國立臺灣大學公共衛生學院流行病學與預防醫學研究所

碩士論文

Graduate Institute of Epidemiology and Preventive Medicine

College of Public Health

National Taiwan University

Master Thesis

寇斯與隨機漫步統計模式於動態複雜型排序資料：以糞便免疫潛血濃度為例

Cox and Random Walk Statistical Models for Dynamics of Intractable Ordinal Data: An Example of Fecal Hemoglobin Concentration

彭思敏

Szu-Min Peng

指導教授：陳秀熙　博士

Advisor：Hsiu-Hsi Chen, Ph.D.

中華民國 104 年 05 月

May, 2015

國立臺灣大學碩士學位論文
口試委員會審定書

論文中文題目

寇斯與隨機漫步統計模式於動態複雜型排序資料：以糞便免疫潛血濃度為例

論文英文題目

Cox and Random Walk Statistical Models for Dynamics of Intractable Ordinal Data: An Example of Fecal Hemoglobin Concentration

本論文係 彭思敏 君（學號 R02849033）在國立臺灣大 學流行病學與預防醫學研究所完成之碩士學位論文，於民國 104 年 05 月 29 日承下列考試委員審查通過及口試及格，特此證明。

口試委員：陳秀熙（簽名）

（指導教授）

張淑惠　立政民

致謝

能夠完成這篇論文，我首先要特別感謝我的指導教授陳秀熙老師。謝謝陳老師在學生的兩年碩士中給予的一切幫助與指導，不論是在課業上時常以生活中的範例來解釋生物統計與流行病學，或是在做論文時的耐心指導以及幫助學生做修改等等的鼓勵，感謝陳老師花費了許多的精力與時間教導學生。感謝口試委員張淑惠老師、丘政民老師以及林明薇老師能夠在學生論文口試時撥冗參與，並給予了許多地指正與建議，使我能夠再次以更完善的角度檢視這份論文，同時在相關知識方面帶給我很多的收穫，最後得到了老師們一致的肯定，學生由衷的感恩。

此外還要感謝嚴明芳老師、陳立昇老師、范靜媛老師還有邱月暇老師，謝謝老師們在學生做論文當中給予的種種指導，不厭其煩地傾聽我的問題，從統計到程式編寫上，都一一點出學生的錯誤並且提供很多建議使我能正確地修正論文。同時感謝 533 的所有學長、姊們，在我提出問題時總是陪伴在旁邊給予幫忙，時常地給予鼓勵、分享經驗使得我在碩士兩年間可以快速地進入狀況。

兩年在流預所的日子裡，要感謝同樣是成大幫的小萱，不管是快樂的、忙碌的、無趣的都一起經歷了，並且同樣期待之後也是連體嬰般的職場生活。也謝謝生統碩二的大家，佳純、良珂一起回家不孤單，隔壁桌的芸婕常常一起討論一起進步，碩研室的大家一起生活，做論文再忙也有你們相挺，就算偶爾疲憊也會被激勵出更多的腎上腺素。謝謝因為進入 533 認識的古玫生，加油打氣總是來的剛剛好。感謝管家阿祐打理我的身體健康，總是在忙碌時給予溫暖支持，但也同時要求我投資自己的健康。最重要的還要感謝家人們，給予不匱乏的物資以及精神鼓勵，讓我得以在這兩年內心無旁騖的讀書做研究，感謝你們給的單純幸福。

最後，僅將此論文與各位分享，歡喜相聚，祝福滿滿。

彭思敏　謹致

于臺灣大學流行病學與預防醫學研究所

中華民國 104 年 7 月

中文摘要

背景

糞便潛血濃度(f-Hb) 已證實對於大腸直腸癌的發生率以及死亡率具有極佳的預測力。因此對於在族群篩檢而言，f-Hb 在篩檢時之重複測量數值以及其動態變化對於族群的風險而言亦具有其重要的角色。然而，在運用族群篩檢資料發展描述 f-Hb 變化的模型時，由於其序位型資料特性以及資料中所包含的相關性、設限以及截切等特性，使得模型的建構極為困難。本研究利用有吸收性境界 (absorbing barrier) 之隨機漫步模型(random walk model) 將上述特性納入考量建構描述族群 f-HB 動態變化之模式。

目的

本篇論文第一個目的為利用存活分析的模式評估在不同篩檢組別(正常、大腸腺瘤、大腸直腸癌症) f-Hb 的差異表現，並分別估計並得到族群發生大腸腺瘤以及大腸直腸癌症的糞便潛血濃度數值中位數($f\text{-}Hb_{50}$)，以及其不同的臨界值。本片論文第二個主要的目的為應用隨機漫步模型來量化 f-Hb 濃度的動態變化，並加以考慮在族群發生大腸腺瘤以及大腸直腸癌症時的最大上界值(即觸及吸收境界)的情況。

方法

我們首先利用傳統的單因子變異數分析以及存活分析針對 f-Hb 在不同篩檢組別(正常、大腸腺瘤、大腸直腸癌症)平均數或是中位數的差異進行檢定。接著運用寇斯等比例風險模型(Cox proportional hazards regression model)控制可能的影響變項，並且將資料中的相關特性納入考慮，以序位方式對 f-Hb 數值進行排序，估計各組別(正常、大腸腺瘤、大腸直腸癌症)之對比風險值。配合無母數排序的方法，吾人可以在上述三個組別中計算其糞便潛血濃度數值中位數($f\text{-}Hb_{50}$)，並且分別估

計得到族群發生大腸腺瘤以及大腸直腸癌症的 f-Hb 之臨界值。

在建構動態隨機模型方面，藉由運用隨機漫步模型，並發展基於該模型的漸進分佈(asymptotic distribution) 和多項分佈(multi-nominal distribution) 來描述 f-Hb 重複測量資料變化的進程，並估計 f-Hb 在三種不同的疾病狀態下的數值升高機率(p) 以及降低(q)。進一步可以利用估計得到的機率估計值，計算各組別(大腸直腸癌症或大腸線瘤病患)相對應的賭徒破產機率(即觸及吸收境界之機率)。

結果

利用經過自然對數轉換後的 f-Hb 所作的變異數分析結果中，顯示出三個組別的糞便潛血濃度平均數值達到顯著性的差異 ($F=104324$, $p<0.001$, $R^2=0.142$)，無母數方法檢定的結果顯示同樣顯著差異 ($p<0.001$)。

利用寇斯比例風險模型分析在將其他解釋變相納入調整後(性別、年齡、家族病史以及篩檢工具廠牌)，以篩檢無疾病的人當作比較組，其結果顯示癌症組的風險比是 0.181 (0.178, 0.184)，大腸腺瘤組的風險比為 0.204 (0.202, 0.205)。此估計結果顯示大腸直腸癌個案以及大腸腺瘤個案具有較高的 f-Hb 數值，表示在大腸直腸癌篩檢計畫中，檢測出的糞便潛血濃度越高的人，其後續發展成為大腸腺瘤或大腸直腸癌之風險亦較高。

利用隨機漫步模型結合邏輯斯迴歸所估計得到的結果得到 f-Hb 淨上升機率(drift rate, p-q) 在癌症病患中最高，大腸腺瘤病患次之，最低為無大腸相關疾病的篩檢族群。已僅考慮前進與後退機率的隨機漫步邏輯斯迴歸中為例，在由模型估計的前進機率(p)與後退機率(q)在癌症組中分別為 0.733 及 0.267，在大腸腺瘤組算得的前進與後退機率分別為 0.575 和 0.425，在篩檢後沒有被診斷為大腸疾病的病患的前進機率為 0.358，後退機率為 0.642；因此 f-Hb 上升機率在癌症及線瘤組別中皆為大於 0 的正值，而在正常人則為負值。此外，若與正常族群相較，利用模型與估計結果可以計算癌症族群的在 f-Hb 之上升勝算比為正常族群的 4.92 倍；而

在大腸腺瘤的族群中，此一勝算比是正常人的 2.43 倍。利用模型估計結果計算賭徒破產機率時，若對於癌症設定 f-Hb 值 400 μg/g 為吸收狀態；而大腸腺瘤則以 300 μg/g 為吸收狀態；正常篩檢族群的吸收狀態則訂在 20 μg/g。計算出來的結果在癌症族群中達到吸收狀態機率為 0.867，高於大腸腺瘤組的 0.455，而正常組別則是最低的，其吸收機率幾乎為 0。當假定每個人的起始濃度(x) 為 1 時，平均而言，癌症人期望走 740 步到達 400 μg/g，大腸腺瘤組則須走 893 步到達 300 μg/g。對正常族群而言，達到 f-Hb 濃度 0 μg/g 之吸收狀態的期望步數則為 7.05 步。

結論

本研究運用了寇斯風險比例模式以及建立了隨機漫步迴歸模型以分析具有極端值以及右偏特性的序位資料，模型中亦將由於 f-Hb 值極低而造成的不可量測(左設限)資料，以及遺失值皆納入模型建構之考量。此外，本研究所建構之模型亦包含了多階段疾病特性。

本研究運用所建構的模型於全國大腸直腸癌症篩檢資料，估計了相較於正常族群下，大腸直腸癌族群以及大腸腺瘤族群之高 f-Hb 濃度的風險對比值，同時利用族群 f-Hb 中位數定義各族群之 f-Hb 臨界值。運用隨機漫步模型架構，本研究藉由對於各族群之 f-Hb 上升與下降之估計值結合其淨上升機率以及到達吸收狀態所需步數之計算釐清 f-Hb 隨著時間升高或是降低時有多少破產機率(即有多少達到吸收狀態的機率)，並且估算走到吸收狀態需要的期望步數。本文中的研究結果所建立的新指標，將有助於發展大腸直腸癌族群篩檢計畫決策以及監測規劃。

關鍵字；隨機漫步、賭徒破產、大腸直腸癌症篩檢、糞便潛血、化學免疫法。

ABSTRACT

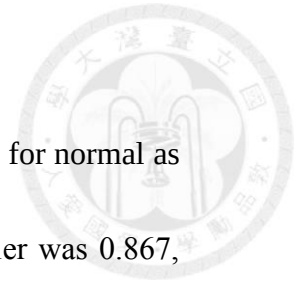
Background As fecal hemoglobin concentration (f-Hb) is a good predictor for colorectal cancer (CRC) incidence and mortality, the dynamics of f-Hb is therefore of great interest in the face of large population-based screening data on periodical examination of f-Hb. Modeling the evolution of f-Hb is intractable as it is an ordinal property and often involves with correlated, censoring, truncating, and dynamic movement with absorbing barriers in the province of the random walk model.

Aims This thesis was first to assess the values of f-Hb across three groups (normal, adenoma, and CRC), estimate the effective median f-Hb concentration ($f\text{-}Hb_{50}$) and its threshold when the adenoma and CRC were detected. The second aim was to apply the random walk model to quantify the dynamic change of f-Hb considering the upper limit because of occurrence of adenoma and CRC.

Methods Conventional survival analysis was employed to test the difference in the mean (or median) value of f-Hb across three groups. The Cox proportional hazards (PH) regression model, making allowance for correlated property, was applied to estimating the hazard ratio (HR) of reaching the ranking of f-Hb across three groups after controlling for relevant covariates. The non-parametric method was used to estimate effective median value of f-Hb ($f\text{-}Hb_{50}$) and the threshold value of f-Hb to hit colorectal adenoma and CRC.

To consider the dynamic (stochastic) property, a random walk model with asymptotic distribution and multi-nominal distribution was further developed to elucidate the evolution (repeated measurement) of f-Hb data to estimate the forward probability (p) and backward probability (q) by three types of diseases status. These parameters were also exploited for calculating the gambler's ruin probabilities of hitting adenoma and CRC.

Results The result of ANOVA shows that the differences in the mean value of f-Hb across three groups were statistically significant. The result of Cox PH regression after adjusting for other covariates (gender, age, family history and brand), compared to the normal group, the HR of the CRC group was 0.181 (0.178, 0.184) and the adenoma group was 0.204 (0.202, 0.205), which suggest that screenee who had higher f-Hb may have higher probability to be diagnosed with disease. The estimated results on the random walk logistic regression model is that the drift rate (p-q) was the highest in the CRC patients followed by adenoma, and the lowest in subjects free of colorectal neoplasia. With the random walk logistics regression model merely considering forward (p) and backward probability, the calculation probabilities gave 0.733 and 0.267 for patents diagnosed as CRC, 0.575 and 0.425 of p and q for patients diagnosed as adenoma, and 0.358 and 0.642 of p and q for the normal subjects. Compared with the normal group, the odds ratio of moving forward was 4.923 for CRC and 2.426 for



adenoma. If we set 400 μg/g for CRC, 300 μg/g for adenoma and 20 μg/g for normal as the absorbing barrier the gambler's ruin probability of reaching the barrier was 0.867, which was higher than 0.455 of adenoma whereas the ruin probability for the normal subject was very low. If the initial value (x) was set 1 it takes, on average, 740 steps for CRC, 893 steps for adenoma, and 7.05 steps for normal to reach absorbing barrier.

Conclusions The thesis has applied the Cox PH regression model and developed a random walk regression model to accommodate the ordinal data with long tail distribution at extremely high value, undetectable circumstance at extremely low value, and missing values and also in relation to multi-state outcome. These proposed models have been applied to nationwide population-based screening for CRC with FIT to estimate the hazard ratio for CRC and adenoma as opposed to the normal subjects, also to estimate the f-Hb_{50} and threshold of developing CRC and adenoma, and get a better understanding of how f-Hb moves forward and backward with time, providing what is the chance of having gambler's ruin (reaching to the barriers of f-Hb) and how many steps are expected to be taken before ruining. These findings provide a new insight into policy-making for colorectal cancer screening and also the surveillance of early-detected colorectal cancer.

Keywords：Random walks, gambler's ruin, colorectal cancer, screening, fecal hemoglobin, FIT.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. Introduction

Modelling ordinal data on quantitative biomarker such as fecal hemoglobin concentration (f-Hb) is very intractable partly because of correlated measurements and partly because of incomplete information (censoring and truncation) problem. In addition, absorption barrier (the upper limit value) also adds to the complexity of such a kind of data.

Very few studies have been conducted before to deal with these statistical issues. One of studies using a random walk model has been conducted to assess the dynamics of score after the administration of endoscopy (Hopper & Young, 1988). However, this study has not evaluated the questions with a formal assessment of such a dynamic ordinal data using the theory of random walk model to report the drift of outcome with unrestricted barrier and the ruin probability with gambler's algorithm (Cox & Miller, 1965).

We are motivated by the recent research on fecal immunological test (FIT) that is widely used in population-based screening for early detection of colorectal cancer and effective in reducing mortality. The application of FIT has extended from qualitative test to quantitative test based on faecal hemoglobin (f-Hb) concentration. The former is to set a cutoff to classify the participants into positive and negative ones. The latter is to make use of quantitative f-Hb from 0 to upper limit of f-Hb concentration. The recent

researches have also demonstrated the quantitative use of baseline faecal hemoglobin (f-Hb) concentration for predicting incident colorectal neoplasia (Chen, Yen, Chiu, Liao, & Chen, 2011**;** Chen et al., 2013) and also colorectal cancer mortality (Chen et al., 2013).

These findings have raised the interest of using quantitative faecal hemoglobin as an ordinal outcome to compare three groups of the underlying population, consisting of free of CRC neoplasia, colorectal adenoma, and colorectal cancer. However, modelling ordinal data such as f-Hb is not straightforward as the distribution is by no means normal distribution and fraught with considerable heterogeneity, including the extreme right values of f-Hb, the outliers of the distribution, and the extreme left undetectable f-Hb that can be treated as left-censored value. To tackle these issues, we treat the order of f-Hb as the outcome of time to event with ranking statistics and apply a Cox proportional hazards regression model to model the difference of f-Hb across three groups (free of CRC neoplasia, colorectal adenoma, and colorectal cancer) with adjustment for other possible covariates.

The first aim of this thesis was to first assess the value of f-Hb across three groups classified by the status of colorectal neoplasia, normal, colorectal adenoma, and colorectal cancer based on a Cox proportional hazards regression model making allowance for left censoring of undetectable f-Hb and interval censoring of f-Hb of

interval cancer. The second major aim of this thesis was to apply the random walk model to quantify the dynamic change of f-Hb considering the absorbing barriers because of occurrence of colorectal adenoma and colorectal cancer.

II. Literature Review

2.1 Theory of Random Walk Model

The random walk is a stochastic process in discrete time. Define a simple random walk as fallow: each jump is +1 with probability p, -1 with probability q, and 0 (no jump) with the probability 1-p-q .

That is,

$$p_{ij} = \begin{cases} p & if\ j = i+1 \\ q & if\ j = i-1 \\ 1-p-q & if\ j = i \end{cases} \qquad (2.1.1)$$

, with $p_{ij} = Pr\{X_n = j | X_{n-1} = i\}$. Where X_n is the position immediately after n jumps, i.e. at time n, $X_n = X_0 + Z_1 + Z_2 + \cdots + Z_n$, Z_i is the moves of in ith jump and $\{Z_i\}$ is a sequence of independently and identically distributed random variables.

There are several types of random walk model that are described as follows.

(1) Unrestricted

We suppose the particle starts at the origin. Also, we assume at time n, the particle reaches the point k. Thus, it has to make r_1 positive jumps, r_2 negative jumps, and r_3 zero jumps. Hence, we have

$$Pr\{X_n = k\} = \sum \frac{n!}{r_1! r_2! r_3!} p^{r_1} (1-p-q)^{r_3} q^{r_2} \qquad (2.1.2),$$

over the value of r_1, r_2 and r_3 satisfying $r_1 - r_2 = k$ and $n = r_1 + r_2 + r_3$.

By the central limit theorem, i.e. if n is large, X_n will be approximately normally

distributed with mean $n\mu$ and variance $n\sigma^2$, with $\mu = p - q$ and $\sigma^2 = p + q - (p-q)^2$. Thus, we can have an approximation equation

$$P(j \le X_n \le k) \cong \Phi\left(\frac{k+c-n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{j-c-n\mu}{\sigma\sqrt{n}}\right) \qquad (2.1.3),$$

c=1/2 or c=1 according to the following condition: $p + q < 1$ or $p + q = 1$.

(2) Two Absorbing Barriers

Suppose the particle ceases when it reaches either $-b$ or a $(a, b > 0)$. We say that absorption occurs at state a (or state $-b$). Define $f_{ja}^{(n)}$ as the probability that the particle is absorbed at a at exactly time n. $f_{ja}^{(n)}$ is also the probability that an unrestricted particles, that is,

$$f_{ja}^{(n)} = P(-b < X_1 < a, \dots, -b < X_{n-1} < a, X_n = a | X_0 = j),$$

$$n = 1,2, \dots \qquad (2.1.4),$$

with the initial value condition $X_0 = j$ when n=0.

Next, we can use the generating function

$$F_{ja}(s) = \sum_{n=0}^{\infty} f_{ja}^{(n)} s^n = F_j(s) \qquad (2.1.5),$$

after the substitution of a trial solution, $F_j(s) = \lambda^j$, the two solutions are

$$\lambda_1(s), \lambda_1(s) = \frac{1 - s(1-p-q) \pm \left[\{1 - s(1-p-q)\}^2 - 4pqs^2\right]^{1/2}}{2ps} \qquad (2.1.6),$$

and

$$\lambda_1 = \frac{q}{p} > \lambda_2 = 1 \qquad (p < q),$$

$$\lambda_1 = 1 > \lambda_2 = \frac{q}{p} \qquad (p > q), \qquad (2.1.7)$$

$$\lambda_1 = 1 = \lambda_2 \qquad (\mathrm{p} = \mathrm{q}).$$

Ruining probability then can be calculated by

$$F_{ja}(s) = \frac{\{\lambda_1(s)\}^{j+b} - \{\lambda_2(s)\}^{j+b}}{\{\lambda_1(s)\}^{a+b} - \{\lambda_2(s)\}^{a+b}} \qquad (2.1.8),$$

set s=1 and let the particle starts at origin then

$$\mathrm{P}(\text{absorption occurs at a}) = F_{0a}(1) = \frac{1-\left(\frac{q}{p}\right)^b}{1-\left(\frac{q}{p}\right)^{a+b}} \qquad (2.1.9)$$

and $\mathrm{P}(\text{absorption occurs at} - \mathrm{b}) = F_{0,-b}(1) = 1 - F_{0a}(1)$. From the formula derived in the Cox and Miller (1965), denote N as the time to absorption, we have the probability distribution of N

$$\mathrm{P}(\mathrm{N} = \mathrm{n}) = f_{0a}^{(n)} + f_{0,-b}^{(n)} \qquad (2.1.10), \text{ and}$$

its generating function

$$\mathrm{E}(s^N) = F_{0a}(s) + F_{0,-b}(s) \qquad (2.1.11).$$

From the Wald's identity, the expected number of steps to absorption is

$$\mathrm{E}(\mathrm{N}) = \begin{cases} \frac{(a+b) - ae^{\theta_0 b} - be^{\theta_0 a}}{e^{-\theta_0 a} - e^{\theta_0 b}} & (\mu \neq 0) \\ \frac{ab}{\mu\sigma^2} & (\mu = 0) \end{cases} \qquad (2.1.12),$$

$\theta_0 = {}^{2\mu}/_{\sigma^2}$ if the steps follow normal distribution.

(3) Two Reflecting Barriers

Suppose the particle starts in the state j and that the state 0 and state a (a>0) are reflecting barriers. Suppose we have $X_0 = j$, and

$$X_n = \begin{cases} X_{n-1} + Z_n \\ a \\ 0 \end{cases} \qquad (2.1.13).$$

Let $p_{jk}^{(n)}$ be the probability that the particle occupies the state k at time n having started in the state j. Assume there is a limiting equilibrium distribution of the state occupation probabilities the we have as $n \to \infty, p_{jk}^{(n)} \to \pi_k$ (k=0,1,…,a). Hence we obtain the truncated geometric distribution

$$\pi_k = \frac{1-p/q}{1-p/q^{a+1}} \left(\frac{p}{q}\right)^k \qquad (k = 0, \dots, a) \qquad (2.1.14).$$

2.2 Re-analysis of Hopper et al study

As mentioned earlier, one of important papers that applies a random walk model for evaluating clinical trials involving serial observations. In a clinical trial, when the status of patients during and after treatment is recorded, analysis of such information will be more convincing. Applications of semi-Markov models have been restricted to diseases with no reverse transitions. The methods of non-parametric inference for these compartmental processes were based on the martingale theory through counting processes.

The alternative is to use the simple random walk that is a stochastic process in discrete time and can be used to deal with the cases where the multistate aspect of disease status may be summarized by an ordinal measure on which patients may improve or regress throughout the clinical trial. With a numerical maximization routine, this method can provide a suitable statistical inference about the efficacy of different treatment regimes. The random walk model was applied to re-analyze the data on two examples.

(1) Example 1: Symptom and endoscopy measures of treatments for peptic oesophagitis

A double-blind trial was conducted on 59 patients with peptic oesophagitis, the goal is to study the efficacy of two treatments (30 controls, 29 Pyrogastone). Scores were recorded on a six-point scale, and recorded at the same epochs (endoscopy scores: 4 weeks for 3 times; symptomatic scores: 2 weeks for 5 times). Table 2.1 gives the estimate of two scores.

The authors used the two logistic models to estimate the change of these two scores, and chosen the most fitted one with log-likelihood.

In the endoscopy scores case, they estimated r=0 in control group.

Here the re-estimation using the unrestricted normal approximation gives the following estimates: $\mu = -0.249$ and $\sigma^2 = 0.227$ for case group in symptomatic scores, and $\mu = -0.162$ and $\sigma^2 = 0.176$ for the control group.

We also calculated the probability $\mathrm{P}(X_6 < 0) = \Phi\left(\frac{0+0.5-6\times(-0.249)}{0.227\times\sqrt{6}}\right) = \Phi(3.59) = 0.9998$, $\mathrm{P}(X_6 < -0.5) = \Phi\left(\frac{-0.5+0.5-6\times(-0.249)}{0.227\times\sqrt{6}}\right) = \Phi(2.69) = 0.9964$.

Table 2.2 shows the results of the probability $\mathrm{P}(X_n < -0.5)$.

Regarding the application to absorbing barriers on symptomatic scores example, we can obtain the ruin probability of different start position j, from 0 to 6 as absorbing state. The ruin probabilities are given in Table 2.3.

For the application on reflecting barriers (state 0 and state 6) on symptomatic scores example, we can obtain the limiting equilibrium distribution π_k given in Table 2.4.

(2) Example 2: Stool frequency as a measure of treatment for colitis

A randomized double-blind trial compared the effect of two drug treatments, bacitracin or vancomycin. Stool frequencies were recorded on eight successive days for 18 patients in each treatment group, and were categorized into 10 levels (level 1 as an absorbing barrier).

From day 0 to day 7, the mean improvement in bacitracin was 2.73±5 0.56 levels, compared to 3.61 ± 0.38 on vancomycin, (P > 0.20).

Table 2.5 shows the estimates of random walk model parameters (with standard errors), and log-likelihood, for bacitracin and vancomycin treatment groups using stool frequency level data.

The results of analysis with the random walk suggested that patients in the bacitracin group show only 58 percent (comparison of E-values) of the improvement in resolution of diarrhoea. The fit of the four models could have different suggestion, while the changes in log-likelihood were not significant. Thus the inference of this example should be carefully.

It should be noted that very few literatures proposed the random walk model to elucidate the dynamics of such an ordinal data like quality of life. Even the paper proposed the random walk model for dealing with the drift of probabilities. There is lacking of formal assessment of computing the ruin probability for reaching the absorbing barrier and the expected steps (time) taken for reach the boundary of the best improved and the worst unimproved states, which will be my major goal of my thesis.

III. Materials

Data on Colorectal Cancer Screening Data

Data we used here are derived from the Taiwanese Nationwide Colorectal Cancer Screening Program using fecal immunochemical test (FIT) as a tool. Details on the planning and implementation of the screening program were reported elsewhere (Chiu et al., 2015). Briefly, the nationwide screening program launched in 2004 was provided to residences of Taiwan aged between 50 to 69 years with a two-year screening interval. The target population consisted of a residency of 5417699 subjects with a staggered entry with the goal of 20% coverage rate set for the initial 5 years. During the study period between January 1, 2004 and December 31, 2009, there were 1160895 attendees with a coverage rate of 21.4% and a repeat screening rate of 28.3%. The fecal hemoglobin concentration of attended were detected by the OC Sensor method by using two brands of commercial kits. A positive test was defined for the given test and those with positive result were referred for confirmatory diagnosis using colonoscopy as a major method. Individual information such as sex, age, family history, and the outcomes of colorectal neoplasm derived from the report confirmatory diagnosis and cancer registry including non-advance adenoma, advanced adenoma (defied as large than 10mm or with villous component) and colorectal cancer were also collected.

Attendees with missing or unidentifiable FIT values or those using unspecified

method for the measurement of fecal hemoglobin were excluded from analysis. The basic characteristics of demographic distribution are listed in Table 3.1 and Table 3.2.The dataset consist of 1031314 screenees and 1265305 repeated measures used for the following analysis.

IV. Methodology

In the thesis, we present analysis of fecal hemoglobin (f-Hb) concentration from the application of conventional statistical approach to the development of new random walk model to demonstrate how f-Hb concentration was heterogeneous with three categories of colorectal neoplasia including normal, adenoma (including non-advanced adenoma and advanced adenoma) and colorectal cancer.

4.1. One-way analysis of variance

Instead of treating the disease status of colorectal neoplasia as the outcome, we treat f-Hb as the outcome of interest and the disease status as the independent variable and test the difference in f-Hb across three categories of disease status with the traditional statistical method, one-way analysis of variance. The null hypothesis is set by

$$H_0: \mu_0 = \mu_1 = \mu_2$$

where μ_0, μ_1, μ_2 represent the mean value of normal, colorectal adenoma, and colorectal cancer. The drawback of using one-way ANOVA is that the result is easily affected by the tail distribution of extreme value.

4.2. Survival Analysis for fecal hemoglobin concentration

It is very interesting in the thesis to consider f-Hb concentration as the ranking data that permits us to consider the use of survival analysis to assess the difference of f-Hb across three or four disease groups with the adjustment for other covariates.

4.2.1 Kaplan-Meyer Method

We therefore first applied the conventional nonparametric method, the Kaplan-Meyer method, to evaluate whether there are differences between colorectal neoplasms, followed by deriving the cumulative distribution curve of f-Hb among different states.

4.2.2 Cox Proportional Hazards Regression Model

Second, we treated the f-Hb of each screenee as the time to event and the disease status as a covariate in Cox proportional hazards regression model. In contrast to survival time, the smaller the f-Hb, the higher the hazard ratio and the lower the risk for developing colorectal neoplasm. By using the method of ties proposed by Breslow (1974), we can deal with the problem of left censoring data with ties resulting from the undetectable f-Hb level.

The reason here that we were not using the exact method for handling the ties was because the population screens cohort contents millions participants, and the sample size was too large for using the exact method. By asymptotic property, the method of ties proposed by Breslow would be expected to be the same as the exact method.

The maximum likelihood estimator of hazard λ_0 in terms of β is given at the same f-Hb concentration (denote f_i) by

$$\widehat{\lambda_1} = {m_i}\Big/\Big((f_i - f_{i-1})\sum_{i\in R_i}\exp\big(\beta' Z_i\big)\Big) \qquad (4.2.1)\ ,$$

where m_i is the number of screenees at f_i while R_i is the set of screenees who were not withdrawn between (0, f_i), i.e. whose f_i higher then f_{i-1}. Z_i here denoted the covariates we used. The underlying cumulative distribution is estimated by

$$\hat{F}(f_i) = \prod_{i=1}^{l}\big(1 - m_i \cdot \ln\sum_{i\in R_i}\exp\big(\beta' Z_l\big)\big) \qquad (4.2.2).$$

Hence the log-likelihood function would be

$$\ln\big(L(\beta)\big) = \sum_{i=1}^{k}\big(\beta' s_i - m_i \cdot \ln\sum_{i\in R_i}\exp\big(\beta' Z_i\big)\big) \qquad (4.2.3)\ ,$$

where s_i is the sum of Z_i over the number at f_i .

Besides, in order to take into account the correlation as a result of repeated screen in population-based screening, we used the method proposed by Lin and Wei (1989), and requested the robust sandwich estimate for the covariance matrix.

4.2.3 Interval Cancers censored at f-Hb

Because interval cancer patients did not have information on f-Hb when diagnosed, which is defined as the censored data, we computed the faecal hemoglobin concentration of interval cancer cases from random samples of the prevalence screen-detected cancer cases and subsequent screen-detected cancer cases by the stratum of gender and age using the cold-deck method, one of conventional methods for dealing with missing data (Rubin, 1987).

4.3. Random Walk Model

It should be noted that although the equation (2.2.1) can be thought to delineate the random process of the dynamic change of f-Hb, the empirical data as indicated in the section of material do not permit us to directly apply this equation to get the estimate of random sum. Most of repeated screens only included two rounds of screen. Based on Markov property, we assume the change of f-Hb from f-Hb at baseline (measured at first screen, i.e. initial location) after n step for each one screenee is equivalent to n jumps based on any of the change of f-Hb between the value of two successive screens (including first screen and second screen) within the same individual or across individual. By using this assumption, define three possibilities of the change among n jumps denoted by the random variable X where X=1, -1, and 0 represent forward movement, backward movement, and no movement to depict the change of f-Hb between (*j*-1) th and *j* th screen. The forward (p), backward probability (q), and no movement (r=1-p-q) of drift are defined by by

$$\begin{cases} \text{p} & \text{, if } \text{fHb}_\text{j} - \text{fHb}_{\text{j}-1} > 0 \text{ (move forward)} \\ \text{q} & \text{, if } \text{fHb}_\text{j} - \text{fHb}_{\text{j}-1} < 0 (move\ backward) \\ \text{r} & \text{, fHb}_\text{j} - \text{fHb}_{\text{j}-1} = 0 \text{ (no movement)} \end{cases} \tag{4.2.4}$$

.

The random variable X among n jumps follows a multinomial distribution denoted as: $\text{X}\sim\text{Multinomial}(\text{n}, \text{p}, \text{q})$.

4.3.1 Unrestricted Random Walk Model

Supposed that sample size (n) is large enough, with the asymptotic property, we proposed to use the normal distribution as the limiting distribution of X_n when estimating the forward and backward probability.

$$X_n \xrightarrow{a} \text{Normal}(n\mu, n\sigma^2) \qquad (4.2.5)\ ,$$

with $\mu = p - q$ and $\sigma^2 = p + q - (p - q)^2$.

Again, we assumed the steps have identical and independent distribution, hence the step of *j*th jump (X_j) follows normal distribution with mean μ, variance σ^2.

The likelihood function is

$$L = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X_j - \mu)^2}{2\sigma^2}\right) \qquad (4.2.6)\ ,$$

and the log-likelihood function is

$$\ln(L) = \sum_{j} -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(X_j - \mu)^2}{2\sigma^2} \qquad (4.2.7)\ ,$$

where n is the number of jumps.

When analysis, we classified the screenees into three groups by their disease statuses: cancer, adenoma and normal.

4.3.2 Random Walk Logistic Regression Model

The ith jump between *j*th and (*j*+1)th screen is denoted by the random variable X_j,

$$X_j = \begin{cases} 1 & , \text{if } fHb_j - fHb_{j-1} > 0 \\ 0 & , \text{if } fHb_j - fHb_{j-1} = 0 \\ -1 & , \text{if } fHb_j - fHb_{j-1} < 0 \end{cases} \quad (4.2.8)$$

Again, $X \sim \text{Multinomial}(n, p, q)$

To model the effect of disease status on the probabilities of movement, we proposed the generalized logistic regression model for estimating the forward, backward, and no movement. We treated the disease status as a covariate that is incorporated into the generalized logistic regression model, through which we can model the moving probabilities among different states in the same time.

Generalized logistic regression model：

$$\text{logit}(p_i) = log\left({p_i}/{1 - p_i}\right) = \alpha_0 + \alpha_1 \cdot SDC_i + \alpha_2 \cdot Advadenoma_i + \alpha_3 \cdot Nonadvdenoma_i + \alpha_4 \cdot IC \quad (4.2.9),$$

$$\text{logit}(q_i) = log\left({q_i}/{1 - q_i}\right) = \beta_0 + \beta_1 \cdot SDC_i + \beta_2 \cdot Advadenoma_i + \beta_3 \cdot Nonadvdenoma_i + \beta_4 \cdot IC \quad (4.2.10),$$

To simplify the generalized logistic regression model, we proposed six scenarios listed as follows.

(i) Combine adenoma group and also the cancer group, that is, let $\alpha_2 = \alpha_3$, and $\alpha_1 = \alpha_4$.

(ii) Combine cancer group, let $\alpha_1 = \alpha_4$.

(iii) Combine adenoma group, let $\alpha_2 = \alpha_3$.

(iv) Combine adenoma group and also the cancer group, that is, let $\alpha_2 = \alpha_3$, and $\alpha_1 = \alpha_4$, and estimates parameters by the two logistic regression models.

(v) Combine the cancer group, let $\alpha_1 = \alpha_4$, and estimates parameters by the two logistic regression models.

(vi) Combine the adenoma group, let $\alpha_2 = \alpha_3$, and estimates parameters by the two logistic regression model.

In the model (i), (ii), and (iii), we combined q and r into q and did estimation based only on the first logistic regression model (4.2.9). In the model (iv), (v), and (vi) we used both regression models (4.2.9) and (4.2.10) and then estimated p, q, and r by different states.

In addition to the analysis of data on screenees who had participated more than one time, we also considered including data on prevalent screenees (who participated in screening once only).

In the prevalence case, we assume who diagnosed as cancer or adenoma would move forward, and those who had screening results as normal cases would either stay on or move backward. Thus we can define the steps of prevalence cases,

$$X_{i0} = \begin{cases} 1 & , \text{ if the ith prevalence case was cancer or adenoma} \\ 0 & , \text{ if the ith prevalence case was normal} \end{cases} \qquad (4.2.11)\ .$$

Noted in the prevalent screen, there are absence of interval cancers. As the results show no movement probability (r) for cancer and adenoma group is relative low, we only used the same logistic regression model (4.2.9) , and set *q=1-p* in the following analysis

We have three scenarios when including prevalent cases.

(i) Combine adenoma group and the cancer group, that is, let $\alpha_2 = \alpha_3$, and $\alpha_1 = \alpha_4$.

(ii) Combine cancer group, let $\alpha_1 = \alpha_4$.

(iii) Combine adenoma group, let $\alpha_1 = \alpha_4$,

After setting up the logistic model for prevalent cases, we can define the probabilities for each state from the coefficients in the regression model.

As regards the estimates based on only the regression model (4.2.9), we have

$$\text{(SDC)} \quad p_1 = \frac{\exp(\alpha_0 + \alpha_1)}{1 + \exp(\alpha_0 + \alpha_1)} \qquad (4.2.12),$$

$$\text{(Adv adenoma)} \quad p_2 = \frac{\exp(\alpha_0 + \alpha_2)}{1 + \exp(\alpha_0 + \alpha_2)} \qquad (4.2.13),$$

$$\text{(Non Adv adenoma)} \quad p_3 = \frac{\exp(\alpha_0 + \alpha_3)}{1 + \exp(\alpha_0 + \alpha_3)} \qquad (4.2.14),$$

$$\text{(IC)} \quad p_4 = \frac{\exp(\alpha_0 + \alpha_4)}{1 + \exp(\alpha_0 + \alpha_4)} \qquad (4.2.15),$$

$$\text{(Normal)} \quad p_5 = \frac{\exp(\alpha_0)}{1 + \exp(\alpha_0)} \qquad (4.2.16),$$

$$\text{and} \quad q_i = 1 - p_i \ , \ i = 1,2,3,4,5 \qquad (4.2.17).$$

Regarding the estimates based on both regression models (4.2.9) and (4.2.10), we have

$$\text{(SDC)} \quad p_1 = \frac{\exp(\alpha_0 + \alpha_1)}{1 + \exp(\alpha_0 + \alpha_1) + \exp(\beta_0 + \beta_1)} \ ,$$

$$q_1 = \frac{\exp(\beta_0 + \beta_1)}{1 + \exp(\alpha_0 + \alpha_1) + \exp(\beta_0 + \beta_1)} \qquad (4.2.18),$$

$$\text{(Adv adenoma)} \quad p_2 = \frac{\exp(\alpha_0 + \alpha_2)}{1 + \exp(\alpha_0 + \alpha_2) + \exp(\beta_0 + \beta_2)} \ ,$$

$$q_2 = \frac{\exp(\beta_0 + \beta_2)}{1 + \exp(\alpha_0 + \alpha_2) + \exp(\beta_0 + \beta_2)} \qquad (4.2.19)$$

$$\text{(NonAdv adenoma)} \ p_3 = \frac{\exp(\alpha_0 + \alpha_3)}{1 + \exp(\alpha_0 + \alpha_3) + \exp(\beta_0 + \beta_3)} \ ,$$

$$q_3 = \frac{\exp(\beta_0 + \beta_3)}{1 + \exp(\alpha_0 + \alpha_3) + \exp(\beta_0 + \beta_3)} \qquad (4.2.20),$$

$$\text{(IC)} \quad p_4 = \frac{\exp(\alpha_0 + \alpha_4)}{1 + \exp(\alpha_0 + \alpha_4) + \exp(\beta_0 + \beta_4)} \ ,$$

$$q_4 = \frac{\exp(\beta_0 + \beta_4)}{1 + \exp(\alpha_0 + \alpha_4) + \exp(\beta_0 + \beta_4)} \qquad (4.2.21),$$

$$\text{(Normal)} \quad p_5 = \frac{\exp(\alpha_0)}{1 + \exp(\alpha_0) + \exp(\beta_0)} \ ,$$

$$q_5 = \frac{\exp(\beta_0)}{1 + \exp(\alpha_0) + \exp(\beta_0)} \quad (4.2.22),$$

and $r_i = 1 - p_i - q_i$, $i = 1,2,3,4,5$ (4.2.23).

Then we can have the likelihood function given k screenee:

for analyses based only on (4.2.9) and also based on (4.2.9) and (4.2.10). Assuming the probabilities applied to first screen are the same as the change of f-Hb at successive screens as indicated in the equation (4.2.4). The likelihood function based on the data on first screen is given as follows.

$$L = \sum_{i=1}^{k} \sum_{j=1}^{n} p^{\sum x_{1i}} \cdot q^{\sum x_{2i} + \sum x_{3i}} \quad (4.2.24),$$

for analyses based on (4.2.9) and (4.2.10),

$$L = \sum_{i=1}^{k} \sum_{j=0}^{n_i} p^{\sum x_{1ij}} \cdot q^{\sum x_{2ij}} \cdot r^{\sum x_{3ij}} \quad (4.2.25),$$

Where

$$x_{1i} = \begin{cases} 1, if\ X_i = 1 \\ 0\ , o.w. \end{cases}$$

$$x_{2i} = \begin{cases} 1, if\ X_i = -1 \\ 0\ ,\ o.w. \end{cases} \quad (4.2.26),$$

$$x_{3i} = \begin{cases} 1, if\ X_i = 0 \\ 0\ ,\ \ \ o.w. \end{cases}$$

The likelihood function for the n jumps of subsequent screens as indicated above was also derived in a similar manner.

With the random walks model and the regression equations we set up, we can estimated the coefficients of variables and calculated the probabilities movement in random walk model.

4.3.3 Gambler's ruin and expected number of game

After the estimation of the probabilities of movement, we can further apply the gambler's ruin theorem. The gambler's ruin problem is the random walks with absorbing barriers 0 and N. A gambler starts out with x f-hb, and he wins 1 unit with probability *p* and lose 1 unit with probability *q=1-p*. The gambler stops when he has a state of 0 or N .

Following the formal derivation of processes for the two absorbing barriers by Cox and Miller (1965), here we use alternative way of deriving the ruin probability.

We are interesting in the computation of probability V_x that the player will be ruined after commencing with x. At the end of the first game (first step analysis), he will has (x+1) if he wins the game with p (V_{x+1}), or he will has (x-1) if he loses the game with q (V_{x-1}). Thus, we have

$$V_x = qV_{x-1} + pV_{x+1} \quad , \quad 0 < x < N \qquad (4.2.27)$$

$$\Rightarrow \quad p(V_{x+1} - V_x) = q(V_x + V_{x-1}) \, ,$$

$$\Rightarrow \quad V_{x+1} - V_x = {}^{q}/_{p} \, (V_x + V_{x-1}) \quad .$$

By recursive method, we have

$$V_{x+1} - V_x = \left(q/p\right)^x (V_1 - 1) \ , \quad 0 < x < N \qquad (4.2.28),$$

Let

$$V_x - 1 = V_x - V_0 = (V_x - V_{x-1}) + (V_{x-1} - V_{x-2}) + \cdots + (V_1 - 1)$$

$$= \begin{cases} \dfrac{1 - \left(q/p\right)^x}{1 - \left(q/p\right)} (V_1 - 1) & p \neq q \\ x(V_1 - 1) & p = q \end{cases} \qquad (4.2.29) .$$

The absorbing barrier leading to V_N ,

$$V_x = \begin{cases} 1 - \dfrac{1 - \left(q/p\right)^x}{1 - \left(q/p\right)^N} & p \neq q \\ 1 - \dfrac{x}{N} & p = q \end{cases} \qquad (4.2.30) .$$

Furthermore, let D_x denote the expected time until a gambler who starts with x, say 1 (f-hb) is ruined.

The boundary conditions are $D_0=0$, $D_N=0$. By first-step analysis,

$$D_x = q(D_{x-1} + 1) + p(D_{x+1} + 1) = 1 + qD_{x-1} + pD_{x+1} \qquad (4.2.31)$$

$$\Rightarrow \ p(D_{x+1} - D_x) = q(D_x - D_{x-1}) - 1$$

Let $M_x = D_x - D_{x-1}$

$$\Rightarrow \qquad pM_{x+1} = qM_{x-1} \qquad (4.2.32)$$

Again by the recursive method, we have

$$M_x = \left(\frac{q}{p}\right)^{x-1} M_1 - \frac{1}{p} \sum_{j=0}^{x-2} \left(\frac{q}{p}\right)^j \qquad (4.2.33)$$

Also we have the initial condition $M_1 = D_1 - D_0 = D_1$

$$\Rightarrow \quad D_k = \sum_{j=1}^{k} M_j = \sum_{j=1}^{k} \left[\left(\frac{q}{p}\right)^j D_1 - \frac{1}{p} \sum_{i=0}^{j-2} \left(\frac{q}{p}\right)^j \right]$$

$$=$$

$$\begin{cases} \frac{1-\left(q/p\right)^k}{1-\left(q/p\right)} \left[D_1 - \frac{1}{p-q} \right] - \frac{k}{p-q} & (p \neq q) \\ k\left(D_1 - (k-1)\right) & (p = q) \end{cases} \qquad (4.2.34)$$

⇨ With $D_N = 0$,

$$D_1 = \begin{cases} \frac{N}{p} \left(\frac{1}{1-\left(q/p\right)^N} \right) - \frac{1}{p-q} & (p \neq q) \\ N-1 & (p = q) \end{cases} \qquad (4.2.35)$$

Thus we can calculate the expected number of game (D_x) until the gambler that starts at \$x is ruined.

$$D_x = \begin{cases} \frac{1-\left(q/p\right)^x}{1-\left(q/p\right)} \left[D_1 - \frac{1}{p-q} \right] - \frac{x}{p-q} & (p \neq q) \\ x(N-x) & (p = q) \end{cases} \qquad (4.2.36)$$

V. Results

5.1 One-way analysis of variance

Table 3.1 shows the descriptive results of f-Hb by disease status and other characteristics such as gender, age, family history, and brand type. The similar findings are shown when interval cancer is added (Table 3.2). Table 5.1.1 shows the frequencies of all repeated screens. Figure 5.1.1-5.1.6 shows the distribution of original f-Hb and also the corresponding ones with log transformation. These figures also show the results with and without considering undetectable f-Hb (including 0) in the normal group. The undetectable problem is considered by left censoring with the Breslow tie method in the Cox proportional hazards regression model. It can be seen that the log transformation renders the positive skewed distribution go toward a normal shape.

The analysis of variance for the log transformation of f-Hb (adding 0.5 unit to the right) shows that the difference in the mean value of f-Hb across three groups were statistically significant. (Table 5.1.2, $p<0.001$, $R^2=0.142$). The similar findings were noted when the non-parametric analysis was performed (Table 5.1.3).

5.2 Cox Proportional Hazards Regression Model

The results of univariable analysis are listed in Table 5.2.1 showing significant differences in the f-Hb concentration between categories of colorectal neoplasm, with disease-free case (normal group) as the reference group, the hazard ratio (HR) of the colorectal cancer group was 0.197 (0.194, 0.20), and the HR of the adenoma group was 0.213 (0.212, 0.215).

The results of multivariable analysis also show that men generally had higher f-Hb concentration than women (HR=0.948, (0.944, 0.951)), the old age group also had higher f-Hb concentration than the young age group. The effect of family history was significant in univariable analysis (HR=1.051, (1.036, 1.067)) but not significant in multivariable analysis (HR=1.012, (0.997, 1.027)). After adjusting for other covariates (gender, age, family history and brand), compared to the normal group, the HR of the cancer group was 0.181 (0.178, 0.184) and the adenoma group was 0.204 (0.202, 0.205). This model clearly clarifies that those who had been diagnosed with colorectal cancer tended to have higher f-Hb level in screening, as the adenoma group does. This indicates that screenee who had higher f-Hb may have higher probability to be diagnosed with disease. Table 5.2.2 shows the similar findings estimated by the accelerated failure time model.

Figures 5.2.1 and 5.2.2 show the cumulative figure with the non-parametric

method for f-Hb. We found the computation of f-Hb for interval cancer with the cold-deck method got the curve corrected (Figure 5.2.1 and Figure 5.2.2). Based on the nonparametric method we can also assess the f-Hb_{50} of CRC was 142 μg Hb/g, f-Hb_{50} of adenoma was 66 μg Hb/g, and f-Hb_{50} of normal near 0 μg Hb/g. The threshold value was 600 μg Hb/g for CRC and 400 μg Hb/g for adenoma. (Figure 5.2.2). Figures 5.2.3-5.2.5 show the corresponding curves by gender and age groups for cancer. The conspicuous difference was noted in the Figure of adenoma by gender (Figure 5.2.6).

5.3 The Random Walk Model

We used the faecal hemoglobin concentration of screenees as the repeated measures, the f-Hb change from last time over than 0 with probability *p*, less than 0 with probability *q*, and the staying probability is *r*. Tables 5.3.1 and 5.3.2 display the basic distribution of the steps about fecal hemoglobin concentration among all the states.

By assuming the normal distribution of each step and applying the central limit theorem, the unrestricted estimates for three groups are listed in Table 5.3.3. It can be clearly seen that the highest forward probability was noted for the colorectal cancer group, followed by the colorectal adenoma group, and the least for the

normal group.

Following the random walk model, we can model the probabilities of movement among different states in the same time.

Table 5.3.4 shows the estimated the corresponding regression coefficients of the logistic regression model of the forward probability (p). The results suggest patients diagnosed as CRC were more likely to move forward than those diagnosed as adenoma (α_1 (=1.592 (95% CI: 1.407~1.776)) > α_2 (=0.886 (95% CI: 0.836~0.937))) but the normal subjects were more likely move backward as the regression coefficient was negative (α_0=-0.583 (95% CI: -0.592~-0.574)).

Table 5.3.5 shows the calculation of forward (p) and backward (q) probability given the estimated regression coefficients gave 0.733 (95% CI: 0.697~0.768) and 0.267 (95% CI: 0.232~0.303) of p and q for patents diagnosed as CRC, 0.575 (95% CI:0.563~0.587) and 0.425 (95% CI:0.413~0.437) of p and q for patients diagnosed as adenoma, and 0.358 (95% CI:0.356~0.360) and 0.642 (95% CI:0.640~0.644) of p and q for the normal subjects. The drift (p-q) was positive for CRC and adenoma and negative for normal subjects. Compared with the normal group, the odds of moving forward was 4.923 for CRC and 2.426 adenoma. If we set 400 μg/g f-Hb for CRC, 300 μg/g f-Hb for adenoma and 20 μg/g for normal as the absorbing barrier the gambler's ruin probability of reaching the barrier was

0.867, which was higher than 0.455 of adenoma according to the equation of ruin probability whereas the ruin probability for the normal subject was very low. If the initial value (x) was set 1 it takes, on average, 740 steps for CRC, 893 steps for adenoma, and 7.05 steps for normal to reach absorbing barrier. This means it spent around 2.03 years for CRC to evolve from 1 to 400 and 2.44 years for adenoma from 1 to 300 and only 7 days from 1 to 0 for the normal subjects.

Table 5.3.6 shows the corresponding results when the adenoma was further classified as non-advanced adenoma and advanced adenoma. Again, all cases including CRC, advanced adenoma, and non-advanced adenoma show positive drift and only the normal group show negative drift. It is very interesting to note that the positive drift in the advanced adenoma group was even remarkable than the CRC group. Table 5.3.7 shows the corresponding ruin probability was up to 98.1% for advanced adenoma, 86.7% for CRC, and 17.4% for non-advanced adenoma. Again the ruin probability for the normal group was still very low. The expected number of steps taken to reach the absorbing barrier were 740.67 for CRC, 386 for advanced adenoma, and 1051.33 for non-advanced adenoma.

Tables 5.3.8 and 5.3.9 shows the similar findings when colorectal cancers were classified by two detection modes, screen-detected cases and interval cancers.

Table 5.3.10 gives the estimated regression coefficients based on the random walk model with the two logistic regression models for estimating the probabilities of forward, backward, and no movement. The sizes of regression coefficients with forward drift (αs) were larger than those of backward drift (βs) whereas the opposite was noted for the normal group. Table 5.3.11 shows the similar findings for the probabilities of forward, backward, and no movement, which are similar to those shown in Table 5.3.4.

Tables 5.3.12 and 5.3.13 show the corresponding results when the adenoma was further classified as non-advanced adenoma and advanced adenoma. Again, all cases including CRC, advanced adenoma, and non-advanced adenoma show positive drift and only the normal group show negative drift or no movement. Tables 5.3.14 and 5.3.15 shows the similar findings for the random walk model with the two logistic regression models when colorectal cancers were classified by two detection modes, screen-detected cases and interval cancers.

Tables 5.3.16-5.3.21 give the estimated results of coefficients, the probabilities of movement, ruin probability, and the expected number of steps taken to reach absorbing barrier including prevalent screen-detected cases. The distribution of f-Hb among prevalent screen-detected cases may be representative

of the long-run equilibrium distribution. The results with the consideration of prevalent cases show a remarkable contrast across three groups.

Tables 5.3.22-5.3.23 show the corresponding estimated results as above by considering gender as the covariate. Males had higher forward probability than females without considering prevalent screen data whereas the opposite was noted when prevalent screen was included but there was not much difference.

VI. Discussion

Novelties of empirical findings and methodology

The innovation of this current thesis can be specified from the two perspectives, methodological, and practical aspects. The development of good methodology provides an unbiased evaluation of the association between the disease status and the outcome of ordinal data. The application of the developed methodology to f-Hb concentration obtained from FIT in population-based screening also offers useful information that aid health decision-makers in designing an even delicate screening policy for personalized preventive strategies. Both are discussed as follows.

Advance in methodological development

As far as the former is concerned, the first is pertaining to the evaluation of the dynamics of the ordinal property of biomarker such f-Hb that is measured from fecal immunological test (FIT) and widely used for population-based colorectal cancer screening. Such an evaluation even in the well-known blood pressure has been very rare. The most intractable argument is that, in addition to the skewed property of such an ordinal data, undetectable f-Hb that corresponds to left-censored characteristics in the language of survival, the dynamic of f-Hb during the repeated FIT test, multi-state outcome of colorectal neoplasia, and the relationships of the upper and lower limit (two

absorbing barriers) to the status of colorectal neoplasia render the elucidation of the dynamics of f-Hb very intractable.

We began with a simplified statistical approach with the Cox proportional hazards regression model that regard the f-Hb concentration as the dependent variable of time to event and the disease status (including normal, adenoma, and CRC) as the main independent variable of interest making allowance for age, gender, family history, and brand of FIT. To further get a better understanding of dynamics of f-Hb making use of a large population-based screening data, we proposed a random walk regression model to estimate the forward (p) and backward (q) probability or no movement (r) in order to calculate ruin probability and the expected steps to reach the absorbing barrier given disease status including colorectal adenoma and cancer.

New empirical findings on f-Hb used for screening and surveillance policy of colorectal cancer screening program

Instead of regarding f-Hb concentration as the covariates and three or four multi-state colorectal neoplasia as dependent variables, we applied the Cox proportional hazards regression model and simple random walk regression model to relate the disease status to the dynamics of f-Hb. Modelling the dynamics of f-Hb in this manner may not only elucidate the disease progression of colorectal neoplasia but also provide a

new insight into how the median and threshold of f-Hb when colorectal adenoma and colorectal cancer were reached. In addition, the introduction of disease status and personal attributes (such as age and gender) into the two logistic regression models corresponding to the forward probability and the backward probability enables one to calculate the ruin probability and the expected number of steps to reach the upper limit of f-Hb for colorectal adenoma and colorectal cancer.

Several novel findings were noted in the current thesis

(1) The statistically significant differences in f-Hb concentration across three groups, normal, colorectal adenoma, and colorectal cancer are demonstrated to indicate the quantitative value of the administration of FIT after controlling for demographic features and family history.

(2) The effective median f-Hb concentration ($f\text{-}Hb_{50}$) was 142 μg Hb/g for CRC and 66 μg Hb/g for adenoma.

(3) The threshold f-Hb concentration was 600 μg Hb/g for CRC and 400 μg Hb/g for adenoma.

(4) The odds ratio of raising f-Hb (forward) as opposed to depreciating f-Hb (backward) derived from the simple random walk regression model was five times for CRC and two-and-half times for colorectal adenoma compared with the normal group.

(5) The probability of reaching 400 μg Hb/g after long-run transition was 86.7% for

CRC and reaching 300 μg Hb/g after long-run transition was 45.5% for adenoma but very low for the normal subjects to reach 20 μg Hb/g (Table 5.3.5).

(6) The expected steps taken to reach 400 μg Hb/g were estimated as 544 steps for the patients diagnosed as CRC and 515 steps for those diagnosed as colorectal adenoma to reach 300 μg Hb/g (Table 5.3.5).

The statistical issues of f-Hb recorded in FIT

There are several statistical concerns over f-Hb of FIT. First, although f-Hb is regarded as a quantitative measure for detecting possible colorectal cancer and adenoma its statistical property shows a skewed distribution. Therefore, using the continuous data with mean value as an indicator for the severity of biomarker seems inadequate. The ranking statistics may be better than the mean one. This prompts us to apply the Cox proportional hazards regression model to treat f-Hb concentration as an order statistics with the partial likelihood function for the ranking of a successive f-Hb concentration in order to assess the effect of disease status on the ranking of f-Hb. The second statistical issue of f-Hb is pertaining to the undetectable f-Hb at very extreme low value that is often recorded as 0. In the language of survival, these undetectable cases were treated as censored cases with ties based on the Breslow tie method. This method is rather robust and asymptotic to large samples given too many ties. The other reason of failing to use

the exact method is because too many ties render computationally intensive and intractable. Another solution to the undetectable f-Hb we had not done here is the E-M algorithm. We can construct a regression model for those screenees with actual f-Hb and their personal characteristics so that we can estimate the f-Hb for the undetectable one by using the E-M algorithm.

The second concern over the values of f-Hb is relevant to interval cancers. Because it is not possible to know the exact value of f-Hb for interval cancer when some cases missed at screen but surfaced to clinical phase the direct use of f-Hb measured at previous screen for interval cancers is not correct. The cold deck method was used for filling the missing values of f-Hb for these interval cancers. Figure 5.2.1 and 5.2.2 show the remarkable contrast between the uncorrected curves and the corrected ones for CRC. This again underscores the complexity of the statistical property of f-Hb.

Third, as the transition time regarding the evolution of f-Hb at prevalent screen was different from that at subsequent screen. The length bias would be more likely to happen at first screen because those with the long sojourn time lingering from the low f-Hb to high f-Hb were more likely to be detected at first screen than subsequent screens. The combined use of both detection modes using the Cox proportional hazards regression model seems inadequate. This can be clearly seen in the random walk regression model. The forward and back probabilities using data on subsequent screen

and interval cancer was substantially different from those including prevalent screen cases. It stands to reason that prevalent screen-detected cases have experienced long travelling from low f-Hb to high f-Hb compared with subsequent screen-detected cases. The forward and backward probabilities have almost reached the equilibrium distribution at first screen whereas the subsequent screen-detected may not have sufficient follow-up time to reach the equilibrium distribution as seen at first screen. This may account for why the forward and backward probabilities after including prevalent screen-detected cases were more distinct than those only including subsequent screen-detected and interval cancers.

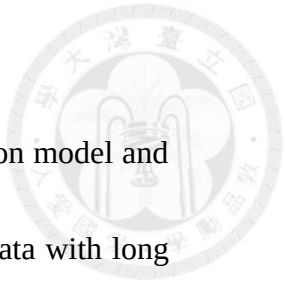
Implications for population-based screening for CRC

The empirical findings here provide a new insight into policy-making for CRC screening and surveillance of early CRC detected with FIT.

For example, the median f-Hb_{50} and the threshold of f-Hb can be used for identifying high risk of directly receiving colonoscopy. The ruin probability also can be used for assessing how much time the subject would be taken to reach then boundary of high f-Hb in order to assess the baseline risk of underlying population. The fewer the steps taken, the higher the risk for CRC.

Limitations

There are several concerns over the methodology and applications of the proposed methods. Although the proposed Cox proportional hazards regression model and the random walk regression model can accommodate the unique characteristics of ordinal data presented here, it still needs the assumption of modelling such an ordinal data. For example, a simple random walk model used here has the same assumption of incremental independence used in the Weiner process (simple Brown motion) and the variance proportional to duration. The relaxation of such an assumption using O-U process can be considered in the future. The alternative may consider the development of Markov ordinal regression model as done by the Mandel, Gauthier, Guttmann, Weiner, and Rebecca (2007) study. The second concern over the empirical data is that as the repeated screens rate only cover around one-third of subjects participating in the first screen whether the non-participants in the subsequent screen may affect the results is not known.



In conclusion, we have applied the Cox proportional hazards regression model and developed a random walk regression model to accommodate the ordinal data with long tail distribution at extremely high value, undetectable (left-censored) circumstance at extremely low value, and missing values and also in relation to multi-state outcome. The proposed models have been applied to nationwide population-based screening for CRC with FIT to estimate the hazard ratio for colorectal cancer and adenoma as opposed to the normal subjects, also to estimate the effective median f-Hb and threshold of developing CRC and adenoma, and get a better understanding of how f-Hb moves forward and backward with time and what is the chance of having gambler's ruin (reaching to the barriers of f-Hb) and how many steps are expected to be taken before ruining. These findings provide a new insight into policy-making for colorectal cancer screening and also surveillance of early-detected colorectal cancer.

REFERENCE

BOOKS

Cox D.R., Miller H.D. (1965). The theory of stochastic processes. London. *Chapman and Hall.*

Rubin D. B. (1987). Multiple Imputation for Nonresponse in Surveys. Hoboken, NJ, USA. *John Wiley & Sons, Inc.* DOI: 10.1002/9780470316696

JOURNAL ARTICLE

Breslow N. (1974). Covariance analysis of censored survival data. *Biometrics*. 30(1), 89-99.

Chen L. S., Yen A. M. F., Chiu S. Y. H., Liao C. S., Chen H. H. (2011). Baseline faecal occult blood concentration as a predictor of incident colorectal neoplasia: longitudinal follow-up of a Taiwanese population-based colorectal cancer screening cohort. *Lancet Oncology.* 12: 551–558. DOI: 10.1016/S1470-2045(11)70101-2.

Chen L. S., Yen A. M. F., Fraser C. G., Chiu S. Y. H., Fann J. C.Y., Wang P. E., Lin S. C. , Liao C. S., Lee Y. C., Chiu H. M., Chen H. H. (2013). Impact of faecal

haemoglobin concentration on colorectal cancer mortality and all-cause death. *BMJ Open.* 3:e003740. DOI: 10.1136/bmjopen-2013-003740.

Chiu H. M., Chen S. L. S., Yen A. M. F., Chiu S. Y. H., Fann J. C.Y., Lee Y. C., Pan S. L., Wu M. S., Liao C. S., Chen H. H., Koong S. L., and Chiou S. T. (2015). Effectiveness of Fecal Immunochemical Testing in Reducing Colorectal Cancer Mortality From the One Million Taiwanese Screening Program. *Cancer*. DOI: 10.1002/cncr.29462.

Hopper J. L., Young G. P. (1988). A random walk model for evaluating clinical trials involving serial observations. Statistics in Medicine. 7, 581-590. DOI: 10.1002/sim.4780070505.

Lin D.Y. and Wei L.J. (1989). The robust inference for the Cox proportional hazards model. *Journal of the American Statistical Association.* 84(408), 1074-1078. DOI: 10.2307/2290085.

Mandel M., Gauthier S. A., Guttmann C. R. G., Weiner H. L., Rebecca. (2007). Estimating Time to Event From Longitudinal Categorical Data: An Analysis of Multiple Sclerosis Progression. *Journal of the American Statistical Association.* 102(480), 1254–1266. DOI: 10.1198/016214507000000059.

APPENDIX

i. Figure

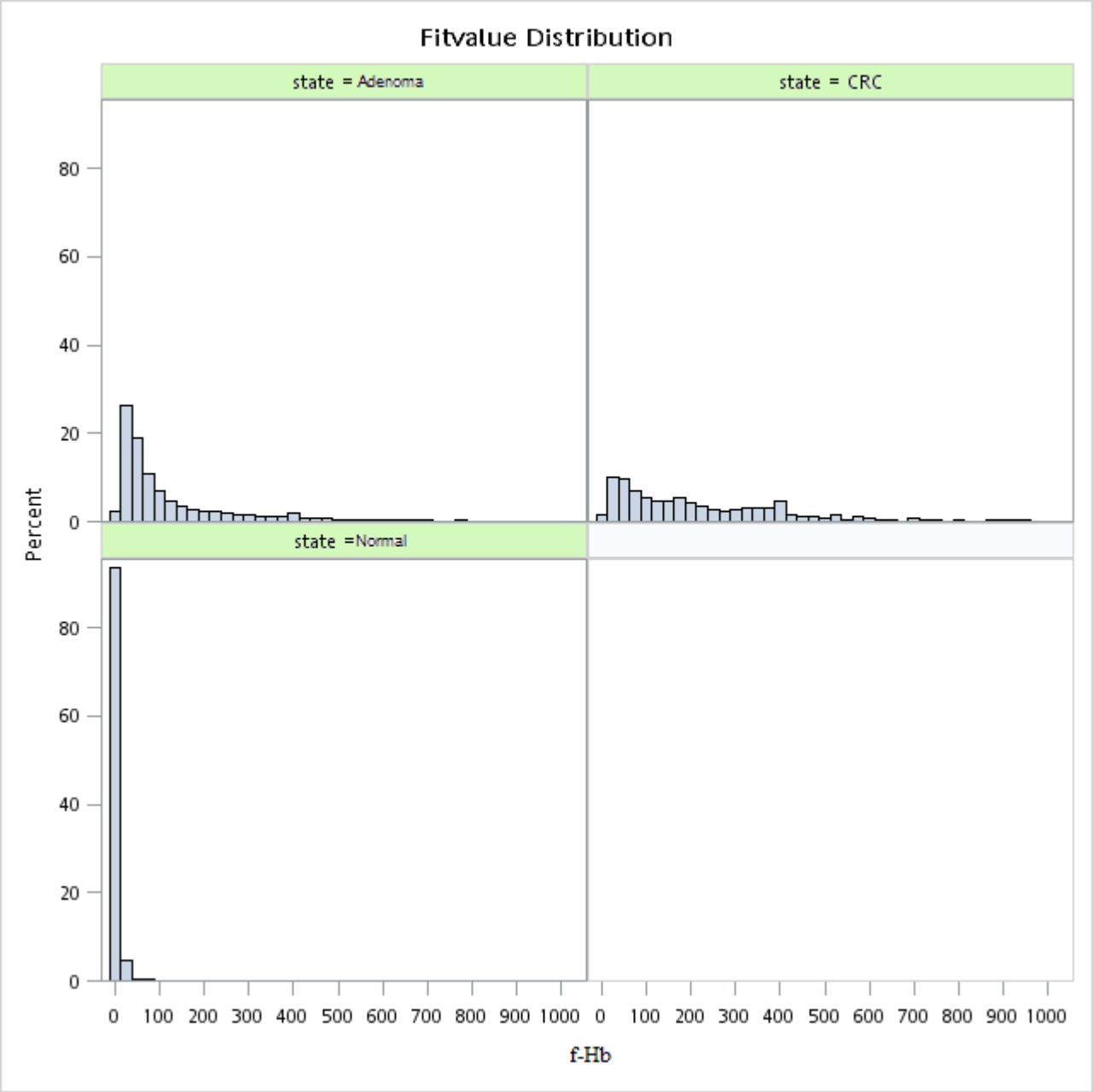


Figure 5.1.1 Histogram of original f-Hb by three disease statuses (normal, adenoma, and colorectal cancer)

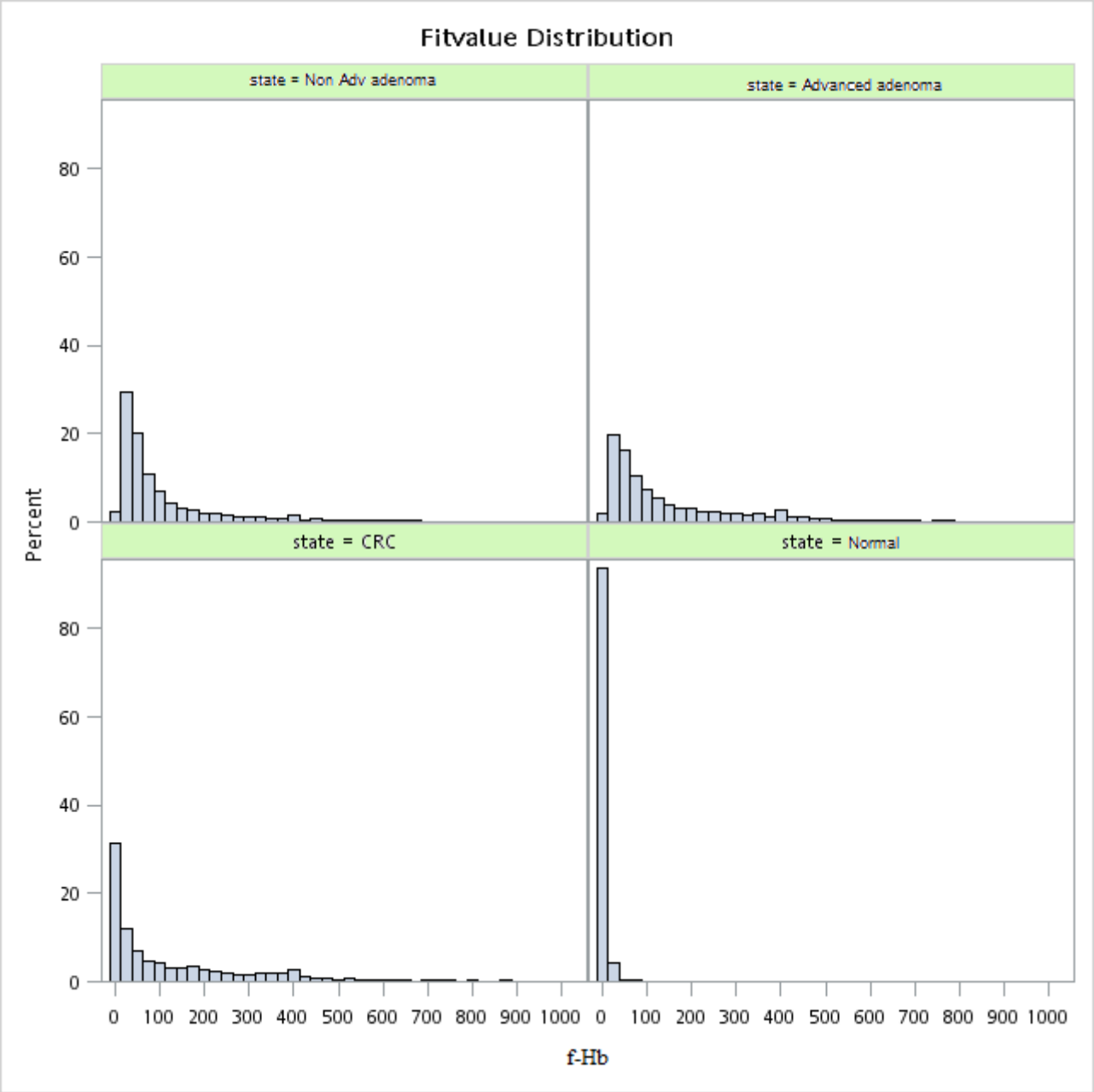


Figure 5.1.2 Histogram of original f-Hb by four disease statuses (normal, non-advanced adenoma, and advanced adenoma, and colorectal cancer)

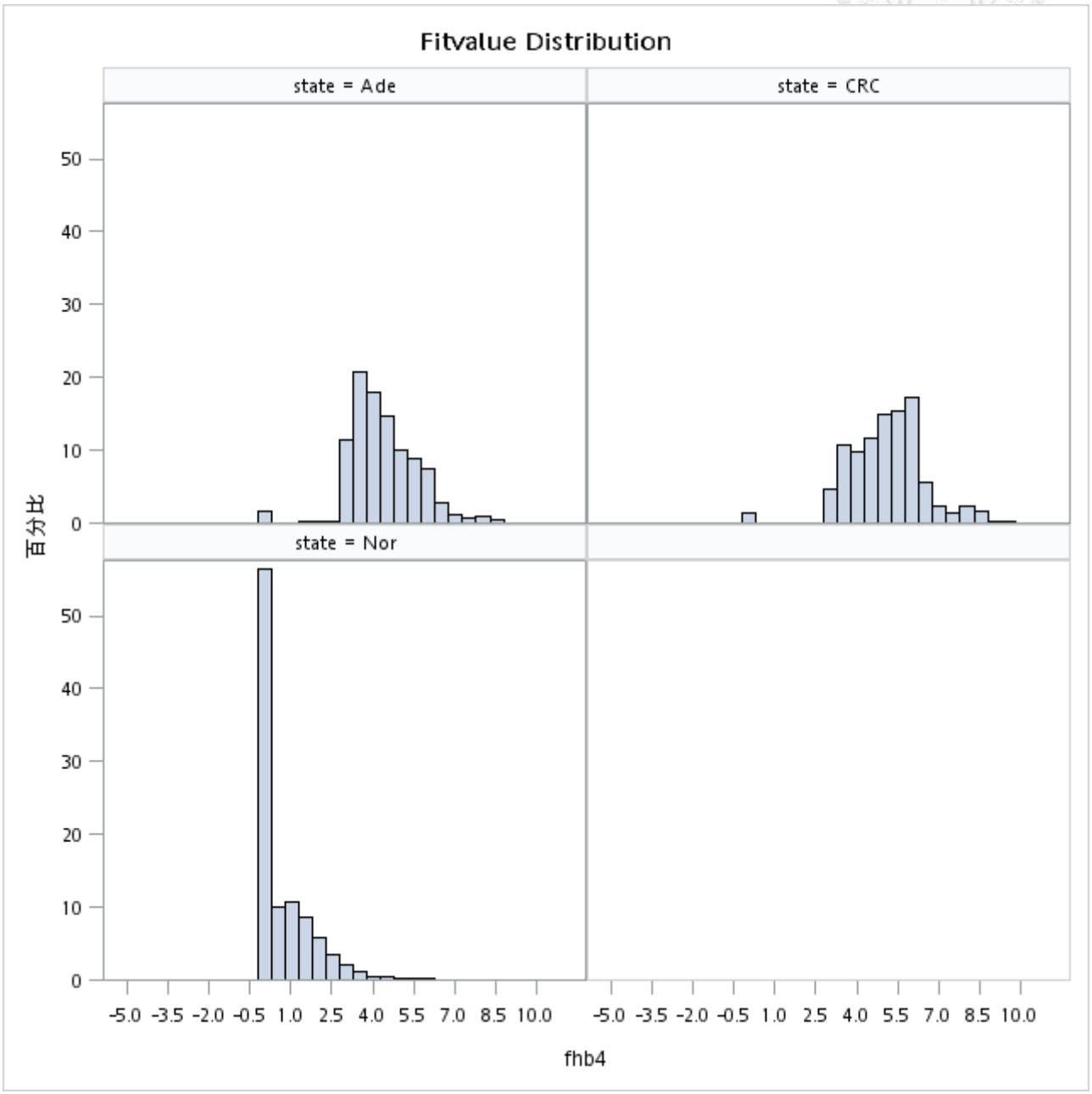


Figure 5.1.3 Histogram of ln(f-Hb) (adding 0.5 unit to the right) by disease status before IC interpolation

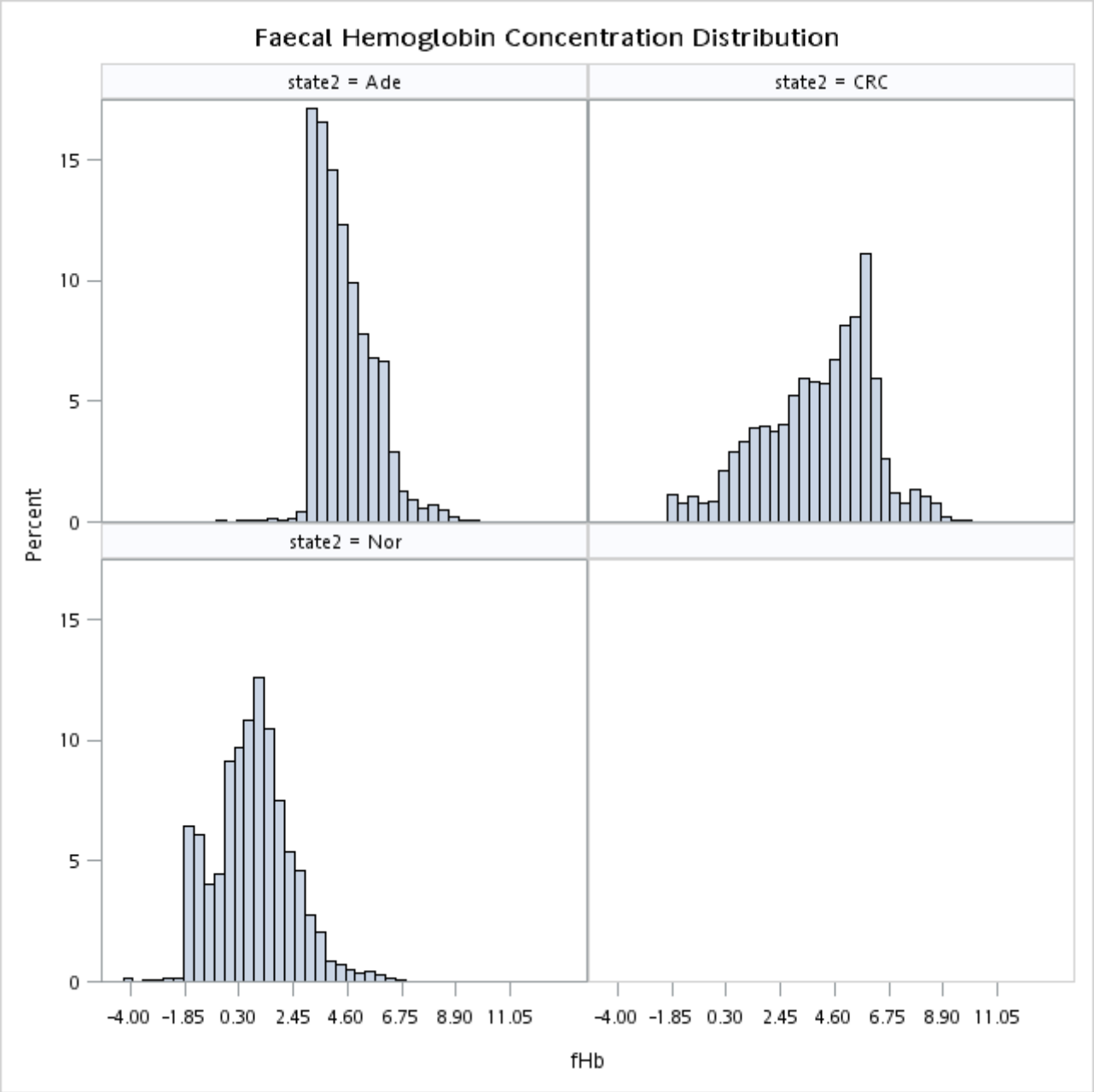


Figure 5.1.4 Histogram of ln(f-Hb) (excluding undetected cases) by disease status before IC interpolation

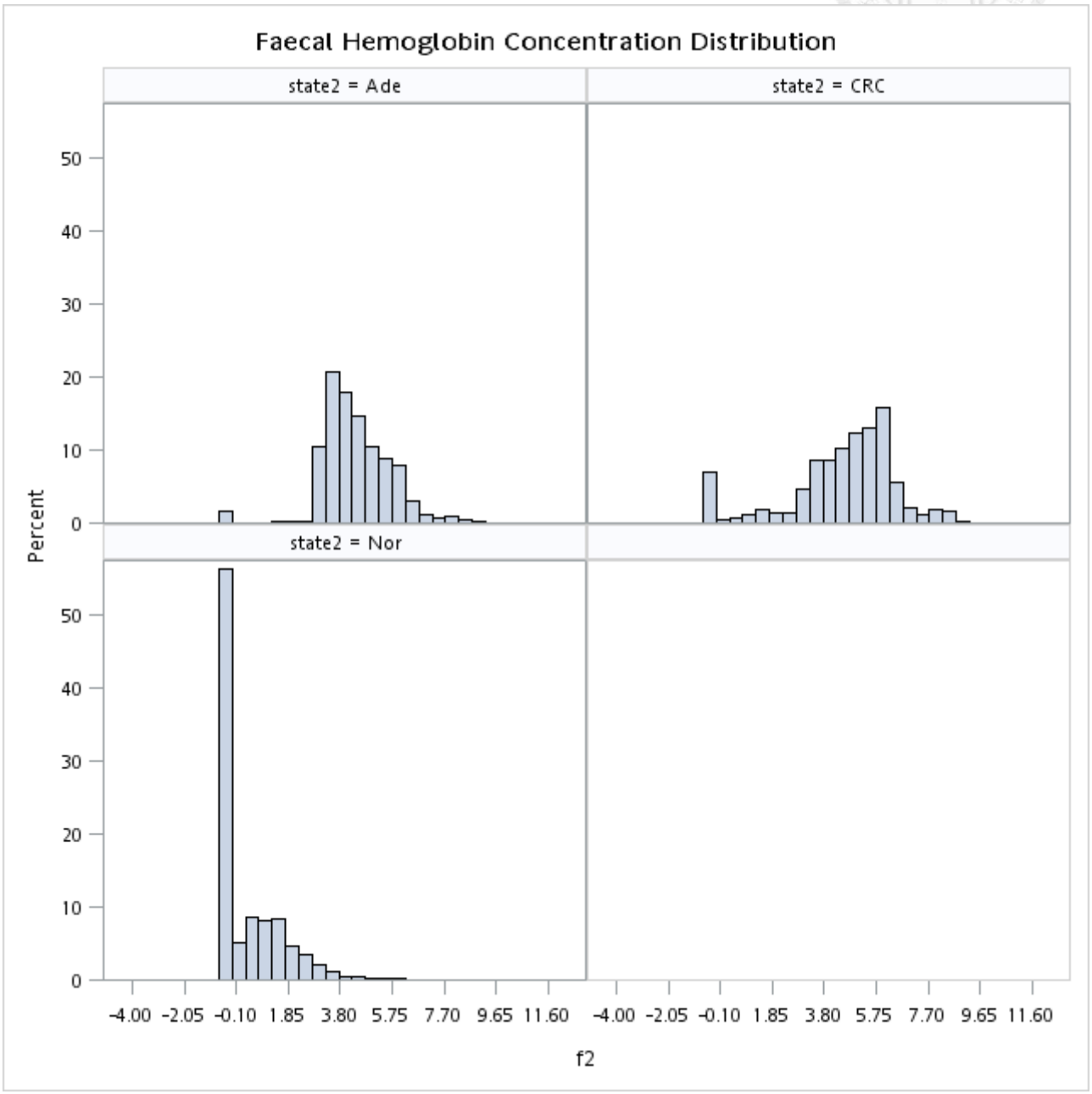


Figure 5.1.5 Histogram of ln(f-Hb) (adding 0.5 unit to the right) by disease status after IC interpolation

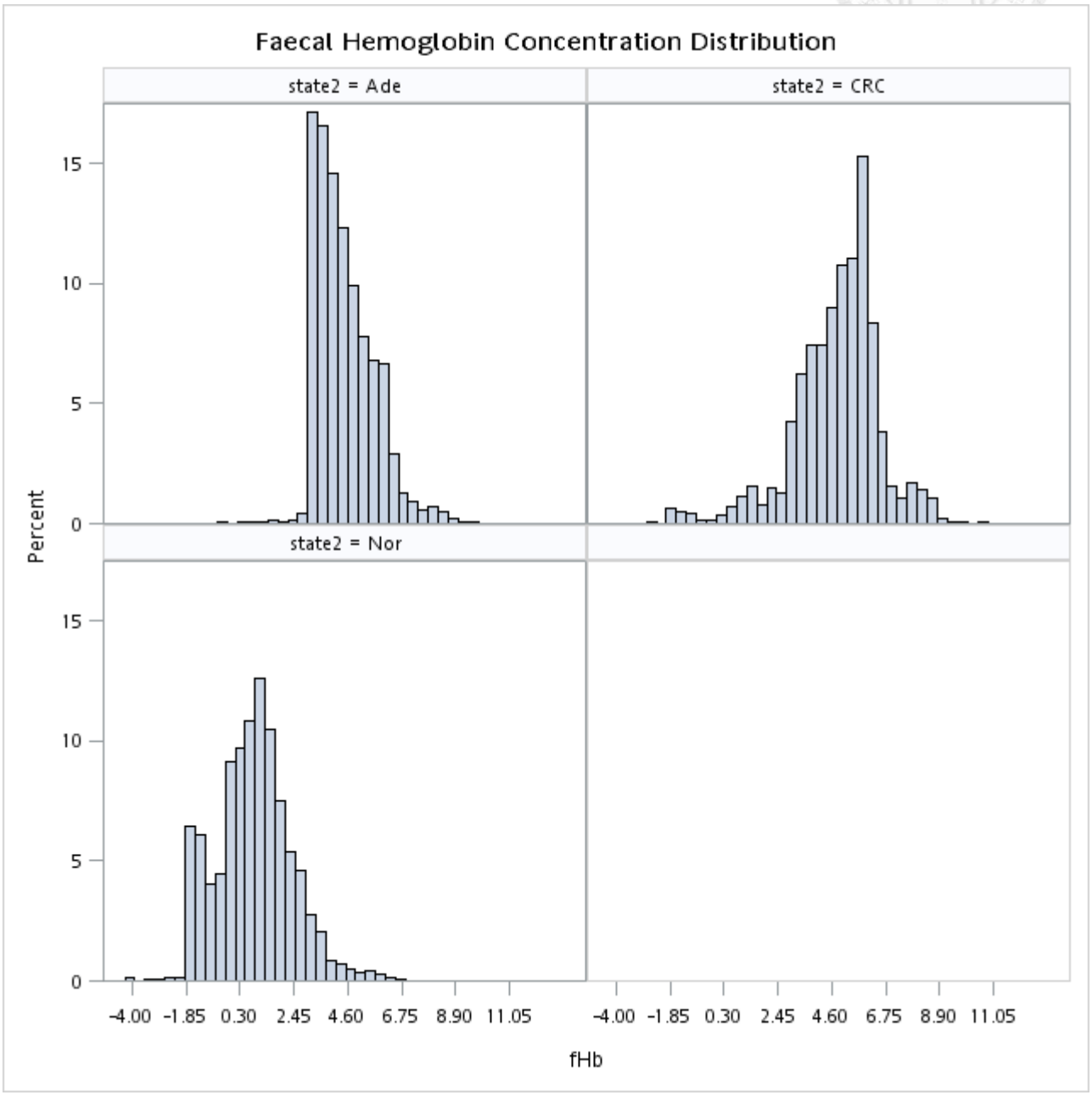


Figure 5.1.6 Histogram of ln(f-Hb) (excluding undetected cases) by disease status after IC interpolation

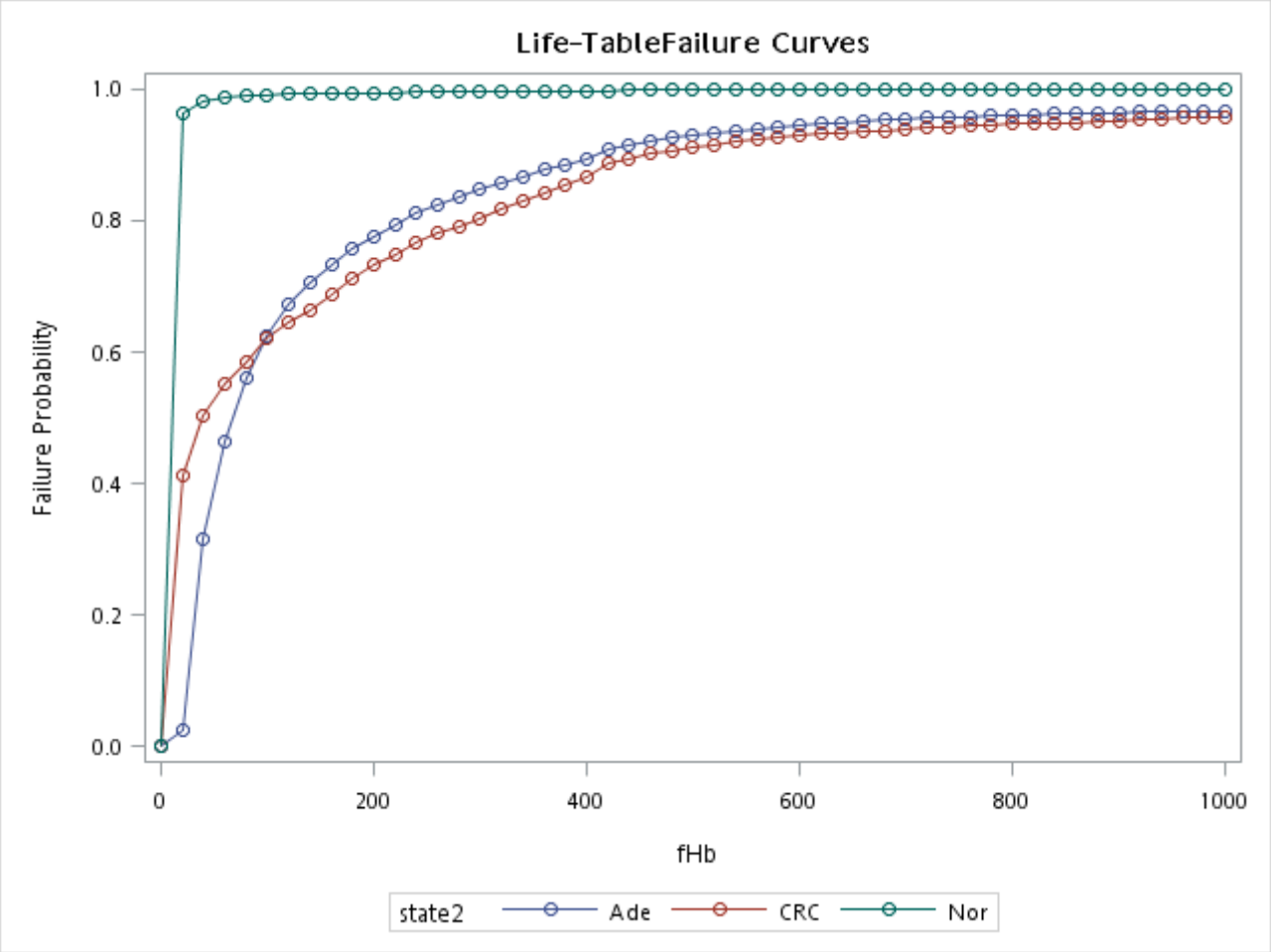


Figure 5.2.1 Cumulative distribution of f-Hb by different states before IC interpolation

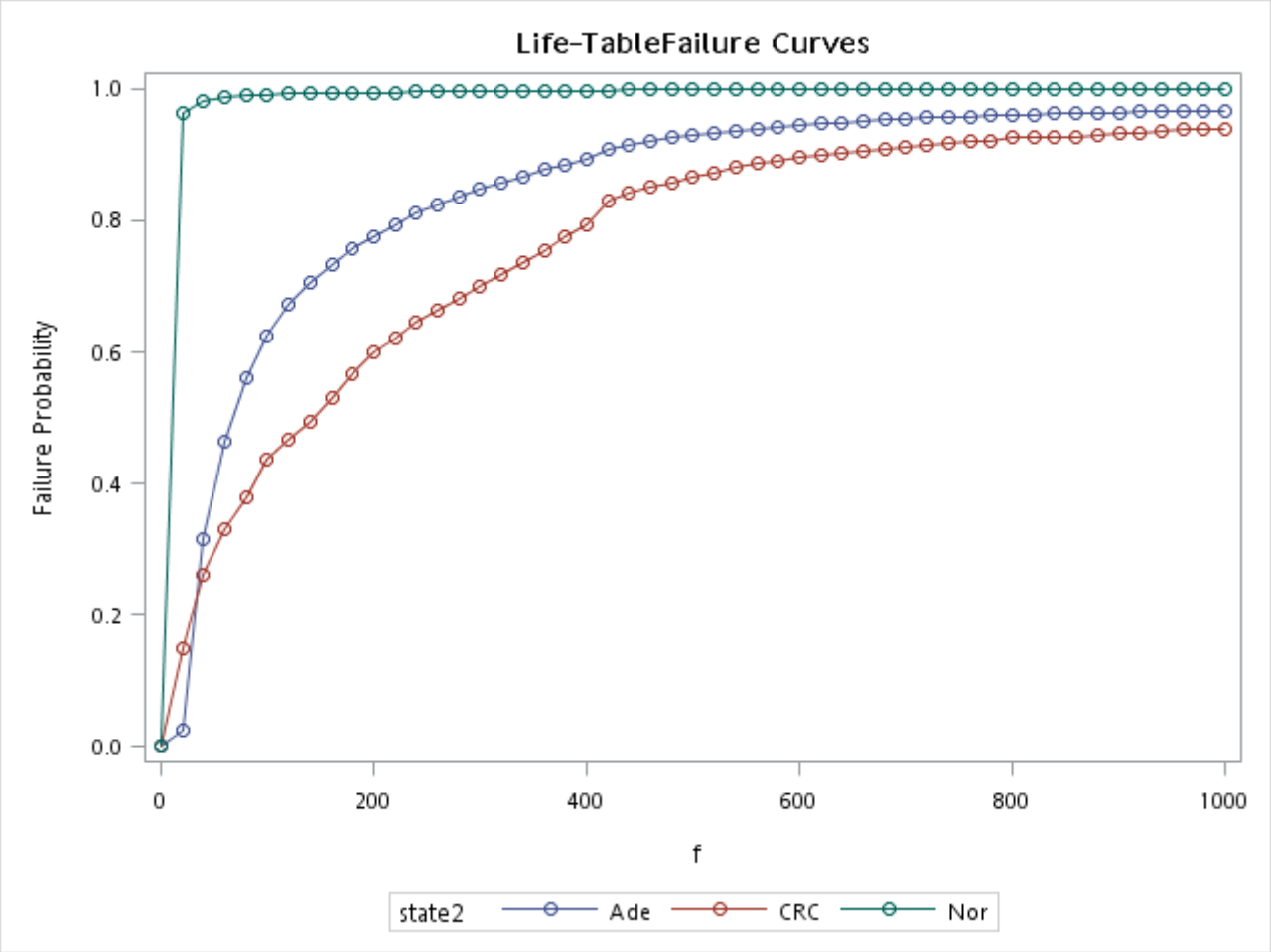


Figure 5.2.2 Cumulative distribution curve of f-Hb by different states after IC interpolation

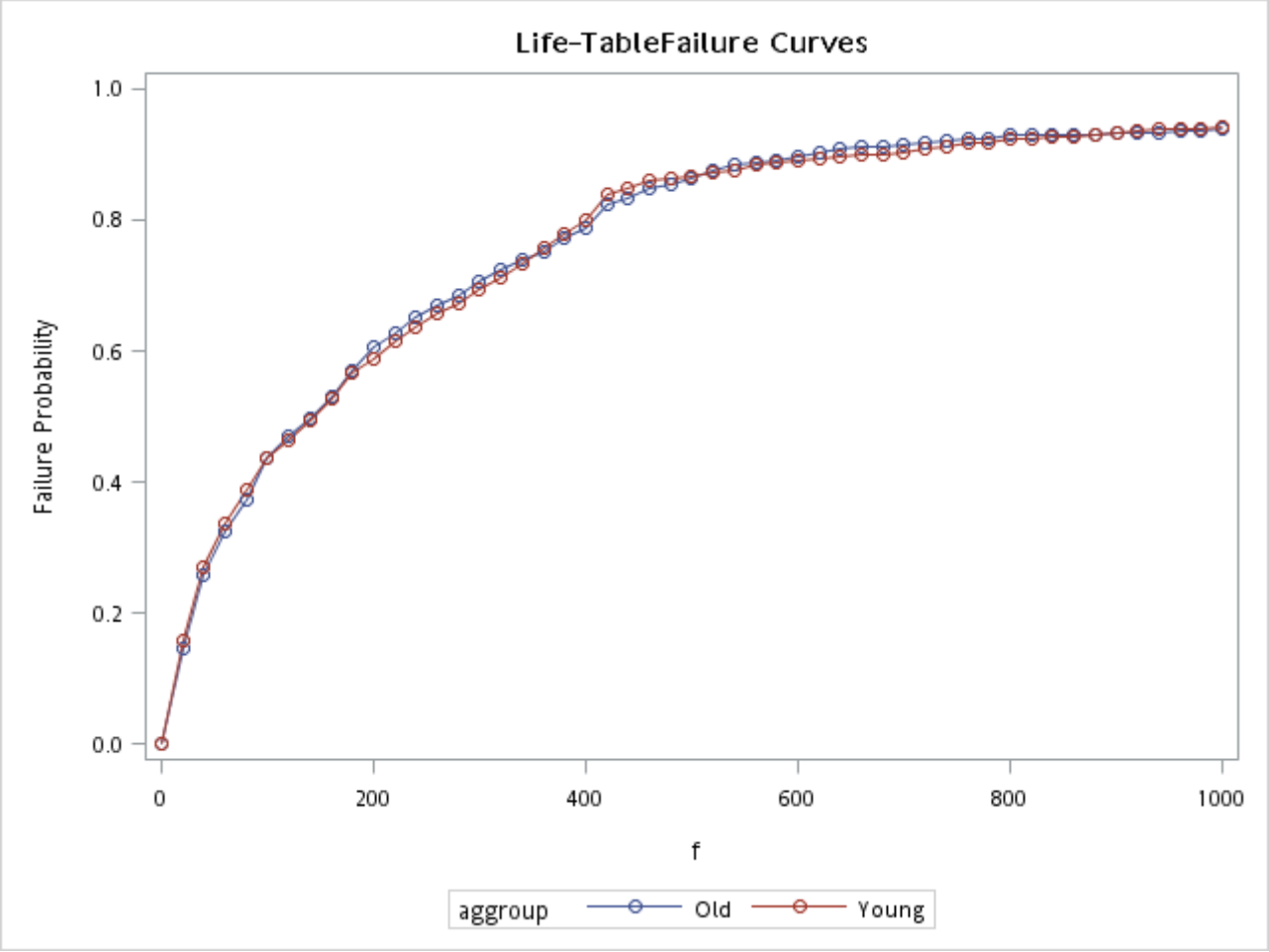


Figure 5.2.3 Cumulative distribution curve of f-Hb among age groups of cancer patients

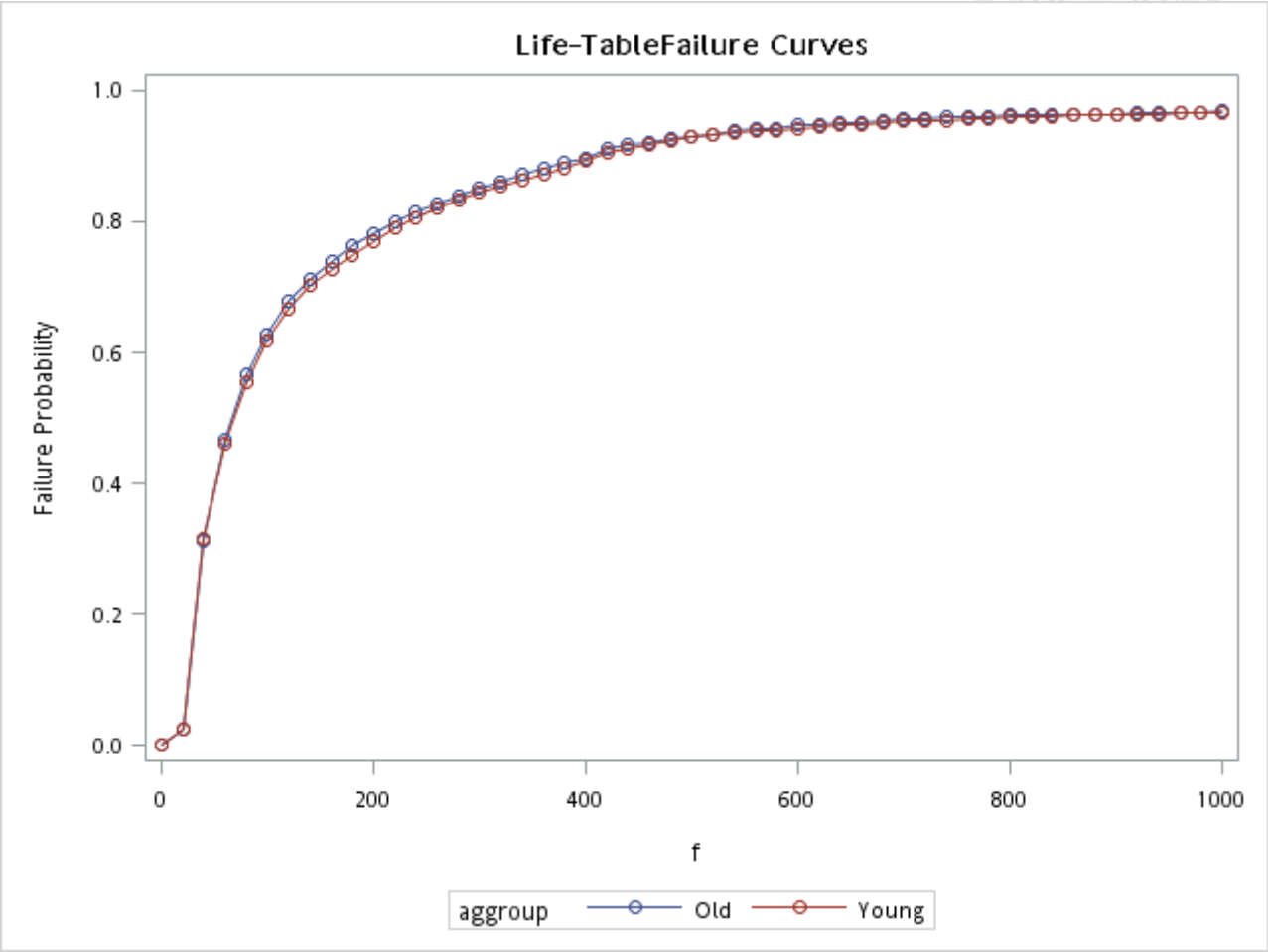


Figure 5.2.4 Cumulative distribution curve of f-Hb among age groups in adenoma patients

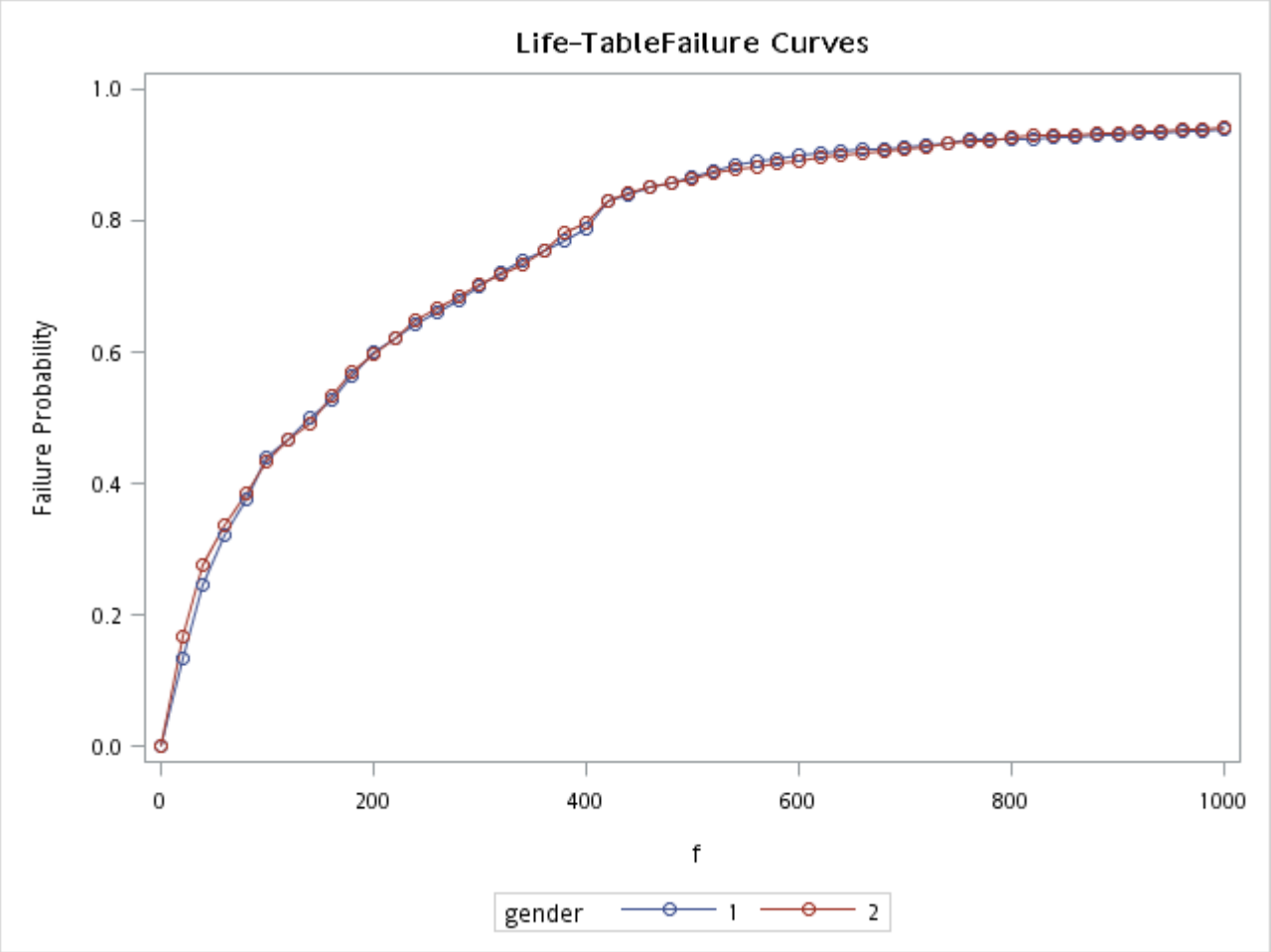


Figure 5.2.5 Cumulative distribution curve of f-Hb among gender groups of cancer patients

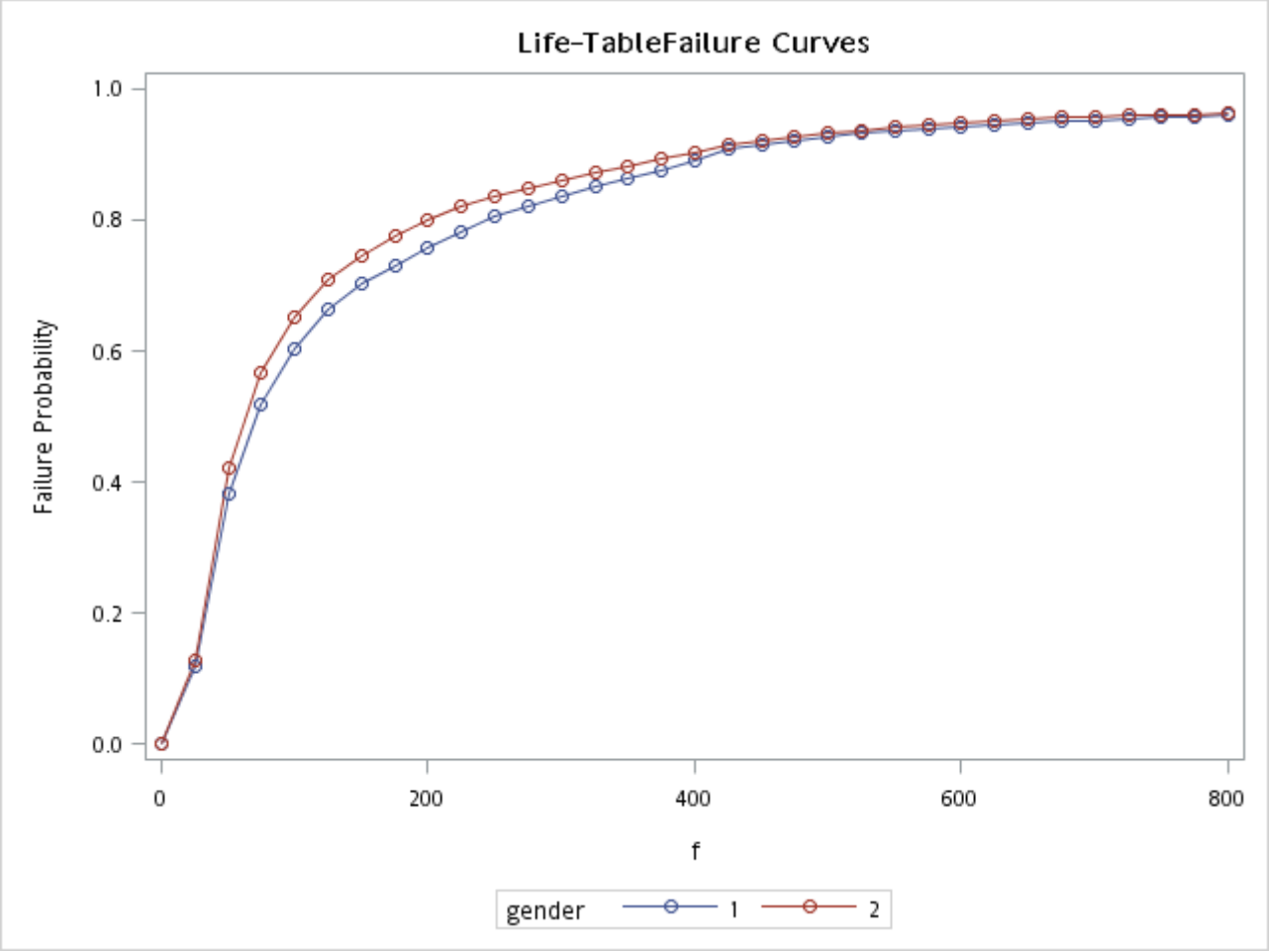


Figure 5.2.6 Cumulative distribution curve of f-Hb among gender groups of adenoma patients

ii. Table

Table 2.1 Estimated results of re-analysis of symptom and endoscopy measures of treatments for peptic oesophagitis

	symptomatic scores		endoscopy scores	
estimate	Case	Control	Case	Control
P	0.02(0.007)	0.02(0.007)	0	0.449(0.026)
Q	0.269(0.026)	0.182(0.021)	0.109(0.015)	0.551
E	-0.25(0.025)	-0.16(0.022)		
Test result	Z=2.7, p<0.02		chisq=13.3 (df=2), p<0.001	

Table 2.2 The results of the probability of symptom score after the movement of n step

	symptomatic scores	
Step n	Case	Control
2	0.9396	0.9038
4	0.9859	0.9674
6	0.9964	0.9880
8	0.9990	0.9954
10	0.9997	0.9982

Table 2.3 The results of ruin probabilities with different absorbing states

	Case group		Control group	
Start at j	Absorbed at 0	Absorbed at 6	Absorbed at 0	Absorbed at 6
1	1	0	1	0
2	1	0	0.9999	0.0001
3	0.9996	0.0004	0.9987	0.0013
4	0.9945	0.0055	0.9879	0.0121
5	0.9257	0.0743	0.8901	0.1099

Table 2.4 Estimated results of limiting equilibrium distribution (π_k) with reflecting barriers (state 0 and state 6) on symptomatic scores example

	symptomatic scores	
k	Case	Control
0	0.9257	0.8901
1	0.0688	0.0978
2	0.0005	0.0107
3	0.0004	0.0012
4	0	0.0001

Table 2.5 The results on the estimates of random walk model parameters (with standard errors), and log-likelihood for bacitracin and vancomycin treatment groups

Model	Bacitracin		Vancomycin		Log-likelihood
	p	Q	p	q	
1	0.235(0.031)	0.386(0.032)	same as bacitracin		-349.867
2	0.236(0.031)	0.365(0.038)	same as bacitracin	0.415(0.046)	-349.415
3	0.268(0.038)	0.383(0.03)	0.185(0.038)	same as bacitracin	-348.438
4	0.269(0.041)	0.384(0.041)	0.184(0.041)	0.381(0.045)	-348.437

Table 3.1 The descriptive results of f-Hb by disease status and other characteristics of visits (screens) for each individual

Variable		N (%)	Median of fHb	Mean of fHb	STD of fHb	Interquartile Range
State	CRC	4574 (0.361)	39 (74.1)*	327.299 (386.838)*	3735.94 (4058.8)*	217.9 (264.6)*
	Adenoma	15604 (1.233)	66.4 (68)*	233.59 (237.362)*	2017.64 (2033.64)*	140.4 (142.5)*
	Normal	1245127 (98.405)	0.04 (2.4)	8.123 (16.208)	400.085 (565.02)	2.4 (4.8)
Gender	Male	471412 (37.257)	0.2	16.07	646.702	2.8
	Female	793893 (62.743)	0	9.676	405.782	2.4
Age group	50~54	396274 (31.318)	0	8.796	178.184	2.2
	55~59	360334 (28.478)	0	11.678	550.729	2.4
	60~64	250973 (19.835)	0.2	13.049	449.627	2.8
	65~69	257724 (20.369)	0.25	16.639	776.188	3.2
Family history	Yes	17055 (1.348)	0	11.652	147.805	2.0
	No	1248250 (98.652)	0.2	12.063	512.231	2.5
Brand	Brand 1	970461 (76.698)	0	6.136	88.762	1.6
	Brand 2	294844 (23.302)	2	31.549	1041.95	5.0
Overall		1265305	0.2	12.058	509.056	2.5

*excluded undetected cases

Table 3.2 Basic characteristics table of f-Hb after adding the value of f-Hb interval cancer with interpolation

Variable		N (%)	Median of fHb	Mean of fHb	STD of fHb	Interquartile Range
State	CRC	4574 (0.361)	142.7 (158.7)	399.327 (423.984)	1321.430 (1357.78)	318.2 (321.8)
	Adenoma	15604 (1.233)	66.4 (68)	233.59 (237.362)	2017.640 (2033.64)	140.4 (142.5)
	Normal	1245127 (98.405)	0.04 (2.4)	8.123 (16.208)	400.085 (565.017)	2.4 (4.8)
Gender	Male	471412 (37.257)	0.2	16.624	647.830	2.8
	Female	793893 (62.743)	0	9.761	306.174	2.4
Age group	50~54	396274 (31.318)	0	9.031	179.615	2.2
	55~59	360334 (28.478)	0	11.987	551.632	2.4
	60~64	250973 (19.835)	0.2	13.584	456.984	2.8
	65~69	257724 (20.369)	0.25	16.601	614.835	3.2
Family history	Yes	17055 (1.348)	0	12.458	153.343	2.0
	No	1248250 (98.652)	0.2	12.316	466.698	2.5
Brand	Brand 1	970461 (76.698)	0	6.634	102.348	1.6
	Brand 2	294844 (23.302)	2	31.027	942.622	5.0
Overall		1265305	0.2	12.318	463.884	2.5

*excluded undetected cases

Table 5.1.1 Interval cancer frequency in all repeated measures

	Interval Cancer	**Screen detected cancer**			**All cancer**
		Prevalence	Subsequence	All	
Numbers	1807	1891	876	2767	4574
Total	1265305	1265305	1265305	1265305	1265305
Rate (%)	0.143	0.149	0.069	0.219	3.616

Table 5.1.2 The results of ANOVA table

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	343442.748	171721.374	104324	<.0001
Error	1.27E+06	2082744.965	1.646		
Corrected Total	1.27E+06	2426187.714			

*$R^2 = 0.141557$

Table 5.1.3 The non-parametric analysis of f-Hb across three disease status

Method		Chi-Square Statistic	DF	P-value
Kruskal-Wallis Test		58206.937	2	<.0001
Median Analysis	One-Way	18792.669	2	<.0001
Savage Analysis	One-Way	206459.438	2	<.0001

Table 5.2.1 The estimated hazard ratio of reaching f-Hb using Cox proportional hazards regression model：$h(fHb) = h_0(fHb) \times \exp(\boldsymbol{\beta X})$

		Univariable	Multivariable
		HR	aHR
Variable			
State	CRC	0.197 (0.194,0.20)	0.181 (0.178,0.184)
	Adenoma	0.213 (0.212,0.215)	0.204 (0.202,0.205)
	Normal (ref)	1	1
Gender	Male	0.918 (0.915,0.921)	0.948 (0.944,0.951)
	Female (ref)	1	1
Age group	50~54 (ref)	1	1
	55~59	0.968 (0.965,0.972)	0.982 (0.978,0.986)
	60~64	0.910 (0.906,0.914)	0.931 (0.927,0.935)
	65~69	0.874 (0.870,0.877)	0.896 (0.892,0.900)
Family history	Yes	1.051 (1.036,1.067)	1.012 (0.997,1.027)
	No (ref)	1	1
Brand	Brand 1	1.558 (1.553,1.564)	1.624 (1.618,1.630)
	Brand 2 (ref)	1	1

Table 5.2.2 The estimated hazard ratio of reaching f-Hb using the Accelerated failure time model：
$\ln(fHb) = \boldsymbol{X\beta} + \sigma\boldsymbol{\varepsilon}$

		Univariable	Multivariable
		coefficients	coefficients
Intercept			1.743 (1.733,1.754)
State	CRC	3.876 (3.822,3.93)	3.884 (3.225,3.282)
	Adenoma	3.281 (3.252,3.31)	3.254 (3.225,3.282)
	Normal (ref)	0	0
Gender	Male	0.318 (0.308,0.328)	0.152 (0.143,0.161)
	Female (ref)	0	0
Age group	50~54 (ref)	0	0
	55~59	0.109 (0.097,0.122)	0.046 (0.034,0.057)
	60~64	0.251 (0.237,0.265)	0.143 (0.131,0.156)
	65~69	0.364 (0.351,0.378)	0.206 (0.193,0.218)
Family history	Yes	0.169 (0.124,0.214)	0.081 (0.04,0.121)
	No (ref)	0	0
Brand	Brand 1	-0.294 (-0.303-0.284)	-0.417 (-0.426,-0.408)
	Brand 2 (ref)	0	0
Scale			1.789 (1.787,1.724)
Shape			0.559 (0.558,0.56)

Table 5.3.1 Number of jumps distribution among states

Number of jumps	States				
	Non Advanced Adenoma	Advanced Adenoma	Screen Detected Cancer (SDC)	Interval Cancer (IC)	Normal
0	8157	3303	2041	1499	816132
1	3370	684	291	132	165050
2	745	82	44	12	26649
3	96	9	3	2	2982
4	1	0	0	0	30
Total	12369	4078	2379	1645	1010843

Table 5.3.2 Step distribution of f-Hb among state

	Forward	Backward	No movement
SDC	2363	56	10
IC	81	81	0
Adv Adenoma	4071	88	19
Non adv Adenoma	10856	2199	254
Normal	81464	83215	62735

Table 5.3.3 The estimated parameters on the use of random walk model assuming normal approximation

Group	μ	σ	p	q	r
CRC	0.484 (0.411,0.556)	0.865 (0.814,0.916)	0.733 (0.594,0.872)	0.249 (0.153,0.345)	0.018 (0,0.131)
Adenoma	0.196 (0.172,0.22)	0.957 (0.94,0.974)	0.575 (0.532,0.619)	0.379 (0.341,0.418)	0.045 (0.011,0.079)
Normal	-0.008 (-0.011,-0.004)	0.851 (0.848,0.853)	0.358 (0.353,0.364)	0.366 (0.36,0.371)	0.276 (0.272,0.28)

Table 5.3.4 Estimated regression coefficients and their 95% Cis with the random walk regression model considering three disease statuses, normal, colorectal adenoma, and colorectal cancer

Coefficient	Estimate	SE	95% CI	
			Lower	Upper
α_0	-0.583	0.004	-0.592	-0.574
α_1	1.592	0.094	1.407	1.776
α_2	0.886	0.026	0.836	0.937

*Loglikelihood = -152790.2

Model: $\text{logit}(p_i) = -0.583 + 1.592\text{CRC}_i + 0.886\text{Adenoma}_i$

Table 5.3.5 Estimated forward (p) and backward (q) probability, the odds ratio of p/q, ruin probability, and the expected steps based on the estimated parameters from Table 5.3.4

State group	p (95% CI)	q (95% CI)	OR	f-Hb (μg/g)	Ruin probability	Expected steps	
						D_1	D_X
Cancer	0.733 (0.697,0.768)	0.267 (0.232,0.303)	4.923	400	0.867	543.76	740.75
Adenoma	0.575 (0.563,0.587)	0.425 (0.413,0.437)	2.426	300	0.455	514.89	893.37
Normal	0.358 (0.356,0.360)	0.642 (0.640,0.644)	1.000	20	$1.9\text{x}10^{-5}$	3.53[a]	7.05[a]

*a: go to 0 μg/g

Table 5.3.6 Estimated regression coefficients and their 95% Cis with the random walk regression model considering four disease statuses, normal, colorectal non-advanced adenoma, advanced adenoma, and colorectal cancer

Coefficient	Estimate	SE	95% CI	
			Lower	Upper
α_0	-0.583	0.005	-0.592	-0.574
α_1	1.592	0.095	1.406	1.778
α_2	2.553	0.102	2.353	2.754
α_3	0.679	0.029	0.621	0.736

*Loglikelihood = -152571.4

Model: $\text{logit}(p_i) = -0.583 + 1.592\text{CRC}_i + 2.553\text{AdvAdenoma}_i + 0.679\text{Adenoma}_i$

Table 5.3.7 Estimated forward (p) and backward (q) probability, the odds ratio of p/q, ruin probability, and the expected steps based on the estimated parameters from Table 5.3.6

State group	p (95% CI)	q (95% CI)	OR	f-Hb (μg/g)	Ruin probability	Expected steps	
						D_1	D_X
Cancer	0.733 (0.696,0.769)	0.267 (0.231,0.304)	4.923	400	0.867	543.73	740.67
Adv Adenoma	0.878 (0.769,0.856)	0.122 (0.101,0.144)	12.906	300	0.981	340.50	386.83
Non adv Adenoma	0.524 (0.510,0.538)	0.476 (0.462,0.490)	1.974	300	0.174	551.72	1051.31
Normal	0.358 (0.356,0.360)	0.642 (0.640,0.644)	1	20	$1.9\text{x}10^{-5}$	3.53[a]	7.05[a]

*a: go to 0 μg/g

Table 5.3.8 Estimated regression coefficients and their 95% Cis with the random walk regression model considering four disease statuses, normal, colorectal adenoma, screen-detected colorectal cancer (SDC), and interval cancer (IC)

Coefficient	Estimate	SE	95% CI	
			Lower	Upper
α_0	-0.583	0.004	-0.591	-0.575
α_1	2.168	0.136	1.901	2.435
α_2	0.887	0.026	0.835	0.938
α_3	0.583	0.160	0.270	0.896

*Loglikelihood = -152760.2

Model: $\text{logit}(p_i) = -0.583 + .2.168\text{SDC}_i + 0.887\text{Adenoma}_i + 0.583\text{IC}_i$

Table 5.3.9 Estimated forward (p) and backward (q) probability, the odds ratio of p/q, ruin probability, and the expected steps based on the estimated parameters from Table 5.3.8

State group	p (95% CI)	q (95% CI)	OR	f-Hb (μg/g)	Ruin probability	Expected steps	
						D_1	D_X
Cancer	0.830 (0.762,0.867)	0.170 (0.133,0.208)	8.756	400	0.958	580.49	577.79
Adenoma	0.575 (0.563,0.587)	0.425 (0.413,0.437)	2.426	300	0.455	514.86	893.27
IC	0.500 (0.422,0.578)	0.500 (0.422,0.578)	1.793	300	0.007	299	596
Normal	0.358 (0.356,0.360)	0.642 (0.640,0.644)	1.000	20	$1.9\text{x}10^{-5}$	3.53[a]	7.05[a]

*a: go to 0 μg/g

Table 5.3.10 Estimated regression coefficients and their 95% Cis with the random walk regression model considering three disease statuses, normal, colorectal adenoma, and colorectal cancer with two logistic regression models considering forward (p), backward(q), and no movement (r)

Coefficient	Estimate	Stderr	95% CI	
			Lower	Upper
α_0	0.261	0.005	0.252	0.271
α_1	3.434	0.279	2.886	3.981
α_2	2.280	0.063	2.157	2.404
β_0	0.283	0.005	0.273	0.292
β_1	2.333	0.282	1.781	2.886
β_2	1.843	0.064	1.717	1.969

*Loglikelihood = -253418.9

Model: $\text{logit}(p_i) = 0.261 + 3.434\text{CRC}_i + 2.280\text{Adenoma}_i$

$\text{logit}(q_i) = 0.283 + 2.333\text{CRC}_i + 1.843\text{Adenoma}_i$

Table 5.3.11 Estimated forward (p), backward (q) probability, staying probability (r) the odds ratio of p/q, ruin probability, and the expected steps based on the estimated parameters from Table 5.3.10

State group	p (95% CI)	q (95% CI)	r (95% CI)	OR	f-Hb (μg/g)	Ruin probability	Expected steps	
							D_1	D_X
Cancer	0.733 (0.695,0.768)	0.249 (0.214,0.285)	0.018 (0.009,0.029)	3.010	400	0.884	533.94	714.10
Adenoma	0.575 (0.563,0.587)	0.379 (0.367,0.392)	0.045 (0.040,0.051)	1.551	300	0.565	493.02	816.59
Normal	0.358 (0.356,0.360)	0.366 (0.364,0.368)	0.276 (0.274,0.278)	1.000	20	0.082	17.78[a]	33.92

*a: go to 0 μg/g

Table 5.3.12 Estimated regression coefficients and their 95% Cis with the random walk regression model considering four disease statuses, normal, colorectal non-advanced adenoma, colorectal advanced adenoma, and colorectal cancer two logistic regression models considering forward (p), backward(q), and no movement (r)

State group	Coefficient	SE	95% CI	
			Lower	Upper
α_0	0.261	0.005	0.251	0.272
α_1	3.433	0.304	2.838	4.028
α_2	3.438	0.216	3.015	3.861
α_3	2.102	0.062	1.981	2.223
β_0	0.282	0.005	0.272	0.293
β_1	2.332	0.315	1.715	2.950
β_2	1.251	0.232	0.796	1.705
β_3	1.876	0.063	1.753	1.999

*Loglikelihood = -253197.5

Model: $\text{logit}(p_i) = 0.261 + 3.433\text{CRC}_i + 3.438\text{AdvAdenoma}_i + 2.102\text{Adenoma}_i$

$\text{logit}(q_i) = 0.282 + 2.332\text{CRC}_i + 1.251\text{AdvAdenoma}_i + 1.876\text{Adenoma}_i$

Table 5.3.13 Estimated forward (p), backward (q) probability, staying probability (r) the odds ratio of p/q, ruin probability, and the expected steps based on the estimated parameters from Table 5.3.12

State group	p (95% CI)	q (95% CI)	r (95% CI)	OR	f-Hb (μg/g)	Ruin probability	Expected steps	
							D_1	D_X
Cancer	0.733 (0.695,0.768)	0.249 (0.213,0.286)	0.018 (0.008,0.030)	3.010	400	0.884	533.94	714.08
Adv Adenoma	0.878 (0.856,0.898)	0.101 (0.082,0.120)	0.022 (0.013,0.031)	8.887	300	0.987	333.12	370.17
Non adv Adenoma	0.524 (0.510,0.538)	0.427 (0.413,0.440)	0.049 (0.013,0.027)	1.255	300	0.336	534.64	968.43
Normal	0.358 (0.356,0.360)	0.366 0.364,0.368)	0.276 (0.274,0.278)	1.000	20	0.082	17.78[a]	33.93

*a: go to 0 μg/g

Table 5.3.14 Estimated regression coefficients and their 95% Cis with the random walk regression model considering four disease statuses, normal, colorectal adenoma, screen-detected colorectal cancer (SDC), and interval cancer (IC) with two logistic regression models considering forward (p), backward(q), and no movement (r)

State group	Coefficient	SE	95% CI	
			Lower	Upper
α_0	0.261	0.006	0.250	0.272
α_1	3.210	0.257	2.706	3.715
α_2	2.282	0.068	2.148	2.415
α_3	9.307	6.789	-3.999	22.614
β_0	0.282	0.006	0.272	0.293
β_1	1.440	0.265	0.921	1.959
β_2	1.844	0.069	1.709	1.980
β_3	9.286	6.787	-4.016	22.588

*Loglikelihood = -253380.4

Model: $\text{logit}(p_i) = 0.261 + 3.21\text{SDC}_i + 2.282\text{Adenoma}_i + 9.307\text{IC}_i$

$\text{logit}(q_i) = 0.282 + 1.44\text{SDC}_i + 1.844\text{Adenoma}_i + 9.286\text{IC}_i$

Table 5.3.15 Estimated forward (p), backward (q) probability, staying probability (r) the odds ratio of p/q, ruin probability, and the expected steps based on the estimated parameters from Table 5.3.14

State group	p (95% CI)	q (95% CI)	r (95% CI)	OR	f-Hb (μg/g)	Ruin probability	Expected steps	
							D_1	D_X
Cancer	0.830 (0.791,0.866)	0.144 (0.111,0.180)	0.026 (0.014,0.039)	5.893	400	0.970	468.15	548.39
Adenoma	0.575 (0.563,0.588)	0.379 (0.367,0.392)	0.045 (0.040,0.051)	1.551	300	0.565	493.02	816.57
IC	0.500 (0.242,0.688)	0.500 (0.243,0.686)	0.000 (0,0.057)	1.022	300	0.007	299	596
Normal	0.358 (0.356,0.360)	0.366 (0.364,0.368)	0.276 (0.274,0.278)	1.000	20	0.082	17.78[a]	33.93

*a: go to 0 μg/g

Table 5.3.16 Estimated regression coefficients and their 95% Cis with the random walk regression model considering three disease statuses, normal, colorectal adenoma, and colorectal cancer based on all detection modes (including prevalent screen)

Coefficient	Estimate	SE	95% CI	
			Lower	Upper
α_0	-2.469	0.004	-2.476	-2.462
α_1	5.280	0.085	5.113	5.447
α_2	4.232	0.022	4.189	4.275

*Loglikelihood =-293795.3

Model: $\text{logit}(p_i) = -2.469 + 5.28\text{CRC}_i + 4.232\text{Adenoma}_i$

Table 5.3.17 Estimated forward (p) and backward (q) probability, the odds ratio of p/q, ruin probability, and the expected steps based on the estimated parameters from Table 5.3.16

State group	p (95% CI)	q (95% CI)	OR	f-Hb (μg/g)	Ruin probability	Expected steps	
						D_1	D_X
Cancer	0.943 (0.934,0.952)	0.057 (0.048,0.066)	195.557	400	0.996	422.93	447.31
Adenoma	0.854 (0.848,0.859)	0.146 (0.141,0.152)	69.142	300	0.971	350.04	408.90
Normal	0.078 (0.078,0.079)	0.922 (0.921,0.922)	1.000	20	$2.6\text{x}10^{-9}$	1.19[a]	2.37[a]

*a: go to 0 μg/g

Table 5.3.18 Estimated regression coefficients and their 95% Cis with the random walk regression model considering four disease statuses, normal, colorectal non-advanced adenoma, advanced adenoma, and colorectal cancer based on all detection modes (including prevalent screen)

Coefficient	Estimate	SE	95% CI	
			Lower	Upper
α_0	-2.469	0.004	-2.476	-2.462
α_1	5.280	0.084	5.114	5.445
α_2	6.107	0.098	5.916	6.299
α_3	3.956	0.023	3.912	4.001

*Loglikelihood = -293371.4

Model: $\text{logit}(p_i) = -2.469 + 5.28\text{CRC}_i + 6.107\text{AdvAdenoma}_i + 3.956\text{Adenoma}_i$

Table 5.3.19 Estimated forward (p) and backward (q) probability, the odds ratio of p/q, ruin probability, and the expected steps based on the estimated parameters from Table 5.3.18

State group	p (95% CI)	q (95% CI)	OR	f-Hb (μg/g)	Ruining probability	Expected steps	
						D_1	D_X
Cancer	0.943 (0.934,0.952)	0.057 (0.048,0.066)	196.381	400	0.996	422.93	447.31
Adv adenoma	0.974 (0.969,0.979)	0.026 (0.021,0.031)	449.293	300	0.999	306.83	313.87
Non adv adenoma	0.816 (0.809,0.822)	0.184 (0.178,0.191)	52.244	300	0.949	366.20	447.72
Normal	0.078 (0.078,0.079)	0.922 (0.921,0.922)	1.000	20	0	1.19[a]	2.37[a]

*a: go to 0 μg/g

Table 5.3.20 Estimated regression coefficients and their 95% Cis with the random walk regression model considering three disease statuses, normal, colorectal adenoma, screen-detected colorectal cancer (SDC), and interval cancer (IC) (including prevalent screen)

Coefficient	Estimate	SE	95% CI	
			Lower	Upper
α_0	-2.469	0.003	-2.476	-2.462
α_1	6.047	0.125	5.803	6.292
α_2	4.232	0.022	4.190	4.275
α_3	2.469	0.157	2.161	2.777

*Loglikelihood = -293646.2

Model: $\text{logit}(p_i) = -2.469 + 6.047\text{SDC}_i + 4.232\text{Adenoma}_i + 2.469\text{IC}_i$

Table 5.3.21 Estimated forward (p) and backward (q) probability, the odds ratio of p/q, ruin probability, and the expected steps based on the estimated parameters from Table 5.3.20

State group	p (95% CI)	q (95% CI)	OR	f-Hb (μg/g)	Ruining probability	Expected steps	
						D_1	D_X
SDC	0.973 (0.966,0.979)	0.027 (0.021,0.034)	425.976	400	0.999	410.11	420.54
Adenoma	0.854 (0.848,0.859)	0.146 (0.141,0.152)	69.142	300	0.971	350.04	408.90
IC	0.500 (0.423,0.577)	0.500 (0.423,0.577)	11.821	300	0.993	299	596
Normal	0.078 (0.078,0.079)	0.922 (0.921,0.922)	1.000	20	0.007	1.19[a]	2.37[a]

*a: go to 0 μg/g

Table 5.3.22 Estimated regression coefficients and their 95% Cis with the random walk regression model considering three disease statuses, normal, colorectal adenoma, colorectal cancer (CRC), besides that, making allowance for gender (covariate)

Coefficient	Estimate	Stderr	95% CI	
			Lower	Upper
α_0	-0.597	0.005	-0.608	-0.587
α_1	1.586	0.097	1.396	1.775
α_2	0.878	0.026	0.826	0.930
α_3	0.043	0.010	0.024	0.062

*Loglikelihood =-152779

Model: $\text{logit}(p_i) = -0.597 + 1.586\text{CRC}_i + 0.878\text{Adenoma}_i + 0.043\text{Gender}_i$

Table 5.3.23 Estimated forward (p) and backward (q) probability, the odds ratio of p/q, ruin probability, and the expected steps based on the estimated parameters from Table 5.3.22

State group	Gender	p (95% CI)	q (95% CI)	OR	f-Hb (μg/g)	Ruining probability	Expected steps	
							D_1	D_X
Cancer	Male	0.737 (0.7,0.774)	0.263 (0.226,0.3)	4.882	400	0.873	540.53	731.94
	Female	0.729 (0.691,0.766)	0.271 (0.234,0.309)		400	0.861	546.71	748.84
Adenoma	Male	0.580 (0.568,0.593)	0.420 (0.407,0.432)	2.406	300	0.476	510.85	878.78
	Female	0.570 (0.557,0.582)	0.430 (0.418,0.443)		300	0.430	519.41	909.99
Normal	Male	0.365 (0.361,0.368)	0.635 (0.632,0.639)	1	20	3.1×10^{-5}	3.7[a]	7.4[a]
	Female	0.355 (0.352,0.357)	0.645 (0.643,0.648)		20	1.5×10^{-5}	3.45[a]	6.89[a]

*a: go to 0 μg/g

Table 5.3.24 Estimated regression coefficients and their 95% Cis with the random walk regression model considering three disease statuses, normal, colorectal adenoma, colorectal cancer (CRC), besides that, taking gender as covariate (including prevalence screen)

Coefficient	Estimate	Stderr	95% CI	
			Lower	Upper
α_0	-2.410	0.006	-2.423	-2.398
α_1	5.311	0.085	5.144	5.477
α_2	4.268	0.022	4.225	4.311
α_3	-0.164	0.013	-0.189	-0.139

*Loglikelihood = -293557

Model: $\text{logit}(p_i) = -2.41 + 5.311\text{CRC}_i + 4.268\text{Adenoma}_i - 0.164\text{Gender}_i$

Table 5.3.25 Estimated forward (p) and backward (q) probability, the odds ratio of p/q, ruin probability, and the expected steps based on the estimated parameters from Table 5.3.24

State group	Gender	p (95% CI)	q (95% CI)	OR	f-Hb (μg/g)	Ruining probability	Expected steps	
							D_1	D_X
Cancer	Male	0.939 (0.929,0.949)	0.061 (0.051,0.071)	202.495	400	0.996	424.78	451.24
	Female	0.948 (0.939,0.956)	0.052 (0.044,0.061)		400	0.997	420.88	442.97
Adenoma	Male	0.845 (0.839,0.850)	0.155 (0.15,0.161)	71.356	300	0.966	353.71	417.57
	Female	0.865 (0.86,0.87)	0.135 (0.130,0.140)		300	0.976	345.45	398.20
Normal	Male	0.071 (0.07,0.072)	0.929 (0.928,0.93)	1	20	0	1.17[a]	2.33[a]
	Female	0.082 (0.081,0.083)	0.918 (0.917,0.919)		20	0	1.2[a]	2.39[a]

*a: go to 0 μg/g