

CANCER ONSET PREDICTION BY APPLYING EXTREME VALUE THEORY

Lisa Slachmuylders

Student ID: 00805775

Promotor: Prof. Dr. Stijn Luca

Copromotor: Dr. Louis Coussement

Tutor(s): Drs. Bastiaan A. Van Velthoven

A dissertation submitted to Ghent University in partial fulfilment of the requirements for the degree of Master of Science in Statistical Data Analysis.

Academic year: 2025 - 2026

ACKNOWLEDGEMENTS

I would like to thank my promotor, Prof. Stijn Luca, for giving me the opportunity to carry out this thesis within the Biostat group. This project brought together biostatistics and bioinformatics, and I am grateful to have had the opportunity to work at this intersection.

I am grateful to Drs. Bastiaan Van Velthoven (member of the Biostat group) and Dr. Louis Coussement (member of the BioBix lab) for their guidance throughout this project.

Bastiaan guided me through the statistical concepts underlying extreme value theory, and Louis provided the GECCO and SEER datasets - that made this work possible, and helped me navigate the bioinformatics challenges I encountered along the way. Throughout the project, they were both always available to help with methodological development, practical issues, theoretical questions, and, ultimately, the proofreading of this thesis.

Their support made this project not only possible, but genuinely enjoyable, and I am very grateful to them both.

CONTENTS

Acknowledgements	i
Contents	iii
Abstract	iv
1 Research objectives	1
2 Introduction	2
2.1 Modelling age at diagnosis	2
2.2 Parametric regression models	3
2.2.1 Survival framework	3
2.2.2 AFT models	3
2.3 Extreme value models	5
2.3.1 Asymptotic models	5
2.3.2 Generalized Extreme Value distribution	6
2.3.3 Generalized Pareto distribution	7
2.4 Genetic predictors	11
2.4.1 Genetic susceptibility to CRC	11
2.4.2 Genome-wide association studies	11
2.4.3 Polygenic risk scores and age at diagnosis	12
2.4.4 GECCO	12
3 Methodology	13
3.1 Modelling age-at-diagnosis	13
3.1.1 Data preprocessing	13
3.1.2 Empirical tail characterization	13
3.1.3 AFT framework	14
3.1.4 Generalized Pareto distribution	15
3.1.5 Model comparison and selection	18
3.2 Application to GECCO cohort	18
3.2.1 Data preparation	18
3.2.2 Model formulation	19
3.2.3 Model Selection	19

3.2.4	Model estimation	20
3.2.5	Validation	21
3.2.6	CRC-specificity of the genetic scale score	22
4	Results	23
4.1	Modelling age-at-diagnosis distribution	23
4.1.1	Exploratory data analysis	23
4.1.2	Empirical tail characterization	24
4.1.3	AFT regression models	24
4.1.4	Extreme value model	26
4.1.5	Model comparison and selection	29
4.2	Application to GECCO cohort	29
4.2.1	GECCO analysis dataset	29
4.2.2	Initial data analysis	30
4.2.3	Model selection	30
4.2.4	Model estimation	31
4.2.5	Validation	34
4.2.6	CRC-specificity of the genetic scale score	35
5	Discussion	37
	References	41
	Appendix A Supplementary data	45
A.1	Supplementary tables	45
A.2	Supplementary figures	51

ABSTRACT

Colorectal cancer (CRC) is increasingly diagnosed at younger ages, with early-onset CRC (EO-CRC) rising despite declining incidence among older adults. These early diagnoses form the lower tail of the age-at-diagnosis distribution, where observations are sparse and may be difficult to predict accurately using models fitted to the full distribution. Extreme value theory (EVT) provides a framework for directly modelling this tail behaviour.

Using SEER data, a Weibull accelerated failure time (AFT) model fitted to the full distribution was compared with a generalized Pareto distribution (GPD) fitted to threshold exceedances below age 50. The Weibull systematically underestimated early-diagnosis quantiles by one to two years, whereas the GPD closely tracked empirical quantiles across the tail, supporting the use of EVT for direct modelling of the lower tail.

The GPD framework was then applied to the GECCO consortium to investigate whether genetic burden contributes to earlier diagnosis. Eleven genetic variants were forward-selected by AIC as covariates on the scale parameter, and were summarized into a genetic scale score. This score was used to compute theoretical endpoint and expected age at diagnosis across representative percentiles of its distribution. Genetic burden was associated with progressively earlier EO-CRC diagnosis: individuals in the 90th percentile of the genetic scale score had a theoretical endpoint age 8.8 years younger (95% CI: 6.6, 11.0) and an expected age 3.3 years younger (95% CI: 2.5, 4.4) than those in the 10th percentile.

Applied to an external test cohort, the model retained good calibration, and the genetic scale score discriminated early-onset from late-onset CRC (AUC = 0.618, 95% CI: 0.549, 0.688). Separately, an interaction test confirmed that the genetic effect was specific to CRC cases rather than age in general.

These findings demonstrate that EVT offers a more accurate characterization of the lower tail of the CRC age-at-diagnosis distribution than conventional bulk models, and that genetic burden measurably alters the timing of diagnosis within the early-onset range, with potential relevance for personalized screening.

CHAPTER 1

RESEARCH OBJECTIVES

Colorectal cancer (CRC) remains a major public health concern and is a leading cause of cancer-related mortality worldwide. Population-based screening has reduced CRC mortality by preventing cancer through the detection and removal of premalignant lesions and by enabling cancers to be diagnosed at earlier stages. However, the incidence of early-onset (EO-) CRC, commonly defined as diagnosis before age 50, has increased despite declining incidence among older adults (Siegel et al., 2026b; Lee et al., 2026), contributing to the lowering of the recommended screening age from 50 to 45 years in the US in 2021 (Patel et al., 2022). Current screening strategies remain largely based on population-level incidence and average risk, despite substantial heterogeneity in CRC susceptibility. Unlike some other common cancers, CRC has no single dominant risk factor. Inherited syndromes account for only a small proportion of cases, while most CRC arises from the combined effects of genetic, lifestyle, environmental, and host-related factors (Siegel et al., 2026b). This multifactorial aetiology motivates personalized screening approaches that integrate multiple sources of risk to tailor screening recommendations to individuals.

Most existing models estimating CRC risk or survival use age as a predictor. Relatively little attention has been given to modelling the age-at-diagnosis distribution, despite its direct relevance to screening initiation. Accurate characterization of the lower tail of this distribution is particularly important because it captures individuals diagnosed with early-onset disease, many of whom are outside the target population for routine screening.

The primary objective of this thesis is to develop a statistical framework to model the lower tail of the CRC age-at-diagnosis distribution. To achieve this, population-based data from the Surveillance, Epidemiology, and End Results (SEER) Program are used to compare two parametric regression modelling approaches for estimating lower-tail quantities: conventional models, fitted to the full distribution, and extreme value models, fitted to the lower tail of age-at-diagnosis distribution.

The second objective is to investigate whether genetic variants increase early-onset probability by altering the lower tail of the age-at-diagnosis distribution. Using data from the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO), the prior identified modelling framework is extended to incorporate genetic covariates and quantify their combined contribution to the age-at-diagnosis distribution. The GECCO cohort is partitioned into independent training and test sets, with the training set used for model development and the test set used to evaluate generalizability of the estimated distributional shifts.

CHAPTER 2

LITERATURE REVIEW

2.1 Modelling age at diagnosis

CRC models conventionally treat age as a covariate, where it serves as a predictor to outcomes such as disease occurrence or survival (Archambault et al., 2022; Su et al., 2023; Chen et al., 2023; Jeon et al., 2018; Patel et al., 2022; Vratana et al., 2023). This project instead models age at diagnosis as the response variable in a case-only cohort. The motivation differs fundamentally from conventional risk prediction: rather than asking whether an individual will develop CRC, the objective is to model when diagnosis occurs among those who eventually do. Modelling the conditional (given CRC) age-at-diagnosis distribution allows to evaluate if covariates, such as genetic variants, alter the timing of diagnosis. Modelling age as an outcome is less common in cancer studies, but has precedent in related settings: population-level models of age-specific disease incidence (Armitage et al., 1954; Belikov, 2017; Belikov et al., 2021), models for longevity (Belzile et al., 2021), and age-of-onset disease studies more broadly (Martin et al., 2021; Ojavee et al., 2022).

Age of onset cannot be observed directly, so age at diagnosis is used as a proxy. However, routine screening advances diagnosis relative to disease onset, introducing lead-time bias that varies across the age distribution (Vratana et al., 2023). Routine screening was largely absent among individuals with EO-CRC during the study period, reducing the influence of screening-related lead time. Consequently, age at diagnosis is expected to be a more reliable proxy for age of onset in the lower tail than at older ages, although some differences remain due to familial screening, recent changes in screening recommendations, and potential biological differences in early-onset disease (Mauri et al., 2019; Tang et al., 2024; Patel et al., 2022). In this thesis, the outcome is referred to as age at diagnosis, while acknowledging that it more closely reflects age of onset in EO-CRC than in late-onset (LO).

To model the age-at-diagnosis distribution, data from the SEER Program are used. SEER has collected population-based cancer incidence data since 1973 and now covers approximately 45.9% of the US population, recording patient demographics, cancer type, disease stage, morphology, primary tumour site, tumour sequence number, etc. (National Cancer Institute, 2026). This project uses a subset of these records, with cases diagnosed between 2000 and 2022.

2.2 Parametric regression models

2.2.1 Survival framework

The outcome of interest, age at diagnosis, is positive and subjected to administrative right censoring in the SEER database. Ages are recorded for patients diagnosed before age 90, whereas those diagnosed at 90 years or older are grouped into a single “90+” category. So the observed outcome is $\min(T, c)$, where T denotes the true age at diagnosis and $c = 90$ the administrative cut-off. Because the censoring mechanism is imposed by the registry rather than the disease process, it is non-informative. This observation structure naturally motivates the use of a survival modelling framework, which provides likelihood-based methods for analysing right-censored positive outcomes (Kleinbaum et al., 2012).

Regression methods for censored data fall into two classes: semi-parametric and parametric. The semi-parametric approach, represented by the Cox model, estimates covariate effects while leaving the baseline distribution unspecified. The parametric approach assumes an underlying distribution and estimates its parameters by maximum likelihood. Both methods accommodate censoring and covariates, but only the parametric approach yields a fully specified distribution. This is required here, because the bulk model is evaluated in the early-onset tail, where the data is sparse, and quantiles there are typically obtained by extrapolation. As an unspecified baseline cannot be evaluated, the semi-parametric approach cannot supply tail quantiles (Kleinbaum et al., 2012). The accelerated failure time (AFT) model is therefore selected as the parametric regression framework for the bulk distribution.

2.2.2 AFT models

An AFT regression model specifies the outcome directly as a function of covariates on the log scale as:

$$\log T_i = \mu + \sum_j \beta_j x_{ij} + \sigma \varepsilon_i \quad (2.1)$$

with μ the intercept, β the regression coefficients, $\sigma > 0$ a scale parameter, and ε_i an error term with a specified distribution. When exponentiated, the covariates act multiplicatively on T_i . A unit change in covariate j stretches or contracts T by a constant factor $\exp(\beta_j)$. This is the central assumption of the model: the covariate multiplies every quantile of the outcome distribution by the same factor, so its effect is a uniform rescaling of the distribution rather than a change in its shape. The error distribution is fixed in advance, and its spread σ is common across covariate values (Kleinbaum et al., 2012). The distribution assumed for ε fixes the distributional family of T . A normal error gives a lognormal T , a logistic error a log-logistic T , and a Gumbel error a Weibull T (Kleinbaum et al., 2012).

Specifying an AFT model thus reduces to choosing the error distribution, which sets the shape of the fitted outcome distribution while the log-linear covariate structure remains unchanged. Because covariates uniformly rescale the distribution rather than altering its shape, an appropriate distributional assumption is essential. The exponential distribution is unsuitable because its single parameter implies a memoryless distribution with a mode at the origin, whereas age at diagnosis is unimodal and concentrated at older ages (National Cancer Institute, 2026). More flexible distributions are therefore required. Common choices for positive outcomes include the Weibull, gamma, log-normal, log-logistic, and generalized gamma distributions (Kleinbaum et al., 2012). The Erlang distribution, a special case of the gamma whose shape parameter is constrained to integers, can also be included as Belikov (2017) showed that it approximates the age distribution of early-onset incidence for several cancer types, including CRC, well.

The parameters $\theta = (\mu, \sigma)$ are estimated by maximizing the likelihood for censored data, defined as:

$$L(\theta) = \prod_{i=1}^{N_{\text{obs}}} f(t_i | \theta, x_i) \cdot \prod_{k=1}^{N_{\text{cens}}} [1 - F(c | \theta, x_k)], \quad (2.2)$$

with t_i the N_{obs} observed ages, x_i the corresponding covariates, and $c = 90$ the fixed censoring point shared by all N_{cens} censored observations. An observed age contributes its density $f(t_i | \theta, x_i)$, while each censored age contributes $1 - F(c | \theta, x_k)$, the probability that the true age exceeds the cap. Although c is shared, these contributions differ across individuals because the covariates enter the distribution through the scale. Assuming independent observations, the joint likelihood factorizes into this product. The estimate $\hat{\theta}$ that maximizes Eq. 2.2 specifies the full outcome distribution, from which any quantile can be evaluated for a given covariate pattern (Kleinbaum et al., 2012).

All distribution parameters are estimated jointly and numerically to maximise the censored likelihood over the full dataset. So the fit is dominated by the region of highest observed density, and the behaviour of the lower tail depends on whether any parameter is free to shape the tail once the bulk is fixed. The gamma and Erlang only have a shape and scale parameter. A bulk at older ages forces the shape to a large value that flattens the low tail, so the tail follows from the bulk rather than being set on its own. The lognormal is similar: its two parameters both describe the bulk, leaving the lower tail determined by them. The Weibull divides the roles more cleanly, with the scale positioning the bulk and the shape setting the rate of tail decay, so the tail retains some independent control. The generalised gamma adds a third parameter and separates body and tail further (Casella et al., 2002).

Regardless of the distributional family, an AFT model fits a single distribution across the whole range. The cost of fitting one global form is that the lower-tail quantiles are extrapolated under the shape that fits the bulk rather than estimated from the sparse data in the tail, and the constant-rescaling assumption holds the rate of tail decay fixed across

covariate values. A fitted AFT model accommodates censoring, admits covariates, and can estimate lower quantiles (Kleinbaum et al., 2012). These properties place the AFT model as a full-distribution approach to the early-onset tail, to be set against the following tail-specific approach.

2.3 Extreme value models

In contrast to AFT, extreme value theory (EVT) models the tail directly, replacing a global distributional assumption with an asymptotic approximation for extreme observations.

2.3.1 Asymptotic models

Asymptotic models focus on the behaviour of $M_n = \max\{Y_1, \dots, Y_n\}$, where $Y_1, \dots, Y_n$ is a sequence of independent and identically distributed random variables with common distribution function F . In theory, the distribution of M_n can be derived for all values of n :

$$\Pr\{M_n \leq y\} = [F_Y(y)]^n \quad (2.3)$$

Because F_Y is unknown, an extreme value analogue of the central limit theory (CLT) is used. It provides approximate model families for F_Y^n , estimated based on the extreme data alone. As in the CLT, the limiting behaviour of F_Y^n is examined as $n \rightarrow \infty$. For any fixed $y < y_+$, $F(y)^n \rightarrow 0$, so the distribution of M_n collapses to a point mass at y_+ , where $F(y_+) = 1$. This degenerate limit carries no information about the tail. A meaningful limit is recovered by centring and scaling, $M_n^* = \frac{M_n - b_n}{a_n}$. The extremal types theorem 2.1 states the following:

Theorem 2.1. *If there exist sequences of constants $\{a_n > 0\}$ and $\{b_n\}$ such that*

$$\Pr\left\{\frac{M_n - b_n}{a_n} \leq y\right\} \longrightarrow G(y) \quad \text{as } n \rightarrow \infty,$$

for a non-degenerate distribution function G , then G belongs to the Gumbel, Fréchet, or Weibull family, regardless of the parent distribution F_Y .

The Gumbel, Fréchet, or Weibull family distributions are collectively called the extreme value distributions, and are historically typed I, II, and III.

$$\begin{aligned}
 \textbf{Type I: } G(y) &= \exp \left\{ -\exp \left[-\left(\frac{y-b}{a} \right) \right] \right\} \quad -\infty < y < \infty \\
 \textbf{Type II: } G(y) &= \begin{cases} 0 & y \leq b \\ \exp \left\{ -\left(\frac{y-b}{a} \right)^{-\alpha} \right\} & y > b \end{cases} \\
 \textbf{Type III: } G(y) &= \begin{cases} \exp \left\{ -\left[-\left(\frac{y-b}{a} \right) \right]^\alpha \right\} & y < b \\ 1 & y \geq b \end{cases}
 \end{aligned} \tag{2.4}$$

Each type has a location b and a scale parameter $a > 0$, with type II and III adding a shape parameter $\alpha > 0$. The distributional families differ in their tail decay. The Gumbel limit has an unbounded tail with exponential decay, the Fréchet an unbounded tail with polynomial decay, and the Weibull a finite upper endpoint (Coles, 2001).

2.3.2 Generalized Extreme Value distribution

Fitting one of the three families directly forces a prior choice of tail type and ignores the uncertainty in that choice. The Generalized Extreme Value (GEV) distribution removes this step by absorbing the three families into a single parametric form, which redefines the non-degenerate $G(y)$ distribution from theorem 2.1 as

$$G(y) = \exp \left\{ -\left[1 + \xi \left(\frac{y-\mu}{\sigma} \right) \right]^{-1/\xi} \right\} \tag{2.5}$$

defined on $y : 1 + \xi \left(\frac{y-\mu}{\sigma} \right) > 0$, with $-\infty < \mu < \infty$ as a location, $\sigma > 0$ a scale, and $-\infty < \xi < \infty$ a shape parameter. The shape parameter sets the tail type: $\xi > 0$ gives the heavy-tailed Fréchet form, $\xi < 0$ the bounded Weibull form, and the limit $\xi \rightarrow 0$ the Gumbel form, the latter resulting in the Eq. 2.6. The sign of ξ is inferred from the data rather than fixed in advance. The GEV family is used to model the maxima of long sequences/blocks of data, e.g. annual maxima.

$$G(y) = \exp \left\{ -\exp \left[-\left(\frac{y-\mu}{\sigma} \right) \right] \right\}, \quad -\infty < y < \infty \tag{2.6}$$

A characterisation of the extremes involves two components, their frequency and their size. Modelling block maxima fixes the frequency at one event per block and models the size alone. The peaks-over-threshold (POT) approach of the next section models both (Coles, 2001).

2.3.3 Generalized Pareto distribution

The peaks-over-threshold approach

The POT approach treats any observation exceeding a threshold u as extreme. It uses every exceedance rather than one maximum per block, so it retains more of the available data. The object of interest is the conditional survival distribution of the excess $Z = Y - u$ given $Y > u$,

$$\Pr\{Y > u + z \mid Y > u\} = \frac{1 - F_Y(u + z)}{1 - F_Y(u)} \quad z > 0. \quad (2.7)$$

This conditional form is the defining feature of the model. The asymptotic counterpart of the extremal types theorem identifies its limit (Pickands, 1975; Coles, 2001).

Theorem 2.2. *If F_Y satisfies the extremal types theorem 2.1, with GEV limit of shape ξ , then for a sufficiently high threshold u the distribution function of the excess $Z = Y - u$, conditional on $Y > u$, is approximately Generalized Pareto with scale $\sigma_u > 0$ and shape ξ .*

The limiting Generalized Pareto distribution (GPD) function is (Eq. 2.8), and is defined on $\{z : z > 0 \text{ and } 1 + \xi z/\sigma_u > 0\}$.

$$F_Z(z) = 1 - \left[1 + \frac{\xi z}{\sigma_u} \right]^{-1/\xi}, \quad \xi \neq 0 \quad (2.8)$$

$$\sigma_u = \sigma + \xi(u - \mu) \quad (2.9)$$

GPD is the only non-degenerate limiting distribution for rescaled excesses, just as GEV is the only one for block maxima. The shape ξ is shared between the two. Block maxima are approximately $\text{GEV}(\mu, \sigma, \xi)$ distributed, and exceedances above u $\text{GP}(\sigma_u, \xi)$ with scale σ_u (Eq. 2.9) (Coles, 2001). As with GEV, ξ sets the qualitative character of the tail, i.e. tail index. If $\xi > 0$ the distribution of excesses has a heavy, unbounded Pareto tail. The limit $\xi \rightarrow 0$ gives the exponential form Eq. 2.10 and results in a light tail with no upper bound. The case $\xi < 0$ gives a Beta-type distribution with a finite theoretical upper end-point at $Z_{\max} = -\frac{\sigma_u}{\xi}$. The bounded case is the relevant one when the variable has a natural ceiling (Gilleland et al., 2016; Scarrott et al., 2012).

$$F_Z(z) = 1 - \exp\left(-\frac{z}{\sigma_u}\right), \quad z > 0, \quad \xi = 0 \quad (2.10)$$

The GDP provides a model for the conditional exceedance probability, while the unconditional tail probability additionally depends on the probability of exceeding the threshold, $\zeta_u = \Pr\{Y > u\}$, which is typically estimated by the sample proportion of points above threshold u . Writing $Y = u + Z$, the tail probability decomposes as

$$\Pr\{Y > u + z\} = \Pr\{Y > u\} \cdot \Pr\{Y > u + z \mid Y > u\} = \zeta_u [1 - F_Z(z)], \quad (2.11)$$

where F_Z is the excess distribution (Eq. 2.7). Substituting $y = u + z$ and the GPD form of $1 - F_Z$ from Eq. 2.8 gives, for $y > u$,

$$\Pr\{Y > y\} = \begin{cases} \zeta_u \left[1 + \frac{\xi(y-u)}{\sigma_u} \right]^{-1/\xi}, & \xi \neq 0 \\ \zeta_u \left[\exp\left(-\frac{y-u}{\sigma_u}\right) \right], & \xi = 0 \end{cases} \quad (2.12)$$

The resulting expression gives the unconditional tail probability for $y > u$. To obtain a fitted value of Y corresponding to a specified tail probability p , this expression can be inverted. The p -quantile y_p is defined by $\Pr\{Y > y_p\} = p$, resulting in

$$y_p = \begin{cases} u + \frac{\sigma_u}{\xi} \left[\left(\frac{\zeta_u}{p} \right)^\xi - 1 \right], & \xi \neq 0 \\ u + \sigma_u \log\left(\frac{\zeta_u}{p}\right), & \xi = 0 \end{cases} \quad (2.13)$$

Every quantile therefore reduces to three estimated quantities, the exceedance probability ζ_u , the scale σ_u , and the shape ξ , once the threshold u is fixed (Coles, 2001; Scarrott et al., 2012).

Threshold selection

Threshold selection involves a bias-variance trade-off. If chosen too low, the excess distribution may not be well approximated by the limiting GPD, resulting in biased parameter estimates. If chosen too high, the number of exceedances decreases, increasing the variance of the estimates. The standard practice is to select the lowest threshold for which the GPD approximation is adequate. Two complementary visual methods are commonly used to guide this choice. The first is an exploratory diagnostic applied prior to model fitting, while the second assesses parameter stability across a range of candidate thresholds (Coles, 2001).

The first diagnostic method is the mean residual life (MRL) plot, which shows the mean excess over a range of thresholds. A suitable threshold is identified where the graph is linear as the threshold increases. For $Z \sim \text{GPD}(\sigma_u, \xi)$ with $\xi < 1$ (when $\xi \geq 1$ the mean is infinite), the mean of the exceedances is $E(Z) = \sigma_u/(1 - \xi)$. If the GP model holds above a threshold u_0 , it holds above every higher threshold $u \geq u_0$, with the same shape ξ and a scale that changes linearly with u (following from Eq. 2.9). The empirical estimates for the mean excess $e(u)$ is therefore expected to be linear in u above a valid threshold u_0 .

$$e(u) = E(Y - u | Y > u) = \frac{\sigma_{u_0} + \xi u}{1 - \xi}, \quad u > u_0 \quad (2.14)$$

A suitable threshold is the lowest u above which $e(u)$ remains linear. Confidence bands follow from the approximate normality of the sample mean (Coles, 2001).

A second diagnostic method is the parameter stability plot, that fits the GPD at a range of thresholds and tracks the estimates. Because the scale changes linearly with u (Eq. 2.9), it varies with u even when the model is correct. By reparameterizing the scale as Eq. 2.15, the modified scale σ^* becomes constant with respect to u .

$$\sigma^* = \sigma_u - \xi u \quad (2.15)$$

Now both σ^* and ξ should be constant above u_0 , once their sampling error is allowed for, and the threshold is taken as the lowest value above which the estimates remain stable (Coles, 2001; Gilleland et al., 2016).

These graphical methods are standard practice but prone to subjectivity. The MRL plot in particular can be hard to interpret. A further limitation is that more than one threshold may give an adequate fit while implying different tail behaviour, so a fixed choice threshold discards that uncertainty. Threshold selection therefore requires some caution (Coles, 2001; Scarrott et al., 2012).

Non-stationarity and covariates

The classical extreme value theory assumes identically distributed observations. When the tail depends on covariates this assumption fails and there is no general limit theory for the non-stationary case. The pragmatic remedy is to keep the GPD as a template and let its parameters vary (Coles, 2001).

Covariates can enter the scale, the shape, or the threshold. The convention introduces them into the scale alone and holds the shape constant (Davison et al., 1990). The rationale for keeping the shape fixed is twofold. The shape fixes the tail index and is often treated as a property of the data generating process rather than something that covariates can shift. Practically, ξ is notoriously difficult to estimate even in the stationary case. It requires a large number of exceedances for reliable inference, and adding covariates to it introduces additional instability, wide confidence intervals (CIs), and identifiability problems. (Coles, 2001) explicitly notes these difficulties and recommends treating ξ as fixed unless there is strong theoretical justification otherwise.

The scale parameter is where covariate information has the clearest place because it directly controls the excess mean and spread of the exceedance distribution above the threshold. A larger σ_u means exceedances extend further, so if covariates have an effect on the tail, it can change how far values fall beyond the threshold. This can be captured by σ_u . Covariates enter σ_u through a log link, $\sigma_u = \exp(\eta)$, where η is a linear predictor of

the covariates. The log link keeps σ_u positive for any covariate value (Coles, 2001; Davison et al., 1990).

Covariates can, in principle, also enter the threshold itself, so that u varies with the covariate rather than remaining fixed (Scarrott et al., 2012). This is not pursued here. The threshold marks where the GP approximation becomes valid, not a property of the tail distribution itself, so a covariate-varying threshold would draw exceedances from different truncation points and undermine standard threshold diagnostics (Coles, 2001).

Model diagnostics

Goodness-of-fit (GOF) is assessed with probability and quantile plots. Both order the excesses and compare the empirical distribution against the fitted GPD, the first on the probability scale and the second on the scale of the variable. Both plots are approximately linear under an adequate model. A comparison of the fitted density against the exceedances gives a further check (Coles, 2001). Non-stationary models need standardised residuals. The fitted scale depends on covariates, so each excess is drawn from a GPD with a different scale and/or shape. Excesses at covariate values with a larger scale or a more positive shape are systematically larger, and pooling them against one fitted curve would confound these covariate effects with lack of fit. Standardisation removes it. Each excess z_i is mapped through its own fitted parameters, $\hat{\sigma}_{u,i}$ and/or $\hat{\xi}_i$, to a standard exponential scale, $\tilde{z}_i = \hat{\xi}_i^{-1} \log(1 + \hat{\xi}_i z_i / \hat{\sigma}_{u,i})$, which is $\exp(1)$ under a correctly specified model. The probability and quantile plots are then built on $\tilde{z}_i$, restoring a single scale for judging a covariate-dependent fit (Coles, 2001). The residuals can also be plotted against the covariates, where a remaining pattern signals structure the exponential plots alone can miss (Davison et al., 1990).

To summarize the POT approach, only the exceedances above a threshold are modelled, so the tail quantiles are estimated from the tail data rather than extrapolated under a shape fitted to the bulk, as in the AFT approach. This gains a direct tail estimate at the cost of a threshold choice. Most importantly, the GPD is asymptotically justified as the limiting distribution of threshold exceedances, regardless of the parent distribution, whereas the AFT's bulk-fitted shape carries no such guarantee in the tail. A GPD model also admits covariates, estimates the tail quantiles, and returns a finite theoretical endpoint for the modelled distribution when the shape is negative.

2.4 Genetic predictors

After selecting an appropriate modelling framework for the tail characterization, genetic information can be incorporated through covariate effects on the model parameters.

2.4.1 Genetic susceptibility to CRC

The incidence of CRC in adults younger than 50 years has risen steadily over the past three decades, making EO-CRC an increasingly important public health concern. Despite this increase, the underlying causes remain largely unknown (Siegel et al., 2026b; Siegel et al., 2026a; Lee et al., 2026). This growing burden has renewed interest in the biological and genetic factors influencing the timing of CRC onset.

Some differences in tumour characteristics and molecular profiles have been observed across the age spectrum of CRC. Compared to LO-CRC, EO-CRC tumours are more often left-sided or rectal and more likely to present with synchronous metastatic disease (Willauer et al., 2019; Tang et al., 2024; Akimoto et al., 2021). Molecular difference have also been reported. TP53 mutations are consistently reported as more prevalent in EO-CRC (Willauer et al., 2019; Huang et al., 2025), while APC and BRAF mutations are generally less frequent (Willauer et al., 2019; Hamilton et al., 2023), findings for KRAS and mismatch-repair status are mixed across studies (Huang et al., 2025; Tang et al., 2024; Hamilton et al., 2023). While these observations suggest some variation in tumour biology with age at diagnosis, they describe characteristics of CRC once it has developed and do not explain why some individuals develop the disease earlier than others. They nevertheless raise the possibility that genetic susceptibility may also vary with age at diagnosis.

CRC onset timing is influenced in part by inherited susceptibility, where rare hereditary syndromes such as Lynch syndrome and familial adenomatous polyposis increase CRC risk and often result in diagnosis at younger ages. However, these syndromes account for only 16-20% of EO-CRC cases (Siegel et al., 2026b). Most individuals diagnosed before age 50 have no identifiable hereditary syndrome, and known environmental and lifestyle factors likewise fail to fully explain the increasing incidence of EO-CRC (Maas et al., 2026; Syed et al., 2019). Consequently, the determinants of age at CRC onset remain incompletely understood. Common inherited genetic variation has emerged as one plausible contributor, prompting genome-wide association studies (GWAS) to identify common susceptibility loci and investigate whether inherited variation also influences the age at which CRC develops.

2.4.2 Genome-wide association studies

GWAS test hundreds of thousands to millions of common variants across the genome for association with a trait, without requiring prior biological hypotheses, unlike candidate-gene studies. Applied to CRC, successive GWAS efforts have expanded the number of known

susceptibility loci, from an initial set of common variants (Peters et al., 2013) to further discoveries incorporating rare variants (Huyghe et al., 2019), with the most recent multi-ancestry analysis identifying over 200 loci (Fernandez-Rozadilla et al., 2023). Individual genetic variants are typically associated with modest effects on CRC risk. Collectively these associations can explain a growing share of inherited susceptibility, enabling polygenic models that aggregate this information into a single polygenic risk score (PRS).

2.4.3 Polygenic risk scores and age at diagnosis

GWAS-derived PRS summarise inherited genetic susceptibility into a single measure of inherited genetic predisposition. Multiple studies have shown that individuals with higher PRS not only have an increased lifetime risk of CRC, but also tend to be diagnosed at younger ages, making PRS an attractive tool for personalised screening strategies (Archambault et al., 2020; Archambault et al., 2022; Jeon et al., 2018; Chen et al., 2023; Tamlander et al., 2024; Briggs et al., 2022). Incorporating PRS into screening recommendations has therefore been proposed as a means of identifying individuals who may benefit from earlier screening while delaying screening for those at lower inherited risk, potentially improving the efficiency of population-based programmes (Tamlander et al., 2024).

PRS are attractive from both a methodological and practical perspective, as it is an aggregate measure yielding a more parsimonious model than modelling genetic variants separately. PRS are typically constructed by summing risk alleles across susceptibility loci, weighted by effect sizes estimated from GWAS of CRC risk (Marees et al., 2018). Consequently, both the selection of variants and their weights are therefore optimized to predict disease occurrence, not timing of the diagnosis. While several studies have reported stronger associations between PRS and EO-CRC over LO-CRC, indicating that inherited susceptibility is related to age at diagnosis (Archambault et al., 2020; Archambault et al., 2022; Li et al., 2021), conventional PRS remain optimized for disease risk prediction. This thesis instead evaluates susceptibility loci according to their contribution to the age-at-diagnosis distribution among individuals who develop CRC, allowing variant selection to be driven by their association with age at diagnosis rather than overall disease risk.

2.4.4 GECCO

GECCO is an international collaboration established to identify genetic and environmental determinants of CRC through large-scale genome-wide association studies. By combining harmonized genotype and phenotype data across numerous epidemiological studies, it provides one of the largest resources for investigating common genetic variation in CRC (Peters et al., 2013), resulting in the identification of more than 200 CRC susceptibility (Fernandez-Rozadilla et al., 2023). The availability of individual-level genotype data together with age at diagnosis makes GECCO well suited for studying whether common susceptibility variants contribute to inter-individual variation in the age-at-diagnosis distribution.

CHAPTER 3

METHODOLOGY

3.1 Modelling age-at-diagnosis

3.1.1 Data preprocessing

Data were obtained from the NCI SEER program via SEER*Stat 9.0.43.0 under a signed Research Data Use Agreement, using the 21-registry grouping (For cases diagnosed from 2000 through the current data year). Exported variables are listed in Table A.1.

All analyses were performed in R (version 4.5.1). Cases were first restricted to anatomically defined CRC sites, after which the following exclusions were applied. Cases reported only through a death certificate or autopsy were removed, as they record age at death rather than at diagnosis. Patients of age 0 years were removed, as these records likely reflect perinatal pathologies rather than typical patterns of CRC diagnoses. For patients with multiple tumours, only the observation regarding the primary tumour was retained to avoid within-person correlation. Finally, the cohort was restricted to Non-Hispanic white people to align with the genetic data set used in downstream analyses. The final study population included 766,723 individuals diagnosed with a primary CRC tumour between 2000 and 2022.

Given the large sample size, formal GOF tests are highly sensitive, with trivial departures from the null yielding extremely small p-values (Wasserman, 2004). Evaluating model performance in the following sections relied primarily on graphical diagnostics to identify meaningful deviations from the null.

3.1.2 Empirical tail characterization

The lower tail of the age-at-diagnosis distribution was first characterized non-parametrically to establish a distribution-free reference. Empirical quantiles were estimated across lower-tail probabilities ranging from $p = 10^{-6}$ to $p = 0.50$. Right-censoring at age 90 was handled by replacing censored ages with ∞ , preserving the ordering for quantiles up to the median. Uncertainty was quantified using 1,000 non-parametric bootstrap resamples to obtain 95% confidence intervals. Sex-stratified quantiles were also estimated to establish a covariate-specific empirical reference against which the corresponding parametric models could be evaluated. Marginal and sex-stratified empirical estimates were subsequently used to assess model performance.

3.1.3 AFT framework

Distribution selection

In a parametric survival model, the outcome is assumed to follow a fully specified distributional family. The AFT model is expressed on the log scale (Eq. 2.1), with ε_i the error term, σ its scale parameter, and $\exp(\beta)$ a multiplicative time ratio associated with covariate x_i . So, the choice of distribution for ε_i determines the shape of the age-at-diagnosis distribution, including the behaviour of its lower tail (Kleinbaum et al., 2012).

Candidate distributions were Weibull, gamma, Erlang, log-normal, log-logistic, exponential, and generalized gamma. They were fitted to the data using the `fitdistrplus` (version 1.2-6) and `flexsurv` (2.3.2) package. These distributions were assessed by overlaying normalized histograms of observed ages with fitted conditional densities, rescaled to the observed support as $f(t | t < 90) = f(t)/F(90)$. Models were compared using Akaike Information Criterion (AIC), with nested models additionally evaluated by likelihood ratio (LR) tests, and GOF was assessed using censored QQ and PP plots. Final model selection balanced overall fit, tail behaviour, parsimony, and parameter interpretability. Based on these criteria the Weibull distribution was selected (cfr. Results section), resulting in a Weibull AFT model.

Model fit and quantile estimation

The Weibull AFT regression model assumes a Gumbel error distribution, implying $T \sim \text{Weibull}(\lambda_i, k)$, with conditional density, for $t \geq 0$,

$$f(t | x_i) = \frac{k}{\lambda_i} \left(\frac{t}{\lambda_i} \right)^{k-1} \exp \left[- \left(\frac{t}{\lambda_i} \right)^k \right] \quad (3.1)$$

The Weibull parameters relate to the AFT parametrization through $k = 1/\sigma$ and $\lambda_i = \exp(\mu + \beta x_i)$. Covariates therefore act multiplicatively on the time scale by modifying λ_i , shifting the age-at-diagnosis distribution while leaving the shape parameter unchanged.

Model parameters were estimated by maximum likelihood using the `survival` (3.8-6) package. For the intercept-only model, no covariates are included, so the fitted Weibull distribution directly characterizes the age-at-diagnosis distribution among individuals diagnosed with CRC. Consequently, the distributional assessment described above serves as the primary model diagnostic.

Tail quantiles were derived by inverting the Weibull cumulative distribution function (CDF) (Eq. 3.2), which resulted in the Weibull quantile function (Eq. 3.3). By substituting $\lambda = \exp(\hat{\mu})$ and $k = 1/\hat{\sigma}$, quantiles of the age by which a proportion p of CRC diagnoses occurs could be estimated (Eq. 3.4).

$$F(x) = 1 - \exp \left(- \left(\frac{x}{\lambda} \right)^k \right) \quad (3.2)$$

$$Q(p) = \lambda(-\log(1-p))^{1/k} \quad (3.3)$$

$$\hat{q}_p = \exp(\hat{\mu} + \hat{\sigma} \log(-\log(1-p))) \quad (3.4)$$

Confidence intervals were constructed on the log-quantile scale using the delta method to estimate the standard error of $\log \hat{q}_p$. Given the large sample size, the underlying asymptotic normal approximation was considered appropriate, making the delta method a computationally efficient alternative to bootstrap-based CIs. The resulting CIs were then backtransformed to the original quantile scale. The resulting 95% Wald interval on the age scale is given by:

$$\exp(\log(\hat{q}_p) \pm z_{0.975} \cdot \text{SE}[\log(\hat{q}_p)]) \quad (3.5)$$

From Eq. 3.4, it follows that $\log \hat{q}_p = \hat{\mu} + \hat{\sigma} \log(-\log(1-p))$. The linearity between $\log \hat{q}_p$ and $\log(-\log(1-p))$, results in a straight line when Weibull quantile estimates are plotted against $\log(-\log(1-p))$. In the extreme lower tail, where p is small, $\log(1-p) \approx -p$, so $\log(-\log(1-p)) \approx \log(p)$. Thus, a log-probability axis closely approximates the theoretical Weibull transformation and allows assessment of the model fit in the lower tail, where deviations from linearity indicate discrepancies between the empirical and theoretical quantiles.

Assessment of sex variable

Sex was included as a covariate in the Weibull AFT model. Its contribution was assessed by comparing models with and without the sex term using the AIC, an LR test, and graphical comparison of the fitted quantiles, with a greater emphasis on the latter due to large sample size.

3.1.4 Generalized Pareto distribution

Data transformation

The POT framework is formulated to model upper-tail exceedances (Coles, 2001). Since the scope of this project targets the left tail of the age-at-diagnosis distribution, ages (T) were reflected as $Y = -T$, converting lower-tail into upper-tail modelling. All threshold selection and model fitting was performed on this reflected scale, estimated quantiles were back-transformed to the original age scale for reporting and interpretation. For subsequent analyses the extRemes (2.2-1) package was used (Gilleland et al., 2016).

Threshold selection

Before fitting the GPD to the data, a threshold has to be selected. Threshold selection followed the two-step diagnostic approach of (Coles, 2001). The MRL plot was inspected for the lowest value of u above which the empirical curve becomes approximately linear, as predicted by EVT once the GPD approximation holds. As a complementary diagnostic, the modified scale σ^* and shape ξ were plotted against u across a range of candidate

thresholds, and were expected to stabilise above the threshold at which the GPD limit becomes valid (Coles, 2001). In both diagnostics, the threshold reflects a bias-variance trade-off: too low a threshold biases the GPD approximation, while too high a threshold discards exceedances and inflates estimation variance. The threshold was therefore set as the lowest value consistent with approximate linearity in the MRL plot and stability in both parameters, retaining as many exceedances as possible without violating either diagnostic. Once the threshold is fixed, the scale σ_u is written as σ .

Model fit and quantile estimation

Exceedances above the selected threshold ($Z_i = Y_i - u$ with $Y_i = -T_i$) were modelled as $Z \sim \text{GPD}(\sigma, \xi)$, with the generalized Pareto distribution (Eq. 2.8). Parameters were estimated by maximum likelihood estimation (MLE) using the `fevd` function, which numerically maximizes the GP log-likelihood. Standard errors for $\hat{\sigma}$ and $\hat{\xi}$ were obtained from the observed information matrix, resulting in 95% Wald CIs using the normal approximation of the MLE. This approximation requires $\xi > -0.5$. For $-0.5 > \xi > -1$ the MLE is still consistent but loses its standard asymptotic properties, and the reported Wald intervals are only approximate (Smith et al., 1985).

GOF of the stationary GPD was assessed using visual diagnostics (quantile and probability plots) provided by the `plot.fevd` function.

Tail quantiles of age at diagnosis were obtained by inverting the fitted GP CDF. For a probability p , t_p is the age satisfying $\Pr\{T < t_p\} = p$, the proportion of diagnosed cases below t_p . Conditioning on the early-onset threshold u_t ,

$$p = \Pr\{T < t_p\} = \Pr\{T < t_p \mid T < u_t\} \Pr\{T < u_t\} \quad (3.6)$$

Under the reflection $Y = -T$, $T < t_p$ becomes $Y > y_p$ with $y_p = -t_p$, and $T < u_t$ becomes $Y > u$ with $u = -u_t$. Since the excess distribution has a finite upper endpoint, the fitted GPD is expected to have $\xi < 0$, so only the $\xi \neq 0$ form is considered here. It follows from Eq. 2.11–2.12 that

$$\begin{aligned} p &= \Pr\{Y > u\} \Pr\{Y > y_p \mid Y > u\} = \zeta_u [1 - F_Z(z_p)] \\ &= \zeta_u \left[1 + \frac{\xi z_p}{\sigma} \right]^{-1/\xi} \end{aligned} \quad (3.7)$$

where $z_p = y_p - u$ is the excess and $\Pr\{Y > y_p \mid Y > u\} = 1 - F_Z(z_p)$ by definition of the excess distribution (Eq. 2.7). Writing the excess on its own scale removes the dependence

on u , and the quantile $\hat{z}_p$ is then obtained by as

$$\hat{z}_p = \frac{\hat{\sigma}}{\hat{\xi}} \left[\left(\frac{\zeta_u}{p} \right)^{\hat{\xi}} - 1 \right] \quad (3.8)$$

Here, $\zeta_u = n_u/n$ is the empirical exceedance rate. Back-transforming, $\hat{y}_p = \hat{z}_p + u$, and the corresponding age at diagnosis is $\hat{t}_p = -\hat{y}_p = -(\hat{z}_p + u)$. Standard errors for $\hat{z}_p$ were obtained using the delta method, and 95% CIs were constructed as $\hat{z}_p \pm z_{0.975} \cdot \text{SE}[\hat{z}_p]$.

Assessment of sex as a covariate

Mirroring the AFT analysis, both an intercept-only and sex-covariate extensions were fitted. In standard GP regression, covariates enter the scale only, with ξ held constant. Although this is the conventional specification, sex was also allowed to enter the shape parameter to assess evidence for sex-specific tail behaviour and whether this additional flexibility improved model fit or altered estimated tail quantiles. So, three covariate specifications were considered: (A) sex on the scale parameter only, (B) sex on the shape parameter only, and (C) sex on both parameters simultaneously.

$$\begin{aligned} \mathbf{A}: \quad & Z_i \sim \text{GPD}(\sigma_i, \xi), \quad \sigma_i = \exp(\beta_0 + \beta_1 \text{Sex}_i) \\ \mathbf{B}: \quad & Z_i \sim \text{GPD}(\sigma, \xi_i), \quad \xi_i = \gamma_0 + \gamma_1 \text{Sex}_i \\ \mathbf{C}: \quad & Z_i \sim \text{GPD}(\sigma_i, \xi_i) \end{aligned} \quad (3.9)$$

All three extensions were compared to the intercept-only model via AIC and LR tests. Model adequacy across all specifications was assessed using probability integral transform (PIT) residuals, which provide a model-comparison diagnostic that operates on a common scale regardless of whether covariates are present. For each exceedance z_i , the PIT value is $u_i = \hat{F}(z_i; \hat{\sigma}_i, \hat{\xi}_i)$, where parameters are evaluated at that observation's covariate values. Under a correctly specified model, $u_i \sim \text{Uniform}(0, 1)$ by the probability integral transform theorem (Casella et al., 2002). To emphasize deviations in the upper tail, PIT values were transformed via $e_i = -\log(1 - u_i)$. For $U \sim \text{Uniform}(0, 1)$ and $E = -\log(1 - U)$, the CDF of E is, for $v \geq 0$,

$$P(E \leq v) = P(-\log(1 - U) \leq v) = P(U \leq 1 - e^{-v}) = 1 - e^{-v} \quad (3.10)$$

which is the CDF of an $\text{Exp}(1)$ distribution. The transformation also increasingly stretches values as u_i approaches 1, making upper-tail deviations more visible. Sorted values were then plotted against theoretical $\text{Exp}(1)$ quantiles, placing all four model specifications on a common axis for direct visual comparison.

As in the AFT framework, the added benefit of including sex as a covariate was negligible (cfr Results section). Therefore, the intercept-only GPD model was selected for subsequent model comparison and final selection.

3.1.5 Model comparison and selection

Model performance was evaluated by comparing the intercept-only quantiles from the Weibull AFT and EVT models against the marginal empirical quantiles at each probability level of interest. Quantile-based diagnostics were used to place both models on a common empirical scale and assess the magnitude and direction of deviations across the tail. Two complementary visualizations were used: quantile fit plots and quantile difference plots. The latter display $\hat{Q}_{\text{model}}(p) - \hat{Q}_{\text{emp}}(p)$ against p on a log-probability scale, with ribbons showing bootstrap uncertainty from the empirical estimates into the difference. Likelihood-based tests were not used to compare the models, as the Weibull and GPD likelihoods are defined on different data domains and are therefore not directly comparable. Formal GOF tests were likewise not used, as their sensitivity to small deviations can make statistical significance difficult to distinguish from practically meaningful differences in fit in large samples. Model selection between the two frameworks was therefore based on the visual diagnostics.

3.2 Application to GECCO cohort

3.2.1 Data preparation

GECCO data acquisition

Genetic data were obtained from the GECCO consortium. Summary statistics are publicly available through the GWAS Catalog, and individual-level data were accessed through dbGaP Authorized Access (phs001315.v1.p1). The dataset comprised four studies of European ancestry: DACHS (Darmkrebs: Chancen der Verhütung durch Screening) was launched in 2003 as a population-based case-control study in southwest Germany to evaluate endoscopic screening and etiologic risk factors for CRC. EPIC (European Prospective Investigation into Cancer) is an ongoing multicenter prospective cohort study examining diet, lifestyle, genetic, and environmental determinants of cancer. HPFS (Health Professionals Follow-up Study) began in 1986 as a prospective cohort study of 51,529 male health professionals aged 40–75, following health-related exposures and CRC outcomes over time. NHS (Nurses' Health Study) was established in 1976 as a prospective cohort study of 121,700 female registered nurses aged 30–55, with biennial follow-up for CRC and other outcomes (Peters et al., 2013).

Candidate SNP panel construction

Given the large number of genotyped single nucleotide polymorphisms (SNPs) relative to the sample size, a literature-based selection strategy was used to reduce the panel to a workable set of candidate SNPs. Candidate SNPs were drawn from two sources: a meta-analysis of overall CRC risk by Fernandez-Rozadilla et al. (2023) and an early-onset CRC

study by Laskar et al. (2024). This literature-derived panel was cross-checked against the SNPs available in the GECCO cohort, and further filtered by removing empty SNPs, duplicates, and SNPs whose chromosomal position did not match the annotation file. Effect alleles were then harmonised against the BIM reference allele, allowing for strand flips and complementary matches. SNPs with unresolved allele mismatches were excluded from the final panel.

Cohort split

The full cohort was then split into a training set, the EPIC subcohort, and a validation set pooling DACHS, NHS, and HPFS. EPIC served as the primary cohort for model fitting and SNP selection. As it was the largest of the four studies and, unlike the single-sex NHS and HPFS, included both women and men, allowing sex to be modelled within a single study.

Initial data analysis

Within the EPIC subcohort, descriptive analyses were performed, and missingness was evaluated. Subjects with missing genotype data were excluded from the analysis. As the analysis models age at diagnosis among individuals who develop cancer, all subsequent analyses are restricted to cases.

3.2.2 Model formulation

The GDP framework used for SEER was applied unchanged to GECCO. Age was reflected as $Y_i = -T_i$, with exceedances $Z_i = Y_i - u$ modelled as $Z_i | x_i \sim \text{GPD}(\sigma_i, \xi_i)$, where x_i denotes a general covariate set for observation i . Model specification followed the same comparison between stationary and non-stationary formulations, allowing covariates to enter the scale, the shape, or both parameters jointly.

The probability of exceeding the threshold was modelled separately. In a stationary model, the exceedance probability ζ_u is the empirical proportion of cases above the threshold. In a non-stationary model, it is instead modelled as a function of covariates via logistic regression, giving a covariate-dependent exceedance probability $\zeta_i = P(Y_i > u | x_i)$. Combining the logistic and GPD components gives :

$$P(Y_i > u + z_i | \text{CRC}, x_i) = \zeta_i \cdot \left(1 + \xi_i \frac{z_i}{\sigma_i} \right)^{-1/\xi_i} \quad (3.11)$$

which back-transforms to age via $T_i = -(Z_i + u)$.

3.2.3 Model Selection

A provisional model selection was first performed to determine whether the genetic effects were better represented through the scale or shape parameter. Because the final SNP panel had not yet been selected, an aggregate PRS constructed from the full candidate panel was

used as a provisional genetic covariate. Candidate models were compared using LR tests and AIC, relative to each other or to the null model. Since 382 exceedances were available in the EPIC cohort, these likelihood-based comparisons were considered informative for model selection. These comparisons indicated that the genetic effect was better represented through the scale parameter, with a stationary shape parameter (data not shown). SNP selection was therefore performed using this scale model. After selection, the selected SNPs were also included as covariates in the shape parameter to assess whether allowing genetic effects on the shape parameter improved the model.

SNP selection

With the scale-only model specification fixed, SNPs were selected from the candidate panel (Section 3.2.1) by forward stepwise selection using AIC. At each step, the SNP yielding the largest improvement in model fit was added until no further inclusion reduced the AIC. To assess the stability of the selected panel, backward elimination was also performed, beginning with the full candidate model and iteratively removing SNPs until no further reduction in AIC was possible. Because the forward-selected model is nested within the backward-retained model, an LR test was performed to evaluate whether the additional SNPs provided a significant improvement to the fit.

Final specification

Following SNP selection, the comparison between stationary and three non-stationary GPD formulations was repeated using the selected SNPs instead of the provisional PRS. Again, the three non-stationary extensions were considered, allowing covariates to enter the model through the scale parameter via the log link, $\sigma_i = \exp(\phi_0 + \sum_k \phi_k X_{k,i})$, through the shape parameter, or in both parameters jointly. The models were compared using LR tests, either against the stationary model or against one another, where relevant. The scale-only specification was selected as the final model (cfr Results section). The final fit was then evaluated with visual diagnostic checks such as PP, QQ, and simulated QQ plots. It was additionally compared to the stationary model using PIT residuals.

3.2.4 Model estimation

Several quantities were derived from the final fitted model to summarise genetic burden and characterise model predictions. These quantities are functions of the fitted coefficients and therefore vary across individuals according to their genotype.

Genetic scale score

Genetic burden was summarised for each individual by the genetic scale score, corresponding to the estimated individual GPD scale parameter. This was obtained from the fitted linear predictor of the GPD scale model (η_i), referred to as the scale index, $\eta_i = \hat{\phi}_0 + \sum_k \hat{\phi}_k X_{k,i}$,

where $\hat{\phi}_k$ are the fitted SNP coefficients. Applying the log link gives the corresponding genetic scale score $\hat{\sigma}_i = \exp(\eta_i)$. Higher genetic scale scores correspond to larger estimated GPD scale parameters and therefore greater genetic burden. The genetic scale score forms the basis for subsequent model-derived quantities, including theoretical endpoints, expected values and unconditional probabilities of early-onset within CRC patients.

Model derived quantities

Under the fitted non-stationary GPD, each individual's genetic scale score implies a distinct upper endpoint on the exceedance scale, $-\hat{\sigma}_i/\xi$, corresponding to a theoretical age endpoint of $u - \hat{\sigma}_i/|\xi|$, and a distinct expected exceedance, $\hat{\sigma}_i/(1 - \xi)$, corresponding to an expected age-at-diagnosis $u - \hat{\sigma}_i/(1 - \xi)$, conditional on early diagnosis. Both endpoint age and expected age therefore vary continuously with the genetic scale score ($\hat{\sigma}_i$) rather than taking a single population-wide value.

The theoretical endpoint age, $u - \hat{\sigma}_i/|\xi|$, and expected age, $u - \hat{\sigma}_i/(1 - \xi)$, were evaluated across the observed distribution of the genetic scale score among exceedances, with representative estimates reported at the 10th, 25th, 50th, 75th, and 90th percentiles. Uncertainty was quantified with non-parametric bootstrapping, to obtain CIs for the difference in endpoint age and expected age between the 10th and 90th percentile of the genetic scale score, summarizing the overall genetic effect across the observed range of genetic burden.

The corresponding logistic exceedance probability, ζ_i , was evaluated at the same percentiles to characterise early-onset risk within CRC cases.

3.2.5 Validation

Missingness in the pooled test cohort was evaluated separately from the training set (EPIC). To assess whether coefficients estimated on the training cohort are applied within their support, the age-at-diagnosis distributions of the training and test cohorts were compared directly. Validation comprised three complementary analyses. First, the generalizability of the conditional age-at-diagnosis distribution given genotype was evaluated among exceedances. The GPD scale coefficients estimated in EPIC ($\hat{\phi}_0, \hat{\phi}_1, \dots, \hat{\phi}_{11}$) were applied to the pooled test cohort, and model fit was assessed using PIT residuals against a bootstrap $\text{Exp}(1)$ envelope.

Second, the ability of the genetic signal associated with early-onset severity to distinguish early-onset from non-early-onset CRC was evaluated in the test cohort. Because the GPD scale parameter is defined only for exceedances, discrimination was instead assessed using the genetic scale score. The training coefficients were applied to each individual of the test cohort to obtain $(\hat{\sigma}_i)$, which provides an individual-level measure of genetically predicted tail severity for all individuals, irrespective of exceedance status. . Discrimination between

early-onset and non-early-onset individuals was then evaluated using the area under the ROC curve (AUC) with the pROC (1.19.0.1) package (Robin et al., 2011).

Finally, directional consistency of individual SNP effects was assessed by fitting an independent GP model with the same SNP panel to the test-cohort exceedances and comparing the signs of the estimated coefficients with those from the training model.

3.2.6 CRC-specificity of the genetic scale score

To evaluate whether the genetic scale score captured CRC-related variation in age at diagnosis rather than general age-associated genetic structure, a case/control interaction analysis was performed. A genetic scale score was calculated for all individuals using the SNP coefficients estimated in the EPIC case model. GPD models including genetic score, affection status, and their interaction were compared with main-effects models using an LR test. The analysis was performed in EPIC and repeated in the independent test cohort using the EPIC coefficients.

CHAPTER 4

RESULTS

4.1 Modelling age-at-diagnosis distribution

In this section the SEER database is used to evaluate which parametric framework best characterizes the lower tail of the age-at-diagnosis distribution by comparing parametric model quantile estimates with non-parametric estimates.

4.1.1 Exploratory data analysis

After sequential filtering (Tab. A.2), the final study population comprised 766,723 non-Hispanic white individuals diagnosed with a primary CRC between 2000 and 2022. The marginal age-at-diagnosis distribution was left-skewed, with a mode at 65 years, and ranged from 4 until >90 (Tab. A.3). A small secondary peak was visible at age 50, consistent with the US screening recommendation age during the study period (Fig. 4.1a). The censored fraction (age > 90) was small (4.1%).

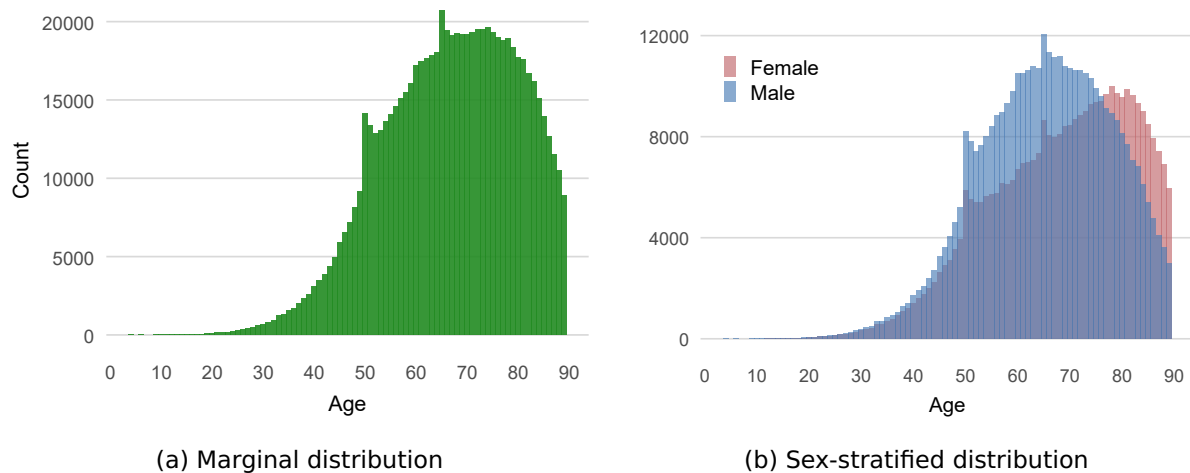


Figure 4.1: Distribution of age at diagnosis in the SEER dataset for subjects diagnosed with CRC (counts).

Age distributions stratified by sex were both left-skewed with a peak at age 50 (Fig. 4.1b). The bulk distributions differed in shape, however, with males reaching a modal diagnosis at 65 and females at 78 years (Tab. A.3). Differences were also present in the lower tail, where males tended to be diagnosed at younger ages than females. Subsequent analyses were therefore conducted both marginally and stratified by sex to assess whether sex-specific modelling was necessary.

4.1.2 Empirical tail characterization

Empirical quantiles, the ages $\hat{q}_p$ by which a proportion p of CRC diagnoses occurred, were estimated across cumulative probabilities from $p = 10^{-6}$ to $p = 0.50$ on the marginal distribution and within each sex stratum. Bootstrap confidence intervals were narrow throughout the central range, widening progressively at the most extreme probabilities as expected from diminishing counts. The minimum observed diagnosis age were captured at the $p = 10^{-6}$ level (Tab. A.4).

Sex-stratified quantiles were closely aligned across most of the probability range, with males diagnosed slightly earlier than females in the tail probabilities (Fig. 4.2). These marginal and sex-stratified empirical estimates served as the reference standard for subsequent model evaluation.

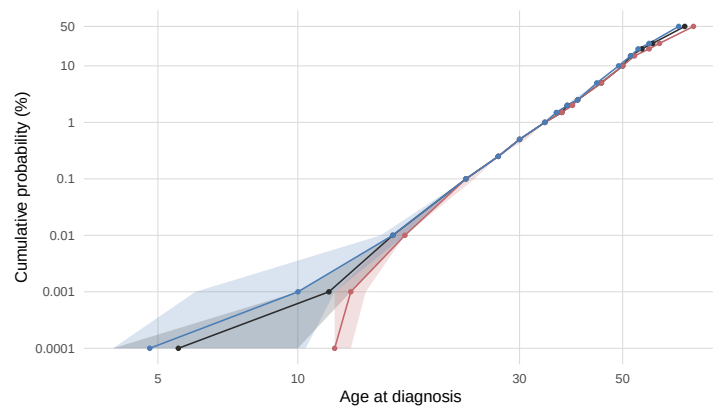


Figure 4.2: Empirical quantiles with 95% bootstrap confidence intervals. Empirical quantiles of the marginal distribution (black), and stratified per gender (pink: female, blue: male). Both axes are on a logarithmic scale: the x-axis shows age at diagnosis and the y-axis shows cumulative probability expressed as a percentage. Shaded bands indicate 95% bootstrap CIs.

4.1.3 AFT regression models

Distribution selection

Candidate distributions were assessed by overlaying conditional density curves on the observed age-at-diagnosis histogram (Fig. A.1). The gamma, Erlang, and lognormal distributions were excluded based on underestimation of the lower tail (Fig. A.1). So, further model comparison was restricted to the Weibull and generalized gamma distribution. AIC and an LR test formally favoured the generalised gamma, and a CDF comparison on logarithmic axes confirmed some improvement in the extreme lower tail (Fig. A.2). This gain was small relative to the cost of an additional parameter, and QQ and PP plots showed no meaningful differences elsewhere (Fig. A.3). Because the objective was to assess whether lower-tail quantities can be estimated reliably by extrapolation from a conventional bulk

model, rather than to identify the best-fitting bulk distribution, the Weibull model was retained for subsequent analyses.

Model fit and quantile estimation

The intercept-only Weibull model describes the marginal age-at-diagnosis distribution. It yielded a characteristic age at diagnosis of $\hat{\lambda} = 73.4$ years and a shape parameter of $k = 1/\hat{\sigma} = 1/0.179 \approx 5.6$. This is well above $k = 1$, the value corresponding to an exponential distribution with a constant hazard, indicating that the hazard of CRC diagnosis increases substantially with age rather than remaining constant across the lifespan. This is consistent with the established age-related epidemiology of CRC.

Tail quantiles were estimated using the Weibull quantile function (Eq. 3.3) (Tab. A.5). They were visually compared to the empirical quantiles with logarithmic axes to better display tail behaviour. The fitted quantiles agreed closely with the empirical ones in the central range. In the lower tail, however, the parametric model sits above the empirical estimates (Fig. 4.3A). So the model assigns more probability to young ages at diagnosis than observed, overestimating the frequency of early diagnosis.

Assessment of sex as a covariate

Next, it was evaluated whether sex improved the fit. Adding sex as a covariate shifted the estimated characteristic age to $\hat{\lambda} = 75.7$ years, with males giving a multiplicative time ratio of $\exp(\hat{\beta}_{\text{male}}) = 0.942$. This corresponds with a diagnosis roughly 6% earlier than women. AIC favoured this model and the LR test was highly significant. However, because the large sample size, the covariate was judged on its quantile fit rather than on formal significance tests (Fig. 4.3). The improvement was not consistent. For women the covariate model tracked the empirical quantiles more closely than the marginal model down to about the 0.1% level, but deviated more strongly beyond it (Fig. 4.3B). For males the central fit was similar but not better as the intercept-only model, with a marginal gain only in the extreme tail (Fig. 4.3C).

Thus, although the empirical distributions differed clearly between men and women (Fig. 4.1b), adding sex as a multiplicative time shift within the AFT framework did not consistently improve the fit, suggesting that their differences could not adequately be captured by a simple rescaling of age. Therefore, the intercept-only Weibull model was retained for subsequent comparisons.

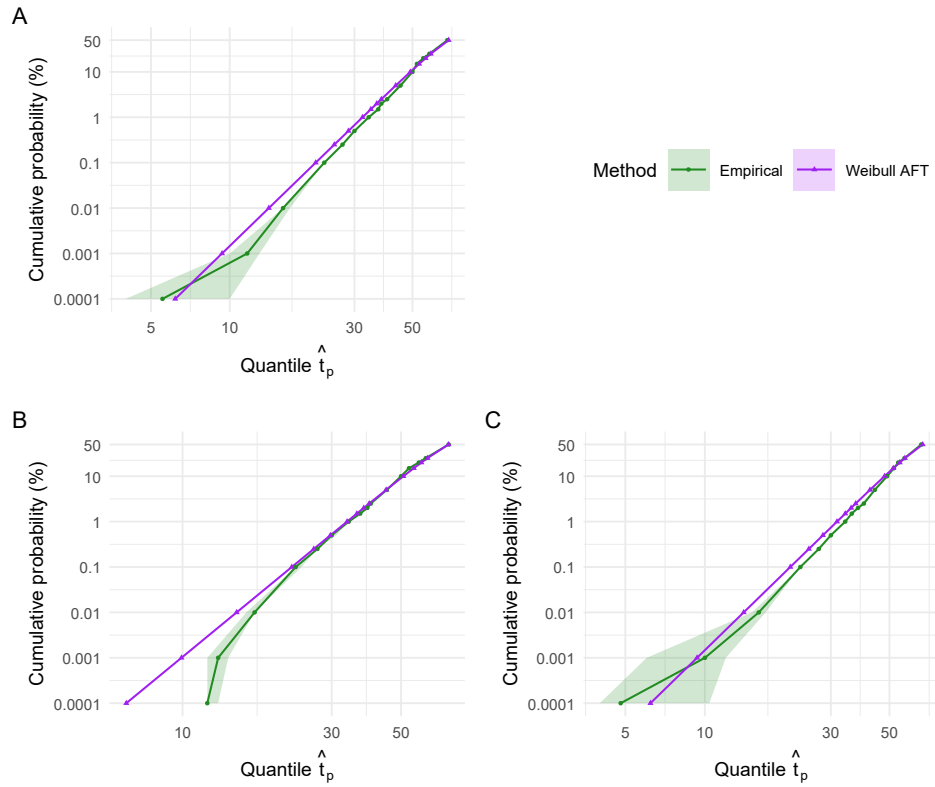


Figure 4.3: Empirical versus Weibull AFT percentiles, by marginal cohort and sex. Shaded bands are 95% delta-method confidence intervals. **A**: The complete cohort is compared to an intercept-only Weibull model, **B & C**: Sex-stratified cohorts are compared to a Weibull model with sex as a covariate, B: female-specific and C: male-specific

4.1.4 Extreme value model

Threshold selection

Candidate thresholds were examined for the reflected age variable Y . The MRL plot shows mean excess values for a range of threshold choices with uncertainty (Fig. 4.4). When evaluating threshold candidates over the full age range, the graph appears to curve from $u = -89$ until approximately $u = -53$, after which there is approximate linearity until $u \approx -15$. This suggests to set $u_0 = -50$. Using this threshold, there are $n_u = 75155$ exceedances that can enter the GPD model. The parameter stability plots (Fig. A.4) were inspected as a complementary check. Across the data-rich thresholds ($u \leq -35$), both the modified scale $\hat{\sigma}^*$ and the shape $\hat{\xi}$ vary smoothly within narrow confidence intervals, and the shape remains negative, confirming a bounded tail in the region of interest. At higher thresholds the estimates become erratic as exceedances thin out, with $\hat{\xi}$ taking occasional positive values that reflect sampling noise from the few youngest cases rather than a genuine change in tail behaviour. Neither parameter is exactly constant; given the large exceedance count, the confidence intervals may be too tight for strict constancy to serve as a discriminating criterion

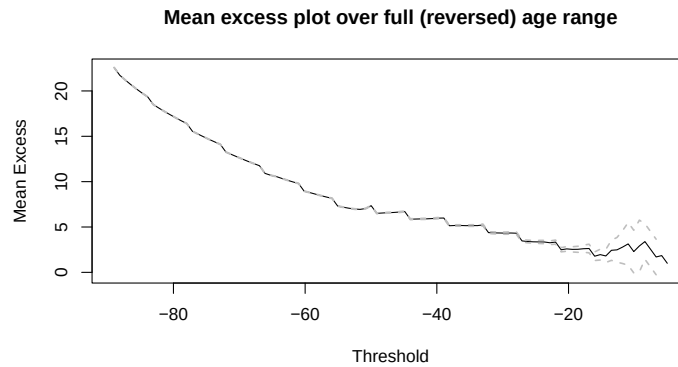


Figure 4.4: Mean residual life plot of reversed age, used to assess threshold stability for the GPD fit.

The threshold was set at $u = -50$, primarily on the basis of the MRL plot, as the lowest value consistent with approximate linearity, with the stability plot confirming a stable negative shape in the reliable region. Within this range, age 50 is additionally supported by screening uptake beginning immediately above this age in this dataset, which would confound the threshold choice with screening effects rather than genuine tail behaviour. Age 50 also coincides with the early-onset definition, so the fitted tail describes precisely these early-onset cases.

Model fit and quantile estimation

Fitting a stationary GPD to the exceedances $Z_i = Y_i - u$ gave a scale of $\hat{\sigma} = 8.73$ (SE 0.030) and a shape of $\hat{\xi} = -0.188$ (SE 0.001). The negative shape parameter places the excess distribution in the type III domain, implying a finite lower bound for age at diagnosis after back-transformation. Using $\hat{z}_+ = -\hat{\sigma}/\hat{\xi}$ gives the upper endpoint of the exceedance distribution, corresponding to a theoretical minimum age of approximately 3.6 years. This value reflects the asymptotic limit of the fitted distribution rather than an expected observation. The estimated shape indicates a relatively short lower tail, with the probability of progressively younger diagnoses declining rapidly towards this bound. The scale parameter, $\hat{\sigma} = 8.73$, sets the spread and expectation of exceedances beyond the threshold.

Diagnostic plots show that the GP stationary fit is adequate across the exceedances. The QQ plot is close to linear, and the PP and density-overlay diagnostics show good agreement (Fig. A.5). This also supports the chosen threshold despite an apparent lack of convergence in the stability plots.

Tail quantiles $\hat{t}_p$ were obtained by inverting the fitted CDF (Eq. 3.8) and back-transforming, with delta-method confidence intervals (Table A.6). Estimated age-at-diagnosis range from 50 years at $p = 0.1$ to approximately 9 years at $p = 10^{-6}$.

Assessment of sex as a covariate

Three covariate specifications (Eqs. 3.9) were evaluated and compared using AIC and LR tests were appropriate. As in the AFT analysis, these tests are interpreted cautiously because of the large n_u . Both AIC and LR tests favoured the covariate models, and the joint model most strongly.

In the joint model, the baseline (female) estimates were $\hat{\beta}_0 = 2.208$ ($\hat{\sigma}_F = \exp(\hat{\beta}_0) \approx 9.1$) and $\hat{\gamma}_0 = -0.213 = \hat{\xi}_F$. The coefficients $\hat{\beta}_1 = -0.062$ and $\hat{\gamma}_1 = 0.033$ give male estimates $\hat{\sigma}_M = \exp(2.208 - 0.062) \approx 8.5$ and $\hat{\xi}_M = \hat{\gamma}_0 + \hat{\gamma}_1 \approx -0.180$. The coefficients are easiest interpreted where each parameter acts. Near the threshold, the GPD is approximately exponential with rate $1/\hat{\sigma}$, so the body of the distribution above u is set by the scale alone. Here $\hat{\sigma}_M$ and $\hat{\sigma}_F$ differ by only about 6% ($\hat{\sigma}_M/\hat{\sigma}_F = 8.55/9.10 \approx 0.94$), so the two sex-specific fits are almost indistinguishable in the region where observations are dense. The shape enters the GPD multiplied by the excess, so its effect scales with how far a point lies above the threshold, so negligible in the bulk and dominant near the endpoint. A small shape contrast $\hat{\gamma}_1 = 0.033$ becomes influential at the theoretical endpoint, $50 - \hat{\sigma}/|\hat{\xi}|$, where it is 7.3 years for females and 2.4 years for males. A difference of nearly five years, driven largely by the shape contrast. The two sex-specific fits thus coincide where the data are informative and diverge at the extreme tail. A similar pattern distinguishes the single-parameter covariate models from the intercept-only fit.

The PIT diagnostic (Fig. 4.5) shows all specifications tracking the diagonal together across the body of the exceedances, separating only in the sparse tail. Since the four fits are equivalent where the data are informative and the covariate alters the fit only in the unconstrained tail, the intercept-only GP model was selected for subsequent comparison, consistent with the AFT analysis.

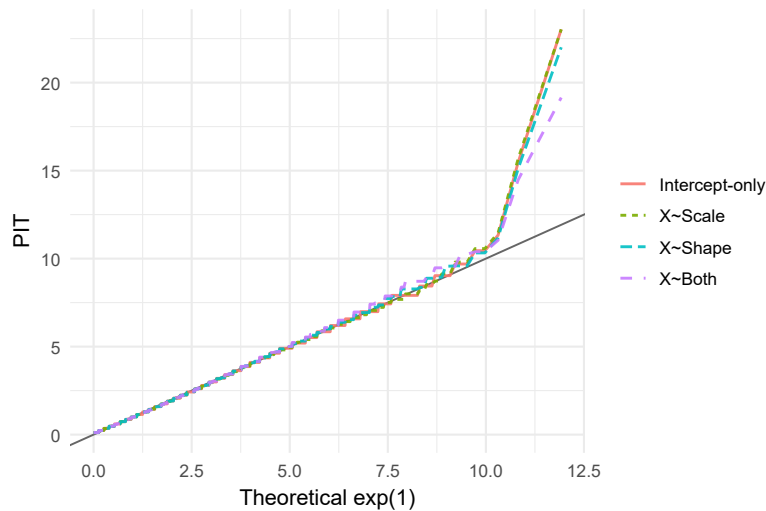


Figure 4.5: Exponential QQ plot of the PIT residuals $-\log(1 - \hat{F}(z))$ for the four GP specifications, plotted against $\text{Exp}(1)$ quantiles. Curves coinciding with the diagonal indicate agreement with the fitted model.

4.1.5 Model comparison and selection

Quantile estimates from the Weibull AFT and GPD models were compared against empirical bootstrap quantiles (Fig. 4.6). From the 10th to the 0.01th percentile, the GPD tracked the empirical quantiles closely, while the Weibull ran one to two years below them. At the most extreme percentile, both models included the reference line but with a wide uncertainty band because of sparse datapoints. Averaged over all probabilities, the GPD sat 0.62 years from the empirical quantiles, against 1.70 for the Weibull. The Weibull underestimated the quantiles for a given low probability, whereas the GPD fitted quantiles fluctuated around the empirical ones.

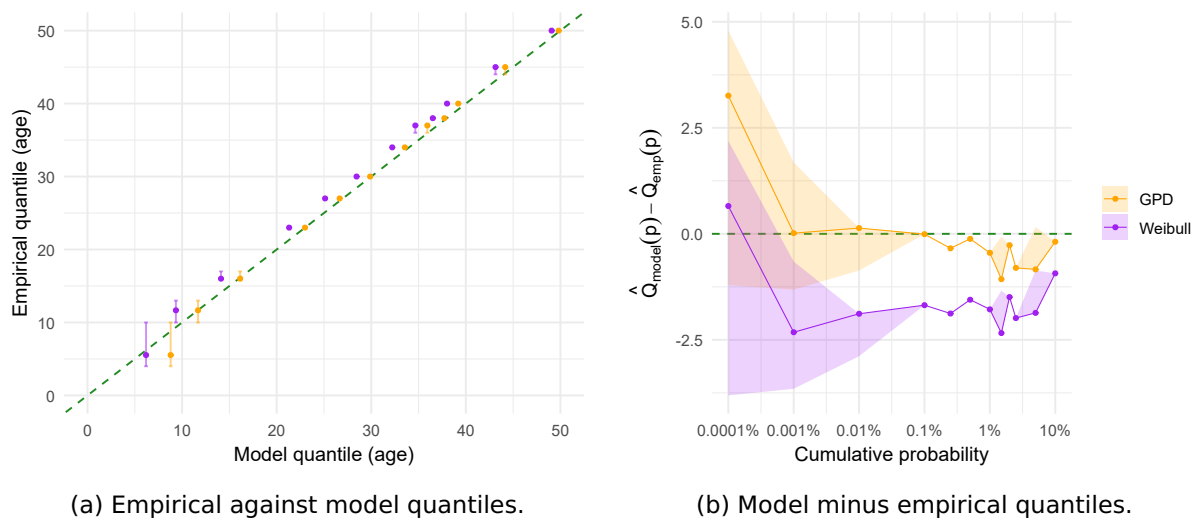


Figure 4.6: Quantile fit of the Weibull and GPD models against the empirical quantiles in the lower tail. Ribbons/bars show 95% bootstrap intervals around the empirical quantiles.

The AFT point estimate departs from the reference line by about 1 to 2 years in one direction across the data-dense range. The GDP point estimates stays around the reference line, including at the most extreme point, where any model extrapolates and the empirical quantile rests on few cases. EVT is therefore the framework to continue with, particularly for the GECCO analyses, which carry far fewer exceedances and no dense empirical tail to expose a bulk model's drift.

4.2 Application to GECCO cohort

In this section, GECCO data are used to extend the GPD model with genetic variants as covariates and evaluate their influence on the lower tail of the age-at-diagnosis distribution, with findings validated in an independent test cohort.

4.2.1 GECCO analysis dataset

The full GECCO cohort included 7,355 participants, of which 3,871 CRC cases, genotyped at 964,193 SNPs. Cross-referencing literature-derived SNP panels with the available genotype

data identified 74 of 205 candidate SNPs. After excluding empty SNPs, duplicates, chromosomal position mismatches, and unresolved allele mismatches, the final analysis panel contained 52 SNPs. The cohort was then split into the EPIC training set and an independent validation cohort pooling DACHS, NHS, and HPFS together.

4.2.2 Initial data analysis

The EPIC subcohort comprised 4,392 participants, including 2,342 controls and 2,127 CRC cases, of the latter 1,137 were female and 990 were male (Tab. A.7). Age-at-diagnosis distributions were left-skewed in both sexes and exhibited similar shapes, with a modal age of 60 years for both (Fig. A.6b). Compared with the SEER cohort, the lower-tail asymmetry was less pronounced and sex-specific differences in the age distributions were smaller. Nevertheless, the distributions retained a clear lower tail, supporting the subsequent extreme value modelling, with model adequacy evaluated using the diagnostics earlier described. Differences between the EPIC and SEER age distributions are expected given their distinct study designs. EPIC is an ongoing prospective cohort study with longitudinal follow-up of enrolled participants (dbGaP, 2016), whereas SEER is a population-based registry that records cancer diagnoses within the general population. These differences in design may influence the observed age-at-diagnosis distribution.

Missingness in the EPIC subcohort was assessed for the outcome and SNP genotypes. The outcome had no missing values. SNP missingness did not exceed 2.7% for any single SNP, and 1.72% of subjects had at least one missing genotype. Complete-case analysis yielded 2,092 cases (973 male, 1,119 female), with a comparable reduction across sexes.

4.2.3 Model selection

Preliminary analysis using a weighted PRS compared covariate effects in the scale, shape, and both parameters jointly. Scale- and shape-only models each improved fit significantly over the stationary model, while the joint model showed no additional improvement over either single-parameter model. As both single parameterisations performed comparably, the scale-only model was used as the conventional choice for SNP selection. Whether the selected SNPs also affected the shape parameter was assessed separately after SNP selection. Forward stepwise selection using AIC identified an 11-SNP panel. As a sensitivity check, backward elimination from the full 52-SNP model retained 51 SNPs. However, with only 382 exceedances, the 51-SNP model exceeds conventional events-per-parameter recommendations, raising concerns about estimation stability. The two models were compared using an LR test, which found no significant improvement in fit from the additional 40 SNPs. The 11-SNP panel was therefore retained as the preferred model, with the SNPs listed in Table A.8.

The selected SNP panel was subsequently evaluated in the scale, shape, and joint parametrisations. Both single-parameter extensions significantly improved fit over the stationary model, null vs scale ($p = 4.049\text{e-}08$), and null vs shape ($p = 1.753\text{e-}10$), whereas allowing the SNPs to enter both parameters did not provide additional improvement over the scale-only specification ($p = 0.222$). The scale-only model was selected as the conventional choice for further analyses. Exponentiated PIT residuals showed improved agreement with the expected distribution in the extreme lower tail relative to the stationary model (Fig. A.7). Standard diagnostic plots indicated no evidence of model inadequacy for the final specification (Fig. A.8).

4.2.4 Model estimation

Estimated model coefficients

Estimated coefficients for the final model (SNPs in scale, constant shape) and their corresponding Wald confidence intervals are reported in Tab. 4.1. Positive coefficients increase the GPD scale parameter, shifting the fitted distribution toward younger diagnosis ages and increasing its spread, whereas negative coefficients shift it toward older ages and reduce spread. To illustrate these effects on the outcome scale, Fig. 4.7 shows the fitted age-at-diagnosis distributions by genotype for each SNP, with the remaining SNPs fixed at their sample mean dosage. For most SNPs, genotype produced visible shifts in the fitted distribution and the corresponding endpoint age.

Table 4.1: Estimated coefficients for the non-stationary GP model coefficients (intercept, ϕ_i , ξ), with the forward-selected 11-SNP panel as covariates in the scale, with their 95% Wald confidence intervals.

Parameter	Estimate	95% lower CI	95% upper CI
Intercept	2.078	1.811	2.345
rs704017_G	0.027	-0.040	0.094
rs16969681_A	0.234	0.165	0.304
rs10821905_A	-0.194	-0.277	-0.112
rs653178_G	-0.083	-0.156	-0.009
rs2070699_A	0.133	0.061	0.205
rs4813802_C	-0.100	-0.167	-0.034
rs1886450_A	0.125	0.047	0.203
rs10936599_A	0.100	0.015	0.185
rs6911915_A	-0.080	-0.145	-0.014
rs497916_A	0.066	-0.008	0.140
rs3087967_A	-0.081	-0.158	-0.003
Shape	-0.576	-0.683	-0.468

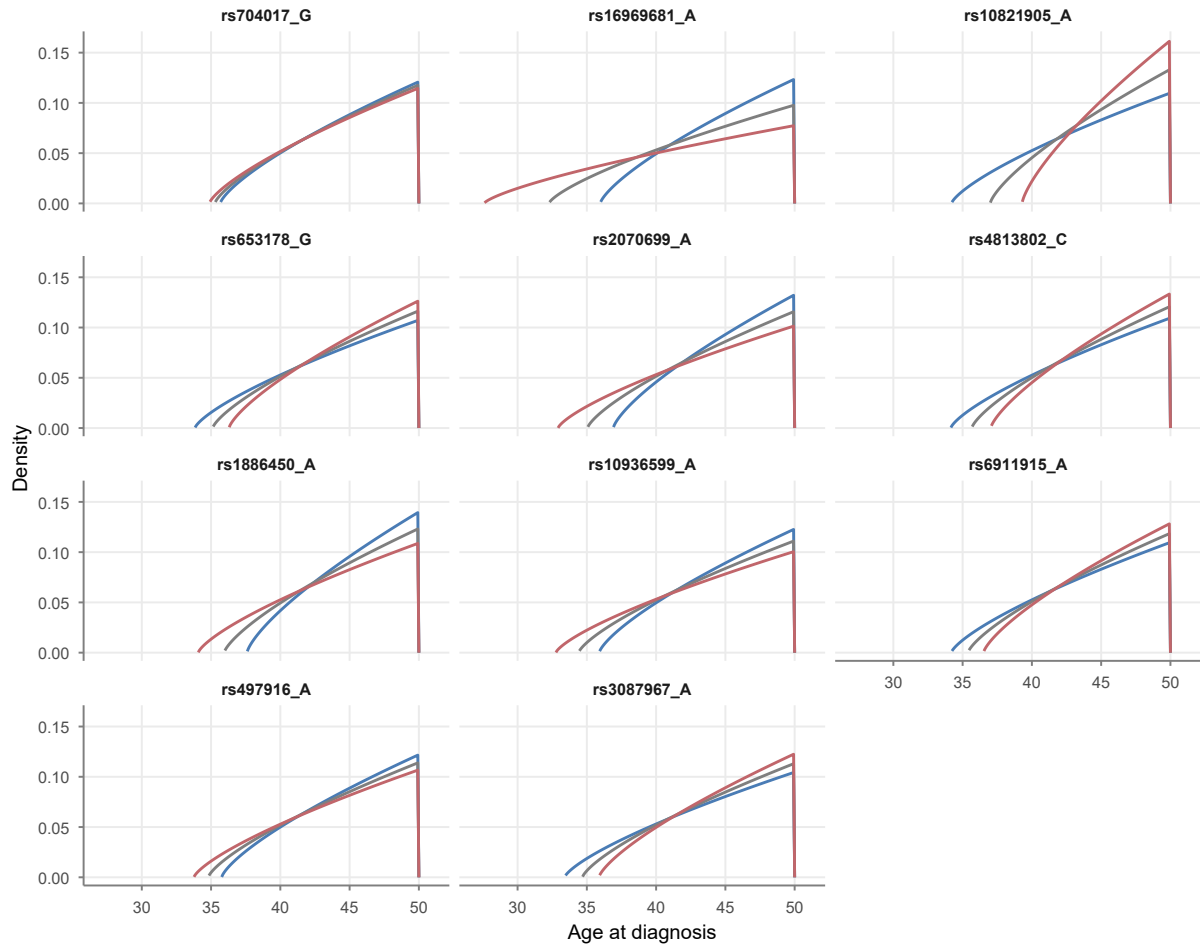


Figure 4.7: Conditional GPD density of age at diagnosis by genotype dosage, for each SNP retained in the forward-selected scale model. For each SNP, the scale parameter was recomputed at dosages 0, 1, and 2 while holding all other SNPs at their sample mean. Blue, grey, and red curves correspond to dosages 0, 1, and 2, respectively.

The fitted model estimated a shape parameter of $\hat{\xi} = -0.57$ (95% Wald CI: -0.68, -0.47), compared with $\hat{\xi} = -0.25$ (95% Wald CI: 0.31, 0.19) for the stationary model. The more negative shape estimate indicates that, after accounting for covariate-dependent variation in the scale parameter, the remaining lower-tail distribution is more strongly bounded than under the stationary model. Because $\hat{\xi}$ lies within the region where the regularity conditions underlying Wald inference no longer hold (Smith et al., 1985), the estimate was further assessed using non-parametric bootstrap resampling. Bootstrap resampling (1,000 iterations) produced a median estimate of -0.63 (95% bootstrap CI: $-0.84, -0.49$). All bootstrap samples converged successfully, and only one estimate (0.1%) was less than or equal to -1 , indicating that the fitted shape parameter was stable under resampling. Although almost all bootstrap estimates lay in the interval $-1 < \xi < -0.5$, where the regularity conditions underlying standard Wald inference do not hold, the close agreement between the Wald and bootstrap CIs suggests that this limitation had little practical impact in the present analysis.

Interpretation of individual SNP coefficients is limited because theoretical endpoint and expected age depend on the combined contribution of all 11 variants. Subsequent analyses therefore summarize each individual's genotype by the fitted genetic scale score $\hat{\sigma}_i = \exp(\hat{\eta}_i)$, combining the estimated effects of all 11 SNPs into a quantity that determines an individual's predicted theoretical endpoint and expected age-at-diagnosis.

Model-derived quantities

The cumulative effect of the selected SNP panel was evaluated at the 10th, 25th, 50th, 75th, and 90th percentiles of the genetic-scale-score distribution. For each percentile, the fitted scale parameter, expected age-at-diagnosis conditional on diagnosis before age 50, theoretical endpoint age, and estimated exceedance probability are reported in Tab. 4.2.

Table 4.2: Fitted scale ($\hat{\sigma}_p$), theoretical endpoint age, expected age at diagnosis conditional on diagnosis before age 50, and exceedance probability ($\hat{\zeta}_p$) for representative individuals at selected percentiles p of the genetic scale score distribution.

Percentile	$\hat{\sigma}_p$	Expected age	Endpoint age	$\hat{\zeta}_p$
10%	6.37	46.0	38.9	0.154
25%	7.27	45.4	37.4	0.166
50%	8.50	44.6	35.2	0.177
75%	9.89	43.7	32.8	0.194
90%	11.51	42.7	30.0	0.203

The expected age-at-diagnosis, conditional on EO-CRC, was 3.3 years lower at the 90th percentile of the scale score compared to the 10th percentile (95% bootstrap CI: 2.5, 4.4). This shift was smaller in magnitude than the corresponding shift in the theoretical endpoint (8.8 years, 95% bootstrap CI: 6.6, 11.0), consistent with the endpoint representing an extreme boundary and the expected value representing a central tendency of the conditional distribution. Both intervals excluded zero, indicating that genetic burden is associated with both a lower typical age at diagnosis and a lower theoretical minimum age.

The estimated exceedance probability also increased across the genetic-scale-score percentiles (Table 4.2). Combining the exceedance probability with the conditional GPD yielded the unconditional probability of diagnosis before age t , within CRC cases (Figure 4.8).

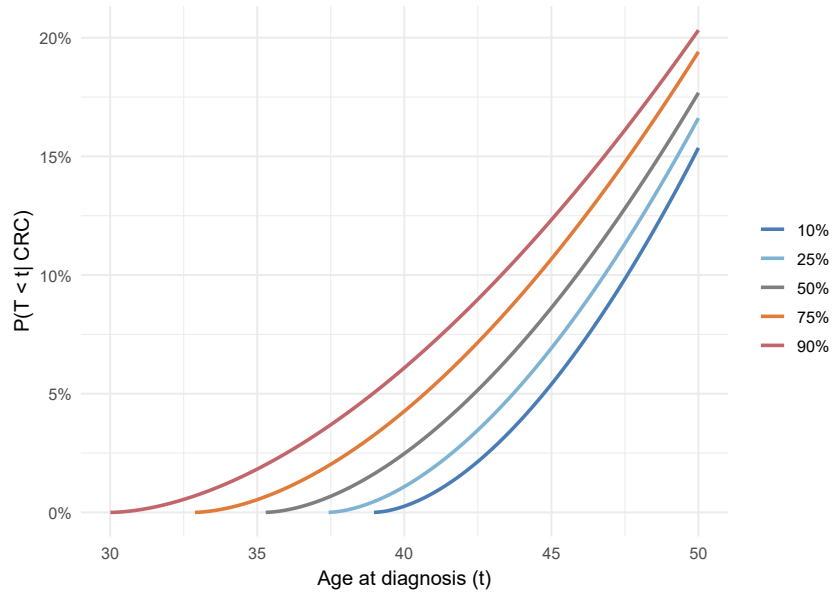


Figure 4.8: Probability of diagnosis before age t , given CRC diagnosis and genetic burden, combining the exceedance probability ($\hat{\zeta}_p$) with the conditional GPD tail for fixed percentiles (10, 25, 50, 75, 90) of the genetic-scale-score distribution.

The cumulative probability curves shifted progressively towards younger ages with increasing genetic scale scores (Figure 4.8). Thus, the association between higher genetic burden and earlier diagnosis observed in the conditional model was not attenuated after incorporating the exceedance probability. Instead, individuals with higher genetic burden were both more likely to be diagnosed before age 50 and, conditional on early diagnosis, to be diagnosed at younger ages.

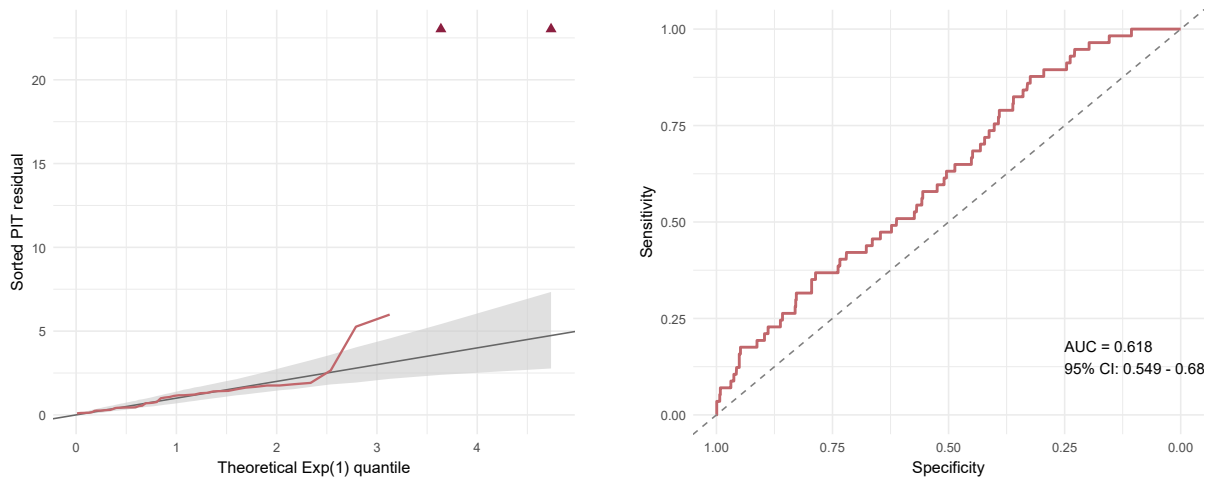
4.2.5 Validation

Missingness in the pooled test cohort was evaluated separately from the training set (i.e. the EPIC subcohort). Roughly 8% of test-cohort subjects had at least one missing SNP dosage, which was higher than the training-set rate, but complete-case exclusion was still applied for consistency with the training set. Support between cohorts was then assessed by comparing age-at-diagnosis distributions (Fig. A.9). The bulk distributions differ between EPIC and the pooled test studies, but the exceedance-based diagnostics reported below depend only on cases below the threshold, so support was additionally compared specifically in that region. The test cohort's age range [26, 49] fell within the training range [28, 49], but resulted in proportionally far fewer exceedances than the training set (57/1603 vs. 382/2092).

PIT residuals from the training GPD scale coefficients tracked the theoretical $\text{Exp}(1)$ diagonal within the bootstrap null envelope for 55 of 57 test exceedances (Fig. 4.9a). The remaining two exceedances, shown as capped points above the plotted range, fell outside the envelope at the uppermost theoretical quantiles. This confirms that the conditional dis-

tribution of age given genotype, estimated in EPIC, generalizes to the pooled test cohort, with departure limited to the two observations.

The genetic scale score discriminated early-onset from late-onset individuals in the pooled test cohort (AUC = 0.618, 95% CI: 0.549-0.688), excluding the null value of 0.5 (Fig. 4.9b). This indicates that a signal selected purely for its association with diagnosis severity among cases also carries some information about whether the threshold is crossed at all.



(a) PIT residuals against the theoretical Exp(1) diagonal, with bootstrap null envelope.

(b) ROC curve for the genetic-scale-score discriminating EO-CRC from LO-CRC individuals.

Figure 4.9: External validation of the frozen model in the pooled GECCO test cohort.

Finally, an independent refit of the 11-SNP GPD on the test cohort showed partial directional stability, with coefficient signs agreeing for 6 of 11 SNPs. One discordant SNP had near-zero coefficients in both models, whereas four showed genuine sign reversals. Despite acceptable overall calibration of the training model, these findings suggest that the combined SNP score is more robust than interpretation of individual SNP effects.

4.2.6 CRC-specificity of the genetic scale score

To assess whether the genetic scale score was associated specifically with CRC onset rather than with age more generally, we first tested whether its effect differed between cases and controls in EPIC. Age at recruitment was used as the tail outcome for controls, since controls have no age at diagnosis, while age at diagnosis was used for cases. Effect modification by case status was assessed using an LR test comparing a model with the “genetic-scale-score x status” interaction to the corresponding main-effects model. The interaction term was significant in EPIC ($\hat{\phi} = 0.665$, $SE = 0.188$, $95\% \text{ CI} = 0.296, 1.0343$, $p = 0.0007$), with a substantially larger genetic-scale-score effect among cases ($\hat{\phi} = 0.817$) than controls ($\hat{\phi} = 0.152$). This supported the interpretation that the genetic scale score was associated with disease-related timing rather than reflecting a general relationship between genetic scale score and age. Then, it was evaluated whether this pattern was replicated in the

independent test cohort. The direction of the effect was consistent, with a larger slope among cases ($\hat{\phi} = 0.639$) than controls ($\hat{\phi} = 0.177$), but “genetic-scale-score x status” interaction was not significant ($\hat{\phi} = 0.462$, $SE = 0.709$, $95\% CI = -0.927, 1.85$, $p = 0.508$). So, the test cohort did not provide independent statistical evidence for effect modification by case status, although its substantially smaller number of exceedances (118 of 2598 versus 867 of 4392 in EPIC) limited the power to detect an effect.

CHAPTER 5

DISCUSSION

Early-onset CRC incidence continues to rise, increasing the need to better understand the factors that influence variation in age at diagnosis among affected individuals (Siegel et al., 2026b). In CRC research, age has predominantly been treated as a predictor rather than as an outcome. Several approaches nevertheless model age directly. Multistage carcinogenesis models use population-level age-specific incidence curves to infer the number of driver mutations required for disease onset, without individual-level covariates (Armitage et al., 1954; Belikov, 2017; Belikov et al., 2021). In genetic studies of Alzheimer's and Parkinson's disease, age of onset has been modelled directly using linear regression or Cox proportional hazards models, typically in case-only analyses (Martin et al., 2021). Ojavee et al. (2022) extend this by modelling age of onset using a Weibull AFT with genetic covariates entering the distribution's parameters, focusing on the bulk of the distribution rather than its tail. EVT provides a complementary approach for extreme ages and has been applied using POT methods to the upper tail of human lifespan (Belzile et al., 2022), without allowing individual-level covariates to modify the GPD parameters. Thus, existing work provides precedents for modelling age with genetic covariates and for applying EVT to extreme ages, but not for combining these approaches to study the lower tail of age at diagnosis in cancer. This study therefore first compares a bulk-distribution model with a POT GPD model for age at diagnosis, before extending the GPD to evaluate the association between genetic burden and earlier diagnosis among EO-CRC cases.

Both the Weibull and generalized gamma provided an adequate description of the bulk age-at-diagnosis distribution, with the generalized gamma achieving a slightly better fit. The Weibull AFT model was nevertheless used as the benchmark bulk model because it is the standard parametric framework for age-of-onset analyses, including genetic applications (Ojavee et al., 2022). Next to that, the slight improvement in bulk fit offered by the generalized gamma does not alter the primary objective of the following comparison, namely evaluating a conventional bulk-distribution model against a tail-specific extreme value model.

Extrapolating a Weibull AFT model fitted to the bulk systematically overestimated the frequency of early diagnosis by 1-2 years, whereas the GPD, fitted directly to threshold exceedances, accurately captured the empirical lower tail. This reflects a general limitation of global parametric models: tail behaviour is constrained by the fit to the full distribution rather than estimated from the extremes themselves, motivating extreme value methods for inference in the tails (Scarrott et al., 2012). These findings provide support extending the GPD with genetic covariates to model the age-at-diagnosis distribution among EO-CRC cases.

Incorporating genetic covariates into the GPD scale parameter revealed a significant association between genetic burden and earlier age at diagnosis among EO-CRC cases. Compared with individuals in the 10th percentile of the genetic scale score, those in the 90th percentile had a fitted theoretical endpoint age 8.8 years younger and an expected age-at-diagnosis 3.3 years younger, indicating that genetic burden is associated with progressively earlier diagnosis within the EO-CRC tail. The same SNP panel was also fitted as a covariate in a separate logistic model for exceedance probability, and ζ_i increased across the genetic scale score, indicating that genetic burden influences both the probability of EO diagnosis within CRC cases and the timing of diagnosis within EO-CRC cases.

The genetic association generalized well to an independent cohort. Calibration of the conditional tail model remained good, with PIT residuals indicating close agreement between observed and expected exceedances, while the genetic scale score also showed modest discrimination between EO- and LO-CRC (AUC = 0.618, 95% CI: 0.549, 0.688). Together, these findings indicate that the learned genetic signal is not confined to the training cohort but captures reproducible variation in age at diagnosis.

The findings are consistent with those of (Archambault et al., 2020), who demonstrated a stronger association between polygenic burden and EO- than LO-CRC. The present analysis extends this observation by modelling a continuous, dose-dependent effect on the age at diagnosis distribution rather than categorizing disease as early or late onset. Proposed PRS-based screening strategies translate genetic burden into personalized screening ages by estimating when individuals reach predefined CRC-risk thresholds (Chen et al., 2023; Tamlander et al., 2024). Their approach is driven by population incidence and becomes increasingly uncertain in younger age groups. The present framework addresses a complementary question by modelling the age-at-diagnosis distribution directly among EO-CRC cases. Although it does not estimate population risk, it characterizes how genetic burden influences disease timing within the clinically relevant early-onset range, suggesting a potential extension for future personalized screening models.

Findings need to be interpreted in the context of the study's limitations. First, the candidate SNP set was limited by the overlap between published CRC susceptibility loci and variants available across the GECCO genotyping platforms. Although more than 200 CRC-associated variants have been reported in recent GWAS, only a subset could be evaluated, as several loci were not represented on the available SNP arrays. In addition, missing genotypes were handled using complete-case analysis rather than conventional genome-wide imputation, resulting in the exclusion of some individuals from the analyses. Future studies would benefit from the inclusion of additional GECCO subcohorts together with standard genotype imputation to improve both genomic coverage and sample size.

Second, external validation was limited by the small number of exceedance cases in the independent cohort, which was sufficient to confirm reproducibility of the genetic-scale-score but not to validate individual SNP effects with adequate precision. In addition, approximately 8% of individuals in the validation cohort had missing data and were excluded from the analysis, rather than being handled through an appropriate missing-data procedure, which may have introduced selection bias. Validation was further complicated by differences in recruitment period between the training and validation cohorts. EPIC reflects a more contemporary population than the pooled validation cohort, which comprises studies initiated several decades earlier. Because EO-CRC incidence has increased markedly over this period, likely owing to changing environmental and lifestyle exposures (Zhao et al., 2023), the effects of genetic variants identified in one cohort may not translate directly to the other. The cohort differences particularly affect the exceedance-probability component, whose baseline depends on the underlying frequency of EO-CRC in the study population. External validation therefore focused on the conditional GPD component, which models age at diagnosis given EO-CRC and is less dependent on differences in baseline incidence between cohorts.

Third, the threshold was fixed at age 50. Although this is also a clinically meaningful cut-off, its robustness was not formally assessed. Sensitivity to neighbouring threshold values and the effect of threshold uncertainty on the parameter estimates were not evaluated. A sensitivity analysis would clarify whether the observed genetic associations are specific to the age 50 threshold or hold more generally across the lower tail.

These limitations also highlight opportunities to refine the genetic component of the proposed framework. The findings of this project motivate to shift variant discovery from disease susceptibility to disease timing. The SNP panel evaluated here was assembled from published CRC susceptibility loci because no comparable age-at-diagnosis resources currently exist (Laskar et al., 2024; Archambault et al., 2020). However, by conditioning on CRC diagnosis the scientific question changes, and variants associated with disease timing are not necessarily the same as those associated with disease occurrence. The scope could extend beyond CRC-risk panels altogether. Early-onset solid cancers have risen in incidence simultaneously, a pattern attributed to shared environmental and lifestyle exposures rather than a single cancer-specific mechanism (Zhao et al., 2023). Whether this shared exposure history reflects shared genetic susceptibility is unclear, but combining variant discovery across early-onset solid cancers, rather than restricting to CRC alone, would be one way to probe this question.

These analyses would benefit from contemporary cohorts assembled within a similar recruitment period and randomly divided into training and validation datasets. The rapid increase in EO-CRC over recent decades suggests that the relationship between inherited susceptibility and disease timing may depend on the environmental context in which those

variants act (Tian et al., 2026). Estimating and validating the model in contemporary cohorts would reduce temporal heterogeneity and improve the transportability of estimated genetic effects to present-day populations.

More broadly, the proposed framework could be extended beyond germline genetic variation to incorporate additional determinants of disease timing. Gene expression, DNA methylation, functional genomic annotations, and other multi-omics features could be included as covariates, while explicitly modelling gene-environment interactions may further clarify the rising incidence of EO-CRC. Recent case-only evidence linking environmental exposures, including herbicides, to EO-CRC supports the importance of considering these interactions (Maas et al., 2026). While incorporating such additional determinants may be valuable from a research perspective, their practical and economic burden could limit their utility for risk stratification compared with a relatively simple blood-based genetic test.

In conclusion, the findings establish GPD as a viable and reproducible framework for modelling age at diagnosis within the clinically relevant early-onset range. By focusing on the lower tail of the distribution, this approach captures genetic associations with disease timing that are not readily identified by bulk-distribution models and provides a foundation for future genetic timing models.

BIBLIOGRAPHY

- Akimoto, Naohiko et al. (2021). "Rising incidence of early-onset colorectal cancer a call to action". In: *Nature Reviews Clinical Oncology* 18 (4), pp. 230–243. ISSN: 17594782. DOI: 10.1038/s41571-020-00445-1.
- Archambault, Alexi N. et al. (2020). "Cumulative Burden of Colorectal Cancer Associated Genetic Variants Is More Strongly Associated With Early-Onset vs Late-Onset Cancer". In: *Gastroenterology* 158 (5), 1274–1286.e12. ISSN: 15280012. DOI: 10.1053/j.gastro.2019.12.012.
- Archambault, Alexi N. et al. (2022). "Risk Stratification for Early-Onset Colorectal Cancer Using a Combination of Genetic and Environmental Risk Scores: An International Multi-Center Study". In: *Journal of the National Cancer Institute* 114 (4), pp. 528–539. ISSN: 14602105. DOI: 10.1093/jnci/djac003.
- Armitage, P and R Doll (1954). "The Age Distribution of Cancer and a Multi-Stage Theory of Carcinogenesis". In: *British Journal of Cancer* 8 (1).
- Belikov, Aleksey V. (2017). "The number of key carcinogenic events can be predicted from cancer incidence". In: *Scientific Reports* 7 (1). ISSN: 20452322. DOI: 10.1038/s41598-017-12448-7.
- Belikov, Aleksey V., Alexey Vyatkin, and Sergey V. Leonov (2021). "The Erlang distribution approximates the age distribution of incidence of childhood and young adulthood cancers". In: *PeerJ* 9. ISSN: 21678359. DOI: 10.7717/peerj.11976.
- Belzile, Léo R. et al. (2021). "Human mortality at extreme age". In: *Royal Society Open Science* 8 (9). ISSN: 20545703. DOI: 10.1098/rsos.202097.
- Belzile, Léo R et al. (2022). "Annual Review of Statistics and Its Application Is There a Cap on Longevity? A Statistical Review". In: DOI: 10.1146/annurev-statistics-040120. URL: <https://doi.org/10.1146/annurev-statistics-040120->.
- Briggs, Sarah E.W. et al. (2022). "Integrating genome-wide polygenic risk scores and non-genetic risk to predict colorectal cancer diagnosis using UK Biobank data: population based cohort study". In: *BMJ* 379. ISSN: 17561833. DOI: 10.1136/bmj-2022-071707.
- Casella, George and Roger L Berger (2002). *Statistical Inference*. Ed. by Carolyn Crockett. 2nd. Duxbury.
- Chen, Xuechen et al. (2023). "Personalized Initial Screening Age for Colorectal Cancer in Individuals at Average Risk". In: *JAMA Network Open* 6 (10), E2339670. ISSN: 25743805. DOI: 10.1001/jamanetworkopen.2023.39670.
- Coles, Stuart (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer London. ISBN: 978-1-84996-874-4. DOI: 10.1007/978-1-4471-3675-0.
- Davison, A C and R L Smith (1990). *Models for Exceedances over High Thresholds*. Tech. rep., pp. 393–442.

- dbGaP (2016). *GECCO: Genetics and Epidemiology of Colorectal Cancer Consortium*. dbGaP Study Accession: phs001315.v1.p1. Accessed: 10Jul2026. URL: <https://dbgap.ncbi.nlm.nih.gov/beta/study/phs001315.v1.p1/#study>.
- Fernandez-Rozadilla, Ceres et al. (2023). "Deciphering colorectal cancer genetics through multi-omic analysis of 100,204 cases and 154,587 controls of European and east Asian ancestries". In: *Nature Genetics* 55 (1), pp. 89–99. ISSN: 15461718. DOI: 10.1038/s41588-022-01222-9.
- Gilleland, Eric and Richard W. Katz (2016). "ExtRemes 2.0: An extreme value analysis package in R". In: *Journal of Statistical Software* 72. ISSN: 15487660. DOI: 10.18637/jss.v072.i08.
- Hamilton, Ashleigh C. et al. (2023). "Distinct Molecular Profiles of Sporadic Early-Onset Colorectal Cancer: A Population-Based Cohort and Systematic Review". In: *Gastro Hep Advances* 2 (3), pp. 347–359. ISSN: 27725723. DOI: 10.1016/j.gastha.2022.11.005.
- Huang, Sida et al. (2025). "Screening History, Stage at Diagnosis, and Mortality in Screen-Detected Breast Cancer". In: *JAMA Network Open* 8 (4). ISSN: 25743805. DOI: 10.1001/JAMANETWORKOPEN.2025.5322.
- Huyghe, Jeroen R. et al. (2019). "Discovery of common and rare genetic risk variants for colorectal cancer". In: *Nature Genetics* 51 (1), pp. 76–87. ISSN: 15461718. DOI: 10.1038/s41588-018-0286-6.
- Jeon, Jihyoun et al. (2018). "Determining Risk of Colorectal Cancer and Starting Age of Screening Based on Lifestyle, Environmental, and Genetic Factors". In: *Gastroenterology* 154 (8), 2152–2164.e19. ISSN: 15280012. DOI: 10.1053/j.gastro.2018.02.021.
- Jones, A. M. et al. (2012). "TERC polymorphisms are associated both with susceptibility to colorectal cancer and with longer telomeres". In: *Gut* 61 (2), pp. 248–254. ISSN: 00175749. DOI: 10.1136/gut.2011.239772.
- Kleinbaum, David G and Mitchel Klein (2012). *Statistics for Biology and Health Survival Analysis A Self-Learning Text Third Edition*. Tech. rep. URL: <http://www.springer.com/series/2848>.
- Laskar, R. S. et al. (2024). "Genome-wide association studies and Mendelian randomization analyses provide insights into the causes of early-onset colorectal cancer". In: *Annals of Oncology* 35 (6), pp. 523–536. ISSN: 15698041. DOI: 10.1016/j.annonc.2024.02.008.
- Law, Philip J. et al. (2019). "Association analyses identify 31 new risk loci for colorectal cancer susceptibility". In: *Nature Communications* 10 (1). ISSN: 20411723. DOI: 10.1038/s41467-019-09775-w.
- Lee, Isaac et al. (2026). "The Increase of Early-Onset Colorectal Cancer: New Insights and Emerging Hypotheses". In: *Cancer Control* 33. ISSN: 15262359. DOI: 10.1177/10732748261432271.
- Li, Shuai and John L. Hopper (2021). *Age dependency of the polygenic risk score for colorectal cancer*. DOI: 10.1016/j.ajhg.2021.02.002.

- Liu, Jimmy Z. et al. (2015). "Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations". In: *Nature Genetics* 47 (9), pp. 979–986. ISSN: 15461718. DOI: 10.1038/ng.3359.
- Lu, Yingchang et al. (2019). "Large-Scale Genome-Wide Association Study of East Asians Identifies Loci Associated With Risk for Colorectal Cancer". In: *Gastroenterology* 156 (5), pp. 1455–1466. ISSN: 15280012. DOI: 10.1053/j.gastro.2018.11.066.
- Maas, Silvana C.E. et al. (2026). "Epigenetic fingerprints link early-onset colon and rectal cancer to pesticide exposure". In: *Nature Medicine*. ISSN: 1546170X. DOI: 10.1038/s41591-026-04342-5.
- Marees, Andries T. et al. (2018). "A tutorial on conducting genome-wide association studies: Quality control and statistical analysis". In: *International Journal of Methods in Psychiatric Research* 27 (2). ISSN: 15570657. DOI: 10.1002/mpr.1608.
- Martin, Eden R., Xiaoyi R. Gao, and Yi Ju Li (2021). "An exploration of genetic association tests for disease risk and age at onset". In: *Genetic Epidemiology* 45 (3), pp. 249–279. ISSN: 10982272. DOI: 10.1002/gepi.22368.
- Mauri, Gianluca et al. (2019). "Early-onset colorectal cancer in young individuals". In: *Molecular Oncology* 13 (2), pp. 109–131. ISSN: 18780261. DOI: 10.1002/1878-0261.12417.
- National Cancer Institute (2026). *Overview of the SEER Program*. URL: <https://seer.cancer.gov/about/overview.html> (visited on 07/15/2026).
- Ojavee, Sven E. et al. (2022). "Novel discoveries and enhanced genomic prediction from modelling genetic risk of cancer age-at-onset". DOI: 10.1101/2022.03.25.22272955.
- Patel, Swati G. et al. (2022). "Updates on age to start and stop colorectal cancer screening: recommendations from the U.S. Multi-Society Task Force on Colorectal Cancer". In: *Gastrointestinal Endoscopy* 95 (1), pp. 1–15. ISSN: 10976779. DOI: 10.1016/j.gie.2021.06.012.
- Peters, Ulrike et al. (2013). "Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis". In: *Gastroenterology* 144 (4). ISSN: 15280012. DOI: 10.1053/j.gastro.2012.12.020.
- Pickands, James (1975). "Statistical inference using extreme order statistics". In: *The annals of Statistics* 3 (1), pp. 119–131.
- Robin, Xavier et al. (2011). "pROC: an open-source package for R and S+ to analyze and compare ROC curves". In: *BMC Bioinformatics* 12, p. 77.
- Scarrott, Carl and Anna Macdonald (2012). "A Review of Extreme Value Threshold Estimation and Uncertainty Quantification". In: *REVSTAT-Statistical Journal* 10 (1), pp. 33–60.
- Siegel, Rebecca L., Nikita S. Wagle, and Ahmedin Jemal (2026a). "Leading Cancer Deaths in People Younger Than 50 Years". In: *JAMA* 335.7, pp. 632–634. DOI: 10.1001/jama.2025.25467.
- Siegel, Rebecca L et al. (2026b). "Colorectal cancer statistics, 2026." In: *CA: a cancer journal for clinicians* 76 (2), e70067. ISSN: 1542-4863. DOI: 10.3322/caac.70067.

- Smith, R and I Weissman (1985). "Maximum likelihood estimation of the lower tail of a probability distribution". In: *Journal Royal Statistical Society* (47), pp. 285–298.
- Su, Yu Ru et al. (2023). "Validation of a Genetic-Enhanced Risk Prediction Model for Colorectal Cancer in a Large Community-Based Cohort". In: *Cancer Epidemiology Biomarkers and Prevention* 32 (3), pp. 353–362. ISSN: 15387755. DOI: 10.1158/1055-9965.EPI-22-0817.
- Syed, Aslam R. et al. (2019). "Old vs new: Risk factors predicting early onset colorectal cancer". In: *World Journal of Gastrointestinal Oncology* 11 (11), pp. 1011–1020. ISSN: 19485204. DOI: 10.4251/wjgo.v11.i11.1011.
- Tamlander, Max et al. (2024). "Genome-wide polygenic risk scores for colorectal cancer have implications for risk-based screening". In: *British Journal of Cancer* 130 (4), pp. 651–659. ISSN: 15321827. DOI: 10.1038/s41416-023-02536-z.
- Tang, Junwei et al. (2024). "Molecular characteristics of early-onset compared with late-onset colorectal cancer: a case controlled study". In: *International journal of surgery (London, England)* 110 (8), pp. 4559–4570. ISSN: 17439159. DOI: 10.1097/JS9.0000000000001584.
- Tian, Ruiyi et al. (2026). "Biological aging and generational shifts in early-onset cancer risk". In: *Nature Medicine*. ISSN: 1546170X. DOI: 10.1038/s41591-026-04448-w.
- Vratanar, Bor and Maja Pohar Perme (2023). "Evaluating cancer screening programs using survival analysis". In: *Biometrical Journal* 65 (7). ISSN: 15214036. DOI: 10.1002/bimj.202200344.
- Wasserman, Larry (2004). *All of Statistics*. Springer.
- Willauer, Alexandra N. et al. (2019). "Clinical and molecular characterization of early-onset colorectal cancer". In: *Cancer* 125 (12), pp. 2002–2010. ISSN: 10970142. DOI: 10.1002/cncr.31994.
- Zhao, Jianhui et al. (2023). "Global trends in incidence, death, burden and risk factors of early-onset cancer from 1990 to 2019". In: *BMJ Oncology* 2 (1). ISSN: 27527948. DOI: 10.1136/bmjonc-2023-000049.
- Zhu, Zhaozhong et al. (2020). "Shared genetic and experimental links between obesity-related traits and asthma subtypes in UK Biobank". In: *Journal of Allergy and Clinical Immunology* 145 (2), pp. 537–549. ISSN: 10976825. DOI: 10.1016/j.jaci.2019.09.035.

APPENDIX A

SUPPLEMENTARY DATA

A.1 Supplementary tables

Table A.1: SEER*Stat variables exported for analysis, mapped to the names used in this study. Variables marked “No” were exported but not used in analysis. SEER defines no CRC-relevant biomarkers; Breslow thickness, PSA, and brain molecular markers pertain to melanoma, prostate, and brain tumours respectively.

SEER*Stat variable	Analysis name	Used
Patient ID	patient_ID	Yes
Age recode with <1 year and 90+	Age_bin	Yes
Single-year age (derived)	Age	Yes
Sex	Sex	Yes
Race recode (W, B, AI, API)	Race	Yes
Race and origin recode (NHW, ...)	Race_origin	Yes
Origin recode NHIA	Origin_NHIA	Yes
Site recode ICD-O-3/WHO 2008	Site	Yes
Year of diagnosis	Year_diagnosis	Yes
Type of Reporting Source	Reporting_source	Yes
Sequence number	Tumor_nr	Yes
Combined Summary Stage (2004+)	Stage04	Yes
Diagnostic Confirmation	Diagn_conf	Yes
Median household income (2023 adj.)	Median_income	No
Rural-Urban Continuum Code	Rural_Urban	No
Grade Clinical (2018+)	Grade_clin18	No
Grade Pathological (2018+)	Grade_pat18	No
Grade Recode (thru 2017)	Grade17	No
Breslow Thickness (2018+)	Breslow18	No
PSA Lab Value (2018+)	PSA18	No
Brain Molecular Markers (2018+)	BMM18	No
PSA Lab Value Recode (2010+)	PSA10	No

Table A.2: Sequential filtering of the SEER dataset based on exclusion criteria described in section 3.1.1

Step	Criterion	Remaining	Removed
Raw extract		1,760,193	
CRC site filter	Anatomically defined CRC sites only	1,426,090	334,103
Age filter	Exclude age at diagnosis = 0	1,426,084	6
Reporting source	Exclude DCO and autopsy-only cases	1,408,871	17,213
Primary tumour	Retain first primary tumour only	1,145,137	263,734
Race/ethnicity	Restrict to Non-Hispanic White	766,723	378,414

Table A.3: Numerical summary of age-at-diagnosis distribution for the full cohort (Marginal) and stratified by sex (Female, Male). Ages are in years.

Group	N	Min	Q1	Median	Mean	Mode
Marginal	766723	4	58	68	67.6	65
Female	366003	12	60	71	69.5	78
Male	400720	4	57	66	65.9	65

Table A.4: Empirical quantiles with 95% bootstrap confidence intervals. Each row corresponds to a cumulative probability level p , reporting the $\hat{q}_p$ quantile of the age-at-diagnosis distribution. Quantiles are shown for the full cohort (Marginal) and stratified by sex (Female, Male).

p	Marginal	Female	Male
1.0e-06	5.53 (4, 10)	12 (12, 13)	4.8 (4, 10.4)
1.0e-05	11.67 (10, 13)	13 (12, 14)	10.01 (6.02, 12.01)
1.0e-04	16 (16, 17)	17 (16, 17)	16 (15.07, 17)
1.0e-03	23 (23, 23)	23 (23, 24)	23 (23, 23)
2.5e-03	27 (27, 27)	27 (27, 27)	27 (27, 27)
5.0e-03	30 (30, 30)	30 (30, 31)	30 (30, 30)
1.0e-02	34 (34, 34)	34 (34, 34)	34 (34, 34)
1.5e-02	37 (36, 37)	37 (37, 37)	36 (36, 36)
2.0e-02	38 (38, 38)	39 (38, 39)	38 (38, 38)
2.5e-02	40 (40, 40)	40 (40, 40)	40 (39, 40)
5.0e-02	45 (44, 45)	45 (45, 45)	44 (44, 44)
1.0e-01	50 (50, 50)	50 (50, 50)	49 (49, 49)
1.5e-01	52 (52, 53)	53 (53, 53)	52 (52, 52)
2.0e-01	55 (55, 55)	57 (56, 57)	54 (54, 54)
2.5e-01	58 (58, 58)	60 (59, 60)	57 (57, 57)
5.0e-01	68 (68, 68)	71 (71, 71)	66 (66, 66)

Table A.5: Estimated lower-tail quantiles $\hat{t}_p$ of age at diagnosis from the Weibull AFT model, with 95% delta-method confidence intervals.

p	$\hat{t}_p$	95% CI
1.0×10^{-6}	6.19	[6.16, 6.22]
1.0×10^{-5}	9.35	[9.31, 9.38]
1.0×10^{-4}	14.12	[14.07, 14.16]
1.0×10^{-3}	21.32	[21.27, 21.37]
2.5×10^{-3}	25.12	[25.07, 25.17]
5.0×10^{-3}	28.45	[28.39, 28.50]
1.0×10^{-2}	32.22	[32.16, 32.27]
1.5×10^{-2}	34.66	[34.61, 34.71]
2.0×10^{-2}	36.51	[36.46, 36.56]
2.5×10^{-2}	38.01	[37.96, 38.07]
5.0×10^{-2}	43.14	[43.09, 43.19]
1.0×10^{-1}	49.07	[49.02, 49.12]

Table A.6: Estimated lower-tail quantiles $\hat{t}_p$ of age at diagnosis with 95% delta-method confidence intervals.

p	$\hat{t}_p$	95% CI
1.0×10^{-6}	8.79	[8.52, 9.06]
1.0×10^{-5}	11.68	[11.46, 11.91]
1.0×10^{-4}	16.14	[15.96, 16.32]
1.0×10^{-3}	22.99	[22.86, 23.13]
2.5×10^{-3}	26.66	[26.54, 26.78]
5.0×10^{-3}	29.88	[29.77, 29.99]
1.0×10^{-2}	33.55	[33.46, 33.65]
1.5×10^{-2}	35.93	[35.85, 36.01]
2.0×10^{-2}	37.73	[37.66, 37.81]
2.5×10^{-2}	39.20	[39.13, 39.27]
5.0×10^{-2}	44.16	[44.12, 44.20]
1.0×10^{-1}	49.81	[49.81, 49.81]

Table A.7: Age at diagnosis by sex, EPIC cases. Counts are prior to complete case exclusion.

Sex	N	Mean	SD	Median	Mode	Min	Max
Female	1137	56.89	8.18	58	60	30	76
Male	990	57.15	8.16	58	60	26	78

APPENDIX A. SUPPLEMENTARY DATA

Table A.8: Forward-selected SNP panel: training-set effect estimates and literature comparison.

SNP	Locus/Gene	$\hat{\phi}$	Reported association	Reference
rs704017	<i>ZMIZ1-AS1 (lncRNA)</i> (10q22.3)	0.027	Associated with CRC	Law et al., 2019
rs16969681	<i>intergenic (SCG5, GREM1)</i> (15q13.3)	0.234	Associated with CRC	Law et al., 2019
rs10821905	<i>intergenic (A1CF, MIX23P2)</i> (10q11.23)	-0.194	Negatively associated with CRC	Fernandez-Rozadilla et al., 2023
rs653178	<i>intergenic (SH2B3, ATXN2)</i> (12q24.12)	-0.083	Associated with multiple immune-related diseases (Crohn's disease, IBD, rheumatoid arthritis, ...)	Liu et al., 2015
rs2070699	<i>intergenic (EDN1)</i> (6p24.1)	0.133	Associated with CRC	Law et al., 2019
rs4813802	<i>intergenic (CASC20, LINC01713)</i> (20p12.3)	-0.100	Associated with CRC risk	Huyghe et al., 2019
rs1886450	<i>intergenic (KLF5, KLF12)</i> (13q22.1)	0.125	Associated with CRC	Lu et al., 2019
rs10936599	<i>TERC</i> (3q26.2)	0.100	Negatively associated with Telomere length , Associated with several cancer types including CRC	Jones et al., 2012
rs6911915	<i>DCBLD1</i> (6q22.1)	-0.080	Negatively associated with CRC	Fernandez-Rozadilla et al., 2023
rs497916	<i>BCL9L, CXCR5</i> (11q23.3)	0.066	Associated with early-onset asthma, Associated with CRC	Zhu et al., 2020; Fernandez-Rozadilla et al., 2023
rs3087967	<i>POU2AF2</i> (11q23.1)	-0.081	Negatively associated with CRC	Law et al., 2019

Table A.9: GPD scale-model coefficients for the CRC-specificity interaction model, EPIC and independent test cohort.

Parameter	Estimate	95% lower CI	95% upper CI
<i>EPIC</i>			
ϕ_0 (intercept)	1.978	1.867	2.088
ϕ_1 (control slope)	0.152	-0.051	0.355
ϕ_2 (case/control diff.)	-0.170	-0.281	-0.058
ϕ_3 (interaction)	0.665	0.296	1.034
Shape (ξ)	-0.299	-0.351	-0.247
<i>Test cohort</i>			
ϕ_0 (intercept)	1.856	1.608	2.105
ϕ_1 (control slope)	0.177	-0.788	1.142
ϕ_2 (case/control diff.)	0.155	-0.145	0.455
ϕ_3 (interaction)	0.462	-0.927	1.852
Shape (ξ)	-0.330	-0.458	-0.201

A.2 Supplementary figures

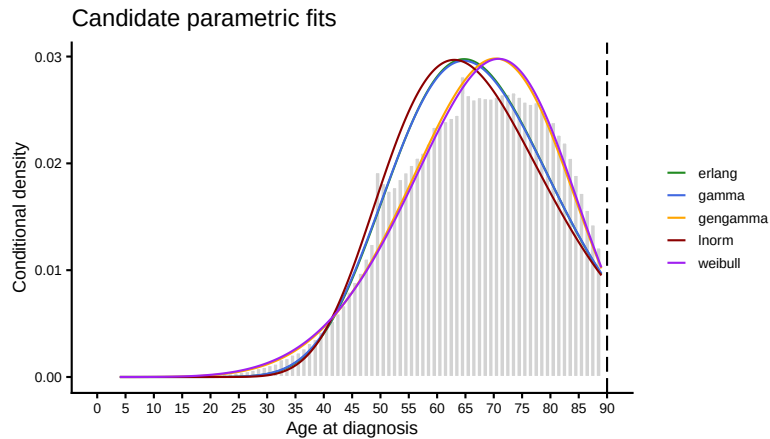


Figure A.1: **Candidate parametric fits to the age-at-diagnosis distribution** Bars show the age-at-diagnosis distribution for observed cases. Curves show the conditional densities $f(x | x < 90) = f(x)/F(90)$ for five candidate distributions fitted by maximum likelihood with right-censoring at age 90. The dashed vertical line marks the censoring threshold.

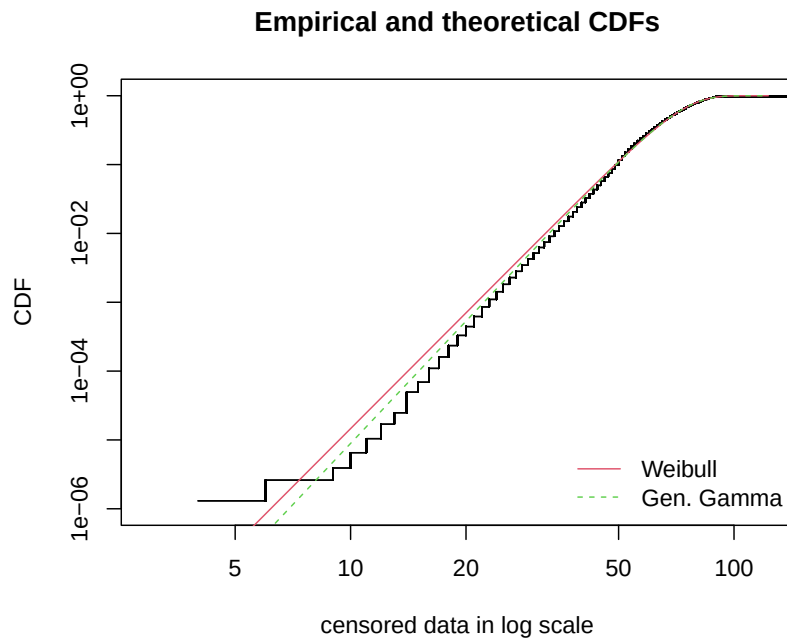


Figure A.2: Empirical and fitted CDFs on logarithmic axes, showing cumulative probabilities down to 10^{-6} . The log scale magnifies the extreme lower tail, where the generalised gamma stays marginally closer to the empirical curve. The two are otherwise interchangeable.

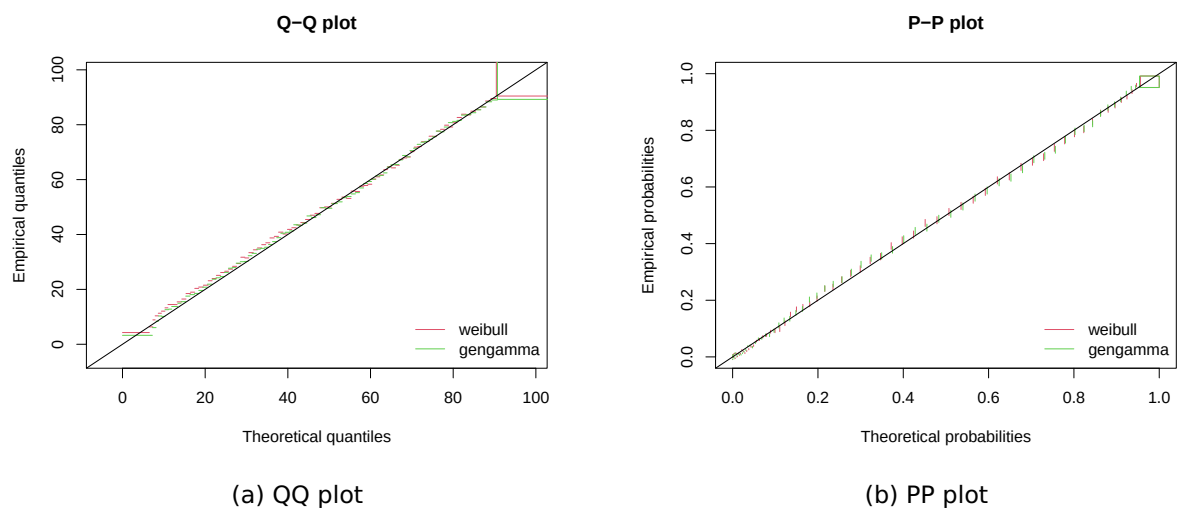


Figure A.3: Goodness-of-fit diagnostics for the Weibull and generalised gamma fits to age at diagnosis. Neither plot shows a meaningful difference between the two distributions.

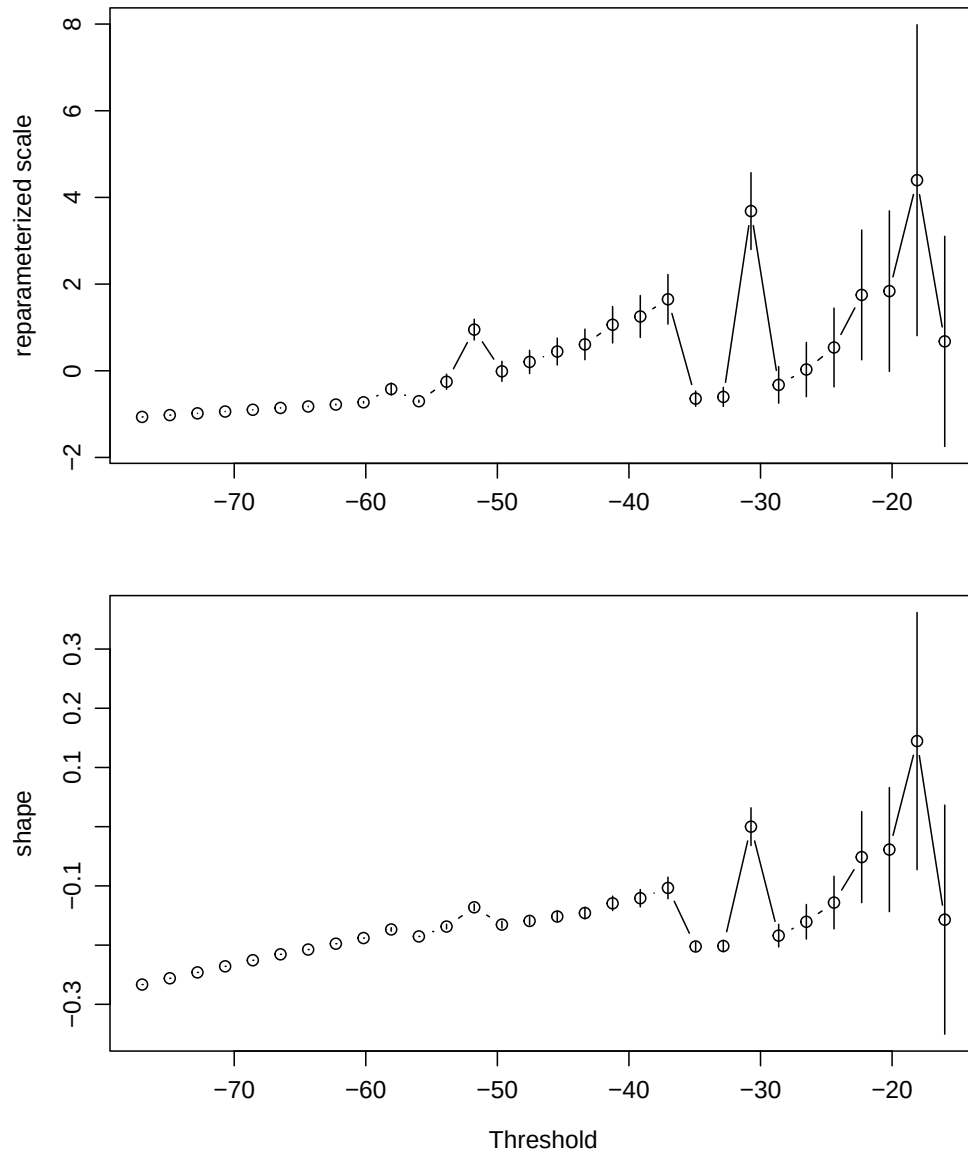
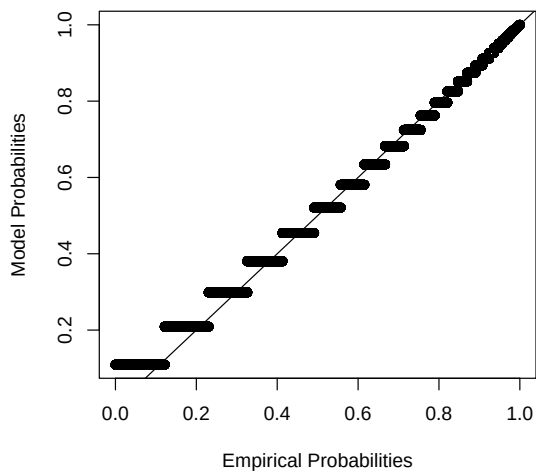
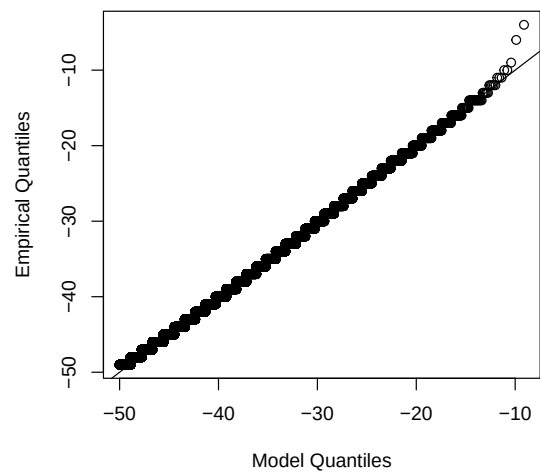


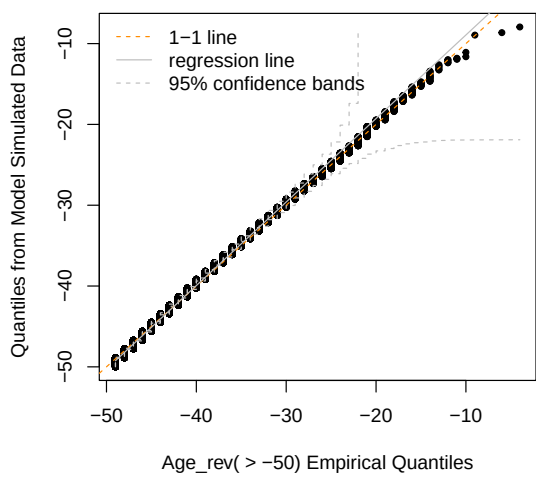
Figure A.4: Stability of the generalised Pareto parameters across candidate thresholds for the reflected age variable. The modified scale (top) and shape (bottom) are estimated at a range of thresholds with 95% confidence intervals; the chosen threshold is the lowest value above which both estimates remain approximately constant, where the asymptotic GPD approximation holds.



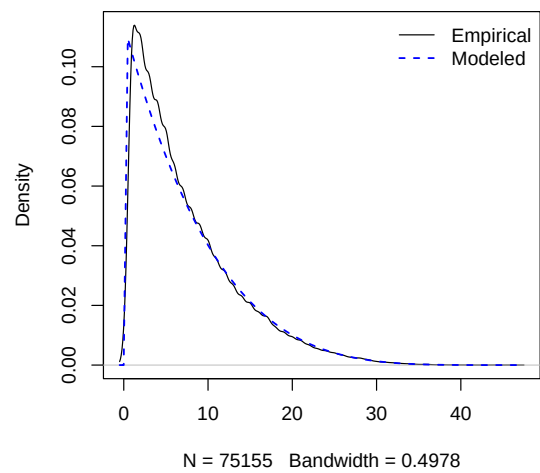
(a) PP plot



(b) QQ plot



(c) QQ plot, model-simulated vs. observed quantiles



(d) Density plot

Figure A.5: GPD diagnostic plots at threshold $u = -50$.

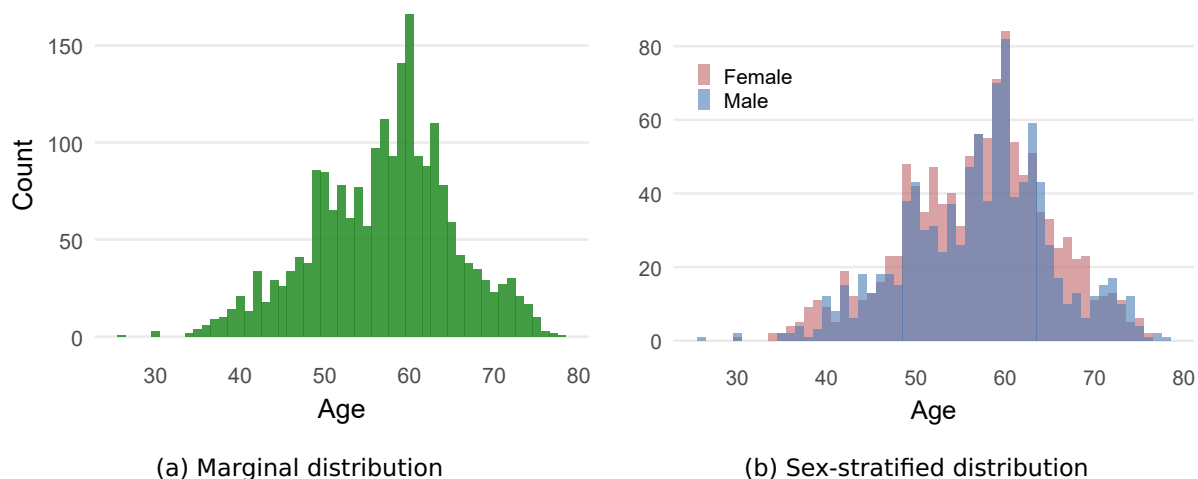


Figure A.6: Distribution of age at diagnosis in the EPIC cohort for subjects diagnosed with CRC (counts).

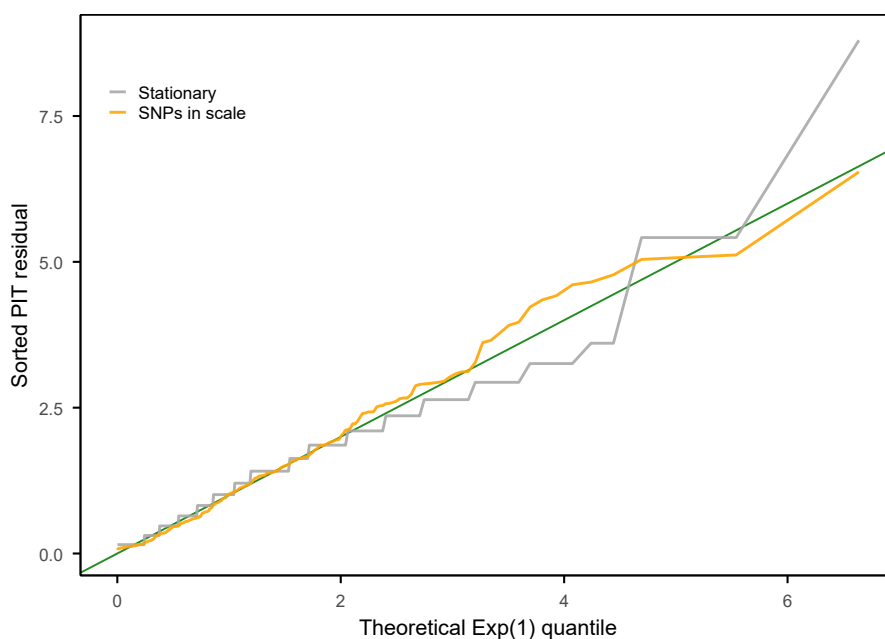
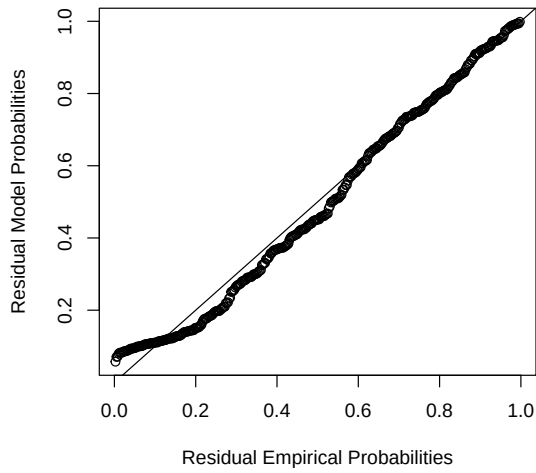
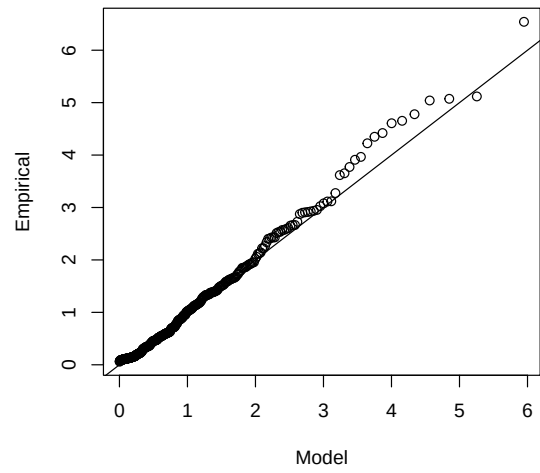


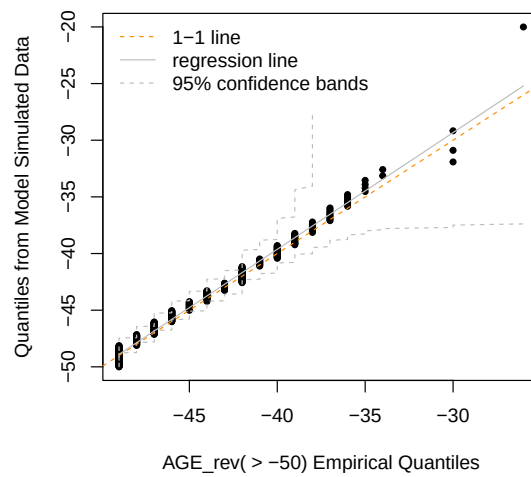
Figure A.7: Exponential QQ plot of sorted PIT residuals against theoretical Exp(1) quantiles, comparing the stationary model against the scale-only model with 11 SNPs.



(a) PP plot



(b) QQ plot



(c) QQ plot, model-simulated vs. observed quantiles

Figure A.8: GPD diagnostic plots for a 11-SNP scale model, fitted on the GECCO EPIC training cohort at threshold $u = -50$.

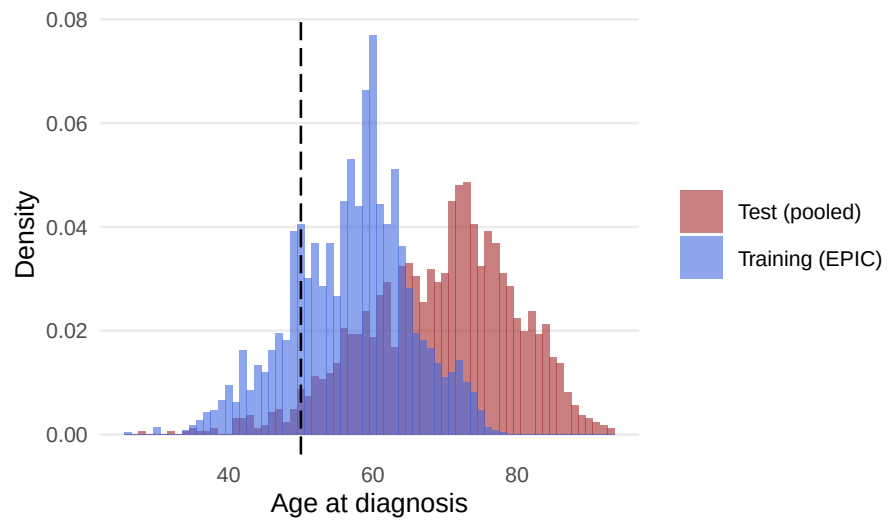


Figure A.9: Density histograms of age at diagnosis in the EPIC training cohort and the pooled test cohort. The dashed line marks the exceedance threshold at age 50.