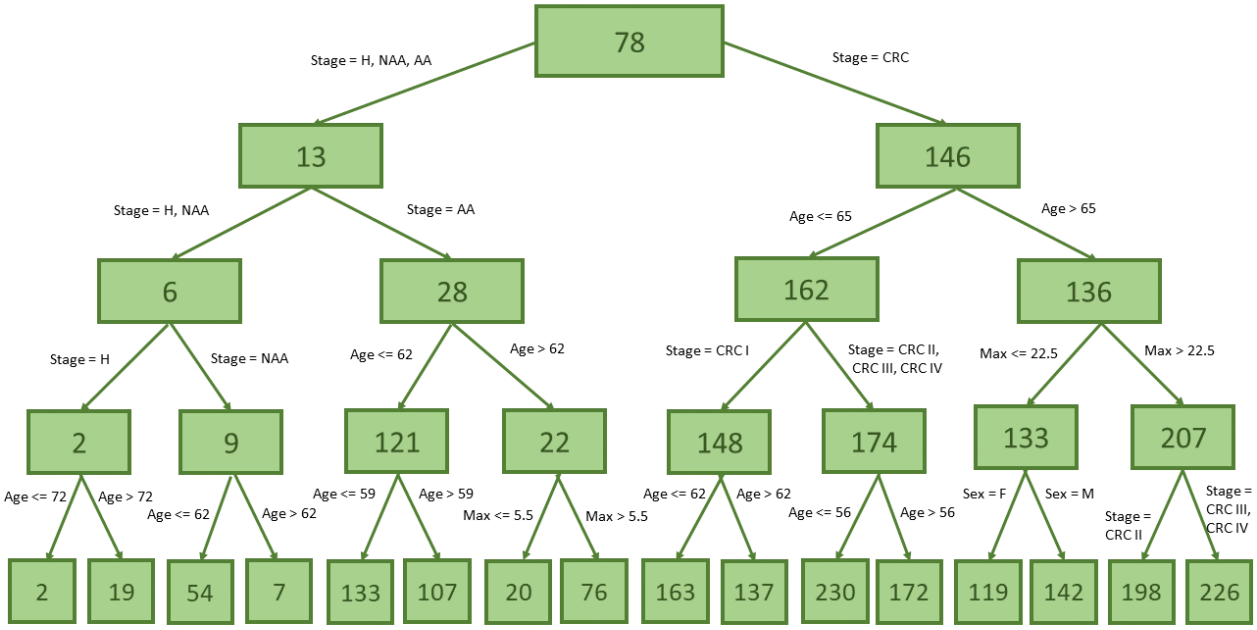


Figure 8: Fixed-effect part of the trained MeCART model. The predicted Hb values are displayed in the green boxes.

Tables 8 and 9 show the results for the random effects and the threshold component. As expected, the threshold value for the healthy stage is high, thus a lot of zeros are simulated for people in the healthy stage. For CRC, on the other hand, the threshold is small, meaning that most of the bloodvalues are simulated using the MEml model. The thresholds of the adenoma stages are in between the thresholds of the healthy and CRC stage, with the non-advanced adenoma stage simulating more zeros than the advanced adenoma stage.

Table 8: The parameter values of the bimodal distribution from which the random effects are drawn. Every stage has its own bimodal distribution, with NAA representing non-advanced adenoma, AA advanced adenoma and CRC colorectal cancer.

	Healthy	**NAA**	**AA**	**CRC**
ξ_1	0	0	-100	-150
θ_1	50	50	50	45
ρ_1	1	1	1	2
ξ_2	0	0	0	0
θ_2	15	15	20	20
ρ_2	15	15	10	4
p	0.75	0.75	0.75	0.5

Table 9: Parameter values for the thresholds after calibration. T_H, T_{NAA}, T_{AA} and T_{CRC} represent the threshold for the stage healthy, non-advanced adenoma, advanced adenoma and CRC, respectively.

T_H	0.982
T_{NAA}	0.665
T_{AA}	0.344
T_{CRC}	0.038

Figure 9 shows how close the results from MISCAN-Colon are to the calibration targets if we use the parameter values from Tables 8 and 9. Overall, we observe that the output from MISCAN-Colon is quite close to the calibration targets. In particular for the number of detected cancers, we see that MISCAN-Colon detects almost the same number of cancers as found in the data. For the number of positive FITs, negative colonoscopies and detected adenomas we see that the output from MISCAN-Colon is close to the calibration targets for the age groups 60-64, 65-69 and 70-74. For the age group 55-59, MISCAN-Colon simulates more positive FITs, negative colonoscopies and detected adenomas than observed in the data. For the age group 75-79, on the other hand, MISCAN-Colon simulates less positive FITs, negative colonoscopies and detected adenomas than observed in the data.

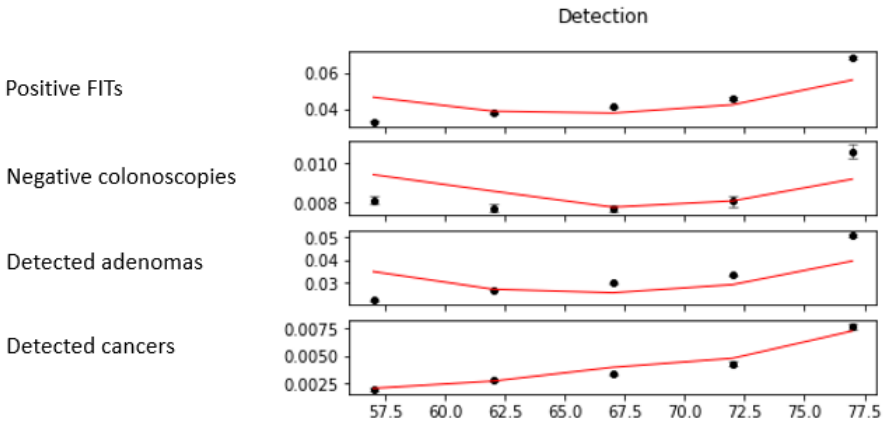
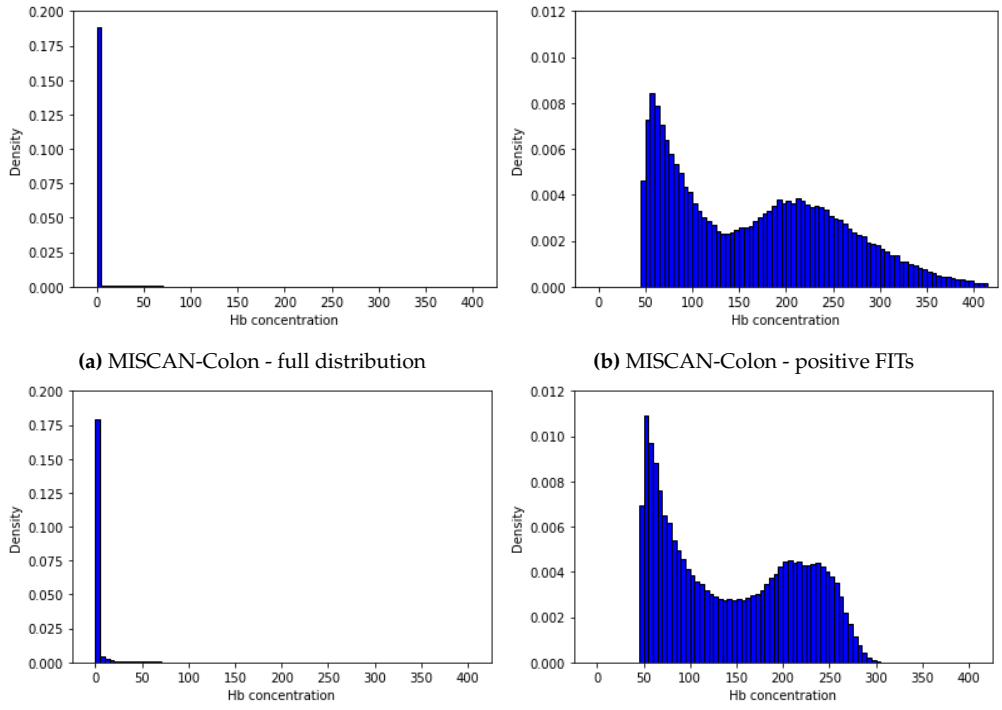


Figure 9: Results of the calibration run with 1e7 individuals, with the black dots representing the calibration targets and in red the results from MISCAN-Colon. The calibration targets are calculated as the number of positive FITs, negative colonoscopies, detected adenomas and detected CRC divided by the total number of people screened. The calibration targets are calculated per age group with groups: 55-59, 60-64, 65-69, 70-74 and 75-79. The age is shown on the horizontal axis and the targets are displayed at the mean age of the group.

If we compare Figure 10a with Figure 10c, we see that the number of zeros simulated in MISCAN-Colon is comparable to the number of zeros in the data set. Furthermore, we observe in Figure 10b that our Hb simulation model is able to capture the two-peak pattern that is present in the data (Figure 10d). The two peaks in the simulated distribution have the same locations as the peaks in the observed distribution, however the height of the first peak is smaller in the simulated distribution. In addition, we observe that the simulated distribution has more Hb values above 300 in comparison to the observed distribution.



(a) MISCAN-Colon - full distribution

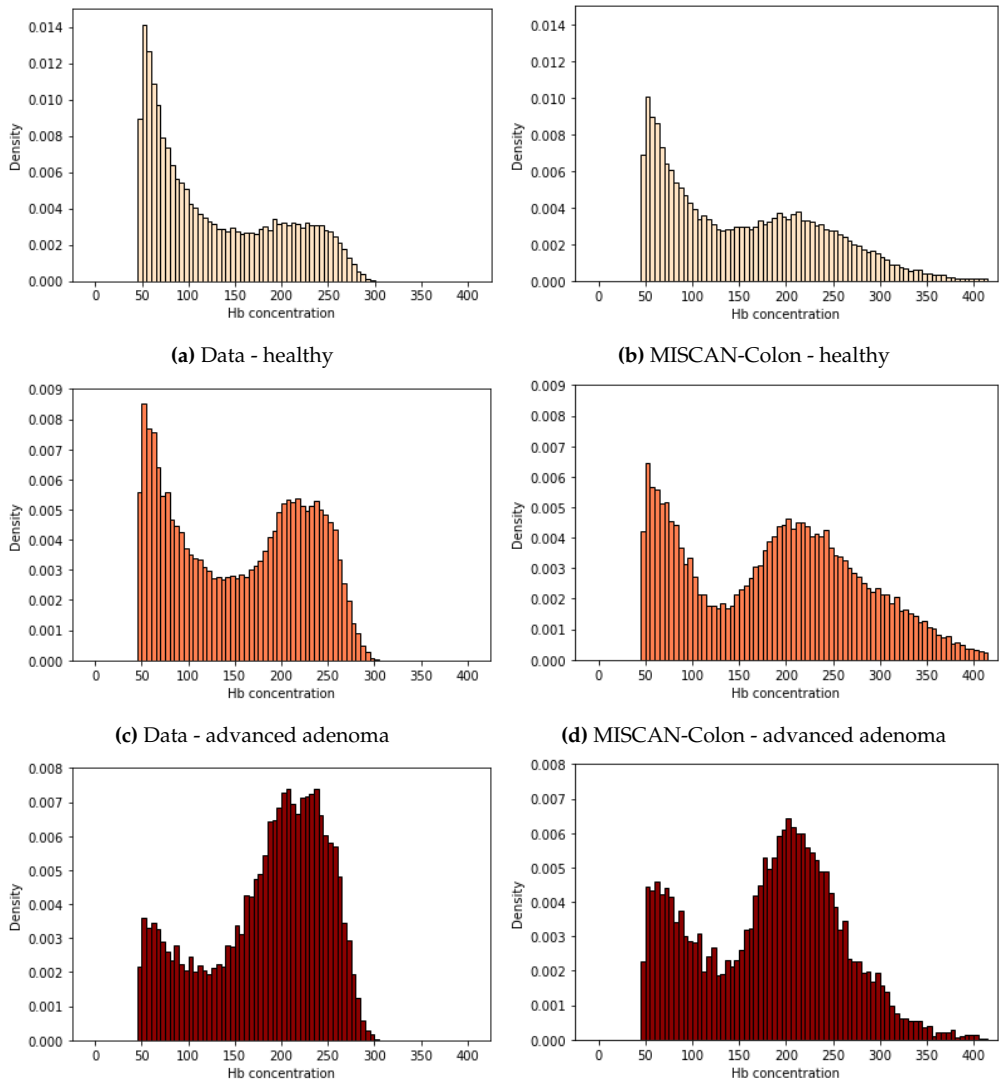
(b) MISCAN-Colon - positive FITs

(c) Data - full distribution

(d) Data - positive FITs

Figure 10: Distributions of the haemoglobin concentrations (Hb) simulated by MISCAN-Colon and distributions of the Hb concentrations obtained from the Dutch CRC screening programme. 1e6 individuals are simulated with MISCAN-Colon. The left figures show the full distribution, while the right figures show only the Hb values above the cut-off of 47 $\mu g/g$.

The Hb simulation model in MISCAN-Colon is able to distinguish between the different stages (Figure 11 and Section D in the Appendix). Similar to the distribution of the observed data, we see that the worse the lesion, the lower the first peak of the distribution and the higher the second peak. Moreover, the extrema of the simulated distribution have the same location as the extrema of the observed distribution, roughly at Hb values 50, 150 and 225. However, there are also some differences. For the healthy and adenoma stages we see that the first peak of the simulated distribution is smaller than in the observed distribution. For the CRC stage, on the other hand, the first peak is higher in the simulated distribution. In addition, for every stage, we see that the right tail of the simulated distribution is too fat in comparison to the observed distribution.



(a) Data - healthy

(b) MISCAN-Colon - healthy

(c) Data - advanced adenoma

(d) MISCAN-Colon - advanced adenoma

(e) Data - CRC

(f) MISCAN-Colon - CRC

Figure 11: Distributions of the haemoglobin concentrations (Hb) simulated by MISCAN-Colon and distributions of the Hb concentrations obtained from the Dutch CRC screening programme for the stages healthy, advanced adenoma and CRC. 1e6 individuals are simulated with MISCAN-Colon. The figures show only the Hb values above the cut-off of 47 μg/g.

8 Conclusion and discussion

In this study, we propose to use a mixed-effect machine learning (MEml) model to simulate haemoglobin (Hb) concentrations in MISCAN-Colon. The MEml model was explored with tree-based machine learning algorithms, *K*-nearest neighbours and artificial neural network. We compared the performance of the MEml models to the performance of a mixed-effect zero-inflated negative binomial (ZINB) model. We conclude that the MEml models outperform the mixed-effect ZINB model sig-

nificantly. The MEml model with the best predictions uses a *K*-nearest neighbours algorithm, however the differences with the other MEml models are small. Since a regression tree can visualize the acquired knowledge in a more comprehensible way and is better in dealing with large data sets than *K*-nearest neighbours, we developed a simulation model in MISCAN-Colon using a mixed-effect regression tree.

Our results support the findings of Ngufor et al. (2019) that the MEml model is able to address data issues like non-linearity, variable interactions, and dependencies between groups of observations. Whereas Ngufor et al. (2019) did not find a significant difference between the performance of the MEml model and linear mixed-effect methods, we observed that the MEml models outperform the mixed-effect ZINB model significantly.

The key finding that the MEml model outperforms the mixed-effect ZINB model suggests that there is a non-linear relationship between the features and the Hb concentration. By visualizing the regression tree we were able to discover new insights into the relationships between the features and the Hb concentration. For example, for individuals with an adenoma or CRC, we found a negative correlation between age and Hb value. There is no literature that confirms this relationship, however O'Connell, Maggard, Livingston, & Cifford (2004) give one possible explanation. They state that CRC appears to be a more aggressive disease in younger people. It would be interesting to further investigate if younger people with an adenoma or CRC are indeed bleeding more in comparison to the older population.

The regression tree also showed that higher Hb values found in previous screening rounds predict a higher Hb value in the current screening test. This confirms that two people with a negative result in the previous screening round do not necessarily have the same risk of getting CRC. Therefore, personalised screening based on previous Hb concentrations shows potential to reduce the harms and increase the benefits of the screening programme.

Another important finding to take away from this thesis is that accuracy is not the only important factor when developing a model. For our application of the MEml model in the MISCAN-Colon model, interpretability and computation time were also essential. Since the accuracy of the MEml models was close, we made the decision to implement a mixed-effect regression tree which is based on a set of simple and transparent rules. By using a regression tree instead of *K*-nearest neighbours, we obtained not only a model that could predict the Hb concentration relatively fast, but also a model that improves our understanding of the relationships between different variables. In addition, *K*-nearest neighbours would require us to store all the training data, which is not pre-

ferred. For other applications of the MEml model, interpretability might be less important and therefore another machine learning algorithm could be more suitable.

This study also has some limitations. First, we used an imputed data set for the training and testing of the mixed-effect models since we had no knowledge of the stage of individuals with a negative FIT. If the imputed values deviate too much from the actual values, this could influence the trained models substantially. One possibility to decrease the uncertainty in the imputed data set would be to generate several imputation sets using the MICE package and take an average over the imputed data sets. This could lead to differences in the regression tree that we implemented in MISCAN-Colon.

Second, choosing the optimal hyperparameters for the MEml models is an important part of model tuning. We performed a grid search to find the best hyperparameters, however, due to long computation time, the grids used did not contain a large number of possible hyperparameters. Consequently, the steps between the different hyperparameters in the grid were fairly large. In the future, it would be better to perform the hyperparameter tuning in two rounds. In the first round, we find the best hyperparameter in a grid comparable to the one we used. Then, in the second round, we use a grid with smaller steps around the best hyperparameter found in round one. A more elaborate grid search could improve the accuracy of each MEml model and cause the results to be in favor of a different machine learning algorithm than *K*-nearest neighbours.

Finally, this study was mainly focused on the fixed-effect part of the mixed-effect model. While we did change the random effect from a normal distribution to a bimodal distribution, we did not consider other distributions for the random effects. It would be interesting to examine if there are other distributions that are more suitable for the simulation of Hb values in MISCAN-Colon.

References

Beshah, T., & Hill, S. (2010). Mining road traffic accident data to improve safety: role of road-related factors on accident severity in ethiopia. *AAAI Spring Symposium: Artificial Intelligence for Development*, *24*, 1173–1181.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Brezigar-Masten, A., & Masten, I. (2012). Cart-based selection of bankruptcy predictors for the logit model. *Expert Systems with Applications*, *39*(11), 10153–10159.

Chang, L.-Y. (2005). Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety science*, *43*(8), 541–557.

Chatterjee, S., & Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.

Cochran, W. G. (1977). *Sampling techniques*. 3rd ed. New York: Wiley.

Compton, C. C., & Greene, F. L. (2004). The staging of colorectal cancer: 2004 and beyond. *CA: a cancer journal for clinicians*, *54*(6), 295–308.

Cooper, K., Squires, H., Carroll, C., Papaioannou, D., Booth, A., Logan, R. F., … Tappenden, P. (2010). Chemoprevention of colorectal cancer: systematic review and economic evaluation. *NIHR Health Technology Assessment programme: Executive Summaries*.

Cottet, V., Jooste, V., Fournel, I., Bouvier, A., Faivre, J., & Bonithon-Kopp, C. (2012). Long-term risk of colorectal cancer after adenoma removal: a population-based cohort study. *Gut*, *61*(8), 1180–1186.

Cover, T. M., & Thomas, J. A. (1991). Elements of information theory. *Wiley series in telecommunications*.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Diebold, F. X., & Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business & economic statistics*, *20*(1), 134–144.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American statistical association*, *56*(293), 52–64.

Ferlay, J., Colombet, M., Soerjomataram, I., Dyba, T., Randi, G., Bettio, M., ... Bray, F. (2018). Cancer incidence and mortality patterns in europe: Estimates for 40 countries and 25 major cancers in 2018. *European Journal of Cancer*, *103*, 356–387.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Friedman, J. H., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York.

Gini, A. (2020). Microsimulation models to inform colorectal cancer screening decisions: From validated tools to tailoring recommendations. *PhD thesis*.

Grobbee, E. J., Schreuders, E. H., Hansen, B. E., Bruno, M. J., Lansdorp-Vogelaar, I., Spaander, M. C., & Kuipers, E. J. (2017). Association between concentrations of hemoglobin determined by fecal immunochemical tests and long-term development of advanced colorectal neoplasia. *Gastroenterology*, *153*(5), 1251–1259.

Habbema, J., Van Oortmarssen, G., Lubbe, J. T. N., & Van der Maas, P. (1985). The miscan simulation program for the evaluation of screening for disease. *Computer methods and programs in biomedicine*, *20*(1), 79–93.

Haghani, S., Sedehi, M., & Kheiri, S. (2017). Artificial neural network to modeling zero-inflated count data: application to predicting number of return to blood donation. *Journal of Research in Health Sciences*, *17*(3), 392.

Hajjem, A., Bellavance, F., & Larocque, D. (2010). Generalized mixed effects regression trees. *Mixed Effects Trees and Forests for Clustered Data*, *34*.

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting*, *13*(2), 281–291.

Kinreich, S., Meyers, J. L., Maron-Katz, A., Kamarajan, C., Pandey, A. K., Chorlian, D. B., ... others (2019). Predicting risk for alcohol use disorder using longitudinal data with multimodal biomarkers and family history: a machine learning study. *Molecular psychiatry*, 1–9.

Kuipers, E. J., & Grobbee, E. J. (2020). Personalised screening for colorectal cancer, ready for take-off. *Gut*, *69*(3), 403–404.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, *34*(1), 1–14.

Loeve, F., Boer, R., van Oortmarssen, G. J., van Ballegooijen, M., & Habbema, J. D. F. (1999). The miscan-colon simulation model for the evaluation of colorectal cancer screening. *Computers and Biomedical Research*, *32*(1), 13–33.

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2012). *Generalized linear models: with applications in engineering and the sciences* (Vol. 791). John Wiley & Sons.

Ngufor, C., Van Houten, H., Caffo, B. S., Shah, N. D., & McCoy, R. G. (2019). Mixed effect machine learning: a framework for predicting longitudinal change in hemoglobin a1c. *Journal of Biomedical Informatics*, *89*, 56–67.

O'Connell, J. B., Maggard, M. A., Livingston, E. H., & Cifford, K. Y. (2004). Colorectal cancer in the young. *The American journal of surgery*, *187*(3), 343–348.

Perveen, S., Shahbaz, M., Saba, T., Keshavjee, K., Rehman, A., & Guergachi, A. (2020). Handling irregularly sampled longitudinal data and prognostic modeling of diabetes using machine learning technique. *IEEE Access*, *8*, 21875–21885.

Ridout, M., Demétrio, C. G. B., & Hinde, J. (1998). Models for count data with many zeros. In *Proceedings of the xixth international biometric conference* (Vol. 19, pp. 179–192).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Sela, R. J., & Simonoff, J. S. (2012). Re-em trees: a data mining approach for longitudinal and clustered data. *Machine learning*, *86*(2), 169–207.

Seni, G., & Elder, J. F. (2010). Ensemble methods in data mining: improving accuracy through combining predictions. *Synthesis lectures on data mining and knowledge discovery*, *2*(1), 1–126.

Simon, K. (2016). Colorectal cancer development and advances in screening. *Clinical interventions in aging*, *11*, 967.

Stroup, W. W. (2012). *Generalized linear mixed models: modern concepts, methods and applications*. CRC press.

Toes-Zoutendijk, E., van Leerdam, M. E., Dekker, E., Van Hees, F., Penning, C., Nagtegaal, I., ... others (2017). Real-time monitoring of results during first year of dutch colorectal cancer screening program and optimization by altering fecal immunochemical test cut-off levels. *Gastroenterology*, *152*(4), 767–775.

Touzani, S., Granderson, J., & Fernandes, S. (2018). Gradient boosting machine for modeling the energy consumption of commercial buildings. *Energy and Buildings*, *158*, 1533–1543.

Trevisani, L., Zelante, A., & Sartori, S. (2014). Colonoscopy, pain and fears: Is it an indissoluble trinomial? *World journal of gastrointestinal endoscopy*, *6*(6), 227.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, *16*(3), 219–242.

Van Hees, F., Zauber, A. G., Van Veldhuizen, H., Heijnen, M. A., Penning, C., de Koning, H. J., ... Lansdorp-Vogelaar, I. (2015). The value of models in informing resource allocation in colorectal cancer screening: the case of the netherlands. *Gut*, *64*(12), 1985–1997.

Van Rossum, L. G., Van Rijn, A. F., Laheij, R. J., Van Oijen, M. G., Fockens, P., Van Krieken, H. H., ... Dekker, E. (2008). Random comparison of guaiac and immunochemical fecal occult blood tests for colorectal cancer in a screening population. *Gastroenterology*, *135*(1), 82–90.

Wah, Y. B., Nasaruddin, N., Voon, W. S., & Lazim, M. A. (2012). Decision tree model for count data. In *Proceedings of the world congress on engineering* (Vol. 4).

Winawer, S. J. (1999). Natural history of colorectal cancer. *The American Journal of Medicine*, *106*(1), 3–6.

Worthley, D. L., Cole, S. R., Esterman, A., Mehaffey, S., Roosa, N., Smith, A., ... Young, G. P. (2006). Screening for colorectal cancer by faecal occult blood test: why people choose to refuse. *Internal medicine journal*, *36*(9), 607–610.

Yau, K. K., Wang, K., & Lee, A. H. (2003). Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *45*(4), 437–452.

Zhao, J., Feng, Q., Wu, P., Lupu, R. A., Wilke, R. A., Wells, Q. S., ... Wei, W. (2019). Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction. *Scientific reports*, *9*(1), 1–10.

Appendices

A Feature set extended mixed-effect ZINB model

Table 10: Possible features to include in the extended mixed-effect ZINB model.

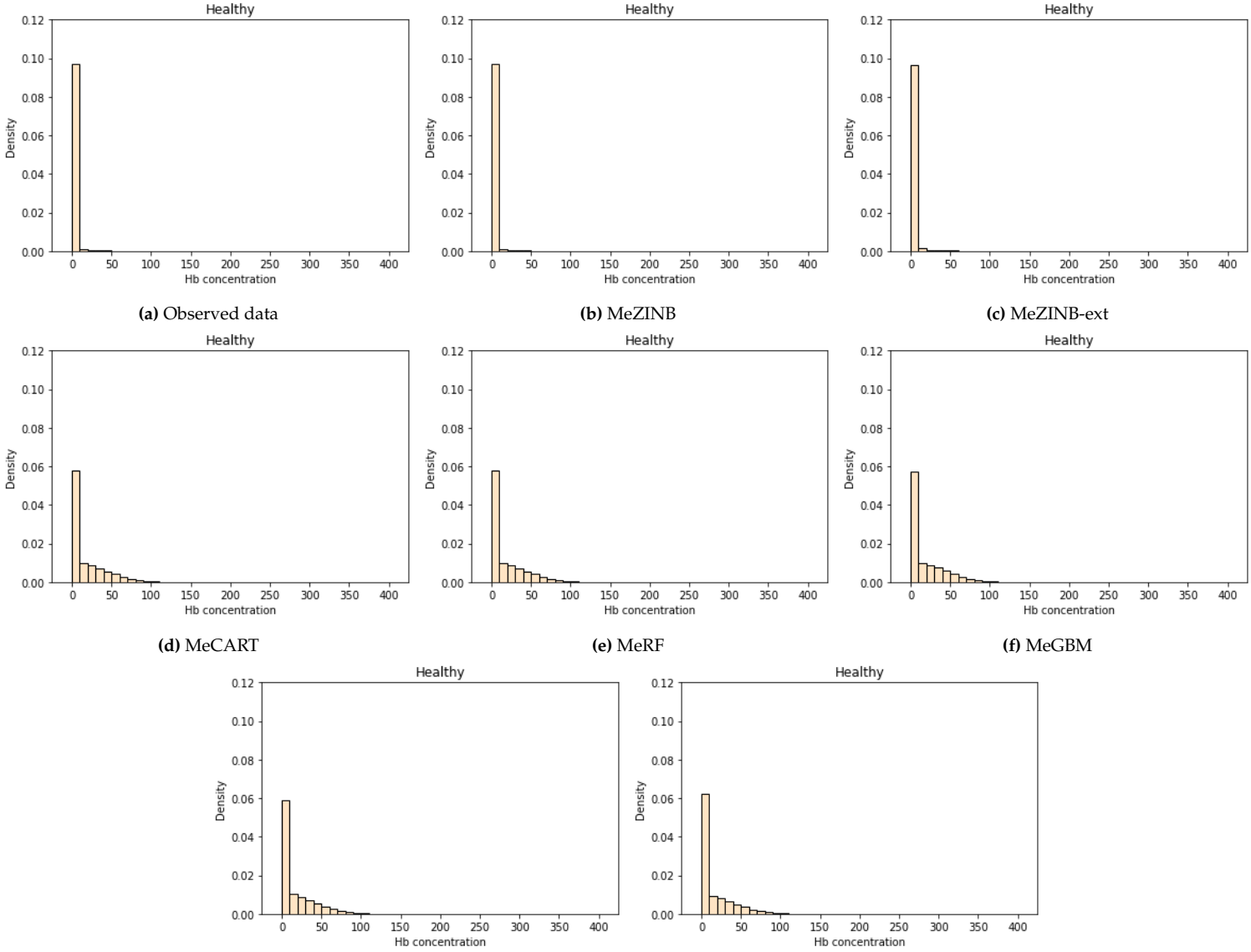
	Variable
1	age
2	sex
3	stage
4	prev
5	seqnr
6	max
7	age^2
8	age^3
9	age*sex
10	age*stage
11	age*prev
12	age*seqnr
13	age*max
14	sex*stage
15	sex*prev
16	sex*seqnr
17	sex*max
18	stage*prev
19	stage*seqnr
20	stage*max
21	prev*seqnr
22	prev*max
23	seqnr*max
24	$seqnr^2$
25	$seqnr^3$
26	max^2
27	max^3

B Calibration targets

Table 11: The calibration targets for the number of positive FITs, negative colonoscopies, detected adenomas and detected CRC divided by the total number of people screened (in %). The calibration targets are calculated per age group.

	55-59	**60-64**	**65-69**	**70-74**	**75-79**
Positive FITs	3.27	3.76	4.11	4.58	6.88
Negative colonoscopies	0.81	0.77	0.77	0.81	1.06
Detected adenomas	2.26	2.71	3.01	3.34	5.05
Detected CRC	0.20	0.27	0.33	0.43	0.77

C Distribution plots mixed-effect models

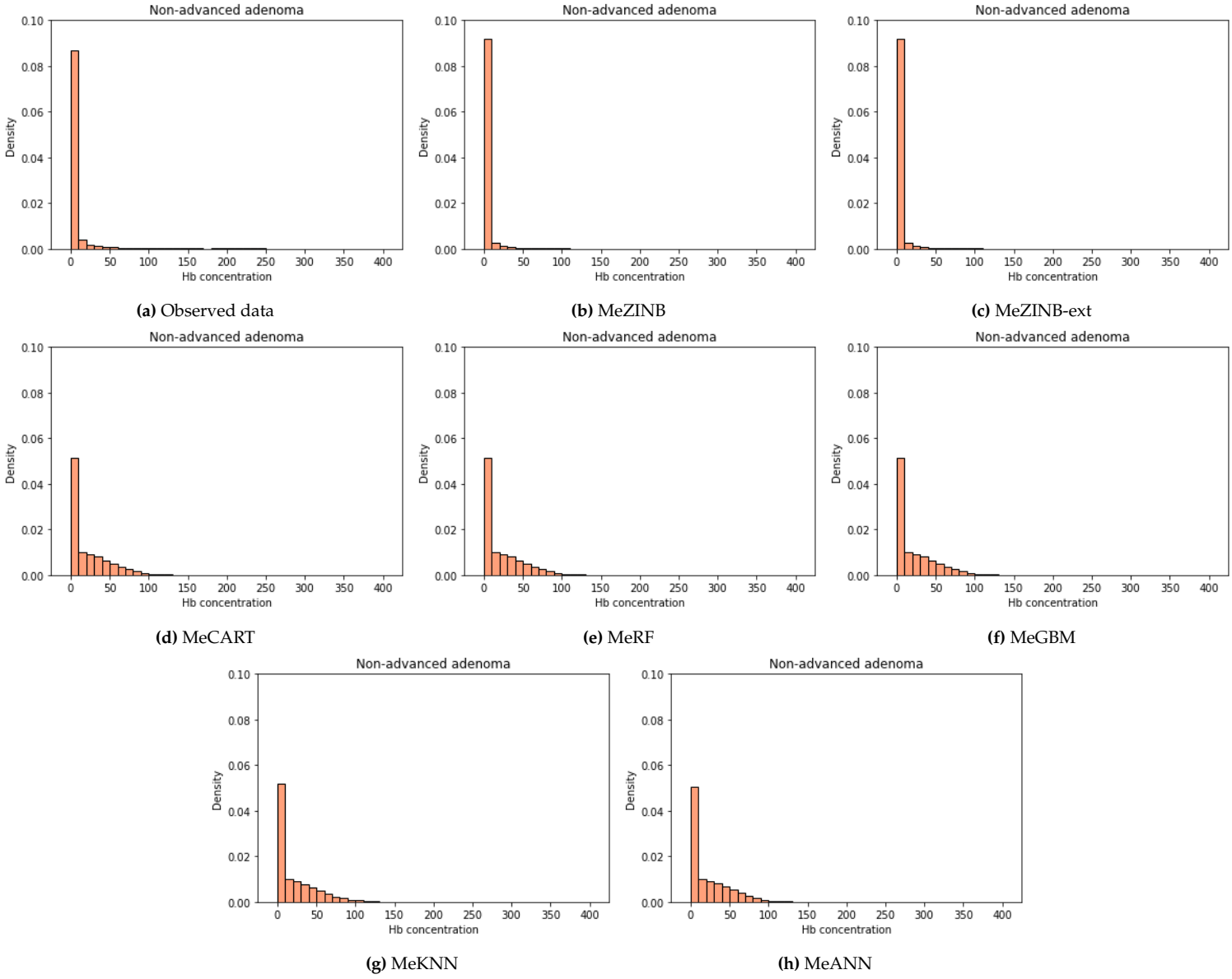


(a) Observed data **(b)** MeZINB **(c)** MeZINB-ext

(d) MeCART **(e)** MeRF **(f)** MeGBM

(g) MeKNN **(h)** MeANN

Figure 12: Distribution of the haemoglobin (Hb) concentrations obtained from the Dutch CRC screening programme and distributions of the Hb concentrations predicted by the different mixed-effect models. The Hb distributions are shown for the stage healthy.

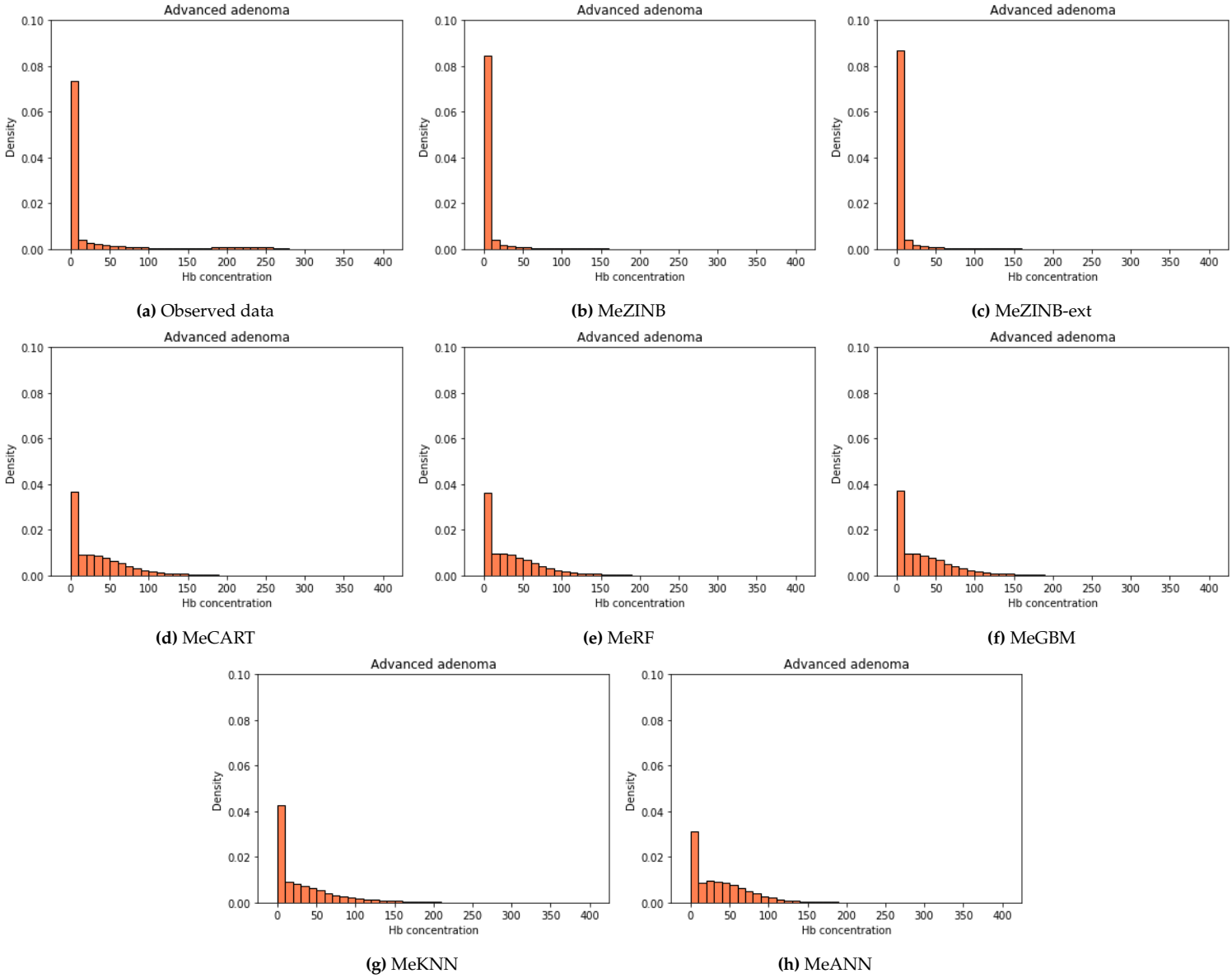


(a) Observed data **(b)** MeZINB **(c)** MeZINB-ext

(d) MeCART **(e)** MeRF **(f)** MeGBM

(g) MeKNN **(h)** MeANN

Figure 13: Distribution of the haemoglobin (Hb) concentrations obtained from the Dutch CRC screening programme and distributions of the Hb concentrations predicted by the different mixed-effect models. The Hb distributions are shown for the stage non-advanced adenoma.

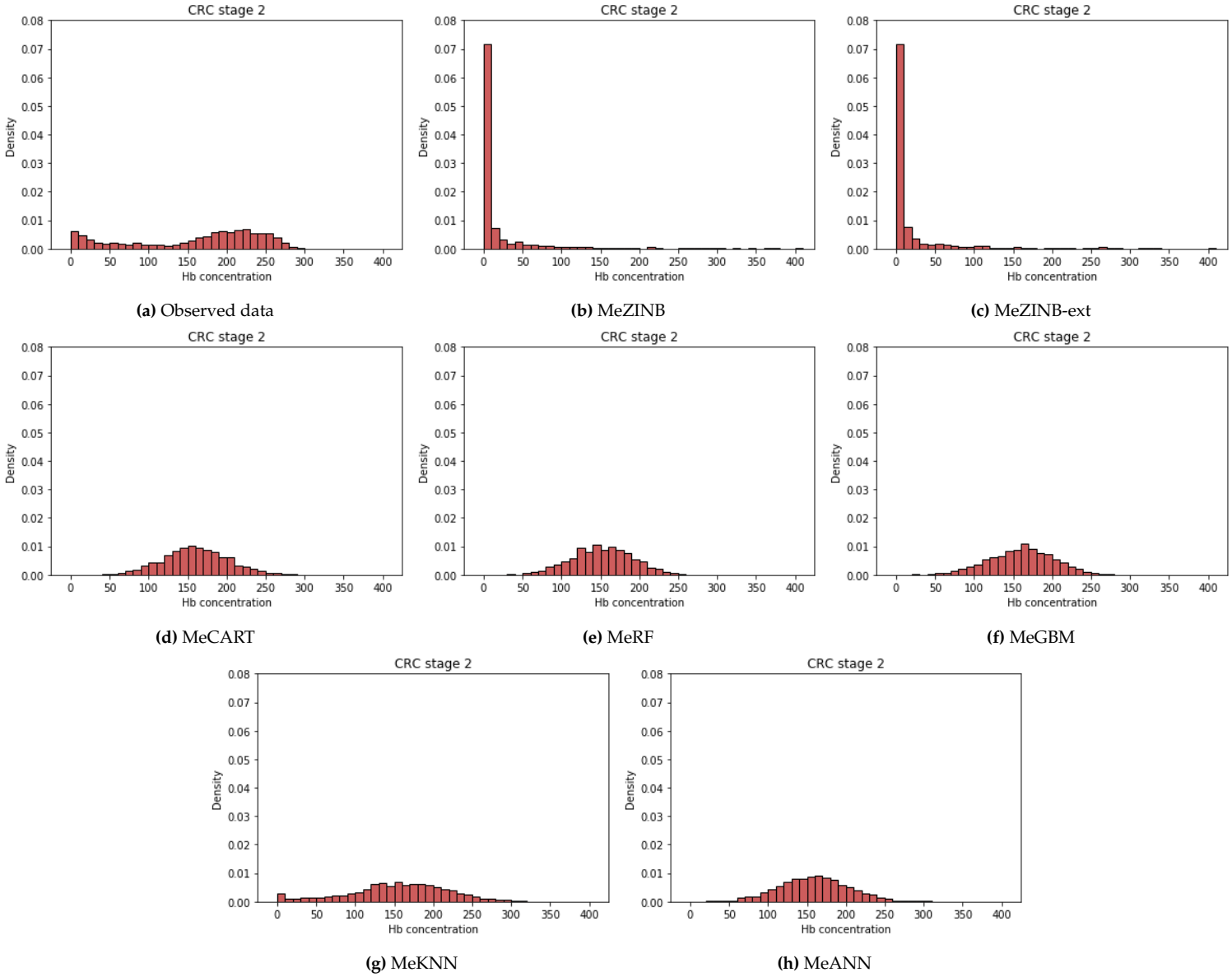


(a) Observed data **(b)** MeZINB **(c)** MeZINB-ext

(d) MeCART **(e)** MeRF **(f)** MeGBM

(g) MeKNN **(h)** MeANN

Figure 14: Distribution of the haemoglobin (Hb) concentrations obtained from the Dutch CRC screening programme and distributions of the Hb concentrations predicted by the different mixed-effect models. The Hb distributions are shown for the stage advanced adenoma.

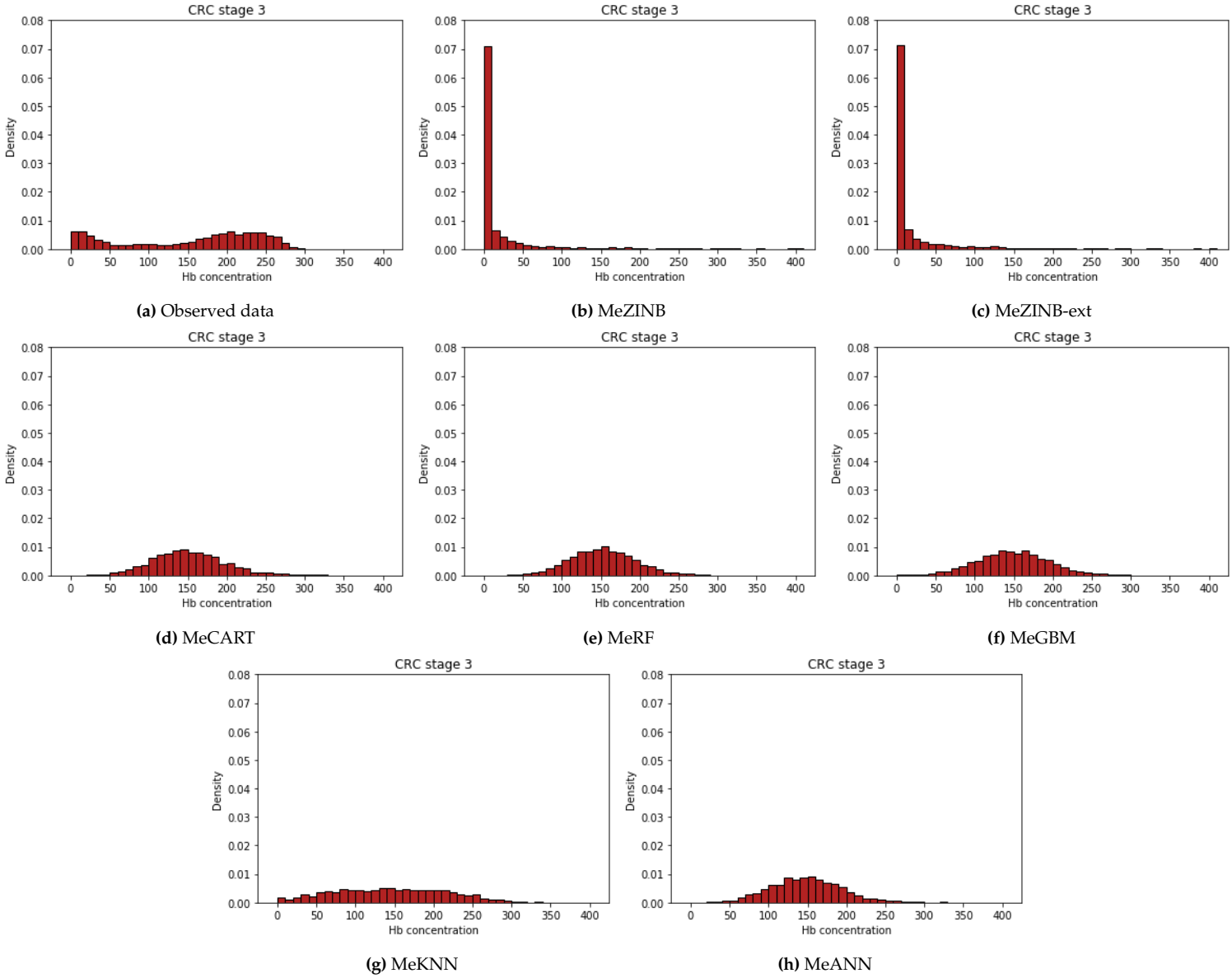


(a) Observed data **(b)** MeZINB **(c)** MeZINB-ext

(d) MeCART **(e)** MeRF **(f)** MeGBM

(g) MeKNN **(h)** MeANN

Figure 15: Distribution of the haemoglobin (Hb) concentrations obtained from the Dutch CRC screening programme and distributions of the Hb concentrations predicted by the different mixed-effect models. The Hb distributions are shown for the stage colorectal cancer stage II.

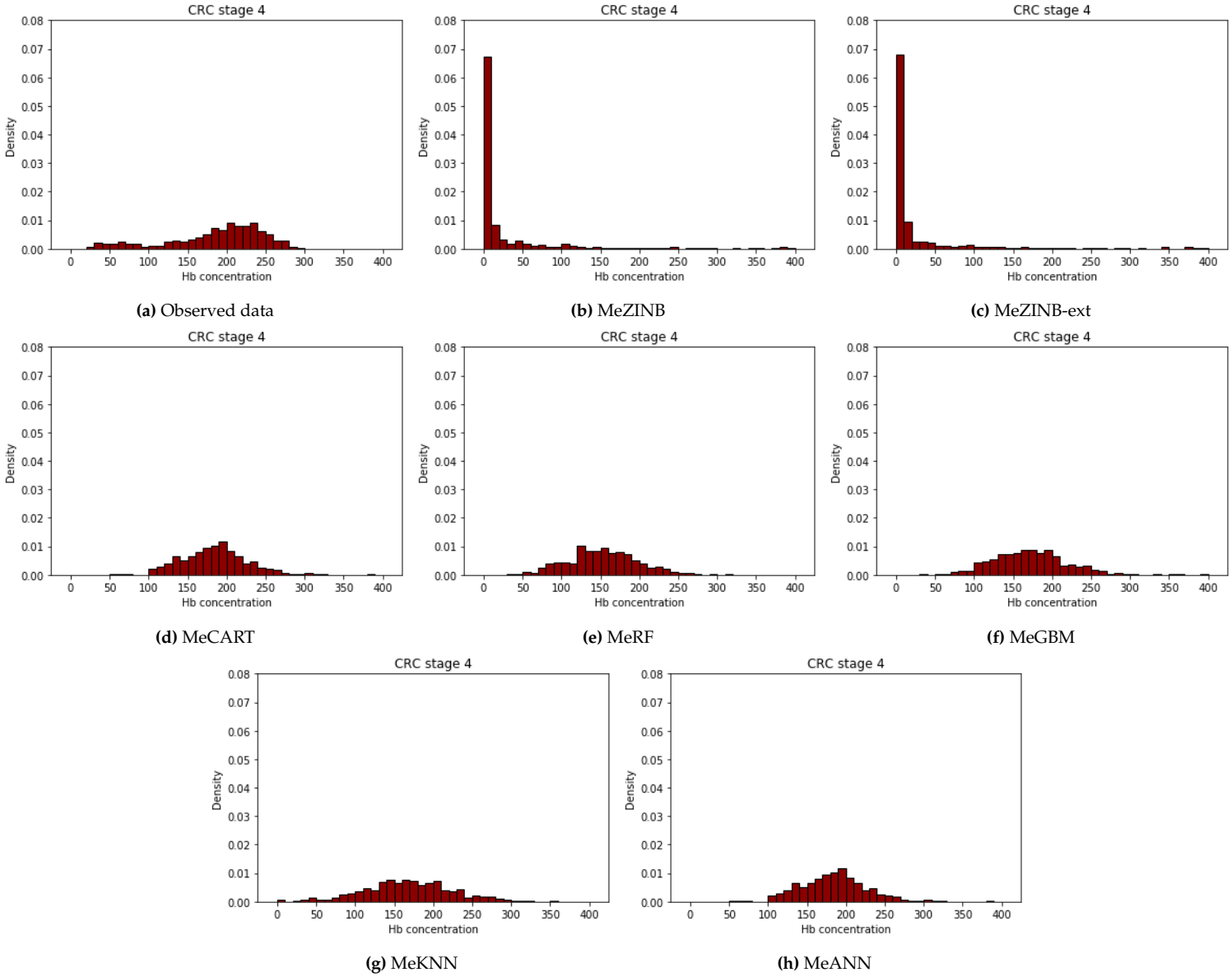


(a) Observed data **(b)** MeZINB **(c)** MeZINB-ext

(d) MeCART **(e)** MeRF **(f)** MeGBM

(g) MeKNN **(h)** MeANN

Figure 16: Distribution of the haemoglobin (Hb) concentrations obtained from the Dutch CRC screening programme and distributions of the Hb concentrations predicted by the different mixed-effect models. The Hb distributions are shown for the stage colorectal cancer stage III.



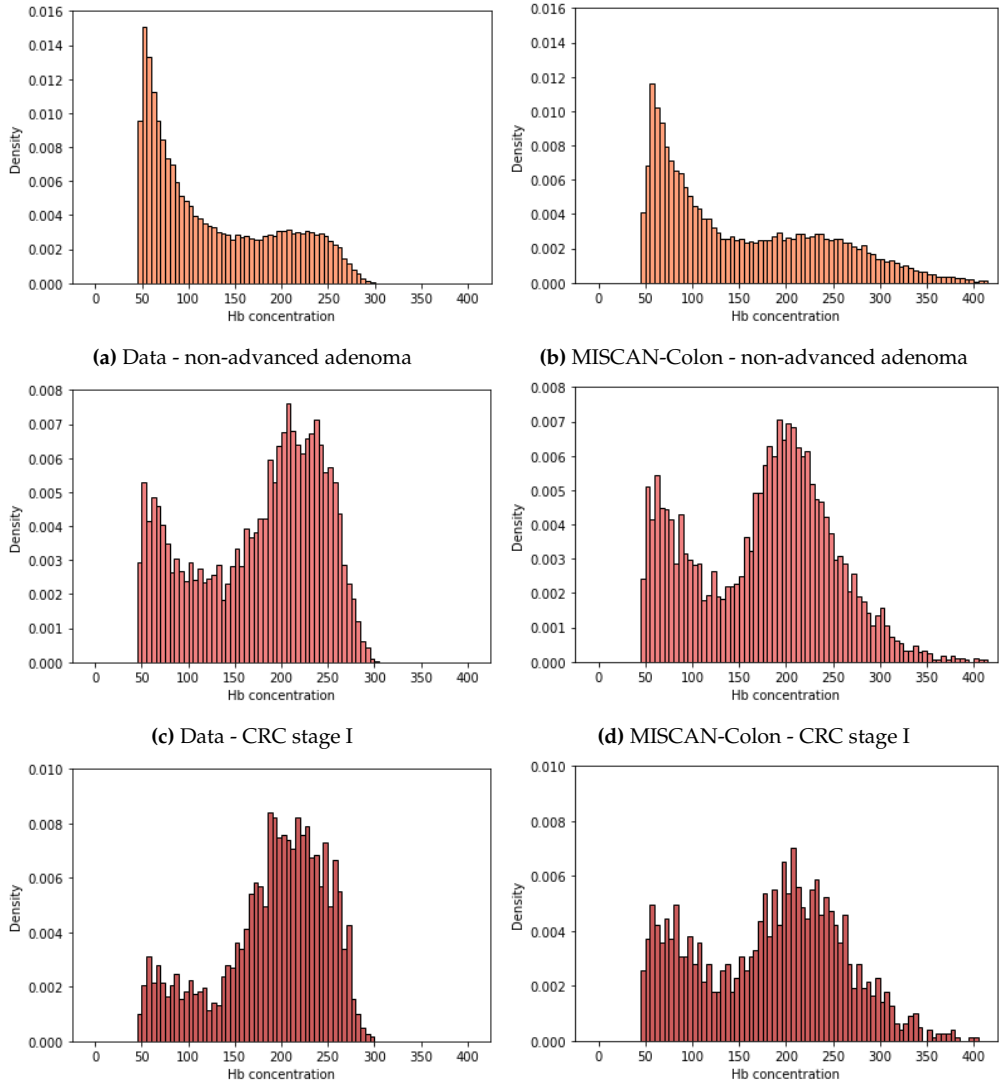
(a) Observed data **(b)** MeZINB **(c)** MeZINB-ext

(d) MeCART **(e)** MeRF **(f)** MeGBM

(g) MeKNN **(h)** MeANN

Figure 17: Distribution of the haemoglobin (Hb) concentrations obtained from the Dutch CRC screening programme and distributions of the Hb concentrations predicted by the different mixed-effect models. The Hb distributions are shown for the stage colorectal cancer stage IV.

D Distribution plots MISCAN-Colon



(a) Data - non-advanced adenoma

(b) MISCAN-Colon - non-advanced adenoma

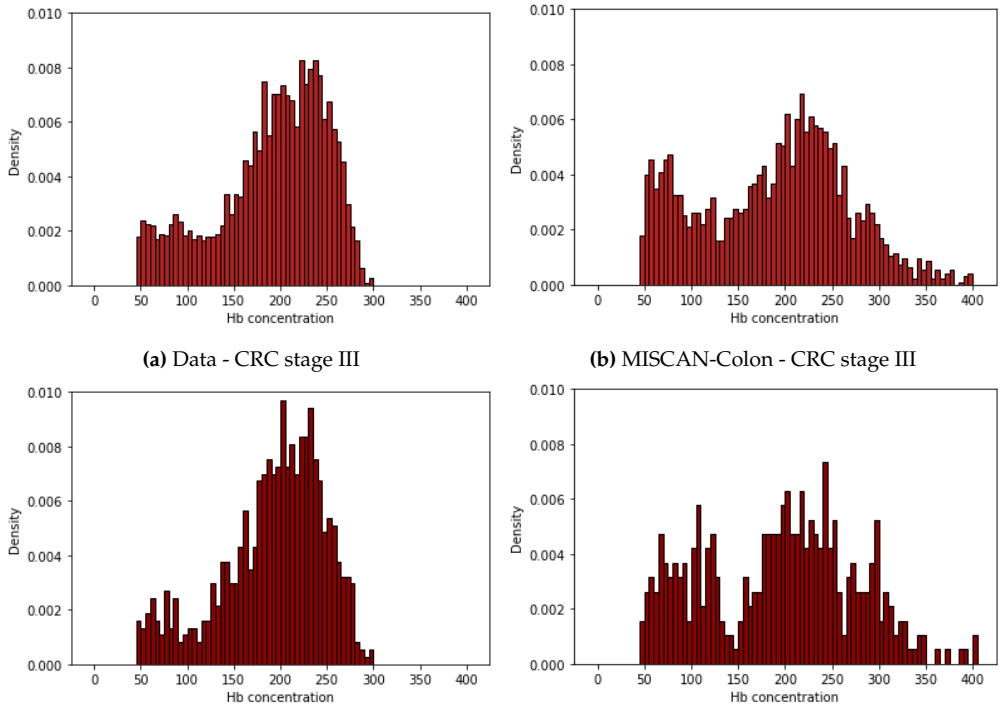
(c) Data - CRC stage I

(d) MISCAN-Colon - CRC stage I

(e) Data - CRC stage II

(f) MISCAN-Colon - CRC stage II

Figure 18: Distributions of the haemoglobin (Hb) concentrations obtained from the Dutch CRC screening programme and distributions of the Hb concentrations simulated by MISCAN-Colon for the stages non-advanced adenoma, CRC stage I and II. 1e6 individuals are simulated with MISCAN-Colon. The figures show only the Hb values above the cut-off of 47 $\mu g/g$.



(a) Data - CRC stage III

(b) MISCAN-Colon - CRC stage III

(c) Data - CRC stage IV

(d) MISCAN-Colon - CRC stage IV

Figure 19: Distributions of the haemoglobin (Hb) concentrations obtained from the Dutch CRC screening programme and distributions of the Hb concentrations simulated by MISCAN-Colon for the stages CRC stage III and IV. 1e6 individuals are simulated with MISCAN-Colon. The figures show only the Hb values above the cut-off of 47 μg/g.