ERASMUS UNIVERSITY ROTTERDAM

ERASMUS SCHOOL OF ECONOMICS

Simulating Haemoglobin Concentrations for MISCAN-Colon Using Black-box Machine Learning as a Step Towards Personalised Colorectal Cancer Screening[1]

Author

Yoëlle Kilsdonk (513530)

Supervisors

E. P. O'Neill (EUR)

dr. I. Lansdorp-Vogelaar (EMC)

R. van den Puttelaar (EMC)

D. van den Berg (EMC)

Second assessor

dr. O, Vicil (EUR)

August 4, 2022

[1]The views stated in this thesis are those of the author and not necessarily those of the supervisors, second assessor, Erasmus School of Economics, Erasmus University Rotterdam or Erasmus Medical Centre.

Abstract

Keywords— MISCAN, Machine Learning

Table of contents

List of Figures

List of Tables

1 Introduction

Colorectal cancer (CRC) is one of the leading causes of cancer-related deaths Western countries (Loeve et al., 1999; Sung et al., 2021; Torre et al., 2015), and it is expected that the absolute number of cases will increase as a result of aging and growth of populations. Fortunately, considerable research finds that a large proportion of CRC cases and deaths could be prevented by screening. But, how do we determine which screening policies are optimal? Clinical trials often only last a couple of years, while policy makers are most interested in the (cost-)effectiveness of screening strategies over a lifetime. For example, to answer the question 'can we reduce CRC mortality through changes in the current policy?', one would have to follow individuals throughout their whole lives. Also, in order to compare amongst screening policies, one would have to simultaneously implement and evaluate multiple policies using a real-life population. Since both of these scenarios are infeasible in practice, the Erasmus Medical Center (EMC) developed the MISCAN-Colon (MIcrosimulaten SCreening ANalysis) model – a microsimulation model for the evaluation of CRC screening.

The current implementation of the MISCAN-Colon model simulates a positive or negative faecal immunochemical test (FIT) result based on the sensitivity (true positive rate) and specificity (true negative rate) of the FIT. Recently, however, the Public Health department of EMC explored an extension of MISCAN-Colon to evaluate the benefits of personalised screening strategies, where instead of simulating FIT outcomes the model would simulate haemoglobin concentrations in a patient's stool (van Duuren et al., 2022). In this thesis, we employ black-box machine learning methods to realise this MISCAN extension.

Most machine learning methods rely on the assumption of independently identically distributed observations, which is likely to be violated in healthcare data due to correlations within individuals[2]. To overcome this issue, Ngufor et al. (2019) propose an approach which incorporates random-effects in machine learning algorithms for efficient analysis of longitudinal data. Based on this approach, van den Berg (2021) finds that mixed-effect machine learning (MEml) models significantly outperform the current proposed method to simulate haemoglobin concentrations. The optimal MEml model was chosen to be a decision tree due to its interpretability, as more complicated models attained similar performance. It is unclear, however, whether the increase in predictive accuracy found in van den Berg (2021) is specifically due to the inclusion of random-effects, or due to the use of machine learning methods in general. Therefore, this research investigates the contribution of the inclusion of random-effects to the predictive accuracies of black-box machine learning methods. We implement artificial neural networks (ANNs) and support vector regressions (SVRs) both with and without mixed-effects, using the approach of Ngufor et al. (2019). This leads to the following research questions:

RQ1a Does the introduction of random-effects in machine learning models lead to better performance, i.e., do MEml models outperform 'regular' machine learning models?

RQ1b Which model is best suited for predicting the haemoglobin concentration based on the data

[2]In this case, patients with negative FITs participate in multiple rounds, which allows for such correlation.

set provided by EMC?

RQ2 How well does the model from Q1b perform as simulation model in MISCAN-Colon?

The data for this research is provided by the EMC from the Dutch national CRC screening program from 2014-2020. For each of the 3.2 million individuals in the data set, a maximum of four screening rounds are available. We only include individuals who participate in two or more consecutive rounds, and those who participate in only one round in total. One of the variables in this data set is the current stage of CRC in an individual, which is imputed using the multiple imputation via chained equations approach by Van Buuren and Oudshoorn (1999).

This research consists of two phases, the first being outside of MISCAN-Colon, where we predict haemoglobin concentrations using four different models. These models are trained, validated, and tested using the longitudinal data set provided by the EMC. Based on phase one, we answer RQ1a and RQ1b. In phase two, we implement the most promising model in MISCAN-Colon, and calibrate this model such that the simulated haemoglobin concentrations resemble the observed concentrations of real-life Dutch population screening data as closely as possible. We then answer RQ2.

The remainder of this research is structured as follows. We provide background information on colorectal cancer MISCAN-colon in Section 2, along with an overview of ANNs and SVRs in healthcare literature. In Section 3 we describe the data and the data imputation method. We present our methodology in Section 4, followed by our results and conclusion in Sections 5 and 6 respectively.

2 Literature

2.1 Colorectal cancer

Colorectal cancer (CRC) is the development of cancer from the colon or rectum, which usually starts as a benign adenoma (i.e., a noncancerous tumor), and is one of the most commonly diagnosed and most deadly cancers worldwide (Torre et al., 2015; Sung et al., 2021). Specifically, according to the Dutch Rijksinstituut voor Volkgezondheid en Milieu five percent of people will develop CRC in the Netherlands. Nearly nine in ten cases of that 5% occur in people older than 55. Risk factors for CRC include age, gender, genetics, environment, diet, physical activity, and smoking (Botteri et al., 2008; Thanikachalam and Khan, 2019). Moreover, the worldwide burden of CRC is expected to further increase due to, *inter alia*, the rapid growth and aging of the population (Jiang et al., 2022; Winawer, 2007), which is a testament to the importance of optimising screening procedures.

Only a small percentage of adenomas become cancerous (Strum, 2016). Therefore, we distinguish between progressive and non-progressive adenomas, where (non-)progressive adenomas do (not) develop into CRC, as shown in Figure 1. The most common form of CRCs are colorectal adenocarcinomas, with a prevalence of over 95% (Thrumurthy et al., 2016).

We also distinguish between clinical and preclinical stages, where preclinical indicates that the cancer is not yet diagnosed. Preclinical cancer can then progress from stage I to stage IV, where symptoms may

develop in each stage, which in turn may lead to disease diagnosis (Compton and Greene, 2004). Once the cancer has been diagnosed as a result of symptoms, it is referred to as clinical.

Figure 1 shows the progression of CRC in five stages. In stage 0, the adenoma has not grown beyond the mucosa (i.e., the inner lining) of the colon or rectum. Stage I is when the cancerous adenocarcinoma has grown beyond the mucosa without spreading to the lymphatic system or distant organs. In stage II the adenomacarcinoma has invaded the colonic or rectal wall, with possible infection of nearby organs. Finally, in stages III and IV, the metastatic adenocarcinoma has spread to lymph nodes and distant organs.

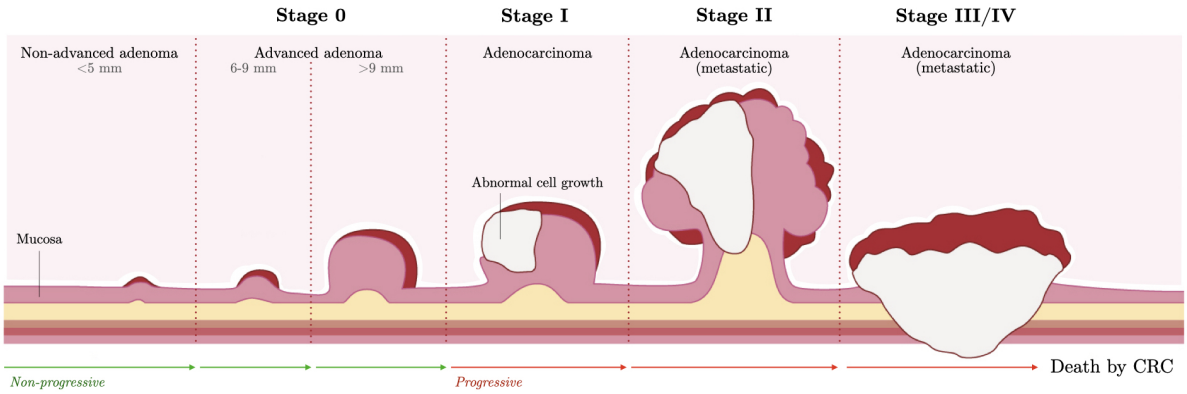


Figure 1: Progression of colorectal cancer in stages

2.1.1 Screening

The effect of screening is twofold. First, research indicates that over 90-95% of CRCs develop from (benign) adenomas (Bronner and Haggitt, 1993; Morson, 1974). Hence, early detection and removal might aid in CRC prevention (Loeve et al., 1999). Second, early detection of an (a)symptomatic cancer may result in an improvement in prognosis. Specifically, a large body of literature finds that screening results in a reduction in mortality, as cancers can be detected at an early and curable stage (Jiang et al., 2022; Levin et al., 2008; Toribara and Sleisenger, 1995; Whitlock et al., 2012).

Screening tests can be subdivided into two categories: stool-based tests and visual exams. The guaiac-based fecal occult blood test (gFOBT) and fecal immunochemical test (FIT), e.g., belong to the first category, in which the stool is tested for haemoglobin. If these tests report a high haemoglobin concentration, this could be an indicator for the presence of CRC. The two most common visual exams are (flexible) sigmoidoscopy, and colonoscopy, which investigate the structure of the colon and rectum for abnormal tissue. According to the review by Ding et al. (2022), colonoscopies are most effective in reducing CRC-related deaths at an approximate 68% decrease (Brenner et al., 2014). As for the stool-based tests, the FIT reduces CRC-related deaths by 22% on average, which is approximately 7% more effective than the gFOBT test (Hewitson et al., 2008; Zorzi et al., 2015). The FIT also has a higher participation rate and positivity rate compared to gFOBT in CRC screening programs, while reporting fewer false negatives (Mousavinezhad et al., 2016). Moreover, the FIT is relatively close in effectiveness compared to flexible sigmoidoscopies while being considerably less invasive, with reported

mortality reduction of approximately 28%, (Holme et al., 2013). When screening with a test (other than a colonoscopy) leads to abnormal test results, the general advice is to proceed with a follow-up colonoscopy (Ding et al., 2022).

In the Netherlands, each person between the age of 55-75 is asked to participate in the population screening for CRC once every two years since January of 2014[3]. The participants receive a FIT, which is sent back to the laboratory after taking a stool sample. In the event of an aberrant result, health care professionals can make a referral for a colonoscopy and treatment. If any abnormalities are present during the colonoscopy, small amounts of tissue can be removed for analysis (i.e., a biopsy) and abnormal growths or adenomas can be identified and removed. This way, CRC can be detected at an early stage. Figure 2 shows that, according to the Integraal Kankercentrum Nederland, patients diagnosed with CRC through the Dutch population screening had a more favorable stage distribution than patients without screening. Also, patients who were diagnosed through population screening were more likely to receive less invasive treatments.

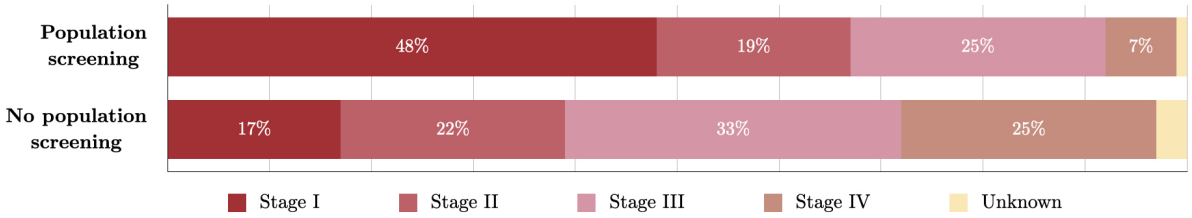


Figure 2: Distribution of diagnosed cancers in patients with, and without screening (source: https://iknl.nl/)

Unfortunately, screening is not a silver bullet in healthcare, as it could lead to, e.g., overdiagnosis or false positives, while also being costly and invasive. Welch and Black (2010) provide a summary of current evidence that early detection leads to overdiagnosis in breast, lung, and prostate cancer. Overdiagnosis – defined as the diagnosis of a medical condition or disease that would not cause symptoms or death during a patient's lifetime – is associated with long-term psychosocial harm, lower quality of life, and unwanted/unnecessary usage of (follow-up) tests, treatment, and healthcare facilities (Barton et al., 2001; Brodersen and Siersma, 2013; Jenniskens et al., 2017; van der Steeg et al., 2011). On the other hand, Brasso et al. (2010) and Wardle et al. (2003) find no adverse psychological effects due to cancer screening, although they do not specifically investigate the effects of overdiagnosis. That said, overdiagnosis could be particularly harmful if it leads to unnecessary treatments, each of which comes with their specific risk[4].

Given the previously stated disadvantages to screening, it is clear that policy makers must continually evaluate the trade-off between harms and benefits to attain the most efficient screening policies. A large body of literature indicates that *personalized* screening may aid in achieving such optimized policies, e.g., for diseases such as colorectal-, prostate-, and breast cancer (Frampton et al., 2016; Pashayan et al., 2011; Schröder et al., 2009). Moreover, Grobbee et al. (2017) suggest that FIT-based programs can be

[3]For more information see: https://www.rivm.nl/darmkanker.
[4]For an assessment of operative risk in CRC surgery, we refer to Fazio et al. (2004) and Hanley (2005).

improved upon by using a screening policy with person-specific intervals and -thresholds depending on previous haemoglobin concentrations in a person's stool. However, since personalised screening necessitates policymakers and health care providers to make decisions on, i.a., what tools to use to identify risk levels and at which risk levels screening or prevention programs are warranted, the possibilities of feasible personalized screening policies are limitless. Recently, van Duuren et al. (2022) addressed this problem using the adapted version of Habbema et al.'s (1985) MISCAN (MIcrosimulaten SCreening ANalysis) microsimulation model, called MISCAN-Colon, where one can overlay screening scenarios on a simulated population *before* real-life implementation. The current implementation of the MISCAN-Colon model only simulates a positive or negative FIT result based on the sensitivity and specificity, but was extended by van Duuren et al. (2022) with a prototype module which returns haemoglobin concentrations in a person's stool.

2.2 Methods

This paper extends (part of) the research of van Duuren et al. (2022), using black-box machine learning methods instead of their linear mixed-effects model. However, as mentioned previously, oftentimes healthcare data is longitudinal, with (possibly) repeated measurements over different intervals of time, which could cause correlations within patients. Unfortunately, most machine learning methods are not robust to such correlations.

One possible solution to this problem could be to employ 'regular' machine learning models while explicitly modeling the interpatient correlation through inclusion of time-specific variables (e.g., current number of test, previous haemoglobin concentrations, maximum haemoglobin concentrations). However, the nature of this data suggests that better estimation may be possible if the information of the repeated measurements would be included at the level of the algorithm itself. In this section, we provide an overview of the literature on machine learning – specifically artificial neural networks (ANNs), support vector machines (SVMs) and eXtreme Gradient Boosting (XGBoost) – in longitudinal health data.

2.2.1 Artificial neural networks

The trajectory of cancer is clearly non-linear, highly variable and dependent on a large variety of factors, most of which are not understood to this day. The flexibility of ANNs can be used to effectively address these problems. Another important property of ANN, with respect to our application, is their suitability for prediction of non-negative variables (Haghani et al., 2017; Sakthivel and Rajitha, 2017). Moreover, Haghani et al. (2017) shows that ANNs outperformed Poisson regression, negative binomial regression, zero-inflated Poisson regression, and zero-inflated negative binomial regression in their research to predicting the number of return to blood donations using zero-inflated data.

However, ANNs, just as SVMs, make the implicit assumption of independently identically distributed data, which is often violated in longitudinal data. While certain ANNs have been successfully adjusted to account for temporal trends (e.g., Choi et al.'s recurrent neural networks), longitudinal data often also contains unequal time intervals between measurements, and an unequal number of observations per

individual. To account for these specific data characteristics, Xiong et al. (2019) propose a new type of ANN called the mixed effects neural network model, which adapts mixed effects within a deep neural network architecture for gaze estimation, based on eye images. This model is person-specific, and uses few calibration samples to eliminate the person-specific bias in longitudinal data. In the field of Alzheimers, Tandon et al. (2006) introduce another mixed effects neural network to accurately model the nonlinear course of the disease. Their model generalizes a linear mixed effects model by incorporating a general non-linear function of the input variables. This model is shown to be much more accurate and effective compared to standard ANNs and linear mixed effects models. Lastly, Mandel et al. (2021) propose a generalized neural network mixed effects model, which is structured as a generalized linear mixed model (GLMM), where the linear fixed effect is replaced by a feed-forward ANN and a random effect component is added. They use this approach to predict depression and anxiety levels of schizophrenic patients using longitudinal data.

2.2.2 Support vector machines

In an attempt to merge longitudinal data with machine learning, Luts et al. (2012) propose a mixed-effects least squares support vector machine (LS-SVM) classifier using regression modeling and a prediction step. The research by Cheng et al. (2014) provides analytical expressions of confidence and prediction intervals of mixed-effects LS-SVM approaches, such as this one. An alternative approach to modeling longitudinal data using SVM is proposed by Chen and DuBois Bowman (2011). They construct a support vector classifier based on linear combinations of features from different cross-sectional time-points to make predictions, using an expectation-maximization algorithm.

Another branch of literature focuses on the generalisation of SVM to SVR. For example, the longitudinal SVM classifier by Chen and DuBois Bowman (2011) is extended by Du et al. (2015) to perform regression. Another SVR model suitable for longitudinal data is the semiparametric mixed-effects least squares support vector regression (LS-SVR) model by Seok et al. (2011). This model shows slightly improved performance and prediction over 'standard' LS-SVR using pharmacokinetic and pharmacodynamic data. Finally, Cho (2010) propose a mixed-effects LS-SVR where a random-effect term is added to the optimization function of LS-SVR to include random effects in the model. While all aforementioned methods make use of a different premise, each of them inherit the benefits of the non-linear kernel property of the SVM.

2.2.3 XGBoost

In short, XGBoost is an ensemble method developed by Chen and Guestrin (2016), which creates boosted trees using sequentially built shallow decision trees.

XXX – Insert literature on longitudinal data using XGBoost

In this study, we follow the analytic framework by Ngufor et al. (2019), which integrates the random-effects structure of GLMM in non-linear machine learning models, compatible with longitudinal data.

While their paper only shows interpretable tree based mixed-effect machine learning models, the framework can easily be extended to other (common) machine learning models. A more extensive explanation of their method is presented in Section 4.2.

2.3 MISCAN-Colon

As previously mentioned, MISCAN-Colon allows for the evaluation of different screening policies by comparing their costs and effectiveness, as well as assessing the risk of false positives and overdiagnosis on a simulated population (Loeve et al., 1999).

The model simulates individual life histories in which several colorectal lesions can emerge, and produces incidence and mortality rates in the simulated population using information on the epidemiology and natural history of the disease as input combined with screening- and demography characteristics. By comparing the simulated life histories with, and without screening, MISCAN-Colon can evaluate the costs and benefits of a specific screening strategy.

The MISCAN-Colon model can be decomposed into three parts: demography, natural history and screening. Figure 3 shows an exemplified version of these three parts, using a fictive individual named Robin. The upper line, referred to as the demography part, simulates the life of Robin without cancer who dies at 87 years old of other causes than CRC. The middle line simulates Robin's life *with* cancer, but without screening, which adds a natural history to the demography part. In this scenario, Robin dies at 72 due to CRC. The bottom line simulates Robin's life when screening is overlayed, with 15 life years gained as a result.

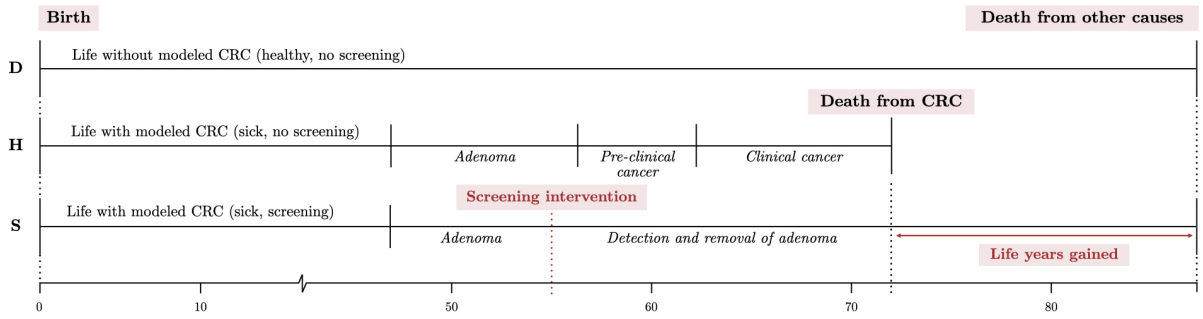


Figure 3: Simulations from the MISCAN-Colon model, where the upper bar shows the demography part (D), the middle bar adds the natural history (H) to D, and the lower bar adds both screening (S) and H to D

Three remarks on Figure 3. First, the survival of a lesion after diagnosis depends on the stage of the cancer (and other risk factors). Thus, screening does not ensure that an individual survives from CRC. The possible prognoses after a positive test result for CRC screening are: delay in moment of death, no change in moment of death, or premature death by complications of treatment. Second, the figure only shows examples of individuals with *one* lesion for simplicity, but the MISCAN-Colon model also allows for the modelling of zero or multiple lesions. New lesions that appear after clinical diagnosis of CRC are accounted for in the simulated survival. Third, Figure 3 only shows lethal progressive adenomas, but it

is also possible that an individual develops non-lethal adenomas which would never result in death of an individual.

3 Data

The data for this research is obtained from the Dutch CRC screening program in 2014-2020. For each individual who participated in the biennial screening a maximum of four rounds of data are available. This analysis exclusively focuses on those who participated in one round only, or multiple consecutive rounds.

Given that this research explores, i.a., 'regular' machine learning models while within individual correlation might be present, we introduce additional variables to allow for as much individual variation as possible. First, we include a lagged dependent variable (previous haemoglobin concentrations), as Grobbee et al. (2017) find that an undetectable haemoglobin concentration two years ago decreases the current risk of having CRC. We also include the minimum and maximum haemoglobin value per individual over all FITs prior to the current time of screening.

Table 1: Original variables in the data set provided by the Erasmus Medical Centre

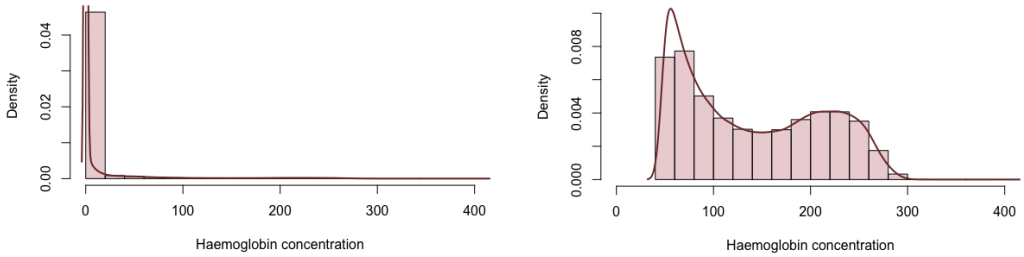
Variable	**Description**	**Range**
`Age`	Age of respondent at time of screening	55 – 77
`Birth year`	Year of birth	1938 – 1963
`FIT number`[1]	Indicator for sequence number of the FIT	1 – 4
`Haemoglobin current`	Haemoglobin value found in FIT in current round	0 – 437.1
`Haemoglobin max`	Maximum obtained haemoglobin value over all tests at time of screening	0 – 47.0
`Haemoglobin min`	Minimum obtained haemoglobin value over all tests at time of screening	0 – 47.0
`Haemoglobin previous`	Haemoglobin value of previous round	0 – 47.0
`Haemoglobin threshold`	Threshold value used to determine the unit of the bloodtest result	275, 47
`ID`	Personal identification number	1 – 3,710,672
`Result`	Indicator for result of screening bloodtest	0 (Favourable, 96.1%), 1 (Unfavourable, 3.9%)
`Round`	Indicator for current round	1 – 4
`Sex`	Gender of respondent	0 (Male, 48.0%), 1 (Female, 52.0%)
`Stage`[2]	Stage of cancer at time of screening	1 (Healthy, 0.8%), 2 (Non-advanced adenoma, 1.1%), 3 (Advanced adenoma, 1.7%), 4 (Colorectal cancer, 0.3%), NA (Unknown, 96.1%)

Notes: [1]`Fit number` is one-hot encoded, such that the resulting dummy variables are equal to one for the current FIT, and zero otherwise.[2]`Stage` is only available for individuals where cancer has been detected, and is unknown otherwise.

After data pre-processing (described in Appendix A), the data set contains 6,796,731 observations for 3,170,234 individuals of which almost 52% are female. In total, 803,651 individuals participated in one round only, 1,108,079 individuals participated in two consecutive rounds, 1,118,003 individuals participated in three consecutive rounds, and 140,501 individuals participated in all four rounds. Table

1 shows an overview of all variables included in the data set, along with descriptive statistics. Figure 4a shows the distribution of haemoglobin concentration in the data set over all observations, and Figure 4b shows the distribution of haemoglobin concentrations amongst observations with positive FITs. Clearly, the dependent variable is heavily zero-inflated. We can also distinguish a bimodal distribution in the positive FITs, with the largest peak between $[47; 80]$ and a second peak around $[180; 260]$.

Figure 4: Haemoglobin concentration densities and histogram



(a) Complete data set (in μg/g) (b) Exclusively positive FITs (in μg/g)

3.1 Missing values

The `stage` variable is unknown for individuals with negative (favourable) FIT outcomes in phase one, as these individuals do not undergo follow-up procedures. In our data set 96.1% of observations report favourable FIT outcomes, such that `stage` is only known 3.9% of the time.

Most statistical procedures are designed for complete data, and ANNs and SVRs are no exception to this rule. SVRs are less sensitive to missing data than ANNs, as this method only relies on a subset of observations: the support vectors. However, missing data can be problematic in a non-linear setting such as ours, as observations have a more local influence on the marginal with non-linear kernels (Stewart et al., 2018). To our knowledge, there are no adaptations to SVRs which allow for missing values to this degree, where the missingness is at random. There are adaptations to ANNs to account for missing values, e.g., the combination of deep networks with probabilistic mixture modes by Śmieja et al. (2018). This method is based on the premise that instead of calculating the activation function on a single data point, the first hidden layer in the network computes the *expected* activation of neurons. However, given that such methods are not available for SVRs, using this adapted ANN would invalidate the comparison in performance between the ANNs and SVRs, as both methods would be based on different input data. Thus, we can either delete or impute `stage`.

If we only delete observations without reported stages, the resulting data set exclusively contains individuals above the cut-off value of 47 μg/g, which is not only unrepresentative for the Dutch population, but will also likely result in poor predictive performance when the models are used in MISCAN-Colon, where individuals below the cut-off do occur. A similar problem of decreased performance might also prevail when deleting the variable in its entirety, as previous internal research by EMC shows that `stage` is a strong predictor of haemoglobin concentrations. Additionally, as the purpose of screening is to identify the current stage of cancer in an individual, we opt to impute `stage`.

We include two additional data sets – the '15 threshold' and 'MISCAN simulation' data set – in an

attempt to increase the accuracy of the imputations. The '15 threshold' data set, provided by the EMC from the Dutch national CRC screening program, contains a total of 16,591 individuals who participated in the first round of 2014[5] with known `stage`. The threshold for whether one should be admitted to the surveillance program was set to 15 micrograms of blood per gram of feces instead of 47 μg/g for a subset of these individuals. Thus, this data set contains real-life data on the current stage of individuals with `current haemoglobin` below 47 μg/g, in contrast to the original data set, which only reports `stage` for observations over 47 μg/g.

Given that the '15 threshold' data set is relatively small in size compared to the original data set, we also perform a population simulation run in MISCAN-Colon. Specifically, we simulate two million individuals from 2014-2020, with the same sex ratio as the original data set. This 'MISCAN simulation' data set consists of 3,076,778 observations, where the current `stage` is always known and `haemoglobin current` is always unknown. Table 2 in Appendix A.1 reports descriptive statistics for both additional data sets. The combination of all three data sets results in 9,890,100 observations in total, of which 6,533,768 `stage` observations and 3,076,778 `haemoglobin current` observations are missing.

There are two major iterative approaches for multiple imputation in general missing data patterns: joint modeling and the fully conditional specification. Joint modeling assumes joint multivariate normality of all variables, which is inapt for imputing categorical variables, and therefore unsuitable for this analysis. In contrast, the fully conditional specification does not rely on multivariate normality, and applies a multivariate imputation model variable by variable using a collection of conditional densities per incomplete variable (Van Buuren, 2018).

A popular data imputation method amongst the fully conditional specification is Multiple Imputation via Chained Equations (MICE), which is an often used and recommended method in healthcare literature (Ambler et al., 2007; Baneshi and Talei, 2011; Chowdhury et al., 2017; Faris et al., 2002; Jolani et al., 2015). We employ MICE to impute `stage`, using `haemoglobin current`, `result`, `age`, and `sex`. To this end, we assume that the missing observations are missing at random, which means that there might be systematic differences between the missing and observed stages, but these can be entirely explained by other observed variables (Bhaskaran and Smeeth, 2014). In this case, the missingness of `stage` is a direct result of the test outcome of the FIT.

In each iteration of MICE we first impute `haemoglobin current`, and then impute the corresponding `stage`. Specifically, in step one, we replace all missing values in the data set with a random draw from the data as temporary place holder. In step two, we set the place holder back to missing only for the variable we wish to impute. In step three, we replace these missing values using an appropriate imputation method (e.g., sampling, predictive mean matching, linear regression or logistic regression) using (part of) the remaining variables in the data set. Steps two and three are then repeated until all missing variables are filled, at which point we completed one full cycle. We perform ten cycles in total, as per recommendation of Raghunathan et al. (2002). The observed data combined with the imputed values at the end of the tenth cycle constitute one imputed data set. This process is repeated to create 5 imputed data sets, such that a total of 5×10 iterations are performed. The final distribution of all five

[5]The threshold was set to 47 μg/g for all individuals from round two in 2014 onward.

imputed versions of `stage` are then compared to the `stage` distribution in the 'MISCAN simulation' data set. The imputed variable which most closely compares to the MISCAN `stage` variable is then used as replacement for the *stage* variable in the original data set. As a final step, all observations from the '15 threshold' and 'MISCAN simulation' are removed. Appendix A.1 provides a more detailed explanation of the MICE algorithm specific to this paper.

4 Methodology

4.1 Machine learning

4.1.1 Artificial neural networks

ANNs, developed by Lippmann (1987), are inspired by the human brain, mimicking the way that biological neurons signal to one another. ANNs are comprised of (1) an input layer, (2) possibly one or more hidden layers, and (3) an output layer. The input variables are related to the output variable(s) through a network of interconnected nodes, with associated weight and threshold. If the output of any individual node is above the specified threshold value, that node is activated, sending data to the next layer of the network. The optimal values for these weights are estimated when the ANN is fitted, such that a predetermined loss function is minimized – the tweedie loss function in our case. The input layer of the ANN consists of p nodes, where p is equal to the number of explanatory variables. In our setting, the output node $\hat{f}(x)$ represents the predicted haemoglobin concentration.

To advance from one layer to another, the ANN uses activation functions $h(\cdot)$, with the sum of the weights and the intercept, referred to as the bias, as input. We compare the identity activation function $h(x) = x$, and the rectified linear unit (ReLu) activation function $(x) = \max(0, x)$[6].

Hornik et al. (1989) show in their universal approximation theorem that an ANN with at least one hidden layer, and a large enough number of neurons, can approximate any finite-dimensional Borel measurable function up to any arbitrary accuracy. In other words, an ANN with zero hidden layers can only represent linear functions, whereas we can approximate *any* function with a continuous mapping with finite spaces using an ANN with one hidden layer. In practice, however, a network with multiple hidden layers can be more efficient. Therefore, I consider ANNs with both one, and two hidden layers. In case of an ANN with two hidden layers, with H nodes in the first layer and L nodes in the second,

[6]This comparison is made on a subset of the data. Eventually, only one activation function is used. A necessary condition is that the output is strictly non negative.

the values at each node are calculated as follows:

$$z_h^1 = g\left(\sum_{j=1}^{p} w_{hj}^1 x_j\right) \qquad \forall h \in \{1,\ldots,H\},$$
$$z_l^2 = g\left(\sum_{h=1}^{H} w_{lh}^2 z_h^1\right) \qquad \forall l \in \{1,\ldots,L\},$$
$$\hat{f}(x) = g\left(\sum_{l=1}^{L} w_l^3 z_l^2\right),$$

where x_j represents each of the input regressors, z_i^j represents the i^{th} node of the j^{th} hidden layer, and w_{ik}^j is the weight of node k on node i in hidden layer j. Figure 5 shows an example of an ANN with two hidden layers. We use 8-fold[7] cross-validation to determine the number of layers, and nodes in each layer.

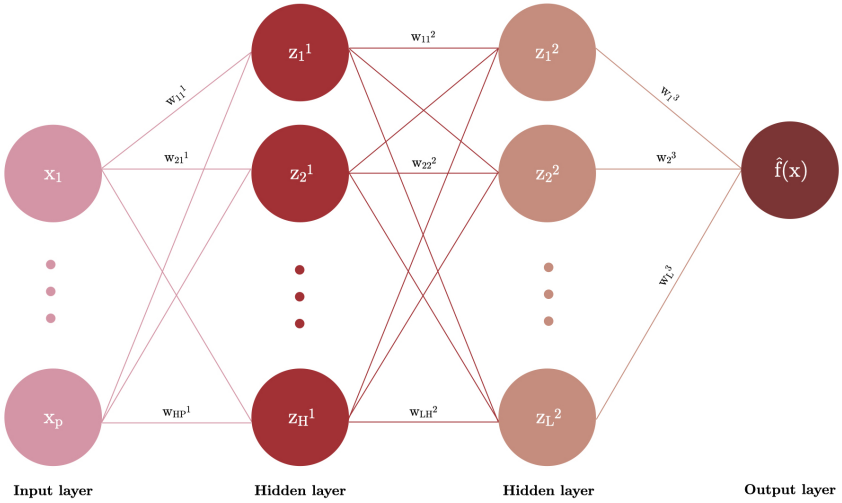


Figure 5: Example of an artificial neural network with two hidden layers and one output node

One of the risks of ANNs is that they tend to overfit on the training data. To mitigate overfitting, we use the efficient early stopping regularization (Prechelt, 1998). We also explore other regularization terms (Lasso or Ridge) and dropout (Srivastava et al., 2014) as options to minimize overfitting in each ANN.

4.1.2 Support vector regression

The second algorithm is based on SVMs, which separate binary classified data using a hyperplane as decision boundary such that the margin between the classes is maximised (Cortes and Vapnik, 1995). To this end, the input variables x_j are transformed into an m-dimensional feature space using a non-linear mapping, after which the SVM algorithm searches for the optimal separating hyperplane represented by a set of support vectors – the data points on either side of the hyperplane that are closest to the hyperplane.

[7] Eight folds are chosen to efficiently parallelize across four CPU's.

The use of a kernel implicitly maps the input vector to higher dimensional feature spaces, where the problem becomes a linear surface that fits the data, which allows for SVMs to handle highly non-linear data. SVMs model non-perfectly separable data through the introduction of soft margins, where some slack is allowed for observations to be on the wrong side of the margin.

The sparse solution and relatively easy generalization of SVMs lend themselves to adaptation to regression problems (Awad and Khanna, 2015). SVRs use the same principle as the SVMs, but predicts discrete values. The basic idea behind SVR is to find the best line within a threshold value ε. To this end, we introduce the ε-tube – also referred to as a ε-insensitive region around the function – which reformulates the optimization problem to find the tube that best approximates the continuous-valued function, while balancing model complexity and prediction error. More specifically, SVR looks for the flattest tube that contains most of the training instances.

The fit time complexity of SVR is more than quadratic with the number of samples, thus for large data sets, Linear SVR is preferred – which provides a faster implementation than SVR, as it only considers the linear kernel. Consequently, this is the employed method for this research.

4.1.3 XGBoost

For our third method, we consider the scalable XGBoost algorithm. The idea is that the performance of current trees is improved upon by making more accurate predictions for observations for which the predictions of the previous trees are incorrect, such that each tree is dependent on its predecessor.

Adopting the notation of Chen and Guestrin (2016), the XGBoost algorithm minimizes a negative log-likelihood loss function that measures the difference between the prediction $\hat{y}_i$ and the true outcome y_i for each individual, using a regularization term $\Omega(f_k)$. Specifically, the XGBoost regression algorithm minimizes the following objective function

$$\mathcal{L}(\hat{y}_i) = \sum_{i=1}^{N} l(y_i, \hat{y}_i) + \sum_{k=1}^{T} \Omega(f_k), \tag{1}$$

where $l(\hat{y}_i, y_i)$ is a differentiable convex training loss function and the regularization term equals $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2$ for $f_k \in (f_1, ..., f_T)$. The predicted values $\hat{y}_i$ are obtained by sequentially building shallow decision trees f_k, where the importance weights of every observation are updated proportional to their misclassification error in the previous tree. Each f_k corresponds to an independent tree structure q with leaf weights ω. The additional regularization term $\Omega(f_k)$ penalizes the complexity of regression tree f_k and consequently aids in the reduction of over-fitting. Given that the dependent variable in this research is heavily zero-inflated, we use the Tweedie loss function described in Yang et al. (2018) for $l(y_i, \hat{y}_i)$, which can be written as

$$l\left(y_i, \hat{y}_i, p\right) = \sum_{i=1}^{N} -y_i \frac{\hat{y}^{(1-p)}}{1-p} + \frac{\hat{y}^{(2-p)}}{2-p},$$

where p is the Tweedie power parameter[8]. Instead of using traditional optimization methods to minimize

[8]Tweedie distributions are a family of distributions that include gamma, normal, Poisson and their combinations. The power parameter allows the user to specify which mean-variance relation to use.

the objective function in Equation 1, Chen and Guestrin (2016) propose to train the model in an additive manner. Let $\hat{y}_i^{(t)}$ bbe the i^{th} instance at the t^{th} iteration, we can rewrite the objective function as

$$\mathcal{L}^{(t)} = \sum_{i=1}^{N} l\left(y_i, \hat{y}_i^{(t-1)} + f_t\left(\mathbf{x}_i\right)\right) + \Omega\left(f_t\right), \tag{2}$$

where the algorithm greedily adds a tree f_t that most improves the model according to Equation 1. The rewritten objective function in Equation 2 is then optimized using second-order Taylor approximation. Thereafter, we can calculate the optimal weight of each leaf and the corresponding optimal value for fixed tree structures q[9].

4.2 Mixed-effects machine learning

In a general GLMM framework, the model assumes that the responses y_{it} for a single subject i, conditional on an (assumed iid normal) subject-specific risk factor γ_i, are independent and follow a distribution from the exponential family with mean: $E(y_{it}|\gamma_i) = \mu_{it} = h(\eta_{it})$, where $\eta_{it} = \beta' x_{it} + \gamma_i$, where $h^{-1}(\cdot) = g(\cdot)$ represents the link function and β represents the vector of population fixed-effect coefficients. The GLMM assumes a parametric distribution and imposes restrictive linear relationships between the link function $g(\cdot)$ and the covariates. Machine learning algorithms do not make *a priori* assumptions on the distribution, but, as mentioned before, they do often implicitly make the iid assumption.

Ngufor et al. (2019) propose a MEml framework, which estimates the fixed-effects component ($\beta' x_{it}$) using machine learning algorithms. Thus η_{it} is now defined as

$$\eta_{it} = f(x_i) + \gamma_i, \tag{3}$$

with estimated dependent variable

$$y_i = f(x_i) + \gamma_i + \varepsilon_i, \tag{4}$$

where the function $f(\cdot)$ is unknown, and must be estimated. While Ngufor et al. (2019) use only tree based algorithms to estimate $f(\cdot)$, they state that any supervised learning algorithm can be used. In turn, this research contributes to the existing literature by using both ANNs and SVRs in this MEml framework. The proposed MEml models are estimated using the expectation-maximization approach, in which the random effects in Equation 3 and the population-level effects in 4 are alternatively estimated. In essence, we first initialize the random effects $\hat{\gamma}_i = 0$, and use this $\hat{\gamma}_i$ to compute $y_{it}^* = y_{it} - \hat{\gamma}_i$. We then train our machine learning model to estimate $\hat{f}(x_{it})$ in Equation 4 using y_{it}^*. Finally, we estimate γ_i in Equation 3 using $\hat{f}(x_{it})$. This process repeats until convergence[10].

[9] For a full mathematical formulation of these values, we refer to Chen and Guestrin (2016).
[10] For more details on the estimation procedure, we refer to Ngufor et al. (2019).

4.3 Tuning

As with most machine learning methods, the performance of both ANNs and SVRs are dependent on proper tuning. Due to the large dimensionality of the parameter grid, we cross-validate the hyperparameters of the different models using a Bayesian search (Bergstra et al., 2013). This method first explores the parameter space and then performs a guided search in (seemingly) promising subspaces in terms of cross-validated accuracies. The Hyperopt method can be seen as an exploration/exploitation strategy, that starts by exploring the performance across the candidate hyperparameter space, and subsequently randomly exploits the most promising subspace of hyperparameters. For the same number of iterations, this method can lead to better hyperparameter settings than the ones of random search.

For computational efficiency, Putatunda and Rama (2018) introduce Randomized Hyperopt. This method first randomly samples a predetermined fraction $\rho \in [0, 1]$ from the validation train fold without replacement, and then performs a Hyperopt iteration on this sampled fold. In their application, they show that the loss in performance is limited, while drastically decreasing computation time, allowing for more Hyperopt iterations. We employ Randomized Hyperopt with eight folds.

For the ANNs, we tune the number of hidden layers, dropout rate, early stopping, number of neurons, batch size, and the learning rate. We do not consider weight decay since we already account for overfitting with early stopping and dropout. For the SVRs we tune the kernel, degree of non-linearity, regularization parameter, and ε.

In addition, normalization might be necessary, as Jayalakshmi and Santhakumaran (2011) show that the performance of NN is contingent on normalization of the explanatory variables. For SVMs, Herbrich and Graepel (2000) show that normalisation of the feature vectors leads to increased performance as well. We consider four distinct normalization schemes: no normalization, min-max normalization, standardization, and robust standardization using the median and $25\% - 75\%$ interquantile range.

4.4 Forecasting

When making predictions outside of MISCAN, the models use the age, sex, birth year, stage, and FIT sequence number at time t, and we use the haemoglobin difference, maximum, minimum, stage, and previous haemoglobin value y_{t-1}^{Hb} at time $t-1$, to predict $\hat{y}_t^{Hb}$.

It is inappropriate to randomly split the data into train and test sets due to the longitudinal nature of the data set. Consequently, we create two distinct groups of individuals to create the train, validation and test set. Group 1 is for training and validation, and group 2 is for testing. Group 1 contains 480,000 individuals (due to computation time, distributed over eight cores), and group 2 contains the remaining 2,703,824 individuals for testing. Thus, each ID occurs only once between these two groups.

4.5 Class imbalance

Since the data is zero-inflated, and therefore highly unbalanced, we use the state-of-the-art rebalancing technique SMOTE-NC (Chawla et al., 2002) for the training data of group 1 only, along with either ENN (Wilson, 1972), Tomek Links (Tomek, 1976), or NearMiss, depending on computational feasibility within

time constraints. Naturally, we do not perform oversampling on the validation and test sets, as these sets should conform to the original class distribution.

https://datascience.stackexchange.com/questions/69085/smote-for-regression smote for regression in R en anders smogn in python https://github.com/nickkunz/smogn

4.6 Performance measures

The root mean squared error (RMSE), mean absolute error (MAE), and median absolute error (MedAE) are used to assess individual predictions. We use either the Diebold-Mariano (DM) test, or model confidence sets, to test for significant differences between models.

5 Results

6 conclusion

References

Ambler, G., Omar, R. Z., and Royston, P. (2007). A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome. *Statistical Methods in Medical Research*, 16(3):277–298.

Awad, M. and Khanna, R. (2015). Support Vector Regression. In *Efficient Learning Machines*, pages 67–80. Springer.

Baneshi, M. and Talei, A. (2011). Multiple Imputation in Survival Models: Applied on Breast Cancer Data. *Iranian Red Crescent Medical Journal*, 13(8):544.

Barton, M. B., Moore, S., Polk, S., Shtatland, E., Elmore, J. G., and Fletcher, S. W. (2001). Increased patient concern after false-positive mammograms. *Journal of General Internal Medicine*, 16(3):150–156.

Bergstra, J., Yamins, D., and Cox, D. D. (2013). Hyperopt: A Python Library for Optimizing the Hyperparameters of Machine Learning Algorithms. In *Proceedings of the 12th Python in Science Conference*, volume 13, page 20. Citeseer.

van den Berg, D. (2021). Simulation of haemoglobin concentrations in MISCAN-Colon using a mixed-effect machine learning model. Master's thesis, Erasmus University Rotterdam.

Bhaskaran, K. and Smeeth, L. (2014). What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology*, 43(4):1336–1339.

Botteri, E., Iodice, S., Bagnardi, V., Raimondi, S., Lowenfels, A. B., and Maisonneuve, P. (2008). Smoking and Colorectal Cancer: A Meta-analysis. *Journal of the American Medical Association*, 300(23):2765–2778.

Brasso, K., Ladelund, S., Frederiksen, B. L., and Jørgensen, T. (2010). Psychological distress following fecal occult blood test in colorectal cancer screening–a population-based study. *Scandinavian Journal of Gastroenterology*, 45(10):1211–1216.

Brenner, H., Stock, C., and Hoffmeister, M. (2014). Effect of screening sigmoidoscopy and screening colonoscopy on colorectal cancer incidence and mortality: systematic review and meta-analysis of randomised controlled trials and observational studies. *British Medical Journal*, 348.

Brodersen, J. and Siersma, V. D. (2013). Long-Term Psychosocial Consequences of False-Positive Screening Mammography. *The Annals of Family Medicine*, 11(2):106–115.

Bronner, M. P. and Haggitt, R. C. (1993). The Polyp-Cancer Sequence: Do All Colorectal Cancers Arise from Benign Adenomas? *Gastrointestinal Endoscopy Clinics of North America*, 3(4):611–622.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Chen, S. and DuBois Bowman, F. (2011). A Novel Support Vector Classifier for Longitudinal High-dimensional Data and its Application to Neuroimaging Data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(6):604–611.

Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM.

Cheng, Q., Tezcan, J., and Cheng, J. (2014). Confidence and prediction intervals for semiparametric mixed-effect least squares support vector machine. *Pattern Recognition Letters*, 40:88–95.

Cho, D.-H. (2010). Mixed-effects LS-SVR for longitudinal data. *Journal of the Korean Data and Information Science Society*, 21(2):363–369.

Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. (2016). Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *Machine Learning for Healthcare Conference*, pages 301–318. Proceedings of Machine Learning Research.

Chowdhury, M. H., Islam, M. K., and Khan, S. I. (2017). Imputation of Missing Healthcare Data. In *20th International Conference of Computer and Information Technology*, pages 1–6. IEEE.

Compton, C. C. and Greene, F. L. (2004). The Staging of Colorectal Cancer: 2004 and Beyond. *CA: A Cancer Journal for Clinicians*, 54(6):295–308.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273–297.

Ding, H., Lin, J., Xu, Z., Chen, X., Wang, H. H., Huang, L., Huang, J., Zheng, Z., and Wong, M. C. (2022). A Global Evaluation of the Performance Indicators of Colorectal Cancer Screening with Fecal Immunochemical Tests and Colonoscopy: A Systematic Review and Meta-Analysis. *Cancers*, 14(4):1073.

Du, W., Cheung, H., Johnson, C. A., Goldberg, I., Thambisetty, M., and Becker, K. (2015). A Longitudinal Support Vector Regression for Prediction of ALS Score. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1586–1590. IEEE.

Faris, P. D., Ghali, W. A., Brant, R., Norris, C. M., Galbraith, P. D., Knudtson, M. L., Investigators, A., et al. (2002). Multiple imputation versus data enhancement for dealing with missing data in observational health care outcome analyses. *Journal of Clinical Epidemiology*, 55(2):184–191.

Fazio, V. W., Tekkis, P. P., Remzi, F., and Lavery, I. C. (2004). Assessment of operative risk in colorectal cancer surgery: the Cleveland Clinic Foundation colorectal cancer model. *Diseases of the Colon & Rectum*, 47(12):2015–2024.

Frampton, M., Law, P., Litchfield, K., Morris, E., Kerr, D., Turnbull, C., Tomlinson, I., and Houlston, R. (2016). Implications of polygenic risk for personalised colorectal cancer screening. *Annals of Oncology*, 27(3):429–434.

Grobbee, E. J., Schreuders, E. H., Hansen, B. E., Bruno, M. J., Lansdorp-Vogelaar, I., Spaander, M. C., and 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., and 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., and 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.

Hanley, J. A. (2005). Analysis of Mortality Data from Cancer Screening Studies: Looking in the Right Window. *Epidemiology*, pages 786–790.

Herbrich, R. and Graepel, T. (2000). A PAC-Bayesian Margin Bound for Linear Classifiers: Why SVMs work. In Leen, T., Dietterich, T., and Tresp, V., editors, *Advances in Neural Information Processing Systems*, volume 13. MIT Press.

Hewitson, P., Glasziou, P., Watson, E., Towler, B., and Irwig, L. (2008). Cochrane Systematic Review of Colorectal Cancer Screening Using the Fecal Occult Blood Test (Hemoccult): An Update. *Journal of the American College of Gastroenterology*, 103(6):1541–1549.

Holme, Ø., Bretthauer, M., Fretheim, A., Odgaard-Jensen, J., and Hoff, G. (2013). Flexible sigmoidoscopy versus faecal occult blood testing for colorectal cancer screening in asymptomatic individuals (Review). *Cochrane Database of Systematic Reviews*.

Hornik, K., Stinchcombe, M., and White, H. (1989). Multilayer Feedforward Networks Are Universal Approximators. *Neural networks*, 2(5):359–366.

Jayalakshmi, T. and Santhakumaran, A. (2011). Statistical Normalization and Back Propagation for Classification. *International Journal of Computer Theory and Engineering*, 3(1):1793–8201.

Jenniskens, K., De Groot, J. A., Reitsma, J. B., Moons, K. G., Hooft, L., and Naaktgeboren, C. A. (2017). Overdiagnosis across medical disciplines: a scoping review. *BMJ Open*, 7(12):e018448.

Jiang, Y., Yuan, H., Li, Z., Ji, X., Shen, Q., Tuo, J., Bi, J., Li, H., and Xiang, Y. (2022). Global pattern and trends of colorectal cancer survival: a systematic review of population-based registration data. *Cancer Biology & Medicine*, 19(2):175.

Jolani, S., Debray, T. P., Koffijberg, H., van Buuren, S., and Moons, K. G. (2015). Imputation of systematically missing predictors in an individual participant data meta-analysis: a generalized approach using mice. *Statistics in Medicine*, 34(11):1841–1863.

Levin, B., Lieberman, D. A., McFarland, B., Andrews, K. S., Brooks, D., Bond, J., Dash, C., Giardiello, F. M., Glick, S., Johnson, D., et al. (2008). Screening and Surveillance for the Early Detection of Colorectal Cancer and Adenomatous Polyps, 2008: A Joint Guideline From the American Cancer Society, the US Multi-Society Task Force on Colorectal Cancer, and the American College of Radiology. *Gastroenterology*, 134(5):1570–1595.

Lippmann, R. (1987). An Introduction to Computing with Neural Nets. *IEEE ASSP magazine*, 4(2):4–22.

Loeve, F., Boer, R., van Oortmarssen, G. J., van Ballegooijen, M., and 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.

Luts, J., Molenberghs, G., Verbeke, G., van Huffel, S., and Suykens, J. A. (2012). A mixed effects least squares support vector machine model for classification of longitudinal data. *Computational Statistics & Data Analysis*, 56(3):611–628.

Mandel, F., Ghosh, R. P., and Barnett, I. (2021). Neural networks for clustered and longitudinal data using mixed effects models. *Biometrics: A Journal of the International Biometric Society*.

Morson, B. (1974). The polyp-cancer sequence in the large bowel. *Journal of the Royal Society of Medicine*, 67:451–457.

Mousavinezhad, M., Majdzadeh, R., Sari, A. A., Delavari, A., and Mohtasham, F. (2016). The effectiveness of FOBT vs. FIT: A meta-analysis on colorectal cancer screening test. *Medical Journal of the Islamic Republic of Iran*, 30:366.

Ngufor, C., van Houten, H., Caffo, B. S., Shah, N. D., and McCoy, R. G. (2019). Mixed Effect Machine Learning: A framework for predicting longitudinal change in hemoglobin A1c. *Journal of Biomedical Informatics*, 89:56–67.

Pashayan, N., Duffy, S. W., Chowdhury, S., Dent, T., Burton, H., Neal, D. E., Easton, D. F., Eeles, R., and Pharoah, P. (2011). Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. *British Journal of Cancer*, 104(10):1656–1663.

Prechelt, L. (1998). Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767.

Putatunda, S. and Rama, K. (2018). A Comparative Analysis of Hyperopt as Against Other Approaches for Hyper-Parameter Optimization of XGBoost. In *Proceedings of the 2018 International Conference on Signal Processing and Machine Learning*, pages 6–10.

Raghunathan, T. E., Solenberger, P. W., and Van Hoewyk, J. (2002). IVEware: Imputation and Variance Estimation Software. *University of Michigan*.

Sakthivel, K. and Rajitha, C. (2017). A Comparative Study of Zero-inflated, Hurdle Models with Artificial Neural Network in Claim Count Modeling. *International Journal of Statistics and Systems*, 12(2):265–276.

Schröder, F. H., Hugosson, J., Roobol, M. J., Tammela, T. L., Ciatto, S., Nelen, V., Kwiatkowski, M., Lujan, M., Lilja, H., Zappa, M., et al. (2009). Screening and Prostate-Cancer Mortality in a Randomized European Study. *New England Journal of Medicine*, 360(13):1320–1328.

Seok, K. H., Shim, J., Cho, D., Noh, G.-J., and Hwang, C. (2011). Semiparametric mixed-effect least squares support vector machine for analyzing pharmacokinetic and pharmacodynamic data. *Neurocomputing*, 74(17):3412–3419.

Śmieja, M., Struski, Ł., Tabor, J., Zieliński, B., and Spurek, P. (2018). Processing of missing data by neural networks. *Advances in Neural Information Processing Systems*, 31.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

van der Steeg, A., Keyzer-Dekker, C., De Vries, J., and Roukema, J. (2011). Effect of abnormal screening mammogram on quality of life. *Journal of British Surgery*, 98(4):537–542.

Stewart, T. G., Zeng, D., and Wu, M. C. (2018). Constructing support vector machines with missing data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(4):e1430.

Strum, W. B. (2016). Colorectal Adenomas. *New England Journal of Medicine*, 374(11):1065–1075.

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249.

Tandon, R., Adak, S., and Kaye, J. A. (2006). Neural networks for longitudinal studies in Alzheimer's disease. *Artificial Intelligence in Medicine*, 36(3):245–255.

Thanikachalam, K. and Khan, G. (2019). Colorectal Cancer and Nutrition. *Nutrients*, 11(1):164.

Thrumurthy, S. G., Thrumurthy, S. S., Gilbert, C. E., Ross, P., and Haji, A. (2016). Colorectal adenocarcinoma: risks, prevention and diagnosis. *British Medical Journal*, 354.

Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions Systems, Man and Cybernetics*, 6:769–772.

Toribara, N. W. and Sleisenger, M. H. (1995). Screening for Colorectal Cancer. *New England Journal of Medicine*, 332(13):861–867.

Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global Cancer Statistics, 2012. *CA: A Cancer Journal for Clinicians*, 65(2):87–108.

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. CRC press.

Van Buuren, S. and Oudshoorn, K. (1999). *Flexible multivariate imputation by MICE*. TNO Prevention and Health.

van Duuren, L. A., Ozik, J., Spliet, R., Collier, N. T., Lansdorp-Vogelaar, I., and Meester, R. G. (2022). An Evolutionary Algorithm to Personalize Stool-Based Colorectal Cancer Screening. *Frontiers in Physiology*, page 2515.

Wardle, J., Williamson, S., Sutton, S., Biran, A., McCaffery, K., Cuzick, J., and Atkin, W. (2003). Psychological Impact of Colorectal Cancer Screening. *Health Psychology*, 22(1):54.

Welch, H. G. and Black, W. C. (2010). Overdiagnosis in cancer. *Journal of the National Cancer Institute*, 102(9):605–613.

Whitlock, E. P., Lin, J. S., Liles, E., Beil, T. L., and Fu, R. (2012). Screening for Colorectal Cancer: A Targeted, Updated Systematic Review for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, 157(2):120–134.

Wilson, D. L. (1972). Asymptotic Properties of Nearest Neighbor Rules Using Edited Data. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421.

Winawer, S. J. (2007). Colorectal cancer screening. *Best Practice & Research Clinical Gastroenterology*, 21(6):1031–1048.

Xiong, Y., Kim, H. J., and Singh, V. (2019). Mixed Effects Neural Networks (MeNets) With Applications to Gaze Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, Y., Qian, W., and Zou, H. (2018). Insurance Premium Prediction via Gradient Tree-Boosted Tweedie Compound Poisson Models. *Journal of Business & Economic Statistics*, 36(3):456–470.

Zorzi, M., Fedeli, U., Schievano, E., Bovo, E., Guzzinati, S., Baracco, S., Fedato, C., Saugo, M., and Dei Tos, A. P. (2015). Impact on colorectal cancer mortality of screening programmes based on the faecal immunochemical test. *Gut*, 64(5):784–790.

A Data

The data cleaning procedure is as follows. We first delete variables which are inane to this papers analysis (e.g., information on the morphology and topography of a cancer), and variables which possibly contain patient sensitive information (e.g., participation date, patient pseudonym, and invitation date). We then remove individuals with invalid or missing entries, individuals who returned to the data set after a positive FIT, and individuals younger than 55 or older than 77 in round 1 of 2014. The final data set only includes individuals who participated in two or more consecutive rounds and those who participated in one round at most.

With respect to data engineering, we first transform `result` to attain three categories: favourable, unfavourable, and missing. Here, 'unfavourable' contains all observations with '*unfavourable*' and '*unfavourable (unreliable)*' as result, and 'favourable' contains only observations with '*favourable*' as result. The remaining observations are cast to 'missing', and are deleted from the data set. Hereafter, given that the results of the FIT are based on two thresholds: 275 ng/ml and 47 μg/g, we multiply observations where `haemoglobin current` is based on 275 as threshold by $\frac{47}{275}$, such that all haemoglobin values are represented in the same unit. Finally, we create the following variables: `haemoglobin previous`, `haemoglobin max`, `haemoglobin min`, and we perform one-hot-encoding to `FIT number` and `stage`. More detailed descriptions of each of the variables in the final data set are shown in Table 1.

A.1 MICE

This research employs Multiple Imputation via Chained Equations (MICE) to impute missing values in the stage variable of the original data set. To run this algorithm we create a data set consisting of two data sets from the Dutch screening program and a simulated population run in MISCAN-Colon. Table 2 shows a subset of the combined data. Note that the the '15 threshold' data set contains information on all variables at all times, while the simulated population never contains information on `haemoglobin current` and the original data set only contains information on `stage` 3.8% of the time. Table 3 shows descriptive statistics for each of the additional data sets.

The MICE iterations are as follows:

1 First replace all missing values with placeholders. In this case, all missing values are replaced by a random draw of data (with replacement) within each respective variable.

2.1 Remove the placeholder of `haemoglobin current`.

2.2 Use random sampling to impute all missing values in `haemoglobin current`.

3.1 Remove the placeholder of `stage`.

3.2 Using the newly imputed `haemoglobin current` in combination with `result`, `age` and `sex`, perform predictive mean matching to impute `stage`.

Once these steps are complete, we have completed one full cycle of MICE. We perform 5 $\times$ 10 cycles, after which we are left with five distinct imputed `stage` variables. We then compare the distribution of

stages in each of these imputed variables to the distribution of stages in the 'MISCAN simulation' data set, and select the imputed variable which most closely matches the distribution in the MISCAN `stage` to replace the original `stage`. Finally, we drop the '15 threshold' and 'MISCAN simulation' data sets.

Table 2: Multiple Imputation via Chained Equations exemplified

Step 0

ID	Result	Age	Sex	Hb	Stage
471	Negative	68	Female	0	NA
471	Negative	70	Female	20.0	NA
471	Positive	72	Female	307.1	4
⋮					
151	Negative	73	Male	37.3	1
152	Positive	73	Female	47.7	2
⋮					
MI1	Negative	65	Male	NA	1
MI1	Negative	58	Male	NA	2

Step 1

ID	Result	Age	Sex	Hb	Stage
471	Negative	68	Female	0	2
471	Negative	70	Female	20.0	2
471	Positive	72	Female	307.1	4
⋮					
151	Negative	73	Male	37.3	1
152	Positive	73	Female	47.7	2
⋮					
MI1	Negative	65	Male	37.3	1
MI1	Negative	58	Male	47.7	2

Step 2.1

ID	Result	Age	Sex	Hb	Stage
471	Negative	68	Female	0	2
471	Negative	70	Female	20.0	2
471	Positive	72	Female	307.1	4
⋮					
151	Negative	73	Male	37.3	1
152	Positive	73	Female	47.7	2
⋮					
MI1	Negative	65	Male	?	1
MI1	Negative	58	Male	?	2

Step 2.2

ID	Result	Age	Sex	Hb	Stage
471	Negative	68	Female	0	2
471	Negative	70	Female	20.0	2
471	Positive	72	Female	307.1	4
⋮					
151	Negative	73	Male	37.3	1
152	Positive	73	Female	47.7	2
⋮					
MI1	Negative	65	Male	20.8	1
MI1	Negative	58	Male	42.6	2

Step 3.1

ID	Result	Age	Sex	Hb	Stage
471	Negative	68	Female	0	?
471	Negative	70	Female	20.0	?
471	Positive	72	Female	307.1	4
⋮					
151	Negative	73	Male	37.3	1
152	Positive	73	Female	47.7	2
⋮					
MI1	Negative	65	Male	20.8	1
MI1	Negative	58	Male	42.6	2

Step 3.2

ID	Result	Age	Sex	Hb	Stage
471	Negative	68	Female	0	1
471	Negative	70	Female	20.0	1
471	Positive	72	Female	307.1	4
⋮					
151	Negative	73	Male	37.3	1
152	Positive	73	Female	47.7	2
⋮					
MI1	Negative	65	Male	20.8	1
MI1	Negative	58	Male	42.6	2

Notes: This table represents an exemplified version of one full cycle of Multiple Imputation via Chained Equations. The data set consists of individuals from the original, the '15 threshold' and the 'MISCAN simulation' data set, denoted by 47*, 15* and MI* as `ID` preface, respectively. The red numbers in Step 1 are obtained from a random draw with replacement from the full data set. The red numbers in Step 2.2 and 3.2 are obtained through predictive mean matching using all variables except the one that will be imputed (i.e., excluding the variable with a question mark in Step 2.1 and 3.1, respectively). `Hb` represents `haemoglobin current`. For more information on each variable see Table 1. The numbers in this table are for illustrative purposes only.

Table 3: Descriptive statistics of additional data sets required for performing MICE

Variable	**Data set**	
	MISCAN simulation	15 threshold
`Age`	$[55; 77]$	$[56; 76]$
`Haemoglobin current`	–	$[0; 292.8]^*$
`Haemoglobin threshold`	–	$[15, 45, 88, 275]$
`Sex`	0 (Male, 48%), 1 (Female, 52%)	0 (Male, 58.2%), 1 (Female, 47.8%)
`Stage`	1 (Healthy, 84.3%), 2 (Non-advanced adenoma, 8.3%), 3 (Advanced adenoma, 7.0%), 4 (Colorectal cancer, 0.4%)	1 (Healthy, 20.4%), 2 (Non-advanced adenoma, 28.9%), 3 (Advanced adenoma, 42.8%), 4 (Colorectal cancer, 7.9%)

Notes: *The '15 threshold' data set contains five observations with `haemoglobin current` below the (lowest) threshold of 15 μg/g, which were not deleted since `stage` was known.