Name student: Duco Twan Mülder

Student ID number: 434081

Predicting Colorectal Cancer Screening Outcomes: an Application of Advanced Ordered Outcome Models and Machine Learning

Master Thesis

to achieve the university degree of

Master of Science

Econometrics and Operations Research

Business Analytics and Quantitative Marketing

submitted to

ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economics

Supervisor: dr. E.P. O'Neill

Second assessor: prof. dr. P.J.F. Groenen

Supervisor from Erasmus Medical Center: dr. R.G.S. Meester

Rotterdam, October 14, 2021

Abstract

Accurate prediction of who is at risk of adverse outcomes can help improve cancer screening programmes. In this research, we have investigated whether we can improve the performance of the previously applied predictive models by including knowledge on the stages of disease development in the Dutch screening programme for colorectal cancer (CRC). We implement existing ordinal models as well as provide a new finite mixture approach, constituting our finite mixture generalised ordered logit model. This general model has - to the best of our knowledge - not yet been described in the academic literature. This study confirms that age, gender, and previously measured hemoglobin concentrations can predict future colonoscopy outcomes. This indicates that risk stratification can increase the effectiveness of CRC screening programmes. Additionally, we find that our finite mixture model outperforms the standard ordered logit model in terms of ordinal classification. Moreover, we implement a random forest algorithm that exceeds the predictive power of regression models at the cost of interpretability. All our models can be used to compute individual risk scores. We conclude that these risk scores can be used in a clinical study to evaluate the effectiveness of risk stratification in CRC screening.

Keywords: colorectal cancer • screening • ordinal models • continuation ratio • generalised ordered logit • finite mixture models • machine learning • risk stratification

Contents

1 Introduction 1

1.1 Outlook . . . 2

2 Literature Review 3

2.1 Colorectal cancer disease development . . . 3

2.2 Cancer screening and risk stratification . . . 4

2.3 Factors driving CRC . . . 4

2.4 Predicting binary colorectal cancer screening outcomes . . . 5

2.5 Regression models for predicting categories . . . 6

2.6 Modelling heterogeneity . . . 7

3 Data 8

3.1 Study population . . . 8

3.2 Data on colonoscopy outcomes . . . 9

3.3 Ordered categories . . . 9

4 Methodology 12

4.1 Baseline model . . . 12

4.2 Ordered logit . . . 12

4.3 Continuation ratio . . . 14

4.4 Finite mixture ordered logit . . . 15

4.5 Finite mixture generalised ordered logit . . . 17

4.5.1 Generalised ordered logit . . . 18

4.5.2 Mixture approach in gologit . . . 18

4.5.3 Restrictions . . . 19

4.6 Parameter estimation . . . 20

4.6.1 Expectation-Maximisation algorithm . . . 20

4.6.2 Standard errors . . . 21

4.7 Random forest . . . 22

4.8 Model evaluation . . . 22

4.8.1 Tests for regression coefficients . . . 22

4.8.2 Discriminatory ability . . . 23

4.8.3 Ordinal classification performance . . . 24
4.8.4 Brier scores . . . 25
4.8.5 Akaike Information Criterion . . . 26
4.8.6 Risk stratifying ability . . . 26

5 Results 27
5.1 Category specific effects . . . 27
5.2 Parameters of the finite mixture models . . . 29
5.3 Goodness of fit . . . 31
5.4 Numeric performance summary statistics . . . 32
5.4.1 Discriminatory ability . . . 32
5.4.2 Ordinal classification performance . . . 34
5.4.3 Brier scores . . . 34
5.4.4 Akaike Information Criterion . . . 35
5.5 Risk stratifying ability . . . 35

6 Conclusion & Discussion 37
6.1 Limitations . . . 38
6.2 Implications for future research and practice . . . 39

References 41

7 Appendix 46

Acronyms

AA Advanced adenoma

AIC Akaike information criterion

AMAE Average mean absolute error

AN Advanced neoplasia

ANN Artificial neural network

AUC Area under the curve

BART Bayesian Additive Regression Trees

cdf Cumulative distribution function

CI confidence interval

CRC colorectal cancer

EM Expectation maximisation

FIMGOL Finite mixture generalised ordered logit

FIT Fecal immunochemical test

FMOL Finite mixture ordered logit

FSB Facilitaire Samenwerking Bevolkingsonderzoeken (Dutch center for screening programmes)

gologit Generalised ordered logit

Hb Hemoglobin

MAE Mean absolute error

MC Medical center

MMAE Maximum mean absolute error

NA Non-adenomatous lesion

NAA Non-advanced adenoma

pdf Probability distribution function

RIVM Rijksinstituut voor Volksgezondheid en Milieu (Dutch national institute for public health and the environment)

ROC Receiver operating characteristic

For readability purposes, we will reexplain any acronym used in each section. Each section can therefore be read separately. Vectors will be written in lowercase, bold letters. Matrices will be written in uppercase, bold letters.

1 Introduction

Colorectal cancer (CRC) is one of the most common types of cancer in the Netherlands. In 2019, 12,900 people were diagnosed with CRC and 4,826 people died as a consequence of CRC. Treatments to CRC are effective, provided that a patient is diagnosed at an early stage. For that reason, the Dutch government has initiated a national screening programme for CRC. The first evaluations conclude that the programme has been successful in detecting CRC and its precursors at an early stage and thereby causing CRC mortality to decrease in the future (Lansdorp-Vogelaar et al., 2019).

Currently, all Dutch citizens aged between 55 and 75 are biennially invited to participate in the screening programme. Participants collect a small amount of stool and send a sample to the screening organisations. Screening is then based on the Fecal Immunochemical Test (FIT), which measures the level of hemoglobin (Hb) in the stool. A participant is invited for further research - an endoscopic examination in the form of a colonoscopy - when the stool contains more than a fixed cut-off level of Hb. Whether someone is asked to undergo a colonoscopy solely depends on the outcome of the most recent FIT-concentration. Age, gender, and previously measured FIT-concentrations are not taken into consideration.

Research however shows that age, gender, and historical FIT-concentrations contain predictive power for the outcomes of colonoscopies. Meester (2021) uses a standard logistic regression model to predict whether an advanced lesion is found in the endoscopic examination in the third Dutch screening round for individuals who tested below the cut-off value in the first two screening rounds. Based on the area under the Receiver Operating Characteristic (ROC) curve, one can conclude that age, gender, and the two previously measured FIT-concentrations are well able to predict whether CRC is detected in test round three.

In addition to detection, prevention of CRC is an important aspect of the screening programme. As malignant tumours develop from harmless polyps, the detection and most often instant removal of polyps during a colonoscopy, assures that the screening programme also prevents CRC cases. The existing predictive models however, solely predict a binary outcome variable for advanced stages of polyps. To draw conclusions on whether a colonoscopy was useful, it is desirable to also include information about CRC prevention effects through the detection of other lesions. Instead of a binary outcome model, we therefore opt for a multinomial modelling approach.

Rather than solely focusing on the predictive performance, we also wish to gain insight in the behaviour of the disease development and its interaction with previously measured FIT-values.

Understanding the path towards malignant lesions is essential in many simulation models that are used to determine optimal cancer screening strategies (Loeve et al., 1999). We are particularly interested in whether previously measured FIT-values are able to describe and predict the stage of a lesion.

In this research, we aim to model the progression of colorectal lesions and improve the accuracy of previously applied predictive models. We therefore pose the following research question:

> **Research Question:** Can we improve the performance of the previously applied predictive models by including knowledge on the stages of disease development in the Dutch screening programme for colorectal cancer?

By including knowledge on the stages of disease development, we hope to be able to increase the detection rate of CRC and its precursors and decrease the number of unnecessary colonoscopies. Finally, our models should be able to compute individual risk scores. These risk scores can then be used to personalise the screening procedure; the screening recurrence and/or the Hb cut-off value could be altered based on the predicted risk.

1.1 Outlook

This research uses advanced regression and machine learning techniques to improve the prediction accuracy of previously applied models. We propose new ordinal cluster models that, to the best of our knowledge, have not hitherto been described in the academic literature. Our general implementation of the new ordered models can be used in all other applications for which the order of a multinomial outcome variable is relevant.

Our innovative methodological research is supported by its direct application in predicting colonoscopy outcomes. Our models reaffirm that the values of previously measured negative FITs are predictive for future colonoscopy outcomes. Particularly interesting to clinicians is that our models show that FIT-values are well able to predict whether someone will obtain a FIT above the cut-off value in future, but that they are not indicative for the progression of an advanced lesion.

In addition to advanced regression techniques, we apply machine learning methods in predicting colonoscopy outcomes. Our random forest model outperforms the predictive performance of regression techniques in terms of the Area under the ROC-curve (AUC), at the cost of interpretability.

Our models provide additional evidence that risk stratification can improve the screening programme for CRC. Though further research in the application of more advanced machine learning

techniques could provide more evidence, we conclude that policy makers should consider risk personalisation in the screening for CRC at this point in time. We finally provide suggestions for both methodological as well as clinical further research.

With this research question, this paper continues with a review of the relevant literature regarding predictive modelling in CRC screening. We will then describe the data as provided by Erasmus Medical Center (Erasmus MC). Subsequently, we will describe the existing ordinal methodology and introduce our new models as generalisations of the existing frameworks. We will then evaluate our models based on several numeric performance statistics and visualisations. Finally, we will draw conclusions and provide suggestions for further research.

2 Literature Review

In this section, we review the existing literature regarding screening for colorectal cancer (CRC). We start by describing the disease development and how CRC screening affects the public health, followed by a description of the risk factors of CRC. We then review the previously applied models to predict colonoscopy outcomes. We subsequently introduce ordinal models to predict stages of lesions, rather than a binary outcome variable. Finally, we explain how we can model the existing heterogeneity in our data according to the existing literature.

2.1 Colorectal cancer disease development

The development of CRC is generally a process in which a benignant lesion develops into a malignant tumour. CRC mainly develops from a lesion known as an adenoma; a type of polyps arising from the lining of the intestine (Cooper et al., 2010). Cottet et al. (2012) classify adenomas into non-advanced and advanced adenomas, based on their size and histopathological features. Adenomas are not harmful, as long as they do not progress into advanced stages with abnormal tissue growth, also known as Advanced Neoplasia (AN) (Markowitz and Winawer, 1999). The removal of an adenoma could however prevent it from growing into CRC. Because CRC only causes symptoms until the tumour reaches a considerable size, patients are oftentimes diagnosed at a late stage in the absence of a screening programme (Simon, 2016). Mandelblatt et al. (1996) conclude that a late diagnosis rapidly deteriorates the prognosis of a patient.

Because patients benefit from the detection of the precursors of CRC, we wish to include the development of the disease in our models. The stages follow a natural ordering, starting with small adenomas growing into tumours. Our models should take this natural pathway into account.

2.2 Cancer screening and risk stratification

Screening programmes are proven to be effective in decreasing CRC incidence and CRC mortality (Schreuders et al., 2015). Particularly for the Dutch screening programme, Lansdorp-Vogelaar et al. (2019) conclude that the screening programme prevents one out of five CRC cases and one out three CRC deaths. Cruzado et al. (2013) show that screening programmes can also be cost-effective, as early detection can decrease the costs of treatment. Screening can therefore have a large positive impact on the public health (Meester et al., 2015).

Despite the proven effectiveness of the Dutch screening programme in early CRC detection, one can identify multiple disadvantages of cancer screening. First, screening programmes are costly operations that are not cost-effective in all cases (Rodgers et al., 1990). Second, intensive screening could lead to overdiagnosis: detecting a condition that would never have caused any symptoms (Esserman et al., 2013). Overdiagnosis could lead to psychological stress and is particularly harmful if it leads to unnecessary treatments. Finally, particularly in the case of colorectal cancer screening, patients with a false-positive Fecal Immunochemical Test (FIT) carry the burden of an unnecessary colonoscopy that contains some health risks on its own (de Wijkerslooth et al., 2012). Policy makers therefore continuously weigh the harms and benefits of screening programmes.

For that reason, screening organisations continuously evaluate the effectiveness of a programme and seek for improvement (Lansdorp-Vogelaar et al., 2019). A potential improvement of the CRC screening programme is risk stratification (Auge et al., 2014). This implies that the intensity of the screening depends on an individual's estimated risk level. For CRC screening, this means that the recurrence of invitations or the cut-off value of the FIT could be altered. Despite its potential benefits, risk stratification is not yet applied in large-scale CRC screening programmes (Schreuders et al., 2015).

If the models in this study can successfully distinguish participants with a higher risk on lesions from participants with a lower risk, the screening intensity could be personalised based on the estimated risk. This personalisation potentially increases the effectiveness of the screening programme, making the programme more valuable to the public health.

2.3 Factors driving CRC

Both the CRC incidence as well as the speed of growth of adenomas are affected by a wide range of factors next to age and gender. Crawford et al. (2012) for example find a higher CRC incidence in rural and socially deprived areas in the UK. Next to social-demographic factors, Strum (2016)

finds a positive association between CRC incidence and the consumption of alcohol, meat, and the presence of obesity as well as protective factors such as consumption of fruit, vegetables, and regular use of painkillers such as ibuprofen and diclofenac. Finally, molecular genetic studies show that the presence of certain heritable single nucleotide polymorphisms increases the risk on CRC (Peters et al., 2013). Data on these risk factors could increase the accuracy of risk prediction. However, to assure a high participation rate, policy makers have chosen to avoid comprehensive surveys about personal information and screen based on the FIT only (Toes-Zoutendijk et al., 2017).

The probability of getting CRC as well as the speed at which a lesion grows therefore depend on several social-demographic, behavioural, and genetic factors that are not known by the screening organisations. As it is clear that we do not observe all factors that influence CRC screening outcomes, we aim to improve the previously applied prediction models by incorporating this unobserved heterogeneity in our frameworks.

2.4 Predicting binary colorectal cancer screening outcomes

Researchers have shown that the detection of AN, a collection of highly developed polyps and tumours, in a colonoscopy can be predicted using historical values of the FIT. Meester (2021) has implemented a logistic regression model based on age, gender, and two previously measured FITs in the Dutch context. This logistic model has been implemented for several variable transformations and interactions. Finally, the model has been optimised with respect to the area under the Receiver Operating Characteristic (ROC) curve. The final model yields an Area Under the of the ROC-Curve (AUC) of 0.77 (95% Confidence Interval (CI): 0.76-0.78). The model therefore clearly outperforms random assignment, which would yield an AUC of 0.5.

Similarly, Cooper et al. (2018) predict the same binary outcome variable - whether AN is detected in a colonoscopy - for a British dataset. In contrast to Meester (2021), this model includes the most recent FIT-concentration and uses dummies for first-time invitees and previous non-responders. Cooper et al. (2018) implement both a logistic regression model as well as an Artificial Neural Network (ANN). Though the ANN outperforms the regression model in terms of the AUC, a large share of interpretability is lost in this method; the model fails to explain why individuals are assigned to a particular risk level.

In this research, we will start by reconstructing the model as implemented by Meester (2021). We will extend the binary approach to a multinomial ordered framework. Though machine learning algorithms could yield better predictions, Cooper et al. (2018) describe major interpretability

disadvantages. As interpretability is essential in medical applications, we choose to examine the performance of machine learning algorithms in this context, but to use advanced regression techniques as our main approach.

Both Meester (2021) and Cooper et al. (2018) construct their main performance metric around the ROC-curve. An ROC-curve is a graph with the false positives rate on the x-axis and the true positives rate on the y-axis at different threshold settings (Gu et al., 2009). By its construction, the area under this curve measures the ability of a test to distinguish between those with and those without the disease (Cooper et al., 2018). As the ROC-curve is only applicable for a binary classifier, the previously applied models cannot be compared to categorical models directly. Hand and Till (2001) have developed a multinomial equivalent to the ROC-curve, but it is not yet applicable to ordinal outcome data.

To compare our models to the previously applied binary outcome model, we will show how our models can be reduced back to binary classifiers and evaluate the AUC. The AUC however fails to reflect the value of our ordinal approach. We will therefore introduce new performance metrics to evaluate the ordinal classification performance in our methodology section.

2.5 Regression models for predicting categories

While the previously applied models predict a binary outcome variable, we aim to model the stage of lesions that can be found in a colonoscopy. As the stages follow upon each other, ordinal models are suitable for this application. These models utilise the ordinal nature of the data by describing various modes of stochastic ordering and hence eliminate the need for assigning scores or otherwise assuming cardinality instead of ordinality (McCullagh, 1980).

McCullagh (1980) modelled the cumulative probabilities of the ordered outcomes as a monotonic increasing transformation of a linear predictor onto the unit interval, assuming a logit or probit link function. This multinomial ordered logit model has become a widely used standard ordinal model. The framework, also called the proportional odds model, is however limited in the sense that it assumes that the effect of explanatory variables on category shifts is the same for all category shifts (Williams, 2016). This assumption, known as the parallel assumption, is often violated. Modellers have therefore developed several generalisations of the ordered logit model in which the parallel assumption is relaxed.

Both the Generalised Ordered Logit (gologit) model from Peterson and Harrell Jr (1990) as well as the continuation ratio model as specified by Feinberg (1980) relax the parallel assumption of the

proportional odds model. The gologit model selectively relaxes the parallel lines assumptions to allow for non-proportional odds for (a subset of) the explanatory variables (Peterson and Harrell Jr, 1990). The continuation ratio model also relaxes the parallel assumption and is based on conditional incremental cut points, with outcomes at a given level discarded after being compared to higher levels (Mcgowan et al., 2000). The continuation ratio model therefore has the structure of a discrete hazard model and is suitable to applications for which the outcomes follow a path over time (Ananth and Kleinbaum, 1997).

Because clinicians do not know whether the blood emission increases in parallel to the growth of the lesion size, the parallel assumption might be violated in our application. We therefore implement a generalised version of an ordered model. As the lesions in our application also grow over time, we will start by relaxing the parallel assumption through the continuation ratio model. For its generality, we will use the gologit model as a starting point for a mixture model to allow for unobserved heterogeneity.

2.6 Modelling heterogeneity

A popular way of modelling unobserved heterogeneity is by using a finite mixture model. Li (2018), Deng et al. (2006) and Tuma and Decker (2013) show how finite mixture models are able to determine similarly behaving segments in a dataset and improve the prediction accuracy of the model in the contexts of traffic analysis, biostatistics, and marketing respectively. Finite mixture regression models can therefore be a powerful method when observations are influenced by unobserved factors.

In our context, we wish to implement a finite mixture model for ordered data as proposed by Everitt and Merette (1990). It is supposed that the population is split into C classes and each class has its own data-generating process (Boes and Winkelmann, 2006). This implies that we relax the distributional assumption of the standard ordered model and its implied homogeneity. The segmenting approach is particularly useful in this application, as practitioners finally wish to use our results to define risk groups. Finite mixture ordered models can be implemented using the Expectation-Maximisation algorithm from Dempster et al. (1977).

In this study, we aim to combine the existing gologit approach with the existing finite mixture modelling techniques, hence creating a finite mixture generalised ordered model. This new model is, to the best of our knowledge, a new model in the econometric literature.

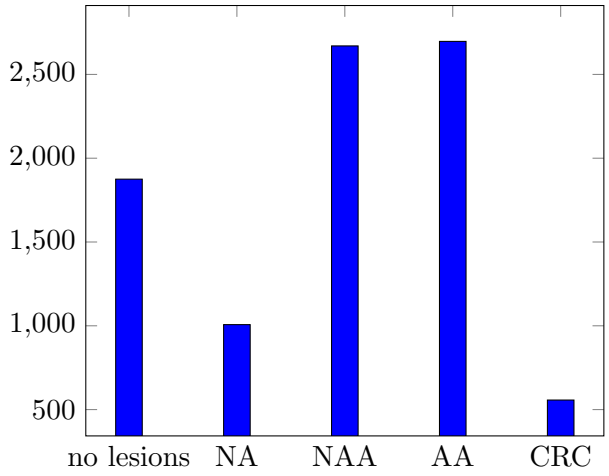
3 Data

In this research, we will use the data on the Dutch screening for Colorectal Cancer (CRC) as executed by the Dutch National Institute for Public Health and the Environment (*Rijksinstituut voor Volksgezondheid en Milieu, RIVM*). The data will be provided by the Erasmus Medical Center (Erasmus MC), which has acquired the data through the Dutch Center of Screening programmes (*Facilitaire Samenwerking Bevolkingsonderzoeken, FSB*). We will start by introducing the study population. We continue by describing the data on colonoscopy outcomes. Finally, we will describe the four naturally ordered categories in this application.

3.1 Study population

In the time period 2013 until 2018, $N_1 = 3,436,106$ people participated in one round of the screening programme, $N_2 = 1,555,752$ people participated in two rounds and $N_3 = 265,881$ people participated in three rounds. In this research, we aim to predict the outcome of endoscopic examination after the third test round, based on age, gender, and the previous two values of the Fecal Immunochemical Test (FIT) for individuals who tested below the cut-off value in the first two rounds. This means that for this research, we will only use the 265,881 observations who participated in all three test rounds.

The gender and age of the participants are known: 52% of the participants who consequently participated when they received an invitation were female and the average ages of the participants in the first three rounds were 64.1 years, 66.1 years and 69.0 years respectively. The follow-up research based on a positive test as a percentage of the number of participants per round declined from 5.1% in the first round, to 3.5% and 3.3% in the second and third round respectively. The decrease in percentage of positive FITs is in line with experts' expectations, as more cases of CRC can be detected in primary screening rounds.



Notes: This figure reports the distribution of the five outcome categories: no lesions, non-adenomatous lesions (NA), non-advanced adenomas (NAA), advanced adenomas (AA), and colorectal cancer (CRC). Note that the categories are ranked on the severity of the lesion.

Figure 1: Distribution of colonoscopy outcomes after a positive FIT in round three

3.2 Data on colonoscopy outcomes

Out of all 265,881 participants in screening round three, colonoscopy was performed on 8,806 people (3.3%) and five different outcomes can be distinguished:

1. No lesions found in colonoscopy;
2. Non-Adenomatous lesions (NA), including serrated polyps and hyperplastic polyps;
3. Non-Advanced Adenoma (NAA);
4. Advanced Adenoma (AA);
5. Colorectal Cancer (stage I and stage II) (CRC).

Note that the enumeration above is ranked based on the progression of the lesion. This means that the severity of the diagnosed lesion is increasing by category; category five represents having CRC. The higher the category level, the closer the polyp is to becoming a malignant tumour, hence causing CRC. If desired for modelling or interpretation purposes, we can combine categories or split categories two (serrated polyp, hyperplastic polyp) and five (CRC stage I, CRC stage II) into smaller subsets. We have plotted the colonoscopy results after test round three in Figure 1.

3.3 Ordered categories

The ordinal models in this research allow for combinations and division of categories as reported in Figure 1 as desired. We therefore compare our ordinal models for multiple category instances

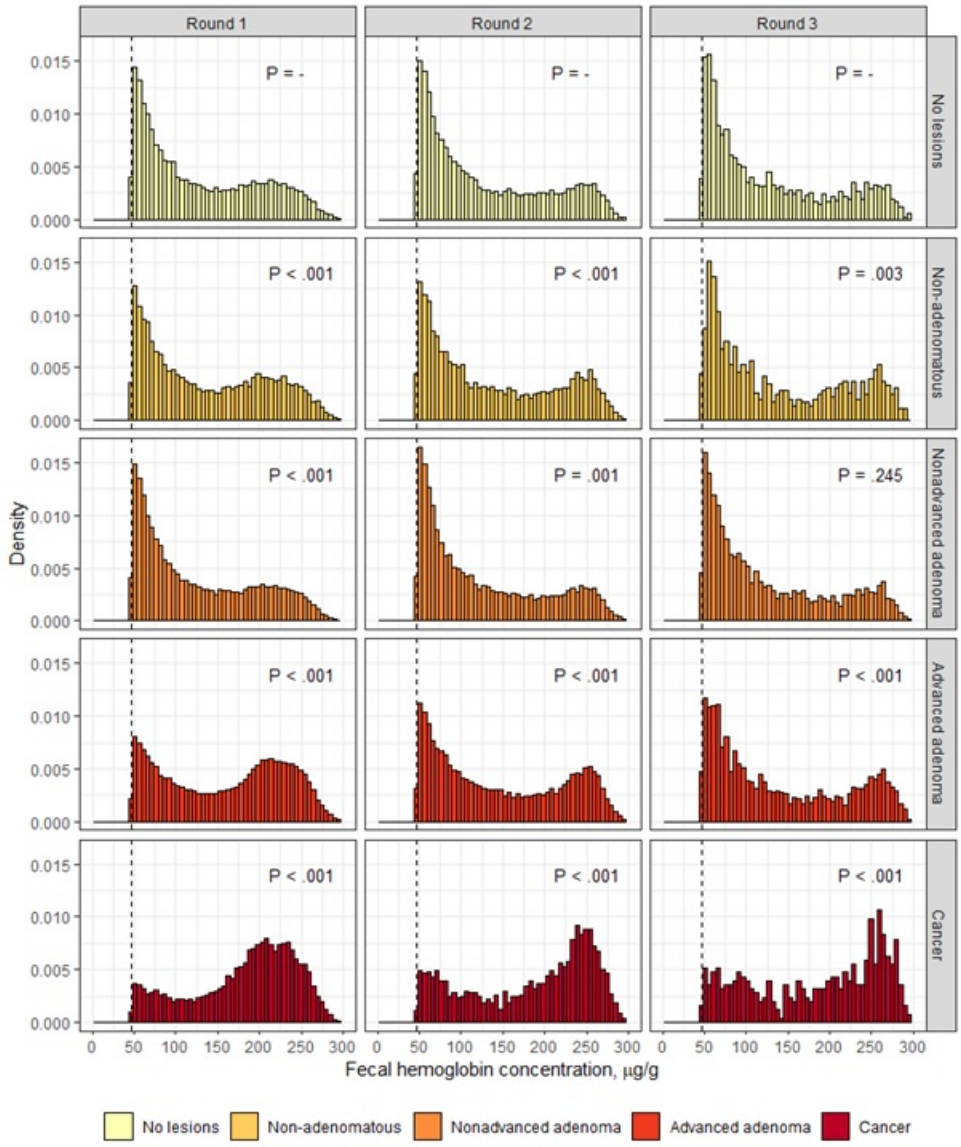
ranging from three to five categories. However, in this report we present the results of our models using four categories:

1. No positive FIT-test in round three, hence no colonoscopy;
2. Positive FIT-test, no advanced lesions detected (including NA and NAA);
3. Positive FIT-test, AA detected;
4. Positive FIT-test, CRC detected.

Our choice of four categories makes our results easy and valuable to interpret. Clinicians consider AA and CRC detection as relevant findings in terms of malignity. This means that our second proposed category contains those participants with a false-positive FIT-result. The split between AA and CRC is useful to understand whether we are able to predict the progression of an advanced lesion. To obtain this insight, we will have to investigate the effect of measured hemoglobin (Hb) concentrations on the probability of changing from category three to four.

We base our choice of four categories not only on reasons for interpretability, but also on our knowledge on the distribution of the Hb concentration of positive FITs per outcome category. We have included plots of these distributions per round and per outcome category in Figure 2 (Meester, 2021).

By inspection of the distribution of blood values per detected lesion, we observe, in line with our expectation, that the distribution changes the most for more severe lesions. This is confirmed by performing a Kolmogorov-Smirnov test (Smirnov et al., 1948). The distribution of FIT-values per outcome and the corresponding p-values of the Kolmogorov-Smirnov test are also included in Figure 2. Based on visual inspection, we see a large change in distribution between the AA and CRC category. This constitutes another reason to keep AA and CRC as separate categories in our proposed frameworks. Because the distribution of NA and NAA seem relatively similar, we choose to combine those categories in the second category. This combination also solves potential problems due to the small size of the NA category.



Notes: The p-value represents the statistical difference with the category 'No Lesions' according to a Kolmogorov-Smirnov test.

Figure 2: Hemoglobin concentrations in the first, second, and third round among persons with a positive FIT (hence undergo colonoscopy) per detected lesion (Meester, 2021)

4 Methodology

In this research, we aim to improve the predictions of colonoscopy outcomes by including knowledge on the stages of the disease development. This means that our models take the multi-class, naturally ordered colonoscopy outcomes as a dependent variable. To avoid a loss of information, we model the colonoscopy outcomes as an ordered dependent variable. In this section, we will first introduce the previously applied binary outcome model. We then describe the ordered logit model and generalise this model to constitute a new ordered outcome model that - to the best of our knowledge - has not been implemented before. Finally, we will provide numeric statistics and graphs to evaluate the performance of our models.

4.1 Baseline model

We will compare the new models in this research to the binary model as previously applied by Meester (2021). The good performance of this model has shown that previously measured hemoglobin (Hb) concentrations as measured by the Fecal Immunochemical Test (FIT) contain predictive power for future colonoscopy outcomes. Similar to Meester (2021), we split the continuous FIT-values into six categorical variables. Though this increases the amount of parameters to be estimated, it allows for non-linearities, hence potentially improving the performances. The discretisation also increases the interpretability of the coefficients to clinicians and practitioners. Additionally, we will evaluate whether including polynomial and interaction terms of the explanatory variables improves the performance of the model.

4.2 Ordered logit

We will model the colorectal cancer (CRC) screening outcome. Let y_i be this observed colonoscopy outcome variable for individual i and let $\mathbf{y}$ be the N x 1 vector containing the outcome categories of all N observations. Ordered response models are usually motivated by an underlying continuous but latent process (Boes and Winkelmann, 2006). In this research, one could think of the growth of the lesion - driven by socio-demographic, genetic and behavioural factors - as this unobserved latent process y_i^* for individual i. Let $\mathbf{x}_i$ be a K x 1 vector of explanatory variables for individual i. In this research, we will use age, gender, and two previously observed negative FIT-values as the explanatory variables. We model the latent process y_i^* as a function of the K explanatory variables:

$$y_i^* = \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i, \tag{1}$$

where $\boldsymbol{\beta}$ is a K x 1 vector of parameters and ϵ_i the random error term belonging to observation i and β_0 an intercept value. For all observations N, we obtain in matrix notation:

$$\mathbf{y}^* = \boldsymbol{\beta}_0 + \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2}$$

Here, $\boldsymbol{\beta}_0$ is the N x 1 vector of β_0, $\mathbf{X}$ is the K x N matrix with vectors $\mathbf{x}_i$ and $\boldsymbol{\epsilon}$ the N x 1 vector containing the error terms ϵ_i .

We observe and finally wish to model the outcome variable y_i, which can be one out of J categories. We model y_i as a function of the latent variable y_i^*, depending on unobserved boundary parameters $\alpha_0, ..., \alpha_J$:

$$\begin{cases} y_i = 1 & \text{if } \alpha_0 < y_i^* \leq \alpha_1; \\ y_i = j & \text{if } \alpha_{j-1} < y_i^* \leq \alpha_j; \\ y_i = J & \text{if } \alpha_{J-1} < y_i^* \leq \alpha_J. \end{cases} \tag{3}$$

We can then model $\mathrm{P}[y_i = j|\mathbf{x}_i]$, the probability of y_i being in outcome category j as follows:

$$\begin{aligned} \mathrm{P}[y_i = j|\mathbf{x}_i] &= \mathrm{P}[\alpha_{j-1} < y_i^* \leq \alpha_j] \\ &= \mathrm{P}[\alpha_{j-1} < \beta_0 + \mathbf{x}_i'\boldsymbol{\beta} + \epsilon_i \leq \alpha_j] \\ &= \mathrm{P}[\alpha_{j-1} - (\beta_0 + \mathbf{x}_i'\boldsymbol{\beta}) < \epsilon_i \leq \alpha_j - (\beta_0 + \mathbf{x}_i'\boldsymbol{\beta})] \\ &= F\left(\alpha_j - (\beta_0 + \mathbf{x}_i'\boldsymbol{\beta})\right) - F\left(\alpha_{j-1} - (\beta_0 + \mathbf{x}_i'\boldsymbol{\beta})\right) \text{ for } j = 2, .., J-1, \text{ and} \\ \mathrm{P}[y_i = 1|\mathbf{x}_i] &= \mathrm{P}[y_i^* \leq \alpha_1] = F\left(\alpha_1 - (\beta_0 + \mathbf{x}_i'\boldsymbol{\beta})\right) \text{ and} \\ \mathrm{P}[y_i = J|\mathbf{x}_i] &= \mathrm{P}[\alpha_{J-1} < y_i^*] = 1 - F\left(\alpha_{J-1} - (\beta_0 + \mathbf{x}_i'\boldsymbol{\beta})\right), \end{aligned} \tag{4}$$

where $F(\cdot)$ denotes the cumulative distribution of ϵ, where we set $\alpha_0 = -\infty$ and $\alpha_J = +\infty$ such that $F(-\infty) = 0$ and $F(+\infty) = 1$. For identification purposes, we impose $\beta_0 = 0$.

In the ordered logit model, we then use the logistic function as the cumulative density function (cdf) $F(.)$ for ϵ_i:

$$F(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}) = \frac{\exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta})}{1 + \exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta})}. \tag{5}$$

By taking the derivative using the chain rule, we obtain the partial effect with respect to $x_{k,i}$:

$$\frac{\mathrm{P}[y_i \leq j|\mathbf{x}_i]}{\delta x_{k,i}} = \beta_k \left(f(\alpha_{j-1} - \mathbf{x}_i'\boldsymbol{\beta}) - f(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta})\right). \tag{6}$$

Here $f(\cdot)$ represents the probability density function (pdf) corresponding to the cdf, in case of the logistic function, it holds that:

$$f(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}) = \boldsymbol{\beta} F(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}) \left(1 - F(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta})\right). \tag{7}$$

We can compute odds ratios as follows:

$$\frac{\mathrm{P}[y_i \leq j|\mathbf{x}_i]}{\mathrm{P}[y_i > j|\mathbf{x}_i]} = \exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}). \tag{8}$$

We can estimate the standard ordered model by maximum likelihood optimisation. The likelihood function that needs to be maximised equals:

$$\ln L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N}\sum_{j=1}^{J} I[y_i = j] \ln\left(F(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}) - F(\alpha_{j-1} - \mathbf{x}_i'\boldsymbol{\beta})\right), \tag{9}$$

where parameter vector $\boldsymbol{\theta}$ summarises the vector of boundary parameters $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{J-1})'$ and the vector of regression coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)'$. $I[y_i = j]$ is a binary indicator function that equals one when individual i belongs to outcome category j and zero otherwise. This means we estimate $(J-1)+K$ parameters in this model.

Note that the estimated parameter $\boldsymbol{\beta}$ has the same value for all categories j. In other words, the relationship between the explanatory variables $\mathbf{x}_i$ and the odds of a response being in the next higher order category $j+1$, is the same regardless of which categories are to be compared. This feature is called the parallel assumption (Grilli and Rampichini, 2014). As we have not been able to observe a gradual shift in blood values per detected lesion for positive FITs in Figure 2, we have reason to believe that the parallel assumption might neither hold for negative FITs. We therefore relax this assumption and generalise the ordered logit model into the continuation ratio model.

4.3 Continuation ratio

To allow the parameters to vary per shift in outcome category, we implement the continuation ratio model as initially proposed by Feinberg (1980). We can construct the continuation ratio model by altering the odds ratio of the ordered logit model in (8). In the ordered logit model, the numerator contains the probability that individual i belongs to the first j categories. We replace this probability by the conditional probability that individual i belongs to category j, given that i belongs to category j or a category higher than j. The odds ratio then becomes:

$$\frac{\mathrm{P}[y_i = j|y_i \geq j, \mathbf{x}_i]}{\mathrm{P}[y_i > j|\mathbf{x}_i]} = \exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_j). \tag{10}$$

Similar to the standard ordered model, we set $\beta_{0,j} = 0$ for all categories j for identification purposes and set $\alpha_0 = -\infty$ and $\alpha_J = +\infty$ such that $F(-\infty) = 0$ and $F(+\infty) = 1$. This implies that we use a binary logit model to model the choice between $y_i = j$ and $y_i > j$ given that $y_i \geq j$:

$$\mathrm{P}[y_i = j|y_i \geq j, \mathbf{x}_i] = \frac{\exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_j)}{1 + \exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_j)}. \tag{11}$$

Note that the parameter vector $\boldsymbol{\beta}_j$ now contains a subscript j. This means that the estimated parameters now depend on the outcome category j and that we have relaxed the parallel assumption.

We can use the conditional probabilities to compute the unconditional outcome probabilities:

$$\begin{aligned}
\mathrm{P}[y_i = j|\mathbf{x}_i] &= \mathrm{P}[y_i = j|y_i \geq j, \mathbf{x}_i] \prod_{l=1}^{j-1} (1 - \mathrm{P}[y_i = l|y_i \geq l, \mathbf{x}_i]) \\
&= \frac{\exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_{1,j})}{1 + exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_{1,j})} \prod_{l=1}^{j-1} \frac{1}{1 + \exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_{1,l})} \text{ for } j = 2, .., J-1, \text{ and} \qquad (12) \\
\mathrm{P}[y_i = J|\mathbf{x}_i] &= 1 - \sum_{l=1}^{J-1} \mathrm{P}[y_i = l|\mathbf{x}_i].
\end{aligned}$$

We can use these unconditional probabilities to construct the log-likelihood function:

$$\begin{aligned}
\ln L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \sum_{j=1}^{J-1} I[y_i = j] \ln \left(\frac{\exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_{1,j})}{1 + \exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_{1,j})} \prod_{l=1}^{j-1} \frac{1}{1 + \exp(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_{1,l})} \right) \\
+ I[y_i = J] \ln \left(1 - \sum_{l=1}^{J-1} \mathrm{P}[y_i = l|\mathbf{x}_i] \right). \qquad (13)
\end{aligned}$$

Here, parameter vector $\boldsymbol{\theta}$ summarises $(J-1)$ x 1 parameter vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{J-1})'$ and $(J-1)*K$ x 1 parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_{J-1}')' = (\beta_{1,1}, ..., \beta_{1,K}, ..., \beta_{J-1,1}, ..., \beta_{J-1,K})'$. With this log-likelihood function, we can estimate the model parameters through standard maximum likelihood optimisation.

4.4 Finite mixture ordered logit

The previously discussed models ignore the unobserved heterogeneity in the dataset. A finite mixture model creates a finite number of clusters within the data and then computes optimal logistic regression parameters per cluster (Boes and Winkelmann, 2006), constituting the Finite Mixture Ordered Logit (FMOL) model in our context. Let C be the number of clusters in the dataset, let $\alpha_{j,c}$ be the boundary parameter j belonging to cluster c, and let $\boldsymbol{\beta}_c$ be the K x 1 parameter vector belonging to cluster c. Similar to (4), we then obtain:

$$\begin{aligned}
\mathrm{P}[y_i = j|\mathbf{x}_i] &= \sum_{c=1}^{C} \pi_c \left(F(\alpha_{j,c} - \mathbf{x}_i'\boldsymbol{\beta}_c)) - F(\alpha_{j-1,c} - \mathbf{x}_i'\boldsymbol{\beta}_c) \right) \text{ for } j = 2, .., J-1, \text{ and} \\
\mathrm{P}[y_i = 1|\mathbf{x}_i] &= \sum_{c=1}^{C} \pi_c \mathrm{P}[y_i^* \leq \alpha_{1,c}] = \sum_{c=1}^{C} \pi_c F(\alpha_{1,c} - \mathbf{x}_i'\boldsymbol{\beta}_c), \text{ and} \qquad (14) \\
\mathrm{P}[y_i = J|\mathbf{x}_i] &= \sum_{c=1}^{C} \pi_c \mathrm{P}[y_i^* > \alpha_{J-1,c}] = 1 - \sum_{c=1}^{C} \pi_c F(\alpha_{J-1,c} - \mathbf{x}_i'\boldsymbol{\beta}_c),
\end{aligned}$$

where π_c is the population probability of belonging to cluster c and can be interpreted as the relative cluster sizes. Similar to the standard ordered model, for all clusters c we set $\beta_{0,c} = 0$ for identification purposes and set $\alpha_{0,c} = -\infty$ and $\alpha_{J,c} = +\infty$ such that $F(-\infty) = 0$ and $F(+\infty) = 1$. Note that the parameters within the logistic function now depend on cluster c.

Similar to (6), the partial effect with respect to $x_{k,i}$ can be obtained by taking the derivative:

$$\frac{\mathrm{P}[y_i \leq j|\mathbf{x}_i]}{\delta x_{k,i}} = \sum_{c=1}^{C} \pi_c \beta_{c,k} \left(f(\alpha_{j-1} - \mathbf{x}_i'\boldsymbol{\beta}_c) - f(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_c) \right), \tag{15}$$

in which $f(\cdot)$ again represents the pdf as specified in (7).

We define the log-likelihood of the model:

$$\ln L(\boldsymbol{\theta}, \boldsymbol{\pi} \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N} \sum_{j=1}^{J} I[y_i = j] \ln \left\{ \sum_{c=1}^{C} \pi_c F\left((\alpha_{j,c} - \mathbf{x}_i'\boldsymbol{\beta}_c) - F(\alpha_{j-1,c} - \mathbf{x}_i'\boldsymbol{\beta}_c) \right) \right\}, \tag{16}$$

where $\boldsymbol{\pi}$ is a C x 1 vector containing the cluster probabilities π_c and parameter vector $\boldsymbol{\theta}$ summarises both $(J-1) * C$ x 1 parameter vector $\boldsymbol{\alpha} = (\alpha_{1,1}, ..., \alpha_{J-1,1} ..., \alpha_{1,C}, ..., \alpha_{J-1,C})'$ and $C * K$ x 1 parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_C')'$ parameters. This means that we have to estimate $C*(J+K-1)$ parameters in $\boldsymbol{\theta}$ and C parameters in $\boldsymbol{\pi}$. The total amount of parameters to estimate therefore equals $C * (J + K)$.

Note that in (16), we have a summation within the logarithmic term. This makes the optimisation of the log-likelihood non-trivial. Rather than estimating the model through standard numerical optimisation, we therefore estimate the model using an Expectation-Maximisation (EM) algorithm as proposed by Dempster et al. (1977). For an EM-algorithm, we use the complete log-likelihood function rather the standard log-likelihood function. In the complete log-likelihood function, we assume that the unobserved individual class membership is known. This results in the following function where $I[c_i = c]$ is a binary indicator function for individual class membership and N x 1 vector $\mathbf{c}$ contains these individual class memberships:

$$\ln L_{com}(\boldsymbol{\theta}, \boldsymbol{\pi} \mid \mathbf{y}, \mathbf{X}, \mathbf{c}) = \\ \sum_{i=1}^{N} \sum_{j=1}^{J} I[y_i = j] \sum_{c=1}^{C} I[c_i = c] \left\{ \ln \pi_c + \ln \left(F(\alpha_{j,c} - \mathbf{x}_i'\boldsymbol{\beta}_c) - F(\alpha_{j-1,c} - \mathbf{x}_i'\boldsymbol{\beta}_c) \right) \right\}. \tag{17}$$

As we however cannot observe class membership, we cannot maximise this function directly. The EM-algorithm consists out of two steps in which we iteratively compute class membership in the expectation-step (E-step), followed by parameter estimation in the maximisation-step (M-step).

In the E-step, we take the expectation of (17) with respect to the class membership given the observed data and the current fit of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. By applying Bayes theorem, Boes and Winkelmann (2006) show that this expectation yields the posterior probability $p_{i,c}$ that individual i belongs to class c:

$$p_{i,c}\left(\mathbf{y_i}, \mathbf{x}_i; \boldsymbol{\theta}^{(q)}, \boldsymbol{\pi}^{(q)}\right) = \frac{\pi_c^{(q)}\left(F(\alpha_{j,c}^{(q)} - \mathbf{x}_i'\boldsymbol{\beta}_c^{(q)}) - F(\alpha_{j-1,c}^{(q)} - \mathbf{x}_i'\boldsymbol{\beta}_c^{(q)})\right)}{\sum_{s=1}^{C}\pi_s^{(q)}\left(F(\alpha_{j,s}^{(q)} - \mathbf{x}_i'\boldsymbol{\beta}_s^{(q)}) - F(\alpha_{j-1,s}^{(q)} - \mathbf{x}_i'\boldsymbol{\beta}_s^{(q)})\right)}, \tag{18}$$

where the superscript (q) indicates the q-th iteration of the algorithm.

In the M-step, we subsequently maximise (17) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$, where we replace the unobserved $I[c_i = c]$ by the computed $p_{i,c}$ from the E-step. Estimation of $\boldsymbol{\pi}$ can simply be done by taking the averages $\frac{1}{N}\sum_{i=1}^{N} p_{i,c}$. Maximisation of $\boldsymbol{\theta}$ can be done for each class separately. In other words, we estimate C simple logit models while weighing the data (Boes and Winkelmann, 2006).

The EM-algorithm then iterates between the E-step and the M-step. After each iteration of both an E-step and an M-step, we compute the log-likelihood. When the difference between two consecutive values of the log-likelihood is below a stopping condition, we stop iterating.

Due to the iterative character, the algorithm is not guaranteed to attain a global optimum. We will use multiple starting values to avoid this problem. Dempster et al. (1977) show that when multiple starting values are used, the algorithm typically performs well.

4.5 Finite mixture generalised ordered logit

We have now generalised the standard ordered logit model in two ways. We have first generalised the model with respect to the outcome categories in the continuation ratio model. Then, we have generalised the ordered logit model with respect to clusters in our finite mixture model. In this section, we will propose a method in which both generalisations are combined. The parameter estimates are then able to vary per outcome category j as well as per cluster c. This new model is - to the best of our knowledge - new in the econometric literature.

Starting again from the ordered model, we will first describe a different way to generalise the ordered logit model with respect to the outcome categories j. Rather than the continuation ratio model, we will use the generalised ordered logit (gologit) model as proposed by Peterson and Harrell Jr (1990) in our finite mixture approach. We choose the gologit model over the continuation ratio model for its generality and because it allows us to impose the required parameter restrictions without changing the log-likelihood function.

4.5.1 Generalised ordered logit

Similarly to the continuation ratio model, we now relax the parallel assumption, such that the value of $\boldsymbol{\beta}$ could differ per outcome category j. This means that the $\boldsymbol{\beta}$ parameters from (4) will again get a subscript j to indicate the category of interest. The probability that individual i belongs to category j now equals:

$$\begin{aligned}
\mathrm{P}[y_i = j|\mathbf{x}_i] &= \mathrm{P}[\alpha_{j-1} \leq y_i^* \leq \alpha_j|\mathbf{x}_i] \\
&= F(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_j) - F(\alpha_{j-1} - \mathbf{x}_i'\boldsymbol{\beta}_{j-1}) \text{ for } j = 2, .., J-1 \text{ and} \\
\mathrm{P}[y_i = 1|\mathbf{x}_i] &= \mathrm{P}[y_i^* \leq \alpha_1|\mathbf{x}_i] = F(\alpha_1 - \mathbf{x}_i'\boldsymbol{\beta}_1) \text{ and} \\
\mathrm{P}[y_i = J|\mathbf{x}_i] &= \mathrm{P}[\alpha_{J-1} < y_i^*|\mathbf{x}_i] = 1 - F(\alpha_{j-1} - \mathbf{x}_i'\boldsymbol{\beta}_{J-1}).
\end{aligned} \tag{19}$$

Similar to the continuation ratio model, we set $\beta_{0,j} = 0$ for all categories j for identification purposes and set $\alpha_0 = -\infty$ and $\alpha_J = +\infty$ such that $F(-\infty) = 0$ and $F(+\infty) = 1$. Again, we estimate the model by maximum likelihood. The likelihood function in this case equals:

$$\ln L(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N}\sum_{j=1}^{J} I[y_i = j] \ln(F(\alpha_j - \mathbf{x}_i'\boldsymbol{\beta}_j) - F(\alpha_{j-1} - \mathbf{x}_i'\boldsymbol{\beta}_{j-1})), \tag{20}$$

where parameter vector $\boldsymbol{\theta}$ summarises $(J-1)$ x 1 parameter vector $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{J-1})'$ and $(J-1)*K$ x 1 parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_{J-1}')' = (\beta_{1,1}, ..., \beta_{1,K}, ..., \beta_{J-1,1}, ..., \beta_{J-1,K})'$. In this model, we therefore have to estimate $(J-1)*(K+1)$ parameters.

Though it is likely that the observations in our dataset are subject to unobserved heterogeneity, the gologit model estimates the same parameters $\boldsymbol{\beta}_j$ for all individuals i. For that reason, we will now introduce our new model that allows for cluster-specific parameters within the gologit model.

4.5.2 Mixture approach in gologit

The Finite Mixture Generalised Ordered Logit (FIMGOL) model combines the gologit approach from Peterson and Harrell Jr (1990) and a finite mixture approach, as already presented in the finite mixture ordered logit model. To the best of our knowledge, a model as such has not been described nor implemented in the academic literature hitherto.

In order to combine the gologit model and the FMOL model, the $\boldsymbol{\beta}_j$ parameter in the log-likelihood from (20) gets an extra subscript c, as we allow the parameter value to be both outcome- and cluster-specific. This means that the log-likelihood now equals:

$$\ln L(\boldsymbol{\theta}, \boldsymbol{\pi} \mid \mathbf{y}, \mathbf{X}) = \sum_{i=1}^{N}\sum_{j=1}^{J} I[y_i = j] \ln\left\{\sum_{c=1}^{C} \pi_c \left(F(\alpha_{j,c} - \mathbf{x}_i'\boldsymbol{\beta}_{j,c}) - F(\alpha_{j-1} - \mathbf{x}_i'\boldsymbol{\beta}_{j-1,c})\right)\right\}, \tag{21}$$

where $\boldsymbol{\theta}$ summarises the $C*(J-1)$ x 1 boundary parameter vector $\boldsymbol{\alpha} = (\alpha_{1,1}, ..., \alpha_{J-1,1} ..., \alpha_{1,C}, ..., \alpha_{J-1,C})'$ and the $C*(J-1)*K$ x 1 parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_C')' = (\boldsymbol{\beta}_{1,1}', ..., \boldsymbol{\beta}_{1,J-1}', ..., \boldsymbol{\beta}_{C,1}', ..., \boldsymbol{\beta}_{C,J-1}')'$. This means that parameter vector $\boldsymbol{\theta}$ now contains $C * (J-1) * (K+1)$ parameters. Similar to the finite ordered logit model, we also need to estimate the C parameters in $\boldsymbol{\pi}$, making a total of $C * (J * (K+1) - K)$ parameters.

Similar to estimating the finite mixture model, estimation can be performed using the EM-algorithm from Dempster et al. (1977). We define the complete log-likelihood by assuming observed clusters and obtain:

$$\ln L_{com}(\boldsymbol{\theta}, \boldsymbol{\pi} \mid \mathbf{y}, \mathbf{X}, \mathbf{c}) = \sum_{i=1}^{N} \sum_{j=1}^{J} I[y_i = j] \sum_{c=1}^{C} I[c_i = c] \{\ln \pi_c + \\ + \ln \left(F(\alpha_{j,c} - \mathbf{x}_i' \boldsymbol{\beta}_{j,c}) - F(\alpha_{j-1,c} - \mathbf{x}_i' \boldsymbol{\beta}_{j-1,c}) \right). \quad (22)$$

In the E-step of the algorithm, we take the expectation with respect to the class membership and in the M-step, we maximise the complete log-likelihood from (22) with respect to the logistic parameters for all j categories and c classes. We again apply Bayes theorem to obtain the following posterior class membership probabilities $p_{i,c}$

$$p_{i,c}\left(\mathbf{y_i}, \mathbf{x}_i; \boldsymbol{\theta}^{(q)}, \boldsymbol{\pi}^{(q)}\right) = \frac{\pi_c^{(q)} \left(F(\alpha_{j,c}^{(q)} - \mathbf{x}_i' \boldsymbol{\beta}_{j,c}^{(q)}) - F(\alpha_{j-1,c}^{(q)} - \mathbf{x}_i' \boldsymbol{\beta}_{j-1,c}^{(q)}) \right)}{\sum_{s=1}^{C} \pi_s^{(q)} \left(F(\alpha_{j,s}^{(q)} - \mathbf{x}_i' \boldsymbol{\beta}_{j,s}^{(q)}) - F(\alpha_{j-1,s}^{(q)} - \mathbf{x}_i' \boldsymbol{\beta}_{j-1,s}^{(q)}) \right)}, \quad (23)$$

where the superscript (q) again denotes the q-th iteration of the algorithm. The iterative EM-algorithm computes the log-likelihood as specified in (21) after each EM-iteration and reaches convergence when the difference between two consecutive values is negligible. Again, we use multiple starting values to reach a global optimum.

4.5.3 Restrictions

In both the log-likelihood function of the ordered logit model as specified in (13) as well as in the regular and complete log-likelihood functions of the finite mixture ordered logit model in (16) and (17) respectively, we obtain strictly positive terms in the logarithmic terms of the equations by construction. Because parameters in the FIMGOL model can vary per category outcome j and per cluster c, we might have:

$$F(\alpha_{j,c} - \mathbf{x}_i' \boldsymbol{\beta}_{j,c}) \leq F(\alpha_{j-1,c} - \mathbf{x}_i' \boldsymbol{\beta}_{j-1,c}), \quad (24)$$

which could lead to negative probabilities, hence negative terms in the logarithmic term in the (complete) log-likelihood. To properly define our model, we impose the following restriction:

$$\boldsymbol{\beta}_{j,c} \geq \boldsymbol{\beta}_{j+1,c} \forall (j,c). \tag{25}$$

We will now prove that this restriction is sufficient in our model:

$$\begin{aligned}
\boldsymbol{\beta}_{j,c} \geq \boldsymbol{\beta}_{j+1,c} \forall (j,c) &\iff \boldsymbol{\beta}_{j,c} \leq \boldsymbol{\beta}_{j-1,c} \forall (j,c) && (26)\\
&\implies -\mathbf{x}_i' \boldsymbol{\beta}_{j,c} \geq -\mathbf{x}_i' \boldsymbol{\beta}_{j-1,c} \forall (j,c) && (27)\\
&\implies \alpha_{j,c} - \mathbf{x}_i' \boldsymbol{\beta}_{j,c} > \alpha_{j-1,c} - \mathbf{x}_i' \boldsymbol{\beta}_{j-1,c} \forall (j,c) && (28)\\
&\iff F(\alpha_{j,c} - \mathbf{x}_i' \boldsymbol{\beta}_{j,c}) > F(\alpha_{j-1,c} - \mathbf{x}_i' \boldsymbol{\beta}_{j-1,c}) \forall (j,c) \; \square && (29)
\end{aligned}$$

In this proof, (27) follows from (26) and the fact that the explanatory variables in our application are positive. Notice that we obtain a strict inequality in (28), because by construction of the gologit model, $\alpha_{j,c} > \alpha_{j-1,c} \; \forall (j,c)$. We obtain equivalence in (29) as the logistic function $F(\cdot)$ is monotonically increasing. As our proposed restriction leads to (29), we can never have negative terms in the logarithmic term of the (complete) log-likelihood.

Intuitively, our restriction implies that we impose that the parameter value always decreases for higher category shifts. As we will see in the results of the continuation ratio model, our data are not likely to behave in a different pattern. For other applications of our model, one could reverse the order of categories in the ordinal set-up and assess the model in both directions.

Within each maximisation step of the EM-algorithm, we first compute the parameters for the highest category shift $\boldsymbol{\beta}_{J-1,c}$ for each cluster c. We subsequently optimise values which we add to the parameter for lower category shifts $j = 1, ..., J-2$ after taking the exponential transformation. By taking the exponential transformation of the optimised values (that could take any value in the real number space), we adhere to the restriction as specified in (25).

4.6 Parameter estimation

We can estimate the parameters of the ordered logit and the continuation ratio model by maximising the log-likelihood function. The estimation of the finite mixture models is performed by our own implementation of an EM-algorithm as proposed by Dempster et al. (1977).

4.6.1 Expectation-Maximisation algorithm

The EM-algorithm iterates between taking the expectation with respect to the class membership given the current fit of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ (E-step) and maximising the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ (M-step). We

have summarised our two implementations of the EM-algorithm in Algorithm 1. For the numerical optimisation in the M-step of the algorithm, the Nelder-Mead optimisation routine as installed in the `Optim` package of R.4.0.3 will be used (Nelder and Mead, 1965).

Algorithm 1 Summary of the Expectation Maximisation algorithm

1: Initialise $p_{i,c}$, $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ and set $L \leftarrow 100$
2: Compute the log-likelihood L_{new} using (16) or (21)
3: **while** $L > 0.0001$ **do**
4: $L_{old} \leftarrow L_{new}$
5: E-step:
- Compute individual cluster probabilities $p_{i,c}$ using (18) or (23)
6: M-step:
- Compute π_c by taking $\frac{1}{N}\sum_{i=1}^{N} p_{i,c}$ for all c
- Optimise $\boldsymbol{\theta}$ with respect to the complete log-likelihood (17) or (22) using numerical optimisation
7: Update current log likelihood L_{new} using (16) or (21)
8: $L \leftarrow L_{new} - L_{old}$
9: **end while**
10: Return optimal values $p_{i,c}$, $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$

Notes: This set-up of the EM-algorithm can be used for both the finite mixture ordered logit as well as the finite mixture generalised ordered logit model. The first mentioned equations in lines 2, 5, 6, and 7 of the algorithm belong to the former model, the second belong to the latter.

4.6.2 Standard errors

To estimate the standard errors, one can use the property that the variance of maximum likelihood estimation can be estimated by inverting the Hessian matrix of the log-likelihood (Cameron, 1988). In our finite mixture models, the numerical computation of the Hessian matrix is impeded by the standard probability assumptions of non-negativity of the cluster probabilities $\boldsymbol{\pi}$ and the property that $\sum_{c=1}^{C} \pi_c = 1$. We therefore reparameterise the vector $\boldsymbol{\pi}$:

$$\gamma_c = \log(\pi_c) - \log(\pi_C) \text{ for } c = 1, ..., C-1. \tag{30}$$

We can obtain our original parameters by taking the inverse of our reparametrisation:

$$\pi_c = \frac{\exp(\gamma_c)}{1 + \sum_{l=1}^{C-1} \exp(\gamma_l)} \text{ for } c = 1, ..., C-1, \qquad \pi_C = \frac{1}{1 + \sum_{l=1}^{C-1} \exp(\gamma_l)}$$

Using this reparametrisation, we can numerically approximate the Hessian of the log-likelihood by the Richardson method as installed in the `NumDeriv` package of R.4.0.3 (Gilbert et al., 2006). We can obtain the standard errors by taking the square root of the diagonals of the inverted Hessian matrix.

4.7 Random forest

Though our research focuses on advanced regression techniques in the context of ordinal outcomes, we wish to compare and contrast our methods to a machine learning method that is known for its high performance in complex prediction problems. For that purpose, we construct a random forest for our application.

A random forest as initially proposed by Ho (1995) establishes predictions based on the predictions of multiple decision trees. Decision trees create predictions using a tree-based structure, with the observations in branches and the predictions in the leaves (Quinlan, 1986). A random forest repeatedly selects a random sample with replacement of the training set and fits a tree for each set. After training, predictions can be made by taking the average of the predictions from all individual regression trees. By their construction, random forests are highly flexible models, making them particularly suitable for complex predictive tasks.

Random forests are known for their notable predictive performance and often outperform regression techniques in predictive tasks (Prinzie and Van den Poel, 2008). On the other hand, one should realise that by using a random forest, we lose a large share of interpretability. In clinical applications like ours, interpretability can be essential to practical use.

4.8 Model evaluation

In this research, we wish to improve the performance of the predictive models for outcomes of the Dutch screening programme for CRC. We will therefore evaluate, compare, and test our proposed models based on multiple performance measures and graphics. We will train our models with 70% of the available observations and compute the performance measures on the other 30% in the test set.

4.8.1 Tests for regression coefficients

To gain insight in the behaviour of the disease development and its interaction with previously measured FIT-values, we have constructed our models such that we can estimate the effect of the explanatory variables per outcome category shift. To understand the pathway of precursors to CRC, it is of clinical interest to know how Hb concentrations indicate a particular lesion. For that reason, we will test whether the parameters among category shifts differ significantly. We therefore wish to test the null hypothesis $H_0 : \beta_{j,k} = \beta_{m,k}$ where $\beta_{j,k}$ represents the parameter belonging to explanatory variable x_k for a category shift from j to $j+1$. As we compute parameter estimates using a maximum likelihood approach, we know that the estimators are asymptotically

normally distributed under certain standard assumptions (Wilks, 1962). Under normality, we can test whether we find a significant difference between individual parameters using an asymptotic z-test:

$$\frac{\hat{\beta}_{j,k} - \hat{\beta}_{m,k}}{\sqrt{\hat{\text{Var}}(\hat{\beta}_{j,k} - \hat{\beta}_{m,k})}} \sim N(0,1), \tag{31}$$

where $\hat{\beta}_{j,k}$ equals the parameter estimate of category shift j to $j = 1$ for the k-th explanatory variable. We can compute the variance in the denominator $\hat{\text{Var}}(\hat{\beta}_{j,k} - \hat{\beta}_{m,k}) = \text{Var}(\hat{\beta}_{j,k}) + \text{Var}(\hat{\beta}_{m,k}) - 2 \cdot \text{Cov}(\hat{\beta}_{j,k}, \hat{\beta}_{m,k})$ from the estimated co-variance matrix of the estimated coefficients.

Next to testing the difference between parameters for each variable individually, we wish to jointly test whether there exists a significant difference between all Hb parameters of shift $j \to j+1$ and shift $m \to m+1$. We therefore test null hypothesis: $H_0 : \boldsymbol{\beta}_j = \boldsymbol{\beta}_m$, where $\boldsymbol{\beta}_j$ represents the parameter vector of all Hb parameters belonging to the category shift from j to $j+1$.

We will now describe a Wald-procedure to test this hypothesis. Let q be the number of parameters we wish to compare and let $\mathbf{R}$ be a q x $2q$ matrix:

$$\mathbf{R} = \begin{bmatrix} -1 & \dots & 0 & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & -1 & 0 & \dots & 1 \end{bmatrix} = \begin{bmatrix} -I_q & I_q \end{bmatrix}, \tag{32}$$

where I_q represents a q x q identity matrix. We then stack the parameters in a $2q$ x 1 parameter vector $\boldsymbol{\theta}' = (\boldsymbol{\beta}'_j, \boldsymbol{\beta}'_m)$ and let $\mathbf{r}$ be a q x 1 vector of zeros. Testing $H_0 : \boldsymbol{\beta}_j = \boldsymbol{\beta}_m$ is then equivalent to testing $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{r}$. Under normality of the maximum likelihood estimates, we then have:

$$(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r})'[\mathbf{R}\hat{\boldsymbol{\Sigma}}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r}) \sim \chi^2_{2q}, \tag{33}$$

where $\hat{\boldsymbol{\theta}}$ equals the estimated parameters and $\hat{\boldsymbol{\Sigma}}$ equals the estimated co-variance matrix belonging to $\hat{\boldsymbol{\theta}}$.

4.8.2 Discriminatory ability

Though our models are able to perform predictions over multiple categories, clinical research so far has focused on a binary outcome; the presence of Advanced Neoplasia (AN). In our set-up, clinicians consider the two most severe categories - Advanced Adenoma (AA) and CRC - as AN. To assess whether our methods are able to improve the predictive performance of the existing models, we will evaluate to what extent our ordinal models can predict AN.

To evaluate the binary predictive performance, we will compute the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) or also known as the concordance statistic. The ROC-curve is a two-dimensional depiction of classifier performance that demonstrates the diagnostic ability when the thresholds are varied (Fawcett, 2006). The AUC then measures the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006). We will use a bootstrap method to compute a confidence interval around the values for the AUC.

4.8.3 Ordinal classification performance

As the AUC is defined for binary classifiers only and multinomial generalisations are not applicable to ordinal outcome data, the metric fails in assessing to what extent our models are able to describe the ordinal process of the disease development. To compare the ordinal performance of our methods, we will use various measures of the Mean Absolute Error (MAE).

Cruz-Ramírez et al. (2014) define the MAE as the mean absolute deviation of a classifier in terms of the ordinal scale:

$$MAE_j = \frac{1}{n_j} \sum_{i=1}^{n_j} e_j(x_i), \tag{34}$$

where n_j equals the number of observations in class j and $e_j(x_i)$ equals the deviation of the classification on the ordinal scale. It equals the absolute difference between the true outcome y_i and the predicted outcome $\hat{y}_i$. As the metric is defined for an ordinal classifier, we have to translate the probabilistic output of our models to a classification. For that purpose, one can use a classification scheme of choice. The classification scheme that we have used can be found in Algorithm 2.

Using the probabilistic output of ordinal outcome models is common practice to multiple medical scoring rules Feldmann and Steudel (2000). Note that our classification scheme has a tree based structure. A classification scheme as such is also used by Wu et al. (2014) to construct simple decision rules for classifying whether someone is at high risk of CRC. The reasoning behind the set-up of our algorithm is that we first wish to verify whether an observation is in the most progressed categories before predicting less severe lesions, as these have the highest priority in CRC screening programmes (Mandelblatt et al., 1996).

Algorithm 2 Classification Scheme

Initialise thresholds parameters t_{CRC}, t_{AA} and t_{NA}
if $\hat{\mathrm{P}}[y_i = \mathrm{CRC}|\mathbf{x}_i] > t_{CRC}$ **then** $\hat{y}_i \leftarrow \mathrm{CRC}$
else
    if $\hat{\mathrm{P}}[y_i = \mathrm{CRC}|\mathbf{x}_i] + \hat{\mathrm{P}}[y_i = \mathrm{AA}|\mathbf{x}_i] > t_{AA}$ **then** $\hat{y}_i \leftarrow \mathrm{AA}$
    else
        if $\hat{\mathrm{P}}[y_i = \mathrm{CRC}|\mathbf{x}_i] + \hat{\mathrm{P}}[y_i = \mathrm{AA}|\mathbf{x}_i] + \hat{\mathrm{P}}[y_i = \mathrm{NA}|\mathbf{x}_i] > t_{NA}$ **then** $\hat{y}_i \leftarrow \mathrm{NA}$
        else
            $\hat{y}_i \leftarrow$ NO_COLO
        end if

Notes: This scheme translates the probabilistic output of our ordinal models to a categorical classification. Note that CRC refers to Colorectal Cancer, AA to Advanced Adenoma, NA to Non-Advanced adenomas and NO_COLO to a prediction with a negative FIT.

One can either initialise the threshold parameters of the algorithm based on rules of thumb or optimise them with respect to an objective function of choice. Similar to Cruz-Ramírez et al. (2014), we choose to optimise the threshold parameters with respect to the Average Mean Absolute Error (AMAE) and Maximum Mean Absolute Error (MMAE):

$$AMAE = \frac{1}{J}\sum_{j=1}^{J} MAE_j, \qquad MMAE = \max_j\{MAE_j; j = 1, ..., J\}.$$

Note that we take the average and maximum over the outcome categories J. By construction, the AMAE and MMAE give relatively more weight to observations in minority outcome categories, as the outcome categories are of equal importance in the objective, irrespective the number of observations per outcome category. This is an attractive property in our application, where the outcome categories with (severe) lesions are of key interest, yet in the minority. We will report the optimal values of the AMAE and MMAE of our models to compare the ordinal performance.

4.8.4 Brier scores

As our models output probabilistic predictions rather than strict classification, we wish to find an adaptation of the mean squared error to evaluate the accuracy. For that purpose, we use a proper scoring rule as defined by Brier et al. (1950). The Brier score measures the mean squared difference between the predicted probability outcome assigned to the possible outcomes and the

actual outcome:

$$Brier = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{J}\Big(\hat{\mathrm{P}}[y_i = j] - I[y_i = j]\Big)^2,$$

where $I[y_i = j]$ again represents an indicator function that equals one if $y_i = j$ and zero otherwise. Though the Brier score is able to evaluate the accuracy of probabilistic predictions, it does not take the ordinal nature of our outcome variable into account. That is, $\hat{\mathrm{P}}[y_i = k]$ is penalised for all $j \neq y_i$ irrespective the distance between outcome categories on the ordinal scale $|k - j|$.

To give more insight in the quality of the ordinal probabilistic predictions, we will also report a Ranked Brier Score as proposed by Weigel et al. (2007). The Ranked Brier score measures how well our probabilistic predictions match the observed outcomes:

$$RankedBrier = \sum_{k=1}^{C} Brier(k),$$

where $Brier(k)$ is the Brier score for the event that the outcome lies in the first k categories. By construction, the Ranked Brier score penalises high probability predictions further from the true outcome category harder than those that are closer to the true outcome.

4.8.5 Akaike Information Criterion

In both our finite mixture models, we will have to evaluate to what extent the burden of additional parameters outweighs the increase in log-likelihood. For that purpose, we compute the Akaike Information Criterion (AIC) for our models:

$$AIC = 2 * G - 2 * \log L(\hat{\boldsymbol{\theta}}|\mathbf{y}, \mathbf{X}), \tag{35}$$

where G equals the total number of parameters to be estimated in the model and $\hat{\boldsymbol{\theta}}$ is a G x 1 vector summarising all the estimated parameter values (Akaike, 1974).

4.8.6 Risk stratifying ability

In this research, we aim to improve the predictive models for colonoscopy outcomes with the final goal of personalising CRC screening. To personalise screening, our models need to be able to stratify observations, ideally segregating observations with a high risk from observations with a low risk on advanced lesions. We aim to visualise the risk stratifying ability of our models.

Based on the probability output of our models, we assign an individual risk score to our observations. This risk score is a (weighted) sum of the probabilities of having a lesion. In this research, we compute the risk score by adding the probabilities of the two most severe categories (AA and

CRC). Practitioners could alter the weights as desired or optimise them in simulation models such as the one proposed by Loeve et al. (1999). We then order the observations based on the risk score and divide them in percentiles, with the first percentile having the lowest mean assigned risk score and the 100-th percentile having the highest mean assigned risk score. For each percentile, we compute the actual relative risk by dividing the number of detected advanced lesions by the population average. If we plot the relative risks of the test data, we ideally see substantial differences between the first and last percentiles.

Next to plotting the relative risk, we also create a scatter plot with the ordered risk percentiles, with the assigned probability risk on the x-axis and the percentage of detected lesions within a percentile on the y-axis. Evaluating the distance between the observations and the 45°-line provides an indication of the goodness of fit of the models.

5 Results

In this research, we aim to improve the performance of the previously applied predictive models by including knowledge on the stages of disease development in the Dutch screening programme for colorectal cancer (CRC). In order to do so, we have included the stages of the disease development in innovative ordinal models. In this section, we will interpret and test the estimated parameters, report the performance of our models in terms of numeric performance statistics as well as visualise how these models are able to perform risk stratification. We will compare our models to the binary outcome model from Meester (2021). We have reconstructed this model and obtained similar results. We have reported the exponential transformation of the coefficients of the baseline model in Table 4 in the appendix.

5.1 Category specific effects

The first ordinal models that we have fitted are the ordered logit model and the continuation ratio model. We have reported the exponential transformation of the two models' coefficients in Table 1. Note directly that the continuation ratio model allows parameters to vary per shift in category, leading to a triple number of parameters to be estimated.

From Table 1, we learn that the estimated parameters of the ordered logit model closely resemble the parameters of the first category shift of the continuation ratio model. The parameters decrease in value for higher category shifts and lose significance in the last category shift. Note that the asterisks indicate whether the parameters significantly differ from zero.

Table 1: Model specifications of the ordered logit and continuation ratio model

	Ordered Logit	Continuation Ratio		
Variable		$1 \rightarrow 2$	$2 \rightarrow 3$	$3 \rightarrow 4$
Age	0.993	0.994	1.007	1.030
Male sex	1.243***	1.233***	1.040	0.902
First Hb concentration				
0	Ref.	Ref.	Ref.	Ref.
0.1–10	1.912***	1.907***	1.552***	0.715*
10–20	3.353***	3.287***	1.919***	0.748
20–30	4.355***	4.220***	2.411***	0.978
30–40	4.870***	4.806***	1.935***	0.449
40–47	6.641***	6.172***	4.563***	0.681*
Second Hb concentration				
0	Ref.	Ref.	Ref.	Ref.
0.1–10	3.173***	3.127***	1.916***	0.876
10–20	4.471***	4.406***	1.990***	0.831
20–30	5.799***	7.389***	2.886***	0.449**
30–40	5.750***	5.518***	2.514***	0.607
40–47	5.985***	5.743***	2.604***	0.527

Notes: *** p-value <0.001, ** p-value <0.01, *p-value <0.05

In addition to testing whether the individually parameters significantly differ from zero, we have tested whether the parameters of the hemoglobin (Hb) concentrations among category shifts differ from each other. That is, we have tested whether we have found a significant difference among the effect of Hb-values on shifts $1 \rightarrow 2$, $2 \rightarrow 3$, and $3 \rightarrow 4$.

We have first tested the individual differences in parameters using the asymptotic z-test as described in (31). This means that we test the null hypotheses $H_0 : \beta_{j,k} = \beta_{m,k}$ for $j = 1, m = 2$ and $j = 2, m = 3$ for the ten parameters belonging to the Hb-values. We found significant differences for all of the estimated parameters under a 0.001 significance level. The precise values of the z-statistics can be found in Table 5 in the appendix. We therefore reject H_0 for all the tested instances and conclude that the parameters among category shifts differ significantly.

Additionally, we have tested whether all parameters belonging to the Hb values jointly differ among category shifts using a Wald-procedure as specified in (33). We therefore test two null hypotheses. Let $\boldsymbol{\beta}_{j,Hb}$ be the 10 x 1 vector of the parameters belonging to the Hb-values for a category shift from j to $j + 1$ First, we test $H_0 : \boldsymbol{\beta}_{1,Hb} = \boldsymbol{\beta}_{2,Hb}$. The value of the test-statistic becomes 199.92, which is significant under a 0.001 significance level. We therefore reject H_0 and

have statistical evidence that $\boldsymbol{\beta}_{1,Hb} \neq \boldsymbol{\beta}_{2,Hb}$. Second, we test $H_0 : \boldsymbol{\beta}_{2,Hb} = \boldsymbol{\beta}_{3,Hb}$. The value of the test-statistic then becomes 187.69, which is also significant under a 0.001 significance level. We therefore reject H_0 and have statistical evidence that $\boldsymbol{\beta}_{2,Hb} \neq \boldsymbol{\beta}_{3,Hb}$.

The value of the parameters, their significance and the performed statistical tests show that our explanatory variables mostly have a significant effect on the first category shift. As the second category means that a participant is assigned a colonoscopy, but no advanced lesions are found, we can conclude that the previous values of the Fecal Immunochemical Test (FIT) contain the strongest explanatory power in predicting which participants will get a positive FIT later in life rather than predicting the final outcome of the colonoscopy. The tests also indicate that the variables contain little predictive power in differentiating between Advanced Adenomas (AA) and CRC.

5.2 Parameters of the finite mixture models

Subsequently, we have accounted for potential heterogeneity by allowing the parameters to vary per cluster in the finite mixture ordered logit (FMOL) model. The exponential transformation of the coefficients of the finite mixture ordered logit model can be found in Table 2.

As we now allow for clusters, the number of parameters to be estimated is multiplied by the number of clusters, with respect to the ordered logit model. From the model specifications, we see that the differences in parameters among clusters are subtle. This means that the data do not contain highly different clusters.

We have generalised the FMOL model by allowing the parameters to vary per category in the Finite Mixture Generalised Ordered Logit (FIMGOL) model. Note that this final generalisation leads to a rapid increase in the number of parameters to be estimated. For practical purposes, we therefore exclude the model specifications from this section, but include the model specification for two, three, and four clusters in Tables 6, 7, and 8 respectively in the appendix.

As mentioned before, we have imposed the restriction that the parameters have to decrease with higher category shifts. This for example means that the parameter corresponding to category shift $1 \rightarrow 2$ is larger than the parameter corresponding to the category shift $2 \rightarrow 3$ by construction. This is indeed the case for all clusters in all models. From the continuation ratio model as specified in Table 1, we have already seen that this pattern holds for all significant variables. From the results of this model, we can deduct that the burden of the restriction is relatively minor in this application.

Table 2: Model specifications of the finite mixture ordered logit model for two, three, and four clusters

	Two clusters		Three clusters			Four clusters			
Age	0.918	0.893	0.896***	0.889***	0.883***	0.892	0.893	0.889	0.888
Male sex	1.081***	1.062***	1.078	1.064	1.039	1.056	1.074	1.053	1.066
First Hb concentration									
0	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
0.1–10	1.921***	1.863***	1.823**	1.819**	1.779**	1.820**	1.821	1.825	1.811
10–20	3.589***	3.618***	3.343***	3.395**	3.311**	3.295**	3.300	3.339	3.291
20–30	5.783***	6.013**	5.595***	5.610***	5.500***	5.641***	5.670	5.649	5.587
30–40	8.467***	8.336***	6.675***	6.858***	6.590***	7.360***	6.557	7.532	7.516
30–47	11.018**	8.273***	10.341**	10.365**	10.023**	10.900**	11.015	10.9811	10.613
Second Hb concentration									
0	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
0.1–10	3.225***	3.121***	3.309***	3.268**	3.311**	3.200**	3.258	3.235	3.202
10–20	4.785***	4.419***	5.058***	5.008**	5.042**	4.940**	4.958	5.004	4.972
20–30	7.003***	6.796***	7.414***	7.327**	7.400**	7.598**	7.613	7.621	7.626
30–40	7.717***	6.987***	7.947***	7.809***	7.975***	8.016***	7.979	8.059	8.034
30–47	8.014***	6.940***	9.136***	9.069***	9.133***	9.900***	9.811	10.060	9.802
Intercept 1→2	3.498	3.406	3.463	3.413	3.349	3.434	3.441	3.417	3.406
Intercept 2→3	4.608	4.416	4.469	4.415	4.340	4.443	4.451	4.428	4.416
Intercept 3→4	6.472	6.289	6.378	6.318	6.238	6.375	6.389	6.356	6.355
cluster probabilities	0.547	0.453	0.518	0.259	0.223	0.284	0.263	0.210	0.243

In the FIMGOL models with two and three clusters, we again observe that the differences in parameters among clusters are present yet subtle. The decrease in parameter values for higher category shifts is however substantial, particularly for the parameters corresponding to the FIT-values of category shift $3 \rightarrow 4$.

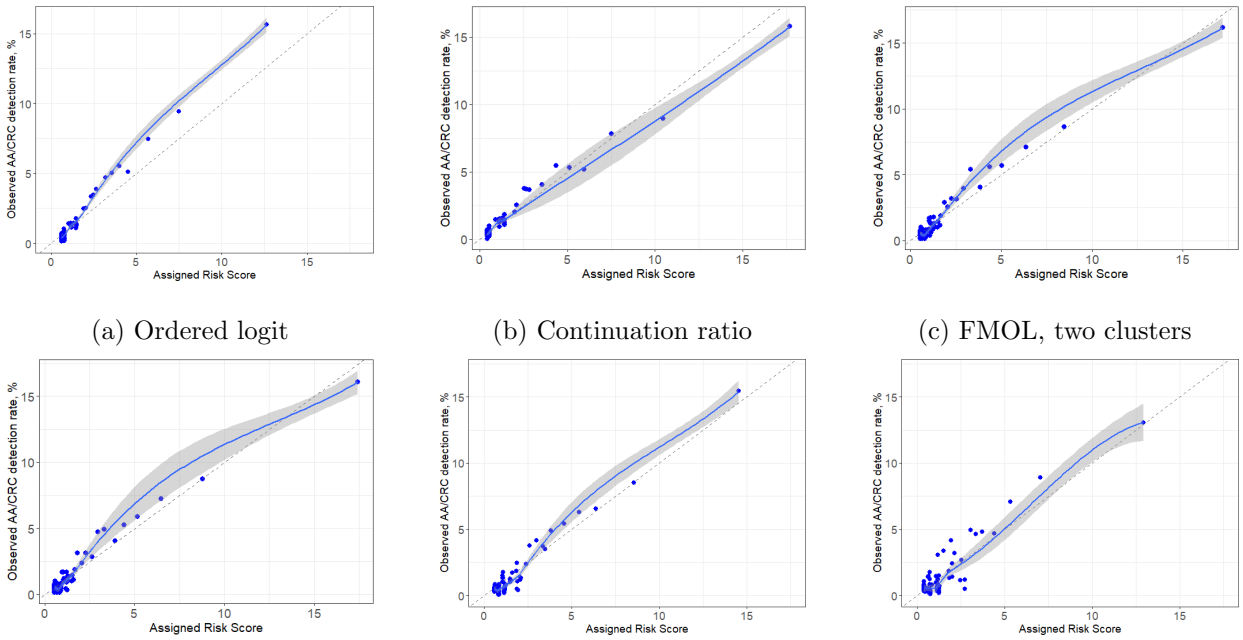
The FIMGOL model with four categories shows larger differences in parameter values. Particularly, the parameter values for cluster four take double the value of the parameters in cluster one (see Table 8 in the appendix). Note that the cluster probability of clusters two and four are however less than one percent. This means that the parameters of clusters two and four have limited influence on the model's probabilistic predictions. We can therefore conclude that four clusters are more than necessary in this application of the generalised finite mixture model. As we believe that we do not need more than three clusters in this application, we will therefore focus on models up until three clusters in this paper.

In addition to the finite mixture models with multiple clusters, we have fitted a model with one cluster. As expected, we have obtained the ordered logit model for the FMOL model and the generalised ordered logit (gologit) model for the FIMGOL model.

5.3 Goodness of fit

To evaluate the goodness of fit of our ordinal probabilistic predictions, we have split the population in percentiles and ordered those according to the mean assigned risk score. Though any other (weighted) sum of predicted probabilities can be chosen as a risk score, we have used the probability on Advanced Neoplasia (AN) as a risk score. We subsequently plot the observed detection rate of AN relative to the risk score assigned by our models. We can evaluate the goodness of fit by inspecting the distance to the 45°- line. The goodness of fit graphs of our proposed models can be found in Figure 3.

From Figure 3, we see that adding complexity to ordinal models increases the goodness of fit. We see that the standard ordered logit model underestimates the risk, particularly for high-risk percentiles. The mixture models outperform the ordered logit model in terms of fit. Note that our new models particularly outperform the ordered logit models for high-risk percentiles. In cancer screening, correct estimation of high risk individuals is of key importance.



(a) Ordered logit (b) Continuation ratio (c) FMOL, two clusters

(d) FMOL, three clusters (e) FIMGOL, two clusters (f) FIMGOL, three clusters

Notes: Plots of the observed Advanced Neoplasia (AN) detection rate over the mean assigned risk score of percentiles ordered by the estimated risk. Each blue dot represents one percentile. We have included goodness of fit plots zoomed in on percentiles with a risk score below five percent in Figure 7 of the appendix.

Figure 3: Goodness of fit plots of the ordinal models

Though it is essential to correctly estimate the risk of high-risk individuals, we also desire proper goodness of fit for the other observations. Our models also fit to these percentiles properly. As this is hard to judge from Figure 3, we have included goodness of fit plots zoomed in on percentiles with a risk score below five percent in Figure 7 of the appendix.

5.4 Numeric performance summary statistics

We have evaluated all models in terms of the area under the receiver operator curve (AUC), the Average Mean Absolute Error (AMAE), the Maximum Mean Absolute Error (MMEA), the Brier statistic, Ranked Brier statistic, and the Akaike Information Criterion (AIC). A summary of the performance statistics can be found in Table 3, where we have printed the best values of each performance statistic in bold.

5.4.1 Discriminatory ability

To compare our models to the binary logit model, we have evaluated whether the ordinal models are also able to predict the occurrence of AN. Likewise Meester (2021), all lesions other than AA and CRC are not considered as AN and will be discarded as a relevant outcome in the binary set-up.

Table 3: Summary of the performance statistics of the evaluated models

Model	AUC	AMAE	MMAE	Brier	Ranked Brier	AIC
Binary Logit	0.778	-	-	0.180	-	21224
Ordered Logit	0.779	1	1.415	**0.078**	0.043	57479
Continuation Ratio	0.783	0.953	**1.384**	0.079	0.043	**57128**
FMOL						
2 clusters	0.778	0.946	1.428	0.079	0.043	57808
3 clusters	0.778	**0.945**	1.405	**0.078**	0.043	57914
4 clusters	0.768	**0.945**	1.404	**0.078**	0.043	57975
FIMGOL						
2 clusters	0.758	0.964	1.471	0.079	0.043	58100
3 clusters	0.727	0.991	1.489	0.079	0.043	58699
4 clusters	0.678	1	1.481	0.079	0.043	69981
Random Forest	**0.791**	0.955	1.831	0.081	0.046	-

Notes: A report of the Area Under the ROC-curve (AUC), Average Mean Absolute Error (AMAE), Maximum Mean Absolute Error (MMAE), Brier-score, Ranked Brier-score and the Akaike Information Criterion (AIC). The statistics of the best performing ordinal model has been printed in bold, except for the Ranked Brier score, for which the differences are negligible. The reported statistics are computed on a test-set of 30% of the available data. Note that FMOL refers to the Finite Mixture Ordered Logit model and FIMGOL to the Finite Mixture Generalised Ordered Logit model.

To evaluate the binary predictive performance, we sum the individual's predicted probabilities that AA or CRC is detected. We use this sum of probabilities in our computations of the AUC.

In terms of the AUC, we observe that our ordinal models reaffirm that age, gender, and FIT-values are well able to predict colonoscopy outcomes. The values of the AUC of our models attain similar results as the binary model from Meester (2021). As expected, the ordered logit model in its binary form yields the same AUC as the binary logit model; the models are equivalent when the ordered logit is reduced to a binary set-up. Neither the FMOL nor the FIMGOL approaches are able to improve the binary predictive performance.

Where adding complexity to advanced regression models does not seem to improve the AUC of the models, we do see improvement for the random forest model. The flexibility of the random forest model and the power of a decision tree ensemble method allow for better predictions at the

cost of interpretability.

5.4.2 Ordinal classification performance

We evaluate the ordinal classification performance of our models by an analysis of the Mean Absolute Errors (MAE). As the outputs of all our models are probabilistic, we have used category specific thresholds for the prediction of a particular category. We have computed the optimal values of the thresholds with the AMEA and the MMAE as an objective function. In Table 3, we only report the optimal values of AMEA and MMAE. The optimal parameter values, hence the thresholds in our classification framework, can be found in tables 9 and 10 in the appendix.

We observe that the FMOL models outperform all other models in terms of AMAE. As the AMAE values are below one, the predicted classes by the cluster models are - averaged over the number of classes - correct or in the adjacent category. Note that for the ordered logit and the FIMGOL model with four clusters, the optimal AMAE equals one. For these models, we observe optimal classification thresholds such that the model always predicts class two. By construction of the AMAE, the optimization routine pushes the optimal solution to one of the classes in the center of the ordinal scale, when using AMAE as an objective.

The MMAE on the contrary is not influenced by a preference for classes in the center of the ordinal scale. Based on the MMAE, the continuation ratio model performs best. Note that for all models, the MAE is smaller than two for all categories. This means that under these models, the average predicted class for participants with CRC is at least in the second class in the ordinal scale. This implies that by these models, participants with CRC are on average assigned to an outcome category in which a participant is assigned a colonoscopy.

Note that based on the AMAE and the MMAE, the random forest model seems to be outperformed by the regression models. The relatively large optimal MAEs are caused by the fact that the random forest model assigns relatively more probability mass to the zero category (no colonoscopy). Due to the flexibility of the random forest model, the predictions are heavily influenced by the class imbalance in this application. Class reduction techniques could improve the ordinal performance of the random forest model.

5.4.3 Brier scores

We have evaluated the Brier score and the Ranked Brier score to evaluate the probabilistic predictions of our models. In this case, it should be noted that we cannot directly compare the Brier scores of the binary outcome model to the ordinal models, as the number of categories - two

for the former, four for the latter - differs among the models.

For both the Brier and the Ranked Brier score, we only observe marginal differences, due to a large class imbalance in the application of this research. As advanced lesions are detected in less than two percent of the participants only, the models will typically assign most probability mass to the categories without lesions. Though the Brier scores provide little ground for model comparison in this application, the metrics can be used in other applications of our models.

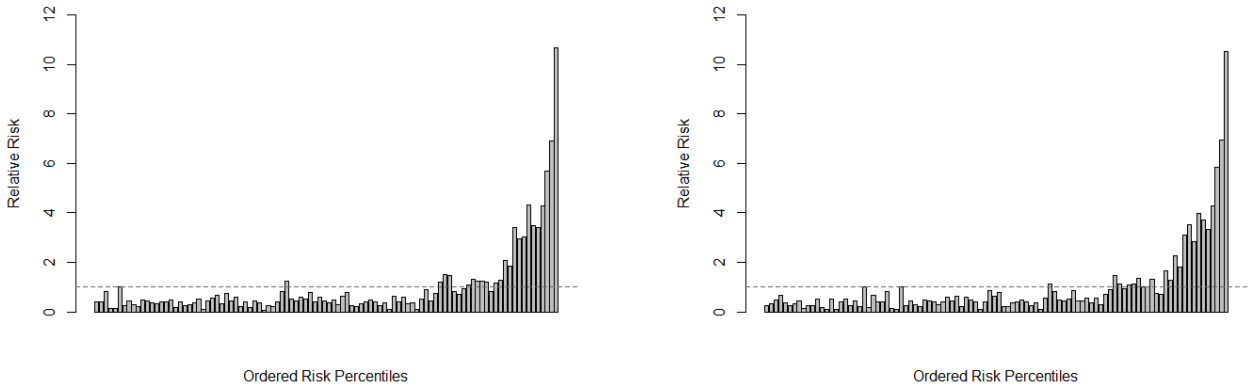
5.4.4 Akaike Information Criterion

Because the AIC is based on the log-likelihood function and the number of parameters in a model, we can use this metric to evaluate the amount of clusters. In this application, we observe that models with fewer clusters yield lower AIC values and are hence preferred over models with more clusters. Because the random forest model is not based on the log-likelihood, we cannot compute the AIC for this model.

5.5 Risk stratifying ability

The final goal of a predicting colonoscopy outcomes is to personalise screening for CRC. We therefore want our models to be able to differentiate between high-risk and low-risk individuals. To evaluate the risk stratifying ability of our models, we have plotted the relative risk of the percentiles ordered on the assigned risk score from our models. Ideally, our models find the highest relative risk in the top percentiles and the lowest in the bottom percentiles.

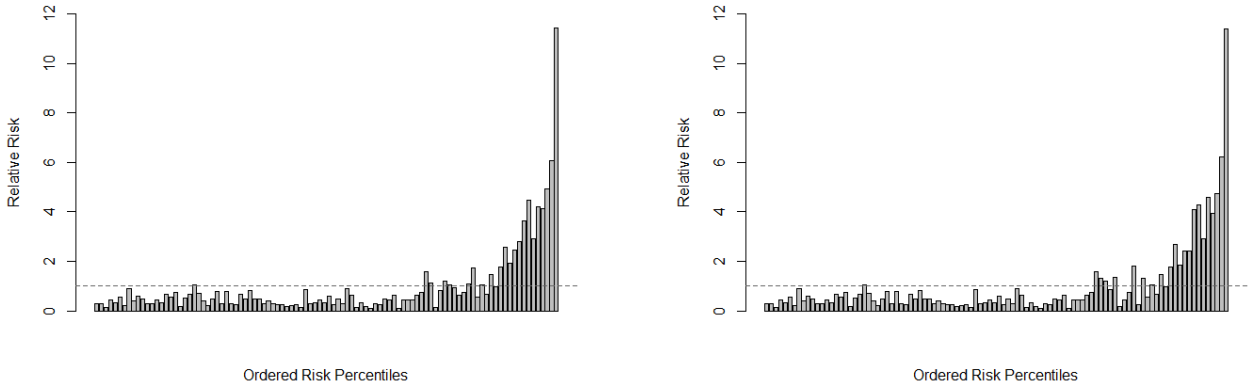
We observe that the ordered models are well able to stratify the risk among risk groups. Our most basic ordinal model as well as the continuation ratio model find percentiles in which the risk on colorectal cancer exceeds the population average more than 10 times. The relative risk plots of the ordered logit model and the continuation ratio model can be found in Figure 4.



(a) Ordered logit model (b) Continuation ratio model

Figure 4: Relative risk plots of percentiles ordered by the predicted risk score

Adding complexity marginally improves the risk stratifying ability of our models. We have included the relative risk plots of the finite mixture ordered logit model in Figure 5. The relative risk in the top 1% now exceeds the population average by 11 times. Also note that the first percentiles for which the relative risk exceeds one occur among higher ordered percentiles than for the standard ordered logit model. For completeness, we have include relative risk plots of the binary outcome model and the generalised cluster model in Figure 8 and Figure 9 of the appendix respectively.



(a) Two clusters (b) Three clusters

Figure 5: Relative risk plots of percentiles ordered by the predicted risk score for the finite mixture ordered logit model

Whereas adding complexity to the regression models does not alter the relative risk plot tremendously, our random forest model does give a different result. Our random forest model is able to isolate high-risk percentiles in unseen data with a risk that exceeds the population average by more than 28 times. We have plotted the relative risks of the ordered risk percentiles of the random forest model in Figure 6.

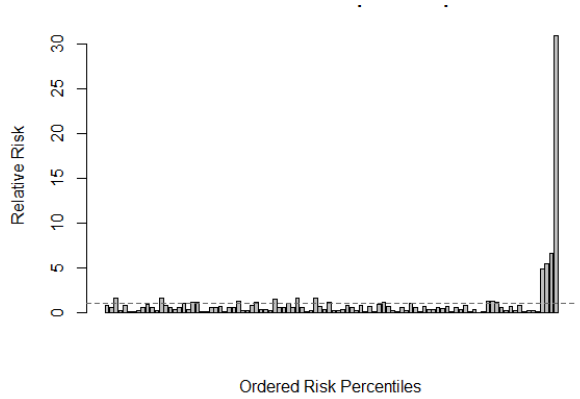


Figure 6: Relative risk plots of percentiles ordered by the predicted risk score for the random forest model

6 Conclusion & Discussion

In this research, we have investigated whether we can improve the performance of the previously applied predictive models by including knowledge on the stages of disease development in the Dutch screening programme for colorectal cancer (CRC). We conclude that though regression models are not able to improve the binary discriminatory ability, our application of the random forest model does improve the predictive performance, at the cost of interpretability. Our study therefore reaffirms that age, gender, and previously measured hemoglobin (Hb) concentrations as measured by the Fecal Immunochemical Test (FIT) are highly predictive for future colonoscopy outcomes. Our conclusion is therefore that risk stratification based on historical FITs can improve the effectiveness of the Dutch screening programme.

Our proposed ordinal models are new to the econometric literature. The ordinal models show that we can include the disease development in predictive models and that the stages of the disease development can be modelled successfully. Though our models do not outperform the previously applied models in terms of binary predictions, we have shown that our mixture approaches improve the goodness of fit and ordinal classification with respect to the standard ordered logit model. Our implementations of finite mixture models are ready to use for applications outside this context.

In addition to our conclusion that risk stratification based on historical FITs is possible, we have also gained a deeper insight in the mechanisms behind the predictive ability of models based on FIT-concentrations. Our implementation of the continuation ratio model shows that historical FIT-values are predictive for the future outcome of the FIT, but that they are less able to predict the progression of advanced lesions at the follow-up colonoscopy. This result is useful to clinicians investigating the pathway of a polyp to CRC as well as to modellers who wish to incorporate this pathway in simulation models. Improved simulation models help policy makers in optimising the screening strategy with respect to the test frequency and Hb cut-off value of the FIT.

Next to our regression models, we have experimented with the application of machine learning techniques. Our application of the random forest model outperforms the existing models in terms of the Area Under the Receiver Operator Characteristic (ROC) Curve (AUC) and is able to detect and isolate high-risk participants. By construction of the model, the increase in predictive power comes at the cost of interpretability, which is essential in any medical context.

6.1 Limitations

The first limitation in this research is that we are not able to observe lesions among observations which do not obtain a positive FIT in the third round. For that reason, we can only take the lesions observed during a colonoscopy after a positive FIT as a dependent variable. Though methodologically interesting, performing colonoscopies on participants without a positive FIT is clinically undesirable and costly in general, and therefore difficult to accomplish.

Another way to cope with this issue would be to include data on interval cancers; cancers that are missed in the screening programme, but detected after a patient experiences symptoms. Asymptomatic lesions would however still be missed in this set-up. A separate study to the previous FITs of participants with interval cancers could give more insights in the performance of the screening programme and the FIT.

A second limitation in our research is the limited number of screening rounds. The Dutch government has initiated the screening programme for CRC in 2014. This means that at the time of this research, we have data on three screening rounds. We can therefore only use data on FITs of the first two rounds as explanatory variables. With the number of rounds increasing, one could approach our predictive task using longitudinal data methods. Our current models do not yet allow for a longitudinal approach, but can be extended to include a time-varying component.

Finally, our models show that the improvement of a finite mixture approach on the binary

discriminatory ability is limited. With only a limited amount of predictor variables, our models have not revealed highly different clusters. Including more patient information - such as genetic data or demographic information - could widen the differences between clusters, hence increase the relevance of a finite mixture approach (Grobbee et al., 2017).

6.2 Implications for future research and practice

Our first and main suggestion for further research comprises a clinical study that could provide conclusive evidence that risk stratification improves the CRC screening programme. Our modelling study shows that historical FIT-concentrations can predict colonoscopy outcomes, but it is impossible to design a study that can conclusively show that personalising the screening procedure indeed increases the CRC detection rate without clinical validation.

In a clinical study, one could split the participants in two groups. The first group is the control group and follows the biennial screening programme as it is performed now. For the other group, a risk stratification approach is applied; one could use our predictive models to assign a risk score to participants, based on age, gender, and previous FITs. Researchers could then order the participants based on the predicted risks and split the participants of the test group into a number of subsets. For the subsets with relatively high risk, researchers should intensify the screening. The screening could be intensified by increasing the test-frequency and/or by lowering the FIT cut-off value for high-risk individuals. For the subsets with relatively low risk, researchers could do the opposite and for example reduce the test frequency. Comparing the advanced lesions detection rate, false-positive rate, and AUC of the test and control group could prove that risk stratification improves the effectiveness of the screening programme.

In addition to clinical studies, our models could be the basis for further methodological research. One could use and test our models in any other application with a multinomial dependent outcome variable with relevant ordering. In our implementation of the finite mixture models, we have used an iterative EM-algorithm to estimate the parameters of the model. Further research using for example a Majorise-Minimisation (MM) algorithm could decrease the computation time of estimating the parameters in our models (Hunter and Lange, 2004). A reduction of the computation time will increase the practical value of our models, certainly in applications with more parameters and/or categories than in our case.

Further research in applying more advanced machine learning techniques could further improve the prediction of colonoscopy outcomes. Janitza et al. (2016) for example show promising perfor-

mances of ordinal random forests. Other techniques such as ordinal neural networks as proposed by Cheng et al. (2008) and ordinal Bayesian Additive Regression Trees (BART) as proposed by Kindo (2016) potentially improve the performance of the previously applied models. Better predictive models could subsequently increase the returns of risk stratification in CRC screening.

Our research shows that machine learning algorithms improve the accuracy of the predictive models at the cost of interpretability. As interpretability is quintessential in medical applications, research towards more interpretable machine learning techniques could be valuable additions to the current existing models. Birbil et al. (2020) for example propose two algorithms for interpretation and boosting of tree-based ensemble methods. Interpretable machine learning algorithms could provide valuable prediction models to further increase the effectiveness of risk stratification.

To conclude, we emphasise again that this research provides additional ground for risk personalisation in screening programmes for CRC. We believe that - both in the Dutch context as well as in other screening programmes - CRC detection rates can increase while decreasing the number of unnecessary colonoscopies using risk personalisation. We hope to inspire researchers and policy makers with this hopeful conclusion in the combat against cancer.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: a review of methods and applications. *International journal of epidemiology*, 26(6):1323–1333.

Auge, J. M., Pellise, M., Escudero, J. M., Hernandez, C., Andreu, M., Grau, J., Buron, A., López-Cerón, M., Bessa, X., Serradesanferm, A., et al. (2014). Risk stratification for advanced colorectal neoplasia according to fecal hemoglobin concentration in a colorectal cancer screening program. *Gastroenterology*, 147(3):628–636.

Birbil, S. I., Edali, M., and Yuceoglu, B. (2020). Rule covering for interpretation and boosting. *arXiv preprint arXiv:2007.06379*.

Boes, S. and Winkelmann, R. (2006). Ordered response models. *Allgemeines Statistisches Archiv*, 90(1):167–181.

Brier, G. W. et al. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Cameron, T. A. (1988). A new paradigm for valuing non-market goods using referendum data: maximum likelihood estimation by censored logistic regression. *Journal of environmental economics and management*, 15(3):355–379.

Cheng, J., Wang, Z., and Pollastri, G. (2008). A neural network approach to ordinal regression. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1279–1284. IEEE.

Cooper, J. A., Parsons, N., Stinton, C., Mathews, C., Smith, S., Halloran, S. P., Moss, S., and Taylor-Phillips, S. (2018). Risk-adjusted colorectal cancer screening using the fit and routine screening data: development of a risk prediction model. *British journal of cancer*, 118(2):285–293.

Cooper, K., Squires, H., Carroll, C., Papaioannou, D., Booth, A., Logan, R., Maguire, C., Hind, D., and Tappenden, P. (2010). Chemoprevention of colorectal cancer: systematic review and economic evaluation. *Health technology assessment (Winchester, England)*, 14(32):1–206.

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

Crawford, S. M., Sauerzapf, V., Haynes, R., Forman, D., and Jones, A. P. (2012). Social and geographical factors affecting access to treatment of colorectal cancer: a cancer registry study. *BMJ open*, 2(2).

Cruz-Ramírez, M., Hervás-Martínez, C., Sánchez-Monedero, J., and Gutiérrez, P. A. (2014). Metrics to guide a multi-objective evolutionary algorithm for ordinal classification. *Neurocomputing*, 135:21–31.

Cruzado, J., Sánchez, F. I., Abellán, J. M., Pérez-Riquelme, F., and Carballo, F. (2013). Economic evaluation of colorectal cancer (crc) screening. *Best practice & research Clinical gastroenterology*, 27(6):867–880.

de Wijkerslooth, T. R., de Haan, M. C., Stoop, E. M., Bossuyt, P. M., Thomeer, M., Essink-Bot, M.-L., van Leerdam, M. E., Fockens, P., Kuipers, E. J., Stoker, J., et al. (2012). Burden of colonoscopy compared to non-cathartic ct-colonography in a colorectal cancer screening programme: randomised controlled trial. *Gut*, 61(11):1552–1559.

Dempster, A. P., Laird, N. M., and 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.

Deng, W., Chen, H., and Li, Z. (2006). A logistic regression mixture model for interval mapping of genetic trait loci affecting binary phenotypes. *Genetics*, 172(2):1349–1358.

Esserman, L. J., Thompson, I. M., and Reid, B. (2013). Overdiagnosis and overtreatment in cancer: an opportunity for improvement. *Jama*, 310(8):797–798.

Everitt, B. S. and Merette, C. (1990). The clustering of mixed-mode data: a comparison of possible approaches. *Journal of Applied Statistics*, 17(3):283–297.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874.

Feinberg, S. E. (1980). *The analysis of cross-classified categorical data*. Springer Science & Business Media.

Feldmann, U. and Steudel, I. (2000). Methods of ordinal classification applied to medical scoring systems. *Statistics in medicine*, 19(4):575–586.

Gilbert, P., Gilbert, M. P., and Varadhan, R. (2006). The numderiv package.

Grilli, L. and Rampichini, C. (2014). Ordered logit model. *Encyclopedia of quality of life and well-being research*, pages 4510–4513.

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.

Gu, Q., Zhu, L., and Cai, Z. (2009). Evaluation measures of the classification performance of imbalanced data sets. In *International symposium on intelligence computation and applications*, pages 461–471. Springer.

Hand, D. J. and Till, R. J. (2001). A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Hunter, D. R. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, 58(1):30–37.

Janitza, S., Tutz, G., and Boulesteix, A.-L. (2016). Random forest for ordinal responses: prediction and variable selection. *Computational Statistics & Data Analysis*, 96:57–73.

Kindo, B. P. (2016). Bayesian ensemble of regression trees for multinomial probit and quantile regression.

Lansdorp-Vogelaar, I., van Leerdam, M. E., Kuipers, E. J., de Koning, H., Bonfrer, H. M., van Ballegooijen, M., van Velthuysen, M.-L. F., Thomeer, M. G., van Veldhuizen, H., Nagtegaal, I., Ramakers, C., van Kemenade, F. J., Dekker, E., Spaander, M. C., Buskermolen, M., Toes-Zoutendijk, E., Kooyker, A. I., and Opstal-van Winden, A. W. (2019). Landelijke monitoring en evaluatie van het bevolkingsonderzoek naar darmkanker in nederland.

Li, G. (2018). Application of finite mixture of logistic regression for heterogeneous merging behavior analysis. *Journal of Advanced Transportation*, 2018.

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.

Mandelblatt, J., Andrews, H., Kao, R., Wallace, R., and Kerner, J. (1996). The late-stage diagnosis of colorectal cancer: demographic and socioeconomic factors. *American Journal of Public Health*, 86(12):1794–1797.

Markowitz, A. J. and Winawer, S. J. (1999). Screening and surveillance for colorectal cancer. In *Seminars in oncology*, volume 26, pages 485–498.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

Mcgowan, M. J. et al. (2000). Mj: ordinal outcomes with the continuation ratio model. In *Proceedings of the Northeast SAS Users Group Conference*. Citeseer.

Meester, R. G., Doubeni, C. A., Zauber, A. G., Goede, S. L., Levin, T. R., Corley, D. A., Jemal, A., and Lansdorp-Vogelaar, I. (2015). Public health impact of achieving 80% colorectal cancer screening rates in the united states by 2018. *Cancer*, 121(13):2281–2285.

Meester, R. G. S. (2021). Risk prediction based on quantitative fecal immunochemical test results. a population-based prediction model. *Working paper*.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The computer journal*, 7(4):308–313.

Peters, U., Jiao, S., Schumacher, F. R., Hutter, C. M., Aragaki, A. K., Baron, J. A., Berndt, S. I., Bézieau, S., Brenner, H., Butterbach, K., et al. (2013). Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology*, 144(4):799–807.

Peterson, B. and Harrell Jr, F. E. (1990). Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39(2):205–217.

Prinzie, A. and Van den Poel, D. (2008). Random forests for multiclass classification: Random multinomial logit. *Expert systems with Applications*, 34(3):1721–1732.

Quinlan, J. R. (1986). Induction of decision trees. *Machine learning*, 1(1):81–106.

Rodgers, A. et al. (1990). The uk breast cancer screening programme: an expensive mistake. *Journal of public health medicine*, 12(3/4):197–204.

Schreuders, E. H., Ruco, A., Rabeneck, L., Schoen, R. E., Sung, J. J., Young, G. P., and Kuipers, E. J. (2015). Colorectal cancer screening: a global overview of existing programmes. *Gut*, 64(10):1637–1649.

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

Smirnov, N. et al. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of mathematical statistics*, 19(2):279–281.

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

Toes-Zoutendijk, E., van Leerdam, M. E., Dekker, E., Van Hees, F., Penning, C., Nagtegaal, I., van der Meulen, M. P., van Vuuren, A. J., Kuipers, E. J., Bonfrer, J. M., et al. (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.

Tuma, M. and Decker, R. (2013). Finite mixture models in market segmentation: A review and suggestions for best practices. *Electronic Journal of Business Research Methods*, 11(1).

Weigel, A. P., Liniger, M. A., and Appenzeller, C. (2007). The discrete brier and ranked probability skill scores. *Monthly Weather Review*, 135(1):118–124.

Wilks, S. S. (1962). Mathematical statistics.

Williams, R. (2016). Understanding and interpreting generalized ordered logit models. *The Journal of Mathematical Sociology*, 40(1):7–20.

Wu, H.-C., Chang, C.-J., Lin, C.-C., Tsai, M.-C., Chang, C.-C., and Tseng, M.-H. (2014). Developing screening services for colorectal cancer on android smartphones. *Telemedicine and e-Health*, 20(8):687–695.

7 Appendix

Table 4: Model specifications of the reconstructed binary logit model

Variable	Odds Ratio	p-value
Age	0.999	<0.001
Male sex	1.248	0.96
First Hb concentration		
0	Ref.	
0.1–10	2.443	<0.001
10–20	4.424	<0.001
20–30	6.239	<0.001
30–40	6.326	<0.001
40–47	11.464	<0.001
Second Hb concentration		
0	Ref.	<0.001
0.1–10	4.391	<0.001
10–20	6.105	<0.001
20–30	9.025	<0.001
30–40	8.501	<0.001
40–47	8.916	<0.001
Intercept	-5.323	<0.001

Table 5: Values of the z-statistics of the differences in parameters of the continuation ratio model

	z-statistics	
Variable	Shift 1	Shift 2
First Hb concentration		
0	Ref.	Ref.
0.1–10	38.58***	35.16***
10–20	44.52***	19.16***
20–30	20.58***	11.35***
30–40	23.69***	8.07***
40–47	4.83***	11.43***
Second Hb concentration		
0	Ref.	Ref.
0.1–10	61.42***	26.99***
10–20	57.92***	17.76***
20–30	29.36***	19.87***
30–40	27.59***	13.44***
40–47	18.42***	9.47***

Notes: *** p-value <0.001, ** p-value <0.01, *p-value <0.05

Table 6: Model specification of the generalised finite mixture ordered logit model with two clusters

	Cluster 1			Cluster 2		
Variable	$1 \to 2$	$2 \to 3$	$3 \to 4$	$1 \to 2$	$2 \to 3$	$3 \to 4$
Intercept	3.418	3.913	4.300	3.478	4.040	4.506
Age	0.908	0.844	0.665	0.915	0.859	0.689
Male sex	1.233***	1.040***	0.902	1.030***	1.007**	0.994
First Hb concentration						
0	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
0.1–10	1.813***	1.780***	1.398	1.807***	1.787**	1.418
10–20	3.908***	3.093***	2.236	3.184***	3.124**	2.188
20–30	4.400***	4.182***	3.533	4.442***	4.323**	4.003
30–40	4.867***	4.595***	3.824	4.644***	4.411**	4.206
40–47	6.817***	6.607***	5.134	6.951***	6.663**	6.174
Second Hb concentration						
0	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
0.1–10	3.129***	3.075***	2.509	3.151***	3.121**	2.684
10–20	4.337***	4.257***	3.219	4.352***	4.175**	3.025
20–30	5.922***	5.644***	3.213	5.841***	5.726**	3.531
30–40	5.859***	5.592***	3.850	5.833***	5.608**	3.596
40–47	6.455***	6.340***	4.637	7.221***	6.752**	4.408
Cluster probabilities	0.621			0.379		

Notes: *** p-value <0.001, ** p-value <0.01, *p-value <0.05

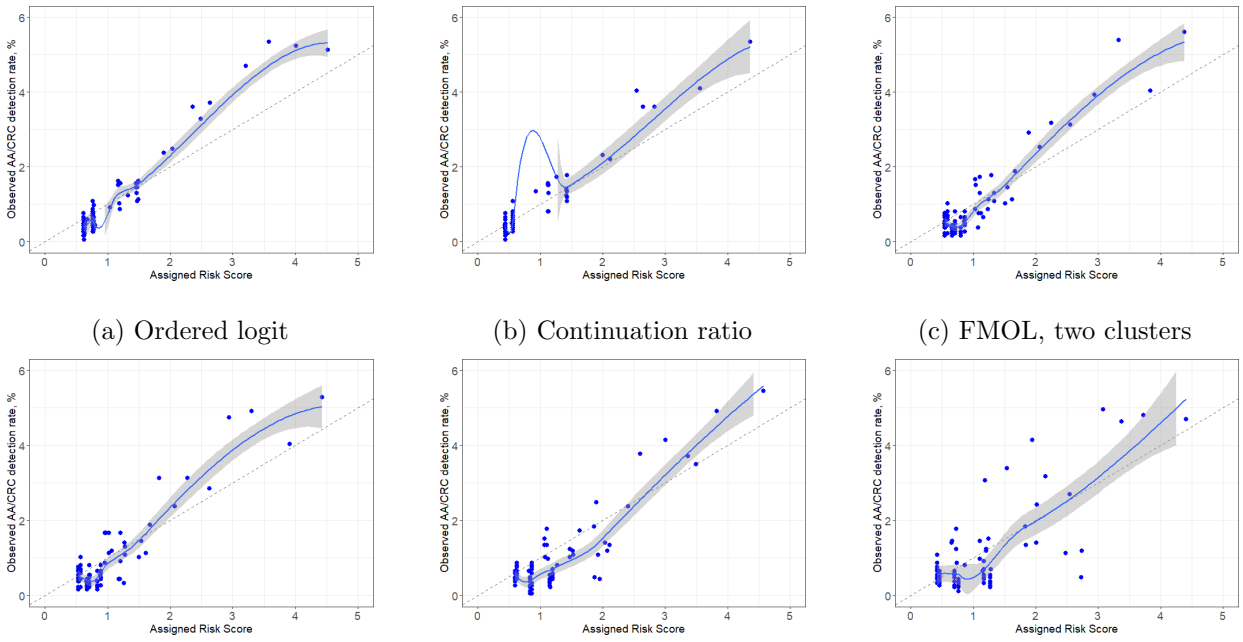
Table 7: Model specification of the generalised finite mixture ordered logit model with three clusters

	Cluster 1			Cluster 2			Cluster 3		
Variable	$1 \rightarrow 2$	$2 \rightarrow 3$	$3 \rightarrow 4$	$1 \rightarrow 2$	$2 \rightarrow 3$	$3 \rightarrow 4$	$1 \rightarrow 2$	$2 \rightarrow 3$	$3 \rightarrow 4$
Intercept	1.949	2.067	2.179	2.127	2.227	2.338	2.059	2.137	2.256
Age	0.727	0.624	0.443	0.745	0.639	0.449	0.739	0.632	0.473
Male sex	0.775***	0.641***	0.483	0.818***	0.673**	0.480	0.796***	0.657	0.453
First Hb concentration									
0	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
40–47	1.396***	1.244***	0.892	1.465***	1.350*	1.015	1.424***	1.299**	0.844
40–47	2.575***	2.303***	1.952	2.621***	2.470**	2.284	2.614***	2.400*	2.113
40–47	4.185***	3.822***	3.547	4.469***	4.308**	3.924	4.373***	4.001*	3.750
40–47	5.023***	4.287***	4.089	5.079***	4.863**	4.602	4.982***	4.774*	4.456
40–47	8.348***	7.584***	6.947	8.569***	7.830**	5.060	8.139***	7.907*	7.529
Second Hb concentration									
0	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
0.1–10	2.743***	2.693***	2.147	2.863***	2.656**	2.262	2.807***	2.760**	2.370
10–20	4.213***	4.095***	3.670	4.307***	4.137*	3.831	4.266***	4.145**	3.669
20–30	5.645***	5.501***	4.619	5.772***	5.306*	4.739	5.650***	5.552**	4.910
30–40	6.118***	5.615***	5.012	6.419***	6.133**	5.558	6.420***	6.136*	5.503
40–47	8.469***	8.082***	6.144	9.143***	8.674**	8.357	8.860***	8.215**	7.803
π_c	0.909			0.041			0.050		

Notes: *** p-value <0.001, ** p-value <0.01, *p-value <0.05

	Cluster 1			Cluster 2			Cluster 3			Cluster 4		
Variable	1 → 2	2→3	3→ 4	1 → 2	2→3	3→ 4	1 → 2	2→3	3→ 4	1 → 2	2→3	3→ 4
Intercept	1.477	1.723	2.119	1.171	1.230	1.614	0.833	0.853	1.446	1.645	1.780	2.449
Age	0.729	0.643	0.507	0.678	0.564	0.430	0.644	0.545	0.460	0.731	0.625	0.525
Male sex	0.874	0.711	0.682	0.801	0.698	0.643	0.710	0.589	0.505	0.956	0.749	0.673
First Hb concentration												
0	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
0.1—10	1.544	1.426	1.401	1.541	1.476	1.381	1.417	1.329	1.171	1.650	1.609	1.533
10—20	3.389	3.323	2.874	3.496	3.276	3.154	3.036	2.937	2.839	3.625	3.315	3.004
20—30	6.234	6.088	5.899	10.947	10.061	9.645	9.846	9.145	8.316	10.039	9.266	8.575
30—40	10.138	9.772	8.738	18.689	18.361	17.813	10.790	10.164	8.255	17.124	16.407	15.407
40—47	8.831	8.663	7.978	16.906	15.692	14.797	14.245	13.752	12.020	18.732	17.922	16.782
Second Hb concentration												
0	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.	Ref.
0.1—10	3.298	3.195	2.725	3.464	3.403	3.193	3.415	3.185	3.060	3.587	3.360	3.132
10—20	5.668	5.096	4.361	6.181	5.796	5.237	6.146	5.973	5.879	5.978	5.850	5.246
20—30	8.960	8.844	7.291	13.129	12.130	11.833	13.697	12.577	11.519	12.583	11.391	11.206
30—40	12.447	10.743	8.623	19.445	19.337	18.226	12.678	12.237	10.424	17.261	17.010	14.697
40—47	15.903	15.094	12.001	18.179	17.805	14.208	21.726	20.374	17.261	33.740	29.593	27.674
Cluster probabilities		0.227			0.004			0.772			0.001	

Table 8: Model specification of the generalised finite mixture ordered logit model with 4 clusters



(a) Ordered logit (b) Continuation ratio (c) FMOL, two clusters

(d) FMOL, three clusters (e) FIMGOL, two clusters (f) FIMGOL, three clusters

Figure 7: Goodness of fit plots of the ordinal models for the percentiles with a risk score $< 5\%$

Table 9: Optimal classification boundary values with Average Mean Absolute Error as an objective

Model	t_{crc}	t_{aa}	t_{na}	AMAE	MMAE
Ordered Logit	1.020	-1.522	-0.072	1	2
Continuation Ratio	0.002	0.067	0.117	1.142	1.470
Finite Mixture Cluster					
2 clusters	0.044	-0.033	0.109	0.946	1.668
3 clusters	0.044	-0.035	0.114	0.945	1.668
4 clusters	0.044	-0.035	0.114	0.945	1.667
Generalised Cluster					
2 clusters	0.046	-0.037	0.117	0.964	1.516
3 clusters	0.063	-0.054	0.015	0.991	1.411
4 clusters	1.067	-2.491	-0.230	1	2
Random Forest	0.009	0.029	0.131	1.078	2.008

Notes: The boundary values can be used to translate the probabilistic output of our models to a categorical classification, using algorithm 2.

Table 10: Optimal classification boundary values with Maximum Mean Absolute Error as an objective

Model	t_{crc}	t_{aa}	t_{na}	AMAE	MMAE
Ordered Logit	0.002	0.083	0.099	1.151	1.415
Continuation Ratio	0.002	0.083	0.103	1.153	1.384
Finite Mixture Cluster					
2 clusters	0.002	0.086	0.097	1.156	1.428
3 clusters	0.002	0.088	0.098	1.159	1.405
4 clusters	0.002	0.083	0.096	1.159	1.404
Generalised Cluster					
2 clusters	0.003	0.084	0.097	1.304	1.471
3 clusters	0.001	0.082	0.095	1.348	1.489
4 clusters	0.002	0.087	0.103	1.348	1.481
Random Forest	0.003	0.089	0.099	1.123	2.008

Notes: The boundary values can be used to translate the probabilistic output of our models to a categorical classification, using algorithm 2.

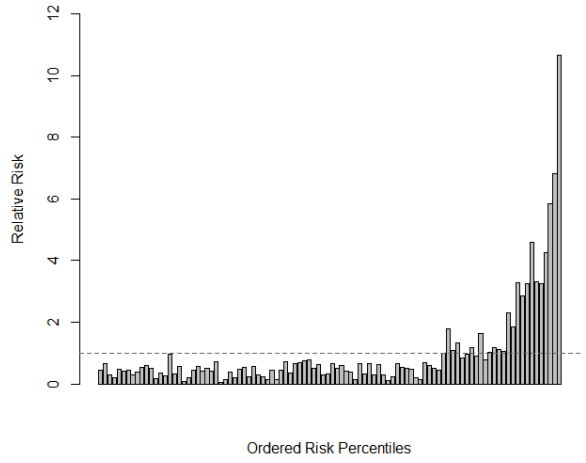
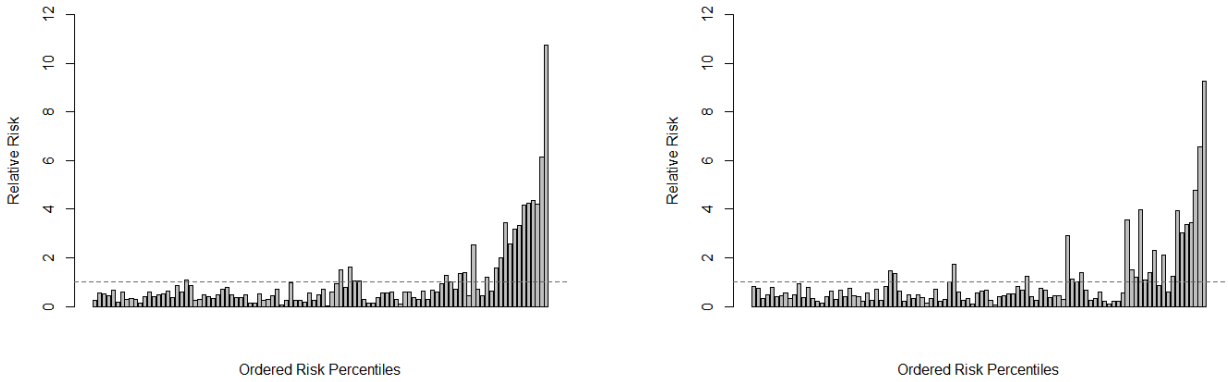


Figure 8: Relative risk plot of the binary outcome model



(a) Two clusters (b) Three clusters

Figure 9: Relative risk plots of the finite mixture generalised ordered logit model