Multiscale Modeling of Esophageal Adenocarcinoma

Kathleen Mary Curtius

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2015

Reading Committee:

E. Georg Luebeck, Chair

Hong Qian

John Inadomi

Program Authorized to Offer Degree:
Department of Applied Mathematics

University of Washington

Abstract

Multiscale Modeling
of Esophageal Adenocarcinoma

Kathleen Mary Curtius

Chair of the Supervisory Committee:
Professor E. Georg Luebeck
Department of Applied Mathematics

Over the past three to four decades, esophageal adenocarcinoma (EAC) incidence has increased dramatically in the Western world due to causes that are not well understood. Current screening strategies for early detection aim to identify individuals with Barrett's esophagus (BE), an intestinal metaplasia that develops in the lower esophagus as an important first step in the progression to EAC. However, current approaches for prevention of EAC by screening and surveillance programs have achieved minimal success in reducing mortality and paradoxically yield underdiagnosis and overdiagnosis.

In order to better understand these issues, we consider the influences of critical processes at multiple spatial scales in an effort to bridge molecular, cellular and tissular knowledge to population-level data related to BE and the progression of BE to EAC. Specifically, the mathematical framework presented here cohesively models biological mechanisms that include epigenetic drift, cellular dynamics, clonal growth, crypt structured organization in BE, spatial propagation of premalignant and malignant lesions, surveillance through biopsy and imaging, and clinical interventions. With the multiscale modeling approach, we can better understand the role and impact that different levels of data have on clinical outcomes. Our modeling aim is to ultimately improve the efficacy of screening and surveillance to reduce EAC mortality.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First, I would like to sincerely thank my advisor, Georg, for his immense guidance, support, and countless hours spent meeting with me on these research projects. His helpful discussions have taught me the true meaning of collaboration.

I would also like to thank the other members of my supervisory committee, John, Suresh, Hong, and Vladimir, who constitute a group of scientists that I aspire to emulate in your rigor and creativity. You have each contributed in unique ways to my research and the projects would not have progressed without your influences.

To my colleagues at the Fred Hutchinson Cancer Research Center, I thank Bill Hazelton, Jihyoun, Bill Grady, Tom Vaughan, Tom Paulson, and Chao-Jen for providing a wonderful environment for furthering the collective goal of understanding and curing cancer. Thank you to Anup Dewanji for our intensive mathematical discussions and derivations scribbled on Fred Hutchinson café napkins. Additionally, I would like to acknowledge Ken Wang at Mayo Clinic for his openness to help an inquisitive mathematician.

During my PhD, I have been fortunate to be a part of the Cancer Intervention and Surveillance Modeling Network (CISNET) consortium within the National Cancer Institute. I have greatly benefited from the meetings and conference calls, and enjoyed the social engagements with the CISNET Esophagus group. I am very grateful to have had the unique ability to share my thoughts, misgivings, questions, and opinions with a group of such influential cancer modelers, physicians, and policy-makers. Thank you for listening to and incorporating the mathematical side of our efforts these past years. Here I will also thank the National Science Foundation for their generous support in a Graduate Research Fellowship, grant no. DGE-0718124, during my graduate career at the University of Washington.

Further, I am very grateful for my wonderful experience in the Department of Applied Mathematics at the University of Washington. Thank you to the inspiring faculty for your prestigious instruction and humor. Thank you also to the Applied Math graduate students for their constant motivation and help in my research. We are all in this together. Specifically, I would like to thank my officemate, Natalie, for being the friend I needed and respected in the high and low times during graduate school.

To my family, Mom, Dad, Mike, Eileen, and Tim, thank you for your constant love and open minds when I have seemed to be working in a foreign, mathematical universe. My mom Mary has been my spiritual hero since I first learned how to count. As an intelligent, strong woman, you believed in me through any times of my own self-doubt and I admire you. I also thank my dear friends here in Seattle, California, and beyond who encouraged me to be the best version of myself in every respect.

Finally, I would like to express my dearest appreciation for my fiancé Ben. Thank you for cooking all of our meals and supporting me in a way that I did not know was possible. Our connection is more beautiful than words can express.

to Ben,

my love and partner

NOTATION AND ABBREVIATIONS

EAC	:	Esophageal Adenocarcinoma
BE	:	Barrett's Esophagus
ESCC	:	Esophageal Squamous Cell Carcinoma
NS	:	Normal Squamous
GEJ	:	Gastroesophageal Junction
ND	:	No Dysplasia
HGD	:	High Grade Dysplasia
LGD	:	Low Grade Dysplasia
CISNET	:	Cancer Intervention and Surveillance Modeling Network
TSCE	:	Two-Stage Clonal Expansion
MSCE	:	Multistage Clonal Expansion
MSCE-EAC	:	Multistage Clonal Expansion for EAC
b-d-m	:	birth-death-mutation
MSCE–1	:	Multistage Clonal Expansion approximation for malignant progression
CRC	:	Colorectal Cancer
PDE	:	Partial Differential Equation
ODE	:	Ordinary Differential Equation
PGF	:	Probability Generating Function
$S_{MSCE}(t)$	:	Survival from cancer from MSCE process by time t
$P_{MSCE}(t)$	:	Probability of cancer occurrence from MSCE process by time t

$h_{_{MSCE}}(t)$	:	Hazard for symptomatic cancer detection from MSCE process by time t
$f_{_{MSCE}}(t)$	:	Density function for cancer detection from MSCE process by time t
IVP	:	Initial Value Problem
IC	:	Initial Condition
MLE	:	Maximum Likelihood Estimate
MCMC	:	Markov Chain Monte Carlo
GaC	:	Gastric Cancer
PaC	:	Pancreatic Cancer
$S_{_{EAC}}(t)$	:	Survival from cancer from MSCE-EAC process by time t
$P_{_{EAC}}(t)$	:	Probability of cancer occurrence from MSCE-EAC process by time t
$h_{_{EAC}}(t)$	:	Hazard for symptomatic cancer detection from MSCE-EAC process by time t
$f_{_{EAC}}(t)$	:	Density function for EAC detection from MSCE-EAC process by time t
$\Lambda_{_{EAC}}(t)$	:	Cumulative hazard for cancer detection from MSCE-EAC process by time t
OCT	:	Optical Coherence Tomography
SSA	:	Stochastic Simulation Algorithm
SAL	:	Step Anticipation Leap
EET	:	Endoscopic Eradication Therapy
EMR	:	Endoscopic Mucosal Resection
RFA	:	Radiofrequency Ablation

CE-D	:	Complete Eradication of Dysplasia
CE-IM	:	Complete Eradication of Intestinal Metaplasia (BE)
NNT/death	:	Number Needed of Treatments to prevent one death
EGD	:	Esophagogastroduodenoscopy ("upper endoscopy")
FPP	:	Filtered Poisson Process
CI	:	(MCMC estimate) Credible Interval

Chapter 1

Introduction

1.1 The Problem of Esophageal Adenocarcinoma

Esophageal cancer is currently the fastest growing cancer in the Western world [1,2]. Among US males, it is also the seventh leading cause of cancer death [3]. The two major types of esophageal cancers are esophageal squamous cell carcinoma (ESCC), which typically occurs in the mid-portion of the tube-like esophagus, and esophageal adenocarcinoma (EAC), which occurs at the distal end of the esophagus near the junction to the stomach. While the incidence of most cancers including ESCC have been declining in the US, the incidence rate of EAC has increased 5-6 fold over the past four decades (see Fig. 1.1) [4,5]. Although lifestyle factors (eg., abdominal obesity, smoking, alcohol use) have been implicated as contributors to this trend, the observed increase of this magnitude remains largely unexplained [6].

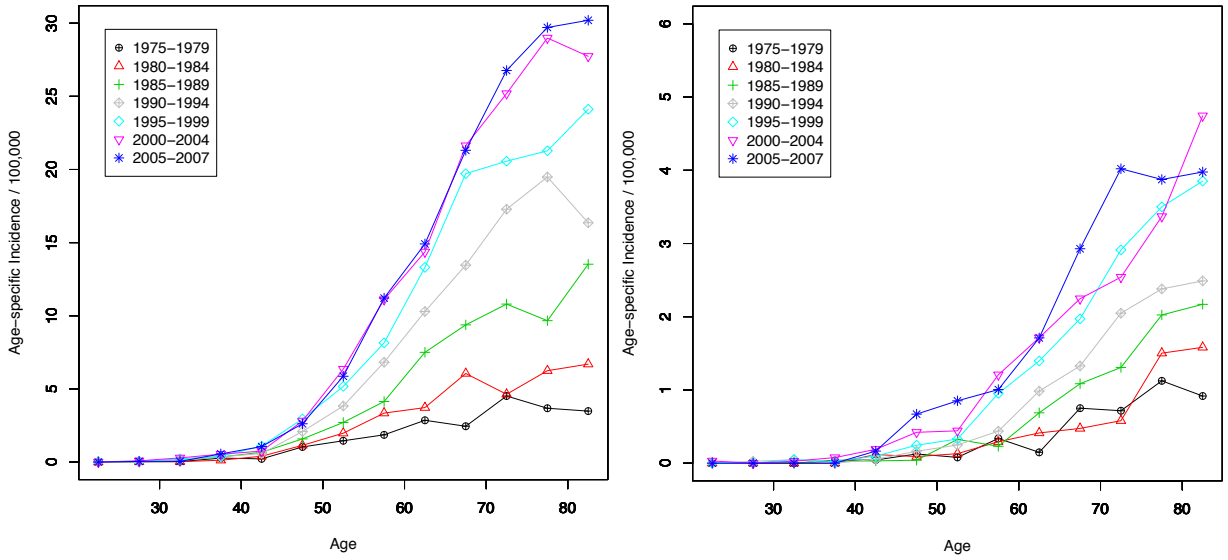

Figure 1.1: **Age-specific EAC incidence rates in the US.** For all stages of EAC cancers combined, age-specific incidence rates per 100K persons are plotted for males (left panel) and females (right panel) in the US, Caucasian race. The incidence curves are stratified into 5-year calendar periods to depict the dramatic 5-6 fold increase in incidence rates for both genders since 1975. Data available from Surveillance, Epidemiology, and End Results (SEER) Program [7].

EAC arises primarily (if not exclusively) in Barrett's esophagus (BE), a metaplastic tissue alteration in the esophageal lining. In BE, a layer of columnar cells replace an area of normal, stratified squamous epithelium and form crypts that are similar to intestinal crypts (Fig. 1.2). BE is frequently associated with chronic symptoms of gastroesophageal reflux disease (GERD), which exposes the distal esophagus to bile salts and stomach acid. Adenocarcinomas develop within BE in a multistage process involving morphologic changes in the tissue. BE progresses from nondysplastic (ND) columnar epithelium, through low grade and then high grade dysplasia (a precancerous and pre-invasive cell type detected under the microscope), and finally form small malignant cell populations that progress to invasive

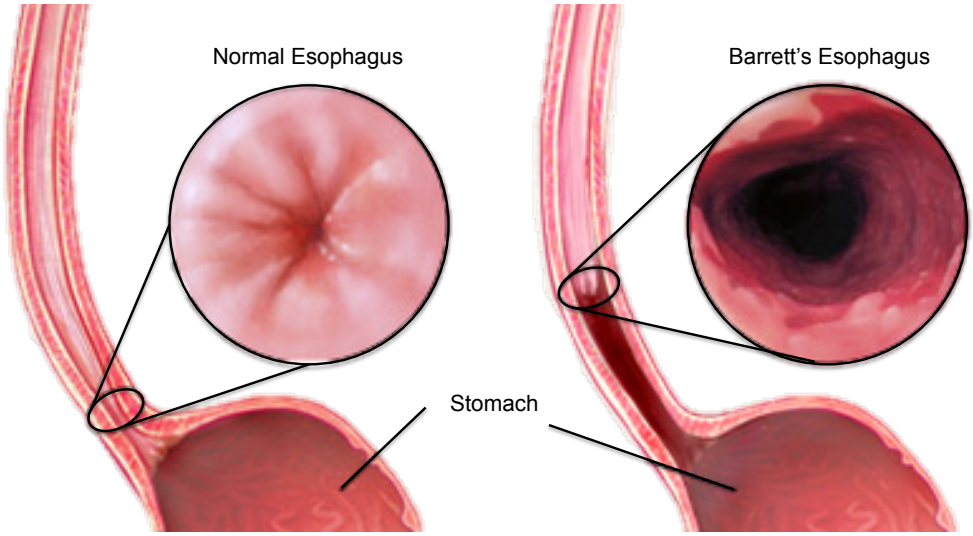

Figure 1.2: **Normal lower esophagus and Barrett's esophagus.** In BE patients, a certain length of normal squamous epithelium in the lower esophagus (left inset) is replaced by the salmon-colored, intestinal metaplasia (right inset) at the distal end of the esophagus before the gastroesophageal junction (GEJ). Modified image from [8].

cancer. Because BE is asymptomatic, the majority of BE patients remain undiagnosed and thus most EAC cases are diagnosed at an advanced stage. This is an unfortunate reality because mortality associated with EAC is very high (about one half of diagnosed patients die within a year of diagnosis [9]). Overall, underdiagnosis is a major concern because more than 90% of EAC cases arise in patients who are not periodically screened or in surveillance programs [10, 11].

The main concept of screening is that early detection of a disease offers the opportunity to change its prognosis; compared with incidental cancers, the prognosis is greatly improved for BE patients with high grade dysplasia (HGD) and/or cancer that is detected at an early stage. Therefore, gastroenterologists focus on identifying BE patients on screening endoscopy and advising these patients to undergo long-term, periodic endoscopic surveillance

with multiple biopsies of the BE to detect advanced dysplasia or early cancer. However, on average only 0.2-0.5% of people with BE develop EAC each year [12]. Thus, the majority of BE patients who undergo regular surveillance will not develop EAC in their lifetimes, indicative of overscreening.

Herein lies the balancing act performed during *risk stratification* - identify who is most 'at risk' of progressing to EAC and suggest effective surveillance and treatment strategies for the high risk groups. With this motivation, the overall public health goal of the mathematical modeling presented in this thesis is to help improve the efficacy of screening and surveillance for BE and EAC. Although clinical endpoints like cancer incidence and prevalence of HGD are reported on a population level, many important biological processes on smaller physical and temporal scales occur with significant inter-individual heterogeneity during disease progression from normal tissue to incident cancer. Due to the biological and clinical nature of screening patients at various times during their lives, such details at tissular and cellular levels provide vital information to determine screening outcomes obtained by different modalities and protocols. In order to model these aspects accurately, we develop a multiscale framework for modeling events on population, individual, tissue, cell, and molecule levels at appropriate time scales (see Fig. 1.3).

In Chapter 2, we introduce the theory of multistage carcinogenesis and then develop the mathematical tools used to analyze a general multistage clonal expansion (MSCE) model. With the MSCE model, we show that cancer incidence curves harbor information about hidden processes of tumor initiation, premalignant clonal expansion, malignant transformation, and even some limited information on tumor growth before clinical detection (see [13] for full publication). Through calibration of the mathematically derived hazard functions to incidence data, our analyses of the incidences of three digestive tract cancers show that the age-specific incidence curves- upon adjustments for secular trends- are well approximated by a model that explicitly incorporates the stochastic growth kinetics of premalignant clones, the appearance of a first viable malignant clone, and a constant time delay corresponding to the mean sojourn time of a malignant clone.

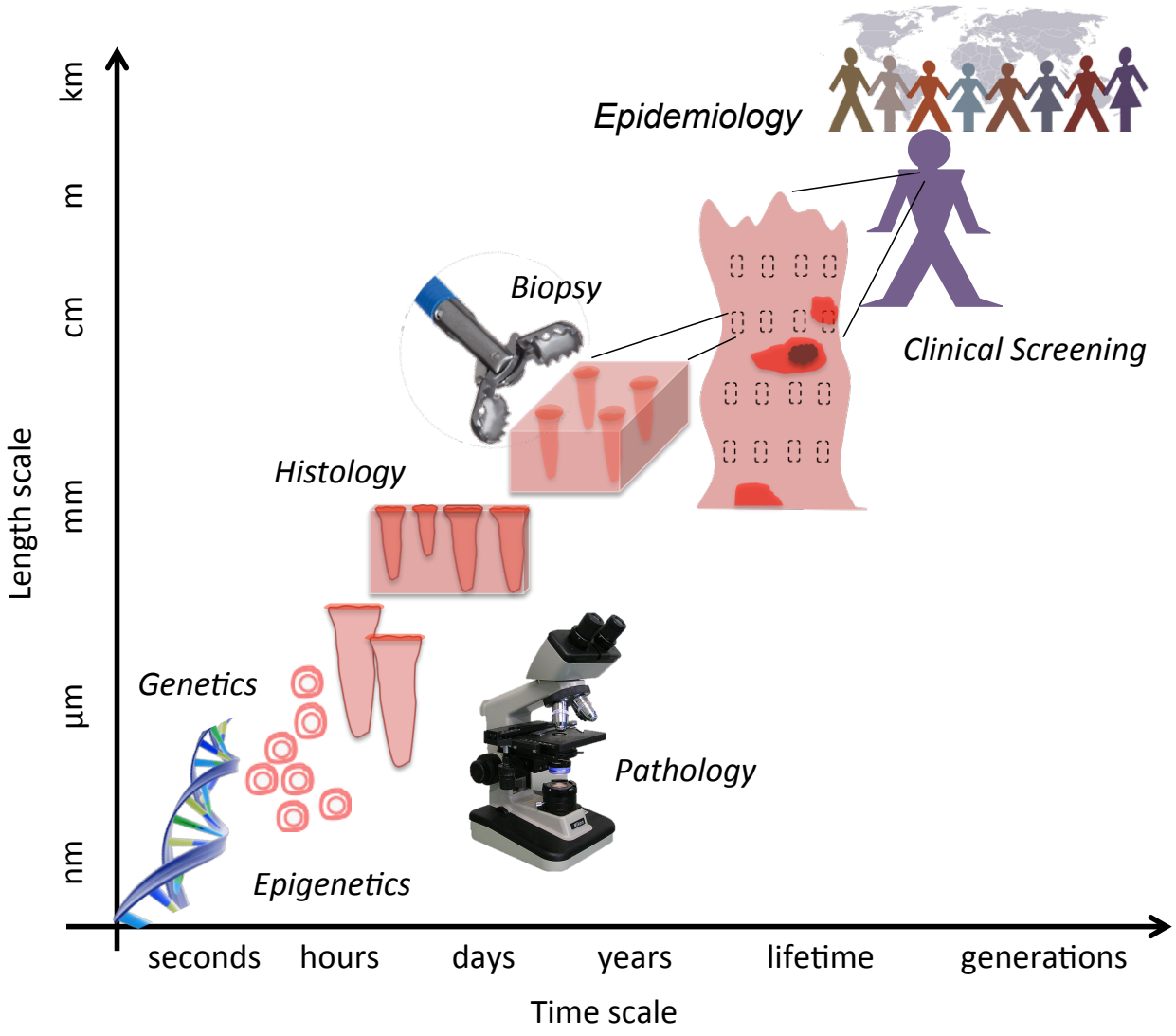

Figure 1.3: **Multiscale modeling of EAC.** Levels of size, from molecules to populations, undergo dynamics at different time scales and each influence the public health concerns surrounding fast-rising EAC incidence rates and associated high mortality.

With the methodology introduced in Chapter 2, we further extend the MSCE model in Chapter 3 to apply to the more specific problem of EAC. Here we develop the MSCE-EAC cell module, composed of three components - incorporation of gastroesophageal reflux into the BE conversion process, derivation of the MSCE-EAC hazard function, and likelihood estimation of model parameters. To address the question of predicted future incidence trends in the US, we calibrate the MSCE-EAC model to clinical and epidemiologic data including EAC incidence from the Surveillance, Epidemiology, and End Results (SEER) registry from 1975 to 2010. We then project EAC incidence and mortality to year 2030 (see [6] for full publication). This analysis was performed in parallel with two independent modeling groups as part of the Base Case I comparative modeling project within the Cancer, Intervention, and Surveillance Modeling Network (CISNET) Consortium (see next section for more details). Overall, the future projected increases in cumulative EAC deaths and incidence give rise to significant concerns about the burden to society.

Because current biopsy-based endoscopic screening involves the removal of multiple esophageal tissue samples for histological analysis, the multiscale framework requires details about the BE tissue structure and specified screening protocols. Chapter 4 develops two additional modules that, together with the cell module, constitute the multistage clonal expansion for EAC (MSCE-EAC) screening model. The MSCE-EAC screening model is used for screening BE patients *in silico* to evaluate the effects of biopsy sampling, diagnostic sensitivity, and treatment on disease burden. Our framework seamlessly integrates relevant cell-level processes during EAC development from Chapters 2-3 with a spatial screening process to provide a clinically relevant model for detecting dysplastic and malignant clones within the crypt-structured BE tissue. With this computational approach, we retain spatio-temporal information about small, unobserved tissue lesions in BE that may remain undetected during biopsy-based screening but could be detected with high-resolution imaging. This allows evaluation of the efficacy and sensitivity of current and future screening protocols to detect neoplasia (defined as dysplasia and early preclinical EAC) in the esophageal lining. We demonstrate the clinical utility of this model by predicting three important clinical outcomes:

(1) the probability that small cancers are missed during biopsy-based screening, (2) the potential gains in neoplasia detection probabilities if screening occurred via high-resolution tomographic imaging, and (3) the efficacy of endoscopic eradicative treatments that result in the curative depletion of metaplastic and neoplastic cell populations in BE in terms of the long-term impact on reducing EAC incidence (see [14] for full publication).

To further explore this final point on treatment for premalignant lesions, we utilize the MSCE-EAC screening model in Chapter 5 to predict the long-term impact of successive eradicative treatments and recurrent disease on EAC incidence and mortality reduction. Although endoscopic eradication of BE is effective in reducing EAC incidence and mortality in BE patients, we find that the resources needed to achieve these results increase substantially as patients with lower severity of disease are selected for treatment. Diminished efficiency and the large number of persons needed to treat to avert one death (NNT/death) suggest that overall cost is likely to be prohibitive for treatment of BE patients other than those with HGD. This analysis was performed as the second part, Base Case II, of the comparative modeling project within the CISNET esophagus group (see [15]).

Rather than focusing on fixed treatment strategies, Chapter 6 introduces a methodology for optimal screen design based on achieving maximal diagnostic yields during a screen. Our aim was to calculate optimal screening age recommendations that would avoid fruitless over screening throughout patient lifetimes. For the MSCE-EAC model structure, we derive optimality criteria for both single screen scenarios and adaptive screen scenarios, i.e., those that are conditional on the outcome of the previous screen for optimization of a next screening time. Chapter 6 presents a few example results in context of the MSCE-EAC model. This methodology provides a tool for quantifying the potential for the greatest screen yield during BE surveillance, with explicit incorporation of disease risk and patient-specific outcomes at a given screening time.

Although such screens may reveal that a patient has BE, the duration of time a patient has lived with this precursor (BE dwell time) is generally not known when the patient is first diagnosed because BE is asymptomatic. This is particularly unfortunate because BE

dwell time is expected to be a strong predictor for EAC. Our mathematical modeling of EAC suggests that the life-time risk of developing EAC could vary widely depending on how long a patient has lived with BE. Thus, biomarkers that drift differentially with age between normal esophageal tissue and BE have the potential to greatly improve EAC risk estimation. In Chapter 7, we evaluate over 450,000 CpG dinucleotides, genome-wide, in tissue samples of 40 BE patients to identify a robust subset of CpGs that undergo age-related methylomic drift allowing us to estimate a patient's BE dwell time. To do so, we use a unique combination of longitudinal and cross-sectional tissue samples to generate data that allow us to construct a BE-specific molecular clock. We identify 70 CpGs that are hypomethylated in normal squamous (NS) tissue but show significant differential increases in methylation in BE (vs. NS) with age, consistent with the drift observed in the longitudinal samples (10 BE patients). Using a Bayesian model that incorporates longitudinal methylomic drift rates, patient age, and the observed paired BE methylation levels, we estimate patient-specific BE onset times using Markov Chain Monte Carlo (MCMC). Independent application of this method to a cohort of 22 familial BE (FBE) patients as well as a group of 10 BE patients with neoplasia, specifically HGD and/or EAC, will reveal potential differences in the 3 data sets' estimated onset times. As part of current work, our analyses are testing the conjecture that methylomic drift occurs in BE compared to NS tissue and hence allows for an estimation of how long a BE patient has lived with BE. Thus, we are testing if methylomic drift may be used as a quantitative biomarker candidate for HGD/EAC risk in BE patients (see [16]).

In summary, we develop a mathematical, multiscale framework that captures overall EAC incidence while addressing questions regarding the efficacy of screening protocols, the potential improvements possible with new screening techniques, the optimality of screening times, and the potential of a new risk predictor to stratify high risk patients for EAC based on patient-specific epigenetic data.

1.2 Cancer Intervention and Surveillance Modeling Network (CISNET) Consortium

Generally, mathematical modeling of clinical outcomes can be used to make projections by systematically integrating available data. Despite its potential, one common criticism of modeling is that independent modeling efforts often yield disparate results. These differences can primarily be attributed to differing model inputs and model structures, and often a lack of transparency in model assumptions, all impeding wider acceptance of simulation modeling results. The Cancer Intervention and Surveillance Modeling Network (CISNET) is a National Cancer Institute (NCI) funded consortium of mathematical and computational modelers who work closely together to address many of these limitations using a comparative modeling approach. Comparative modeling aims to improve each of the participating models by providing an environment where experts in modeling methodology, cancer control, and clinical management can easily and openly communicate intricate modeling details and results with independent groups. This context allows for an iterative process where some common calibration targets are shared but other aspects of the individual models are free to differ. After each model is calibrated and applied independently, the predictions are shared and analyzed cooperatively in biyearly meetings, providing a transparent setting for iterative improvement of the models, and enhancing our understanding of the natural history of a particular type of cancer. After model development and refinement, CISNET researchers can then use their models to evaluate the impact of potential interventions on trends in cancer incidence and mortality, helping to optimize cancer control strategies. CISNET models have been utilized by the U.S. Preventive Services Task Force for breast [17] and colorectal cancer (CRC) screening guidelines [18], in formulating draft recommendations for lung cancer screening [19], and by the Centers for Medicare and Medicaid Services (CMS) to compare the effectiveness of CRC screening strategies [20, 21].

The CISNET consortium focuses on 5 cancer sites: breast, colorectal, lung, prostate, and esophagus. The esophagus group is composed of 4 academic institutions that have

developed 3 independent models of EAC. In Chapters 3 and 5, we present our multiscale modeling results from Base Case I and Base Case II analyses, respectively. We compared our results (referred to as the *FHCRC model* results in those publications) to the two other mathematical models that were developed independently by participants in the CISNET esophagus group: the Esophageal AdenoCarcinoma Model (EACMo) from the Massachusetts General Hospital (Boston, MA) - *MGH model*, and University of Washington (Seattle, WA) and the Microsimulation Screening Analysis model from Erasmus University Medical Center (Rotterdam, The Netherlands) - *UW-MISCAN model*. Full descriptions of the MGH model and the UW-MISCAN model can be found on the Esophagus section of the CISNET website [22] and in the Appendix of [6], but will be not be discussed in detail here. Briefly, while we employ a multiscale framework for modeling multiple levels of detail, the MGH and UW-MISCAN models are empirically-based microsimulations of natural histories.

Chapter 2

The Multistage Clonal Expansion Model with Applications

2.1 Modeling Motivation: The Multistage Nature of Cancer

Carcinogenesis, the transformation of normal cells into cancer cells, has been shown to be a multistage process in which stem cells accumulate a series of rate-limiting, (epi)genetic changes. Although it is difficult to know the exact steps of the process, experimental evidence suggests that most malignancies can be analyzed under an initiation-promotion-progression paradigm [23]. Initiation is characterized by genetic (or epigenetic) alterations that produce heritable changes in a cell's phenotype. Initiated cells with such alterations are prone to clonal expansion. The next phase, promotion, is reversible and refers to the outgrowth of initiated cells into premalignant lesions. Progression refers to the conversion and development of malignant cells, their growth into tumors, and finally the onset of the clinical disease.

Multistage models of carcinogenesis provide a quantitative description of these salient phases (possibly consisting of many steps within each) and have majorly contributed to the advancement of cancer research [24–26]. Multistage models are able to both elucidate the

actual biological processes taking place during carcinogenesis and allow the investigation of the effects of carcinogens, risk factors, and other covariates on the stochastic processes involved in cancer initiation, promotion and progression. Thus, multistage models provide a natural framework to evaluate the potential benefits of prevention and intervention strategies designed to reduce cancer risk, like screening for premalignant lesions and eradicative treatment of small malignancies. For a thorough biological rationale and modeling history, we refer the reader to more comprehensive reviews on multistage models [24–27].

2.1.1 Hazard or age-specific incidence function

The standard statistical concept in survival analysis that captures the instantaneous rate of detecting cancers (of a specific type) among individuals who have not been previously diagnosed with that cancer is the hazard function. Let T_C be the random variable for age at cancer detection. Then the hazard function is defined mathematically at time (age) t as

$$h(t) = \frac{1}{\Delta t} \lim_{\Delta t \to 0} \Pr[t < T_C < t + \Delta t | T_C \geq t] \tag{2.1}$$

Thus, the hazard is a theoretical representation of the observed age-specific incidence, as shown for EAC in Fig. 1.1.

The hazard can also be formulated in terms of the survival function, or the probability that cancer has not been detected by time $t : S(t) \equiv \Pr[T_C > t]$,

$$h(t) = \frac{1}{\Delta t} \lim_{\Delta t \to 0} \Pr[T_C < t] - \Pr[T_C > t + \Delta] = \frac{-\dot{S}(t)}{S(t)} \tag{2.2}$$

We will be referring to such hazard and survival functions for cancer throughout the following chapters.

2.1.2 Brief notes on the history of multistage cancer models

To give some context to the modeling developed in this thesis, we provide a brief summary of the key points in multistage modeling history. Beginning in the 1950s, the multistage

theory was originally suggested independently by Muller [28] and Nordling [29]. Nordling noted that the log-log linearity nature of some cancer mortality curves (i.e., log mortality is linearly related to log age) suggested that several mutations were needed for malignant transformation. This prompted Armitage and Doll [30] to incorporate this idea into a stochastic model of carcinogenesis in which a single stem cell develops into a cancer cell by suffering a series of irreversible, heritable, mutation-like (also referred to as "rate-limiting") events. By deriving the hazard (incidence) rate, they showed that the power of age observed in cancer mortality may be related to the number of transformations required to produce a malignancy. Although the Armitage-Doll model is broadly consistent with the epidemiology of several cancers, it predicts a higher number of mutations (stages) required for malignant transformation than experimental evidence has shown [26]. To remedy this, Armitage and Doll proposed a two-stage model that incorporated cell growth dynamics and assumed that intermediate-stage cells grew exponentially [31]. After Armitage and Doll's first modeling efforts, more multistage models by Fisher, Kendall, and others were studied, most of which included cellular proliferation, a critical process in carcinogenesis [32, 33]. For more early history, we refer the reader to a 1978 review by Whittemore and Keller of the mathematical models of carcinogenesis developed since the theory began [26].

In 1971, Knudson proposed a two-stage statistical model to explain the incidence of retinoblastoma in children [34]. Knudson derived the "two-hit" hypothesis from his model, a concept that states that the two alleles of a single gene have to be inactivated to produce a tumor in the retina. As one of the most important contributions of mathematical and statistical modeling to cancer research, Knudson's work led to the concept of tumor suppressor genes, which was confirmed in laboratory experiments soon after [35]. Moolgavkar and Venzon incorporated the ideas of Knudson into a mathematical two-stage model, which is consistent with the incidence of most cancers in both children and adults [36]. Similarly to Kendall and others, Moolgavkar and Venzon modeled the clonal expansion of premalignant cells as a stochastic birth-death process. To obtain analytical expressions for the hazard and survival curves of their model, Moolgavkar and Venzon used the theory of filtered Poisson

processes, a methodology we will employ throughout the chapters and projects presented here. As shown in Moolgavkar and Knudson [37], the now called two-stage clonal expansion (TSCE) model was the first carcinogenesis model to be consistent with the epidemiology of several specific cancers (lung, breast, colon) and also with the data from carcinogen experiments and has since been widely used for the analysis of epidemiological [38–45] and experimental [46–48] data. In addition, mathematical extensions of the Armitage-Doll and TSCE models have been proposed to accommodate features of different cancers (eg., esophagus). In the following sections, we introduce a general formulation of this type of multistage model.

2.2 The Generalized Moolgavkar-Venzon-Knudson (MVK) Model

In the past 60 years, variants of multistage carcinogenesis models have been analyzed that each formulate the stepwise progression to cancer mathematically as a stochastic, multi-type branching process. Here we will focus on the methods and applications of single pathway multistage models but other variants such as multiple pathway and mixed models have also been explored (see text by Tan [25] for additional mathematical and biological background).

To encompass a large set of examples, Little [49] described generalizations of the two-stage model of Moolgavkar, Venzon, and Knudson, as he terms the MVK model, and of the Armitage-Doll multistage model to allow for an arbitrary number of mutational stages. Along with possible mutations occurring at rate $\mu_i(t)$, the cells of type i compete with processes of symmetric division and differentiation/death at stage-specific rates $\alpha_i(t)$, $\beta_i(t)$, respectively. This biological process, referred to as "clonal expansion", is formulated mathematically as a stochastic birth-death-mutation (b-d-m) process and all rates have units of per cell per year. The difference between these two generalized models factors into the cell counts of each type after a mutational event occurs with rate $\mu_i(t), i = 1, ..k-1$. Within the *generalized*

multistage framework, a cell in mutational stage I_i acquires mutations through an "Armitage-Doll" (AD) type of asymmetric cell division (one daughter cell acquires mutation $i+1$ and enters stage I_{i+1}, the other daughter cell dies or differentiates). Within the *generalized MVK* framework, a cell in mutational stage I_i acquires mutation through an asymmetric Poisson-process (PP) division (one daughter cell acquires mutation $i+1$ and becomes type I_{i+1}, the other daughter cell remains in mutational stage I_i). For the analyses provided in this chapter, we focus on the MVK type models, illustrated in Fig. 2.1.

The two-mutation model of Moolgavkar, Venzon, and Knudson, i.e., the two-stage clonal expansion (TSCE) model, corresponds to the case of $k = 2$. To note, Moolgavkar and Venzon first formulated this two-stage model with stochastic clonal expansion of both normal and intermediate cells [36]. Historically, we refer to intermediate stem cells that enter a stage that may undergo clonal expansion as "initiated" cells. However, the authors also introduced the mathematical techniques to analyze a two-stage model with deterministic growth of normal cells and stochastic growth of intermediate cells, which Moolgavkar and Knudson applied to data from epidemiological studies [37]. Here we will focus on the deterministic case that considers the number of normal cells $N(t)$ to be non-random.

The generalized MVK model also includes common extensions of the TSCE model with several "pre-initiation" stages that do not undergo clonal expansion. These models with multiple pre-initiation stages before a single clonal expansion stage correspond to the generalized MVK model in which the intermediate cell division and differentiation/death rates for stages $1, .., k-2$ equal zero before a single clonal expansion at stage $k-1$. When considering colorectal cancer incidence, Luebeck and Moolgavkar [41] show that a k-stage model with $k = 4$, including two rare and one fast event before initiation, was most consistent with the data among a class of models with several pre-initiation stages and a single clonal expansion. Similar biologically-motivated multistage models that require more than one pre-initiation stage with aymmetric divisions to accrue mutations have been studied previously [49–57] and are now generally termed as multistage clonal expansion (MSCE) models.

In the following section, we will present a generalized MVK model-driven investigation

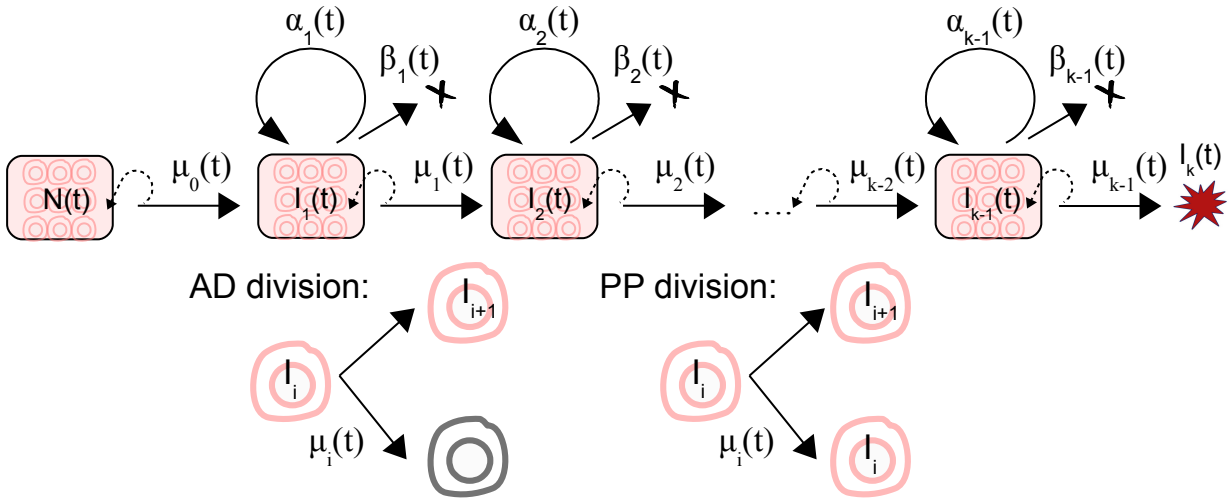

Figure 2.1: **The generalized MVK model.** The generalized MVK model for cancer assumes that at age t there are $N(t)$ susceptible stem cells, each subject to irreversible mutation to a type of cell in stage I_1 at a rate $\mu_0(t)$. The cells with one mutation divide into two such cells at a rate $\alpha_1(t)$, die or differentiate at a rate of $\beta_1(t)$, or symmetrically divide at a rate $\mu_1(t)$ into an equivalent daughter cell in stage I_1 and another cell with a second irreversible mutation now in stage I_2. The cells in stage I_2 are independently subjected to competing events of division, death/differentiation, and mutation with rates $\alpha_2(t), \beta_2(t), \mu_2(t)$, respectively, and so on until at the $(k-1)$th stage the cells that have accumulated $(k-1)$ mutations proceed at a rate $\mu_{k-1}(t)$ to acquire another mutation and enter final stage of cancer I_k. All mutations for the MVK framework are modeled as Poisson process mutations (PP divisions). In contrast, a generalized multistage model would model mutations as Armitage-Doll mutations (AD divisions) and transitions would not include the dotted arrows that retain the cell count in state I_i after undergoing division at rate $\mu_i(t)$.

of cancer incidence in which we analyze SEER-9 [7] incidence data (1975-2008) for three gastrointestinal malignancies: colorectal cancer (CRC), gastric cancer (GaC), and pancreatic cancer (PaC). For the full publication, see [13].

2.3 Impact of Tumor Progression on Cancer Incidence Curves

In theory, any number of extensions and permutations can be formulated within the framework of the generalized MVK model, whether in number of pre-initiated stages and/or number of clonally expanding stages, but it is not fully clear how this choice influences the age-specific incidence curve. Here, we ask the basic question, how do the rate-limiting steps involved in tumor initiation, malignant transformation, and ensuing clonal expansions influence the shape of the cancer incidence curve? Conversely, what can we possibly learn from observed incidence curves about these hidden processes? Previous studies [41, 50–52] identified two characteristic features, or phases, in the incidence curves for colorectal and pancreatic cancers using data from the Surveillance Epidemiology and End Results (SEER) registries [7] when using a 3-stage ($k = 3$) MSCE model with a single clonal expansion for premalignant cells. After adjusting the incidence for secular trends related to birth-cohort and calendar-year (period), the authors were able to identify an exponential phase in the incidence curve beginning in early adult life and extending to approximately the age of 60 and a linearly-increasing trend for later-onset cancers extending beyond the age of 60. In this study we ask the question whether the impact of malignant growth and fitness (defined as clone survival) on observed incidence patterns is actually discernible from incidence data. To address this question, we use a 4-stage generalized MVK model which explicitly incorporates distinct (but overlapping) clonal expansions for premalignant and malignant cells giving rise to a distribution of malignant tumors in a tissue and clinical observation of cancer via a stochastic detection event occurring in a preclinical tumor. In contrast, earlier versions of

the MSCE model assumed that the first malignant cell in a tissue necessarily leads to clinical detection after a possibly random lag-time [41, 50–52] (the 3-stage model by Jeon et al. [50] for EAC will be discussed in more detail in Chapter 3). Malignant transformations, however, are likely to occur in altered cells whose initial survival fitness may be compromised by genomic instability [58] and therefore may be prone to stochastic extinction in spite of higher cell proliferation. This is supported by comparative measurements of cell division rates and net cell proliferation (using DNA labeling and radio-graphic imaging of tumors, respectively) in a variety of carcinomas, showing large differences in the two rates which can only be explained by the frequent death or shedding of tumor cells [59].

We begin by adjusting for period and birth cohort effects, using rigorous likelihood based methods to estimate model parameters for the extended MSCE model, including malignant clonal expansion rates for each cancer type. We then estimate three characteristic times: (1) the mean sojourn times for premalignant clones until occurrence of the first malignant cell regardless of its fate, (2) the analogous mean sojourn time to appearance of the first surviving (persistent) malignant clone; and (3) the mean sojourn time of persistent preclinical cancers from first malignant cell to time of cancer diagnosis.

Combined with a mathematical exploration of the model hazard function (i.e., the model-derived function that predicts the age-specific cancer incidence) our numerical findings support the hypothesis that the initiation of a benign (non-invasive) tumor, its malignant transformation, and persistence constitute major bottlenecks in the progression of a premalignant tumor to cancer. This is consistent with results from evolutionary models which find neoplastic progression to be driven mainly by mutations that confer only slight improvements in fitness [60], while the transition from a non-invasive to an invasive tumor, which expands with a significantly higher growth rate, constitutes a critical, rate-limiting event.

2.3.1 The MSCE model

For ease of notation, we shall hereafter refer to this particular 4-stage generalized MVK model as the multistage clonal expansion (MSCE) model. The MSCE model is a continuous time Markov chain that tracks random stem cell counts of five different types at any continuous time (age) t: normal $N(t)$, pre-initiated $P^*(t)$, initiated (premalignant) $P(t)$, preclinical malignant $M(t)$, and clinical cancer $C(t)$. See Fig. 2.2a for the conceptual framework of the full MSCE model, which can be viewed as a multi-type branching process, along with an illustrated realization of this stochastic process.

Tumor initiation

A hallmark of all previous MSCE models is that tumor initiation requires a number of rate-limiting mutational events before a stem cell can undergo a clonal expansion that results in a premalignant lesion (see Fig. 2.2). For colon and pancreatic cancer, Meza et al. inferred that it takes two rare hits to transform a normal tissue stem cell into an initiated tumor cell that is no longer under homeostatic control and undergoes a (first) clonal expansion [51]. The two significant initial hits may represent biallelic inactivation of tumor suppressor genes, such as *TP53* or *p16* that occur frequently in many cancers, or the *APC* gene in CRC [61]. Inactivation of *TP53* is seen during early development of many digestive tract cancers, including gastric cancer [62], and pancreatic cancer [63, 64].

A mathematical consequence of the two-hit hypothesis for (premalignant) tumor initiation is that the hazard function of the model has a linearly increasing trend for older ages [51]. The presence of such a linear phase in the incidence curves for colorectal and pancreatic cancer could indeed be demonstrated by likelihood-based comparisons of models with two (or more hits) for initiation. Models with single-hit tumor initiations such as the TSCE model do not give rise to a linear phase in the hazard function [41, 51].

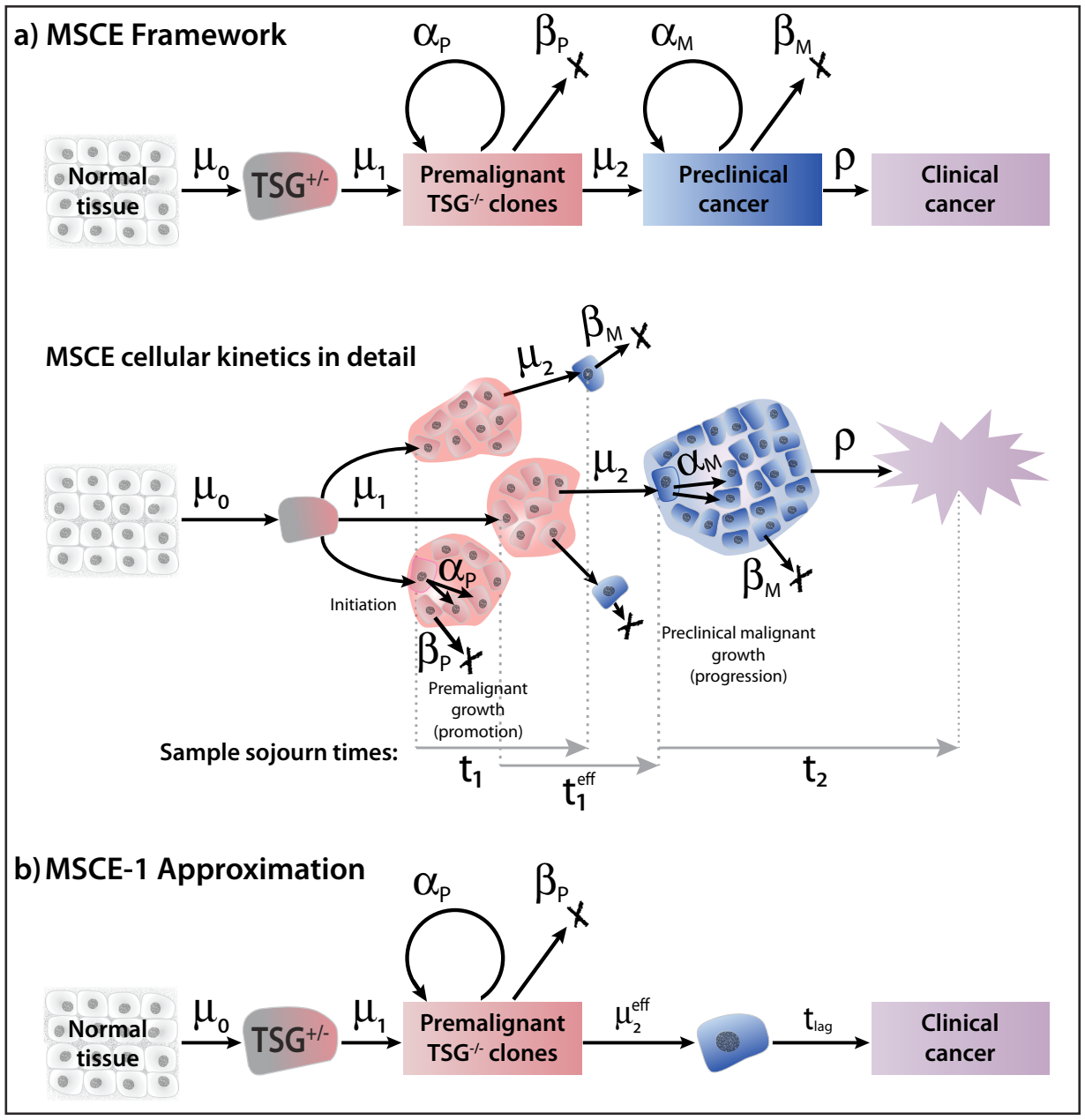

Figure 2.2: **The multistage clonal expansion (MSCE) model.** (Continued on the following page.)

Figure 2.2: **The multistage clonal expansion (MSCE) model.** a) The MSCE model for cancer with two stochastic birth-death-mutation processes for premalignant and malignant cells. The model assumes a 'two-hit' tumor initiation process with Poisson initiation rates μ_0, μ_1 which leads to the stochastic appearance of premalignant progenitor cells in the tissue. In the lower sample MSCE realization, premalignant cells undergo a first clonal expansion described by a birth-death-migration process with cell division rate $\alpha_{_P}$, cell death-or-differentiation rate $\beta_{_P}$, and malignant transformation rate μ_2. Malignant cells, in turn, undergo a second clonal expansion with cell division and death rates $\alpha_{_M}$ and $\beta_{_M}$, respectively. Clinical detection occurs through a size-based detection process with parameter ρ. The sample sojourn time t_1 represents the time from the initiation of a premalignant clone until first malignant transformation. The sample sojourn time t_1^{eff} represents the time from the initiation of a premalignant clone to the first malignant cell in that clone which results in a persistent tumor which escapes extinction. Lastly, the sample sojourn time t_2 represents the time for a persistent tumor to develop from a single malignant cell to detected, clinical cancer. The MSCE model is well approximated by (b) the MSCE-1 Approximation which includes an effective malignant transformation rate μ_2^{eff} and a constant lag-time for tumor progression.

Tumor promotion

Prior to the transition into the initiation-associated linear phase, the MSCE hazard function with constant parameters increases exponentially with a rate that is approximately given by the net cell proliferation rate of premalignant P cells [51]. The transition from the exponential phase to the linear phase occurs around the age of 60 for colorectal and pancreatic cancer. Clonal expansion of P cells is represented by a stochastic birth-death-mutation (b-d-m) process with cell division rate $\alpha_{_P}$, death-or-differentiation rate $\beta_{_P}$, and mutation rate μ_2. The net cell proliferation rate of P cells is given by $\alpha_{_P} - \beta_{_P} - \mu_2$ and the asymptotic probability of extinction of P cells by the ratio $\beta_{_P}/\alpha_{_P}$ [65], which is the probability that a

premalignant cell, together with its progeny, will ultimately become extinct. Premalignant P cells may suffer further mutations with rate μ_2 which transform them into malignant (M) cancer stem cells. Although the premalignant cell population is likely to undergo a complex evolutionary process involving multiple mutations in critical regulatory pathways before acquiring a malignant phenotype [58], only two initial rate-limiting mutations prior to clonal expansion appear necessary to adequately describe the main shape of the incidence curve for the three digestive tract cancers studied here.

Tumor progression and cancer detection

Development of a preclinical tumor in the MSCE model begins with a single (malignant) cell that undergoes clonal expansion and eventually, if the clone survives extinction, progresses to clinically detectable cancer. In contrast to natural history models in which the preclinical state of a tumor is typically assumed to be *screen-detectable*, the preclinical tumor development in the MSCE model starts off with a single malignant cell that undergoes a clonal expansion and eventually, if the clone survives extinction, is detected as cancer. Mathematically, the growth of the malignant tumor is described by a stochastic birth-death process with cell division rate $\alpha_{_M}$ and cell death rate $\beta_{_M}$. For technical reasons, we will refer to all malignant clones, prior to their detection, as "preclinical" clones, even if they are too small to be detected by screening. Screen-detection of premalignancies and/or small malignancies at a certain age will be discussed in Chapter 4 for the case of EAC. Clinical detection of the tumor is treated as a stochastic event with rate ρ per cell per year. This implies that a tumor of size n cells has probability $n\rho\Delta t$ to be detected in a time interval Δt short enough for the tumor to be constant in size. This clinical observation process is a generalization of a birth-death-mutation (b-d-m) process. Note, all analyses reported for this particular exploration are obtained using a fixed value of $\rho = 10^{-7}$. The rationale for this particular value of ρ is that a typical tumor contains about 10^9 cells upon (symptomatic) detection and that only about 1% of the tumor volume is occupied by actively dividing tumor cells [66]. Parameters, such as premalignant and malignant cell proliferation, were estimated with other

values assumed for ρ (in the range of $10^{-6} - 10^{-8}$) and our conclusions did not change (see Appendix A).

2.3.2 From cells to populations: MSCE hazard function

As mentioned previously for cancer multistage model predecessors, the MSCE cell-level description is linked to the population scale by means of the model hazard function, defined as the instantaneous rate of detecting cancers among individuals who have not been previously diagnosed with cancer. In this section we will show that this quantity may be derived from the backward Kolmogorov equations for the stochastic multistage process and solved numerically via a system of coupled ordinary differential equations (ODEs) [13]. For the mathematical development of the MSCE cell model depicted in Fig. 2.2, we first formally introduce the notation for the following random variables of the multi-type branching process

$$
\begin{aligned}
N(t) &= \text{number of renewing normal stem cells in a tissue at time } t \\
P^*(t) &= \text{number of pre-initiated cells at time } t \\
P(t) &= \text{number of premalignant (initiated) cells at time } t \\
M(t) &= \text{number of malignant (preclinical) cells (prior to detection) at time } t \\
C(t) &= \text{number of cancer cells (after detection) at time } t
\end{aligned}
$$

For clinical detection, we will employ the following indicator function, $D(t)$,

$$
D(t) = \begin{cases} 0 \text{ if no cancer detected clinically by time } t \\ 1 \quad \text{otherwise} \quad \text{i.e., } C(\tau) > 0 \text{ for some } \tau \leq t \end{cases}
$$

Beginning with a single pre-initiated (P^*), premalignant (P), or malignant (M) cell at time τ, the probability distributions for number of specific cell type at time $t \geq \tau$, can be derived from the coefficients of each cell type's probability generating function (PGF) Φ_{P^*},

Φ_{P}, or Φ_{M}, respectively. These generating functions are defined as

$$\begin{aligned}\Phi_{M}(y_3,z;\tau,t) &= E[y_3^{M(t)}z^{D(t)}|M(\tau)=1,D(\tau)=0] \qquad (2.3)\\ &= \sum_{k,l} y_3^k z^l \Pr[M(t)=k,D(t)=l|M(\tau)=1,D(\tau)=0]\\ \Phi_{P}(y_2,y_3,z;\tau,t) &= E[y_2^{P(t)}y_3^{M(t)}z^{D(t)}|P(\tau)=1,M(\tau)=0,D(\tau)=0] \qquad (2.4)\\ &= \sum_{j,k,l} y_2^j y_3^k z^l \Pr[P(t)=j,M(t)=k,D(t)=l|P(\tau)=1,M(\tau)=0,D(\tau)=0]\\ \Phi_{P^*}(y_1,y_2,y_3,z;\tau,t) &= E[y_1^{P^*(t)}y_2^{P(t)}y_3^{M(t)}z^{D(t)}|P^*(\tau)=1,P(\tau)=0,M(\tau)=0,D(\tau)=0] \qquad (2.5)\\ &= \sum_{i,j,k,l} y_1^i y_2^j y_3^k z^l \Pr[P^*(t)=i,P(t)=j,M(t)=k,D(t)=l|P^*(\tau)=1,P(\tau)=0,M(\tau)=0,D(\tau)=0]\end{aligned}$$

And the PGF for the entire process Ψ starting from normal cells is the following

$$\begin{aligned}\Psi(y_1,y_2,y_3,z;\tau,t) &= E[y_1^{P^*(t)}y_2^{P(t)}y_3^{M(t)}z^{D(t)}|P^*(\tau)=0,P(\tau)=0,M(\tau)=0,D(\tau)=0] \qquad (2.6)\\ &= \sum_{i,j,k,l} y_1^i y_2^j y_3^k z^l P(i,j,k,l;\tau,t) \qquad (2.7)\end{aligned}$$

where $l=\{0,1\}$ and

$$P(i,j,k,l;\tau,t) = \Pr[P^*(t)=i,P(t)=j,M(t)=k,D(t)=l|P^*(\tau)=0,P(\tau)=0,M(\tau)=0,D(\tau)=0].$$

Backward Kolmogorov equations

The generating functions satisfy the following Kolmogorov backward equations (here the cellular kinetic rates may be age-dependent)

$$\begin{aligned}\frac{\partial\Phi_{M}(y_3,z;\tau,t)}{\partial\tau} &= -\alpha_{M}\Phi_{M}^2(y_3,z;\tau,t)-\beta_{M} \qquad (2.8)\\ &\quad -z\rho\Phi_{M}(y_3,z;\tau,t)+[\alpha_{M}+\beta_{M}+\rho]\Phi_{M}(y_3,z;\tau,t)\\ \frac{\partial\Phi_{P}(y_2,y_3,z;\tau,t)}{\partial\tau} &= -\alpha_{P}\Phi_{P}^2(y_2,y_3,z;\tau,t)-\beta_{P} \qquad (2.9)\\ &\quad +[\alpha_{P}+\beta_{P}+\mu_2]\Phi_{P}(y_2,y_3,z;\tau,t)-\mu_2\Phi_{P}(y_2,y_3,z;\tau,t)\Phi_{M}(y_3,z;\tau,t)\end{aligned}$$

$$\frac{\partial\Phi_{P^*}(y_1,y_2,y_3,z;\tau,t)}{\partial\tau} = -\mu_1\Phi_{P^*}(y_1,y_2,y_3,z;\tau,t)[\Phi_{P}(y_2,y_3,z;\tau,t)-1] \qquad (2.10)$$

$$\frac{\partial\Psi(y_1,y_2,y_3,z;\tau,t)}{\partial\tau} = -\mu_0 N(t)\Psi(y_1,y_2,y_3,z;\tau,t)[\Phi_{P^*}(y_1,y_2,y_3,z;\tau,t)-1] \qquad (2.11)$$

Numerical solutions

To obtain the hazard function, we first solve for the overall survival function (from cancer detection), starting at time 0, which in our notation is

$$S_{MSCE}(t) = 1 - P_{MSCE}(t) = \Pr[D(t) = 0 | P^*(0) = 0, P(0) = 0, M(0) = 0, D(0) = 0] \quad (2.12)$$

$$= \Psi(1, 1, 1, 0; 0, t), \quad (2.13)$$

where $P_{MSCE}(t)$ is the probability of a cancer detection at time t,

$$P_{MSCE}(t) = \Pr[D(t) = 1 | P^*(0) = 0, P(0) = 0, M(0) = 0, D(0) = 0].$$

We will here denote $\Phi_M(1, 0; \tau, t) \equiv \Phi_M(\tau, t)$, $\Phi_P(1, 1, 0; \tau, t) \equiv \Phi_P(\tau, t)$, $\Phi_{P^*}(1, 1, 1, 0; \tau, t) \equiv \Phi_{P^*}(\tau, t)$, and $\Psi(1, 1, 1, 0; \tau, t) \equiv \Psi(\tau, t)$. A dot designates a first derivative with respect to t. The hazard function, i.e., the rate at which cancer is detected in individuals who have not been diagnosed before, is given by

$$h_{MSCE}(t) = -\frac{\dot{S}_{MSCE}(t)}{S_{MSCE}(t)} = -\frac{\dot{\Psi}(0, t)}{\Psi(0, t)} = -\frac{d}{dt}\ln[\Psi(0, t)] \quad (2.14)$$

For fixed t, the boundary value system of coupled PDEs provided in Eqs. (2.8-2.11) can be converted into an initial value problem (IVP) with the change of variables $u = t - \tau$, where u is the "running" time. This redefinition and equations hereafter follow the method used by Crump et al. [67]. Define the following variables for the new IVP: $Y_1(u, t) = \Phi_M(\tau, t), Y_2(u, t) = \dot{\Phi}_M(\tau, t), Y_3(u, t) = \Phi_P(\tau, t), Y_4(u, t) = \dot{\Phi}_P(\tau, t), Y_5(u, t) = \Phi_{P^*}(\tau, t)$, $Y_6(u, t) = \dot{\Phi}_{P^*}(\tau, t), Y_7(u, t) = \Psi(\tau, t), Y_8(u, t) = -\dot{\Psi}(\tau, t)/\Psi(\tau, t)$ with corresponding initial conditions (ICs) $Y_1(0, t) = Y_3(0, t) = Y_5(0, t) = Y_7(0, t) = 1$,

$Y_4(0,t) = Y_6(0,t) = Y_8(0,t) = 0$, and $Y_2(0,t) = -\rho$,

$$\frac{dY_1(u,t)}{du} = \beta_{_M} - (\alpha_{_M} + \beta_{_M} + \rho)Y_1(u,t) + \alpha_{_M}Y_1^2(u,t) \tag{2.15}$$

$$\frac{dY_2(u,t)}{du} = 2\alpha_{_M}Y_1(u,t)Y_2(u,t) - (\alpha_{_M} + \beta_{_M} + \rho)Y_2(u,t) \tag{2.16}$$

$$\frac{dY_3(u,t)}{du} = \beta_{_P} + \mu_2 Y_1(u,t)Y_3(u,t) - (\alpha_{_P} + \beta_{_P} + \mu_2)Y_3(u,t) + \alpha_{_P}Y_3^2(u,t) \tag{2.17}$$

$$\begin{aligned}\frac{dY_4(u,t)}{du} &= 2\alpha_{_P}Y_3(u,t)Y_4(u,t) + \mu_2(Y_4(u,t)Y_1(u,t) + Y_3(u,t)Y_2(u,t)) \\ &\quad - (\alpha_{_P} + \beta_{_P} + \mu_2)Y_4(u,t)\end{aligned} \tag{2.18}$$

$$\frac{dY_5(u,t)}{du} = \mu_1 Y_5(u,t)(Y_3(u,t) - 1) \tag{2.19}$$

$$\frac{dY_6(u,t)}{du} = \mu_1(Y_6(u,t)Y_3(u,t) - Y_6(u,t) + Y_5(u,t)Y_4(u,t)) \tag{2.20}$$

$$\frac{dY_7(u,t)}{du} = \mu_0 N(t)Y_7(u,t)(Y_5(u,t) - 1) \tag{2.21}$$

$$\frac{dY_8(u,t)}{du} = \mu_0 N(t)(Y_8(u,t)Y_5(u,t) - Y_8(u,t) + Y_7(u,t)Y_6(u,t)) \tag{2.22}$$

These 8 coupled ODEs can be solved numerically to obtain the MSCE hazard and survival functions

$$h_{_{MSCE}}(t) = -Y_8(t,t)/Y_7(t,t) \tag{2.23}$$

$$S_{_{MSCE}}(t) = Y_7(t,t) \tag{2.24}$$

As we will show here (see Results), the MSCE model contains an approximation (referred to as MSCE-1) which differs from the original MSCE model in two important aspects: (1) the rate at which P cells suffer a transformation event that gives rise to a detectable cancer is approximated by an 'effective' transformation rate, μ_2^{eff}, and (2) the approximation requires a lag-time to allow for the time from first malignant cell that forms a persistent cancer clone to the time of diagnosis. Furthermore, not all of the MSCE model parameters are identifiable from incidence data - some parameters must be fixed initially in order to achieve parameter identifiability, as discussed in the Results and [13].

While multistage generalizations of the models shown in Fig. 2.2 have also been explored by others [49, 54–57], the general impact of malignant tumor progression on the hazard

function (and the age-specific incidences of the cancers modeled here) has not been fully characterized [41, 50–52], especially in regard to the time scales of premalignant and malignant clonal expansion.

Data

We used data from the Surveillance Epidemiology and End Results (SEER) database by the National Cancer Institute. Incidence data were for all races by single-years of age between 10 and 84 and calendar-year between 1975 and 2008 for males and females in the original nine registries (SEER-9) [7]. Incident cancers were defined using the International Classification of Diseases for Oncology, Third Edition (ICD-O-3) as follows: colorectal cancer (C18-C20); gastric cancer (C16); and pancreatic cancer (C25). Population data were also downloaded for the 9 SEER catchment areas by gender and single-years of age and calendar year. A standard method for adjusting the incidence for secular trends (period and cohort effects) are described in Appendix A.

2.3.3 Results

We find that modeling of cancer incidence data provides new insights into the importance of clonal extinction and clonal growth rates (or doubling times) of premalignant and malignant clones in relation to three underlying time scales in carcinogenesis: the mean sojourn time for premalignant clones until occurrence of the first malignant cell (T_1), the mean sojourn time for premalignant clones until the first surviving malignant clone (T_1^{eff}), and the mean sojourn time of persistent preclinical cancers from first malignant cell to time of diagnosis (T_2). In the following we demonstrate how these time scales contribute to, and are estimable from, the age-specific incidence curves of three digestive tract cancers.

Mathematical properties of MSCE-1 approximation

Impact of malignant clonal expansion: To gain insights into how tumor progression is impacting cancer incidence we begin with a mathematical dissection of the hazard function generated by the MSCE model depicted in Fig. 2.2. For constant rate parameters, this demonstrates that the MSCE model (with a distinct clonal expansion for malignant cells) can be closely approximated by a reduced model (MSCE-1) that adjusts the rate of malignant transformation, μ_2, for non-extinction and further models the outgrowth of persistent malignant clones as a constant time-lag (t_{lag}), which in turn is approximated by the mean sojourn time of a surviving malignant clone, from its inception to detection of cancer, T_2. To better understand the relationship between t_{lag} and the mean sojourn time T_2 in the MSCE model we first show that the hazard function of a model with two consecutive clonal expansions for premalignant and malignant cells is mathematically equivalent to a model with a single clonal expansion of premalignant cells with a time-dependent mutation rate, i.e., replacing $\mu_2 \leftrightarrow \mu_2(1 - S_{_M}(u))$, where $1 - S_{_M}(u)$ is the unnormalized probability of detection of a malignant clone at a time u after the malignant clone is seeded.

We first rewrite Eq. (2.9) from the backward Kolmogorov equations for the 4-stage MSCE model as follows

$$\frac{\partial \Phi_{_P}(\tau,t)}{\partial \tau} = -\alpha_{_P}\Phi^2_{_P}(\tau,t) - \beta_{_P} + [\alpha_{_P} + \beta_{_P} + \mu_2]\Phi_{_P}(\tau,t) - \mu_2\Phi_{_P}(\tau,t)\Phi_{_M}(\tau,t)$$

$$\Rightarrow \frac{\partial \Phi_{_P}(\tau,t)}{\partial \tau} = -\alpha_{_P}\Phi^2_{_P}(\tau,t) - \beta_{_P} + [\alpha_{_P} + \beta_{_P} + \mu_2(1 - \Phi_{_M}(\tau,t))]\Phi_{_P}(\tau,t) \tag{2.25}$$

This has the same form as the backward equation for the 3-stage ($k = 3$) MSCE model that has one clonal expansion stage [41, 50–52] with the mutation rate μ_2 assuming a time-dependent form, i.e,

$$\mu_2 \to \mu_2(1 - \Phi_{_M}(\tau,t)) = \mu_2(1 - S_{_M}(u)) \tag{2.26}$$

where $u = t - \tau$ and $S_{_M}(u)$ is the "survival function" for cancer detection of a preclinical cancer clone a time u since the clone was born (ie. τ is the time that the progenitor cell of

the malignant clone was born). This is defined as (full derivation using Kolmogorov forward equations can be found in [41])

$$S_{_M}(u) = 1 + \frac{1}{\alpha_{_M}} \frac{p_{_M} q_{_M} e^{-p_{_M} u} - q_{_M} p_{_M} e^{-q_{_M} u}}{q_{_M} e^{-p_{_M} u} - p_{_M} e^{-q_{_M} u}}, \tag{2.27}$$

$$p_{_M} = \frac{1}{2}(-(\alpha_{_M} - \beta_{_M} - \rho) - \sqrt{(\alpha_{_M} - \beta_{_M} - \rho)^2 + 4\alpha_{_M}\rho}) \tag{2.28}$$

$$q_{_M} = \frac{1}{2}(-(\alpha_{_M} - \beta_{_M} - \rho) + \sqrt{(\alpha_{_M} - \beta_{_M} - \rho)^2 + 4\alpha_{_M}\rho}). \tag{2.29}$$

Let p_∞ be the probability that a single preclinical clone initiated at time τ eventually becomes detected:

$$\lim_{u\to\infty}(1 - S_{_M}(u)) = 1 - \frac{\alpha_{_M} + p_{_M}}{\alpha_{_M}} \approx 1 - \frac{\beta_{_M}}{\alpha_{_M}} \equiv p_\infty,$$

i.e., as $u \to \infty$, $S_{_M}(u)$ approaches approximately $\beta_{_M}/\alpha_{_M}$, which is the probability of extinction of the preclinical clone, when $\rho \ll 1$. Hence, the u dependent mutation rate $\mu_2(1-S_{_M}(u))$ is bounded between 0 and approximately $\mu_2(1 - \beta_{_M}/\alpha_{_M}) = \mu_2 p_\infty \equiv \mu_2^{eff}$, which takes into account the non-extinction of small preclinical clones before they are detected.

Thus, for fixed t, the PDE from Eq. (2.25) that we wish to integrate over $[0, t]$ can be written as

$$\frac{\partial \Phi_{_P}(\tau, t)}{\partial \tau} = -\alpha_{_P} \Phi_{_P}^2(\tau, t) - \beta_{_P} + \left[\alpha_{_P} + \beta_{_P} + \mu_2^{eff}\left(\frac{1 - S_{_M}(u)}{p_\infty}\right)\right] \Phi_{_P}(\tau, t) \tag{2.30}$$

with the initial condition $\Phi_{_P}(t, t) = 1$.

The main effect of a preclinical tumor progression (the second clonal expansion in the MSCE model) on cancer incidence is a delay (or lag) of time for a cancer clone to grow from a single malignant cell into a detectable tumor, conditional that it does not become extinct. Analytically, this lag time is caused by a steep jump that occurs in the function $\frac{1-S_{_M}(u)}{p_\infty}$ (the time-dependent coefficient in Eq. (2.30)). When $(\alpha_{_M} - \beta_{_M}) \gg (\alpha_{_P} - \beta_{_P})$, the function $\frac{1-S_{_M}(u)}{p_\infty}$ has a steep ascent from near lower bound of 0 to upper bound of 1, with an inflection point u^* at $\frac{1-S_{_M}(u^*)}{p_\infty} = \frac{1}{2} + O(\rho)$. With a series expansion around $\rho = 0$, we may easily compute u^* and a useful approximation.

$$\left.\frac{d^2 S_{_M}(u)}{du^2}\right|_{u=u^*} = 0 \;\Rightarrow\; u^* = -\frac{\ln(-p_{_M}/q_{_M})}{p_{_M} - q_{_M}} = \frac{\ln(-q_{_M}/p_{_M})}{p_{_M}} + \mathcal{O}(\rho). \tag{2.31}$$

This time u^* is the inflection point for the nearly discontinuous function that corresponds mathematically as the appropriate lag time t_{lag}. We now prove that this time is approximately the mean sojourn time of a surviving preclinical tumor, T_2.

Theorem 1. $T_2 = t_{lag} + \mathcal{O}(\rho)$

Proof. Let T be the random sojourn time from transformation of a single malignant cell to tumor detection. The cumulative distribution function for T is given by

$$\Pr[T \leq u] = \frac{1 - S_{_M}(u)}{p_\infty}$$

Since T only takes non-negative values, we can use the following fact

$$E[T] = \int_0^\infty \Pr[T \geq u]\, \mathrm{d}u = \int_0^\infty 1 - \frac{1 - S_{_M}(u)}{p_\infty}\, \mathrm{d}u.$$

Then this mean sojourn time of a surviving tumor is defined as

$$T_2 \equiv E[T] = \int_0^\infty \left(1 - \frac{1 - S_{_M}(u)}{p_\infty}\right) \mathrm{d}u = -\frac{\ln\left(\frac{q_{_M}/(-p_{_M})}{1+q_{_M}/(-p_{_M})}\right)}{(-p_{_M})} = \frac{\ln(-q_{_M}/p_{_M})}{p_{_M}} + \mathcal{O}(\rho). \tag{2.32}$$

[comparing to Eq. (2.31)] $\Rightarrow T_2 = t_{lag} + \mathcal{O}(\rho)$

□

Thus, the time point of inflection u^* of the function $\frac{1-S_{_M}(u)}{p_\infty}$ equals approximately T_2 when ρ is very small (see Chapter 3 for some example consequences of this approximation when applied to EAC). When the difference in magnitudes between the malignant and premalignant proliferation rates becomes more severe, the function $\frac{1-S_{_M}(u)}{p_\infty}$ limits to a discontinuous step function at u^*. We will continue by approximating $\left(\frac{1-S_{_M}(u)}{p_\infty}\right)$ in Eq. (2.30) by a piecewise constant function on $[0, t]$ such that

$$\frac{1 - S_{_M}(u)}{p_\infty} \approx \begin{cases} 0 & \text{if } u \in [0, t_{lag}) \\ 1 & \text{if } u \in [t_{lag}, t] \end{cases} \tag{2.33}$$

where $t_{lag} < t$ will be set equal to T_2. First, we rewrite our ODE for fixed t as before in terms of $u = t - \tau$ and solve on the first interval $[0, t_{lag}]$ with initial condition $\Phi_{_P}(0,t) = 1$:

$$\begin{aligned}
\frac{d\Phi_{_P}(u,t)}{du} &= -\alpha_{_P}\Phi_{_P}^2(u,t) - \beta_{_P} + \left[\alpha_{_P} + \beta_{_P} + \mu_2^{eff} \cdot 0\right]\Phi_{_P}(u,t) \\
\Rightarrow \frac{d\Phi_{_P}(u,t)}{du} &= -\alpha_{_P}\Phi_{_P}^2(u,t) - \beta_{_P} + \left[\alpha_{_P} + \beta_{_P}\right]\Phi_{_P}(u,t) \\
\Rightarrow \Phi_{_P}(u,t) &= 1 \quad \text{for all } u \in [0, t_{lag})
\end{aligned}$$

Next, on $[t_{lag}, t]$, we are solving for a shifted 2-stage survival function with initial condition $\Phi_{_P}(t_{lag}, t) = \Phi_{_P}(0, t - t_{lag}) = 1$. (See solution details in [41])

$$\begin{aligned}
\frac{d\Phi_{_P}(u,t)}{du} &= -\alpha_{_P}\Phi_{_P}^2(u,t) - \beta_{_P} + \left[\alpha_{_P} + \beta_{_P} + \mu_2^{eff}\right]\Phi_{_P}(u,t) \\
\Rightarrow \Phi_{_P}(u,t) &= \left(\frac{q_{_P}^{eff} - p_{_P}^{eff}}{q_{_P}^{eff} e^{-p_{_P}^{eff}(u-t_{lag})} - p_{_P}^{eff} e^{-q_{_P}^{eff}(u-t_{lag})}}\right)^{\mu_1/\alpha_{_P}} \quad \text{for } u \in [t_{lag}, t]
\end{aligned}$$

with

$$p_{_P}^{eff} = \frac{1}{2}(-(\alpha_{_P} - \beta_{_P} - \mu_2^{eff}) - \sqrt{(\alpha_{_P} - \beta_{_P} - \mu_2^{eff})^2 + 4\alpha_{_P}\mu_2^{eff}}), \tag{2.34}$$

$$q_{_P}^{eff} = \frac{1}{2}(-(\alpha_{_P} - \beta_{_P} - \mu_2^{eff}) + \sqrt{(\alpha_{_P} - \beta_{_P} - \mu_2^{eff})^2 + 4\alpha_{_P}\mu_2^{eff}}). \tag{2.35}$$

Using the 3-stage MSCE model hazard form (See [50] for derivation), the mathematical approximation for the MSCE model takes the following form

$$h_{_{MSCE}}(t) \approx h_{_{MSCE-1}}(t) = \mu_0 N \left(1 - \left(\frac{q_{_P}^{eff} - p_{_P}^{eff}}{q_{_P}^{eff} e^{-p_{_P}^{eff}(t-t_{lag})} - p_{_P}^{eff} e^{-q_{_P}^{eff}(t-t_{lag})}}\right)^{\mu_1/\alpha_{_P}}\right), \tag{2.36}$$

where $\mu_2^{eff} \equiv \mu_2(1 - \beta_{_M}/\alpha_{_M})$, $t_{lag} \equiv T_2$, and N is assumed to be a large number of normal tissue stem cells under homeostatic control and thus constant in size.

Therefore, we can compute an analytical approximation to the MSCE model hazard function for incidence by using the form of the hazard of the MSCE-1 model, $h_{_{MSCE-1}}(t)$, and replacing μ_2 with $\mu_2^{eff} \equiv \mu_2(1 - \beta_{_M}/\alpha_{_M})$ and time t with $t - t_{lag} \equiv t - T_2$. This equivalency

can be seen from the derivation above containing the solution of $\Phi_{P}(u,t)$ on $[t_{lag},t]$ with guaranteed survival from $[0,t_{lag}]$. This approximation implicitly assumes that, conditional on non-extinction, the first surviving malignancy in the tissue is eventually the first to be observed. Alongside this altered, effective mutation rate, the time shift by the mean sojourn time of the first surviving tumor, T_2 will push the MSCE-1 type hazard function to the right to coincide with the actual MSCE hazard curve. Fig. 2.3 depicts comparisons of the MSCE-1 analytical approximation and the exact, numerical MSCE hazard solution from Eq. (2.23) in the case of CRC and gastric cancer, fitted separately.

The time-dependence of the conditional mutation rate $\mu_2(1-S_{M}(u))$ has two main effects: 1) it reduces the effective rate of malignant transformation, and 2) it creates a time delay for a malignant clone to grow, conditional on its non-extinction, into a detectable tumor. The latter effect is mainly due to a sharp transition of the conditional mutation rate from zero to its asymptotic value after a time that equals approximately T_2, the mean sojourn time of the malignant clone to detection of cancer (see Theorem 1). Because asymptotically $S_{M}(u) \to \beta_{M}/\alpha_{M}$ as $u \to \infty$, we define $\mu_2^{eff} \equiv \mu_2(1-\beta_{M}/\alpha_{M})$ as the effective malignant transformation rate for the reduced (MSCE-1) model. Therefore, the approximation amounts to

$$\begin{array}{c} \mu_2 \leftarrow \mu_2^{eff} \\ t_{lag} \leftarrow T_2. \\ {\scriptstyle (MSCE-1)} \quad {\scriptstyle (MSCE)} \end{array} \tag{2.37}$$

Time scales of premalignant clonal expansion: Analogous to the derivation for this mean sojourn time of a malignant clone conditioned on survival, we can also compute sojourn times for premalignant clones, as presented in Table 2.1. The first, T_1, refers to the mean time from the initiation of a premalignant cell to the appearance of the first malignant cell from this premalignant cell's progeny. The survival function for this process $S_{P}(u)$ is defined

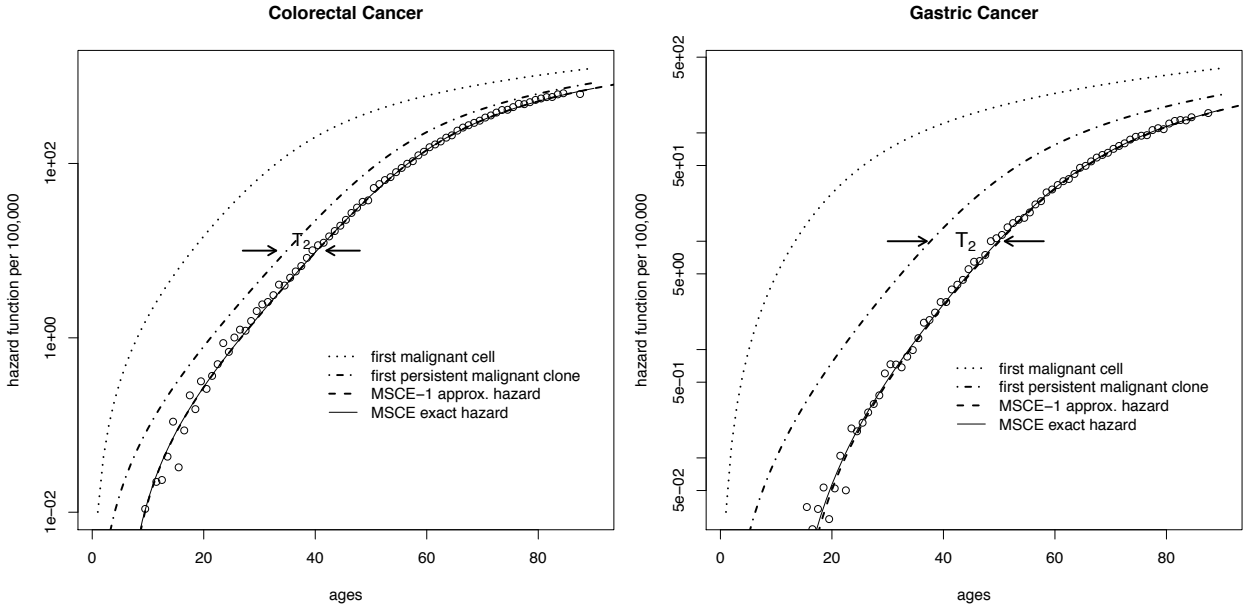

Figure 2.3: **Deconstruction of the MSCE hazard function.** Malignant clone extinction and tumor growth influence the incidence curves (on a logarithmic scale) for a) CRC, and b) gastric cancer. (Plots for pancreatic cancer are similar, but not shown). The SEER data (adjusted for calendar-year and birth-cohort trends) are shown as circles and the overall fit using the MSCE model by the thin solid line. The dotted line on the left shows the underlying hazard for the first malignant cell, regardless of its fate. In contrast, the dash-dotted line shows the hazard for the first ancestor of a persistent (i.e., surviving) malignant clone and the dashed line represents the hazard for the first persistent malignant clone shifted to the right by the mean sojourn time of the malignant clone to cancer detection, T_2.

analogously to Eq. (2.27) with premalignant parameters and with malignant transformation rate μ_2. Next, T_1^{eff} refers to the mean time from the initiation of a premalignant cell until, after it clonally expands, the first transformation of a malignant, preclinical cell conditional that this malignant cell will produce a clone that will survive. The survival function for this process $S_P^{eff}(u)$ is equal to $S_{_P}(u)$ except with malignant transformation rate $\mu_2^{eff} = \mu_2 p_\infty$ replacing μ_2, which as discussed above, guarantees the non-extinction of small preclinical clones before they are detected. Repeating the method used to compute T_2 in Eq. (2.32), we compute these two mean sojourn times as follows

$$T_1 = \int_0^\infty \left(1 - \frac{1 - S_{_P}(u)}{p_\infty}\right) \mathrm{d}u = \frac{\ln(-q_{_P}/p_{_P})}{p_{_P}} + \mathcal{O}(\mu_2) \tag{2.38}$$

$$T_1^{eff} = \int_0^\infty \left(1 - \frac{1 - S_{_P}^{eff}(u)}{p_\infty}\right) \mathrm{d}u = \frac{\ln(-q_{_P}^{eff}/p_{_P}^{eff})}{p_{_P}^{eff}} + \mathcal{O}(\mu_2^{eff}) \tag{2.39}$$

where S_P and S_P^{eff} are survival functions defined analogously to $S_{_M}$, given by

$$S_{_P}^{eff}(u) = 1 + \frac{1}{\alpha_{_P}} \frac{p_{_P}^{eff} q_{_P}^{eff} e^{-p_{_P}^{eff} u} - q_{_P}^{eff} p_{_P}^{eff} e^{-q_{_P}^{eff} u}}{q_{_P}^{eff} e^{-p_{_P}^{eff} u} - p_{_P}^{eff} e^{-q_{_P}^{eff} u}} \tag{2.40}$$

$$S_{_P}(u) = 1 + \frac{1}{\alpha_{_P}} \frac{p_{_P} q_{_P} e^{-p_{_P} u} - q_{_P} p_{_P} e^{-q_{_P} u}}{q_{_P} e^{-p_{_P} u} - p_{_P} e^{-q_{_P} u}} \tag{2.41}$$

with

$$p_{_P} = \frac{1}{2}(-(\alpha_{_P} - \beta_{_P} - \mu_2) - \sqrt{(\alpha_{_P} - \beta_{_P} - \mu_2)^2 + 4\alpha_{_P}\mu_2}), \tag{2.42}$$

$$q_{_P} = \frac{1}{2}(-(\alpha_{_P} - \beta_{_P} - \mu_2) + \sqrt{(\alpha_{_P} - \beta_{_P} - \mu_2)^2 + 4\alpha_{_P}\mu_2}). \tag{2.43}$$

and the analogous equations for $p_{_P}^{eff}, q_{_P}^{eff}$ with μ_2^{eff} replacing μ_2, as provided perviously.

Parameter identifiability and sensitivity

Not all of the MSCE model parameters are identifiable from incidence data - some parameters must be fixed initially in order to achieve parameter identifiability (see Heidenreich et al. [68]). Furthermore, for estimability, the exponential-then-linear character of the multistage hazard function (see Eq. (2.36)) suggests a parametrization that involves the slope of the linear phase

$\lambda \equiv \mu_0 N \mu_1 p_\infty$ and the growth parameter of the exponential phase $g_{_P} \equiv \alpha_{_P} - \beta_{_P} - \mu_2$ [51]. Note, the rates μ_0 and μ_1 cannot be estimated separately because the slope λ depends on their product. Analogous to premalignant growth, we introduce the malignant growth parameter $g_{_M} \equiv \alpha_{_M} - \beta_{_M} - \rho$. In order to identify μ_2 and ρ we find it necessary to fix the cell division rates $\alpha_{_P}$ and $\alpha_{_M}$. Although the product $\alpha_{_M}\rho$ is mathematically identifiable, we were not able to obtain stable estimates and therefore also fixed the (per cell) cancer detection parameter ρ (see Materials and Methods). Otherwise, the biological model parameters were estimated using a Markov Chain Monte Carlo (MCMC) method (see Supplemental Information of [13]).

To explore the dependence of our parameter estimates on the fixed parameters $\alpha_{_P}$, $\alpha_{_M}$, and the cancer detection rate ρ we have also conducted a systematic sensitivity analysis. The results of this analysis (specifically, the ranges of the obtained maximum likelihood estimates for the parameters λ, g_P, g_M, and μ_2^{eff}, assuming constant birth cohort and calendar year effects) for each fixed parameter are given in Table A.1. This analysis (although limited to CRC) shows that the estimates of $g_{_M}$, and therefore the mean sojourn time T_2, vary only slightly when $\alpha_{_P}$ and $\alpha_{_M}$ are perturbed, but may vary up to 20% as the detection rate ρ changes an order of magnitude. Therefore, the dependence of the preclinical cancer sojourn time on ρ is modest.

Low fitness of malignant cells

We use the above results to gain insight into the importance of clonal extinction. Figure 2.3 shows fits obtained with the MSCE model (solid line) to SEER incidence data for a) colorectal cancer, and b) gastric cancers. These fits include adjustments of the model-generated hazard function for secular trends (for details see Appendix A). It is instructive to mathematically 'dissect' the MSCE hazard function to examine the underlying behavior of the incidence curves for the different malignant ancestors. The combined effects of extinction and time for (malignant) tumor growth on incidence can be seen by substituting the 'full' rate μ_2 into the MSCE-1 approximation and ignoring the lag-time, i.e., $t_{lag} = 0$ (dotted line in Fig. 2.3). The higher predicted incidence sans malignant cell extinction or tumor growth

shows that these processes greatly reduce and delay cancer incidence and change the shape of the incidence curve. In comparison, re-introducing the effects of extinction by replacing μ_2 with μ_2^{eff} (without a lag-time) restores the general shape of the incidence curve (dot-dash line) except for cancers occurring too early. Finally, re-introducing the time-lag associated with malignant tumor growth (T_2) in the MSCE-1 approximation accounts for both processes (dashed line) and provides an excellent approximation to the exact incidence curve generated by the full MSCE model (solid line).

Time scales of tumor progression

The MSCE model explicitly models malignant transformations in premalignant tissues of an organ. These tissues may not be uncommon as they may arise independently from a large number of normal ancestor cells. However, our results suggest that most malignant cells and nascent malignancies undergo extinction. The time difference between the appearance of the first malignant cell in a premalignant clone, regardless of its fate, and the first ancestor cell that leads to a stable malignant clone that is bound to turn into symptomatic cancer (unless a patient dies before this happens or an intervention occurs) may be as long as 30-40 years for gastric cancer (see Table 2.1), as long as 20 years for CRC, or as short as 3 years, or less, in the case for pancreatic cancer. It is not clear whether these differences reflect transformation-specific differences in cell survival, exogenous factors, cell senescence, or differences in the degree of genomic instability. Whatever the origin, with the exception of pancreatic cancer, our findings suggest a generally low viability of cancer cells in spite of their aggressive and invasive behavior.

In contrast, the estimated mean sojourn times T_2 of persistent malignant clones vary from 5 to 7 years for CRC, down to less than 1 year for pancreatic cancer (Table 2.1). The latter is consistent with the observation that most pancreatic carcinoma are diagnosed at an advanced metastatic stage. Note, however, that T_1^{eff}, the estimated mean time to the appearance of the first persistent cancer clone (measured from the time the ancestral premalignant cell is born) is somewhat longer for pancreas than colon (52.3 vs. 50.6 years

males	T_1 (95% CI)	T_1^{eff} (95% CI)	T_2 (95% CI)	T_{lag} (95% CI)
CRC	32.6 (30.4 - 36.8)	50.6 (49.7 - 52.1)	5.2 (3.6 - 6.2)	5.4 (3.1 - 7.0)
GaC	17.9 (13.6 - 22.3)	45.8 (43.8 - 47.9)	12.2 (9.8 - 14.7)	9.5 (7.7 -12.0)
PaC	49.1 (40.4 - 52.6)	52.3 (50.9 - 52.9)	0.7 (0.4 - 2.2)	3.0 (0.2 - 6.8)

females	T_1 (95% CI)	T_1^{eff} (95% CI)	T_2 (95% CI)	T_{lag} (95% CI)
CRC	27.5 (25.1 - 30.5)	48.7 (47.6 - 49.9)	6.5 (5.2 - 7.6)	6.5 (5.1 - 8.2)
GaC	20.1 (16.8 - 23.4)	58.6 (56.8 - 60.5)	11.7 (10.1 - 13.2)	10.6 (9.1 -11.7)
PaC	53.2 (44.7 - 56.7)	56.3 (55.2 - 57.1)	0.6 (0.4 - 1.8)	1.7 (0.1 - 4.9)

Table 2.1: **MCMC-based estimates of various tumor promotion and progression time scales**. See text for definitions. Estimates represent medians and 95% credibility intervals of the marginal posterior distribution for each quantity listed. All units are in years.

for males, 56.3 vs. 48.7 in females). This suggests that premalignant precursor lesions in pancreas, such as pancreatic intra-epithelial neoplasia (PanIN), may be present for many years before a stable malignant transformation occurs.

Tumor growth rates

We find highly stable estimates for the net cell proliferation rate g_P of premalignant cells, based on the posterior distributions of the identifiable MSCE model parameters given the observed cancer incidences in SEER (see Supplemental Information of [13]). The reason for this stability appears to lie in the prominence of the exponential phase of the incidence curve and the resulting linear behavior of the log-incidence (see Figs. A.1-A.6). Surprisingly, with the exception of gastric cancers in females, the estimated net cell proliferation rates for premalignant lesions are similar and stay within a range of 0.14 to 0.18 per year, while estimates for the net cell proliferation rate g_M of the malignant lesions are much more variable

and range from 1 per year in gastric cancer to rates as high as 30 per year for pancreatic cancer (see Table 2.2). These values correspond to tumor volume doubling times of 250 days and 8 days, respectively for this group of cancers. While the former is consistent with clinical observations for early gastric carcinoma, which are generally slow growing [69], the latter appears too fast, but not inconsistent with tumor marker doubling times. For example, using the pancreatic tumor marker CA19-9, Nishida et al. [70] estimated doubling times from measurements in patients with inoperable pancreatic cancer in the range of 6 to 313 days. For CRC, the estimated malignant tumor volume doubling times are about 93 days for males and 119 days for females. They too appear at the lower end of the clinical spectrum, but are consistent with the determination by Bolin et al. [71] who followed 27 carcinomas radiographically in the colon and rectum, measuring a median of 130 days with a range of 53 to 1570 days. In spite of considerable uncertainty and variability of the clinical observations, the general agreement of the MSCE model predictions with sparse measurements of tumor doubling times lends support to the claim that carefully collected incidence data harbor quantitative information about the natural history of a tumor, from initiation to promotion to malignant tumor progression.

2.3.4 Discussion

Early models of carcinogenesis recognized the importance of rate-limiting mutations but provided only crude fits to cancer incidence and mortality [72]. Subsequent incorporation of cell proliferation made it possible to account for effects, such as the initiation/promotion effects seen in chemical carcinogenesis [46, 73] or the inverse dose-rate effect for high-linear energy transfer radiation [74], that were more difficult to explain with models that did not include clonal expansion. More recently, multistage extensions of the original two-stage clonal expansion model by Moolgavkar, Venzon and Knudson [36,37] have emerged as useful instruments to explore cancer incidence curves and isolate important secular trends that segregate with birth cohort and/or calendar year (period) from age effects driven by common

males	λ (95% CI) x 10^{-4}	g_P (95% CI)	g_M (95% CI)	μ_2^{eff} (95% CI) x 10^{-6}
CRC	2.13 (2.10 - 2.15)	0.162 (0.160 - 0.164)	2.71 (2.22 - 4.15)	0.73 (0.55 - 0.86)
GaC	0.51 (0.49 - 0.53)	0.140 (0.135 - 0.145)	1.00 (0.80 - 1.30)	3.21 (2.16 - 4.65)
PaC	0.35 (0.34 - 0.36)	0.181 (0.177 - 0.186)	27.7 (7.44 - 49.1)	0.25 (0.21 - 0.32)

females	λ (95% CI) x 10^{-4}	g_P (95% CI)	g_M (95% CI)	μ_2^{eff} (95% CI) x 10^{-6}
CRC	1.57 (1.55 - 1.59)	0.149 (0.147 - 0.151)	2.12 (1.75 - 2.75)	1.56 (1.26 - 1.87)
GaC	0.40 (0.36 - 0.44)	0.100 (0.096 - 0.105)	1.06 (0.91 - 1.25)	2.83 (2.36 - 3.39)
PaC	0.34 (0.33 - 0.35)	0.161 (0.157 - 0.165)	30.0 (9.10 - 50.0)	0.30 (0.26 - 0.36)

Table 2.2: **MCMC-based estimates of MSCE model parameters**. These identifiable parameters are defined as: $\lambda = \mu_0 \cdot N \cdot \mu_1 \cdot p_\infty$, $g_P = \alpha_P - \beta_P - \mu_2$, $g_M = \alpha_M - \beta_M - \rho$, $\mu_2^{eff} = \mu_2 \cdot p_\infty$. Here, we define $p_\infty = 1 - \beta_M/\alpha_M$. Estimates represent medians and 95% credibility intervals of the marginal posterior distribution for each quantity listed. All rate units are in per cell per years.

underlying biological processes [50, 52, 75]. While secular trends are of great interest to epidemiologists and cancer control researchers in understanding the impact of screening, potential exposures to carcinogens (e.g. tobacco smoking), infections, diet, and life-style factors on cancer incidence, in this study we focus on non-specific effects that have their origin in common cell-level processes that drive the age-effect, in particular the impact of malignant tumor progression on the age-specific incidence curve.

Incidence curves are consistent with two types of clonal expansions, slow and fast: Our MSCE model fits to the incidences of three gastrointestinal cancers (CRC, gastric cancer, and pancreatic cancer) yield parameter estimates suggesting that malignant tumor progression is preceded by a prolonged period of premalignant tumor growth characterized by a low rate of net cell proliferation (Tables 2.1 and 2.2). In contrast, malignant tumor growth is estimated to be many-fold faster than premalignant growth. The model distinguishes features of the incidence curves that relate to slow growth of premalignant lesions and fast growth of malignant lesions, and allows estimation of the time period in which tumors sojourn as slowly growing masses before becoming invasive. The effective sojourn time T_1^{eff}, i.e., the time to appearance of the first persistent malignant clone that started with a single premalignant cell, appears to be much longer than estimated from clinical data. For colon, clinical estimates range from 20-25 years [76]. However, this usually refers to the time starting with a small adenoma which must have been already present for some time. It is not known how long adenomas sojourn before they can be observed. A clue can be found in the average time to cancer among patients with familial adenomatous polyposis (FAP), which can be viewed as a lower estimate for the mean sojourn time of an adenoma, as adenomas are likely to form early in life in FAP patients even though the diagnosis of polyposis may not occur until later. From the age distribution of cancer with polyposis in FAP patients (see [76]), which peaks around the age of 40, we conclude that the mean sojourn of an adenoma which has the potential to progress to cancer is likely longer than 40 years since this time generally represents the time to first diagnosis of the cancer – a first passage time in statistical parlance

– and not an average time across all adenomas with neoplastic potential including some that will not turn cancerous in a person's lifetime. Our estimates of 50-55 years for the mean duration of an adenoma developing into a detectable carcinoma are therefore not inconsistent with what can be inferred from the incidence of CRC in FAP.

Identifiability of a malignant progression parameter: Our mathematical analysis shows the approximate equivalency of the hazard functions generated by the MSCE model and a model with a single clonal expansion (MSCE-1) that is adjusted for clonal extinction and delayed by a lag-time representing the mean sojourn time T_2 of the surviving malignancy (see Fig. 2.2). Thus, in practice, only the time-scale associated with malignant tumor progression can be estimated from cancer incidence data but not the full malignant cell kinetics given by the rates of malignant cell division $\alpha_{_M}$, cell death $\beta_{_M}$, and (per cell per year) detection ρ. However, assuming plausible values for the cell division rates ($\alpha_{_M}$) and a (per cell per year) cancer detection rate ρ (see sensitivity analysis), we do obtain estimates for the net cell proliferation rate $g_{_M}$ in malignant tumors that yield tumor volume doubling times that are consistent with clinical observations from radiographic imaging of carcinoma (see Results).

For pancreatic cancer, the estimated sojourn times T_2 for male and female preclinical malignancies are very short, suggesting that the model only captures the short metastatic phase of the development but cannot identify the sojourn of the primary tumor. It is conceivable that non-invasive precursors, such as the PanINs, interact with stromal components such as myofibroblasts that facilitate invasion and metastatic colonization [77]. The resulting colonies may initially grow slowly, perhaps similar to their parental premalignant precursors, but may acquire an aggressive and expansive phenotype at a later time.

Carcinogenesis may well require more than 2 types of clonal expansions. However, as shown by Meza et al. [51] for CRC and pancreatic cancer, the main features of the age-specific incidence curve can almost entirely be explained by the initiation and growth characteristics of premalignant tumors. Here, we posed the follow-up question: what impact does a second

clonal expansion (say, representing malignant tumor growth) have on incidence curves. Our mathematical analysis shows that the impact amounts to a time-translation of the incidence curve which appears to be identifiable in the SEER incidences studied here. This is consistent with the common view that premalignant tumors and malignant tumors result from rather distinct clonal expansions, with markedly different cell kinetics.

Comparison with DNA sequencing studies: For CRC, Jones et al. [78] determined the time required from the founder cell of an advanced carcinoma to the appearance of the metastatic founder cell through comparative lesion sequencing in a small number of subjects. They concluded that it takes on average 2 years for the metastatic founder cell to arise in a carcinoma and an additional 3 years for the metastatic lesion to expand, thus a total of 5 years to the detection of the (metastatic) cancer after the carcinoma forms. Our model-derived estimates for T_2, the mean sojourn time for preclinical CRC (5-7 years), are therefore in good agreement with the estimates for CRC using a molecular clock based on mutational data and evolutionary analysis [78].

Limitations: We previously conducted comparative analyses of incidence data with a variety of models: simple Markov process models without clonal expansion (e.g., the Armitage and Doll model [38, 79]), the two-stage clonal expansion (TSCE) model [40, 45], and with biologically-motivated extensions of the TSCE model [41, 50–52]. Although the latter usually provide superior fits to cancer incidence data compared to the former [41, 51, 52], MSCE models are by no means complete descriptions of the cancer process, but should be considered biologically-motivated schemata that help to identify critical processes and time scales in carcinogenesis. The models lack many clinical and biological features that may or may not be relevant to our understanding of incidence curves. For example, secular trends may also be viewed as acting quite specifically on biological parameters, while in this study we employ a statistical approach (the age-period-cohort model [50, 52, 75]) to effectively adjust cancer incidence for secular trends. Alternatively, in Chapter 3, we perform model selection allowing such secular trends to act on biological parameters in the case of EAC. Moreover,

our analyses assume that all clonal expansions give rise to (mean) exponential growth even though clinical evidence suggests that tumors may slow their growth in a Gompertzian manner due to limited nutrient/oxygen supplies as the tumor develops vasculature [80]. We also did not model effects of tumor dormancy or potential increases in tumor growth rates due to subtle selection effects in the somatic evolution of the tumor [81]. The inferred cell kinetics does represent an average rate which may comprise passenger mutations that confer weak or no selection and possibly driver mutations that are not rate-limiting (or not requisite) but are likely to speed up the growth process, as well as spatial (niche) effects and clonal interference (as suggested by Martens et al. [82]) that have the potential to slow the tumor growth process. While modeling these processes may well improve our fits and alter certain parameter estimates, it is unlikely that such fine-tuning will alter the parameters associated with the basic two (exponential-then-linear) phases of the incidence curves in a significant way. It is remarkable that in its present form the MSCE model identifies mean sojourn times for tumors that are broadly consistent with clinical estimates in spite of the considerable uncertainties of our estimates and ambiguities in clinical observations.

One way to improve the MSCE model and test model assumptions is to incorporate data from screening and imaging of premalignant as well as malignant tumors. Screening for CRC provides information on the number and sizes of adenomatous polyps and screen detected carcinoma. In the case of EAC, screening may include assessment of the presence or absence of dysplasia and/or chromosomal abnormalities in endoscopic biopsies and surveillance for early cancer. In Chapters 3-5, we will expand the MSCE model for the case of EAC and include tissue and screening modules in order to make such improvements and to utilize screening data. Mechanistic models such as the MSCE model may utilize these different outcomes to enhance our understanding of tumor initiation, growth, persistence and preclinical sojourn.

In this study, we demonstrate that the preclinical phase of malignant tumor progression subtly influences the shape of the age-specific incidence curve, leaving a “footprint” that may be identified through likelihood based analyses of incidence data after adjusting for secular

trends. We identify and estimate three characteristic times scales of carcinogenesis: the mean sojourn time from premalignant cell to first malignant cell, T_1; the mean sojourn time from premalignant cell to first malignant ancestor that generates a persistent clone, T_1^{eff}; and the mean sojourn time it takes for persistent tumors to develop from a single malignant cell to clinical cancer, T_2. We conclude that malignant clone extinction and tumor sojourn times play important roles in reducing and delaying cancer incidence and influencing the shape of incidence curves for colorectal, gastric, and pancreatic cancers.

Chapter 3

Stochastic Modeling of EAC Progression & Incidence

In the previous chapter, we developed the multistage clonal expansion (MSCE) model as a general framework to capture the salient features of cancer development. The mathematical analyses of various incidence patterns showed that the age-specific incidence curves - upon adjustments for secular trends - are well approximated by a model that explicitly incorporates the stochastic growth kinetics of premalignant clones, the sporadic appearance of malignant cells within these clones, and a constant time delay corresponding to the mean sojourn time of a malignant clone. In all future chapters, we will focus on the EAC-specific version of the MSCE model, which we will refer to as the multistage clonal expansion for EAC (MSCE-EAC) model. Here, we expand on the natural history of EAC and explicitly derive the extended set of equations that are necessary to solve in order to connect the cell-level processes occurring during EAC progression to population-level data as we did for the general MSCE model in Chapter 2. Then we will apply this set of mathematical tools, which we refer to as the *MSCE-EAC Cell Module*, to EAC incidence and mortality data to explore recent trends and model projections as part of a comparative modeling project within the Cancer Intervention and Surveillance Modeling Network (CISNET) of the National Cancer Institute (NCI).

3.1 MSCE-EAC Cell Module

The multistage clonal expansion for EAC (MSCE-EAC) model assumes that the stepwise progression to cancer, formulated mathematically as a continuous-time Markov process, involves tissue alteration whereby part of the distal normal esophageal squamous epithelium (with variable extent) undergoes metaplastic transformation resulting in a columnar-lined epithelium called Barrett's esophagus (BE). This tissue alteration provides a natural starting point for the cell-level description of the neoplastic progression to EAC, captured by the general MSCE model of Chapter 2. A preliminary multistage model for EAC progression from BE was first described by Jeon et al. [50] and here we extended this model to include the preclinical, stochastic clonal expansion of malignant cells. Once a tissue conversion occurs resulting in BE of size $X(t)$ stem cells at random age T_{BE}, drawn from an onset time distribution with density f_{BE}, the model continues as the MSCE multi-type branching process within a BE tissue comprised of X(t) number of stem cells. The MSCE-EAC model is a continuous time Markov chain that tracks random stem cell counts of five different types at any continuous time (age) t: Barrett's metaplasia $X(t)$, pre-initiated $P^*(t)$, initiated (pre-malignant) $P(t)$, preclinical malignant $M(t)$, and clinical cancer $C(t)$. Figure 3.1 depicts the MSCE-EAC model dynamics with analogous cell kinetic rates to the general MSCE model illustrated in Fig. 2.2a and the additional transition rate $\nu(t)$ from normal squamous tissue to BE. Here we will describe the three main modeling components that constitute the MSCE-EAC cell module: symptomatic gastroesophageal reflux disease (sGERD) component, hazard component, and temporal trends component.

Component 1. Incorporation of GERD symptoms in BE tissue conversion

The model assumes the random onset of BE to be exponentially distributed with rate $\nu(t)$ based on data that suggested the cumulative incidence of BE increases approximately linearly with age [50, 83]. For a previous 3-stage BE/MSCE model with one clonal expansion

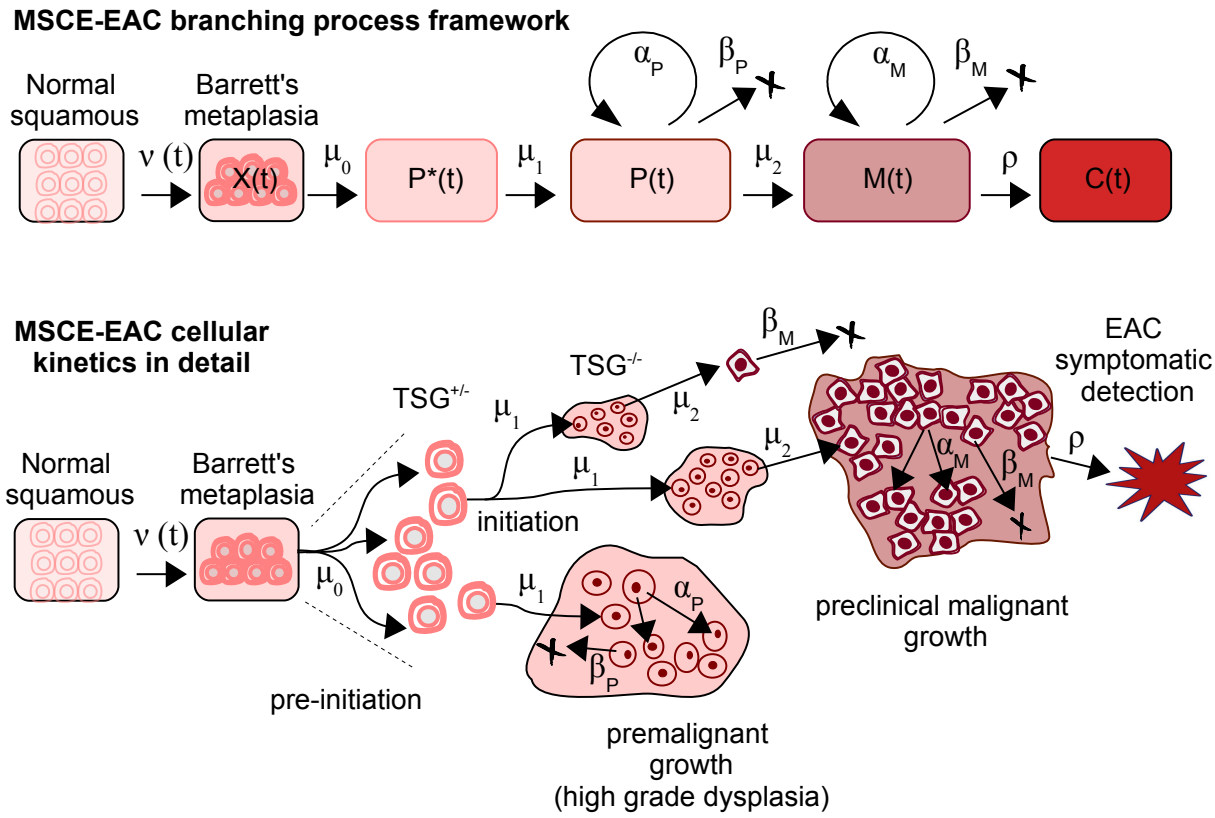

Figure 3.1: **The multistage clonal expansion model for EAC (MSCE-EAC) cell module.** Normal squamous epithelium may transform to BE cells X at a random onset time with rate $\nu(t)$, followed by a 'two-hit' tumor initiation process with Poisson initiation rates μ_0 (creating P^* progeny), and μ_1, which leads to the stochastic appearance of premalignant progenitor P cells in the tissue. Premalignant cells (which we posit to be high grade dysplasia) undergo a first clonal expansion described by a birth-death-migration process with cell division rate $\alpha_{_P}$, cell death-or-differentiation rate $\beta_{_P}$, and malignant transformation rate μ_2. Malignant cells M, in turn, undergo a second clonal expansion by a birth-death-detection process with cell division and death rates $\alpha_{_M}$ and $\beta_{_M}$, respectively. Clinical detection of C cells occurs through a size-based detection process with parameter ρ. TSG, tumor suppressor gene.

stage, Jeon et al. assumed that the rate of BE onset was constant, $\nu(t) = \nu_{BE}$ [50]. However, recent studies and gastroenterology literature support the opinion that symptoms of gastroesophageal reflux disease (GERD) impact the susceptibility of an individual to develop BE metaplasia as a protective response of the tissue to cope with the acidic environment [84–86]. Thus, because GERD increases the risk of BE, we assume that the rate of conversion of normal esophageal tissue to BE metaplasia is GERD-dependent. Here we define symptomatic GERD (sGERD) patients as those with GERD symptoms occurring weekly or more frequently. With this extension, we model the exponential BE rate, $\nu(t)$, as a function of the prevalence of symptomatic GERD in the population, i.e.,

$$\nu(t) = \nu_0 \left((1 - p_{_{sGERD}}(t)) + RR \cdot p_{_{sGERD}}(t) \right), \tag{3.1}$$

where $p_{_{sGERD}}(t)$ is the age-specific prevalence of GERD symptoms, RR is the relative risk of GERD for BE, and ν_0 is a baseline BE rate parameter. Thus, $\nu(t)$ is the rate of BE conversion in a mixed population of people with and without GERD symptoms and includes a baseline rate ν_0 for individuals without GERD and a faster rate for individuals with GERD modeled as $\nu_0 \cdot RR$. (Calibration of the model of GERD prevalence is discussed below). When assuming an exponential distribution for BE onset times, the time-dependent density function $f_{_{BE}}(t)$ is given by

$$f_{_{BE}}(t) = \nu(t) e^{-\int_0^t \nu(s)\,\mathrm{d}s} \tag{3.2}$$

and corresponding cumulative distribution for BE prevalence $F_{_{BE}}$,

$$F_{_{BE}}(t) = \Pr[T_{_{BE}} \leq t] = 1 - e^{-\int_0^t \nu(s)\,\mathrm{d}s}. \tag{3.3}$$

Modeling GERD prevalence

We modeled $p_{_{sGERD}}(t)$ based on data from one study by Ruigomez, et al. [87] for incidence (by 2-year age intervals) of GERD symptoms among children (n=1700), and another study by Ruigomez, et al. [88] on incidence of weekly GERD symptoms among adults (n=1996) with data provided in 10 year intervals. GERD was defined as heartburn and/or regurgitation

experienced at least weekly in these studies. We found that we could achieve excellent fits to these data by using simplified, gender-specific models with three parameters, r_1, r_2, r_3, representing a (slower) transition rate among children, a transition age, and an adult rate for acquiring weekly GERD symptoms, respectively. See Fig. 3.2 for data fits using this simple parametric model. Specifically, the green points in Fig. 3.2 correspond to the three-parameter function $p_{_{sGERD}}(t)$ fit to sGERD data, defined as

$$p_{_{sGERD}}(t) = 1 - \exp(-r_1 \min(r_2, t) - r_3 \min(0, t - r_2)). \tag{3.4}$$

These gender-specific functions for $p_{_{sGERD}}(t)$ are then utilized in the function for the BE rate $\nu(t)$ from Eq. (3.1) along with fixing a value for relative risk RR (we will later set $RR = 5$ based on data from meta-analyses for BE segments greater than 3 cm in length [84]). The parameter ν_0 will then be the one parameter estimated for BE transition density $f_{_{BE}}(t)$ via the MSCE-EAC hazard function, derived in the next section.

Component 2. MSCE-EAC Hazard Function

The MSCE-EAC cell-level description is linked to the population scale by means of the model hazard function, defined as the instantaneous rate of detecting cancers among individuals who have not been previously diagnosed with EAC. This derivation follows the same steps for solving the hazard function as was provided in Chapter 2 for the general MSCE model but represents an extension specific for the case of EAC that includes equations for an 'Armitage-Doll' type initial, one-time transition (AD transition in Fig. 2.1) from normal squamous tissue to BE tissue. Thus we mode a multistage step to signify the onset of BE and then use the generalized MVK model for the following cellular events occurring within BE. We first introduce the notation for the following random variables of the MSCE-EAC

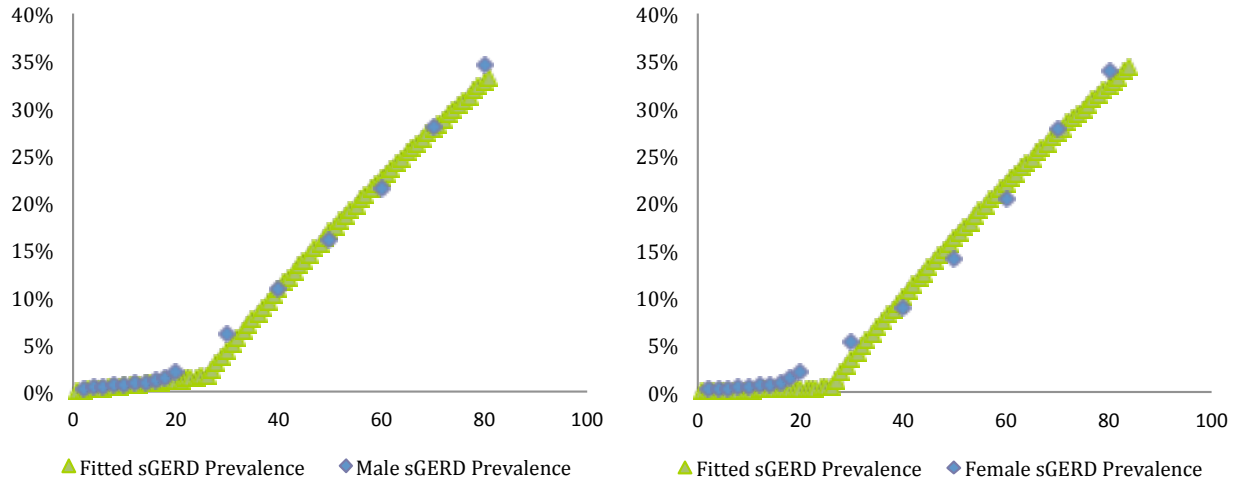

Figure 3.2: **GERD symptom prevalences for males and females.** Gender-specific modeling results for fitted 3-parameter prevalence function (green points), and the sGERD prevalence computed from incidence data (blue points) for (left panel) males, and (right panel) females, all races combined.

multi-type branching process

$$BE(t) = \text{Bernoulli random variable for BE conversion by time } t$$

$$X(t) = \text{number of BE stem cells in a tissue at time } t$$

$$P^*(t) = \text{number of pre-initiated cells at time } t$$

$$P(t) = \text{number of premalignant (initiated) cells at time } t$$

$$M(t) = \text{number of malignant (preclinical) cells (prior to detection) at time } t$$

$$C(t) = \text{number of cancer cells (after detection) at time } t$$

$$D(t) = \text{Bernoulli random variable for clinical detection by time } t$$

Let us consider the probability generating function (PGF) Ψ for the entire process starting at $\tau = 0$, ie. when an individual is born

$$\Psi(y_{BE}, y_1, y_2, y_3, z; t) = \sum_{i,j,k,l,n} y_{BE}^i y_1^j y_2^k y_3^l z^n P(i,j,k,l,n;t),$$

$$\text{where } \; i,n = \{0,1\} \text{ and}$$

$$P(i,j,k,l,n;t)$$

$$= \Pr[BE(t)=i, P^*(t)=j, P(t)=k, M(t)=l, D(t)=n | BE(0)=0, P^*(0)=0, P(0)=0, M(0)=0, D(0)=0].$$

Variables $BE(t), D(t)$ are defined as the following indicator functions corresponding to BE conversion and EAC clinical detection, respectively

$$BE(t) = \begin{cases} 0 \text{ if no BE has not developed by time } t \\ 1 \text{ if BE conversion has taken place by time } t \end{cases}$$

$$D(t) = \begin{cases} 0 \text{ if no cancer detected clinically by time } t \\ 1 \text{ if a malignant cell is detected by time } t. \text{ i.e., } C(\tau) > 0 \text{ for some } \tau \leq t \end{cases}$$

The Chapman-Kolmogorov equations governing the transition probabilities for this multi-stage process include contributions from the initial Armitage-Doll type transition to BE, the two Poisson transitions to initiation, and the two birth-death-migration processes, all of which have been derived previously [41,49,89]. The forward Kolmogorov differential equation for the entire process $\Psi(y_{_{BE}}, y_1, y_2, y_3, z; \tau = 0, t)$ is given by

$$\begin{aligned} \frac{\partial \Psi(y_{_{BE}}, y_1, y_2, y_3, z; t)}{\partial t} &= \nu(t) y_{_{BE}} - \nu(t)\Psi - \mu_0 X(t)(1-y_1) y_{_{BE}} \frac{\partial \Psi}{\partial y_{_{BE}}} - \mu_1 (1-y_2) y_1 \frac{\partial \Psi}{\partial y_1} \\ &+ [\beta_{_P} + \alpha_{_P} y_2^2 - \{\beta_{_P} + \alpha_{_P} + \mu_2(1-y_3)\} y_2] \frac{\partial \Psi}{\partial y_2} \\ &+ [\beta_{_M} + \alpha_{_M} y_3^2 - \{\beta_{_M} + \alpha_{_M} + \rho(1-z)\} y_3] \frac{\partial \Psi}{\partial y_3}, \end{aligned} \tag{3.5}$$

where we have suppressed the dependence on $(y_{_{BE}}, y_1, y_2, y_3, z; t)$ in Ψ for convenience. This six-dimensional PDE may be solved by the method of characteristics but poses numerical issues. Here we will show a more amenable method for solving the generating function using the Kolmogorov backward equations.

Backward Kolmogorov equations

Beginning with an active BE segment (BE), a single pre-initiated (P^*), premalignant (P), or malignant (M) cell at time τ only, we define the following generating functions $\Phi_{_{BE}}$,$\Phi_{_{P^*}}$,

Φ_{P}, or Φ_{M}, respectively,

$$\Phi_{M}(y_3,z;\tau,t) = E[y_3^{M(t)}z^{D(t)}|M(\tau)=1,D(\tau)=0] \tag{3.6}$$
$$= \sum_{k,l} y_3^k z^l \Pr[M(t)=k,D(t)=l|M(\tau)=1,D(\tau)=0]$$

$$\Phi_{P}(y_2,y_3,z;\tau,t) = E[y_2^{P(t)}y_3^{M(t)}z^{D(t)}|P(\tau)=1,M(\tau)=0,D(\tau)=0] \tag{3.7}$$
$$= \sum_{j,k,l} y_2^j y_3^k z^l \Pr[P(t)=j,M(t)=k,D(t)=l|P(\tau)=1,M(\tau)=0,D(\tau)=0]$$

$$\Phi_{P^*}(y_1,y_2,y_3,z;\tau,t) = E[y_1^{P^*(t)}y_2^{P(t)}y_3^{M(t)}z^{D(t)}|P^*(\tau)=1,P(\tau)=0,M(\tau)=0,D(\tau)=0] \tag{3.8}$$
$$= \sum_{i,j,k,l} y_1^i y_2^j y_3^k z^l \Pr[P^*(t)=i,P(t)=j,M(t)=k,D(t)=l|P^*(\tau)=1,P(\tau)=0,M(\tau)=0,D(\tau)=0]$$

$$\Phi_{BE}(y_{BE},y_1,y_2,y_3,z;\tau,t) = E[y_{BE}^{BE(t)}y_1^{P^*(t)}y_2^{P(t)}y_3^{M(t)}z^{D(t)}|BE(\tau)=1,P^*(\tau)=0,P(\tau)=0,M(\tau)=0,D(\tau)=0] \tag{3.9}$$

$$= \sum_{i,j,k,l,n} y_{BE}^i y_1^j y_2^j y_3^l z^n \Pr[BE(t)=i,P^*(t)=j,P(t)=k,M(t)=l,D(t)=n|BE(\tau)=1,P^*(\tau)=0,P(\tau)=0,M(\tau)=0,D(\tau)=0]$$

The generating functions satisfy the following Kolmogorov backward equations

$$\frac{\partial\Phi_{M}(y_3,z;\tau,t)}{\partial\tau} = -\alpha_{M}\Phi_{M}^2(y_3,z;\tau,t) - \beta_{M} \tag{3.10}$$
$$- z\rho\Phi_{M}(y_3,z;\tau,t) + [\alpha_{M}+\beta_{M}+\rho]\Phi_{M}(y_3,z;\tau,t)$$

$$\frac{\partial\Phi_{P}(y_2,y_3,z;\tau,t)}{\partial\tau} = -\alpha_{P}\Phi_{P}^2(y_2,y_3,z;\tau,t) - \beta_{P}$$
$$+ [\alpha_{P}+\beta_{P}+\mu_2]\Phi_{P}(y_2,y_3,z;\tau,t) \tag{3.11}$$
$$- \mu_2\Phi_{P}(y_2,y_3,z;\tau,t)\Phi_{M}(y_3,z;\tau,t)$$

$$\frac{\partial\Phi_{P^*}(y_1,y_2,y_3,z;\tau,t)}{\partial\tau} = -\mu_1\Phi_{P^*}(y_1,y_2,y_3,z;\tau,t)[\Phi_{P}(y_2,y_3,z;\tau,t)-1] \tag{3.12}$$

$$\frac{\partial\Phi_{BE}(y_{BE},y_1,y_2,y_3,z;\tau,t)}{\partial\tau} = -\mu_0 X(t)\Phi_{BE}(y_{BE},y_1,y_2,y_3,z;\tau,t)[\Phi_{P^*}(y_1,y_2,y_3,z;\tau,t)-1] \tag{3.13}$$

$$\frac{\partial\Psi(y_{BE},y_1,y_2,y_3,z;\tau,t)}{\partial\tau} = \nu(\tau)[\Psi(y_{BE},y_1,y_2,y_3,z;\tau,t) - \Phi_{BE}(y_{BE},y_1,y_2,y_3,z;\tau,t)] \tag{3.14}$$

To obtain the hazard function, we first solve for the overall survival function (for EAC cancer detection), starting at time 0, which in our EAC-specific notation for the MSCE model is given by

$$S_{EAC}(t) = 1 - P_{EAC}(t) = \Pr[D(t)=0|BE(0)=0,P^*(0)=0,P(0)=0,M(0)=0,D(0)=0]$$
$$= \Psi(1,1,1,1,0;0,t)$$

where $P_{EAC}(t)$ is the probability of a EAC cancer detection at time t,

$$P_{EAC}(t) = \Pr[D(t) = 1 | BE(0) = 0, P^*(0) = 0, P(0) = 0, M(0) = 0, D(0) = 0]$$

We will here denote $\Phi_M(1, 0; \tau, t) \equiv \Phi_M(\tau, t)$, $\Phi_P(1, 1, 0; \tau, t) \equiv \Phi_P(\tau, t)$, $\Phi_{P^*}(1, 1, 1, 0; \tau, t) \equiv \Phi_{P^*}(\tau, t)$, $\Phi_{BE}(1, 1, 1, 1, 0; \tau, t) \equiv \Phi_{BE}(\tau, t)$, and $\Psi(1, 1, 1, 1, 0; \tau, t) \equiv \Psi(\tau, t)$. A dot designates a first derivative with respect to t. The hazard function, i.e., the rate at which EAC is detected in individuals who have not been diagnosed before, is given by

$$h_{EAC}(t) = -\frac{\dot{S}_{EAC}(t)}{S_{EAC}(t)} = -\frac{\dot{\Psi}(0, t)}{\Psi(0, t)} = -\frac{d}{dt}\ln[\Psi(0, t)]$$

Numerical solutions

For fixed t, this boundary value system of coupled PDEs in Eqs. (3.10-3.14) can be converted into an initial value problem (IVP) with the change of variables $u = t - \tau$, where u is the "running" time. As utilized in Chapter 2, we can use this redefinition and the following method to solve for the hazard function numerically [67]. Define the following variables for the new IVP: $Y_1(u, t) = \Phi_M(\tau, t), Y_2(u, t) = \dot{\Phi}_M(\tau, t), Y_3(u, t) = \Phi_P(\tau, t), Y_4(u, t) = \dot{\Phi}_P(\tau, t), Y_5(u, t) = \Phi_{P^*}(\tau, t), Y_6(u, t) = \dot{\Phi}_{P^*}(\tau, t), Y_7(u, t) = \Phi_{BE}(\tau, t), Y_8(u, t) = \dot{\Phi}_{BE}(\tau, t), Y_9(u, t) = \Psi(\tau, t)$, $Y_{10}(u, t) = \dot{\Psi}(\tau, t)$ with corresponding initial conditions $Y_1(0, t) = Y_3(0, t) = Y_5(0, t) = Y_7(0, t) = Y_9(0, t) = 1, Y_4(0, t) = Y_6(0, t) = Y_8(0, t) = Y_{10}(0, t) = 0$, and $Y_2(0, t) = -\rho$. Then

the equations to solve for our IVP are the following

$$\frac{dY_1(u,t)}{du} = \beta_{_M} - (\alpha_{_M} + \beta_{_M} + \rho)Y_1(u,t) + \alpha_{_M}Y_1^2(u,t) \tag{3.15}$$

$$\frac{dY_2(u,t)}{du} = 2\alpha_{_M}Y_1(u,t)Y_2(u,t) - (\alpha_{_M} + \beta_{_M} + \rho)Y_2(u,t) \tag{3.16}$$

$$\frac{dY_3(u,t)}{du} = \beta_{_P} + \mu_2 Y_1(u,t)Y_3(u,t) - (\alpha_{_P} + \beta_{_P} + \mu_2)Y_3(u,t) + \alpha_{_P}Y_3^2(u,t) \tag{3.17}$$

$$\begin{aligned}\frac{dY_4(u,t)}{du} &= 2\alpha_{_P}Y_3(u,t)Y_4(u,t) + \mu_2(Y_4(u,t)Y_1(u,t) + Y_3(u,t)Y_2(u,t)) \\ &\quad - (\alpha_{_P} + \beta_{_P} + \mu_2)Y_4(u,t)\end{aligned} \tag{3.18}$$

$$\frac{dY_5(u,t)}{du} = \mu_1 Y_5(u,t)(Y_3(u,t) - 1) \tag{3.19}$$

$$\frac{dY_6(u,t)}{du} = \mu_1(Y_6(u,t)Y_3(u,t) - Y_6(u,t) + Y_5(u,t)Y_4(u,t)) \tag{3.20}$$

$$\frac{dY_7(u,t)}{du} = \mu_0 X Y_7(u,t)(Y_5(u,t) - 1) \tag{3.21}$$

$$\frac{dY_8(u,t)}{du} = \mu_0 X(t)(Y_8(u,t)Y_5(u,t) - Y_8(u,t) + Y_7(u,t)Y_6(u,t)) \tag{3.22}$$

$$\frac{dY_9(u,t)}{du} = \nu(u)(Y_7(u,t) - Y_9(u,t)) \tag{3.23}$$

$$\frac{dY_{10}(u,t)}{du} = \nu(u)(Y_{10}(u,t) - Y_8(u,t)) \tag{3.24}$$

Note that first 8 equations, Eqs. (3.15-3.22), are the same as the entire set of Eqs. (2.15-2.22) for the general MSCE model of Chapter 2. With the two new ODEs accounting for the Armitage-Doll transition to BE with non-constant rate $\nu(t)$, we may solve these 10 coupled ODEs numerically to obtain the EAC hazard and survival functions,

$$h_{_{EAC}}(t) = -Y_{10}(t,t)/Y_9(t,t) \tag{3.25}$$

$$S_{_{EAC}}(t) = Y_9(t,t). \tag{3.26}$$

MSCE-EAC convolution formula

For EAC, we follow the derivation of a preliminary BE/MSCE model [50] to alternatively construct the hazard function of the MSCE-EAC model as the combination of two stochastic processes: the random occurrence of BE and the multistage carcinogenesis process arising

in BE. Mathematically, the MSCE-EAC probability density function may be written as a convolution of the BE conversion density f_{BE} (which we will assume to be exponential, defined in Eq. (3.2), in Chapters 3-5) and the MSCE model density after BE onset, f_{MSCE}:

$$f_{EAC}(t) = \int_0^\infty f_{BE}(u) f_{MSCE}(t-u)\,\mathrm{d}u \tag{3.27}$$

$$= \int_0^t f_{BE}(u) f_{MSCE}(t-u)\,\mathrm{d}u, \quad \text{if } u > t, \text{ then } f_{MSCE} = 0 \tag{3.28}$$

First, we can compute the hazard as follows

$$h_{EAC}(t) = -\frac{\dot{S}_{EAC}(t)}{S_{EAC}(t)} = \frac{f_{EAC}(t)}{S_{EAC}(t)} \tag{3.29}$$

$$= \frac{\int_0^t f_{BE}(u) \cdot f_{MSCE}(t-u)\,\mathrm{d}u}{1 - \int_0^t \int_0^s f_{BE}(u) f_{MSCE}(s-u))\,\mathrm{d}u\,\mathrm{d}s} \tag{3.30}$$

$$= \frac{\int_0^t f_{BE}(u) \cdot f_{MSCE}(t-u)\,\mathrm{d}u}{1 - \int_0^t f_{BE}(u) \cdot (1 - S_{MSCE}(t-u))\,\mathrm{d}u} \tag{3.31}$$

where $S_{MSCE} = Y_7(t,t)$ and $f_{MSCE} = -Y_8(t,t)$ are the numerical solutions to Eqs. (3.21-3.22), which were previously defined in Chapter 2. As with the numerical solution for $h_{MSCE}(t)$ provided in Eq. (3.25), Eq. (3.31) does not require the biological rates of the model to be constant.

Analytical Approximation with MSCE-1

With this convolution formula for the EAC hazard function, we will apply the results from Chapter 2 to derive an analytical hazard function for the MSCE-EAC model approximation. With the assumption of constant parameters except for age-dependent $\nu(t)$, we can avoid numerical solvers by using the MSCE-1 approximation from Chapter 2 for the set of ODEs in Eqs. (3.15-3.24). Thus, for t such that $t_{lag} \leq t$,

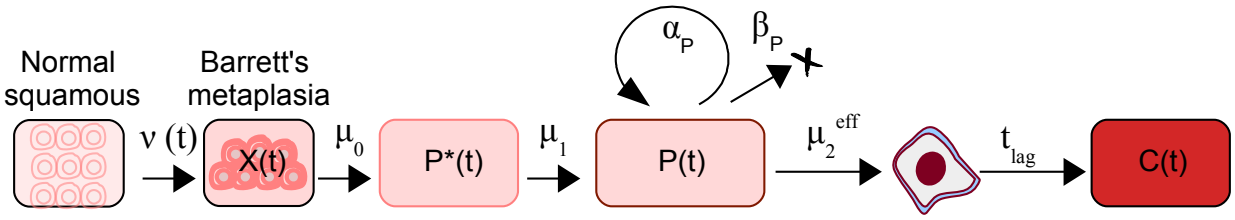

Figure 3.3: **The MSCE-1 approximation for MSCE-EAC.** The MSCE-EAC model is well approximated by the MSCE-1 approximation that includes an effective malignant transformation rate μ_2^{eff} (see Chapter 2) and a constant lag-time for tumor progression. This approximate model does not include a random variable $M(t)$ for the number of malignant cells at time t.

$$h_{EAC}(t) \approx \frac{\int_0^{t-t_{lag}} f_{BE}(u) \cdot f_{MSCE-1}(t-u)du}{1-\int_0^{t-t_{lag}} f_{BE}(u) \cdot (1-S_{MSCE-1}(t-u))\,\mathrm{d}u} \tag{3.32}$$

$$= \frac{\int_0^{t-t_{lag}} f_{BE}(u) \cdot h_{MSCE-1}(t-u) \cdot S_{MSCE-1}(t-u)\,\mathrm{d}u}{1-\int_0^{t-t_{lag}} f_{BE}(u) \cdot (1-S_{MSCE-1}(t-u))\,\mathrm{d}u} \tag{3.33}$$

$$= \frac{\int_0^{t-t_{lag}} f_{BE}(u) \cdot h_{MSCE-1}(t-u) \cdot \left(e^{-\int_0^{t-u} h_{MSCE-1}(s)ds}\right)\,\mathrm{d}u}{1-\int_0^{t-t_{lag}} f_{BE}(u) \cdot \left(1-e^{-\int_0^{t-u} h_{MSCE-1}(s)ds}\right)\,\mathrm{d}u}. \tag{3.34}$$

where we note that $f_{MSCE-1}(t-u) = 0$ for $u > t - t_{lag}$. Again, the 3-stage MSCE-1 model hazard approximates the MSCE hazard (proof in Chapter 2) and was previously defined in Eq. (2.36) as

$$h_{MSCE-1}(t) = \mu_0 X \left(1 - \left(\frac{q_P^{eff} - p_P^{eff}}{q_P^{eff} e^{-p_P^{eff}(t-t_{lag})} - p_P^{eff} e^{-q_P^{eff}(t-t_{lag})}}\right)^{\mu_1/\alpha_P}\right)$$

where

$$
\begin{aligned}
t_{_{lag}} &= T_2 \ \ \text{(see Eq. (2.32))} \\
p_{_P}^{^{eff}} &= \frac{1}{2}\left(-(\alpha_{_P}-\beta_{_P}-\rho) - \sqrt{(\alpha_{_P}-\beta_{_P}-\mu_2^{^{eff}})^2 + 4\alpha_{_P}\mu_2^{^{eff}}}\right) \\
q_{_P}^{^{eff}} &= \frac{1}{2}\left(-(\alpha_{_P}-\beta_{_P}-\rho) + \sqrt{(\alpha_{_P}-\beta_{_P}-\mu_2^{^{eff}})^2 + 4\alpha_{_P}\mu_2^{^{eff}}}\right)
\end{aligned}
$$

See Fig. 3.4 for a comparison of the three hazard functions fitted to SEER incidence that we derived for the MSCE-EAC model: the full numerical hazard (Eq. (3.25)), the convolution hazard (Eq. (3.31)), and the analytical MSCE-1 approximate hazard for EAC (Eq. (3.34)). In the next component, we will describe how the MSCE-EAC model hazard is employed to fit biological model parameters to EAC incidence data.

Component 3. MSCE-EAC likelihood and trends

EAC incidence has increased approximately six-fold in the US since 1975, as reflected in the SEER data [7]. To explore the mechanism of such a dramatic increase, we systematically applied parametric period and cohort trends to the biological parameters of the MSCE-EAC model, and used likelihood methods for model comparison and selection of the best model fit to SEER incidence. Both linear (two parameter) and sigmoidal (three parameter) trends by cohort and/or period were applied to one or more of the five biological processes represented by the MSCE-EAC model: 1) the transition from normal to BE tissue ($\nu(t)$), 2) the rate of two rate-limiting mutations transforming BE stem cells to premalignant cells (assumed to be equal for identifiability, μ_0, μ_1), 3) clonal expansion of premalignant cells ($g_{_P}$ and $\alpha_{_P}$), 4) malignant transformation (μ_2), and 5) clonal growth of malignant tissue ($g_{_M}$ and $\alpha_{_M}$). This analysis identified a highly significant sigmoidal birth cohort effect when modifying the growth rates of premalignant and malignant cells, with rates for malignant growth estimated to be significantly larger than those for premalignant growth (see [90] for more details about model selection). In this section we will provide the sigmoidal parameterization of the birth cohort effect and then the MSCE-EAC likelihood function used to calibrate the model

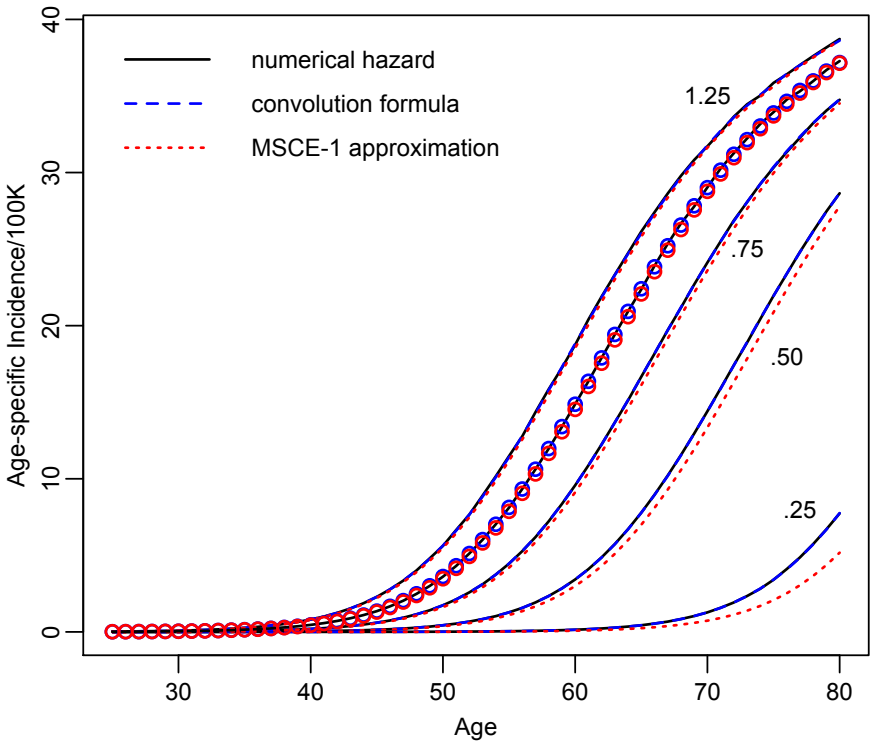

Figure 3.4: **Comparison of three methods to compute the hazard $h_{EAC}(t)$.** Black solid lines are solutions of 10 coupled ODEs, given in Eq. (3.25), blue dashed lines are solutions to the convolution formula, given by Eq. (3.31) involving numerical solutions to $S_{MSCE} = Y_7(t,t)$ and $f_{MSCE} = -Y_8(t,t)$, and red dotted lines are MSCE-1 approximation hazard, given by Eq. (3.34), which includes effective mutation rate μ_2^{eff} and analytical mean sojourn time of malignant tumor T_2. When the MSCE model parameters are constant, the third solution is analytical and requires no numerical solution of ODEs. Curves with circles correspond to actual model fit with biological parameters fit to SEER incidence data for US males of all races, born in 1950. Other four curves examine accuracy of approximation by changing malignant cell proliferation, g_M, by factor indicated. As was derived in Chapter 2, the approximation weakens as g_M approaches g_P.

parameters. The outcomes of such fitted parameters for EAC will be discussed more in the application presented in the following section.

Let b_k represent the birth year for birth cohort, indexed by k. For a person born in birth cohort k, the sigmoidal shape for premalignant (P stem cells) growth rate $g_{P,k}$, cell division rate $\alpha_{P,k}$, and cell death rate $\beta_{P,k}$ are given by

$$g_{P,k} = g_{P,0}\left(g_1 + \frac{2}{1+e^{(-g_2(b_k-g_3))}}\right) \tag{3.35}$$

$$\alpha_{P,k} = \alpha_{P,0}\left(g_1 + \frac{2}{1+e^{(-g_2(b_k-g_3))}}\right) \tag{3.36}$$

$$\beta_{P,k} = \alpha_P - \mu_2 - g_{P,k} \tag{3.37}$$

Similarly, for a person born in birth cohort k, the sigmoidal shape for malignant (M stem cells) growth rate $g_{M,k}$, cell division rate $\alpha_{M,k}$, and cell death rate $\beta_{M,k}$ are given by

$$g_{M,k} = g_{M,0}\left(g_1 + \frac{2}{1+e^{(-g_2(b_k-g_3))}}\right) \tag{3.38}$$

$$\alpha_{M,k} = \alpha_{M,0}\left(g_1 + \frac{2}{1+e^{(-g_2(b_k-g_3))}}\right) \tag{3.39}$$

$$\beta_{M,k} = \alpha_M - \rho - g_{M,k} \tag{3.40}$$

The model estimates values for $g_{P,0}, g_1, g_2, g_3$ (reference year), for each b_k. Assumptions must be made for values of the background cell division rates, $\alpha_{M,0}$ and $\alpha_{P,0}$, and the background malignant proliferation rate $g_{M,0}$. This will be discussed more in the next section.

We use maximum likelihood methods to fit to EAC incidence from SEER for ages 1 to 84 and calendar years 1975-2010 with this "Age-Cohort (AC) model." Explicitly, the expected number of EAC cancers at age a_i, period (calendar year) p_j, and birth cohort $b_k = c_j - a_i$ is

$$\Lambda_{i,j} = PY_{i,j} h_{EAC}(a_i, b_k), \tag{3.41}$$

where $PY_{i,j}$ is the number of person years of age a_i and period p_j (also provided for specified 'at risk' US population by SEER [7]), and the birth cohort specific hazard is

$$h_{EAC}(a_i, b_k) = h_{EAC}(t|t = a_i, b_k) \tag{3.42}$$

The likelihood is then given by

$$\mathcal{L} = \prod_{i,j} \frac{\Lambda_{i,j} \exp -\Lambda_{i,j}}{\mathscr{O}_{i,j}!}, \tag{3.43}$$

where $\mathscr{O}_{i,j}$ is the number of observed EAC cases for age a_i and period p_j (from SEER). As mentioned previously, the analytical, approximate form of $h_{EAC}(t)$ from Eq. (3.34) can be used in this likelihood function to obtain MLE estimates when biological parameters are constant (excluding $\nu(t)$, which may be age-dependent in the convolution during estimation). For the best-fitting model with a sigmoidal birth cohort effect on premalignant and malignant cell proliferation rates, this method of estimation may be used for each birth cohort individually (see Fig. 3.4 for approximation accuracy).

These three components complete our description of the MSCE-EAC cell module that mathematically formulates the multistage clonal expansion process for EAC on the cellular scale and connects this process to the population scale via the MSCE-EAC hazard function, which is amenable to maximum likelihood calibration. In the next section, we will present motivation, methods, and results obtained from fitting the MSCE-EAC cell module to SEER incidence data used to predict future EAC trends.

3.2 CISNET Base Case I: Model Calibration to EAC Incidence

Recent analyses of historical trends in EAC incidence and mortality in the U.S. suggest that EAC incidence continues to rise, although the EAC incidence rate may be beginning to plateau in recent years [5, 91]. Projections of future EAC incidence and mortality would provide important data for health policy makers as they track cancer trends and plan appropriate cancer control policy. In this analysis, we collaborated with two other modeling groups in the Cancer Intervention and Surveillance Modeling Network (CISNET) Esophagus group, as introduced in Chapter 1, in a comparative modeling exercise to make future EAC

incidence and mortality projections (for the complete publication, see [6]). We termed this analysis "Base Case I ." Three independent models (including the MSCE-EAC model) were calibrated to historic EAC incidence and mortality rates between the years 1975-2010 for all males and all females from the SEER 9 database [7]. The models were used to generate independent projections of EAC incidence and mortality to the year 2030. We also outputted the BE prevalence, cancer progression rates in patients with BE, and the average EAC sojourn time (time interval between preclinical EAC and clinically detected EAC); all are important underlying factors that may be subject to secular trends related to period and/or birth cohort. In modeling these trends we developed a better understanding of the observed increases in EAC incidence and mortality. In this section, we will focus on the methodology (continuing from the previous section) and calibration results produced by the MSCE-EAC cell module. However, we will also include a few results that compare the MSCE-EAC outcomes produced by the two other mathematical models (see full publication [6] for inclusion of all results).

3.2.1 Materials and methods

Common calibration targets

All three CISNET models were calibrated to EAC incidence rates from the SEER9 program data for men and women of all races aged 20-84 years in the United States from 1975-2010. The cancer incidence rates are comprised of cancers defined/identified by the International Classification of Diseases for Oncology, third edition (ICD-O-3) histology codes 8140- 8141, 8143- 8145, 8190- 8231, 8260-8263, 8310, 8401, 8480-8490, 8550-8551, 8570-8574, and 8576. Standard mortality statistics are not available for EAC because death certificates do not include the histology of the cancer (esophageal squamous cell cancer (ESCC) vs. EAC). However, incidence-based (IB) mortality data in SEER utilize cancer registry information to link characteristics of the incident cancer (e.g. stage, histology) to individual death certificates [92]. Here we use the survival rates given for the EAC cases compiled in SEER

stratified by gender and age groups, as model inputs to calculate the mortality rates. All three models also use U.S. census data and projections for past (from 1975 onwards) and future (up to 2030) population size [93].

Model outputs

The primary endpoints or model outputs for this study were the projections of overall EAC incidence and mortality rates up to calendar year 2030. For the MSCE-EAC model, SEER incidence rates for years 1975 to 2010 were organized by birth cohorts, based on model selection described in the previous section [90], and used as input in the likelihood function (see Eq. (3.43)) to estimate the biological model parameters. Other predictions included the average time between developing preclinical cancer and cancer diagnosis (EAC mean sojourn time given by T_2 in Eq. (2.32)) and the annual rate of patients diagnosed with BE (without dysplasia) progressing to clinically diagnosed EAC, to be compared with rates in the literature. These progression rates were calculated for age-at-diagnosis with BE at age 60 and with five years of follow-up time. The cancer progression rate measures the percentage of BE patients who advance to EAC annually. Additional intermediate outputs included the prevalences of GERD (see Eq. (3.4)) and BE (see Eq. (3.3)) adjusted to the 2000 U.S. population.

3.2.2 Results

Parameter identifiability and estimates

All cellular kinetic model parameters were estimated using a maximum likelihood method to obtain optimal fits to SEER incidence data. We obtained 95% credible intervals for these estimates using Markov Chain Monte Carlo (MCMC) to sample the posterior distribution of the mddl parameters. All MCMC runs were started with the parameters set at (or near) their respective maximum likelihood estimates (MLEs) and appeared to converge rapidly after a short 1000 cycle burn-in period. The maximum likelihood values were estimated

using the Davidon-Fletcher-Powell gradient search method, available in the R software Bhat package online [94].

During estimation, we compared multiple models by fixing $g_{M,0}$ and detection rate ρ to different values in order to achieve reasonable mean sojourn times and tumor doubling times that are in line with clinical data. We estimated the EAC clinical detection rate $\rho = 10^{-9}$ per cell/year, and malignant cell proliferation rate $g_{M,0} = 0.75$ per cell/year. We fixed $\alpha_{P,0} = 10$ per cell/year, and $\alpha_{M,0} = 150$ per cell/year, and also set the number of stem cells in an average 5 cm BE segment $X = 10^6$. With these assumptions, we obtain the parameter estimates provided in Table 3.1. We will continue to use these parameter estimates for the illustrations and examples presented in the following chapters.

Table 3.1: MSCE-EAC model biological parameters.

Value (95% CI)	Males	Females
ν_0	$3.65\ (3.19 - 4.13) \times 10^{-4}$	$7.48\ (4.87 - 10.29) \times 10^{-5}$
$\mu_0(\mu_1)$	$7.99\ (6.38 - 9.83) \times 10^{-4}$	$7.05\ (6.13 - 12.25) \times 10^{-4}$
μ_2	$4.54\ (3.65 - 6.47) \times 10^{-5}$	$6.89\ (3.16 - 14.28) \times 10^{-5}$
$g_{P,0}$ *	$9.91\ (9.28 - 10.99) \times 10^{-2}$	$1.23\ (1.06 - 1.35) \times 10^{-1}$
g_1 *	$5.09\ (2.75 - 5.90) \times 10^{-1}$	$6.40\ (2.16 - 8.44) \times 10^{-1}$
g_2 *	$5.38\ (4.83 - 5.72) \times 10^{-2}$	$2.98\ (2.47 - 3.44) \times 10^{-2}$
g_3 *	1912.5 (1909.1 - 1914.1)	1945.3 (1923.9 - 1954.4)

All parameter estimates fit to SEER incidence data and have the units of per cell per year. Markov Chain Monte Carlo 95% credible intervals provided beside the maximum likelihood estimates.

*Parameter elements of sigmoidal functions for clonal proliferation provided in Eqs. (3.35-3.40).

EAC incidence and projections

After estimating the MSCE-EAC model parameters, it is straightforward to calculate the predicted EAC incidence for each birth cohort k into the future by computing the hazard function with the estimates as input for any $t > b_k$. See Fig. 3.5 for the age-specific male (upper panel) and female (lower panel) incidence rates until year 2030 for each birth cohort separately, with model predictions produced by the MSCE-EAC cell module. All three CISNET models projected that the male EAC incidence will continue to increase. For later male birth cohorts, incidence results show that the incremental differences between birth cohorts has decreased, indicating a deceleration of the birth cohort effect. For females, the projected incidence rates from all models also showed increasing incidence rates by advancing birth cohorts. However, the slowing trend in the incremental differences between birth cohorts was not as clear, likely due to the relatively low incidence rates for females with greater statistical variance of the estimated parameters resulting in model predictions with greater uncertainty. After aggregating the projections for all cohorts, the total EAC incidence rate for males (upper panel) and females (lower panel) between ages 20-84 (age-adjusted to the 2000 U.S. population) were calculated with the MSCE-EAC cell module and are shown in Fig. 3.6, along with results from the other two models for comparison (mortality rates not shown, see [6]). Despite the differences in approach and mathematical formalism between the three models, all three models yielded small variations in their model fits to SEER and projections to 2030 for total EAC incidence rate. The three models all projected an increase in EAC incidence until 2030. The ranges of incidence rate for all males in 2030 predicted by three models was 8.4-10.1 cases per 100,000 person years. This translates to a 7-10 fold increase in the EAC incidence rate, from 1975 to 2030. For all females, the future incidence rate in 2030 was estimated to be 1.3-1.8 cases per 100,000 person years. From 1975 to 2030, the incidence rate for females was estimated to increase by 8-9 fold.

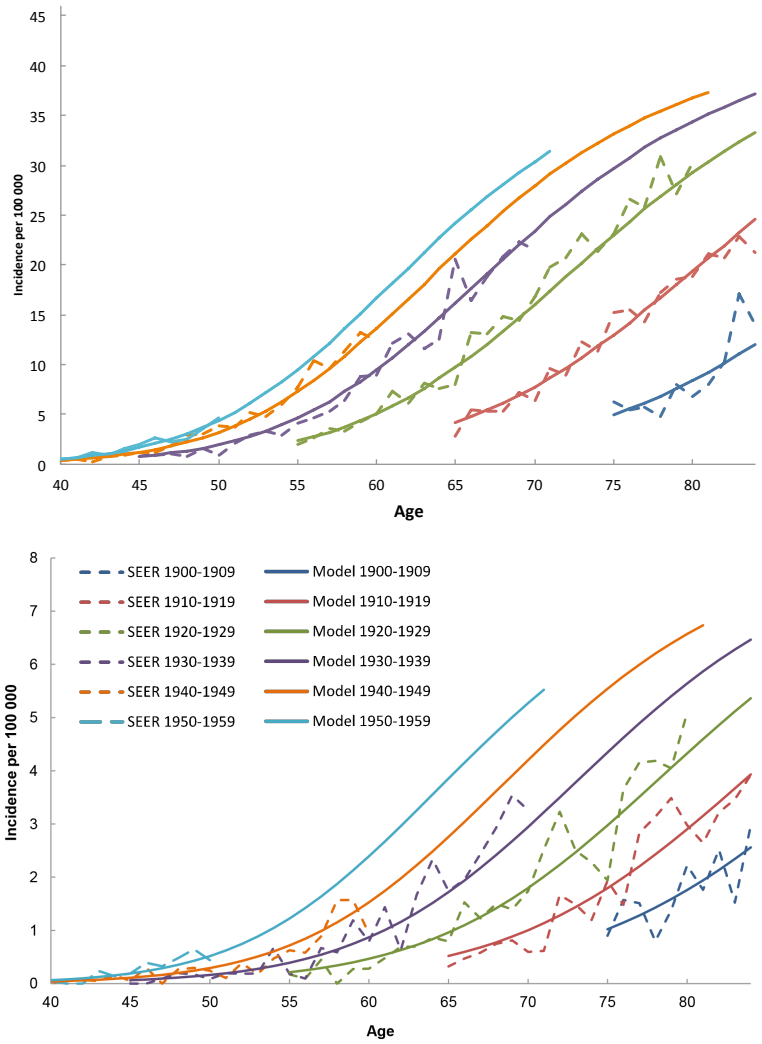

Figure 3.5: **MSCE-EAC calibration to EAC incidence with projections by birth cohort.** Age-specific male (upper panel) and female (lower panel) incidence rates by 10 year birth cohorts, all races combined. Available SEER data shown by dotted lines, MSCE-EAC model fits shown by solid lines, with projections until year 2030. The cohort born in 1959 would be 71 years old in calendar year 2030.

EAC progression rate from BE

The annual rate of BE patients progressing to clinical EAC predicted by the MSCE-EAC model are shown in Fig. 3.7 for males (left panel) and females (right panel), along with re-

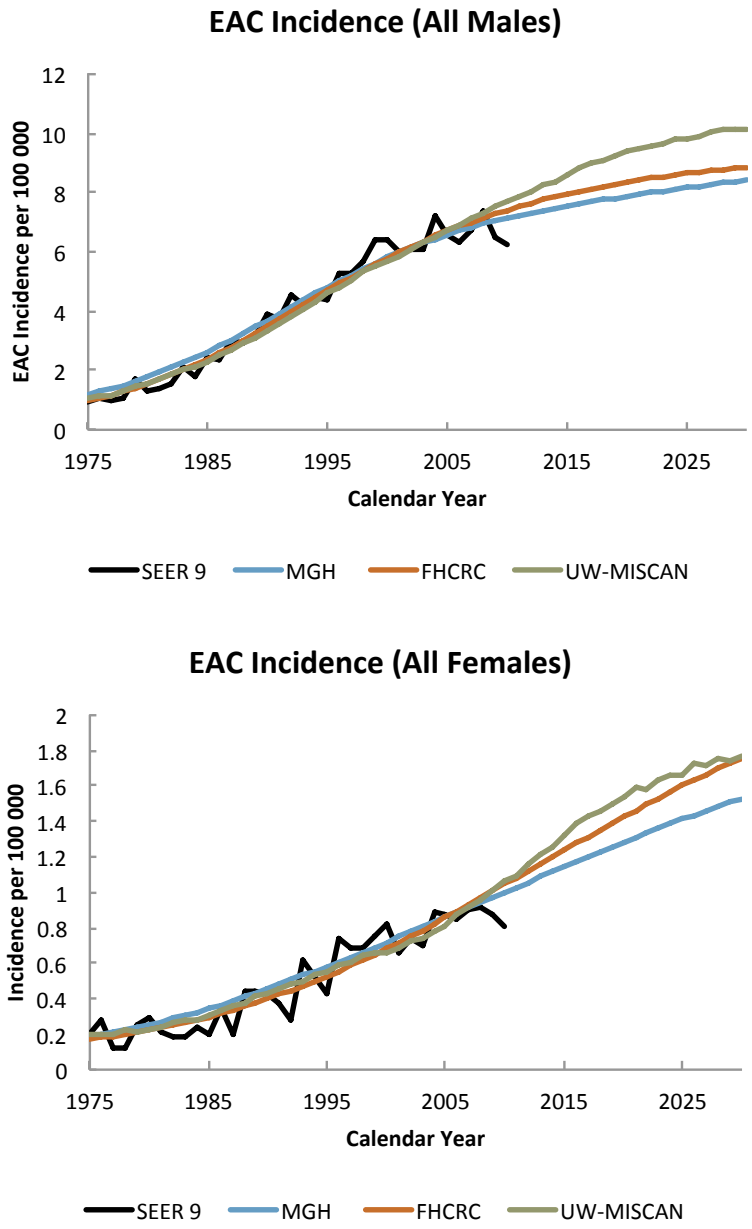

Figure 3.6: **MSCE-EAC incidence projections compared to other models.** Age-specific male (upper panel) and female (lower panel) incidence rates, all races combined. Available SEER data shown by black line, model fits shown by colored solid lines, with projections until year 2030. MSCE-EAC model results denoted by "FHCRC," orange lines.

sults from the other two CISNET models for comparison. The results are shown for patients born between year 1915 to 1970 (these patients will be at age 60 between year 1975 and 2030) who were diagnosed with non-dysplastic BE at age 60 with five years of follow-up time. Again, we consider the premalignant clones (P clones) of the MSCE-EAC cell module to be representative of dysplastic lesions within BE tissue. In order for the MSCE-EAC model to compute this output, the cell module must be supplemented by two other modules, the MSCE-EAC tissue and screening modules because the progression rate depends on the physical size of the neoplastic (dysplastic and early cancer) clones for detectability during endoscopy (see Chapter 4 for full description). Briefly, the tissue module simulates individuals with BE using the cell module parameters for a large sample size (100K for less than .001 standard error on expected values), records their life trajectories and branching process cell counts as defined in the MSCE-EAC cell module until an index screen is performed at age 60, and simulates the spatial position of the clones on a BE segment. Then the screening module takes this realized BE segment, performs a biopsy protocol *in silico* that determines if dysplasia is present, and computes the number of non-dysplastic patients that develop EAC in the first five years of follow-up. The MSCE-EAC tissue and screening modules will be described in detail in Chapter 4, but we will include the results for MSCE-EAC progression rates here for completeness because they were an output for Base Case I . All three models suggested a strong birth cohort effect on the BE-to-EAC progression rates with increasing rates in younger birth cohorts until the cohorts born in 1940, followed by a leveling off in the cohorts born after 1940. For males born after 1940, the ranges of progression rates are 0.10-0.20% per person year. However, the three models predict that in contrast to males, progression rates for females have not yet leveled off.

EAC sojourn time

As described in Chapter 2, the EAC sojourn time T_2 (see Eq. (2.32)) is a useful concept to understand the nature of disease progression and detection. Again, in the MSCE-EAC model, T_2 is the mean time of a surviving malignant tumor, from time of the first malignantly

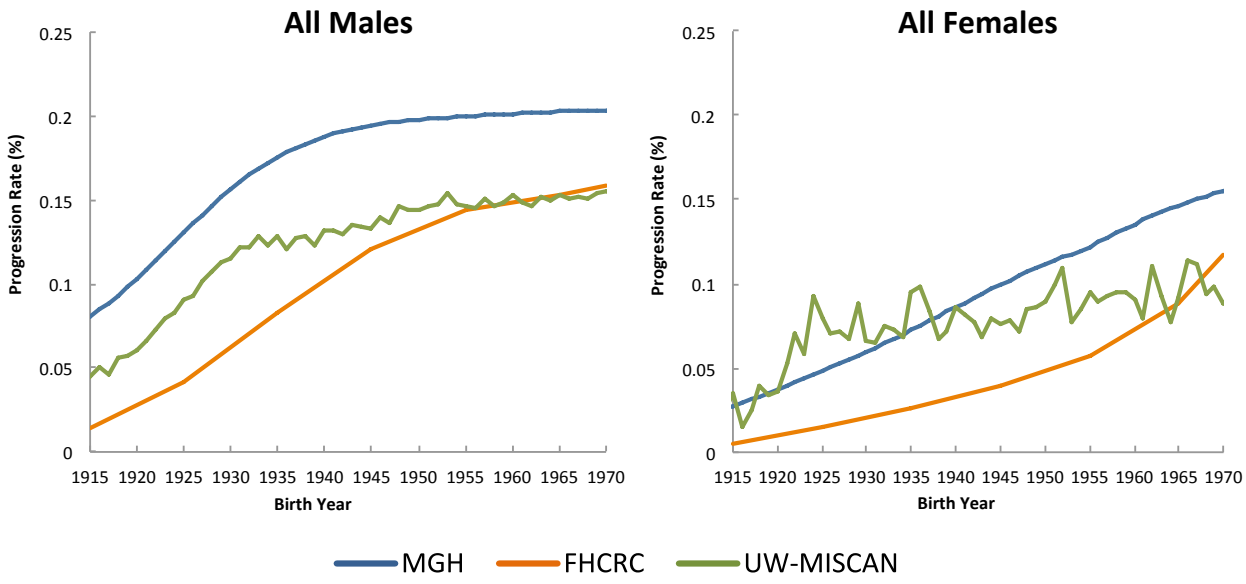

Figure 3.7: **MSCE-EAC progression rate from BE to EAC compared to other models.** Non-dysplastic BE to EAC annual progression rates shown for males (left panel) and females (right panel) by birth year. Progression rates were calculated for age-at-diagnosis with BE at age 60 with 5 years of follow-up time. MSCE-EAC model results denoted by "FHCRC," orange lines.

transformed cell to clinical diagnosis of EAC (ρ event). The capacity of the MSCE-EAC model to predict this important time scale analytically is a strength and will help us to examine the impact of screening and specific interventions in the next two chapters. However, it is difficult to measure in clinical settings since it depends on unobservable events in the disease process, specifically the appearance and progression of preclinical cancer when malignant clones originate from a single progenitor cell and begin growing stochastically at small, undetectable sizes.

The capacity of the models to predict this important time scale is important but the estimates show considerable differences in spite of model calibrations based on the same SEER 9 data (see Fig. 3.8). The other two CISNET models do not similarly model preclinical

cancer beginning with a single malignant cell (see [6] for more details about the other two CISNET models' assumptions and structures). The differences between predictions for EAC sojourn times from the three groups can thus be attributed to fundamental differences in model structure and model-specific constraints. Unlike the other two CISNET models, the MSCE-EAC model predicted that the EAC sojourn time decreases with birth year. We expected this result as consequence of modeling malignant growth with an increasing sigmoid function defined by birth cohort (see Eqs. (3.38-3.40)). Furthermore, this analytical result of the MSCE-EAC model structure explains why the progression rate to EAC was predicted to increase with birth cohort, as seen in Fig. 3.7, due to the decreasing sojourn times of malignancies to become clinical EAC. The average sojourn times predicted by the MSCE-EAC model were 10.2 years for males and 13.3 years for females.

3.2.3 Discussion

To summarize the Base Case I analysis, three independent mathematical models were calibrated to U.S. SEER 9 data, specifically the EAC incidence and mortality rates from 1975 to 2010. In this chapter, we developed all methodology and assumptions needed for the three components that constitute the MSCE-EAC cell module. We used likelihood methods to calibrate the MSCE-EAC hazard function to the SEER 9 data and obtained biological parameter estimates provided in Table 3.1. The MGH and UW-MISCAN groups performed this calibration independently also, based on their model constructions. All three models were then used to generate incidence projections until the year 2030. Although the models differ considerably in structure and design, from biologically-based modeling at the cellular level to empirically-based microsimulations of natural histories, the models' projections (excluding sojourn time estimates) are consistent with one another. As mentioned in Results, the longer MSCE-EAC estimates for malignant clone sojourn times are likely a consequence of defining preclinical cancers as clonal lesions of any size, including smaller and earlier malignancies such as intramucosal carcinoma that may be difficult to detect endoscopically.

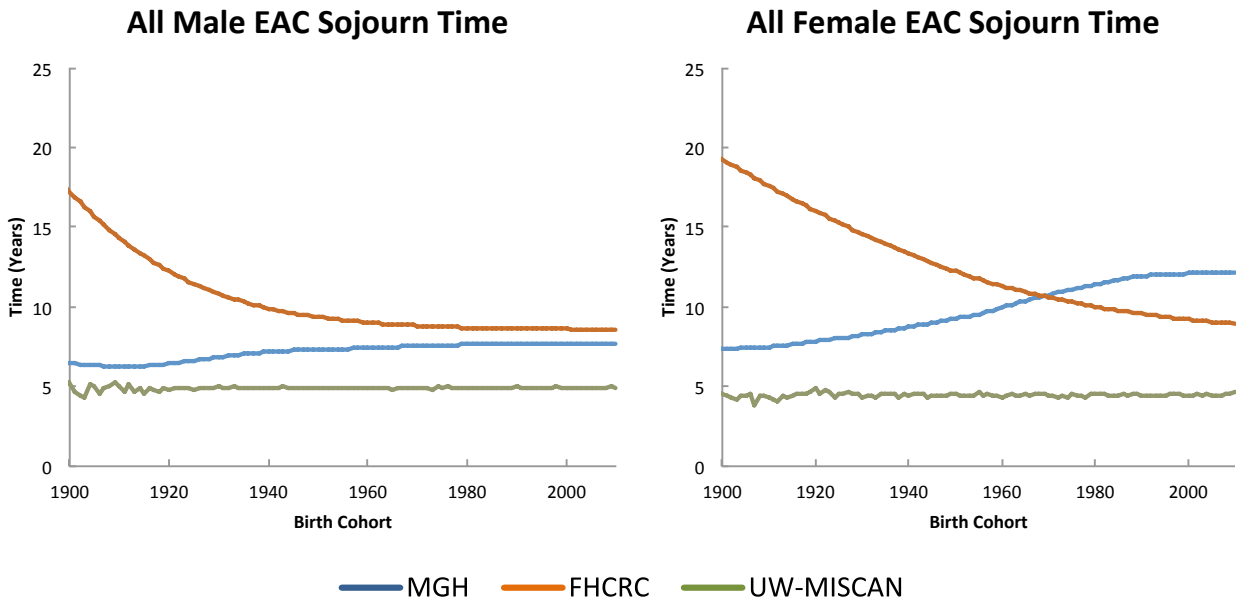

Figure 3.8: **MSCE-EAC mean sojourn times compared to other CISNET models.** Mean EAC tumor sojourn times for males (left) and females (right) of all races, plotted as a function birth cohort for all three CISNET models. The MSCE-EAC model (FHCRC) predicts that diminishing EAC sojourn times by birth cohort causes the increased progression rates for later birth cohorts in Fig. 3.7.

Although results for three outcomes -projected incidence, predicted EAC sojourn times, and progression rates - are provided in this chapter, Kong et. al provides additional comparative modeling results for GERD and BE prevalences, projected mortality rates, and predicted future number of EAC deaths [6].

All three models identify the cancer progression rate (modeled as a birth cohort effect) as an important driver of the observed temporal trends for EAC incidence (Fig. 3.7). Several studies have reported the progression rates of patients with non-dysplastic BE to EAC. Hvid-Jensen et al. reported one of the lower rates of progression, 0.12% per person year to EAC among patients with non-dysplastic BE and LGD [95]. A meta-analysis by Desai et al. reported a progression rate of 0.33% per person year from BE to EAC when only higher

quality, more recent studies were included [96]. A frequently cited progression rate of 0.5% per person year was reported from a meta-analysis that attempted to adjust, using a funnel plot, for publication bias [97]. These reported progression rates range between 0.1% and 0.5% per person year and are consistent with our model estimates for younger cohorts.

Our study has several limitations. First, there are other known risk factors associated with EAC besides those modeled here. For the MSCE-EAC model results, we opted to model the trend in EAC incidence and mortality by varying the biological parameters using a generalization of the age-period-cohort (APC) formalism (the other two CISNET models used similar formulations for risk factors). Thus, varying the transition rates as a proxy for changing risk factor trends did not allow us to investigate the etiology of EAC or to develop a cancer prevention and control policy to reduce the risk of developing EAC. However, the identification of cancer progression as an important driver of EAC trends may provide a focus for future investigation and possible interventions. We have developed a computational framework which allows us to update the analysis as additional clinical evidence on the key risk factor(s) and their impact on EAC emerges. Second, all of our models depict the biological progression following a specific sequence: healthy, absence or presence of GERD symptoms, BE without dysplasia, BE with dysplasia, preclinical cancer, and detected cancer. Although this is the commonly accepted paradigm for EAC carcinogenesis, all EACs may not follow this prescribed sequence or alternative pathways may exist within this paradigm adding heterogeneity which the models do not capture [98, 99].

Despite the limitations, our study has several strengths. We present incidence and mortality projections to 2030 using comparative modeling to investigate EAC trends in multiple birth cohorts. By modeling multiple cohorts separately, our models comprehensively capture incidence and mortality rates with changing age structures as different birth cohorts age. Our comparative modeling approach compares and contrasts the results from independently developed simulation and likelihood-based models using common calibration targets. The approach to resolve differences in model outputs is one of the benefits of comparative modeling, which has been used in other CISNET comparative modeling analyses. When

differences are found, it provides the opportunity to pinpoint the source, which could be a result of error(s) in the model(s), or a consequence of fundamental lapses in knowledge surrounding the natural history of the disease. This approach provides a check for model validity but also the opportunity to identify and discuss gaps in knowledge. This iterative and in many aspects integrative process is perhaps the major strength of comparative modeling and the CISNET consortium, where preliminary model results can be discussed in an open and non-threatening environment conducive to model enhancements that improve risk predictions and therefore credibility.

A statistical analysis on EAC incidence between years 1973-2006 reported that the overall incidence may be plateauing in recent years [91]. A subsequent analysis which included three additional years of incidence data found that the EAC incidence rate continues to increase, although at a decelerating rate [5]. Our projections using three comprehensive computational cancer models also suggest that the incidence and mortality rates of EAC will continue to increase; however, the rate of increase appears to be slowing down for the younger male cohorts. The specific causes of the historical increase in EAC incidence and mortality remain unclear. However, our modeling of potential drivers behind the increasing incidence and mortality trends implicates an enhanced BE-to-EAC progression as significant factor. The future projected increases in cumulative EAC deaths and incidence reflect a significant concern and burden to society. Our findings highlight the importance of public health and cancer control planning with potential interventions to curtail projected EAC morbidity and mortality. The next two chapters will focus on this motive. In Chapter 4, we develop the methodology for the *MSCE-EAC screening model* with some applications and then we present further CISNET comparative modeling results in Chapter 5 with the "Base Case II" analysis focused on the efficacy of surveillance and impact of endoscopic eradication of BE.

Chapter 4

Screening for Neoplasia in Barrett's Esophagus

In the previous chapter, we focused on modeling the cellular events that occur during esophageal adenocarcinoma (EAC) progression and connected the multistage clonal expansion for EAC (MSCE-EAC) model to the population level in order to capture and project EAC incidence in the US population. In this chapter, we present an extension of this method that includes BE patient screening (before symptomatic, incident cancer occurs) that requires another level of detail on the tissue scale due to the specific spatial nature of current endoscopic screening methods (see [14] for the full publication). Screening is targeted toward identifying BE patients who are at the highest risk of developing neoplasia (dysplasia and cancer). Although the risk of BE progressing to EAC is estimated to be low (around 0.2-0.5% per year [12], see Fig. 3.7), clinical evidence suggests that the risk of neoplastic progression in BE varies significantly between individuals depending on age, gender, race/ethnicity, gastroesophageal reflux disease (GERD) and whether or not dysplasia is present in BE.

High grade dysplasia (HGD) occurring in Barrett's esophagus (BE) is generally non-invasive but carries a high risk of progression to esophageal adenocarcinoma (EAC). Low grade dysplasia (LGD) also occurs, but its clinical relevance is less certain. Most patients diagnosed with HGD undergo endoscopic mucosal resection (EMR) or treatment with ra-

diofrequency ablation (RFA) to remove HGD tissue and, in the case of RFA, to reduce the amount of underlying metaplastic BE tissue. Genetic and genomic studies, including longitudinal studies with multiple BE tissue samples from individual patients in the Seattle BE cohort [100], also implicate specific genomic alterations in the neoplastic progression process. Frequently observed alterations in BE include epigenetic silencing or loss of heterozygosity (LOH) of the *P16INK4A* and/or *TP53* tumor suppressor genes [101–104]. Whether these alterations necessarily lead to the clinical presentation of dysplasia and other cellular and architectural changes associated with this diagnosis is presently unknown. However, our working hypothesis with the MSCE-EAC model is that fields of HGD are comprised of clonal populations of premalignant P cells that originate from distinct progenitors in the BE tissue (see Fig. 3.1).

Because dysplasia (in particular HGD) continues to be a widely used clinical predictor for progression to EAC, most BE patients are recommended to undergo periodic endoscopic surveillance with biopsies taken at specified spatial locations in BE to detect neoplastic changes (dysplasia and/or cancer). However, due to the expected large number of adults with BE in the general population ($\sim 1-3\%$ [83, 105]), excessive or ineffective BE screening and surveillance that do not significantly reduce EAC incidence and mortality are a considerable public health concern.

To examine these issues, we propose a computational method called the *multistage clonal expansion for EAC (MSCE-EAC) screening model* that is used for screening BE patients *in silico* to evaluate the effects of biopsy sampling, diagnostic sensitivity, and treatment on disease burden. We developed this mathematical and computational framework to concurrently model the BE-to-EAC progression and endoscopic screening for dysplasia and preclinical cancer prior to asymptomatic diagnosis of EAC. We will present the screening model as three cohesive modules. First, we include the MSCE-EAC cell module described in the previous chapter, which captures key events of the random, GERD-dependent onset of BE, the initiation and stochastic growth of premalignant clones, malignant transformations in premalignant clones, and stochastic growth of malignant clones prior to (symptomatic) cancer

detection. Again, this framework provides a bridge between the cell and population scales and is described in Chapter 3.

The second module is an explicit computational method to efficiently simulate the entire cell model in an individual BE patient until the time of a hypothetical screen. This requires computation of the joint size distribution of premalignant and malignant clones in the BE tissue prior to development of an incident, symptomatic cancer. The method we describe next captures the clonal progression of an idealized, 2D *in silico* tissue composed of intestinal crypts and can generate a variety of spatial patterns (from circular to very diffuse shapes) of both premalignant and malignant clones within the BE segment of a patient.

The third module simulates an endoscopic screen of a patient's BE segment. For a biopsy-based screen, the model mimics the Seattle standard protocol for screening patients with BE, probing the tissue every 1-2 cm with 4 quadrant biopsies for the presence of dysplasia and signs of invasive cancer. We show that the efficacy of this protocol is highly variable and dependent on the sensitivity of detecting neoplastic abnormalities within a biopsy. This sensitivity also affects the number of dysplastic patients predicted to harbor undetected malignancy at the time of screening. The outcomes of the biopsy-based screens are then compared with the model's prediction for screening outcomes when using high-resolution optical coherence tomography (OCT), a new screening technology not yet widely in use. With information about the expected amount of small neoplasms that go undetected during biopsy-based screening, the model explores the potential advantages that image-based screening might offer. Finally, this module simulates ablative treatment of BE patients with detected dysplasia during screening. By explicitly modeling the curative effects of ablative treatment, we gain insights into the critical factors that may limit treatment success.

4.1 Methods

Here we describe the modular design of the multistage clonal expansion for EAC (MSCE-EAC) screening model. The first module is the MSCE-EAC cell module described in depth

in the previous chapter and will not be presented again here. The second module explicitly simulates the cell model outcomes and spatial organization of premalignant and malignant lesions in a Barrett's esophagus (BE) segment. Lastly, the third module simulates an endoscopic screen, whether biopsy-based or image-based, at the tissue level to evaluate the number of BE patients who are positive for neoplasia. These latter tissue level modules require two new spatial parameters that can be calibrated to reproduce published screening prevalences of high grade dysplasia (HGD), and then used to predict other outcomes at the tissue and population levels. Figure 4.1 depicts each level of detail in our multiscale model for screening an individual BE patient: cell, tissue, and organ.

4.1.1 MSCE-EAC Tissue Module

Our previous work with the stochastic, cell-level MSCE-EAC model did not address the expected number and sizes of independent focal lesions of each type in a patient's BE segment at any given age in his/her lifetime, as depicted in Fig. 3.1. However, this knowledge is clinically relevant for effectively monitoring progression to EAC in a BE patient. In this module, we first describe the computational tool developed to obtain stochastic realizations of the number and sizes of premalignant and malignant lesions in a BE patient at any given age. Next, we use these model-derived outcomes of stem cell populations to simulate their spatial configuration as clonal lesions in the BE tissue, which is important given the spatial nature of the biopsy screening protocols.

Because the mathematical complexity of this multistage model makes it difficult to derive tractable analytic size distributions for all cell types through time we resort to direct simulations to track clone number and sizes through an individual's lifetime. Recent advances in stochastic simulation allow further efficiency in computation of cell counts, enabling rapid model testing and examination of many possible scenarios. See Appendix B for the full algorithm and implementation of the MSCE-EAC hybrid simulation of the number of clones and their sizes for all cell types present at time t_s during a hypothetical screening. We call

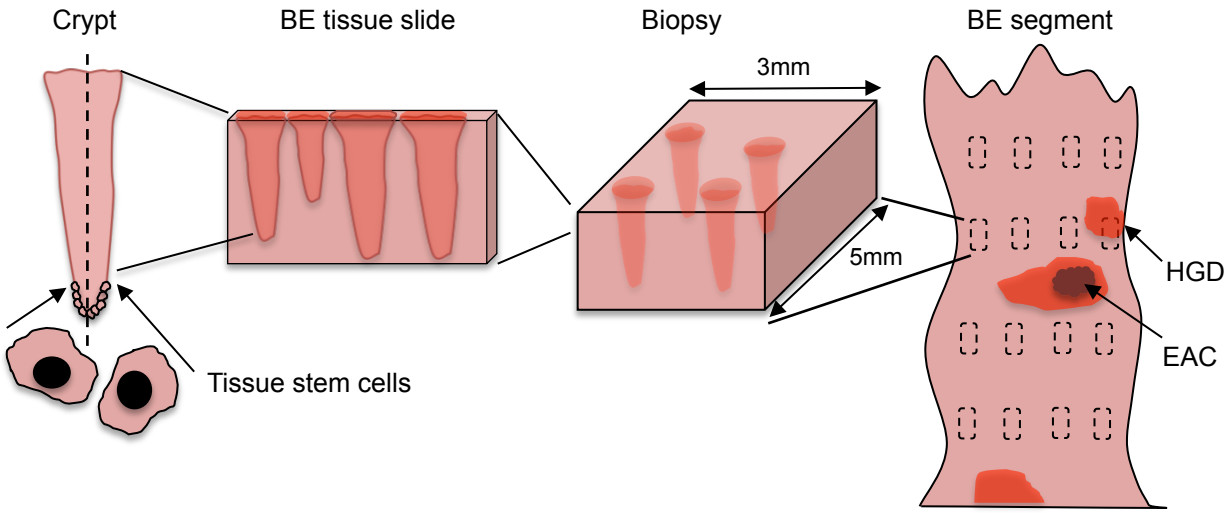

Figure 4.1: **The multiscale nature of BE screening.** Biopsy screening for BE scales from stem cells in the crypt (left) to the BE cylindrical segment of the esophagus depicted (right) with rectangles representing biopsy samples taken during endoscopy via the Seattle biopsy protocol. The BE segment may have dysplasia and/or malignant tissue patches that remain untouched after biopsying. During histological preparation, portions of each biopsy are sliced by microtome and placed on slides for pathologic assessment. Diagnosis is made by microscopic interpretation of crypt and cellular architecture, reflecting the most severe tissue grade found on the slides.

this a 'hybrid' simulation because it employs stochastic simulation when necessary but also makes use of samples from analytical distributions when possible. For the simulation of premalignant (dysplastic) clones, we employ two methods. The first is an exact method, the stochastic simulation algorithm (SSA), first described by Gillespie [106], that simulates every jump in cell count and exponential waiting times between events. The second is a highly efficient approximation to SSA called τ-leaping. Appendix B explains these two methods and describes how the MSCE-EAC simulation uses them cooperatively in a highly efficient approach. The accuracy of both the size distributions generated by the SSA and the τ-leaping method are shown in Fig. B.1 as Q-Q plots for the size distributions of non-extinct premalignant clones compared to the analytical distribution for an independent birth-death-mutation (b-d-m) process.

With cell module parameters as input, the MSCE-EAC hybrid algorithm simulates the multi-type branching process for an individual's cellular progression from birth until time (age) t_s, which can be repeated to generate (synthetic) data for a sample population. In summary, for those individuals who are found to have BE when screened, each patient has a specific number of BE stem cells (X), number of pre-initiated cells (P^*), a number of non-extinct premalignant (P) clones with respective sizes, a number of non-extinct malignant (M) clones with respective sizes and information about the parental P clones from which the M clones originated, and lastly whether the patient is a prevalent, clinical EAC case by time t_s. Note, the stochastic model allows the possibility that the ancestor premalignant clone may go extinct while the malignant clone is still growing at the time of screening t_s. Figure 4.2A shows the random trajectories for a simulated BE patient's clones obtained via this algorithm for the five years of life prior to initial screening at age $t_s = 60$.

Crypt-structured spatial modeling of neoplasia

Thus far, the MSCE-EAC tissue module has employed the cell-level rate parameters, described and fitted to population data in the MSCE-EAC cell module (Chapter 2), to obtain realizations of the multi-type branching process for a BE patient until time of screening. In

this section, we translate the size, in terms of the number of stem cells, in each neoplastic lesion simulated by the hybrid algorithm to the actual 3D size of the neoplasm within the BE segment, which is important for predicting what is detected during an endoscopic screen. This translation requires us to consider the known spatial characteristics of BE tissue biology. The metaplasia that defines BE is made up of a monolayer of epithelial cells that form crypt-like depressions in the underlying connective tissue of the lamina propria that are considered the basic functional unit of BE tissue [107]. These units (or *crypts*) are renewed through asymmetric stem cell divisions that produce differentiated cells, which eventually die or are sloughed off the surface epithelium. Putative stem cells found at or near the crypt base appear to escape this flow. Thus, under homeostatic conditions, the stem cell populations are assumed to remain immortal in the tissue before initiation of dysplasia. However, once a stem cell is initiated (which requires two-rate limiting events in the present form of the MSCE-EAC model), it may undergo stochastic clonal expansion through symmetric divisions that produce two identical initiated stem cells or two cells committed to differentiation, or it may undergo cell death (apoptosis).

Although the mechanisms regulating neoplastic growth across crypt structured tissues are not well understood [108], it is generally assumed that initiated stem cells either spill over and invade neighboring crypts ("top-down" hypothesis) or that bifurcation of crypts may occur in a process called *crypt fission* that leads to clonal expansion at the crypt level ("bottom-up" hypothesis) [108]. Recent studies favor clonal expansion by crypt fission as the dominant mode of generating mutated crypts in the human colon that form adenomas [108]. Our computational model includes the assumption that crypt fission drives the spatial growth of neoplastic lesions in BE.

Due to this spatial structure of crypt openings at the epithelial surface, we can represent the neoplastic clones on a 2D plane even though the crypts in BE constitute a 3D tissue as depicted in Fig. 4.1. We introduce a single adjustable parameter for stem cell area density, σ, which equals the number of stem cells per mm^2. In relation to the cryptal organization of the BE segment, this area density can be defined as $\sigma = c_\sigma \cdot k$, where c_σ is the number of crypts

per mm^2 and k is the number of BE stem cells maintaining each crypt. Although σ is the parameter we will vary in our model, we introduce this representation because experimental estimates are reported in terms of crypts and stem cells in each crypt. Estimates for the crypt density c_σ in BE epithelium appear to show large inter-individual variation. For example, a recent study following a cohort of BE patients found c_σ to range from 3 to 100 (10,000 - 400,000 crypts per $\sim 38cm^2$) [109]. The cause and significance of this variation remains unclear.

Similarly, the stem cell population size k maintaining the crypt's stem cell niche is also highly uncertain according to estimates provided in the literature on epithelial crypts. For colonic crypts, Nicolas and colleagues estimated there are between 8 and 20 stem cells per crypt [110], yet earlier methylation studies reported $\sim$ 64 per crypt [111]. There is also evidence that stem cell numbers may be significantly higher in dysplastic crypts in the colon [112]. Combined, these estimates of c_σ and k provide us with an estimated range of $\sigma \in [24, 6400]$. Because of the high experimental uncertainty for the parameter σ, we allow this to be a variable parameter in our model (see Results for our calibration of σ).

To generate the spatial representation of the clones, each premalignant clone is randomly placed on a hexagonal grid representing the crypt-structured BE tissue. While the overall area occupied by a clone is controlled by the number of stem cells it contains and the stem cell density σ, the spatial appearance of the clones is independently controlled by a diffusivity parameter γ. Briefly, clones are compact when $\gamma > 0$, becoming nearly circular when γ approaches 1. Alternatively, clones are made increasingly more diffuse and branching when $\gamma < 0$. Like σ, the spatial irregularity of neoplasm growth in BE tissue is not well known. Thus, like σ, we will consider γ as a second spatial parameter in our model.

More specifically, we model the probability that crypt j will be the next premalignant occupied area on a grid during crypt fission as a clone grows to the predetermined size

produced by the MSCE-EAC simulation, beginning with a single progenitor P cell, to be

$$p_j(t+1) = \Pr[\text{crypt } j \text{ will be occupied as new daughter neoplastic crypt at time } t+1]$$
$$= \left(\frac{\sum_{i=1}^{6} \mathcal{N}_j(t)}{6} \right)^{\gamma}, \tag{4.1}$$

where $\mathcal{N}_j(t)$ is a list that contains the states (vacant = 0, occupied= 1) of all the neighboring crypts for crypt j at time t. With a total of 6 neighbors on a hexagonal grid, the quantity $\frac{\sum_{i=1}^{m} \mathcal{N}_j(t)}{6}$ equals the mean neighborhood occupation. The simulation generates clones that are compact and nearly circular when $\gamma > 0$ because neighbor probabilities for the potential occupant are exponentiated by a positive number γ; neoplastic crypts with the lowest number of neoplastic neighbors will be selected preferentially to undergo crypt fission in the simulation. When $\gamma < 0$, the clones will grow in an increasingly more diffusive manner since there is a high relative probability, with a maximum proportional to $\left(\frac{1}{6}\right)^{\gamma}$, that a neoplastic crypt with few neighbors, say, on a "branch" of a clone, is chosen to undergo crypt fission.

Under our model, any premalignant ancestor cell may generate a neoplastic clone including a simulated number of premalignant cells and possibly an embedded population of independently expanding malignant cells that originate from malignant transformation(s) with rate μ_2. To accommodate malignant tumor growth, we assume that malignant crypts simply displace the pre-existing premalignant crypts, a circumstance that is often seen in esophagectomy specimens [113].

In summary, for each BE patient, the MSCE-EAC tissue module first obtains the number and sizes of both premalignant and malignant clones (at time t_s) randomly generated by the MSCE-EAC hybrid simulation and then translates these numbers to a spatial configuration of neoplasms within a patient's BE segment. To illustrate this, we use the simulated patient's cellular information provided from the MSCE-EAC simulation shown in Fig. 4.2A, and show in Fig. 4.2B the corresponding 'rolled out' BE cylindrical segment generated with the spatial representation of the clones as described above, using the choices of $\sigma = 3300$ stem cells/mm^2 and $\gamma = -2$. Simulated premalignant and malignant lesions are depicted in dark pink and red, respectively. The black, dotted line rectangles represent the location of biopsies taken

from the tissue under a certain screening protocol that we will explain in the following section.

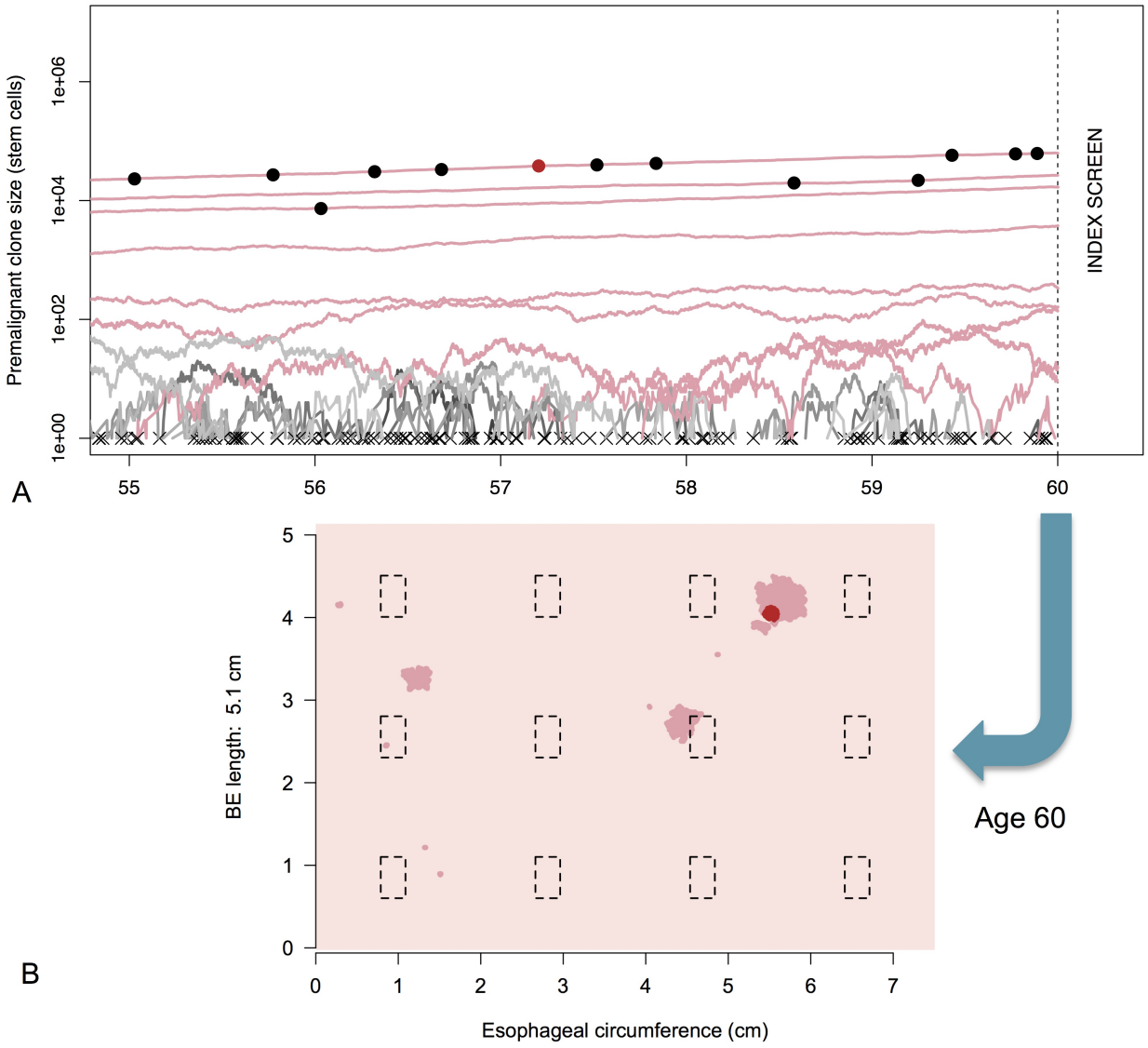

Figure 4.2: **MSCE-EAC tissue module.** (Continued on the following page.)

Figure 4.2: **MSCE-EAC tissue module.** A) Premalignant (dysplastic) stochastic growth trajectories are depicted for a sample male BE patient from the 1930 birth cohort, beginning five years before time of initial screen at age 60. Most dysplastic cell progenies from a single initiated cell went extinct before examination by endoscopy (trajectories depicted by gray lines and an 'x' at time of extinction). The dysplastic clone trajectories that did not go extinct before index screen are shown in dark pink with final sizes intersecting the vertical dotted line at age 60. The dots on the trajectory lines correspond to times of asymmetric division of a dysplastic cell to produce one dysplastic cell and one malignant cell. Malignant transformations produce both clones that quickly went extinct before age 60 (black dots) and non-extinct malignancies (red dot). B) BE segment at time of screening. Same simulated male patient from (A) with BE of length 5.1 cm. Dysplastic clones (dark pink), malignant clone (red), and biopsies (black dashed rectangles) are pictured at time of biopsy-based screening, age 60. Clone diffusivity parameter is $\gamma = -2$, stem cell density parameter is $\sigma = 3300$ stem cells/mm^2.

4.1.2 MSCE-EAC Screening Module

The MSCE-EAC tissue module computes the number of stem cells in each neoplastic clone and generates the shapes of these clones within a BE segment at any given age of a patient. The MSCE-EAC screening module takes this information and performs an endoscopic screen on this simulated BE segment. Here we outline the methodology for generating model predictions related to three specific screening outcomes: (1) the probability that small cancers are missed in HGD patients during biopsy-based screening, (2) the potential gains in neoplasia detection probabilities if screening occurred via high-resolution tomographic imaging, and (3) the efficacy of ablative treatments that result in the curative depletion of metaplastic and neoplastic cell populations in BE in terms of the long-term impact on reducing EAC incidence. These model predictions are described in Results.

Biopsy screens and diagnostic sensitivity

In 1998 the American College of Gastroenterology (ACG) recommended the use of a systematic sampling method known as the Seattle biopsy protocol rather than random biopsy sampling of BE tissue during endoscopic screening. This systematic biopsy protocol specifies 4 jumbo quadrant biopsies ($\sim 15\text{mm}^2$ of tissue each) every 1-2 cm of the BE length to achieve increased sensitivity for detection of dysplasia [114–116]. However, a multi-center study from 34 US states on BE surveillance found that adherence to guidelines was only seen in 51.2% of cases [117]. The authors also found that longer segment BE, which is more time-consuming to biopsy, was significantly associated with reduced adherence even though risk of EAC is considered proportional to the length of the BE segment. In the examples described here, we model endoscopic screening conforming with the Seattle protocol, although the methods may be applied to any protocol of biopsy placement and/or biopsy size during upper endoscopy. Therefore, the model may be utilized for comparing simulated efficacies of different screening protocols. As we will see in the scenarios that follow, for an average esophageal circumference of 75 mm, even the rigorous Seattle biopsy protocol results in large sampling errors due to the fact that biopsies only sample $4 - 6\%$ of the total BE mucosal surface for pathologic assessment.

Nearly half of BE surveillance endoscopies in the US are not performed in adherence to the Seattle biopsy protocol [117] because less tissue is actually sampled than prescribed due to use of smaller forceps or fewer biopsies taken. Specimens for histology are typically fixed in formalin, embedded in paraffin, and tissue profiles are cut at 4-5 μm and stained with haematoxylin and eosin (H&E) [113], but only few histological slides are actually examined for neoplastic changes per biopsy and the precise method of examining specimens is usually not described in publications. All of the aforementioned factors may account for the large range of estimates found for prevalences of dysplasia and/or intramucosal cancer [97, 113, 118–121].

To account for different biopsy protocols, incompletely described histological methods,

and interobserver variation of neoplasia grade, we present results from the computational model for different levels of diagnostic sensitivity based on the minimum number of neoplastic (premalignant/malignant) crypts within a simulated biopsy specimen required for pathologic diagnosis of dysplasia/malignancy among BE patients without prior diagnosis of EAC. As depicted in the simulated BE segment of Fig. 4.2B, some biopsy specimens taken via the Seattle protocol contain varying numbers of neoplastic crypts while missing sizable neoplastic areas or even entire neoplastic clones. Since the pathologist usually assesses less than 1/10 of the actual specimen after histologic sectioning, there is high probability that, even if neoplastic crypts exist in a biopsy specimen, no neoplastic crypts are contained on a histologic slide after sectioning. As a way to capture this sensitivity, we estimated prevalences of neoplastic tissue for a range of biopsy thresholds, i.e., the minimal proportion of neoplastic stem cells in a single biopsy needed for the neoplasm to be detectable at time t_s. Let n_f be the fraction of neoplastic stem cells present out of the total number of BE cells within a single biopsy that are required for a positive diagnosis of neoplasia. Recall that the spatial parameter σ defines the relationship between numbers of cells in each clone and geometric clone size. As an example, if $n_f = 1/2$, then a simulated patient must have at least one biopsy (among all of the biopsies obtained via the Seattle protocol) that contains at least $\frac{15}{2}\sigma$ neoplastic cells (50% of stem cells in a single 15mm^2 biopsy) for a positive diagnosis of neoplasia. Therefore, the *biopsy sensitivity* of a screen simulated by the MSCE-EAC screening model is interpreted as $(1 - n_f) \cdot 100\%$.

For simulations following the Seattle biopsy protocol, the BE segment from the MSCE-EAC tissue module can be visualized as being partitioned into identical rectangular sections, which we will call "biopsy quadrants" with a single biopsy in the center of each quadrant. For example, an average BE segment of length 5 cm and 7.5 cm circumference will have 12 biopsy quadrants, 3 levels of length 5/3 cm with 4 quadrant biopsies each. Furthermore, we assume periodic boundary conditions when placing clones in a random quadrant, as depicted in Fig. 4.3, to simulate how many neoplastic crypts will be located within a biopsy sample. The Results section shows how the spatial characteristics (size and shape) of the clones can

influence the observed prevalence of a neoplasm in a screened population for a range of biopsy sensitivities.

High-resolution imaging screens

Beyond different biopsy-based protocols, the screening module within the MSCE-EAC screening model allows the user to also choose to simulate any screen using high-resolution imaging within the esophagus. The model can perform an optical coherence tomography (OCT) screen in which a positive detection of HGD and/or malignancy occurs if the geometric size of a clone on image is greater than a resolution area threshold, a_{OCT}. Thus, we no longer need to consider neoplastic proportion thresholds as introduced previously for biopsy-based screening, but rather consider a fixed area or caliber threshold of any imaged clone. The results of an imaged-based screen will again depend on parameter σ to determine if the geometric size of a clone is above or below an imaging resolution.

Ablative treatment and impact on EAC incidence

After a simulated screen of BE patients for detection of dysplasia and preclinical EAC at age t_s, the MSCE-EAC screening module can be used to further simulate an intervention such as an ablative treatment using radio frequency. To replicate current practice with radiofrequency ablation (RFA), we simulate RFA treatment on positively screened, non-EAC patients with dysplasia. The MSCE-EAC screening module then projects the EAC incidence, which depends on the stochastic number of stem cells of different types in a BE patient at the time of screening, into the future after an ablative treatment. Ablation is assumed to curatively reduce all clonal populations and the number of BE crypts by certain percentages. As a simple example, we consider the model's EAC incidence predictions after a single ablative treatment when indicated by the presence of high grade dysplasia on future EAC incidence.

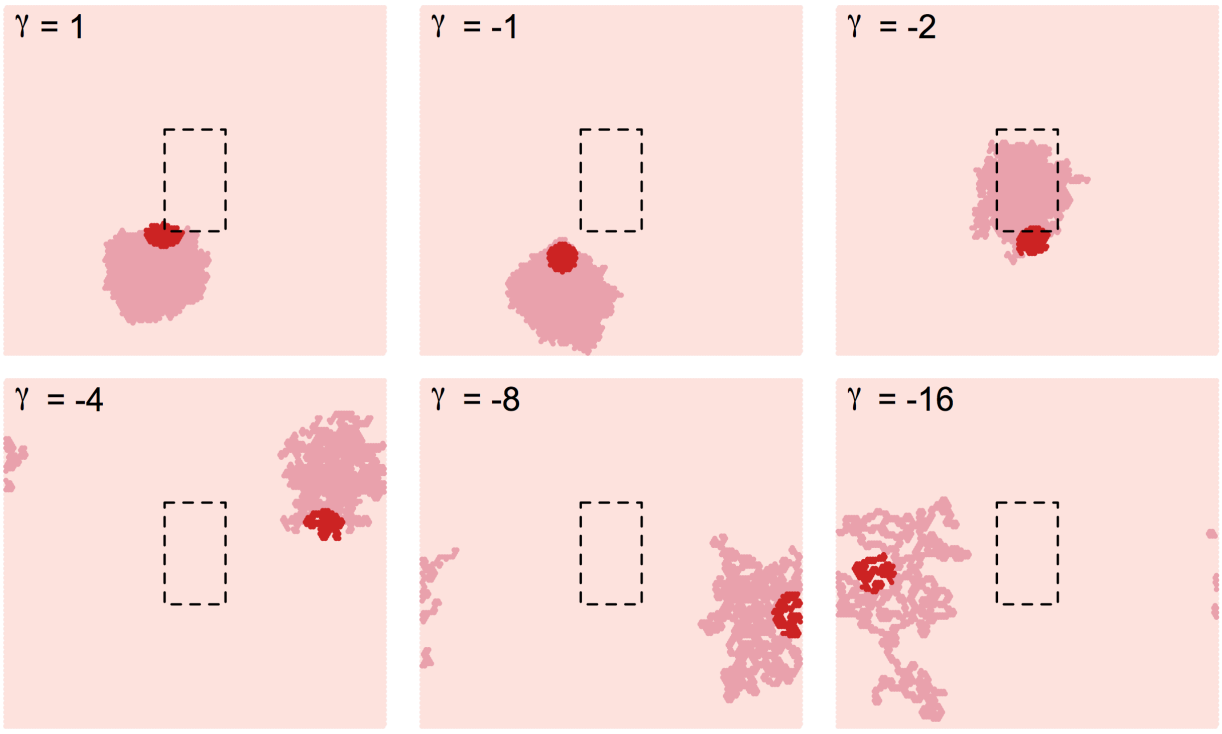

Figure 4.3: **Spatially diffuse clone growth on the BE segment.** Six examples of diffusive clone growth are depicted on a light pink single BE biopsy quadrant for a patient with 5.1 cm BE length. Black dashed rectangles are the jumbo biopsy outlines, dark pink areas are premalignant (dysplastic) clones and red areas are malignant clones. The sample premalignant clone that contains a malignancy is from the realization in Fig. 4.2, containing 62,771 dysplastic stem cells and 5,176 malignant stem cells. Together, this mass of crypts is randomly placed on one of 12 identical biopsy quadrants on this BE segment under the Seattle biopsy protocol with assumption of $\sigma = 3300$ stem cells/mm^2. Values for γ diffusivity parameter of clone growth are shown in top left of each quadrant.

For all times $t > 0$, we can compute the cumulative hazard $\Lambda_{EAC}(t)$,

$$\Lambda_{EAC}(t) = -\ln(S_{EAC}(t)) = -\ln\left(1 - \int_0^t f_{EAC}(s)\,\mathrm{d}s\right) \tag{4.2}$$

For the initial scenario of screening all individuals at time t_s, we derived the MSCE-EAC cumulative hazard function that includes contributions from the subpopulation of individuals found to have BE at time t_s who, immediately following HGD diagnosis, receive treatment at time t_s; and the subpopulation without BE. For any time $t > t_s$ and BE cumulative distribution F_{BE} given in Eq. (3.3), we compute $f_{EAC}(s)$ as follows

$$f_{EAC}(s) = f_{EAC}(s|T_{BE} \leq t_s) \cdot \Pr[T_{BE} \leq t_s] + f_{EAC}(s|T_{BE} > t_s) \cdot \Pr[T_{BE} > t_s]. \tag{4.3}$$

For the screened BE population we follow the method of Jeon et al. [122] to simulate the four possible types of cells present in a patient at screening time t_s^- (where the minus superscript denotes cell populations present prior to any intervention): X = number of BE stem cells in the BE segment, $P^*(t_s^-)$ = number of preinitiated P^* cells, $P(t_s^-)$ = number of initiated, dysplastic P cells (all clones combined), $M(t_s^-)$ = number of malignant, preclinical cancer cells (all clones combined). The MSCE-EAC tissue module simulates realizations of these random variables for each patient up to the instance of screening t_s^-, before intervention occurs. After simulating n independent and identically distributed realizations of individuals (by gender) and performing the Seattle biopsy screening protocol *in silico* on those with BE as described earlier, the screening module provides the vector $A_i = \{X_i, P_i^*(t_s^-), P_i(t_s^-), M_i(t_s^-)\}$ for each patient i with BE, $i = 1, .., n$.

Next, RFA intervention may be simulated for patients diagnosed with dysplasia by introducing the following *post-ablation proportion* vector, $\omega = \{\omega_X, \omega_{P^*}, \omega_P, \omega_M\}$, that describes the cell-type specific depletion of BE tissue. For example, to simulate a perfect ablation of all lesions in BE and of BE metaplasia itself, we set $\omega = \{0, 0, 0, 0\}$. For those patients who have a positive screen (showing HGD), we simulate an RFA treatment by adjusting the patient's (simulated) cell count vector A_i through component-wise multiplication by ω. Thus, the post-RFA numbers of cells in each stage of the MSCE process immediately after

screening and treatment (denoted by time t_s^+) are given by the adjusted cell type vector $\hat{A}_i$

$$\hat{A}_i \equiv \omega \circ A_i = \{\omega_{_X} \cdot X_i, \omega_{_{P^*}} \cdot P_i^*(t_s^-), \omega_{_P} \cdot P_i(t_s^-), \omega_{_M} \cdot M_i(t_s^-)\} \tag{4.4}$$

$$= \{X_i(t_s^+), P_i^*(t_s^+), P_i(t_s^+), M_i(t_s^+)\}. \tag{4.5}$$

BE patients with a negative screen for neoplasia sustain the same (before and after) $A_i \equiv \hat{A}_i$ vector as was computed at time t_s^- since no RFA treatment is performed on these patients. Due to the Markovian renewal property of branching processes, the survival and hazard functions for each screened patient $i = 1, .., n$ for some time $t > t_s$ are computed using the adjusted numbers for each cell type post screen. These survival and hazard functions for the 4-stage MVK model after BE onset are easily computed using the Kolmogorov backward equations for the stochastic multistage process. See Appendix B for the full derivation of all individual contributions to $\Lambda_{_{EAC}}(t)$ defined by Eq. (4.2).

4.1.3 Open source code

The methods outlined in this section are implemented by the comprehensive MSCE-EAC screening model consisting of three modules: cell, tissue, and screening. All necessary tools to employ this method, including examples of user inputs used in the upcoming Results, are available in documented R code at `https://github.com/yosoykit/MSCE_EAC_Screening_Model`.

4.2 Results

Clinical studies that assess the efficacy of screening Barrett's esophagus (BE) patients are naturally limited by the amount of BE tissue that can be sampled for histopathological analysis. To gain insights into how this limitation affects screening efficacy we used the described EAC multiscale method to compute the unobserved proportion of precursor lesions and early cancers, quantities of clinical relevance for early cancer detection. As shown here, the multistage clonal expansion for EAC (MSCE-EAC) screening model predicts all lesions in

BE during a patient's lifetime, including their numbers and sizes. The explicit calculation of the numbers of stem cells in these lesions is functionally dependent on the model parameters that, with the exception of the spatial parameters introduced the Methods section, were previously estimated through model calibrations to EAC incidence data in the US [6]. For each simulated BE patient, we performed a pre-specified screening protocol to ascertain a patient's clinical diagnosis, while also retaining concurrent information on any undiagnosed, potentially detectable lesions.

In the following results, we first calibrated the two unknown spatial parameters (neoplastic lesion shape diffusivity parameter γ and stem cell density parameter σ) to achieve consistency with current literature findings on the prevalence of the most important neoplastic precursor, high grade dysplasia (HGD), in BE patients, without changing the parameters that determine the fits to EAC incidence data reported in [6]. Thus fully calibrated, we then applied the MSCE-EAC screening model to predict three important clinical outcomes, including biopsy and imaging diagnostic sensitivities and the impact of ablative treatment on the risk of developing EAC.

4.2.1 Calibration to HGD prevalence data

In the current epidemiological literature, studies beginning with a biopsy-based index screen of BE patients (i.e., the screen when a patient is first diagnosed with BE) provide widely variable estimates of the prevalence of HGD, ranging from $\sim 2.75 - 8.25\%$ [97, 113, 118–121]. To compare this with model-derived predictions, we simulated an index endoscopic screen on a sample population of patients with BE and computed the prevalences of both pre-malignancy (HGD) and screen-detected (non symptomatic) malignancy. For an illustrative example of the MSCE-EAC screening model outputs, we simulated an index endoscopy for all males and females at screening time (age) $t_s = 60$ in the year 1990 (indicative of index screens from prospective studies that estimate the BE to EAC progression rate). With the BE prevalence given in Eq. (3.3), these results focus on expected screening outcomes in the

subpopulation of individuals found with BE (see Methods). Figures B.2-B.3 depict functions $p_{sGERD}(t)$ and BE prevalence, $F_{BE}(t)$, for simulated males and females, respectively.

Because the detection of a neoplastic lesion may involve both premalignant and malignant cells transformed within the lesion, we first consider the (random) sum of the two cell types to determine the efficacy of the biopsy protocol to detect a neoplastic lesion in BE. The biopsy sensitivity was varied from 10% to 95%, as seen in Figs. 4.4-4.5, to allow for systematic exploration of sensitivity effects (see Methods). If a neoplasm is detected on a biopsy, we doubled the biopsy sensitivity for malignant content because the biopsy is likely under closer inspection by the pathologist.

Parameter estimation and sensitivity

To compute the prevalence of HGD without concurrent malignancy, we adjusted the BE population denominator to not include those with screen-detected cancer on index endoscopy. We then computed the probability of detecting HGD for a range of values for σ and γ. We found that when σ is assumed to be in the range of 2000 to 5000 stem cells per mm^2, the model produces HGD prevalence estimates that are broadly consistent with the range of published estimates for the span of biopsy sensitivities (see Fig. 4.4) [97, 113, 118–121]. In particular, the model estimated an expected range of $2.1\% - 8.7\%$ for HGD prevalence among men and an expected range of $.85\% - 6.1\%$ for HGD prevalence among women when using a density of $\sigma = 3300$ stem cells/mm^2. For each choice of σ, the results were mostly insensitive to the choice of diffusivity parameter γ (see right panel of Fig. 4.4). We present forthcoming results using the cell-level parameters found in Table 3.1, $\sigma = 3300$ ($\sigma \in [2000, 5000]$ shown for sensitivity illustration), and $\gamma = 1$. The model parameters thus reproduce population-level EAC incidence data (obtained through symptomatic detection [6]) and HGD prevalence data (obtained from biopsying BE tissue at specific locations using a standard protocol).

By estimating the same breadth of HGD prevalences as studies have reported [97, 113, 118–121], Fig. 4.4 illustrates our prediction that the predicted HGD detection probability is strongly dependent on n_f, the minimum neoplastic tissue fraction for detection in a single

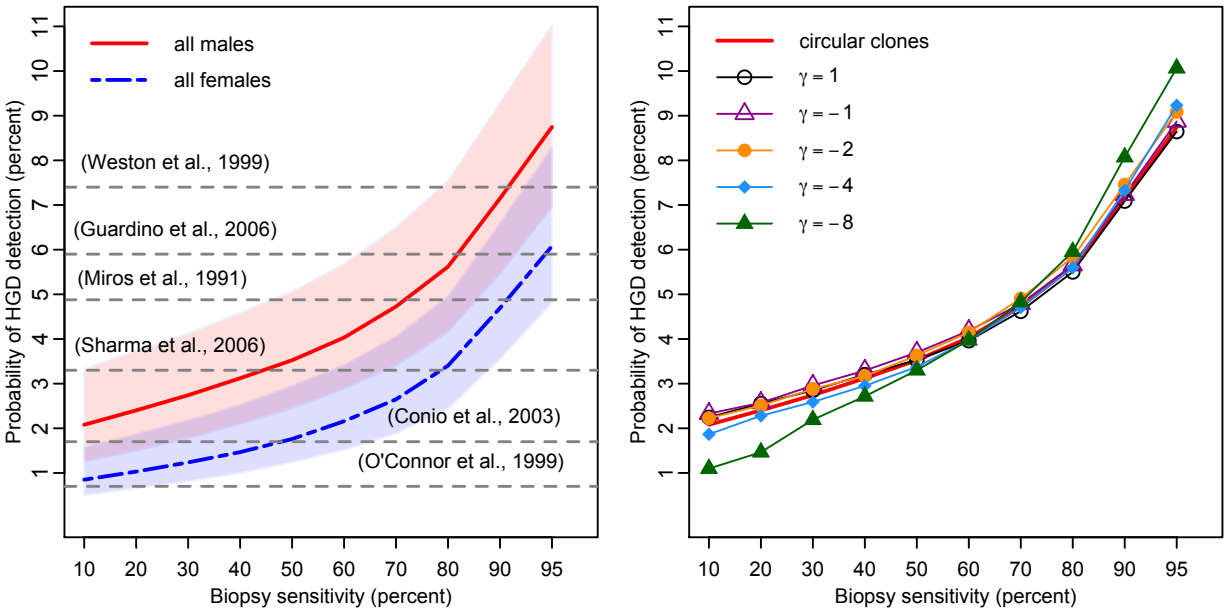

Figure 4.4: **High grade dysplasia prevalences in BE estimated with the MSCE-EAC screening model.** (Left panel) Probability of high grade dysplasia detection among BE patients simulated by the MSCE-EAC screening model for males (red, solid) and females (blue, dash-dotted) at initial screen at age 60 for biopsy sensitivities ranging from $10\% - 95\%$ and assumed density of $\sigma = 3300$ stem cells/mm^2 (shaded regions represent sensitivity of results for $\sigma \in [2000, 5000]$). Since the sensitivity of each study is unknown, some representative literature values for the corresponding probability of HGD detection are depicted as horizontal grey dashed lines at a single percentage level [97, 113, 118–121]. Expected prevalences produced by 100K simulation size of BE patients, shown for males and females. Simulation standard error is less than .001 for all Results. (Right panel) Male HGD prevalences produced by diffusive clone growth on hexagonal BE grid, for diffusivity parameter γ ranging from 1 to -8, versus isotropic, circular clone assumption (red, solid). Solid red curve is the same as that shown for males in left panel. Probabilities of finding HGD for spatially simulated diffusive clone growth are shown for $\sigma = 3300$ stem cells/mm^2 and identical biopsy sensitivities as shown in left panel.

biopsy (see Methods). Specifically, by increasing biopsy sensitivity from 10% to 95%, our model predicted that the prevalence of HGD increases 4-fold. Further, even though most clones remain undetected, the mean number of non-extinct premalignant clones of any size in this cohort was 6.6 per BE patient (See Figs. B.4-B.7). This modeling result highlights the prediction that even with rigorous adherence to the Seattle protocol, the majority of BE patients with concurrent HGD will not be detected nor considered prevalent HGD cases.

4.2.2 Example 1: Missed EAC malignancies in HGD patients

Along with difficulties in first detecting dyplasia present in BE during endoscopic screening, several studies suggest that many BE patients who are diagnosed with HGD without malignancy actually have an undetected cancer that was missed during biopsy screening [123,124]. The MSCE-EAC screening model estimates the probability that a positive HGD patient actually harbors a synchronous, occult malignant clone that is not screen-detected either because it was completely missed in a biopsy sample (e.g. see the small malignancy depicted in Fig. 4.2B) or because it was undetected in a biopsy for a particular biopsy sensitivity, perhaps due to insufficient histologic sectioning. This is an interesting, clinically relevant feature of our modeling. The model predicted the expected fraction of undetected EAC in BE patients diagnosed with HGD to be between $3.2\% - 14.2\%$ for men and $4.3\% - 19.3\%$ for women (see Fig. 4.5). We conclude that the higher probability of missed malignancy in women is due to the lower probability of finding any neoplasia (due to smaller clone sizes, Figs. B.4-B.7) in women during index endoscopy (see Fig. 4.4). These predicted ranges are compared with studies of HGD patients found with concurrent adenocarcinoma, which remained undetected even by rigorous biopsy protocols but are later discovered during resection of the esophagus [125–129]. However, from these esophagectomy studies conducted over the past two decades, the reported prevalence of synchronous malignancy among HGD patients widely varies from $0 - 75\%$. With strict adherence to the Seattle protocol, our model generated a lower estimate of concurrent EAC risk in HGD patients than most published studies, yet it

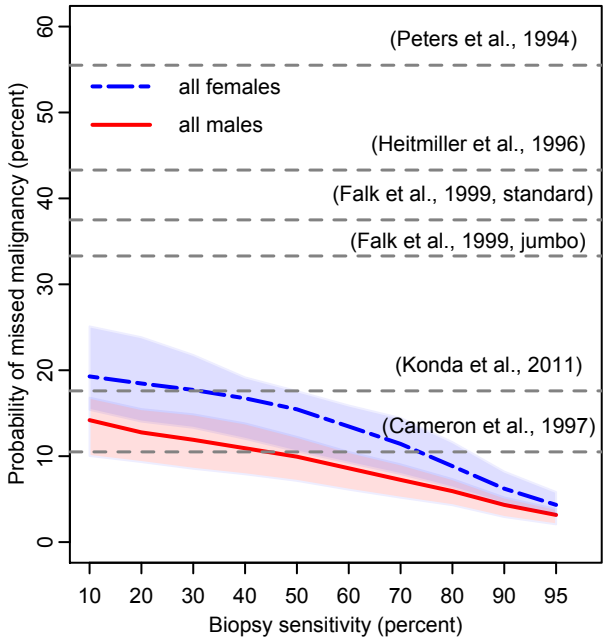

Figure 4.5: **Predicted probability of missed malignancy in positive high grade dysplasia population at index screen.** Percentages of patients diagnosed with HGD during index endoscopy (denominator is population plotted in Fig. 4.4) who concurrently harbored missed, malignant clone(s) present on their BE segments that were not detected on biopsy screen. Since the sensitivity of each study is unknown, literature values for the corresponding probability of missed malignancy are depicted as horizontal grey dashed lines at a single percentage level [125–129]. One publication included percentage of occult malignancy after HGD was diagnosed with either standard or jumbo forcep sizes, as indicated [126]. Expected proportions produced with 100K BE patient simulations each for males (red, solid) and females (blue, dash-dotted) with assumed density of $\sigma = 3300$ stem cells/mm^2 (shaded regions represent sensitivity of results for $\sigma \in [2000, 5000]$).

is consistent with the most recent study by Konda et al. when biopsy sensitivity is low [128]. It is also possible that the studies with high estimates of concurrent malignancy were biased because cancer was suspected in these patients indicating esophagectomy.

4.2.3 Example 2: Predicted HGD prevalence with image-based screening

High-resolution imaging of BE (a technology still in infancy and not yet widely utilized) may provide a benefit through the early detection and endoscopic resection of small premalignant and malignant lesions. The MSCE-EAC screening model can explore the potential quantitative improvements of screening for neoplasia when diagnosed via optical endomicroscopy compared with a less sensitive biopsy protocol.

To this end, we simulated the results from an optical coherence tomography (OCT) screen in which a positive detection of HGD and/or malignancy occurs if the geometric size of a clone on image is greater than a resolution area threshold, a_{OCT} (see Methods). Assuming $a_{OCT} = 1\text{mm}^2$ and the same assumption for stem cell density σ that was used in previous results, the HGD prevalence (excluding incident EAC cases) rose to an expected 27.89% for the BE cohort used in the previous examples (1930 birth year, $t_s = 60$). Therefore, for the range of probabilities of HGD detection shown in Fig. 4.4, the MSCE-EAC screening model estimated an expected 68.7% to 92.8% increase in HGD detection probability using a sensitive imaging technology for screening rather than biopsy-based screening. This modeling exercise reinforces the conclusion that many neoplastic clones of detectable size are being missed with current biopsy protocol screening endoscopies.

4.2.4 Example 3: Predicted EAC incidence after treatment

As a third example demonstrating the utility of the MSCE-EAC screening model, we computed the projected cumulative hazard $\Lambda_{EAC}(t)$ in Eq. (4.2) after a single index screen of BE patients at time $t_s = 60$, removal of screen-detected EAC patients, and subsequent RFA

treatment of HGD positive patients. We explored RFA efficacy under various assumptions about the impact of ablation on cell counts, as specified by the ablation proportion vector ω (see Methods). When comparing to the background incidence (in which no screening occurs), we predicted the effect on EAC cumulative incidence based on a range of RFA effectiveness assumptions (See Fig. 4.6). If patients that were positively detected with HGD at index screen (6% with 60% biopsy sensitivity) receive RFA, the MSCE-EAC screening model predicted that by year 2030, expected EAC cumulative hazard will be reduced by 17.1% if 50% of all BE cell types are effectively removed ($\omega = \{.5, .5, .5, .5\}$) and be reduced by 32.1% if 99% of all BE cell types are effectively removed ($\omega = \{.01, .01, .01, .01\}$). To explore the future influence of missed malignancies, the model predicted that if RFA removed all malignancies ($\omega = \{1, 1, 1, 0\}$) but left behind the HGD tissue, then treatment would only moderately reduce future EAC cumulative hazard by an expected 15.7% before 2030. However, removing the HGD tissue as well as preclinical malignancies ($\omega = \{1, 1, 0, 0\}$) during treatment would create a more significant average reduction in EAC cumulative hazard of an expected 38.7%. The model's predictions of the possible RFA effects on cell populations seem to support the hypothesis that the effectiveness of RFA is determined by its ability to ablate premalignant (dysplastic) tissue.

Interestingly, even the biopsy procedure on all BE patients offers a slight therapeutic effect (EAC cumulative hazard will not return to background) by the mere chance of endoscopically removing, at times, significant amounts of neoplastic tissue in a biopsy specimen, assuming no adverse effects from wounding associated with tissue removal. These results are clearly a simplification of a highly variable and complex clinical procedure, representing only a basic example, but the model is poised to incorporate realistic RFA touch-ups throughout surveillance, as it occurs in current practice, to give increasingly realistic projections. Beyond this example of a one-time treatment, we will apply this treatment methodology in the next chapter to a more complex scenario involving multiple treatments combined with surveillance biopsies to predict the durability of such eradication therapy and its impact on EAC incidence and mortality.

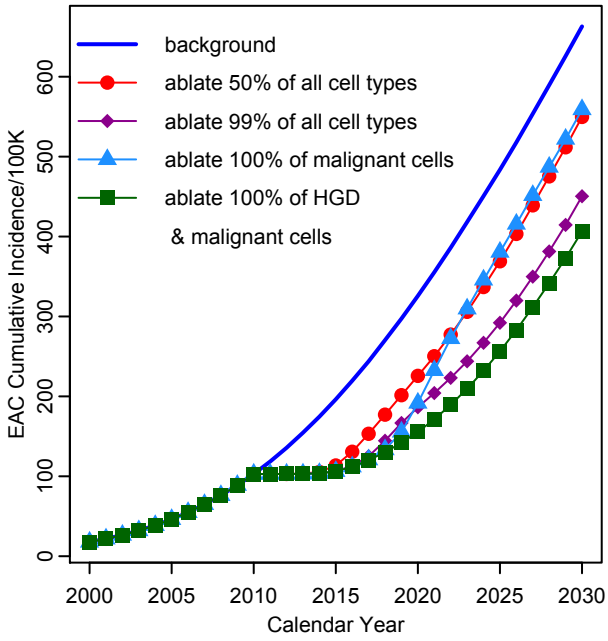

Figure 4.6: **Predicted EAC cumulative age-specific incidence by the MSCE-EAC screening model after RFA treatment of HGD patients at index screen.** EAC cumulative hazard for four different ablation efficiencies of detected HGD patients at screening of 60 year old males in year 2010. Survival for the four cell types following ablation modeled are represented by ablation proportion vector ω, with $\omega = \{1, 1, 1, 1\}$ (no treatment, blue line), $\omega = \{.5, .5, .5, .5\}$ (red, circles), $\omega = \{.01, .01, .01, .01\}$ (purple, diamonds), $\omega = \{1, 1, 1, 0\}$ (light blue, triangles), $\omega = \{1, 1, 0, 0\}$ (green, squares). Scenarios plotted for $\sigma = 3300$ stem cells/mm^2 and 60% biopsy sensitivity, resulting in 6% HGD cases, without detected malignancy, in the non EAC population at screening time (100K BE patient simulation size).

4.3 Discussion

Although few Barrett's esophagus (BE) patients progress to EAC in their lifetime, the cancer burden is considerable due to generally poor treatment outcomes and survival. EAC contributes approximately 4% to all male cancer deaths in the US [3] with a flattening but still increasing trend in mortality according to recent projections based on Surveillance, Epidemiology, and End Results (SEER) data [6]. Because BE is an actionable EAC precursor with a considerable prevalence of 1-3% in the general population [83, 105] (translating into a large number of individuals) and an annual risk of progressing to EAC of approximately 0.2-0.5% per year [12], optimal surveillance for neoplastic alterations in BE and effective treatment strategies are a major challenge to clinicians given the current lack of evidence-based decision tools. Thus we have developed a detailed multiscale model of EAC to better understand the natural history and impact of screening, intervention, and prevention of EAC.

The mathematical framework of our multistage clonal expansion for EAC (MSCE-EAC) screening model describes the step-wise progression and transformation from normal squamous esophageal tissue to a columnar crypt-structured metaplastic tissue in which clonal expansions of dysplastic and malignant cells can occur. Because the description is fully stochastic, it affords predictions of important clinical endpoints that reflect the intrinsic (inter-individual) heterogeneity in the disease process that explains, at least in part, why some individuals progress to cancer in their life-time while others do not.

In contrast to earlier formulations of the multistage clonal expansion models for EAC [13, 50], which analyzed patterns of EAC incidence in the general population, the present model includes two novel modules for exploration of clinical endpoints before symptomatic detection of EAC. The tissue module explicitly computes the number and sizes of neoplastic clones in a BE patient and quantifies their spatial structure within an idealized crypt-structured BE segment at time of screening. With this patient-specific information, we then employ a screening module to perform a screen *in silico* at a specified screening age. As our BE screening examples demonstrate, this model extension makes it possible to explicitly explore

current BE screening efficacy while controlling the operational characteristics of the screening protocol. We show that the detection of high grade dysplasia (HGD) or cancer using the standard (Seattle) biopsy protocol is strongly dependent on the minimum neoplastic tissue fraction needed to be detectable in a biopsy. This sensitivity would be further affected by altering the spacing between biopsy levels and size of the biopsy forceps according to different protocols. Additionally, our MSCE-EAC screening model predicts that over 10% of BE patients screened who receive a diagnosis of HGD with biopsy-based screening also harbor a missed preclinical malignancy with mid-range biopsy sensitivity. We find that the overall efficacy of the biopsy protocol is highly uncertain due to variability in tissue sampling between practitioners and due to considerable uncertainties in the histological assessment of the biopsied tissues.

Our results also suggest that even the best current biopsy protocols may miss between $70\% - 90\%$ of small HGD lesions that are detectable when using high-resolution optical coherence tomography (OCT) imaging at 1mm resolution. While not yet widely available, high-resolution OCT allows a more complete (wide-field) examination of the BE segment. Our results suggest that OCT could surpass the biopsy-based protocols in efficacy to detect neoplastic lesions. However, because quantitative data with OCT are still lacking, the results remain somewhat speculative, but serve to demonstrate the potential gains of OCT screening over the standard biopsy protocol.

Finally, the present framework also allows for the modeling of treatment, such as radiofrequency ablation (RFA). Ablation attempts to remove the intestinal metaplasia together with all neoplastic cells. Assuming that ablation simply decimates the number of BE, dysplastic, and malignant crypts by specific fractions, we computed the residual cancer risk of EAC after RFA (see Fig. 4.6). This 'decimation by fraction' approach also lends itself to modeling the curative effect of multiple RFA 'touch-ups' delivered over a span of time to improve RFA efficacy. From the results derived from simulating an ablative treatment on a population of BE patients found to be positive for HGD during screening, we found that it was crucial to ablate dysplastic and not only preclinical malignant tissue to achieve the most significant

impact on future EAC incidence. Although the example given in this study is somewhat simplistic and does not include the random spatial characteristics of the ablation process, the model framework can accommodate more complex assumptions regarding the biological effects of RFA, including random spatial effects of the ablation 'burn' and localized presence of intestinal metaplasia hidden beneath the neosquamous tissue after RFA treatment.

In summary, the MSCE-EAC screening model introduced in this chapter offers a comprehensive multiscale method to model the neoplastic processes unfolding in BE together with a mechano-spatial modeling of the screening process and treatment. Our results demonstrate the limitations of the standard biopsy-based protocol for the detection of HGD and early cancer due to a highly heterogeneous distribution of dysplastic precursors and malignant foci that can arise in dysplasia. We further demonstrate that these limitations could be overcome by high-resolution OCT imaging which may provide additional biological details and insights into the cancer process, including the growth dynamics of neoplastic clones (in particular their numbers and sizes over time), information that can easily be incorporated into the multiscale description of EAC development and screening presented here.

Chapter 5

Modeling Surveillance and Endoscopic Eradication Therapy

In the previous chapter on screening patients with Barrett's esophagus (BE), we introduced the multistage clonal expansion for EAC (MSCE-EAC) screening model that can be used to evaluate the effects of biopsy sampling, diagnostic sensitivity, and treatment on disease burden. This computational framework allows evaluation of the efficacy and sensitivity of current screening protocols to detect neoplasia (dysplasia and early preclinical EAC) in the esophageal lining. Beyond the results about the role of biopsy and imaging sensitivity involved in a single endoscopic screen, we also examined the potential efficacy of a one-time eradicative treatment performed on a population of BE patients in terms of the long-term impact on reducing EAC incidence. For those patients diagnosed with high grade dysplasia (HGD) at the index screen, we simulated the depletion of metaplastic and neoplastic cell populations in their BE by a prespecified amount. BE patients who are not detected with HGD during endoscopic screen do not receive treatment and retain all cell types (even with possible HGD clones missed during biopsy screening). Then we projected the EAC cumulative hazard for the entire US male population born in 1950, taking into account the progression of cell types of three subpopulations of individuals: 1) those BE patients diagnosed with HGD (contributions of reduced number of BE cell types after treatment), 2)

those BE patients with a negative screen for dysplasia (contributions of all BE cell types), and 3) those individuals who did not develop BE before the index screen (contributions of lifetime risk of EAC from normal squamous).

After examining these basic treatment scenarios, we use the MSCE-EAC screening model in this chapter for a more complex, realistic scenario that examines the impact of endoscopic eradication on EAC incidence and mortality predicted for various long-term surveillance strategies of BE patients based on presence of certain premalignancy or lack thereof. Extending the results of a one-time treatment, we will examine model predictions on the durability of treatment that will involve modeling multiple "touch-ups" during follow-up when recurrences occur. This project was part of a second comparative modeling analysis with the National Cancer Institute's (NCI) Cancer Intervention and Surveillance Modeling Network (CISNET) Esophagus group, termed "Base Case II," and we will include the MSCE-EAC screening model results along with some comparative results in this chapter, all of which are included in [15].

5.1 CISNET Base Case II: Impact of Endoscopic Eradication Therapy on EAC

Most societal guidelines recommend BE patients undergo endoscopic surveillance with tissue biopsy to grade the severity of precursor lesions and detect curable neoplasia [130,131]. While BE with no dysplasia progresses to EAC at a rate of less than 0.5% per year [12,132], BE with high grade dysplasia (HGD) progresses at a rate of 6%-19% per year [133]. In efforts to abate this problem, new techniques for endoscopic eradication of BE such as endoscopic mucosal resection (EMR) and radiofrequency ablation (RFA) have become more widely utilized with the aim of preventing progression to EAC. Current American Gastroenterological Association (AGA) guidelines unequivocally recommend such treatment, termed endoscopic eradicative therapy (EET), only for patients with HGD [134]. The incremental benefit for this therapy

on low grade dysplasia (LGD) and particularly non-dysplastic (ND) BE patients remains uncertain. Recent reports suggesting that RFA decreases cancer incidence among subjects with BE and LGD might prompt increased utilization of eradicative therapy in this lower-risk population [135]. Further, there is a growing evidence base regarding the efficacy and durability of RFA treatment [136–142]. The increasing availability of long-term data affords us the opportunity to analyze the impact of eradicative treatment on EAC incidence and overall mortality reduction using comprehensive and robust simulation models.

As described in Chapter 3 on Base Case I, the CISNET Esophagus group includes three modeling groups who independently developed population-based models for the natural history of BE and EAC that share common calibration targets (Surveillance, Epidemiology, and End Results (SEER) cancer incidence and mortality data) and were previously cross-validated through comparative modeling exercises [6].

The aim of our Base Case II study was to analyze the impact of endoscopic eradication therapy on EAC mortality in a BE population. Specifically, we sought to describe the impact of multiple different strategies utilizing eradication therapy on EAC incidence and mortality and to estimate the number of surveillance endoscopies and treatments required to produce potential clinical benefits.

5.1.1 Methods

In this analysis, we used the MSCE-EAC screening model (also termed the FHCRC model in comparative results) to quantitatively estimate the effectiveness and efficiency of endoscopic ablative therapies and compared our results with those from the other two CISNET models who estimated the same outcomes. As introduced in Chapters 1 and 3, the other two models are the Esophageal AdenoCarcinoma Model (EACMo) from the Massachusetts General Hospital (Boston, MA) (MGH model), and the Microsimulation Screening Analysis model from Erasmus University Medical Center (Rotterdam, The Netherlands) and University of Washington (Seattle, WA) (UW/MISCAN model). The CISNET-EAC models differ

by modeling approach and structure, but all use a common set of calibration data on EAC incidence by age, stage, and calendar year from SEER (1975-2009) (see Chapter 1). From the results of the Base Case I analysis presented in Chapter 3, all three models have been refined through comparative modeling using common population targets from SEER so that all models reproduce EAC incidence and mortality [6]. In this chapter, we apply the models to predict outcomes about surveillance and treatment strategies for BE patients that affect the natural disease progression to clinical cancer. In the case of the MSCE-EAC screening model, this amounts to using the calibrated biological parameters provided in Table 3.1 and the spatial parameters calibrated in Chapter 4 to simulate BE patient lifetime trajectories for each strategy.

All three models assume a step-wise progression from non-dysplastic (ND) BE towards dysplasia and EAC. To name one particular modeling difference, the MGH and UW/MISCAN models include two grades of dysplasia: LGD and HGD, whereas the MSCE-EAC screening model includes a singular grade of dysplasia, HGD. We chose not to model LGD as a separate, previous stage to the premalignant P clone population for several reasons. First, there is no evidence that a distinct clonal population of stem cells maintains LGD that undergoes expansion. Secondly, there is a high level of known intraobserver and interobserver variability with regard to the pathological categorization of epithelial lesions that show borderline features between tissue regeneration, sometimes due to GERD, and LGD [143–146]. Because HGD is distinguished from LGD by a pathologist based primarily on the basis of the degree of architectural and cytological aberrations [143], LGD is more often down-graded to "indefinite" for dysplasia at subsequent exams due to a lack of definite pathologic characteristics. In contrast, the detection of clinically relevant HGD lesions and carcinomas shows a much higher level of consistency in interpretation among pathologists [143]. Thus we focus on modeling HGD, the premalignant, expanding population of cells that are known to procure higher risk for developing EAC.

For this analysis, all three groups modified their original models from Base Case I to include additional modules containing the clinical details of RFA ablation and subsequent

surveillance and management. This section will focus on the methods employed in this analysis for the MSCE-EAC screening model but detailed technical profiles of the other two CISNET models are available online [22].

Simulated population

Hypothetical cohorts for male and female patients, age 50, 60 and 70 when diagnosed with BE, were followed for EAC incidence and mortality until death or age 100. Endoscopic surveillance and eradication therapy were discontinued at age 80. For the MSCE-EAC screening model, the cohorts analyzed were stratified by initial dysplasia status - high grade dysplasia (HGD) or BE with no dysplasia (ND). Cancer risk was dependent on calendar year, birth cohort, age, and sex. Again, each model was calibrated in Base Case I to reproduce the EAC incidence in the SEER registry [6].

Simulated strategies

We modeled and analyzed four strategies described in Table 5.1. In the "Natural History" (NH) strategy, there was no endoscopic screening or surveillance; patients came to medical attention only when a clinical cancer was diagnosed, at which point they would receive standard treatment. When a cancer was diagnosed, survival was modeled according to SEER survival data or survival of esophagectomy, which we will describe further in this section forthcoming. The "Surveillance" (S) strategy refers to scheduled endoscopies previously recommended by numerous societal guidelines prior to the widespread availability of endoscopic eradication therapy. The majority of societal guidelines base the interval of surveillance endoscopy solely on the histological grade of biopsy samples [134]. The two endoscopic eradication treatment (EET) strategies varied by the diagnosed histological stage of patients at which endoscopic eradication is first performed. In the "HGD" strategy, patients with BE underwent endoscopic surveillance until HGD was detected on endoscopic biopsy, at which point the patient underwent treatment (see Fig. 5.1). The other two CISNET models also simulated a third "LGD" strategy in which patients underwent treatment when any dysplasia (HGD or LGD)

was detected on biopsy. The MSCE-EAC screening model does not include results for the strategy "LGD" that considers LGD patients as separate from NDBE patients. Finally, in the "BE" strategy, all BE patients underwent treatment at the start of the simulation regardless of degree of dysplasia (see Fig. 5.2).

To obtain screening outcomes during surveillance (i.e., diagnoses of BE patients at each screen) as was explained in the Methods section of Chapter 4, the MSCE-EAC tissue module explicitly simulates the growth of non-dysplastic BE crypts, HGD crypts, and malignant crypts as a BE patient ages, and thus generates the joint distribution of premalignant and malignant clones sizes at any age before symptomatic cancer is detected in a BE patient. With this information, the MSCE-EAC screening module then simulates a biopsy procedure (image-based screens are also possible with our model) at every surveillance screen to determine a possibly different diagnosis based on the highest grade (non-dysplastic BE, dysplastia, or cancer) of tissue found on biopsy. With this screening simulation, we are able to predict the potential presence (or absence) of malignant cells in biopsies that harbor a sufficiently large number of dysplastic crypts to be subjected to closer examination for the presence of malignant cancer. The clinical results presented here are based on use of the standard (Seattle) biopsy protocol which requires quadrant biopsies every 1-2 cm along the BE segment. Factors contributing to the sensitivity for detection of HGD or cancer lesions include the minimum neoplastic tissue fraction in the biopsy necessary for diagnosis, the spacing between biopsy samples, and the size of biopsy forceps. We found that the general efficacy of biopsy sampling remains uncertain due to considerable uncertainties in the histological assessment of the biopsied tissues [14]. For the Results of Base Case II, we employed a mid-range biopsy detection sensitivity of 40% that reproduces an average prevalence of HGD on index endoscopy as reported in current clinical studies (see [14] and Chapter 4 for details on the definition of diagnostic sensitivity, Fig. 4.4). For the 60 year old males with BE simulated here, this assumption yields a prevalence of 2.8% initial screen-detected cancers and 4.7% initial HGD cases. We also employed a neoplastic clone diffusivity spatial parameter of $\gamma = 1$ (nearly circular clones) because we found that the screening diagnoses

Strategy	NDBE patients	HGD patients
Natural History (NH)	No intervention	No intervention
Surveillance (S)	Surveillance endoscopy with biopsies every 3 years	Surveillance endoscopy with biopsies every 3 months
BE surveillance and HGD treatment (HGD)	Surveillance endoscopy with biopsies every 3 years	Endoscopic eradication therapy followed by surveillance*
BE treatment (BE)	Endoscopic eradication therapy followed by surveillance*	Endoscopic eradication therapy followed by surveillance*

Table 5.1: **Simulated intervention strategies on BE patient cohort.** BE: Barrett's esophagus, ND: No dysplasia, HGD: high grade dysplasia. *All post-treatment surveillance intervals can be found in Table C.1.

were mostly insensitive to choice of γ (determining the degree of compactness vs. branching for neoplastic areas, see Chapter 4 Results).

Modeling endoscopic eradication therapy

After a simulated screen of a BE patient for detection of HGD and preclinical EAC at a specified screening age, the MSCE-EAC screening model simulates an ablative treatment, such as radiofrequency ablation (RFA), as required by the specified strategy given the BE patient's disease state. Treatment with such endoscopic eradicative therapy (EET) attempts to remove the intestinal metaplasia together with all neoplastic cells at risk to become cancer. As described in the Methods section of Chapter 4, the current MSCE-EAC screening model assumes that treatment decimates the number of BE, dysplastic, and malignant crypts by specific fractions controlled by ablative proportion vector ω, such that we modify the size of a patient's BE segment along with any concurrent HGD and/or malignant lesions during

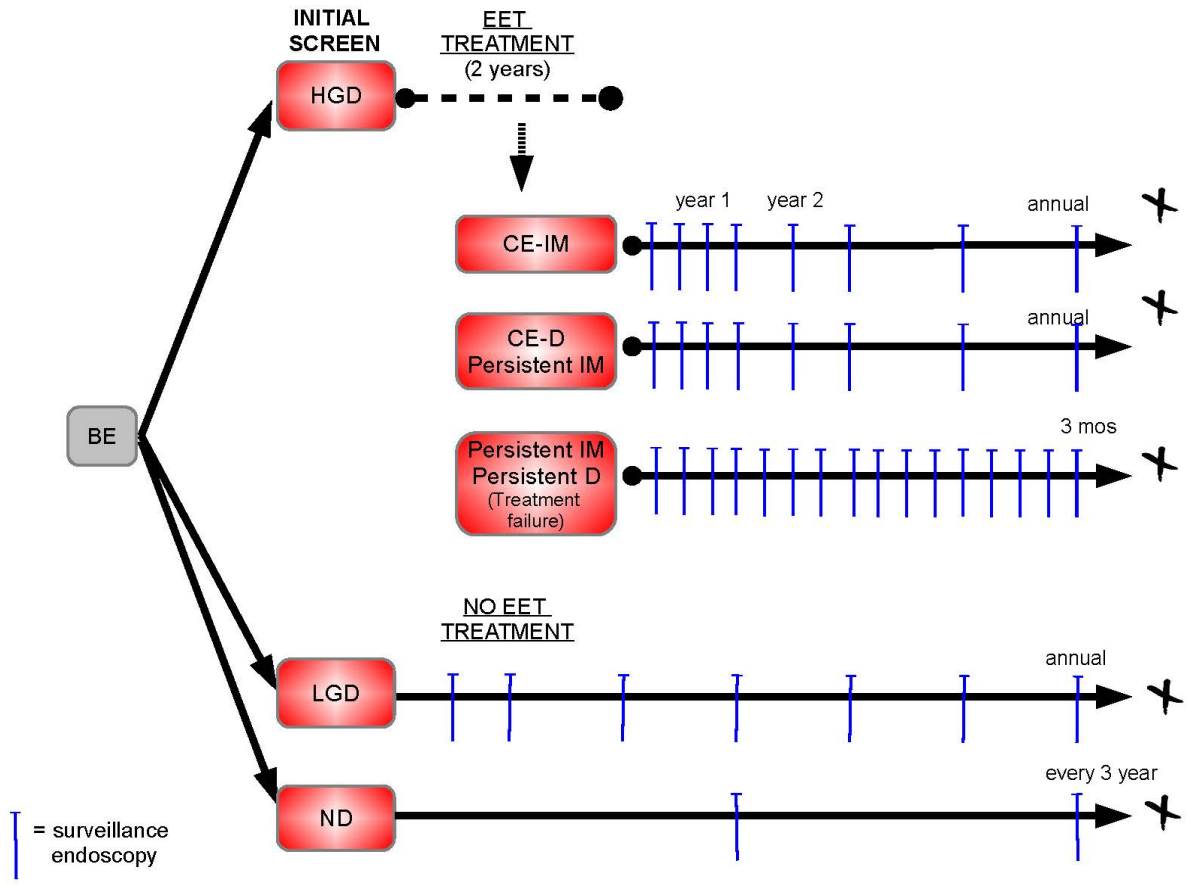

Figure 5.1: **HGD strategy for treatment and surveillance of BE patients.** In the HGD strategy, those detected with HGD at index screen or found to have HGD (de novo from NDBE or recurrent HGD) on any surveillance screen are treated (see Table C.1 for details on surveillance after treatment). The MSCE-EAC screening model does not include a separate diagnosis (nor a separate surveillance protocol) of LGD from NDBE. NDBE patients undergo surveillance as described in Table 5.1.

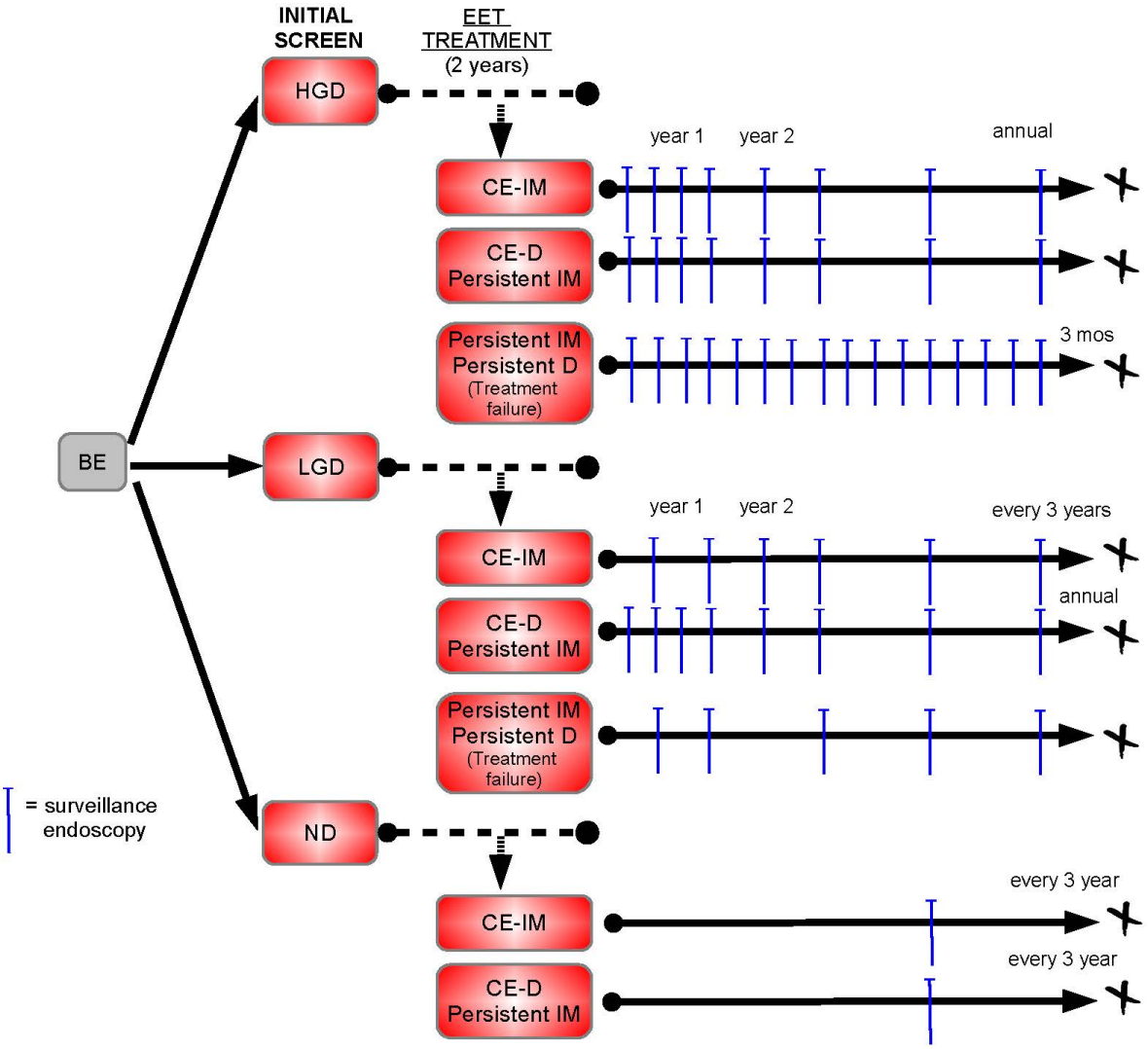

Figure 5.2: **BE strategy for treatment and surveillance of BE patients.** In the BE strategy, all those detected with BE at index screen or recurrent BE on any surveillance screen are treated (see Table C.1 for details on surveillance after treatment). Contrary to the HGD strategy shown in Fig. 5.1, this strategy treats all BE patients regardless if dysplasia is detected. The MSCE-EAC screening model does not include a separate diagnosis (nor a separate surveillance protocol) of LGD from NDBE.

treatment. After treatment, the MSCE-EAC tissue module continues the simulation of the BE cell types that remain until the time of next screen, clinical cancer detection, or death.

The efficacies associated with endoscopic eradication therapy (EET) for BE were based on recently published data [147,148] and expert opinion (see Table 5.2). In the Results here, we assumed an efficacy of 70% removal of all cell types equally ($\omega = \{\omega_X, \omega_{P^*}, \omega_P, \omega_M\} = \{0.3, 0.3, 0.3, 0.3\}$) during an eradicative treatment or touch-up. We calibrated this treatment proportion to published estimate of a 0.15 recurrence rate of HGD after three years of follow-up surveillance for initially diagnosed HGD patients [147, 149, 150]. Initial RFA treatment took place over a two-year period, and was preceded by endoscopic mucosal resection (EMR) in 55% of the patients. Possible outcomes at the end of this period were complete eradication of intestinal metaplasia (CE-IM), complete eradication of dysplasia (CE-D), or treatment failure (persistence of IM and/or D). After treatment failure, patients received endoscopic surveillance at pre-treatment intervals and were not given additional treatments. After treatment success (CE-IM or CE-D), patients were subject to a modified surveillance regimen that included additional endoscopies in the years immediately after the initial treatment period, with later endoscopies following at less frequent but regular intervals. Full details of the post-treatment surveillance strategy for each treatment outcome and pre-treatment state are shown in Appendix C, Table C.1. Following successful treatment, patients could recur to BE, based on the rates in Table 5.2, and those who recurred were also given an additional diagnosis of BE without dysplasia, BE with HGD, or BE with screen-detected EAC after a biopsy procedure simulated by the MSCE-EAC screening module. The probability of recurrence to BE was assumed to be constant over time. Unlike the other two CISNET Markov disease state models, our interpretation is that BE cell types may still exist after treatment, which will drive possible recurrences or first appearances of neoplasia; a BE patient does not re-enter a normal health state (i.e., cured) unless all cells are possibly eradicated with multiple treatments. Patients with recurrences detected during post-treatment surveillance received "touch-up" RFA treatment (defined as circumferential or focal endoscopic RFA performed after the initial treatment period) and were monitored for

Parameter/Definition	Value	Source
Success of therapy in pre-treatment HGD patients		[147]
CE-IM and CE-D	88.89%	
Non CE-IM, CE-D	3.70%	
Non CE-IM and Non CE-D	7.41%	
Success of therapy in pre-treatment NDBE patients		[148]
CE-IM	96.77%	
Non CE-IM	3.23%	
Recurrence rates by baseline histologic grade and grade of recurrence		[151, 152]
Annual recurrence rates to BE after CE-IM		
Pre-treatment NDBE	7%	
Pre- treatment HGD	10%	

Table 5.2: **Common input parameters for effectiveness of EET.** BE: Barrett's esophagus, ND: No dysplasia, HGD: high grade dysplasia , CE: complete eradication, IM: intestinal metaplasia (BE), D: dysplasia, RFA: radiofrequency ablation, EAC: esophageal adenocarcinoma *Expert consensus: panel of experts N Shaheen; S Spechler; J Inadomi; C Hur; J Rubenstein.

further recurrences according to the post-treatment schedule described above. Patients were limited to a maximum of three touch-ups. Again, graphical representations of the simulated treatment strategies can be found in Figs. 5.1-5.2.

Modeling patient survival

Once a malignant lesion is screen-detected, a BE patient may undergo surgery based on his/her surgical candidacy, whether endoscopic mucosal resection or esophagectomy. The

immediate mortality risk of esophagectomy was set to 5%. We utilized data from the SEER registry to model cause specific EAC survival and cure rate trends by stage and age category (ages 50-59, 60-60, 70-84), using the CANSURV program to fit a lognormal survival model to the data while estimating temporal trends on the shape and cure parameters [153]. The all-stage EAC survival curves for each age category were adjusted to account for ablation or surgical resection by fitting the cure model parameters based on a study of 430 patients undergoing ablation and 1586 patients undergoing esophagectomy that were identified in SEER between 1998-2009 [154]. Separate models were developed for EAC cause specific survival by age group, with or without ablation or surgical resection, while accounting for censoring and other cause death by matching cure rates at 2003.5 (midpoint of the 1998-2009 SEER follow-up data from [154]) at age 63.4 (mean patient age for surgical resection), or age 70.5 for ablation. Competing other cause death times were drawn for each patient from US life tables that were downloaded from the Centers for Disease Control website [155] to calculate age-adjusted rates for all race males and females

Outcomes

The main outcomes were presented for a 60 year old male cohort (additional outcomes for female and various ages are shown in Appendix C, see Tables C.3 and C.4). For the MSCE-EAC screening model, expected numbers of outcomes were produced by 100K simulation size of BE patients, for males and females separately. The primary outcomes were EAC incidence and mortality reduction; total numbers of surveillance endoscopies and endoscopic eradicative treatments; numbers of treatments needed to avert one EAC death (NNT/death); life years gained; and complications of endoscopy and treatment. The NNT/death was calculated as the total number of ablative treatments divided by the number EAC deaths averted by a given strategy. We incorporated the total number of treatments needed to prevent one death because multiple treatments were needed per patient. Presenting the results as the number of patients needed to treat would underestimate the overall resources needed. Treatments included the number of sessions of (possibly) joint EMR and RFA treatments. Incremental

results compared the NNT/death for a given strategy to the next-least aggressive strategy by dividing the number of additional treatments by the additional EAC mortality reduction in the more aggressive strategy (Table 5.3).

Sensitivity analysis

We repeated our four simulated strategies with half and twice the base-case assumptions for the durability of successful treatment and for the efficacy of the initial treatment. In addition, we analyzed the effect of halting surveillance after a period of 5 and 10 years of observed good health (no BE recurrence) post-treatment (see Table C.2 for sensitivity analysis parameters).

5.1.2 Results: FHCRC model

EAC incidence and mortality

Without surveillance, 134 EAC cases and 84 EAC deaths were expected to occur in 1,000 60 year old male BE patients (see Table 5.3 for these MSCE-EAC screening model results). Surveillance (S) led to down-staging and an EAC mortality reduction of 20%; however, there was a 36% increase in cancer detection due to overdiagnosis (surveillance-detected EAC that would not have become clinically observed due to death from non-EAC causes). Compared to the surveillance only strategy, the HGD treatment strategy resulted in a decrease in EAC diagnosis of 46% and EAC mortality reduction of 41%. Treating all BE patients at age 60 decreased the number of EAC cases by 81% and the EAC mortality by 74%. The relative impact of the different treatment strategies was consistent across models (See Fig. 5.3 for comparative bar graphs with other two CISNET models).

Resources required

For the MSCE-EAC screening model, 926 total number of treatments (including EMR, ablative treatments and touch-ups) were performed in the HGD strategy (see Table 5.3). Ex-

Results per 1000 BE patients	NH	S	Incremental results vs. NH	HGD	Incremental results vs. S	BE	Incremental results vs. HGD
Number of surveillance endoscopies	0	5799	+5799	8984	+3185	5806	-3178
Number treatments	0	0	+0	926	+926	4902	+3977
Number of EAC cases	134	182	+36%	98	-46%	34	-65%
Number of EAC deaths	84	67	-20%	40	-41%	17	-56%
Life expectancy after diagnosis (years)	19.3	19.4	+0.0	19.7	+0.3	19.8	+0.2

Table 5.3: **Main and incremental results per strategy.** Strategies: NH: Natural History strategy; S: Surveillance strategy,HGD: Endoscopic ablative therapy for HGD diagnosed patients strategy; BE: Endoscopic ablative therapy for all BE diagnosed patients

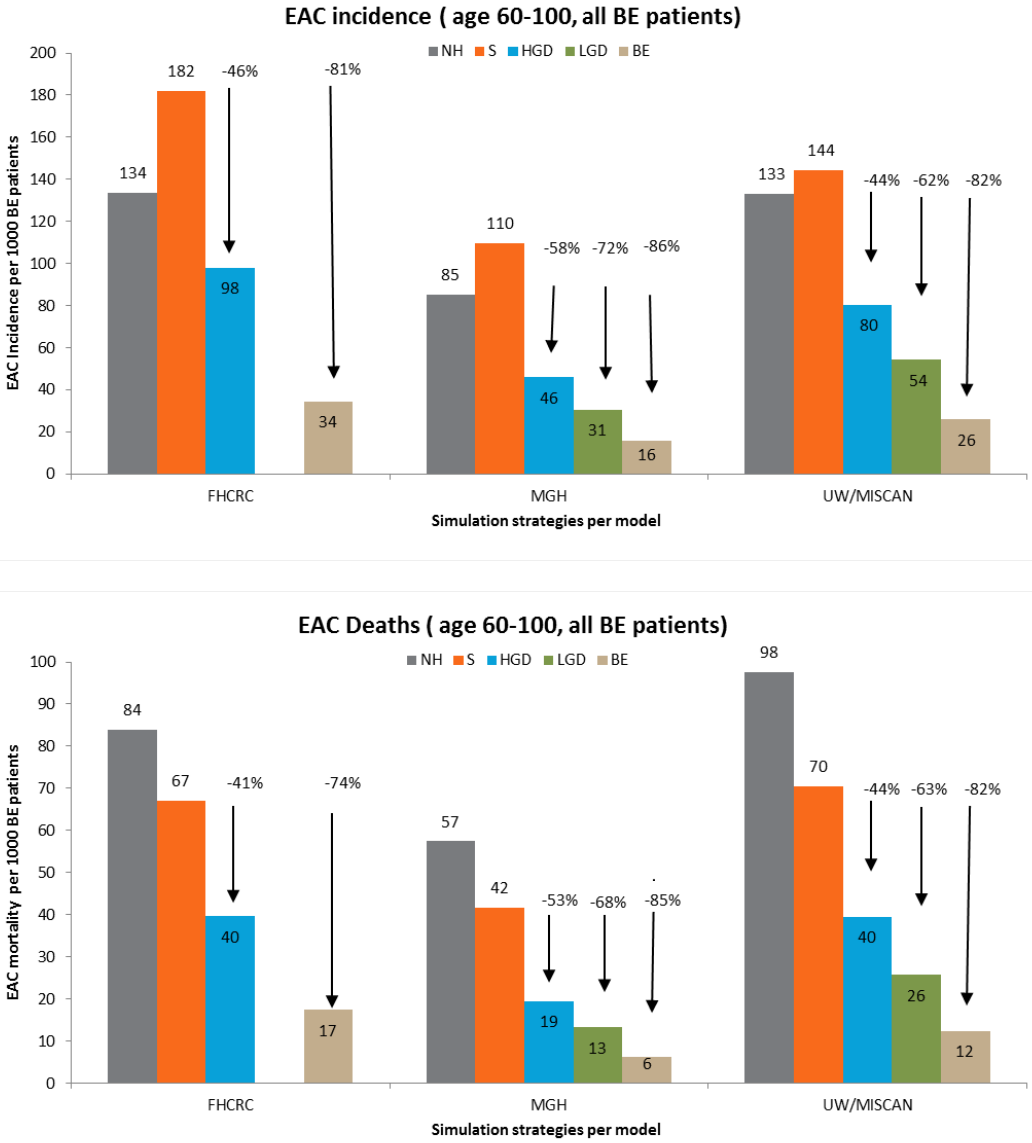

Figure 5.3: **EAC incidence and mortality reductions by strategy** (Upper panel) Predicted EAC incidence per 1000 BE patients per model and strategy. (Lower panel) Predicted EAC deaths per 1000 BE patients. The EAC incidence and mortality reduction percentages are shown for the endoscopic eradicative treatment strategies compared to the strategy including only surveillance and no EET. Stategies NH (gray), S (orange), HGD (blue), BE (tan) shown for all models. MGH and UW/MISCAN results shown for comparison, including a strategy for LGD patients (green).

tending treatment to all BE patients further increased required treatments to 4,902. The number of treatments differed across models, but showed similar patterns for each treatment strategy (see [15] for full description of model result comparisons).

Efficiency of treatment

The significant increase in treatments diminished the efficiency per treatment for the more inclusive strategies. The NNT/death for HGD treatment strategy was 34 (see Fig. 5.4). In this strategy, relatively few treatments were required, resulting in a high mortality reduction (41%). In contrast, the incremental NNT/death for BE treatment strategy compared to HGD treatment strategy was 180 in the MSCE-EAC screening model (see Table C.3 for analogous results for females). We also provide tabular results for incremental NNT/death for males with BE, all races, ages 50 and 70 at index endoscopy (see Table C.4).

Sensitivity analysis

For the HGD treatment strategy, the results of our models were robust to sensitivity analysis. Comparing incremental NNT/death for the BE strategy with the HGD strategy, the results were most sensitive to the durability of successful treatment and for halting surveillance after a period recurrence-free post-treatment surveillance. However, halting post-treatment surveillance after a recurrence-free period of 5-10 years had negligible influence on NNT/death in the HGD strategy (see Appendix C, Fig. C.1).

5.1.3 Discussion

Our study shows that endoscopic eradication of HGD, specifically RFA, could result in substantial reductions in EAC incidence and mortality. However, extending treatment eligibility to patients with lower grades of dysplasia substantially increases the use of eradication therapy while diminishing the incremental effectiveness. A strategy treating all patients with BE (including HGD, LGD and NDBE) results in an unfavorable number needed to treat to pre-

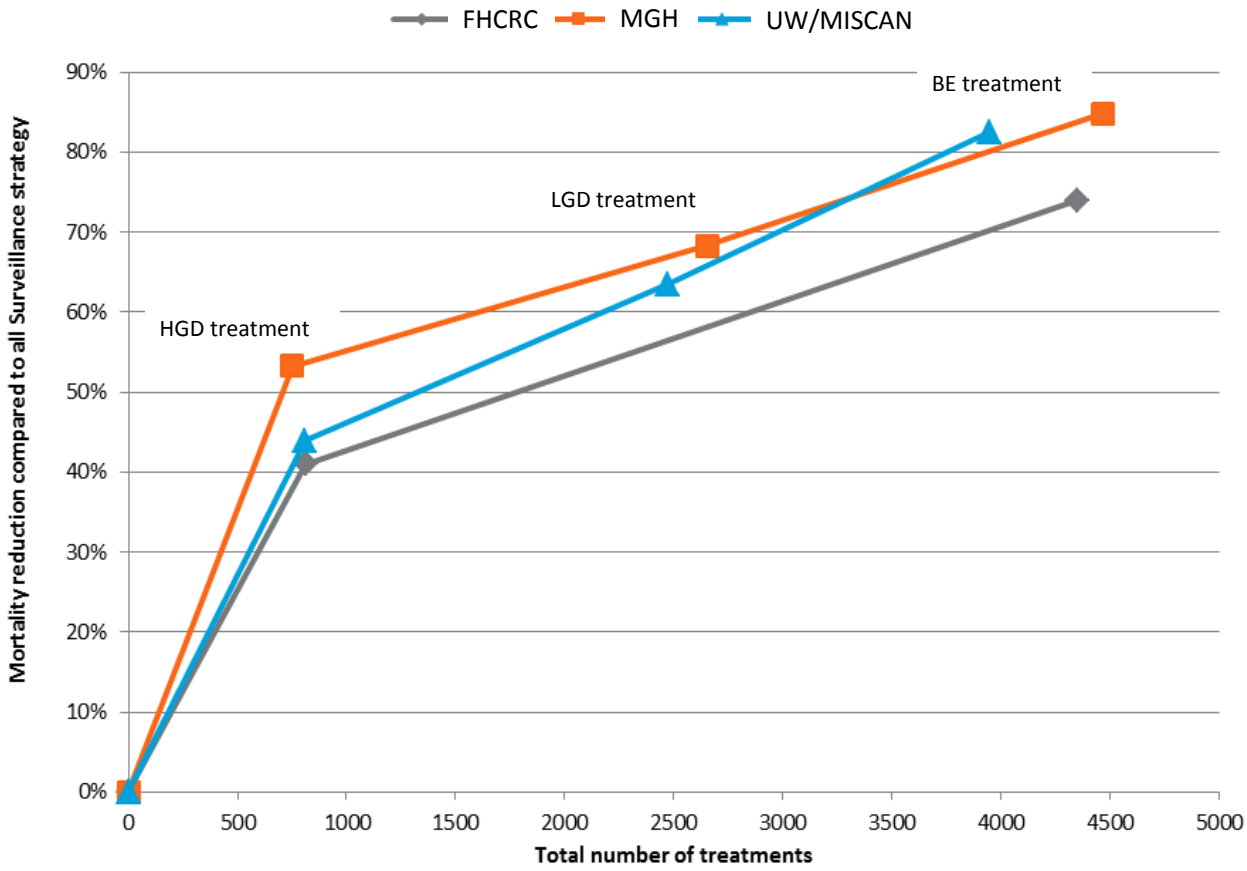

Figure 5.4: **Mortality reduction compared to the total number of treatments per model and strategy.** Results for MGH and UW/MISCAN models shown for comparison, including the LGD strategy.

vent one EAC death. The finding that EET may reduce EAC incidence and mortality is not surprising as the efficacy of the treatment is reported to be high and associated complication rates are relatively low. The more relevant issues to applying this therapy on a population basis are related to healthcare resource utilization, over-screening and over-treatment. Evaluation of the NNT to achieve additional mortality reduction for each strategy demonstrates eradicative therapy for patients with no dysplasia results in diminishing returns. It appears

that the diminishing impact of treatment expansion is due to the likelihood that ND and LGD patients will eventually receive treatment if they develop HGD.

The model results were most sensitive for the duration of successful treatment. Furthermore, all models support the decision to stop offering surveillance to HGD patients five years after successful RFA. All sensitive analyses had a relatively larger impact on the treatment strategies for low grade and absent dysplasia, due to the fact that more patients are being treated with increasingly variable results. Previously published cost-effectiveness studies agree that endoscopic eradication therapy is cost-effective when offered to HGD BE patients [149, 156, 157]. One previous study evaluated the cost-effectiveness of RFA on varying dysplastic grades of BE, concluding that endoscopic ablative therapy was only cost-effective when offered to dysplastic BE patients [149]. Our study similarly showed eradicative treatment is effective for all patients, and that treating patients with less severe or no dysplasia demands a major amount of resources. Prior studies used a single Markov model informed by clinical data available at the time of publication, but were not calibrated to US SEER incidence and mortality data. This study used three simulation models that were independently calibrated to SEER data, which better equips them to assess cancer control strategies and patient guidelines. With the comparative modeling approach using common inputs and outputs, this analysis showed considerable consistency between the models on the relative effectiveness and the NNT/death, demonstrating the robustness of our findings even though there are considerable differences between our models when considering absolute malignant development and mortality among BE patients.

In conclusion, our comparative modeling analyses indicate that EET is an effective means of reducing EAC incidence and mortality. Benefit is predicted to be achieved across all patients with BE; however, the efficiency of eradication is substantially reduced if patients with no dysplasia are treated, and substantially more healthcare resources are required to avert a cancer death in these settings. These findings were consistent across all three esophageal CISNET models and were robust to sensitivity analyses of RFA efficacy and durability. Our results add further evidence to support RFA therapy to patients with HGD, and suggest that

strategies targeting less severe disease will require close scrutiny for cost-effectiveness. Current and future work on this collaborative project includes a full cost-effectiveness analysis for each treatment strategy versus surveillance only. Efficiency of care would be greatly enhanced through improved methods to stratify risk of cancer in lesser forms of dysplasia and, therefore, to better identify individuals who would benefit most from endoscopic therapy.

Chapter 6

Optimal, Adaptive Screen Design

The previous chapter presented expected health outcomes and EAC mortality reductions under fixed surveillance and treatment schedules in current practice. Specifically, we used the MSCE-EAC screening model to quantify the mortality reduction predicted for BE patients who undergo endoscopic eradication therapy. Further, in a cost-benefit analysis with cost defined as the number of treatments required, we compared the relative effectiveness of current treatment strategies for BE patients with high grade dysplastia (HGD) versus the potential benefits of treating all BE patients, with and without dysplasia. In the next two chapters, rather than focusing on "who" is currently benefiting most from treatment during a set surveillance screen schedule, we ask questions regarding "when" should these BE patients be screened. Implicitly, this goal relies on determining when BE patients develop detectable precursor lesions and how long they dwell in a certain detectable disease state. For all cancer screening recommendations, optimal timing of the examinations of susceptible 'at risk' patients for the presence of pre-cancerous or cancerous lesions is important. In the case of EAC, we showed in the last chapter that over-surveillance of BE patients with no evidence of HGD (but possibly indefinite or low grade diagnoses) poses a costly problem as many individuals with BE return to a clinic for upper endoscopies up to every 3 years by current guidelines but never progress to HGD (Table 5.1). To address this problem we propose to derive and to optimize the timing of clinical screens so that the probability is

maximal that an individual is screened within a certain "window" when a premalignancy prior to cancer development may be observed.

The problem of optimal design for screening times via biomathematical modeling has been studied previously. Dewanji et al. defined two optimality criteria for a simple illness-death model as more intuitive alternatives to standard likelihood-based criteria [158]. For a more complicated example, Hanin et al. found optimal screening schedules for preventing metastatic cancer by maximizing an efficiency functional for diagnosing existing metastases through screening versus without screening [159]. In the case of the earlier MSCE model with one clonal expansion stage, Jeon et al. explored solutions for the optimal age to screen and treat patients in order to minimize the lifetime probability of developing colorectal cancer [122]. By imposing a size threshold for adenomas to decide when a screen can detect an adenoma, the authors also optimized a two screen schedule by simulating cell dynamics in patient populations. With this approach, they predicted that undergoing colonoscopies every 10 years starting at age 50, as recommended by the current guidelines, does not significantly reduce cancer risk in comparison to two optimally scheduled colonoscopies before age 80.

Unlike the scheduling results of Hanin et al. [159], this chapter's methodology includes the optimization of screening times before a malignant tumor forms so that appropriate risk stratification, patient-specific screening schedules, and treatment decisions during premalignancy may be achieved. Here we develop a generalized version of the optimality criteria first presented by Dewanji et al. [158] for screening and derive analytical probabilities that allow estimation of optimal screening times for the detection of premalignant lesions and ones that do not require numerical simulations of patient populations [122]. With the model parameters previously inferred for our fully stochastic MSCE-EAC screening model [6, 14], we provide a few example applications for BE patients. The derived probabilities for optimized screening are flexible so that weighing risk of an undesirable disease state (such as development of clinical EAC) before a screening time may be specified in every case. We first derive formulas for the case of a one-time screen and then provide a framework for optimizing adaptive screens, i.e. optimizing time of subsequent screening conditional on the

specific diagnosis of a prior screen.

6.1 Random Variables for Screening and Risk

We aim to derive the time (age) of a screen such that a patient is most probable to be in a certain disease state of interest. One example is, for an individual who has never been screened, we would like to screen him/her in a window that he/she has already developed BE but will not developed HGD. If we "caught" this patient in such a window, there is a possibility that low-level treatment could be administered and protect this patient from developing a more severe diagnosis (this scenario is defined in Strategy 2 in upcoming sections). As we see in the multistage process illustrated in Fig. 3.1, there are various such windows of time that patients are in a particular disease state with certain probability and we seek to maximize the yield of a screen by varying screening time to avoid fruitless overscreening.

First, we define the random variables of interest that represent four clinical points of the multistage process: BE (T_{BE}), premalignancy (T_P^d), which we posit to be HGD, malignancy (T_M^d), and symptomatic cancer (T_C).

$$T_{BE} = \text{time of BE onset}$$

$$T_P^d(t) = \begin{cases} 0 \text{ if no detectable premalignant clone exists in BE at time } t \\ 1 \text{ if there exists at least one detectable premalignant clone in BE at time } t \end{cases}$$

$$T_M^d(t) = \begin{cases} 0 \text{ if no detectable malignant clone exists in BE at time } t \\ 1 \text{ if there exists at least one detectable malignant clone in BE at time } t \end{cases}$$

$$T_C = \text{time of cancer (EAC) spontaneous (symptomatic) detection}$$

To incorporate patient-specific information, one could also differentiate T_{BE} to refer to either T_{BE}^+ or T_{BE}^- for symptomatic GERD BE onset time versus a non-GERD population BE onset time. As in previous chapters, we will derive the forthcoming formulas using a general BE onset distribution $f_{BE}(t)$, but example results will be obtained when we assume the

exponential distribution for BE onset time defined by age-dependent rate $\nu(t)$ characterized by GERD prevalence (first provided in Eq. (3.2)), given by

$$f_{BE}(t) = \nu(t)e^{-\int_0^t \nu(u)\,\mathrm{d}u}. \tag{6.1}$$

Unlike the one-time occurrence, continuous random variables for T_{BE} and T_C, the binary random variables for detectability of premalignancy ($T_P^d(t)$) and malignancy ($T_M^d(t)$) require specific definitions for detectability. If we designate these times to be the time of initiation or malignant transformation of a cell regardless of its fate (i.e., if it will be the ancestor to a clonal progeny that eventually goes extinct or that survives), we may derive analytical probabilities with first-passage time, continuous random variables T_P and T_M, respectively, in a straightforward way using the Kolmogorov forward equations introduced in Chapter 3. However, it is more clinically relevant to define these random variables as binary outcomes for detectable lesions at the time of screening t_s. In the derivations of our screening windows, all detectability criteria will depend on the probability of detection at time t_s of a premalignant lesion that began with a single ancestor P cell initiated at time s, $p_2(s,t_s) = \Pr[T_P^d(t_s)|P(s) = 1]$, and analogously the probability of detection at time t_s of a malignant lesion that began with a single ancestor M cell initiated at time s, $p_1(s,t_s) = \Pr[T_M^d(t_s)|M(s) = 1]$. For example, Jeon et al. defined this detectable screening outcome variable for premalignant, adenomatous polyps in the colon as being at least a certain threshold size, i.e., $p_2(s,t_s) = \Pr[P(t_s) \geq P_t|P(s) = 1]$ where threshold size P_t was set equal to 10^3, 10^4, or 10^5 polyp stem cells to incorporate the instrumental sensitivity of a colonoscopy. In the case of EAC, this definition of size threshold was applied when evaluating the possible improvement in screening sensitivity for HGD and malignant lesions when using optical coherence tomography (OCT) imaging techniques with resolution a_{OCT} in Chapter 4. Alternatively, these variables could be used in conjunction with biopsy screening and we could use the MSCE-EAC screening model to compute (via simulation) these two probabilities given times s and t_s.

We could also define these binary detection variables to be equal to 1 if a clone born at

time s is non-extinct at time t_s (size of clone is ≥ 1) and 0 otherwise. In the case of EAC, this "perfect sensitivity" definition may be useful as progressively better high-resolution imaging technologies or other sensitive screening methods come into clinical practice. For detectability based on this non-extinction criteria at time of screening, we define the probability of detection with a binary outcome variable for malignant clones with the following (see derivation in Appendix B for unconditional birth-death-mutation size distribution),

$$p_1(s, t_s) = 1 - \beta_{_M} \zeta_{_M}(s, t_s) \tag{6.2}$$

where

$$\zeta_{_M}(s, t_s) = \frac{e^{(\alpha_{_M} - \beta_{_M})(t_s - s)} - 1}{\alpha_{_M} e^{(\alpha_{_M} - \beta_{_M})(t_s - s)} - \beta_{_M}}.$$

As discussed in Chapters 2 and 3, the asymptotic version of this probability was derived as $p_\infty \equiv p_1(s, t_s) = 1 - \beta_{_M}/\alpha_{_M}$. We obtain the analogous definition for $p_2(s, t_s)$ with P parameters for the detection probability of premalignant clones,

$$p_2(s, t_s) = 1 - \beta_{_P} \zeta_{_P}(s, t_s) \tag{6.3}$$

where

$$\zeta_{_P}(s, t_s) = \frac{e^{(\alpha_{_P} - \beta_{_P})(t_s - s)} - 1}{\alpha_{_P} e^{(\alpha_{_P} - \beta_{_P})(t_s - s)} - \beta_{_P}}.$$

The asymptotic version of this probability is $p_2(s, t_s) = 1 - \beta_{_P}/\alpha_{_P}$.

Second, we present the two survival functions, derived in previous sections, that will appear throughout the mathematical derivations:

$$\Pr[T_{_C} \geq t_s | T_{_{BE}} = \tau] = S_{_{MSCE}}(t_s - \tau) \tag{6.4}$$

$$\Pr[T_{_C} \geq t_s] = S_{_{EAC}}(t_s) \tag{6.5}$$

Where $S_{_{MSCE}}(t)$ is the survival function for the 4-stage MSCE model after BE onset (see Eq. (2.24)) and $S_{_{EAC}}(t)$ is the survival function for the full MSCE-EAC model (see Eq. (3.26)).

6.1.1 Differentiating importance of events

For some event A (e.g., BE onset) and event B (eg., symptomatic EAC detection), maximizing the probability that a patient is between ages T_A and T_B at time of screening t_s is equivalent to maximizing $\Pr[T_A \leq t_s]$ to ensure that event A has already occured while simultaneously minimizing $\Pr[T_B \leq t_s, T_A \leq t_s]$ to ensure that event B has not yet occurred by t_s. This statement is defined mathematically in the relationship

$$\Pr[T_A \leq t_s \leq T_B] = \Pr[T_A \leq t_s] - \Pr[T_B \leq t_s, T_A \leq t_s]. \tag{6.6}$$

If event B implies event A, the second term on the right hand side of this equation simplifies to $\Pr[T_B \leq t_s, T_A \leq t_s] = \Pr[T_B \leq t_s]$, the absolute risk of event B by time t_s. We will discuss the calculation of risk in the following section.

It may not be a justifiable public health goal to equally weigh the optimizations that event A has occurred and event B has not occurred, as is the case for maximizing the left hand side probability in Eq. (6.6). In general, we may define a positive weight w for event B, and instead maximize the following :

$$\Pr[T_A \leq t_s] - w \cdot \Pr[T_B \leq t_s, T_A \leq t_s] \tag{6.7}$$

For the case of weighing both events equally during optimization of a screen time occurring within specific disease state window, we may set $w = 1$ as in Eq. (6.6). As an alternative to equal weighting, criterion $\mathscr{C}_2$ from Dewanji et al. [158] aimed to maximize time t_s in the window between events A and B (where event B implies event A) with an extra penalty term for event B. In the notation introduced above, this criterion can be written as

$$\Pr[T_A \leq t_s \leq T_B] - \Pr[T_B \leq t_s] = \Pr[T_A \leq t_s] - \Pr[T_B \leq t_s] - \Pr[T_B \leq t_s] \tag{6.8}$$

$$= \Pr[T_A \leq t_s] - 2\Pr[T_B \leq t_s] \tag{6.9}$$

This is the case of $w = 2$ in our generalized methodology introduced here. If $w > 1$, maximization of Eq. (6.7) will yield earlier optimal screening times t_s^* because there will be greater weight to avoid that event B occurs before screening, and vice versa for $w < 1$. As

we will elaborate on in the next section, cancer screening policy makers face a necessary question as to how much weight should be placed on each possible scenario - how heavily to guard against events B, such as EAC cancer development, balanced against the prospect of futile screens when a patient has not yet undergone event A, such as BE onset.

6.1.2 Quantifying risk

We may also quantify a certain penalty for certain screening times by introducing an equation for the risk that some later event B has already occurred given some optimal screening time t_s^*:

$$\text{Risk}_{B}(t_s^*) \equiv R_{B}(t_s^*) = \Pr[T_{B} \leq t_s^*] \tag{6.10}$$

We may wish to assign the risk probability of the unfavorable event B occurring before optimal time t_s to simply be the right term of the difference in Eq. (6.6) as we did in the previous section. In this scenario, we would increase the weight w in Eq. (6.7) and find optimal time t_s^* such that $R_{B}(t_s^*) < \epsilon$, where ϵ is a threshold parameter for the allowable amount of risk of event B. Alternatively, if the risk of possibly missing an undesirable event B before optimal time t_s^* is too small to be of worry, one may decide to reduce the weight w assigned to event B so as to relax the requirement that t_s happens early enough to avoid many B events. This will lead to later optimal t_s^* times and thus raise the risk of event B to some appropriate, or tolerable, level.

We may also be interested in performing a constrained optimization problem with additional risk equations. For example, we will introduce adaptive screen design in this chapter (and will continue to focus on this in Chapter 7) that aims to schedule the following screen for a patient after he/she has developed HGD if we knew when exactly he/she first developed BE at time τ, and that there was no positive diagnosis of HGD at present screening time $t_{s_1}, where \tau < t_{s_1}$. This would then allow for proper treatment such as radiofrequency ablation (RFA) to be performed on the BE segment without many costly screens while the BE segment was still free of HGD. However, we may also want to ensure that such a screening

time is associated with a tolerable level of possible risk of cancer developing before time t_{s_2} with the constraint

$$\mathrm{R}_{C}(t_{s_2}) = \Pr[T_{C} \leq t_{s_2} | T_{BE} = \tau, T_{P}^{d}(t_{s_1}) = 0] < \epsilon \tag{6.11}$$

The incorporation of risk assessment is thus another facet of this optimal screening design methodology.

6.1.3 Incorporating realistic time constraints

The methodology thus far has not incorporated cutoff times due to life expectancy, i.e., the model is such that all persons would develop premalignacy through EAC in a long enough time frame. In this section we introduce realistic time frames for a human lifespan into the aforementioned probabilities so that we do not optimize screening ages that are unrealistically old, i.e., when the patient is more probable to have already died than to have developed the disease. We include US life table estimates for all cause mortality survival probabilities in the US [155] as an independent random variable T_{D} in order to calculate $\Pr[T_{D} \geq t_s]$. We can then compute our criteria quantities of interest for events A and B as before, with an extra variable for time of other-cause death T_{D} that is assumed to be independent of events A and B. Thus, this method allows for random truncation due to death from other causes. Beginning with the criteria for $w = 1$,

$$\begin{aligned} \Pr[T_{A} \leq t_s \leq T_{B}, T_{D} \geq t_s] &= \Pr[T_{D} \geq t_s] \cdot \Pr[T_{A} \leq t_s \leq T_{B}] && (6.12) \\ &= \Pr[T_{D} \geq t_s] \cdot (\Pr[T_{A} \leq t_s] - \Pr[T_{B} \leq t_s, T_{A} \leq t_s]) \,. && (6.13) \end{aligned}$$

For the general weighting criteria, we have the expression

$$\Pr[T_{D} \geq t_s] \cdot (\Pr[T_{A} \leq t_s] - w \cdot \Pr[T_{B} \leq t_s, T_{A} \leq t_s]) \,, \tag{6.14}$$

for some specified $w > 0$. We will denote optimal times derived from maximizing Eq. (6.14) as having satisfied the *optimality criterion* for a given value of weighting factor w.

The optimal screen design methodology presented thus far is general to branching process disease models. In the following sections, we derive the probabilities for the optimality criteria defined in Eq. (6.14) for the MSCE-EAC model. We also provide examples for EAC given the MSCE-EAC screening model parameters derived in previous chapters, most importantly the cellular parameters provided in Table 3.1.

6.2 Optimizing Single Screen Times

In this section, we derive three example probabilities, or *strategies*, for the optimization of a single screen in an individual's lifetime within a certain disease state window.

6.2.1 Strategy 1

This strategy for a single screen aims to optimize the probability of an individual having developed BE before the time of screening but has not yet developed clinical EAC, for specified weight parameter w.

$$\Pr[T_{BE} \leq t_s \leq T_C, T_D \geq t_s] \tag{6.15}$$

$$\text{optimality criterion:} \quad \Pr[T_D \geq t_s] \cdot (\Pr[T_{BE} \leq t_s] - w \cdot \Pr[T_C \leq t_s]) \tag{6.16}$$

$$= \Pr[T_D \geq t_s] \cdot (\Pr[T_{BE} \leq t_s] - w \cdot (1 - S_{EAC}(t_s))) \tag{6.17}$$

Figure 6.1 depicts the contour plot of this optimality criterion for males of all races that are born in 1950. We tested these results for males of birth cohorts between 1930 and 1990 and similar results were found, with optimal screening times t_s^* within 1 year between cohorts. Thus, for the strategies that follow, we will present results for the 1950 birth cohort.

6.2.2 Strategy 2

This strategy for a single screen aims to optimize the probability of an individual having developed BE before the time of screening but has not yet developed detectable HGD, for

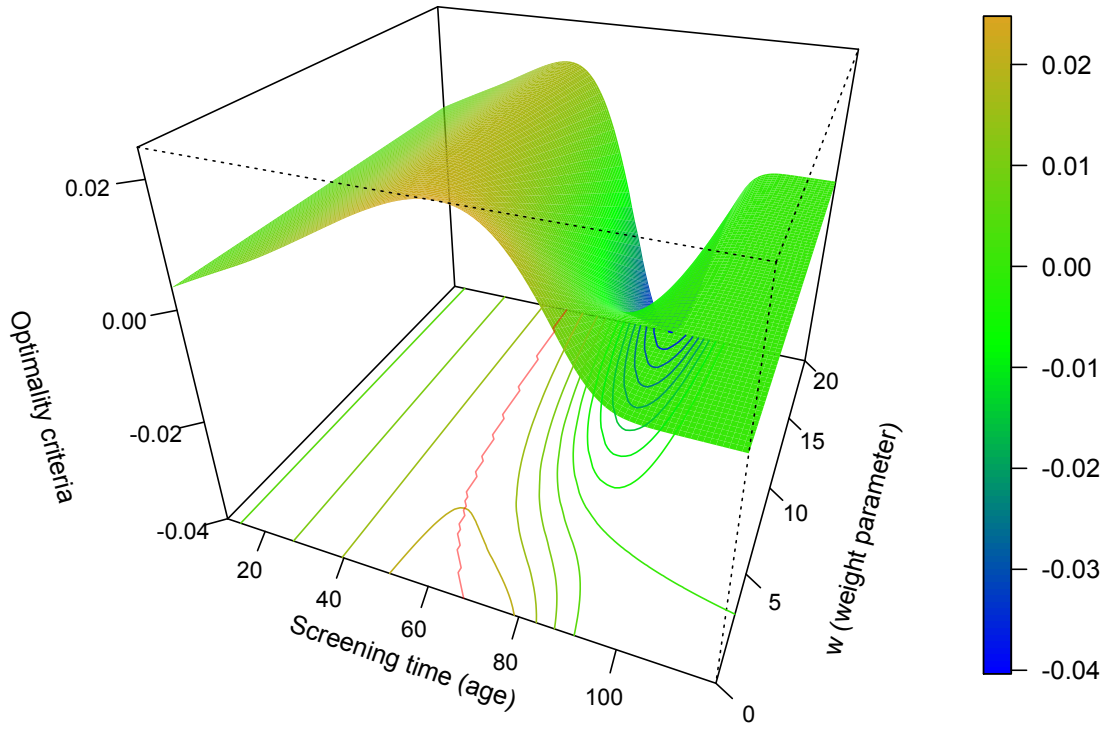

Figure 6.1: **Contour plot of optimality criteria for single screen Strategy 1.** The optimality criteria are plotted for given screening age t_s between ages 10 and 110 and weighting parameters $w \in [0, 20]$. The red line on the 2D contour plot depicts the optimal screen times for males of all races, born in 1950. The range of optimal screening times was [48, 68], with mean time 53 years old. For $w = 1$, the optimal screening time was $t_s^* = 64$ years old.

specified weight parameter w.

$$\Pr[T_{BE} \leq t_s, T_P^d(t_s) = 0, T_D \geq t_s] \tag{6.18}$$

$$\text{optimality criterion:} \quad \Pr[T_D \geq t_s] \cdot (\Pr[T_{BE} \leq t_s] - w \cdot \Pr\left[T_P^d(t_s) = 1\right]) \tag{6.19}$$

$$= \Pr[T_D \geq t_s] \cdot \left(\Pr[T_{BE} \leq t_s] - w \cdot \int_0^{t_s} \Pr\left[T_P^d(t_s) = 1 | T_{BE} = \tau\right] f_{BE}(\tau)\, \mathrm{d}\tau\right) \tag{6.20}$$

We will utilize the filtered Poisson process (FPP) approach to analytically solve for

$\Pr\left[T_P^d(t_s) = 1 | T_{BE} = \tau\right]$. Let $N(\sigma, t_s) = \#$ of initiated P cells from a P^* cell born at time σ that result in a detectable HGD clone at time t_s

$$N(\sigma, t_s) = \sum_l N_l(\sigma, s_l, t_s), \tag{6.21}$$

where $N_l(\sigma, t_s, s_l) = 1$ if initiation at time s_l produces a detectable clone at time t_s, and $s_l \leq t_s$, with probability $p_2(s_l, t_s) = \Pr[T_P^d(t_s) = 1 | P(s_l) = 1]$.
Then $N(\sigma, t_s)$ is a FPP such that

$$N(\sigma, t_s) \sim \text{Poisson}\left(\int_\sigma^{t_s} \mu_1 p_2(s, t_s)\, \mathrm{d}s\right), \tag{6.22}$$

where $p_2(s_l, t_s)$ is defined in Eq. (6.3). Next let $M(\tau, t_s) = \#$ of preinitiated cells after BE onset at time τ that result in a detectable P clone at time t_s. Then we have that

$$M(\tau, t_s) = \sum_i M_i(\tau, \sigma_i, t_s) = 1 \iff N(\sigma_i, t_s) \geq 1 \tag{6.23}$$

with probability $p_3(\sigma_i, t_{s_2}) = 1 - \exp\left[-\int_{\sigma_i}^{t_s} \mu_1 p_2(s, t_s) ds\right]$

Then $M(\tau, t_s)$ is a FPP such that

$$M(\tau, t_s) \sim \text{Poisson}\left(\int_\tau^{t_s} \mu_0 X p_3(\sigma, t_s)\, \mathrm{d}\sigma\right) \tag{6.24}$$

Thus

$$\Pr[T_P^d(t_s) = 1 | T_{BE} = \tau] = 1 - \Pr[T_P^d(t_s) = 0 | T_{BE} = \tau] \tag{6.25}$$

$$= 1 - \exp\left[-\int_\tau^{t_s} \mu_0 X \left(1 - \exp\left[-\int_\sigma^{t_s} \mu_1 p_2(s, t_s)\, \mathrm{d}s\right]\right) \mathrm{d}\sigma\right] \tag{6.26}$$

With this derivation, we have analytically computed all components of the optimality criterion for Strategy 2, given in Eq. (6.20). Figure 6.2 depicts optimal time results for an all race male cohort born in 1950 where the detectability probability of a HGD lesion given initiation time s_l, $p_2(s_l, t_s)$, is defined as a non-extinct progeny of such a clone defined in Eq. (6.3), i.e., the screen has the capability for perfect sensitivity for detecting HGD cells at time of screening.

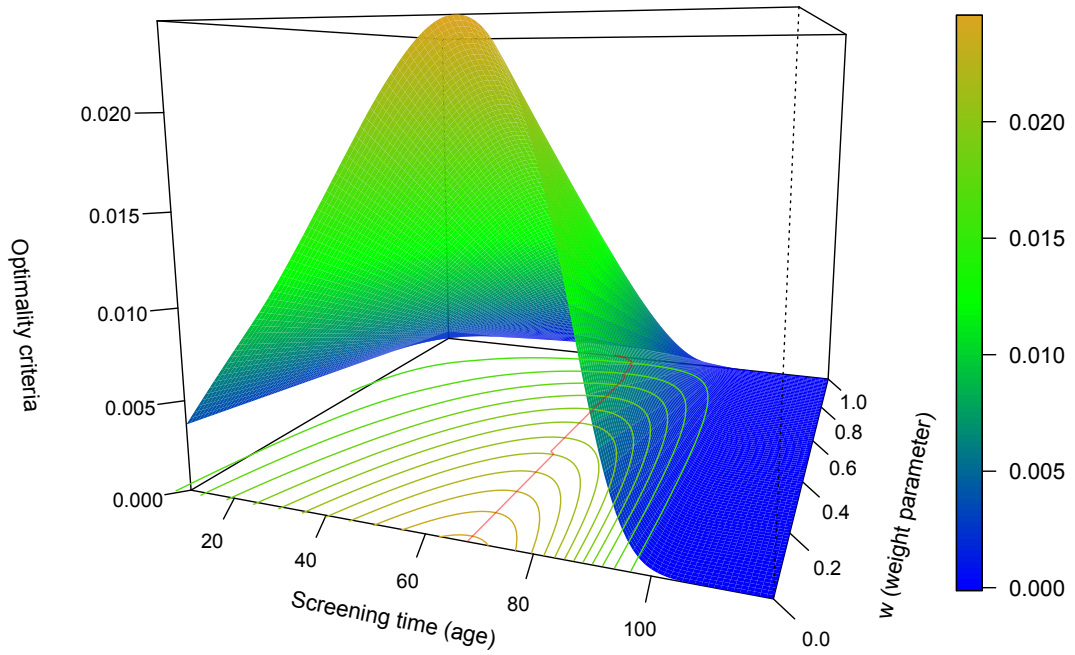

Figure 6.2: **Contour plot of optimality criteria for single screen Strategy 2.** The optimality criteria are plotted for given screening age t_s between ages 10 and 110 and weighting parameters $w \in [0, 1]$. The risk of detecting HGD is closer to that of developing BE than EAC risk is to the risk of BE as in Strategy 1, thus weighting factors $w > 1$ will cause the optimal time chosen to be as early as possible. Therefore we show results for $0 \leq w \leq 1$. The red line on the 2D contour plot depicts the optimal screen times for males of all races, born in 1950. The range of optimal screening times was [59, 68], mean time 67 years old for these values of $w \leq 1$. For $w = 1$, the optimal screening time was $t_s^* = 59$ years old.

6.2.3 Strategy 3

As a final example of optimizing a single screen to occur in a specified disease state window, Strategy 3 aims to optimize the probability of an individual having developed detectable HGD before the time of screening but has not yet developed screen-detectable malignancy, for specified weight parameter w. These optimal ages will reflect the most probable age range

of patients who have developed both BE and detectable HGD but not yet any surviving malignant cells (that could possibly become clinical cancer). This probability and associated optimality criterion are given by

$$\Pr[T_{P}^{d}(t_s) = 1, T_{M}^{d}(t_s) = 0, T_{D} \geq t_s] \tag{6.27}$$

$$\begin{aligned}\text{optimality criterion:} \quad & \Pr[T_{D} \geq t_s] \cdot (\Pr[T_{P}^{d}(t_s) = 1] - w \cdot \Pr[T_{P}^{d}(t_s) = 1, T_{M}^{d}(t_s) = 1]) \\ = \Pr[T_{D} \geq t_s] \cdot \Bigg(& \int_0^{t_s} \Pr[T_{P}^{d}(t_s) = 1 | T_{BE} = \tau] \cdot f_{BE}(\tau)\, \mathrm{d}\tau \\ -w \cdot & \int_0^{t_s} \Pr[T_{P}^{d}(t_s) = 1, T_{M}^{d}(t_s) = 1 | T_{BE} = \tau] \cdot f_{BE}(\tau)\, \mathrm{d}\tau \Bigg)\end{aligned} \tag{6.28}$$

The first integral in this difference was solved for in Strategy 2 (see Eq. (6.26)). Here we solve for the probability in the integrand of the second integral when we assume "perfect sensitivity" for detectability of a malignant lesion and a HGD lesion, given by Eq. (6.2) and Eq. (6.3) for malignant clones and HGD clones, respectively.

$$\Pr[T_{P}^{d}(t_s) = 1, T_{M}^{d}(t_s) = 1 | T_{BE} = \tau] \tag{6.29}$$

To derive this probability, we must define C_1, C_2, C_3, mutually exclusive scenarios of counts of initiations of different types. Scenario C_1 corresponds to the scenario that the P clone that is non-extinct at time t_s is the ancestor of the non-extinct M clone at time t_s. This is most commonly the case in the stochastic realizations since the bigger P clones that remain non-extinct at time of screening have the highest probability to have had μ_2 events (malignant transformation) occur since their initiation. Scenario C_2 accounts for a non-extinct P clone at time t_s that, if it created any malignant progeny before time of screening, those malignancies went extinct. Scenario C_3 accounts for a non-extinct M clone at time t_s whose ancestral P clone went extinct before time t_s. Thus the probability we aim to derive accounts for either event C_1 or C_2 and C_3 together. For each of the three scenarios, we will use filtered Poisson processes techniques and abuse notation slightly by redefining FPP random variables N_1 and N_2 differently but analogously for each scenario, to avoid verboseness. We start with deriving the probability that event C_1 occurs (at least once) with

$u > s$,

$$C_1: \quad \text{Let } N_1(u, t_s, s) = 1 \text{ if } T_{P}^{d}(t_s, s) = 1, T_{M}^{d}(u, t_s, s) = 1 | P(s) = 1, M(u) = 1 \tag{6.30}$$

with probability $p_2^d(s, t_s) \cdot p_1^d(u, t_s) = (1 - \beta_{P}\zeta_P(s, t_s)) \cdot (1 - \beta_{M}\zeta_M(u, t_s))$. Then we have that

$$N_1(t_s, s, \sigma) = \sum_u N_1(u, t_s, s, \sigma) \sim \text{Poisson}\left(p_2^d(s, t_s) \cdot \int_s^{t_s} \mu_2 p_1^d(u, t_s)\, \mathrm{d}u\right) \tag{6.31}$$

and

$$N_2(\sigma, t_s) = \sum_l N_1(t_s, s_l, \sigma) \tag{6.32}$$

so that $N_2(\sigma, t_s)$ is a FPP and the probability generating function (PGF) of $N_2(\sigma, t_s)$ is

$$\Psi(x; \sigma, t_s) = \exp\left[\int_\sigma^{t_s} \mu_1 \cdot \{\Psi(x; s, t_s) - 1\}\, \mathrm{d}s\right] \tag{6.33}$$

where $\Psi(x; s, t_s)$ is the PGF of $N_1(t_s, s, \sigma) = e^{-\lambda(1-x)}$, where $\lambda = p_2^d(s, t_s) \cdot \int_s^{t_s} \mu_2 p_1^d(u, t_s)\, \mathrm{d}u$. Therefore

$$\Pr[C_1 \geq 1] = \Pr[N_2(\sigma, t_s) \geq 1] = 1 - \Pr[N_2(\sigma, t_s) = 0] = 1 - \Psi(x = 0, t_s) \tag{6.34}$$

$$= 1 - \exp\left[\int_\sigma^{t_s} \mu_1 \cdot \left\{e^{-\lambda} - 1\right\}\, \mathrm{d}s\right] \tag{6.35}$$

Similarly,

$$C_2: \quad \text{Let } N_1(u, t_s, s) = 1 \text{ if } T_{P}^{d}(t_s, s) = 1, T_{M}^{d}(u, t_s, s) = 0 | P(s) = 1, M(u) = 1 \tag{6.36}$$

with probability $p_2^d(s, t_s) \cdot (1 - p_1^d(u, t_s)) = (1 - \beta_{P}\zeta_P(s, t_s)) \cdot \beta_{M}\zeta_M(u, t_s)$. Then we have that

$$N_1(t_s, s, \sigma) = \sum_u N_2(u, t_s, s, \sigma) \sim \text{Poisson}\left(p_2^d(s, t_s) \cdot \int_s^{t_s} \mu_2 (1 - p_1^d(u, t_s))\, \mathrm{d}u\right) \tag{6.37}$$

and

$$N_2(\sigma, t_s) = \sum_l N_1(t_s, s_l, \sigma) \tag{6.38}$$

so that $N_2(\sigma, t_s)$ is a FPP and the PGF of $N_2(\sigma, t_s)$ is

$$\Psi_2(x; \sigma, t_s) = \exp\left[\int_\sigma^{t_s} \mu_1 \cdot \{\Psi_2(x; s, t_s) - 1\}\, \mathrm{d}s\right] \tag{6.39}$$

where $\Psi_2(x; s, t_s)$ is the PGF of $N_1(t_s, s, \sigma) = e^{-\lambda_2(1-x)}$, where $\lambda_2 = p_2^d(s, t_s) \cdot \int_s^{t_s} \mu_2(1 - p_1^d(u, t_s))\,\mathrm{d}u$. Therefore

$$\Pr[C_2 \geq 1] = \Pr[N_2(\sigma, t_{s_2}) \geq 1] = 1 - \Pr[N_2(\sigma, t_{s_2}) = 0] = 1 - \Psi_2(x = 0, t_{s_2}) \tag{6.40}$$

$$= 1 - \exp\left[\int_\sigma^{t_s} \mu_1 \cdot \left\{e^{-\lambda_2} - 1\right\}\,\mathrm{d}s\right] \tag{6.41}$$

Lastly, we define

$$C_3: \quad \text{Let } N_1(u, t_s, s) = 1 \text{ if } T_{_P}^d(t_s, s) = 0, T_{_M}^d(u, t_s, s) = 1 | P(s) = 1, M(u) = 1 \tag{6.42}$$

with probability $p_2^d(s,u)\cdot(1-p_2^d(u,t_s))\cdot p_1^d(u,t_s) = (1-\beta_{_P}\zeta_{_P}(s,u))\cdot\beta_{_P}\zeta_{_P}(u,t_s)\cdot(1-\beta_{_M}\zeta_{_M}(u,t_s))$ Then we have that

$$N_1(t_s, s, \sigma) = \sum_u N_1(u, t_s, s, \sigma) \sim \text{Poisson}\left(\int_s^{t_s} \mu_2 \cdot p_2(s,u) \cdot (1 - p_2^d(u, t_s)) \cdot p_1^d(u, t_s)\,\mathrm{d}u\right) \tag{6.43}$$

and

$$N_2(\sigma, t_s) = \sum_l N_1(t_s, s_l, \sigma) \tag{6.44}$$

so that $N_2(\sigma, t_s)$ is a FPP and the PGF of $N_2(\sigma, t_s)$ is

$$\Psi_3(x; \sigma, t_s) = \exp\left[\int_\sigma^{t_s} \mu_1 \cdot \left\{\Psi_3(x; s, t_s) - 1\right\}\mathrm{d}s\right] \tag{6.45}$$

where $\Psi_3(x; s, t_s)$ is the PGF of $N_1(t_s, s, \sigma) = e^{-\lambda_3(1-x)}$, where $\lambda_3 = \int_s^{t_s} \mu_2 \cdot p_2(s,u) \cdot (1 - p_2^d(u, t_s)) \cdot p_1^d(u, t_s)\,\mathrm{d}u$. Therefore

$$\Pr[C_3 \geq 1] = \Pr[N_2(\sigma, t_s) \geq 1] = 1 - \Pr[N_2(\sigma, t_s) = 0] = \Psi_3(x = 0, t_s) \tag{6.46}$$

$$= 1 - \exp\left[\int_\sigma^{t_s} \mu_1 \cdot \left\{e^{-\lambda_3} - 1\right\}\,\mathrm{d}s\right] \tag{6.47}$$

Next, let $M(\tau, t_s)$ =# of preinitiated cells after BE onset at time τ that result in $C_1 \cup (C_2 \cap C_3)$

$$M(\tau, t_s) = \sum_i M_i(\tau, \sigma_i, t_s) \tag{6.48}$$

where $M_i(\tau, \sigma_i, t_s) = 1 \iff C_1 \geq 1 \cup (C_2 \geq 1, C_3 \geq 1)$. Due to the mutual exclusivity of these events, we have that $M_i(\tau, \sigma_i, t_s) = 1$ with probability

$$p_3(\sigma_i, t_s) = \Pr[C_1 \geq 1 \cup (C_2 \geq 1 \cap C_3 \geq 1])] \tag{6.49}$$

$$= \Pr[C_1 \geq 1] + \Pr[C_2 \geq 1] \cdot \Pr[C_3 \geq 1] - \Pr[C_1 \geq 1] \cdot \Pr[C_2 \geq 1] \cdot \Pr[C_3 \geq 1] \tag{6.50}$$

Then $M(\tau, t_s)$ is a FPP such that

$$M(\tau, t_s) \sim \text{Poisson}\left(\int_\tau^{t_s} \mu_2 \cdot p_3(\sigma, t_s)\, \mathrm{d}\sigma\right) \tag{6.51}$$

Finally, we solve for our integrand of interest from Eq. (6.29)

$$\Pr[T_M^d(t_s) = 1, T_P^d(t_s) = 1 | T_{BE} = \tau] \tag{6.52}$$

$$= 1 - \exp\left[-\int_\tau^{t_s} \mu_0 X \cdot p_3(\sigma, t_s)\, \mathrm{d}\sigma\right] \tag{6.53}$$

With this, we have analytically solved for all components of the optimality criterion for Strategy 3 from Eq. (6.28).

6.3 Optimizing Adaptive Screen Times

Rather than finding the optimal time to screen an individual once in his/her lifetime, we may also apply these techniques to derive optimal times to perform a following screen at age t_{s_2}, given the diagnosis at a prior or present screening time t_{s_1}. This adaptive approach allows for patient-specific screening outcome-based risk assessment. In this section, we will provide examples that introduce this methodology, which requires an assessment of the information/diagnosis obtained at the prior screen.

6.3.1 Outcome 1: Strategy 1

For a first outcome, suppose a patient is screened for BE (perhaps he/she has a familial history or GERD symptoms to prompt this decision) and the screen at time t_{s_1} returns

negative for BE. We refer to this scenario as Outcome 1. Analogously to Strategy 1, we aim to screen this individual again at a later age t_{s_2} when he/she is most probable to have developed BE by this time but not yet clinical cancer. For Outcome 1: Strategy 2, we derive the following optimality criterion

$$\Pr[T_{BE} \leq t_{s_2} < T_C, T_D \geq t_{s_2} | T_{BE} > t_{s_1}, T_D > t_{s_1}] \tag{6.54}$$

$$\begin{aligned} \text{optimality criterion:} \quad & \Pr[T_D \geq t_{s_2} | T_D > t_{s_1}] \cdot (\Pr[T_{BE} \leq t_{s_2} | T_{BE} > t_{s_1}] \\ & - w \cdot \Pr[T_C \leq t_{s_2} | T_{BE} > t_{s_1}]) \end{aligned} \tag{6.55}$$

$$\begin{aligned} &= \Pr[T_D \geq t_{s_2} | T_D > t_{s_1}] \cdot \Bigg(\Pr[T_{BE} \leq t_{s_2} | T_{BE} > t_{s_1}] \\ &- w \cdot \frac{1}{\Pr[T_{BE} > t_{s_1}]} \int_{t_{s_1}}^{t_{s_2}} \Pr\left[T_C \leq t_{s_2} | T_{BE} = \tau\right] f_{BE}(\tau) \, \mathrm{d}\tau \Bigg) \\ &= \Pr[T_D \geq t_{s_2} | T_D > t_{s_1}] \cdot \Bigg(\Pr[T_{BE} \leq t_{s_2} | T_{BE} > t_{s_1}] \\ &- w \cdot \frac{1}{\Pr[T_{BE} > t_{s_1}]} \int_{t_{s_1}}^{t_{s_2}} (1 - S_{MSCE}(t_{s_2} - \tau)) f_{BE}(\tau) \, \mathrm{d}\tau \Bigg) \end{aligned} \tag{6.56}$$

Given a range of prior screening times $t_{s_1} \in [10, 30, 50, 60]$, Fig. 6.3 shows these Outcome 1: Strategy 1 probabilities for associated next screening times t_{s_2}, where $t_{s_1} < t_{s_2} \leq 120$. The green diamonds depict the optimal screen times $t^*_{s_2}$ for this example cohort (males of all races, born in 1950). For our choices of t_{s_1}, we found that $t^*_{s_2}$ ranges from [67, 76].

6.3.2 Outcome 1: Strategy 2

This next strategy and derivation is analogous to Strategy 3 from the single screen optimizations. For an individual found without BE at time t_{s_1}, we optimize the probability of an individual having developed detectable HGD before the time of screening but has not yet developed screen-detectable malignancy, for specified weight parameter w. Again, these derivations assume that the binary random variables for neoplastic detection, $T^d_M(t_s)$ and $T^d_P(t_s)$ at some screening time t_s, correspond to a screen with perfect sensitivity that can

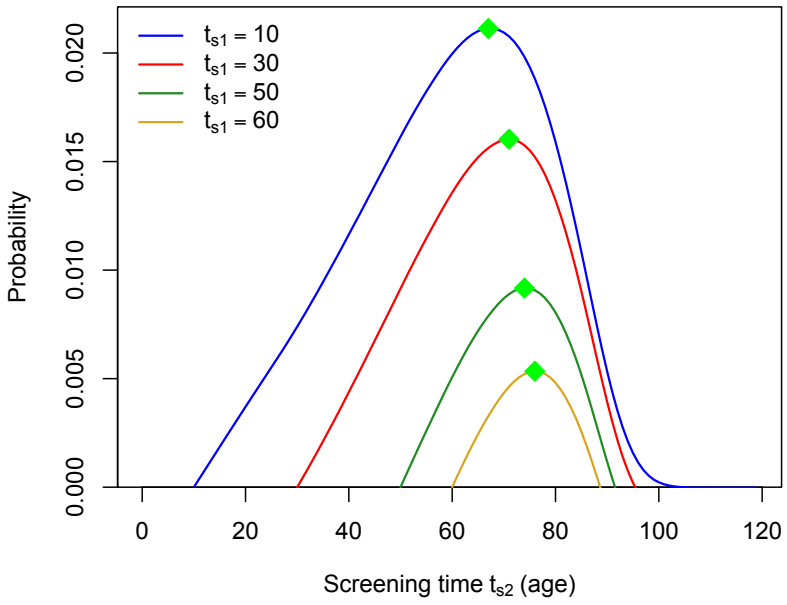

Figure 6.3: **Optimality criteria for Outcome 1: Strategy 1.** The optimality criteria are plotted for each choice of next screening time t_{s_2}, given prior screening age $t_{s_1} \in [10, 30, 50, 60]$ (denoted by color) and weighting parameter $w = 1$, which calculates the true probability of disease state window provided in Eq. (6.54). The optimal screen times $t^*_{s_2}$ for this example cohort males of all races, born in 1950, are plotted by green diamonds and range from [67, 76] for our example choices of t_{s_1}.

detect any non-extinct clone size (see Eqs. (6.2-6.3)).

$$\Pr[T^d_{M}(t_{s_2}) = 0, T^d_{P}(t_{s_2}) = 1, T_{D} \geq t_{s_2} | T_{BE} > t_{s_1}, T_{D} > t_{s_1}] \tag{6.57}$$

$$\begin{aligned} \text{optimality criterion:} \quad & \Pr[T_{D} \geq t_{s_2} | T_{D} > t_{s_1}] \cdot \big(\Pr[T^d_{P}(t_{s_2}) = 1 | T_{BE} > t_{s_1}] \\ & - w \cdot \Pr[T^d_{P}(t_{s_2}) = 1, T^d_{M}(t_{s_2}) = 1 | T_{BE} > t_{s_1}] \big) \end{aligned} \tag{6.58}$$

$$= \Pr[T_{D} \geq t_{s_2} | T_{D} > t_{s_1}] \cdot \left(\frac{1}{\Pr[T_{BE} > t_{s_1}]} \int_{t_{s_1}}^{t_{s_2}} \Pr[T_{P}^{d}(t_{s_2}) = 1 | T_{BE} = \tau] \cdot f_{BE}(\tau) \, \mathrm{d}\tau \right.$$
$$\left. - w \frac{1}{\Pr[T_{BE} > t_{s_1}]} \cdot \int_{t_{s_1}}^{t_{s_2}} \Pr[T_{P}^{d}(t_{s_2}) = 1, T_{M}^{d}(t_{s_2}) = 1 | T_{BE} = \tau] \cdot f_{BE}(\tau) \, \mathrm{d}\tau \right) \tag{6.59}$$

The first integral in this difference was solved for in Strategy 2, and solution is given in Eq. (6.26). As for the probability in the integrand of the second integral, namely,

$$\Pr[T_{P}^{d}(t_{s_2}) = 1, T_{M}^{d}(t_{s_2}) = 1 | T_{BE} = \tau], \tag{6.60}$$

this was also derived previously in Strategy 3 and an expression for it is given in Eq. (6.53) when setting $t_s = t_{s_2}$ throughout the derivation.

6.3.3 Outcome 2

As a second possible outcome, a patient is found to have BE but no neoplasia (HGD/malignancy) at screening time t_{s_1}.

Associated EAC Risk

All choices of a next screening time will have an associated risk of having developed clinical EAC by that time. As introduced in earlier sections, we may also derive the risk of some later disease stage, such as clinical EAC, for the next screening times t_{s_2} that are under consideration for a BE patient to be in a certain earlier disease window. For Outcome 2 at screening time t_{s_1}, this risk for cancer up to the next screening at time t_{s_2} would be defined as

$$\Pr[T_{C} \leq t_{s_2} | T_{BE} < t_{s_1}, T_{P}^{d}(t_{s_1}) = 0, T_{M}^{d}(t_{s_1}) = 0]. \tag{6.61}$$

In general, a doctor will not know how long a patient has harbored undetected BE before some initial screen that first diagnoses him/her, i.e., a doctor will only know $T_{BE} < t_{s_1}$. This is unfortunate because the longer a patient has had BE in his/her esophagus, the higher the associated lifetime risk of EAC will be for this patient. We will explore this problem in the following chapter and propose a model that uses epigenetic data obtained at screening time

t_{s_1} to estimate the patient-specific BE onset age τ, where $\tau \leq t_{s_1}$. If we obtained this BE onset time τ for a patient with Outcome 2 at time t_{s_1}, then his/her associated EAC risk for next screening age t_{s_2} would be more specific than the general risk given in Eq. (6.61) and would be defined as

$$\Pr[T_{_C} \leq t_{s_2} | T_{_{BE}} = \tau, T_{_P}^d(t_{s_1}) = 0, T_{_M}^d(t_{s_1}) = 0]. \tag{6.62}$$

Here we will derive this probability as general lifetime risk without other cause mortality for optimization. Once again, for the scenario with a perfectly sensitive screen that can detect single cells, we may derive this probability using a filtered Poisson approach analogous to the approach used in Strategy 2 for a single screen. Let $N(\sigma, t_{s_2}) = \#$ of initiated P cells from a P^* cell born at time σ that result in a clinically detected EAC clone at time t_{s_2}

$$N(\sigma, t_{s_2}) = \sum_l N_l(\sigma, s_l, t_{s_2}), \tag{6.63}$$

where $N_l(\sigma, t_{s_2}, s_l) = 1$ if initiation at time s_l produces EAC at time t_{s_2}, and $t_{s_1} < s_l < t_{s_2}$, with probability $p_{_C}(s_l, t_{s_2}) = \Pr[T_{_C} < t_{s_2} | P(s_l) = 1]$.
Then $N(\sigma, t_{s_2})$ is a FPP such that

$$N(\sigma, t_{s_2}) \sim \text{Poisson}\left(\int_{\max(\sigma, t_{s_1})}^{t_{s_2}} \mu_1 p_{_C}(s, t_{s_2}) \, \mathrm{d}s \right). \tag{6.64}$$

Clinical EAC cumulative distribution, $p_{_C}(s_l, t_{s_2})$, can be solved via the Kolmogorov equations of the MSCE-EAC branching process, as has been done in previous chapters, from the PDE in Eq. (3.11). Specifically we have that

$$p_{_C}(s, t_{s_2}) = \Pr[T_{_C} < t_{s_2} | P(s) = 1] = 1 - \Pr[T_{_C} \geq t_{s_2} | P(s) = 1] \tag{6.65}$$

$$= 1 - \Phi_{_P}(1, 1, 0; s, t_{s_2}) = 1 - Y_3(s, t_{s_2}) \tag{6.66}$$

where $\Phi_{_P}(1, 1, 0; s, t_{s_2})$ is the EAC survival from a single P cell (see Eq. (3.7)), $Y_3(u, t_{s_2})$ is the associated ODE to solve in the method given in Chapter 3 (see Eq. (3.17)) and $u = t_{s_2} - s$.
Next let $M(\tau, t_{s_2}) = \#$ of preinitiated cells after BE onset at time τ that result in a detectable

P clone at time t_{s_2}. Then we have that

$$M(\tau, t_{s_2}) = \sum_i M_i(\tau, \sigma_i, t_{s_2}) = 1 \iff N(\sigma_i, t_{s_2}) \geq 1 \tag{6.67}$$

with probability $p_3(\sigma_i, t_{s_2}) = 1 - \exp\left[-\int_{\max(\sigma_i, t_{s_1})}^{t_{s_2}} \mu_1 p_{_C}(s, t_{s_2}) ds\right]$

Then $M(\tau, t_{s_2})$ is a FPP such that

$$M(\tau, t_{s_2}) \sim \text{Poisson}\left(\int_\tau^{t_{s_2}} \mu_0 X p_3(\sigma, t_{s_2})\, \mathrm{d}\sigma\right) \tag{6.68}$$

Thus

$$\Pr[T_{_C} \leq t_{s_2} | T_{_{BE}} = \tau, T_{_P}^d(t_{s_1}) = 0, T_{_M}^d(t_{s_1}) = 0] \tag{6.69}$$

$$= 1 - \Pr[T_{_C} > t_{s_2} | T_{_{BE}} = \tau, T_{_P}^d(t_{s_1}) = 0, T_{_M}^d(t_{s_1}) = 0] \tag{6.70}$$

$$= 1 - \exp\left[-\int_\tau^{t_{s_2}} \mu_0 X \left(1 - \exp\left[-\int_{\max(\sigma, t_{s_1})}^{t_{s_2}} \mu_1 p_{_C}(s, t_{s_2})\, \mathrm{d}s\right]\right) \mathrm{d}\sigma\right] \tag{6.71}$$

For example prior screening times $t_{s_1} \in [10, 30, 50, 60]$, Figure 6.4 depicts solutions of Eq. (6.62) for EAC risk for an all race male cohort born in 1950. To note, with these example assumptions on detectability of non-extinct HGD clones, there are other methods besides the filtered Poisson approach that may be used to solve for Eq. (6.71), such as formulating a first-passage time problem. We derive the probability using the filtered Poisson approach for a cohesive methodology. As discussed and shown in Fig. 6.4, the model predicts that associated EAC risk for those that developed BE earlier in life is significantly greater than the risk for those BE patients who developed BE later in life. For the example that a screen is performed at $t_{s_1} = 40$ (purple lines), the male BE patient who developed BE at age 10 (solid line) has a predicted EAC risk by age 85 that was 6.6 times greater than the risk of the BE patient who developed BE at age 40 (dash-dotted line). We will explore using BE onset, $T_{_{BE}}$, as a candidate marker for progression in the next chapter.

Outcome 2: Strategy 1

Next, we aim to find the optimal time at which to screen the BE patient next so that he/she is most probable to have detectable dysplasia but not yet detectable malignancy. At this time,

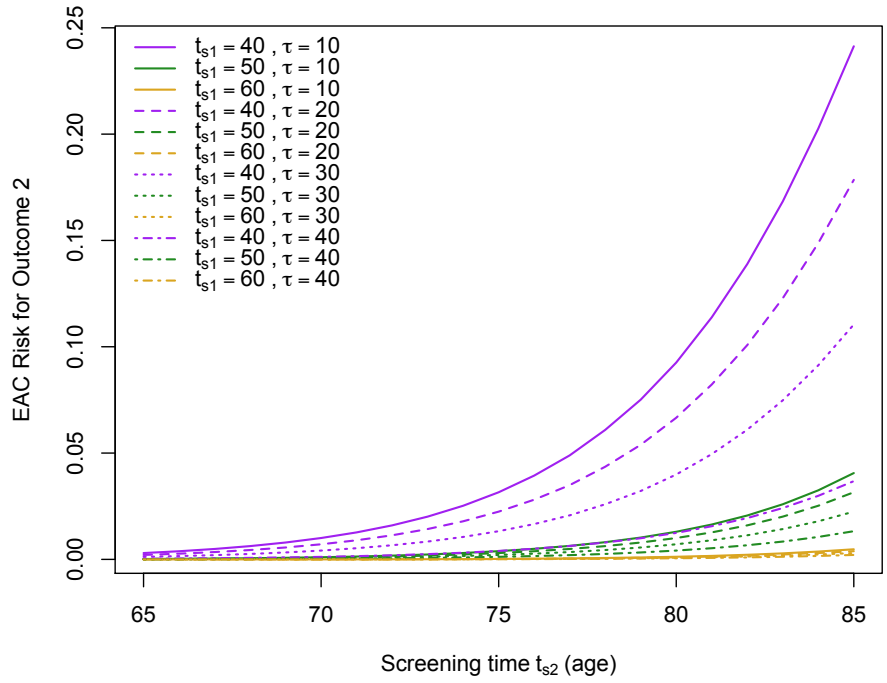

Figure 6.4: **Associated EAC Risk for Outcome 2.** The EAC risk in Eq. (6.62) is plotted for each choice of next screening time t_{s_2} until age 85, given prior screening age $t_{s_1} \in [40, 50, 60]$ (denoted by color) and a BE onset age $\tau \in [10, 20, 30, 40]$ (denoted by line type). The EAC risk varies greatly based on both screening time t_{s_1} for a patient with Outcome 2 and on estimated time of BE onset τ.

the patient can be treated with RFA and/or EMR and greatly reduce his/her cancer risk (see Chapters 4-5 and [14] for modeling examples of treatment impact). To emphasize again, knowing the maximal probability of when the patient will have detectable dysplasia based on a specific screening outcome allows for quantitative assessment of current screening protocols (as introduced in Chapter 5) and addresses the overscreening issue for BE patients. Assuming we have an estimate τ of a patient's BE onset time (possibly derived from epigenetic data),

this problem reduces to optimizing the following probability to determine $t^*_{s_2}$,

$$\Pr[T^d_{M}(t_{s_2}) = 0, T^d_{P}(t_{s_2}) = 1 | T_{BE} = \tau, T^d_{P}(t_{s_1}) = 0, T^d_{M}(t_{s_1}) = 0] \quad (6.72)$$

$$\begin{aligned} \text{optimality criterion:} \quad & \Pr[T^d_{P}(t_{s_2}) = 1 | T_{BE} = \tau, T^d_{P}(t_{s_1}) = 0, T^d_{M}(t_{s_1}) = 0] \\ & - w \cdot \Pr[T^d_{M}(t_{s_2}) = 1, T^d_{P}(t_{s_2}) = 1 | T_{BE} = \tau, T^d_{P}(t_{s_1}) = 0, T^d_{M}(t_{s_1}) = 0] \end{aligned} \quad (6.73)$$

Where w is a weighting variable that equals 1 when Eq. (6.72) is optimized. Again, we can increase or decrease w depending on how much constraint we want to place on keeping the risk of developing malignancy under a certain value, say ϵ. For ease of notation in this strategy, we do not include the other cause mortality term $\Pr[T_D \geq t_{s_2} | T_D > t_{s_1}]$ that can be multiplied to the entire expression of optimality criterion provided in Eq. (6.73) as we have done in previous strategies.

Let us first derive a formula for the first term of this difference based on the MSCE-EAC model construction using again a filtered Poisson process approach,

$$\Pr[T^d_{P}(t_{s_2}) = 1 | T_{BE} = \tau, T^d_{P}(t_{s_1}) = 0, T^d_{M}(t_{s_1}) = 0] \quad (6.74)$$

Let $N(\sigma, t_{s_1}, t_{s_2}) = \#$ of initiated cells from a P^* cell born at time σ result in a detectable clone at time t_{s_2}, not born before t_{s_1}

$$N(\sigma, t_{s_1}, t_{s_2}) = \sum_l N_l(\sigma, t_{s_1}, t_{s_2}, s_l) \quad (6.75)$$

where $N_l(\sigma, t_{s_1}, t_{s_2}, s_l) = 1$ if initiation at time s_l produces a detectable clone at time t_{s_2}, $s_l > t_{s_1}$ with probability $p_2(s_l, t_{s_2}) = \Pr[T^d_{P}(t_{s_2}) = 1, T^d_{P}(t_{s_1}) = 0 | P(s_l) = 1] > 0$ if and only if $s_l > t_{s_1}$.

Then $N(\sigma, t_{s_1}, t_{s_2})$ is a FPP such that

$$N(\sigma, t_{s_1}, t_{s_2}) \sim \text{Poisson}\left(\int_{\max(\sigma, t_{s_1})}^{t_{s_2}} \mu_1 p_2(s, t_{s_2}) \, \mathrm{d}s \right). \quad (6.76)$$

Assuming detectability of P cells based on non-extinction at time of screening, we have the following probability (see Eq. (6.3)),

$$p_2(s, t_{s_2}) = 1 - \beta_P \zeta_P(s, t_{s_2}), \qquad t_{s_1} < s < t_{s_2} \quad (6.77)$$

where

$$\zeta_{P}(s,t_{s_2}) = \frac{e^{(\alpha_{P}-\beta_{P})(t_{s_2}-s)}-1}{\alpha_{P}e^{(\alpha_{P}-\beta_{P})(t_{s_2}-s)}-\beta_{P}} \tag{6.78}$$

The asymptotic version of this probability would be $p_2(s,t_{s_2}) = 1-\beta_{P}/\alpha_{P}$. We will assume the same for malignant clones, with detection probability provided in Eq. (6.2).

Next let $M(\tau,t_{s_2}) = \#$ of preinitiated cells after BE onset at time τ that result in a detectable P clone at time t_{s_2}. Then we have that

$$M(\tau,t_{s_2}) = \sum_{i} M_i(\tau,\sigma_i,t_{s_2}) = 1 \iff N(\sigma_i,t_{s_2}) \geq 1 \tag{6.79}$$

with probability $p_3(\sigma_i,t_{s_2}) = 1-\exp\left[-\int_{\max(\sigma_i,t_{s_1})}^{t_{s_2}} p_2(s,t_{s_2})ds\right]$

Then $M(\tau,t_{s_2})$ is a FPP such that

$$M(\tau,t_{s_2}) \sim \text{Poisson}\left(\int_{\tau}^{t_{s_2}} \mu_0 X p_3(\sigma,t_{s_2})\,\mathrm{d}\sigma\right) \tag{6.80}$$

Thus

$$\Pr[T_{P}^{d}(t_{s_2}) = 1 | T_{BE} = \tau, T_{P}^{d}(t_{s_1}) = 0, T_{M}^{d}(t_{s_1}) = 0] \tag{6.81}$$

$$= 1-\exp\left[-\int_{\tau}^{t_{s_2}} \mu_0 X \left(1-\exp\left[-\int_{\max(\sigma,t_{s_1})}^{t_{s_2}} \mu_1 p_2(s,t_{s_2})\,\mathrm{d}s\right]\right)\,\mathrm{d}\sigma\right] \tag{6.82}$$

For the second term in Eq. (6.73), we will now derive

$$\Pr[T_{M}^{d}(t_{s_2}) = 1, T_{P}^{d}(t_{s_2}) = 1 | T_{BE} = \tau, T_{P}^{d}(t_{s_1}) = 0, T_{M}^{d}(t_{s_1}) = 0] \tag{6.83}$$

We define similar C_1, C_2, C_3 mutually exclusive scenarios of counts of initiations of different types as was utilized in single screen Strategy 3. For each of the three scenarios, we will use filtered Poisson processes techniques and abuse notation slightly by redefining FPP random variables N_1 and N_2 variables differently but analogously for each scenario, to avoid verboseness. We start with deriving the probability that event C_1 occurs (at least once),

$$C_1: \quad \text{Let } N_1(u,t_{s_2},s) = 1 \text{ if } T_{P}^{d}(t_{s_2},s) = 1, T_{M}^{d}(u,t_{s_2},s) = 1 | P(s) = 1, M(u) = 1 \tag{6.84}$$

with probability $p_2^d(s,t_{s_2}) \cdot p_1^d(u,t_{s_2}) = (1-\beta_{_P}\zeta_{_P}(s,t_{s_2})) \cdot (1-\beta_{_M}\zeta_{_M}(u,t_{s_2}))$. Then we have that

$$N_1(t_{s_2},s,\sigma) = \sum_u N_1(u,t_{s_2},s,\sigma) \sim \text{Poisson}\left(p_2^d(s,t_{s_2}) \cdot \int_s^{t_{s_2}} \mu_2 p_1^d(u,t_{s_2})\,\mathrm{d}u\right) \tag{6.85}$$

and

$$N_2(\sigma,t_{s_2}) = \sum_l N_1(t_{s_2},s_l,\sigma) \tag{6.86}$$

so that $N_2(\sigma,t_{s_2})$ is a FPP and the PGF of $N_2(\sigma,t_{s_2})$ is

$$\Psi(x;\sigma,t_{s_1},t_{s_2}) = \exp\left[\int_{\max(\sigma,t_{s_1})}^{t_{s_2}} \mu_1 \cdot \{\Psi(x;s,t_{s_2})-1\}\,\mathrm{d}s\right] \tag{6.87}$$

where $\Psi(x;s,t_{s_2})$ is the PGF of $N_1(t_{s_2},s,\sigma) = e^{-\lambda(1-x)}$ where $\lambda = p_2^d(s,t_{s_2}) \cdot \int_s^{t_{s_2}} \mu_2 p_1^d(u,t_{s_2})\,\mathrm{d}u$. Therefore

$$\Pr[C_1 \geq 1] = \Pr[N_2(\sigma,t_{s_2}) \geq 1] = 1 - \Pr[N_2(\sigma,t_{s_2}) = 0] = 1 - \Psi(x=0,t_{s_2}) \tag{6.88}$$

$$= 1 - \exp\left[\int_{\max(\sigma,t_{s_1})}^{t_{s_2}} \mu_1 \cdot \left\{e^{-\lambda}-1\right\}\,\mathrm{d}s\right] \tag{6.89}$$

Similarly,

$$C_2: \quad \text{Let } N_1(u,t_{s_2},s) = 1 \text{ if } T_{_P}^d(t_{s_2},s) = 1, T_{_M}^d(u,t_{s_2},s) = 0 | P(s)=1, M(u)=1 \tag{6.90}$$

with probability $p_2^d(s,t_{s_2}) \cdot (1-p_1^d(u,t_{s_2})) = (1-\beta_{_P}\zeta_{_P}(s,t_{s_2})) \cdot \beta_{_M}\zeta_{_M}(u,t_{s_2})$. Then we have that

$$N_1(t_{s_2},s,\sigma) = \sum_u N_2(u,t_{s_2},s,\sigma) \sim \text{Poisson}\left(p_2^d(s,t_{s_2}) \cdot \int_s^{t_{s_2}} \mu_2(1-p_1^d(u,t_{s_2}))\,\mathrm{d}u\right) \tag{6.91}$$

and

$$N_2(\sigma,t_{s_2}) = \sum_l N_1(t_{s_2},s_l,\sigma) \tag{6.92}$$

so that $N_2(\sigma,t_{s_2})$ is a FPP and the PGF of $N_2(\sigma,t_{s_2})$ is

$$\Psi_2(x;\sigma,t_{s_1},t_{s_2}) = \exp\left[\int_{\max(\sigma,t_{s_1})}^{t_{s_2}} \mu_1 \cdot \{\Psi_2(x;s,t_{s_2})-1\}\,\mathrm{d}s\right] \tag{6.93}$$

where $\Psi_2(x; s, t_{s_2})$ is the PGF of $N_1(t_{s_2}, s, \sigma) = e^{-\lambda_2(1-x)}$ where $\lambda_2 = p_2^d(s, t_{s_2}) \cdot \int_s^{t_{s_2}} \mu_2(1 - p_1^d(u, t_{s_2}))\,\mathrm{d}u$. Therefore

$$\Pr[C_2 \geq 1] = \Pr[N_2(\sigma, t_{s_2}) \geq 1] = 1 - \Pr[N_2(\sigma, t_{s_2}) = 0] = 1 - \Psi_2(x = 0, t_{s_2}) \tag{6.94}$$

$$= 1 - \exp\left[\int_{\max(\sigma, t_{s_1})}^{t_{s_2}} \mu_1 \cdot \left\{e^{-\lambda_2} - 1\right\}\,\mathrm{d}s\right] \tag{6.95}$$

Lastly, we define

$$C_3: \quad \text{Let } N_1(u, t_{s_2}, s) = 1 \text{ if } T_{P}^{d}(t_{s_2}, s) = 0, T_{M}^{d}(u, t_{s_2}, s) = 1 | P(s) = 1, M(u) = 1 \tag{6.96}$$

with probability $p_2^d(s, u) \cdot (1 - p_2^d(u, t_{s_2})) \cdot p_1^d(u, t_{s_2}) = (1 - \beta_{P}\zeta_{P}(s, u)) \cdot \beta_{P}\zeta_{P}(u, t_{s_2}) \cdot (1 - \beta_{M}\zeta_{M}(u, t_{s_2}))$. Then we have that

$$N_1(t_{s_2}, s, \sigma) = \sum_u N_1(u, t_{s_2}, s, \sigma) \sim \text{Poisson}\left(\int_s^{t_{s_2}} \mu_2 \cdot p_2(s, u) \cdot (1 - p_2^d(u, t_{s_2})) \cdot p_1^d(u, t_{s_2})\,\mathrm{d}u\right) \tag{6.97}$$

and

$$N_2(\sigma, t_{s_2}) = \sum_l N_1(t_{s_2}, s_l, \sigma) \tag{6.98}$$

so that $N_2(\sigma, t_{s_2})$ is a FPP and the PGF of $N_2(\sigma, t_{s_2})$ is

$$\Psi_3(x; \sigma, t_{s_1}, t_{s_2}) = \exp\left[\int_{\max(\sigma, t_{s_1})}^{t_{s_2}} \mu_1 \cdot \{\Psi_3(x; s, t_{s_2}) - 1\}\,\mathrm{d}s\right] \tag{6.99}$$

where $\Psi_3(x; s, t_{s_2})$ is the PGF of $N_1(t_{s_2}, s, \sigma) = e^{-\lambda_3(1-x)}$ where $\lambda_3 = \int_s^{t_{s_2}} \mu_2 \cdot p_2(s, u) \cdot (1 - p_2^d(u, t_{s_2})) \cdot p_1^d(u, t_{s_2})\,\mathrm{d}u$. Therefore

$$\Pr[C_3 \geq 1] = \Pr[N_2(\sigma, t_{s_2}) \geq 1] = 1 - \Pr[N_2(\sigma, t_{s_2}) = 0] = \Psi_3(x = 0, t_{s_2}) \tag{6.100}$$

$$= 1 - \exp\left[\int_{\max(\sigma, t_{s_1})}^{t_{s_2}} \mu_1 \cdot \left\{e^{-\lambda_3} - 1\right\}\,\mathrm{d}s\right] \tag{6.101}$$

Next, let $M(\tau, t_{s_2}) =$# of preinitiated cells after BE onset at time τ that result in $C_1 \cup (C_2 \cap C_3)$

$$M(\tau, t_{s_2}) = \sum_i M_i(\tau, \sigma_i, t_{s_2}) \tag{6.102}$$

where $M_i(\tau, \sigma_i, t_{s_2}) = 1 \iff C_1 \geq 1 \cup C_2 \geq 1, C_3 \geq 1$. Due to the mutual exclusivity of these events, we have that $M_i(\tau, \sigma_i, t_{s_2}) = 1$ with probability

$$p_3(\sigma_i, t_{s_2}) = \Pr[C_1 \geq 1 \cup (C_2 \geq 1 \cap C_3 \geq 1])] \tag{6.103}$$

$$= \Pr[C_1 \geq 1] + \Pr[C_2 \geq 1] \cdot \Pr[C_3 \geq 1] - \Pr[C_1 \geq 1] \cdot \Pr[C_2 \geq 1] \cdot \Pr[C_3 \geq 1] \tag{6.104}$$

Then $M(\tau, t_{s_2})$ is a FPP such that

$$M(\tau, t_{s_2}) \sim \text{Poisson}\left(\int_{\tau}^{t_{s_2}} \mu_2 \cdot p_3(\sigma, t_{s_2})\, \mathrm{d}\sigma\right) \tag{6.105}$$

With these, we have analytically solved for our expression of interest from Eq. (6.83),

$$\Pr[T_M^d(t_{s_2}) = 1, T_P^d(t_{s_2}) = 1 | T_{BE} = \tau, T_P^d(t_{s_1}) = 0, T_M^d(t_{s_1}) = 0] \tag{6.106}$$

$$= 1 - \exp\left[-\int_{\tau}^{t_{s_2}} \mu_0 X \cdot p_3(\sigma, t_{s_2})\, \mathrm{d}\sigma\right], \tag{6.107}$$

and thus have both components to compute the optimality criterion for Outcome 2: Strategy 1 from Eq. (6.73). This is the final example of outcomes and strategies for adaptive screening provided in this chapter although more variations for other outcomes may be computed in a similar fashion.

6.4 Concluding Remarks

In this chapter, we derived probabilities for some example scenarios using the MSCE-EAC model framework in order to introduce the methodology for optimal screen design for maximal screen yield, and included a few results using calibrated model parameters. This framework for optimal design is a generalized method of weighting cost (in terms of occurrence of an unfavorable event) and risk when optimizing a single screen or an adaptive screen based on the outcome of a previous screen. Previously Dewanji et al. considered a single screen scenario. Here we developed a generalization of the optimality criterion introduced by Dewanji et al. and expanded the method to include the adaptive screen design component and

a flexibility in weighting the risk of undesirable events [158]. With these expressions, we may use the MSCE-EAC model to predict the inefficiencies of current surveillance scheduling. The decisions of which clinical windows are most important to perform screens within and the associated permissible risks are left to physician and health economist discretion.

Chapter 7

Barrett's Esophagus: "How long has that been there?"

There has been intense interest in using biomarkers based on genetic and epigenetic alterations in Barrett's esophagus (BE) to more accurately determine the risk of BE progressing to esophageal adenocarcinoma (EAC). Both DNA hypo- and hypermethylation in the promoters of various genes have been implicated during early stages of EAC progression leading to significant changes in transcriptional activity of important tumor suppressor genes and oncogenic pathways [160–164]. At other genomic locations methylation levels appear to gradually drift with age [165]. Recently identified *clock CpGs* provide a novel characterization of a tissue in terms of its biological age, and these markers were used to show accelerated tissue aging in a variety of tissues and tumors [166, 167]. The main goal of this study is to determine whether CpGs that drift differentially between BE and a normal reference tissue can be used to more precisely determine the biological age of a patient's BE tissue to more reliably determine the patient's risk to progress to dysplasia or cancer.

As emphasized throughout previous chapters, the risk of progressing from BE to EAC appears to be relatively low, on average. It is estimated that only 0.2-0.5% of patients with BE develop EAC each year [12], yet the cancer-specific mortality is high with the 5-year survival mark below 15% for invasive EAC [168]. Thus, there is a pressing need to develop

accurate markers that can identify BE that is more likely to progress to EAC in a person's lifetime and BE that is indolent or has low neoplastic potential. The assumption of a constant BE-to-EAC progression rate implies that a BE patient who has had BE for n-times as long as another BE patient of the same age is expected to have an n-fold higher cumulative risk of developing EAC. However, the annual BE-to-EAC progression rate is unlikely constant, but is expected to increase with age of the tissue (dwell time) due to the step-wise accumulation of genetic and epigenetic alterations that drive premalignant and malignant progression in BE over time [164, 169, 170]. Therefore, a longer dwell time for BE is likely to increase the risk for neoplasia and cancer in an exponential manner. Many-fold higher risks result from the participation of clonal expansion processes of neoplastic precursors and malignant tumors as reflected by the exponential increases in the age-specific incidence of EAC in the general population [6].

High-throughput methylation arrays and quantitative sequencing assays have recently become available to quantitatively assess CpG-methylation patterns genome-wide [171–174]. Although, at the molecular level, CpG-methylation is essentially a binary variable (a CpG site is either methylated or unmethylated), degree of methylation is typically reported on the array as a fraction (β-value) of methylated probes relative to the sum of methylated and unmethylated probes ($\beta = M/(M+U)$), representing a population average across individual epigenomes in a tissue sample. Moreover, it has been demonstrated that the methylation state at a specific CpG site can vary considerably from one cell generation to the next due to the processive, error prone nature of the DNA methylation maintenance system during DNA replication [175–177]. To identify CpGs that undergo methylomic drift that can be correlated with BE tissue age, we generated data from formalin fixed paraffin embedded (FFPE) tissue samples from two groups of BE patients: one group of 10 patients each with 2 or more tissue samples at least 5 years apart (data set D1) provides longitudinal information at the individual level, while the second group of 30 patients (data set D2) provides cross-sectional information as well as differential drift information between normal squamous (NS) and BE tissue. The combined statistical analyses of these two data sets, as

described in Materials and Methods, suggests that at least ∼70 hypomethylated CpG sites undergo significant differential methylomic drift in BE vs NS. The observed drift appears relatively uniform across the set of identified CpGs, giving rise to high correlations in the methylation differentials (BE vs NS) between CpGs. Thus, a hallmark of methylomic drift CpGs is that they are highly correlated (as all clocks are that keep chronological time) even if the markers belong to independent functional genomic or metabolomic pathways.

To infer the patient-specific BE onset times from bivariate BE methylation levels of these drift CpGs, we use a Bayesian model which accounts for (CpG-specific) random effects in drift rates and for inter-individual heterogeneity in the onset times. To gain insights into how the age of BE onset influences EAC risk we also use a recently developed mathematical model for EAC incidence to compute standardized life-time risks for the individuals in data set D2 given their predicted BE onset times [6, 14]. Additionally, we apply the methodology developed here to methylation array data from familial BE (FBE) patients (data set D3) as well as to patients who developed high grade dysplasia (HGD) and/or cancer (data set D4). The quantitative comparison of both BE onset times and inferred EAC risks for BE patients without neoplasia (D2) versus familial BE (D3) and BE patients with neoplasia (D4) will suggest if *BE onset* is a powerful event-marker of cancer risk or not. In the following we describe the data and methodology used in this current study.

7.1 Materials and Methods

All CpG-methylation data for this study were generated with the Illumina Infinium HumanMethylation450 beadchip [178, 179] which includes over 485,000 CpG-methylation sites throughout the genome (covering 99% of Reference Sequence (RefSeq) genes (National Center for Biotechnology Information (NCBI), Bethesda, MD, USA)). Data normalization was performed using the Bioconductor *minfi* package, which includes Illumina background level correction, color adjustment and Subset-quantile Within Array Normalization (SWAN) normalization. For linear regressions of the probe-specific methylation fractions on patient age

we used M-values rather than β-values to better account for epigenetic drift that occurs at very low ($< 1\%$) and high levels of methylation. M-values are logit-transformed β-values (where $\beta = M/(M+U)$) allowing for non-linear saturation effects of methylation fractions with age at both ends of the methylation spectrum.

7.1.1 Patient data

We examined levels of DNA methylation at over 450,000 CpG sites in tissue samples from four groups of BE patients (see Tables D.1-D.3 for detailed patient information). The first data set (D1) is unique and consists of longitudinal samples from 10 BE patients ages 33-70 years at index biopsy (mean age = 53.8) with 2 or more tissue biopsies that were collected at least 5 years apart comprising a total of 30 samples. An example of patient data for two particular CpGs that show longitudinal drift for each of these 10 patients' serial sample sets is shown in Fig. 7.1.

The second, cross-sectional data set (D2) includes matched BE and normal squamous (NS) tissue samples from 30 BE patients ages 21-88 years (mean age = 63.4) comprising a total of 60 tissue samples. While the longitudinal patients in D1 provide information on individual BE tissue methylation drift, the aggregated cross-sectional BE methylation values at various ages provide information on a population scale of drift with age. This apparent drift could be caused by various factors and we will focus on the potential influence of varying BE onset ages, i.e., the time at which differential drift in BE first began, as one such factor. Fig. 7.2 denotes two example D2 patients' ages at time of biopsy ($a_1 = 21, a_2 = 80$), and the theoretical consequence their ages will have on the statistical inference of their BE onset ages. The inter-individual heterogeneity of such conditional distributions for BE onset times will thus affect the methylation level data around a population mean drift. An illustration for a single CpG site for the BE samples from D2 is shown in the insert of Fig. 7.2. Note, for the cross-sectional group (D2), the matched BE and NS samples originate from biopsies collected during the same endoscopy.

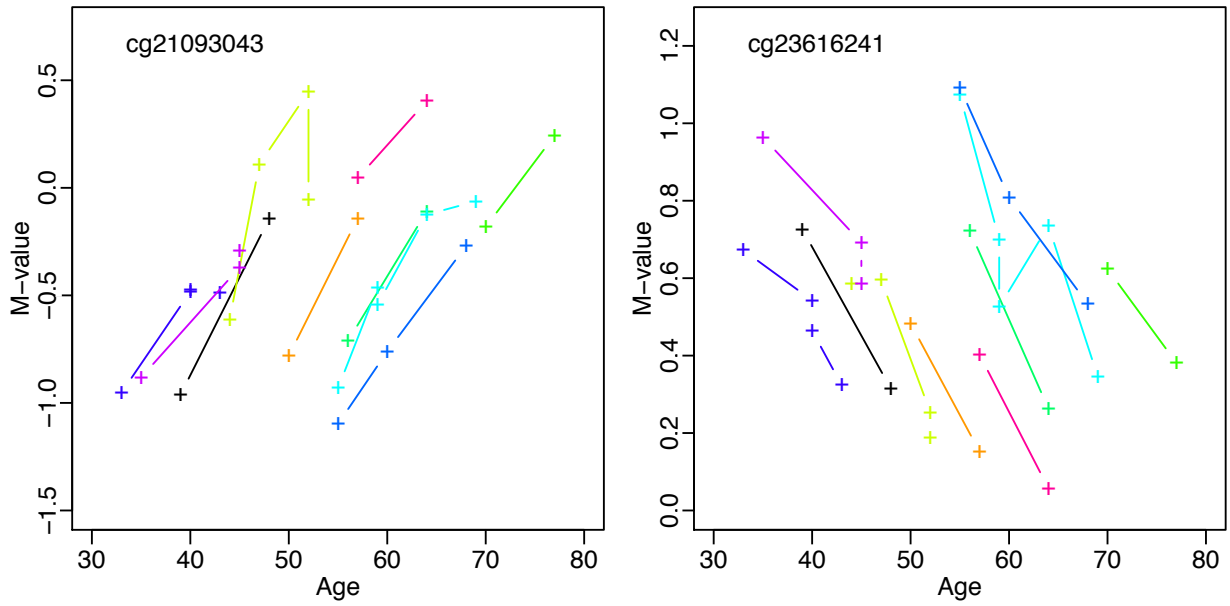

Figure 7.1: **Longitudinal drift from set D1.** Examples of one CpG (cg21093043) that significantly drifts up (left panel, becomes increasingly hypermethylated) and one CpG (cg23616241) that significantly drifts down (right panel, becomes increasingly hypomethylated) among longitudinal data points. Each individual from data set D1 provides serial samples denoted by color.

The third data set (D3) includes BE tissue samples from 22 familial BE (FBE) patients ages with one sample per patient, and the fourth set (D4) includes BE tissues from 10 BE patients with neoplasia (HGD and/or EAC). Familial Barrett's esophagus (FBE) was defined as having a first- or second-degree relative with long-segment BE, adenocarcinoma of the esophagus, or adenocarcinoma of the gastroesophageal junction whose diagnosis was confirmed by review of endoscopy and histology reports [180]. The data also includes gender and age of each patient when the tissue biopsies were collected (see Tables D.1-D.3).

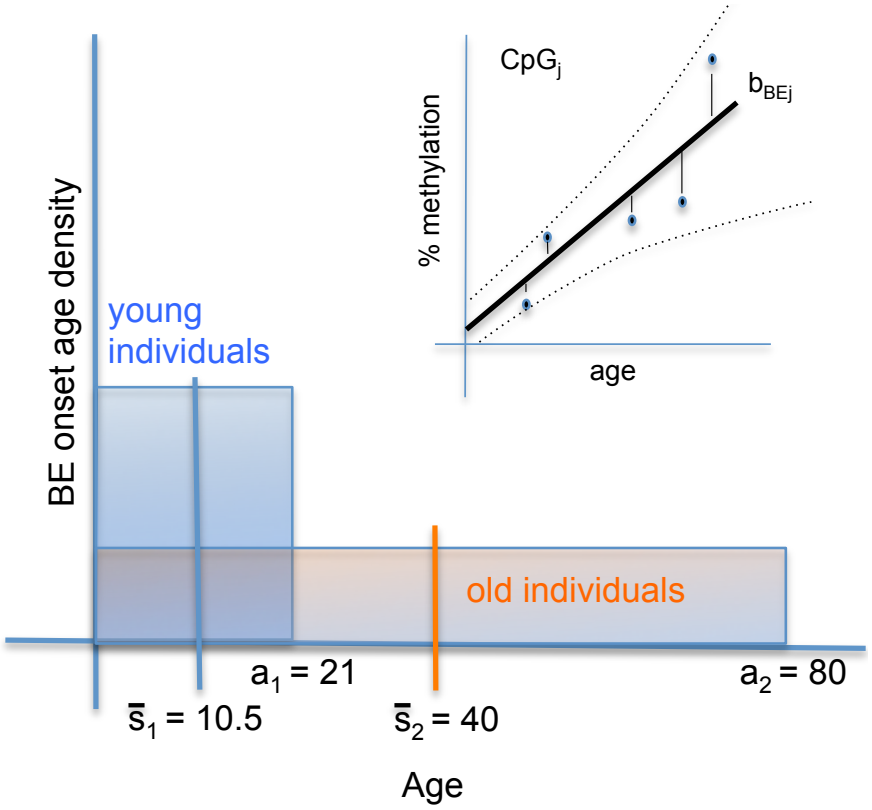

Figure 7.2: **Population drift from set D2.** The heterogeneity with age around the population mean drift slope may be caused by the inter-individual heterogeneity of BE onset times. Illustrational of cross-sectional BE data for a certain CpG_j shown in upper right inset. Assuming uniform distributions for BE onset times (conditional on age), two example patients from data set D2, who had biopsies taken at index endoscopy ages $a_1 = 21$ and $a_2 = 80$, would have hypothetical mean BE onset times at ages $\bar{s}_1 = 10.5$ and $\bar{s}_2 = 40$, respectively.

7.1.2 Identification of markers of differential epigenetic drift

Two concepts have fairly recently emerged that relate alterations in DNA-methylation to biological tissue age. The first is based on discovery of sets of clock-CpGs that may undergo complex age-dependent changes in methylation that correlate strongly with chronological age

and neoplastic progression [166, 167, 181]. The second concept relates to subtle changes in methylation levels due to epigenetic drift as a result of a semi-conserved replication process of DNA-methylation patterns [182–185]. Significantly, some CpG-islands that show very low (hypo-)methylation levels early in life are known to undergo gradual methylation over time, presumably as a result of sporadic *de novo* methylation events during DNA replication, a process commonly understood as *epigenetic* or *methylomic drift* [182, 186–190]. Therefore, to identify potential CpG candidates that may serve as markers for differential tissue aging in the emerging metaplastic tissue of BE patients relative to their normal squamous (NS) tissue, we used linear regression to identify CpGs in BE tissue samples that show significant longitudinal drift in study D1 as well as consistent cross-sectional drift among patients in D2. From this first selection of age-related drift markers, we retained CpGs that are either hypomethylated ($\beta_{NS} < 0.25$) or hypermethylated ($\beta_{NS} > 0.75$) in NS tissue and show significant differential positive or negative drift between NS and BE tissues, respectively, for the markers in these two categories. The following steps summarize our discovery pipeline in more detail.

Step 1: Identify BE drift-CpGs using longitudinal data

To identify CpGs that show consistent drift across all patients in D1, we examined the relationship between incremental changes in methylation levels (M-value) and time since first biopsy for all D1 patients as shown in Fig. 7.1. To this end, we first translated each patient's age by the age of his/her first biopsy and subsequently adjusted all M-values by the patient's predicted M-value at the time of the first biopsy using linear regression. If a patient had only 2 biopsies, we used a connecting line through the two data points. Next, we performed linear regressions on the normalized D1 data in aggregate to identify candidate CpGs that undergo concordant drift between individuals and to determine marker-specific drift rates, $b_j, j = 1, .., M$ from the regression slopes of M markers. To ensure that CpGs that appear to drift incrementally (with time-since-first-biopsy) also drift cross-sectionally with age, we simultaneously tested each CpG for concomitant cross-sectional drift across

all samples in D1 using a nominal p-value = 0.01 while a false discovery rate (FDR) of $q = 0.20$ was applied for the incremental drift analysis. This testing identifies 2,944 CpGs out of a total of 456,385 CpGs that drift upward and 1,774 CpGs that drift downward across the 10 D1 patients. We deliberately kept a permissive FDR of 0.2 to avoid excessive pruning of informative markers due to the relative small number of observations in D1. Note, this selection procedure ignores the inter-individual variability in the estimated drift rates, $b_j, j = 1, .., M$ for M number of CpGs, but generates an approximate "standard clock" rate for each drift CpG so that individual-level methylomic drift can be translated into biological age information. This information will be used for generating patient-specific drift rates, details forthcoming.

Step 2: Identify NS vs BE differential drift in cross-sectional data

Next, we examined which of the candidate CpGs identified in Step 1 show significant differential drift between the matched NS and BE tissues of data set D2. We used Analysis of Covariance (ANCOVA) regression modeling to test whether the methylomic drift rates (i.e., M-value regression slopes) differed between NS and BE tissues among the 30 patients in set D2. We further subdivide the CpGs into two subgroups: those that are essentially hypomethylated in NS tissue with $\beta_{NS} < 0.25$ (475 CpGs), and those that can be considered hypermethylated in NS tissue, i.e., $\beta_{NS} > 0.75$ (335 CpGs) for all NS samples in D2. As we will show, this categorization distinguishes positive and negative methylomic drift in BE tissue, respectively for hypo- and hypermethylated CpGs in the reference NS tissue. This particular choice is less confounded by heterozygous methylation where drift could, at least in principle, occur in opposite directions (e.g., when the paternal allele is unmethylated, but the maternal allele is methylated). Using ANCOVA, we find 83/475 CpGs to differentially drift between BE and NS in the first group (nominal p-values < 0.05), while only 21/335 CpGs appear to differentially drift between the two tissues in the second group. As expected, the majority (70) of the 83 differential, upward drifting CpGs have estimated BE drift rates that are in fact larger than their corresponding NS drift rates, while only 8 out of 21 differ-

ential downward drifting CpGs appear to have estimated BE drift rates that are lower than their corresponding NS drift rates. Thus, we will continue our analysis and selection using the larger subset of 83 positively drifting CpGs.

A principal component analysis (PCA) of residuals from the BE methylation age regression (which are hypothesized to reflect BE tissue age differences) for the selected 83 differential drift CpGs confirms the clustering of these CpGs into a clustered group (70 CpGs) with cross-sectional BE drift rates that are estimated to be higher than those estimated for NS tissue which tend to be flat. Only a few outliers (13 CpGs) show the opposite behavior and likely represent false positives from the initial candidate selection in step 1 (see Fig. 7.3). Additionally, all but one of the 70 CpGs drift positively (mean population drift slope =.033), and all 70 are strongly correlated (mean $r = 0.64$) in their linear regression residuals. The only technically down drifting CpG on the population level of this set is nearly flat (slope = -0.0001), over 150 times larger than the universally negative slopes of the 13 CpGs in the left cluster (black dots in Fig. 7.3). The PCA seemed to cluster the positively drifting CpGs from Step 1 into two groups: one that drift similarly on a population level in D2 (70 green points in Fig. 7.3), and those few that show the opposite behavior (13 black points in Fig. 7.3). Thus, we consider these 70 differential drift CpGs as an admissible subset of CpGs that provide differential tissue age information for a quantitative estimation of BE onset times. Additional information on each of the 70 CpGs is provided in Tables D.4-D.5. In the following we will refer to this specific subset as *BE clock CpGs*.

Fig. 7.4 shows data (M-values) from patient data sets D1 and D2 for several representative BE clock CpGs. We next show how the individual BE onset times can be estimated from the methylomic drift observed in these clock CpGs using a Bayesian model that allows for measurement error and variability in marker-specific BE drift rates.

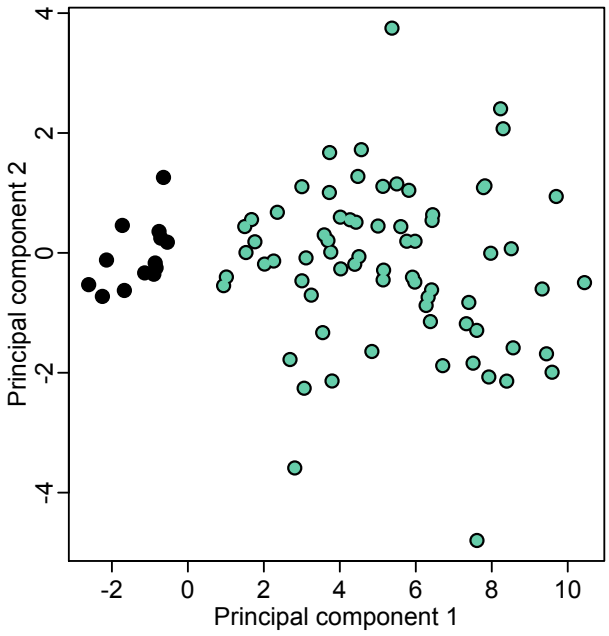

Figure 7.3: **Principal component analysis of methylation residuals for positive drift CpGs.** PCA analysis of $m_{BE} \sim age$ BE regression residuals for a total of 83 differential drift CpGs identified from data set D2 with matched hypomethylated NS tissues.

7.1.3 Bayesian BE clock model for estimating onset times and drift

Our goal is to employ information about the CpG locus-specific aging rates for both BE tissue and normal squamous tissue obtained from the data to predict when a patient first developed BE as a one-time occurrence. Given a patient's current methylation data from tissue samples, a set of intersecting lines representing the patient-specific, linear drift in methylation levels of BE tissue and normal squamous tissue will provide an estimate of the onset time of BE, i.e., when the BE tissue manifested itself by aging at a different rate than the normal squamous esophageal tissue. Here we are assuming methylomic drift is essentially linear (at the logit scale), although there is also evidence that age-associated variation in methylation levels may be better modeled by a function of logarithmic age for

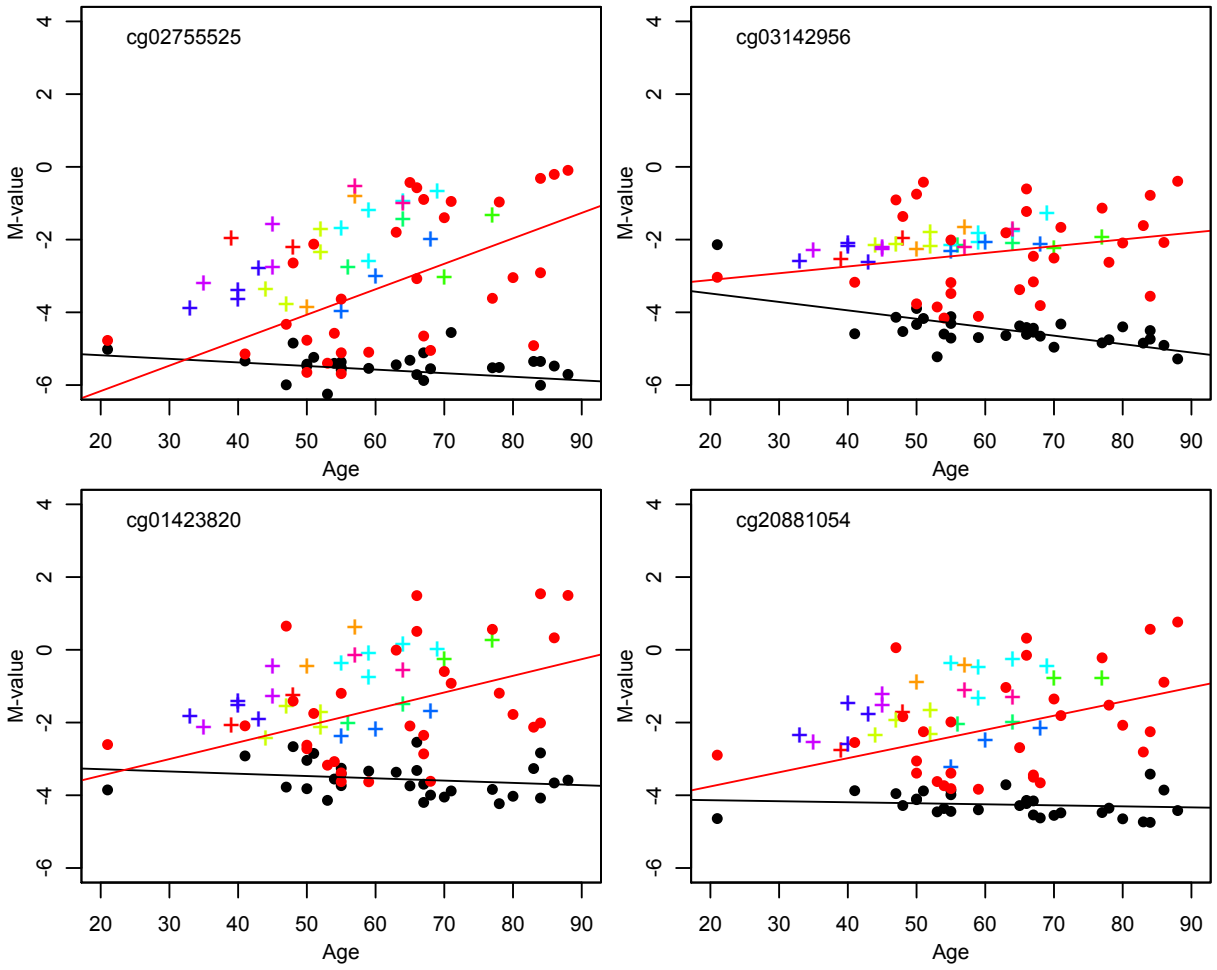

Figure 7.4: **BE clock CpGs.** Coss-sectional patient data D2 for 4 representative BE clock CpGs plotted for SQ M-values (black points) and BE M-values (red points). These four CpGs have the highest p-values for significant individual BE drift from the longitudinal data set D1. BE data for the longitudinal patients in data set D1 are shown with + symbols and different colors distinguishing each patient's samples, as in Fig. 7.1. Corresponding regression lines for cross-sectional data D2 are also plotted.

younger individuals [181].

Here we describe a Markov Chain Monte Carlo (MCMC) method to estimate the transformation times of BE, i.e., the *Bayesian BE clock model* for methylomic aging and the corresponding MCMC likelihood contributions.

For the BE clock CpGs applied to a patient group of size N, the data consists of measurements for $M = 70$ CpG sites for each patient $i, i = 1, .., N$ observed at time (age) $t_i = a_i$. These M measurements are each an observation $y_{BEij}(t_i), j = 1, .., M$, representing CpG_j M-value (logit(β) values) taken from BE cells in patient i. Since the measurements from BE cells over time have differential, increased drift compared to those for squamous cells after the onset of BE, we consider the following linear model for the conditional expected measurements, variable $Y_{BEij}(t_i)$, taken from patient i at time t_i for each CpG, given the onset of BE at time $s_i \leq t_i$,

$$E[Y_{BEij}(t_i)] \;\; = \;\; \alpha_{_{SQj}} + b_{_{SQj}} s_i + b_{ij}(t_i - s_i). \tag{7.1}$$

for $j = 1, .., M$. Thus, given the following parameters - the onset of BE at time $T_{_{BE}} = s_i$, the slope ($b_{_{SQj}}$) and intercept ($\alpha_{_{SQj}}$) of the SQ population regression lines obtained from individuals with matched samples in data set D2, and the patient-specific, CpG-specific BE drift rate b_{ij} - we observe M independent observations for N independent individuals. Allowing for patient-specific drift rates for the BE clock CpGs, we explicitly model the inter-individual differences in BE drift rates between 'slow' and 'fast' aging BE tissues.

Again, the observation from a single patient i, for $i = 1, \ldots, N$, observed at time t_i, is of the form

$$\{y_{BEij}, \quad j = 1, \ldots, M\}. \tag{7.2}$$

The likelihood contribution from a single patient observed at time t_i is then

$$\begin{aligned} &\prod_{j=1}^{M} f(y_{BEij}) \\ &= \prod_{j=1}^{M} f_{_N}(y_{BEij}; \mu_{_{BEij}} = \alpha_{_{SQj}} + b_{_{SQj}} s_i + b_{ij}(t_i - s_i), \sigma_{_{BEi}}) \end{aligned} \tag{7.3}$$

where $f_{_N}$ is the normal density function. In order to apply the MCMC technique, we assume uniform priors $p(s_i)$ for BE onset times s_i (due to the fact that the distribution of BE onset times in the general population is unknown), uniform priors $p(\sigma_{_{BEi}})$ for the standard deviation $\sigma_{_{BEi}}$ on the BE M-value observations, and prior distribution $p(b_{ij})$ for slopes $b_{ij}, j = 1..M$, which will be derived from the longitudinal data as a gamma distribution with empirical mean and shape parameters.

In terms of simulating the BE onset times $s_1, \ldots, s_N$ from the corresponding posterior distribution, let us define the vector $\Psi_i = (s_i, b_{i1}, \ldots, b_{iM}, \sigma_{_{BEi}})$. Samples of Ψ_i under its posterior distribution for patient i will be obtained using MCMC. Note that the posterior distribution of Ψ_i given the observation $\mathbf{y}_\sim$ comprised of patient-specific data of the form in Eq. (7.2), for $i = 1, \ldots, N$, is

$$\pi(\Psi_i|\mathbf{y}_\sim) \propto \text{likelihood} \cdot \text{prior} \tag{7.4}$$

$$= \prod_{j=1}^{M} f_{_N}(y_{BEij}; \mu_{_{BEij}}, \sigma_{_{BEi}}) \cdot p(s_i) \cdot p(b_{ij}) \cdot p(\sigma_{_{BEi}}) \tag{7.5}$$

This Bayesian BE clock model implements MCMC using the R package Bhat available from the CRAN website [94]. To summarize the MCMC output, the posterior parameter means, medians, and other quantiles over the MCMC will give Bayes estimates of the BE onset time s_i for individual $i, i = 1, \ldots, N$, the patient-specific, CpG-specific drift slopes $b_{ij}, j = 1, \ldots, M$, and patient-specific standard deviation parameter, $\sigma_{_{BEi}}$.

We obtain MCMC posterior estimates for all Bayesian BE clock model parameters using chains of length 500K cycles, all of which converged rapidly.

7.1.4 Testing significance of BE dwell time differences

To formally assess differences between different patient groups (say BE with dysplasia versus BE without dysplasia), we will use Bayes factors to statistically test if the BE onset ages estimated for one patient group s_i, $i = 1, .., N_k$, lead to BE dwell times that are significantly different from those of another group with estimated BE onset ages s'_i, $i = 1, .., N_l$ for

$k, l \in \{2, 3, 4\}$. For two specified data sets D_k, D_l, we will compare the average percentage of life until age at biopsy (a_i) during which the patient harbored BE. This quantity is given for two data sets by the following variables,

$$\gamma_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{a_i - s_i}{a_i}, \qquad \gamma_l = \frac{1}{N_l} \sum_{i=1}^{N_l} \frac{a'_i - s'_i}{a'_i} \tag{7.6}$$

Thus, we are interested in testing hypotheses $H_0 : \gamma_k > \gamma_l$ versus $H_1 : \gamma_k \leq \gamma_l$. For this test, we consider data $\mathbf{y}_\sim = \{\mathbf{y_k}, \mathbf{y_l}\}$ comprised of patient-specific observations of the form in Eq. (7.2) and compute the Bayes factor

$$B_{01} = \frac{\Pr[\mathbf{y}_\sim | H_0]}{\Pr[\mathbf{y}_\sim | H_1]} = \frac{\Pr[H_0 | \mathbf{y}_\sim] / \Pr[H_0]}{\Pr[H_1 | \mathbf{y}_\sim] / \Pr[H_1]} = \frac{\Pr[H_0 | \mathbf{y}_\sim] / \Pr[H_0]}{(1 - \Pr[H_0 | \mathbf{y}_\sim]) / (1 - \Pr[H_0])} \tag{7.7}$$

to quantify the evidence in favor of the null hypothesis H_0 and against the alternative H_1 [191]. To compute $\Pr[H_0 | \mathbf{y}_\sim]$, we apply the ergodic theorem and approximate the posterior probability by the fraction of MCMC samples satisfying $\gamma_k > \gamma_l$. The prior $\Pr[H_0]$ is computed similarly except we sample onset times s_i for the two groups of patients being compared directly from the uniform prior distributions $s_i \sim \text{Uniform}(0, a_i)$.

7.2 Results

7.2.1 Bayesian BE clock model estimates for BE patients in D2

First, we will use the Bayesian BE clock model to obtain posterior estimates of parameters for data set D2 (size $N_2 = 30$ patients) with the BE clock CpG set (size $M = 70$). See Materials and Methods for modeling details and CpG selection. Preliminary results show a wide inter-individual variability predicted for median BE onset ages among the 30 patients in D2.

7.2.2 BE onset predictions for familial BE cases in D3

Assuming that familial Barrett's esophagus (FBE) is mainly an inherited phenomenon, Chak et al. performed a study that shows that a certain putative inherited susceptibility gene(s)

influence the development of BE and its associated cancers in at least 7% of patients with these conditions [180]. We will estimate the Bayesian BE clock model parameters for the independent data set D3 (size $N_3 = 22$ patients) with FBE, with age range 39-84 at time of biopsy (mean age = 62.8).

Because a younger age of disease incidence is often considered a surrogate marker for a genetic predisposition, we will test the hypothesis that the FBE patients of data set D3 had been living with their BE for longer than the general BE patients of data set D2, which in our notation translates to $H_0 : \gamma_3 > \gamma_2$ (see Materials and Methods for details).

7.2.3 BE onset predictions for HGD/cancer cases in D4

Next, we will use the Bayesian BE clock model to estimate parameters for data set D4 consisting of 10 patients with diagnosis of neoplasia, defined as either high grade dysplasia (HGD), EAC, or both (see Table D.3 for patient details). We will similarly obtain parameter estimates from the Bayesian BE clock model and then use Bayes factor testing to see if the data suggests that the BE dwell time in neoplastic patients is longer than the dwell times for the general BE data set D2 and also the FBE dwell times obtained from data set D3, which in our notation translates to $H_0 : \gamma_4 > \gamma_2$ and $H_0 : \gamma_4 > \gamma_3$, respectively (see Materials and Methods for details).

7.2.4 Predicted EAC risks for BE patients

With the BE onset time predictions provided in the previous results, we may also associate a particular risk of developing EAC before a certain age. We standardize the risk of developing EAC to represent lifetime risk at age 85 (with oldest patient of the 3 data sets being 88 years old at time of biopsy/diagnosis of BE). We employ the multistage clonal expansion for EAC (MSCE-EAC) model that was previously calibrated to EAC incidence data by birth cohort, to obtain hypothetical EAC risk estimates for each patient [6, 14]. The risk we seek to quantify is the probability of developing EAC at random time T_{EAC} by age 85 conditional

on the onset age of BE, T_{BE}. For each BE patient, with BE onset time estimated to be age s from his/her methylation profile, we can compute the following risk

$$\Pr[T_{EAC} < 85 | T_{BE} = s] = 1 - S_{MSCE}(85 - s), \tag{7.8}$$

where S_{MSCE} is the survival probability for the multistage clonal expansion (MSCE) model from Chapter 2 (see Eq. (2.24)) [6, 13, 14].

7.3 Discussion

A fundamental problem in predicting the risk of esophageal adenocarcinoma (EAC) in patients with BE continues to be the difficulty in assessing the neoplastic potential of the premalignant field in which malignancies arise. Several lines of evidence support the notion that both BE segment length and the duration BE has been present in a patient (i.e., BE dwell time) are important determinants of EAC risk in addition to environmental and possibly genetic risk factors. While endoscopic surveillance with systematic biopsy sampling is the standard clinical care to screen BE patients for neoplastic changes and early cancer, most BE patients never develop esophageal cancer in their lifetime. Priority has therefore been given to novel approaches to assessing the molecular signatures of EAC progression and biomarkers that more precisely define EAC risk at the individual level. In parallel, because chronological age is recognized as one of the strongest predictors of cancer risk, renewed attention has been given to exploring the roles of biological tissue-age and cellular senescence in carcinogenesis.

Mutliple age-related changes in DNA methylation have recently been observed and interpreted to measure biological tissue age with accelerated tissue aging in neoplastic tissues [166, 167]. We therefore asked the question whether individuals who progress to EAC may simply have had BE for a longer period of time compared with individuals who do not, allowing more time for somatic clonal evolution and transformation to run their course. Alternatively, accelerated cancer progression could also be due to faster tissue aging in BE

possibly as a result of increased stem cell turnover. To determine when exactly a patient developed BE we examined methylation array data for differential epigenetic drift in BE versus normal squamous tissue (NS) in 30 patients with matched tissue samples. Additionally, we used methylation data from 10 patients with serial biopsies taken at least 5 years apart comprising a total of 29 separate BE tissue samples. These latter data are unique as they allow an estimation of CpG-specific drift rates which can not be gleaned from the cross-sectional data. Thus, to identify CpGs that drift significantly and differentially between BE and NS tissue, we began our search for CpG dinucleotides that undergo positive or negative longitudinal drift in BE tissue using the data generated from the serial biopsies. We deliberately cast a wide net when adjusting the regression p-values for multiple comparisons (q-value < 0.2) to identify candidate CpGs that drift significantly with age in BE (see Materials and Methods). This rather permissive preselection was then subjected to an Analysis of Covariance (ANCOVA) with two categorical groups (BE and NS), and subsequent Principal Component Analysis (PCA) as described in the Material and Methods to isolate CpGs that drift differentially in BE vs NS tissue and show high correlations in their M-value$\sim$age regression residuals between markers.

This bioinformatic approach allowed us to identify two subgroups of BE-specific drift-CpGs based on biological considerations of the DNA methylation maintenance process. For the first group (undergoing positive drift) we consider sporadic *de novo* methylation as the underlying process that leads to gradual gains in CpG methylation, while the second group (undergoing negative drift) can be considered to arise due to sporadic losses of 5-methylcytosines as a result of active CpG demethylation via DNA glycosylases and/or local suppression of DNA methylation maintenance during DNA synthesis [192,193]. These processes are known to involve distinct sets of enzymes, including DNA methyltransferases (Dnmt's) that are associated with the DNA replication machinery [194]. Epigenetic drift, we hypothesize, is therefore also a reflection of stem cell turnover, tissue repair and mitotic activity in a tissue. Furthermore, in BE these processes may also be modulated by life-style factors, history of gastroesophageal reflux, PPI use and NSAID use [90]. Interestingly, we found little evidence

for differential BE vs NS drift among CpGs that were hypermethylated ($\beta > 0.75$) in normal tissue but showed negative methylation drift in BE (second group).

Although the differential epigenetic drift we observe is likely to occur in genomic regions with no relevance to cancer progression (neutral drift) we hypothesize that at least some markers start to differ from normal tissue levels possibly as a result of tissue transdifferentiation or stem-cell reprogramming in BE [195, 196]. Of note, our set of 70 differential drift CpGs appears enriched for CpGs located on *islands* (76% vs 49% based on the number of annotated CpGs designated to this category by Illumina, see Tables D.4-D.5) [197]. The absence of hitherto identified CpG markers (or their loci) of neoplastic progression in BE among the 70 drift CpGs we identified in this study suggests that these CpGs undergo neutral drift.

Because BE is mainly asymptomatic (over 90% of EAC cases do not even present with a prior history of BE [198] although BE is found pre-operatively in most EAC cases [199]), a direct validation of when a patient first develops BE is presently not feasible. For this reason we are currently attempting to validate our predictions indirectly through two lines of evidence. There have been previous efforts to identify tissue-based indicators that accurately reflect the biological age of a tissue using regularized regression techniques by directly regressing age on the methylation levels of a large number of CpGs to identify subsets of CpGs that are predictors of chronological age [166, 167]. Although we cannot use these techniques in the context of BE because the BE onset times are unknown, we will see if our predictions are at least broadly consistent with the straightforward application of these clock models to estimate absolute tissue-age differences between BE and NS tissue. Specifically, using the published elastic net coefficients by Horvath [167] and by Hannum et al. [166] we will compute the predicted biological age of the BE tissue and subtracted the predicted biological age of the normal squamous (NS) esophageal tissue to arrive at estimates of the BE dwell time for the 30 patients in D2. By subtracting these estimates from chronological age we may obtain corresponding BE onset times that will be tested against with our mean estimates for correlations. Interestingly, our clock uses a distinct set of markers from those of Horvath

and Hannum et al.

Once a patient's BE onset distribution has been estimated from his/her methylomic drift profile, his/her risk of developing EAC can be estimated more precisely. We will use the previously validated MSCE-EAC model that explicitly considers the uncertainty of the timing of the BE onset in the general population and describes, conditional on when BE develops, the stochastic process of neoplastic progression from metaplastic to dysplastic tissue to cancer [6, 14]. Our results will test the theoretical predictions that show a strong dependence of EAC risk on the BE dwell time. Note, however, these EAC risk predictions do not consider the effects of interventions and therefore may be higher than empirical risk estimates. Our current work aims to see if BE onset, as determined by methylomic drift, should be considered a potential biomarker for EAC risk, although further validation via properly powered prospective studies or case-control studies in BE patients will be needed to confirm this. Such studies may provide the requisite data to further test how well BE tissue-age performs in identifying individuals that likely progress to HGD or EAC in their lifetime so that endoscopic surveillance and available interventions can be utilized more effectively.

Appendix A

Impact of Tumor Progression on Cancer Incidence Curves

A.1 Parameter estimates and uncertainty for the MSCE model

A.1.1 Fits to SEER incidence data

In the following we describe the fits of SEER-9 incidence data of CRC, GaC, and PaC using both the MSCE model and its approximation (referred to as MSCE-1) [7]. A comparison shows that the estimated lag-time parameters for the MSCE-1 approximation are in generally good agreement with corresponding estimates for T_2 in the MSCE model assuming biologically plausible values for the cell division rates of premalignant cells ($\alpha_{_P} \in [2, 50]$ per year) and malignant cells ($\alpha_{_M} \in [25, 100]$ per year [200]) and for the observation event rate $\rho \in [10^{-6}, 10^{-8}]$ per cell per year. See Table A.1 for the results of the sensitivity analysis of $\lambda, g_{_P}, g_{_M}, \mu_2^{eff}$ on choices for $\alpha_{_P}, \alpha_M, \rho$.

In general, the exact solution of the MSCE hazard function with two consecutive clonal expansions may be difficult to distinguish from a model with a single clonal expansion plus a time lag. This also means that the parameters of the model in Fig. 2.2, in particular

CRC males	λ x 10^{-4}	g_P	g_M	μ_2^{eff} x 10^{-6}
α_P [2, 50]	[2.13, 2.13]	[0.162, 0.162]	[2.58, 2.59]	[0.15, 3.76]
α_M [25,100]	[2.13, 2.13]	[0.162, 0.162]	[2.52, 2.67]	[0.73, 0.77]
ρ [10^{-6},10^{-8}]	[2.13, 2.13]	[0.162, 0.162]	[2.08, 3.07]	[0.75, 0.76]

CRC females	λ x 10^{-4}	g_P	g_M	μ_2^{eff} x 10^{-6}
α_P [2, 50]	[1.57, 1.57]	[0.149, 0.149]	[2.04, 2.05]	[0.32, 8.09]
α_M [25,100]	[1.57, 1.57]	[0.149, 0.149]	[1.92, 2.17]	[1.61, 1.62]
ρ [10^{-6},10^{-8}]	[1.57, 1.57]	[0.149, 0.149]	[1.61, 2.46]	[1.61, 1.63]

Table A.1: **Sensitivity analysis on MSCE model parameters.** Sensitivity analysis to test the effect of uncertainty in the cell division rates α_P and α_M, and in the detection rate ρ on the estimated MSCE model parameters.

the parameters that pertain to preclinical cancer progression, may be difficult to estimate from cancer incidence data alone. Additional assumptions such as equality of the mutation rates μ_0, μ_1 and μ_2, replacing some of the parameters with measured values, or inclusion of screening data on the number and sizes of preclinical tumors, are required to estimate all relevant model parameters (see [13] for more information on parameter identifiability).

A.1.2 Birth cohort and calendar year trends

We used a modified age-period-cohort (APC) approach to adjust the incidences for secular trends. While the period and cohort effects are modeled non-parametrically, the age effect follows the hazard function of the MSCE model and is therefore parametrically constrained. This finesses a non-trivial identifiably problem of the APC approach and allows us to separate cleanly age, period and cohort effects (see [41, 51, 52]). Briefly, the APC ansatz assumes an incidence function [52]

$$I_{bc}(t) = \Theta_b \, \Theta_c \, h_{_{MSCE}}(t) \tag{A.1}$$

where Θ_b and Θ_c are coefficients that modify the MSCE hazard $h_{_{MSCE}}(t)$. See Main Text and Supplemental Information of [13] for more information on estimated birth cohort and calendar year coefficients.

Figures A.1-A.6 show age-specific incidences of CRC, GaC, and PaC in SEER-9 from 1975-2008 by gender and 5 year calendar periods. The left panels show the raw (unadjusted) incidences whereas the right panels show period and cohort adjusted incidence curves as well as the MSCE model prediction. For CRC, GaC, and PaC both the exponential and the linear behavior of the adjusted curves stand out.

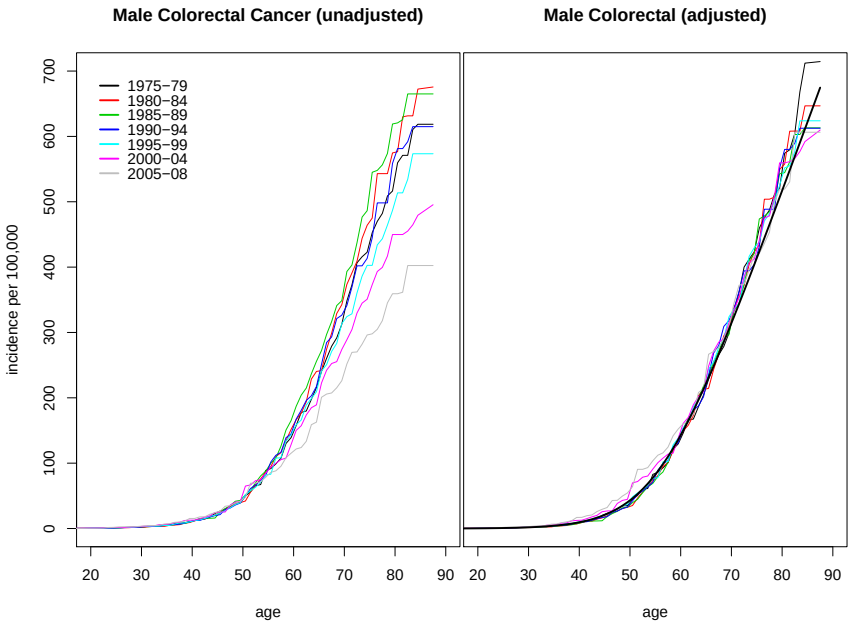

Figure A.1: (Left panel) Unadjusted CRC incidences among males from SEER-9 by 5 year calendar periods. (Right panel) Adjusted incidences using the estimated calendar-year and birth-cohort coefficients, and the MSCE model fit (thick black line).

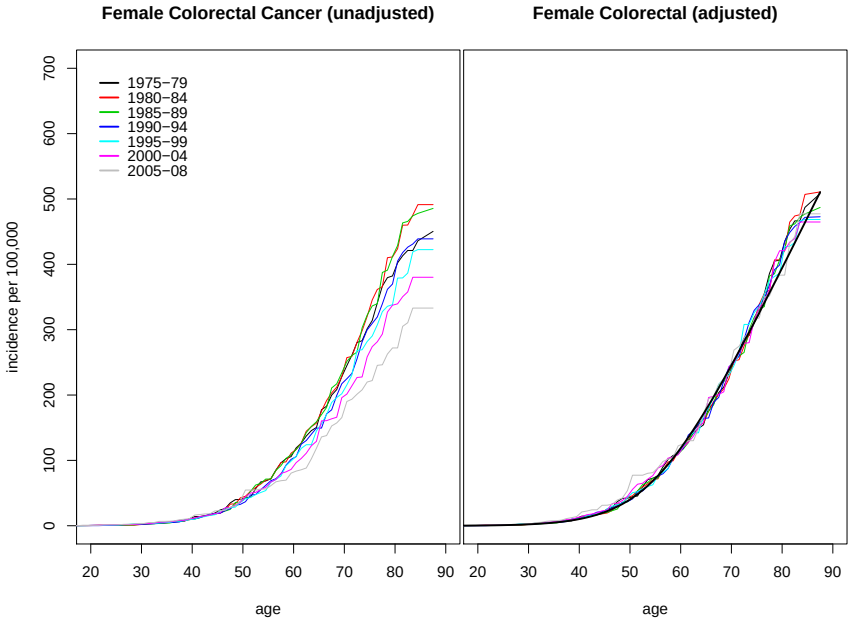

Figure A.2: (Left panel) Unadjusted CRC incidences among females from SEER-9 by 5 year calendar periods. (Right panel) Adjusted incidences using the estimated calendar-year and birth-cohort coefficients, and the MSCE model fit (thick black line).

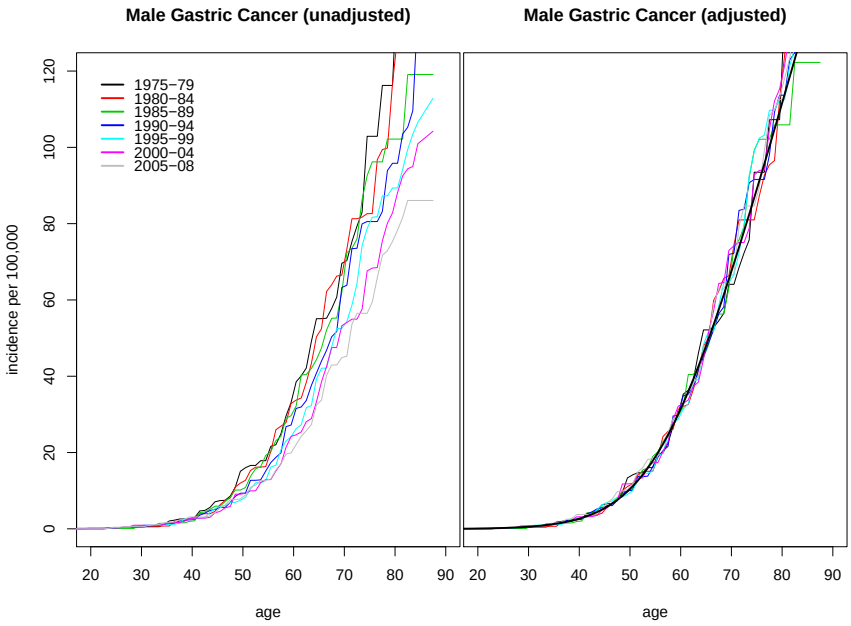

Figure A.3: (Left panel) Unadjusted GaC incidences among males from SEER-9 by 5 year calendar periods. (Right panel) Adjusted incidences using the estimated calendar-year and birth-cohort coefficients, and the MSCE model fit (thick black line).

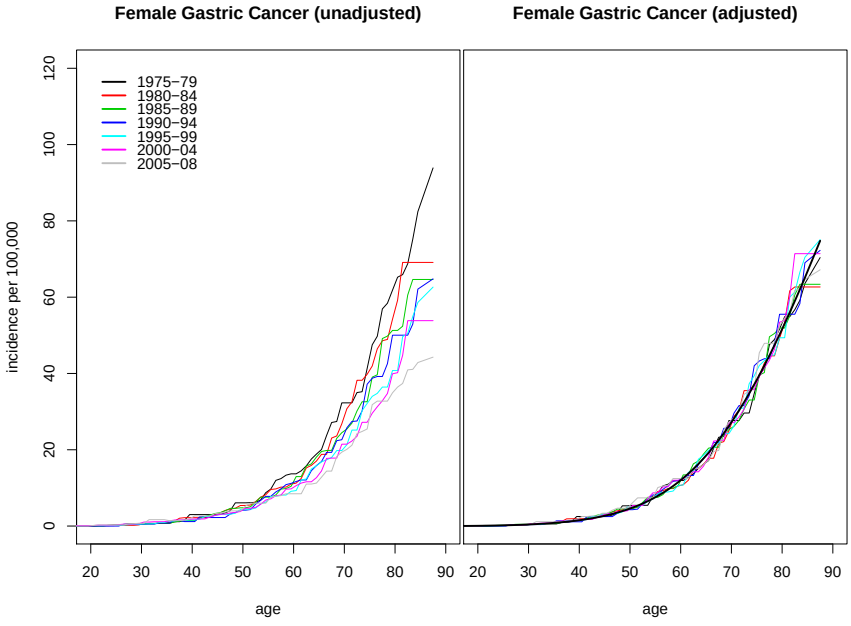

Figure A.4: (Left panel) Unadjusted GaC incidences among females from SEER-9 by 5 year calendar periods. (Right panel) Adjusted incidences using the estimated calendar-year and birth-cohort coefficients, and the MSCE model fit (thick black line).

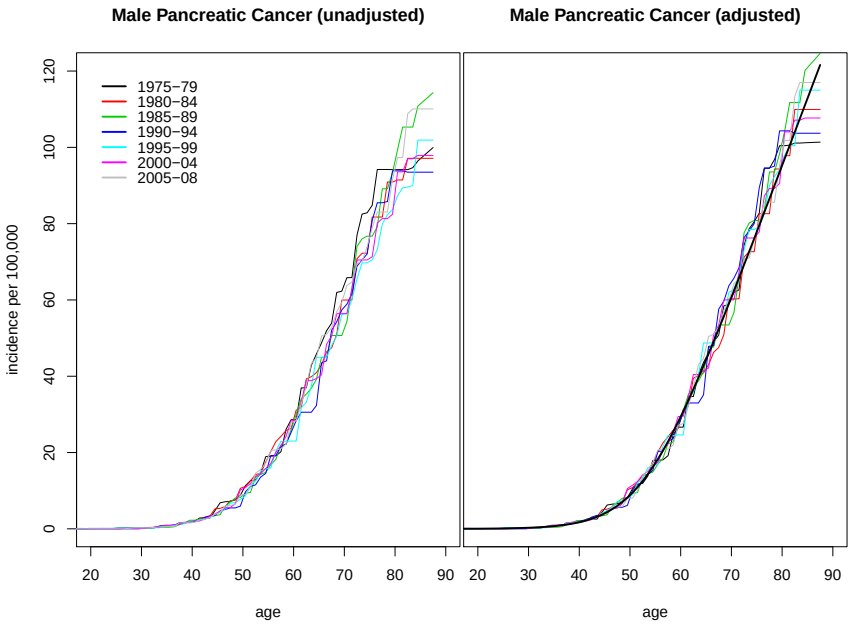

Figure A.5: (Left panel) Unadjusted PaC incidences among males from SEER-9 by 5 year calendar periods. (Right panel) Adjusted incidences using the estimated calendar-year and birth-cohort coefficients, and the MSCE model fit (thick black line).

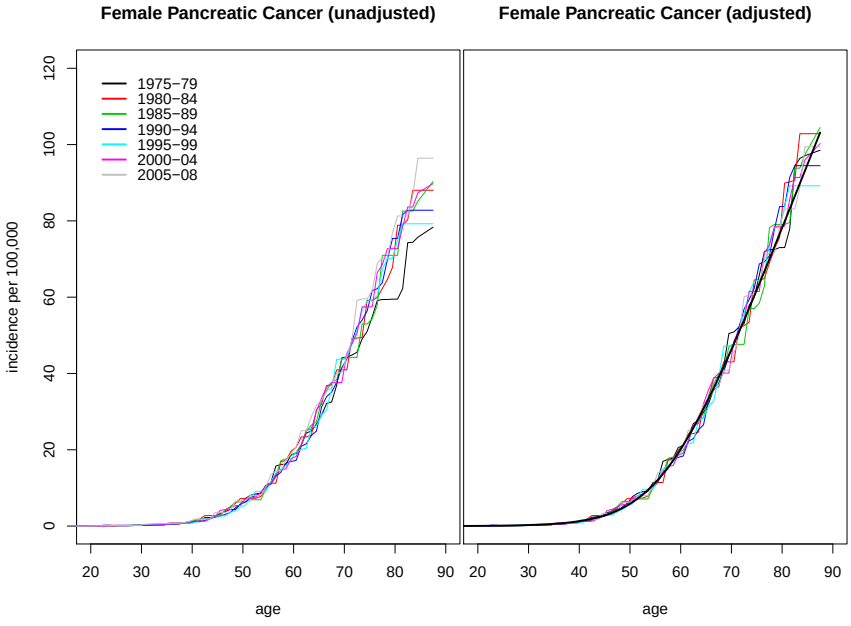

Figure A.6: (Left panel) Unadjusted PaC incidences among females from SEER-9 by 5 year calendar periods. (Right panel) Adjusted incidences using the estimated calendar-year and birth-cohort coefficients, and the MSCE model fit (thick black line).

Appendix B

The MSCE-EAC Screening Model

B.1 MSCE-EAC tissue module: hybrid simulation algorithm

The following steps describe the MSCE-EAC tissue module simulation for a single individual's MSCE-EAC multi-type branching process realization from birth until time of screening t_s

Step 1: Simulate BE onset time

For each individual, generate an age of BE onset, T_{BE} using the inverse cumulative distribution technique with Eq. (3.3). Because the incidence rate of BE in the general population is very small, most random realizations of this onset time will be longer than the human lifespan. The simulation continues only for those individuals who have a T_{BE} onset time within the age range of interest (possibly before a set screening age t_s).

Step 2: Generate BE segment length

Generate the size of a patient's BE segment (measured clinically as the length from gastroesophageal junction (GEJ) to the top of the longest "tongue" of metaplastic tissue) as a random deviate from a length distribution based on clinical studies [97, 113, 118–121]. This length will be translated into a number of BE stem cells, X, which depends on a spatial model

parameter σ for the density of stem cells per mm^2. The MSCE-EAC model assumes that BE stem cells are under tight homeostatic control with zero net growth of the non-dysplastic BE stem cell population (i.e., X remains constant). Mutations and clonal populations occurring during the simulation grow within this fixed BE segment.

Step 3: Generate pre-initiated stem cells P^*

Any of the BE stem cells, total of X, may undergo a Poisson rate-limiting mutation with rate μ_0 during an asymmetric division to produce a BE daughter stem cell and a pre-initiated P^* cell. A P^* cell may arise through inactivation of a single tumor suppressor allele. To model this process up until time of screening t_s, generate a number N_{P^*} of P^* pre-initiation events from a Poisson distribution with mean $\mu_0 \cdot X \cdot (t_s - T_{BE})$, each of which occur at a uniform time τ between T_{BE} and t_s, and save in vector $\vec{\tau}_1$.

Step 4: Generate initiated stem cells P and premalignant clones

Similarly, each P^* cell may undergo a second Poisson rate-limiting mutation with rate μ_1 during an asymmetric division to produce one P^* daughter stem cell and one initiated P cell. A P cell may be a cell with a tumor suppressor gene that has both alleles inactivated, which will allow it to undergo clonal expansion as an independent birth-death-mutation process. Again, for each P^* cell born at time $\tau_{1i}, i = 1, .., N_{P^*}$, generate the number N_P of initiated P progenitor cells from a Poisson distribution with mean $\mu_1 \cdot (t_s - \tau_{1i})$, each of which occurs at a uniformly distributed time between τ_{1i} and t_s, and save in vector $\vec{\tau}_2$.

For each P cell initiation, begin a simulation of the ensuing birth-death-mutation (b-d-m) process to follow the number and times of symmetric divisions, death or differentiation, and malignant transformations that occur in each premalignant clone.

Step 5: Generate preclinical cancer cells M, malignant clones, and clinical EAC

During simulation of premalignant clones, malignant transformations may occur within a particular clone, modeled as asymmetric divisions of a P cell with rate μ_2. For each malignant progenitor M cell born at time τ_3, begin an independent birth-death-detection process that is represented by an analytical solution to the corresponding Kolmogorov equation for the generating function as derived in Eq. (3.5) of Jeon et al. [122]. Thus, the hybrid simula-

tion makes use of previous theoretical results for an analytical distribution to avoid further simulation. We are first interested in knowing whether a malignant clone born at time τ_3 leads to a clinical EAC by time t_s. To generate this potential outcome, we use the 1-stage *survival function* S_M,

$$S_M(u) \equiv \Pr[D(t_s) = 0 | M(\tau_3) = 1] \tag{B.1}$$

where $u = t_s - \tau_3$, $D(t_s)$ is the random variable for clinical detection by time t_s, and $M(\tau_3)$ is the random number of malignant (preclinical) cells at time τ_3 in a malignant clone. Letting $p = 1 - S_M(t_s - \tau_3)$ represent the probability of cancer detection of a particular malignant clone born at time τ_3, draw a Bernoulli random variable with probability p to decide if this clone will be detected as a clinical EAC by time t_s. Draw Bernoulli deviates from detection probability p for each malignant clone generated in a patient, and repeat for every patient in the simulated population to obtain the EAC detection prevalence by time t_s.

For patients in whom a malignant clone born at time τ_3 does not result in clinical EAC by time t_s, use an analytical distribution to generate the size of the malignant clone present at time t_s, conditional on no EAC detection. Jeon et al. [122] derived this conditional size distribution for a birth-death-detection process, which is a shifted geometric distribution, described in more detail forthcoming.

This step completes the MSCE-EAC hybrid simulation of an individual from birth until time (age) t_s which can be repeated to generate (synthetic) data for a sample population. In summary, for those individuals who are found to have BE by time of screening, each patient has a specific X number of BE stem cells, P^* number of pre-initiated cells, a number of non-extinct P clones with respective sizes, a number of non-extinct M clones with respective sizes and information about the parental P clones from which the M clones originated, and lastly whether the patient is a prevalent, clinical EAC case by time t_s.

B.1.1 Implementation of SSA and τ-leap method for P clones

As mentioned in Step 3 of the MSCE-EAC algorithm above, initiated premalignant clones undergo independent birth-death-mutation (b-d-m) processes that we simulate to track cell count and times of malignant transformations. The stochastic simulation algorithm (SSA) is a mathematically exact method to follow each event that occurs during a realization of the continuous time Markov chain beginning with a single cell. Considering an individual premalignant clone of size X_t at time t, we define the intensity function vector $r(X_t) = (\beta_{_P} X_t, \mu_2 X_t, \alpha_{_P} X_t)$ for death/differentiation, malignant transformation, and birth of new P stem cell, where, over a short period of time s, we expect $r_j(X_t)s + o(s)$ events of type j to occur. Due to the Markovian property of the process, we wait an exponential length of time until the next event occurs with intensity $r_0(X_t) = \sum_{j=1}^{3} r_j(X_t) = X_t(\beta_{_P} + \mu_2 + \alpha_{_P})$. Once an exponential time to next event is chosen, we jump to the neighboring state $X_t + v_j$ with probability $r_j(X_t)/r_0(X_t)$, where v_j is the j^{th} component of the state change vector $v = (-1, 0, 1)$ for the b-d-m process. Fortunately, in the case of the P clone process with constant rates, the probabilities $r_j(X_t)/r_0(X_t)$ are constant with respect to the current state X_t so we may generate a number K of events of the three types with probabilities $\left(\frac{\beta_{_P}}{\beta_{_P}+\mu_2+\alpha_{_P}}, \frac{\mu_2}{\beta_{_P}+\mu_2+\alpha_{_P}}, \frac{\alpha_{_P}}{\beta_{_P}+\mu_2+\alpha_{_P}}\right)$ and cumulatively sum each $X_t + v^j$ step for the K chosen events to create a state vector N. Then we generate the K exponential waiting times of the process at once from an exponential with mean $\lambda_t = N(\beta_{_P} + \mu_2 + \alpha_{_P})$ and cumulatively sum these to arrive at a new later time $t_2 > t$.

The SSA works very well when cell count of the P clone is small and the event intensities $r(X_t)$ are fluctuating quickly. In particular, our simulation benefits to use the SSA for the beginning of a P clone's growth from a single cell, when the probability of extinction is high (because $\beta_{_P}$ is only slightly smaller than $\alpha_{_P}$) and most clones are eliminated after a small number K of initial events. However, the SSA can become excruciatingly slow when a P clone becomes very large, i.e. contains a large number of stem cells. Therefore, rather than simulating every event choice and time, we can employ an accelerated but approximate

procedure called the τ-leap method, first introduced by Gillespie and others [201–203]. The goal of this procedure is to advance the cell count by a preselected time increment τ in contrast to the exponential time increments generated in the SSA. To control the loss of accuracy with this approximation, the choice of leap-size τ must satisfy the historically referenced "leap condition" which is large enough that many events occur in that time, but nevertheless small enough that the intensity function value r is likely to change only "infinitesimally" as a consequence of those events. To the extent that this condition is satisfied, the mathematical rationale in replacing Markovian kinetics with Poisson kinetics [204] states that the number of times each independent event j will occur in the set time length τ can be approximated by a Poisson random variable with mean $\omega(t, t+\tau)$ on the interval $(t, t+\tau)$. For the ordinary τ-leap scheme, we assign $\omega(t, t+\tau) = r_j(X_t)\tau$. Thus, we set the intensity of event j equal to the constant $r_j = r_j(X_t)$ and we update the cell count vector $X_{t+\tau} = X_t + \sum_{j=1}^{3} n_j v_j$, where n_j are independent Poisson variates with means $r_j\tau$. Beyond ordinary τ-leaping, advancements have been made in improving accuracy when anticipating changes in the various components of the intensity vector by expanding $r_j(X_{t+\tau})$ in a Taylor series around time t with base value $r_j(X_t)$ to derive linear and quadratic approximations [205].

Selection of τ increments

As mentioned previously, we first set a number K (e.g., $K = 1000$) SSA steps to perform very quickly at the initiation of the P clone (first initiated cell asymmetrically divides from pre-initiated cell) in order to exactly simulate the small clones and capture the early extinction stochastic event correctly. Then, if a P clone is still growing, we switch to a τ-leaping algorithm to speed up the simulation of the larger clones without loss of much accuracy. In search of a balance between computational efficiency and accuracy for our hybrid SSA/τ-leap algorithm, we would like to take advantage of the leap condition by employing τ-leaping when the P clones are large, which will take a very large number of reaction events to change the intensity function "significantly", and the exact SSA when τ is required to be small so that only a few reactions would be leaped over regardless. Recent work by Sehl et

al. [205, 206] and others derived and applied methods to anticipate the largest τ such that the leap condition will be satisfied and accuracy will not be undesirably diminished. This will require us to introduce a small positive constant ϵ, which must be chosen empirically, to denote the acceptable relative change in the intensity function vector r. Adhering to the results of Cao et al. [207], we then chose our increment to be the largest τ such that

$$\left|\frac{d}{dt}r_j(X_t)\right|\tau \leq \epsilon r_j(X_t) \quad \Rightarrow \quad \tau \leq \frac{\epsilon}{\alpha_{_P} - \beta_{_P}} \tag{B.2}$$

holds for all j. Further, Cao et al. [202] explore the problem of negative population sizes which may occur with some probability within the τ-leaping method. In the setup described, this happens extremely rarely since τ-leaping usually only begins for clones with a substantial number of cells, thus they have a low probability of extinction because they are entering the exponential mean growth phase (described in more detail in the following section). Thus, we can reject this choice of τ that produced a negative size and reduce τ, by 1/2 for example, until no negative populations are produced since this is a rare event and will not impede our computational runtime [202].

To obtain better accuracy without compromising speed in simulation time, a recent step anticipation leap (SAL) method has been developed that generalizes the ordinary τ-leaping method by projecting linear and quadratic changes in reaction propensities [205]. However, due to the nature of the birth-death-mutation processes modeled for premalignant P clones in the MSCE-EAC hybrid simulation, the leap condition for all these methods produces a restraint on τ that does not depend on the current size of the clone, as seen in Eq. (B.2). The linear and quadratic extrapolations of the propensity functions do not yield major improvements in accuracy when τ does not depend on the clone size at time t. Therefore, we employ the ordinary τ-leaping scheme in which we set time length $\tau = \frac{\epsilon}{\alpha_{_P} - \beta_{_P}}$, choose ϵ empirically to obtain desirable accuracy with our choice of cellular kinetic parameters, and approximate the number of times each independent event j (either birth, death/differentiation, or mutation) will occur by a Poisson random variable with mean $\omega(t, t+\tau) = r_j(X_t)\tau$ on the interval $(t, t+\tau)$. We may apply the result of the following section that an independent

b-d-m process produces a shifted geometric size distribution for non-extinct clones, given by Eq. (B.10) but with P clone parameters, and enjoys mean exponential growth with rate $\alpha_{_P} - \beta_{_P}$. See Q-Q plot comparison in Fig. B.1 of both SSA and τ-leap algorithms against the true theoretical shifted geometric distribution for the b-d-m process. In practice with the estimated parameters given in Table 3.1, the hybrid SSA/τ-leap algorithm utilizes small values of $\tau < .004$ years, which allows even more accuracy yet still benefits from far less computational time than if we were to use solely SSA type steps.

B.1.2 Malignant size distribution

Expanding on Step 5 of the MSCE-EAC algorithm above, when a malignant progenitor M cell is born at time $\tau_{_3}$, an independent birth-death-detection process begins and we have the analytical solution to the corresponding Kolmogorov equation for the generating function $\Phi_{_M}(y_3, z; \tau, t)$ (from Eq. (3.6)) as derived in Eq. (3.5) of Jeon et al. [122]. Thus, we can make use of previous theoretical results here allowing us to avoid further simulation. We are first interested in knowing whether this malignant clone born at time $\tau_{_3}$ leads to a clinical EAC by time t_s. To generate this potential outcome, we use the 1-stage survival function, where $u = t_s - \tau_{_3}$

$$S_{_M}(u) \equiv \Pr[D(t_s) = 0 | M(\tau_{_3}) = 1] = \Phi_{_M}(1, 0; \tau_{_3}, t_s) = 1 + \frac{1}{\alpha_{_M}} \frac{p_{_M} q_{_M} e^{-p_{_M} u} - q_{_M} p_{_M} e^{-q_{_M} u}}{q_{_M} e^{-p_{_M} u} - p_{_M} e^{-q_{_M} u}} \tag{B.3}$$

with

$$p_{_M} = \frac{1}{2}(-(\alpha_{_M} - \beta_{_M} - \rho) - \sqrt{(\alpha_{_M} - \beta_{_M} - \rho)^2 + 4\alpha_{_M}\rho}) \tag{B.4}$$

$$q_{_M} = \frac{1}{2}(-(\alpha_{_M} - \beta_{_M} - \rho) + \sqrt{(\alpha_{_M} - \beta_{_M} - \rho)^2 + 4\alpha_{_M}\rho}) \tag{B.5}$$

Letting $p = 1 - S_{_M}(t_s - \tau_{_3})$ be the probability of cancer detection of a particular malignant clone born at time $\tau_{_3}$, we first draw a Bernoulli random variable with probability p to decide if this clone will be detected as a clinical EAC by time t_s. The algorithm draws Bernoulli

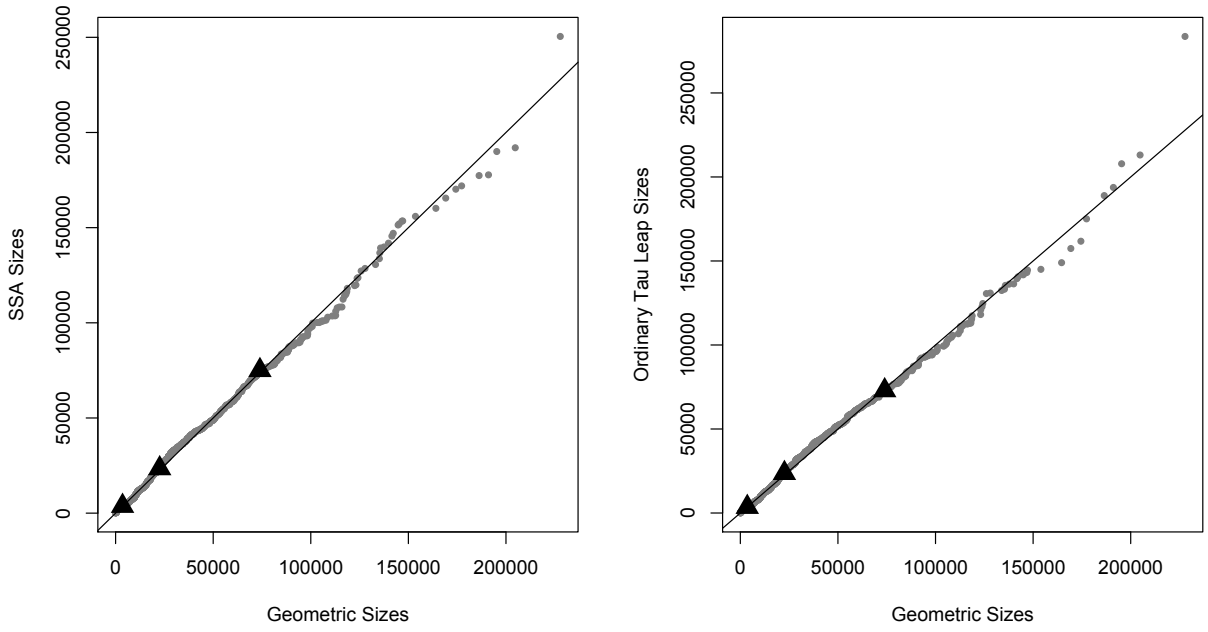

Figure B.1: **Comparison of neoplastic clone growth simulation accuracies using stochastic simulation algorithm (SSA) and τ-leaping.** Quantile-quantile plots of simulated non-extinct birth-death-mutation processes with $\epsilon = .0005$, stopped at time $t = 30$ years after first P cell initiated. (Left panel) SSA sizes vs. geometric distribution sizes. (Right panel) τ-leap sizes vs. geometric distribution sizes. Black triangles denote 10th, 50th, and 90th percentiles. Results shown for 100K simulations for each of the three types. Both SSA and τ-leap reproduce the theoretical geometric distribution very well.

deviates from this 1-stage survival for each malignant clone in a BE patient and provides an EAC detection prevalence by time t_s.

For a patient in whom a malignant clone born at time τ_3 does not result in clinical EAC by time t_s, we would now like to use an analytical distribution to generate the size of the malignant clone present at time t_s, conditional that it did not undergo EAC detection.

Jeon et al. [122] derived this conditional size distribution for a birth-death-detection process, which turns out in fact to be a shifted geometric distribution. Following Eqs. (3.9-3.16) in that paper, we have the size distribution of a malignant clone, given that no clinical cancer develops by time t_s for $n = 0$,

$$P_0 \equiv \Pr[M(\tau_3, t_s) = 0 | D(\tau_3, t_s) = 0, M(\tau_3, \tau_3) = 1] \tag{B.6}$$

$$= \frac{\zeta_M(\alpha_M + p_M)(\alpha_M + q_M)(q_M e^{-p_M u} - p_M e^{-q_M u})}{q_M(\alpha_M + p_M)e^{-p_M u} - p_M(\alpha_M + q_M)e^{-q_M u}}, \tag{B.7}$$

and for $n \geq 1$,

$$P_n \equiv \Pr[M(\tau_3, t_s) = n | D(\tau_3, t_s) = 0, M(\tau_3, \tau_3) = 1] \tag{B.8}$$

$$= \frac{1}{n!} \frac{\partial^n \Phi_M(y_3, 0; \tau_3, t_s)}{\partial y_3^n}\bigg|_{y_3=0} \frac{1}{\Phi_M(1, 0; \tau_3, t_s)} \tag{B.9}$$

$$= (1 - P_0)(1 - \alpha_M \zeta_M)(\alpha_M \zeta_M)^{n-1} \tag{B.10}$$

where

$$\zeta_M = \frac{e^{-p_M u} - e^{-q_M u}}{(q_M + \alpha_M)e^{-p_M u} - (p_M + \alpha_M)e^{-q_M u}}$$

Thus, conditional on a malignant clone remaining undetected by time t_s, we again construct an inverse cumulative function and begin with a uniform random deviate $u \in [0, 1]$. If $P_0 \geq u$, this particular malignant clone in question goes extinct before time t_s. If $u > P_0$, we derive the inverse cumulative function as follows

$$\Pr[M(u) > n] = \sum_{k=n+1}^{\infty} \Pr{}^*[M(u) = k] = \sum_{k=n+1}^{\infty} (1 - P_0)(1 - \alpha_M \zeta_M)(\alpha_M \zeta_M)^{k-1}$$

$$= (1 - P_0)(1 - \alpha_M \zeta_M)(\alpha_M \zeta_M)^n \sum_{k=n+1}^{\infty} (\alpha_M \zeta_M)^{k-(n+1)}$$

$$\Rightarrow \Pr[M(u) \leq n] = 1 - (1 - P_0)(\alpha_M \zeta_M)^n$$

Thus, we may generate a size n from this distribution,

$$n = \frac{\ln\left(\frac{1-u}{1-P_0}\right)}{\ln(\alpha_M \zeta_M)}, \quad \text{with } u > P_0.$$

B.2 MSCE-EAC screening module: EAC incidence projections

Here we derive all components of the general cumulative hazard $\Lambda_{EAC}(t)$, given by Eq. (4.2), in Chapter 4. For the initial scenario of screening all individuals at time t_s, we derived the MSCE-EAC cumulative hazard function that includes contributions from the subpopulation of those individuals found to have BE at time t_s that, immediately following HGD diagnosis, may receive treatment at time t_s; and the subpopulation without BE. For any time $t > t_s$ and BE cumulative density function F_{BE} given in Eq. (3.3), we compute the MSCE-EAC density function $f_{EAC}(s)$ for the general population explicitly as follows

$$f_{EAC}(s) = f_{EAC}(s|T_{BE} \leq t_s) \cdot \Pr[T_{BE} \leq t_s] + f_{EAC}(s|T_{BE} > t_s) \cdot \Pr[T_{BE} > t_s]. \tag{B.11}$$

The convolution formula for the unscreened population contributing to the MSCE-EAC density is given by

$$f_{EAC}(t|T_{BE} > t_s) = \frac{1}{\Pr[T_{BE} > t_s]} \int_{t_s}^{t} f_{BE}(u) f_{MSCE}(t-u)\,\mathrm{d}u \tag{B.12}$$

$$= e^{\int_0^{t_s} \nu(s)\,\mathrm{d}s} \cdot \int_{t_s}^{t} \nu(u) e^{-\int_0^{u} \nu(s)\,\mathrm{d}s} \cdot f_{MSCE}(t-u)\,\mathrm{d}u, \tag{B.13}$$

where $f_{MSCE}(t-u) = -Y_8(t-u, t-u)$ (see Eq. (3.22)) is the numerical solution for the 4-stage multistage model after BE onset.

For the screened BE population, we follow the method of Jeon et al. [122] and consider the 4 possible types of cells present in a patient at screening time t_s^- (where the minus superscript denotes cell populations present prior to any intervention) X = number of BE stem cells in BE segment, $P^*(t_s^-)$ = number of pre-initiated P^* cells, $P(t_s^-)$ = number of initiated, dysplastic P cells (all clones combined), $M(t_s^-)$ = number of malignant, preclinical cells (all clones combined). The MSCE-EAC hybrid simulation records these random variables for each BE patient at the instance of screening t_s^-, before any intervention occurs. After simulating n independent and identically distributed (by gender) individuals and performing the

Seattle biopsy screening protocol *in silico* as described in Chapter 4 Methods, the simulation obtains the vector $A_i = \{X_i, P_i^*(t_s^-), P_i(t_s^-), M_i(t_s^-)\}$ for each patient with BE, $i = 1, .., n$.

As described in Chapter 4, we explore the simulated intervention of radiofrequency ablation (RFA) of dysplasia patients by introducing the following ablation proportion vector, $\omega = \{\omega_{_X}, \omega_{_{P^*}}, \omega_{_P}, \omega_{_M}\}$, to deplete the existing cell types and leave a specified fraction (based on desired effectiveness of RFA treatment) in the esophagus given by ω. We adjust each dysplastic patient's cell count vector A_i through component-wise multiplication by ω. Thus, the post-RFA numbers of cells in each stage of the MSCE process, in simulated patient i, immediately after screening and treatment (denoted by time t_s^+) are given by an adjusted cell type vector $\hat{A}_i$ as follows

$$\hat{A}_i \equiv \omega \circ A_i = \{\omega_{_X} \cdot X_i, \omega_{_{P^*}} \cdot P_i^*(t_s^-), \omega_{_P} \cdot P_i(t_s^-), \omega_{_M} \cdot M_i(t_s^-)\} \tag{B.14}$$

$$= \{X_i(t_s^+), P_i^*(t_s^+), P_i(t_s^+), M_i(t_s^+)\} \tag{B.15}$$

$$\equiv \{\hat{X}_i, \hat{P}_i^*, \hat{P}_i, \hat{M}_i\}. \tag{B.16}$$

BE patients with a negative screen for dysplasia sustain the same (before and after) $A_i \equiv \hat{A}_i$ vector as was computed at time t_s^- since no RFA treatment is performed on these patients. Due to Markovian renewal of the branching process, we may then compute the survival and hazard functions, as in [122], for each screened individual $i = 1, .., n$ for some time $t > t_s$ with contributions from each cell type post screen,

$$S_{_{EAC}}(t - t_s|\hat{A}_i) = S_4(t - t_s)^{\hat{X}_i} S_3(t - t_s)^{\hat{P}_i^*} S_2(t - t_s)^{\hat{P}_i} S_1(t - t_s)^{\hat{M}_i} \tag{B.17}$$

$$h_{_{EAC}}(t - t_s|\hat{A}_i) = \hat{X}_i h_4(t - t_s) + \hat{P}_i^* h_3(t - t_s) + \hat{P}_i h_2(t - t_s) + \hat{M}_i h_1(t - t_s) \tag{B.18}$$

$$\Rightarrow \quad f_{_{EAC}}(t|T_{_{BE}} \leq t_s) \approx \frac{1}{n} \sum_{j=1}^{n} h(t - t_s|\hat{A}_j) \cdot S(t - t_s|\hat{A}_j). \tag{B.19}$$

These survival and hazard functions for the 4-stage MSCE model after BE onset may be easily computed from the Kolmogorov backward equations. The 8 ODEs from Eqns. (3.15-3.22) in Chapter 3 can be solved numerically to obtain the survival and hazard functions we require

$$h_4(t - t_s) = -\frac{Y_8(t - t_s)}{Y_7(t - t_s)} \quad \text{and} \quad S_4(t) = Y_7(t - t_s),$$

$$h_3(t-t_s) = -\frac{Y_6(t-t_s)}{Y_5(t-t_s)} \quad \text{and} \quad S_3(t) = Y_5(t-t_s),$$
$$h_2(t-t_s) = -\frac{Y_4(t-t_s)}{Y_3(t-t_s)} \quad \text{and} \quad S_2(t) = Y_3(t-t_s),$$
$$h_1(t-t_s) = -\frac{Y_2(t-t_s)}{Y_1(t-t_s)} \quad \text{and} \quad S_1(t) = Y_1(t-t_s).$$

We have now derived all necessary components of $\Lambda_{EAC}(t)$ of Eq. (4.2) after a single screen of all individuals at time t_s. See the Results section for an illustrative figure.

B.3 Example simulation details

For the Results presented in Chapter 4, we simulated an index endoscopy for all males and females age $t_s = 60$ in the year 1990 (indicative of index screens from prospective studies that estimate the BE to EAC progression rate). With BE prevalence F_{BE} given in Eq. (3.3), the Results focus on output regarding the subpopulation of individuals found with BE, for whom the MSCE-EAC screening model obtains screening results (see Chapter 4, Methods). See Figs. B.2-B.3 for values $p_{sGERD}(t)$ and $F_{BE}(t)$ used for males and females, respectively.

We generated a BE segment length for each patient from a beta distribution with shape parameters 16/11 and 4, restricted to the range of 1-16 cm for both males and females. BE segments of the simulated patients have an average length of 4.9 cm. Short segment BE (less than 3 cm) comprises 22% of the density and long segment BE (greater than 3 cm) comprises 78%. This BE distribution recreates the proportions of long and short segments recorded for the study patient population in [118]. Figures B.4-B.7 depict the number distributions and long-tailed, Luria-Delbruck type size distributions for the non-extinct premalignant and malignant clones, respectively, present at time $t_s = 60$ for the cohorts of males and females separately. Based on 100K simulation, the mean number of premalignant clones per BE patient, without symptomatic cancer, in this cohort is 6.6, while only about 1% of these dysplastic clones harbor a non-extinct malignancy.

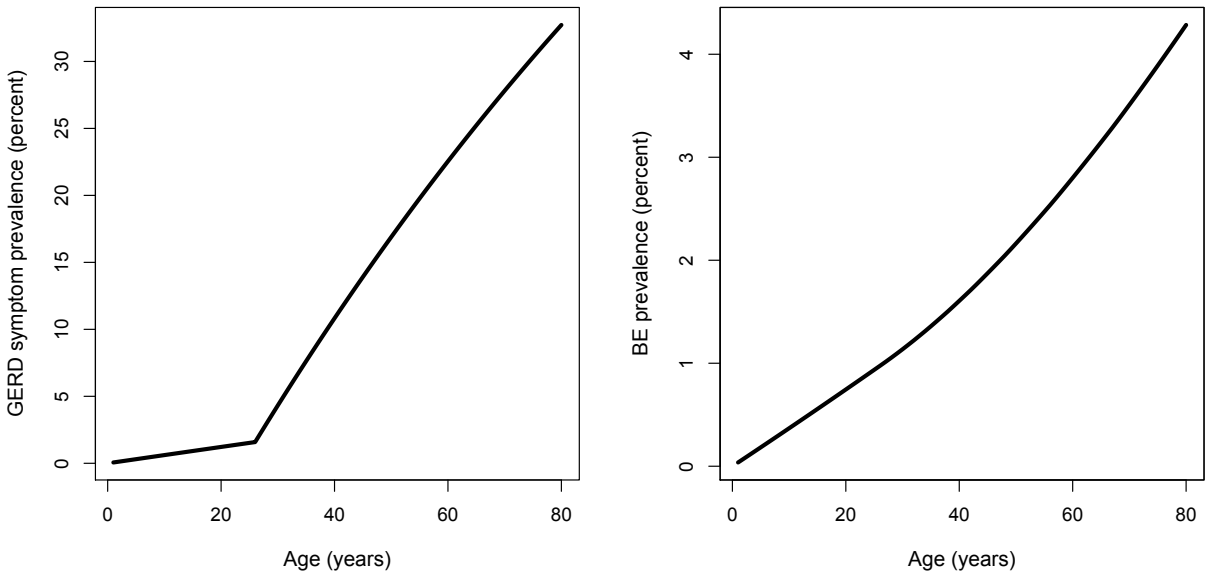

Figure B.2: **Male symptomatic GERD and BE prevalences in the MSCE-EAC screening model.** (Left panel) GERD symptom prevalence $p_{_{sGERD}}(t)$. (Right panel) BE prevalence $F_{_{BE}}(t)$ for males, assuming $RR = 5$ relative risk for symptomatic GERD patients.

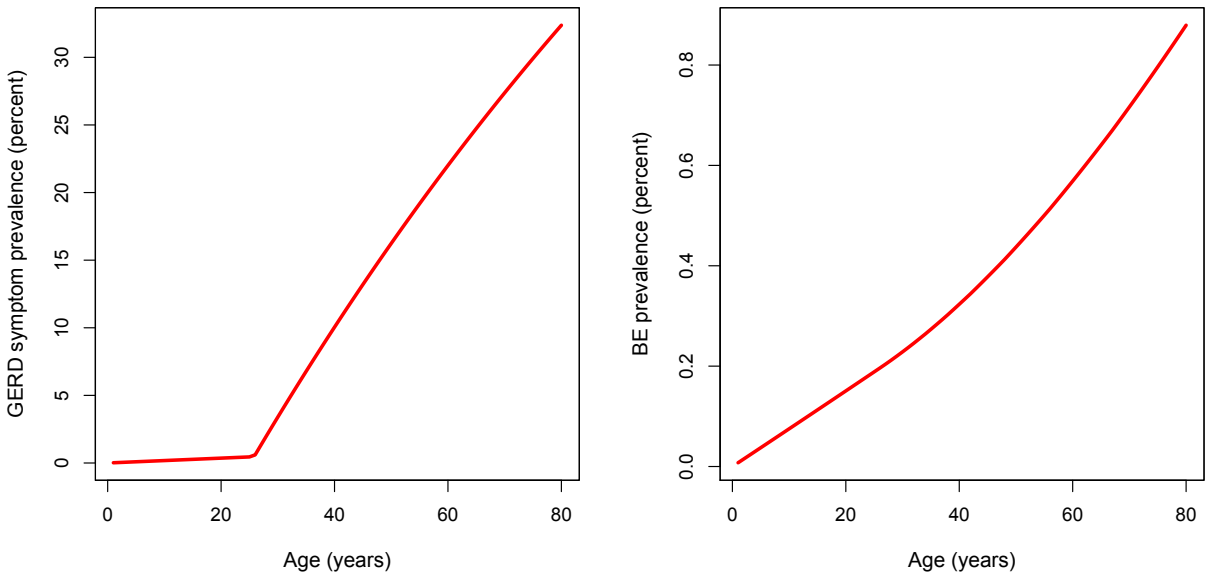

Figure B.3: **Female symptomatic GERD and BE prevalences in the MSCE-EAC screening model.** (Left panel) GERD symptom prevalence $p_{_{sGERD}}(t)$. (Right panel) BE prevalence $F_{_{BE}}(t)$ for females, assuming $RR = 5$ relative risk for symptomatic GERD patients.

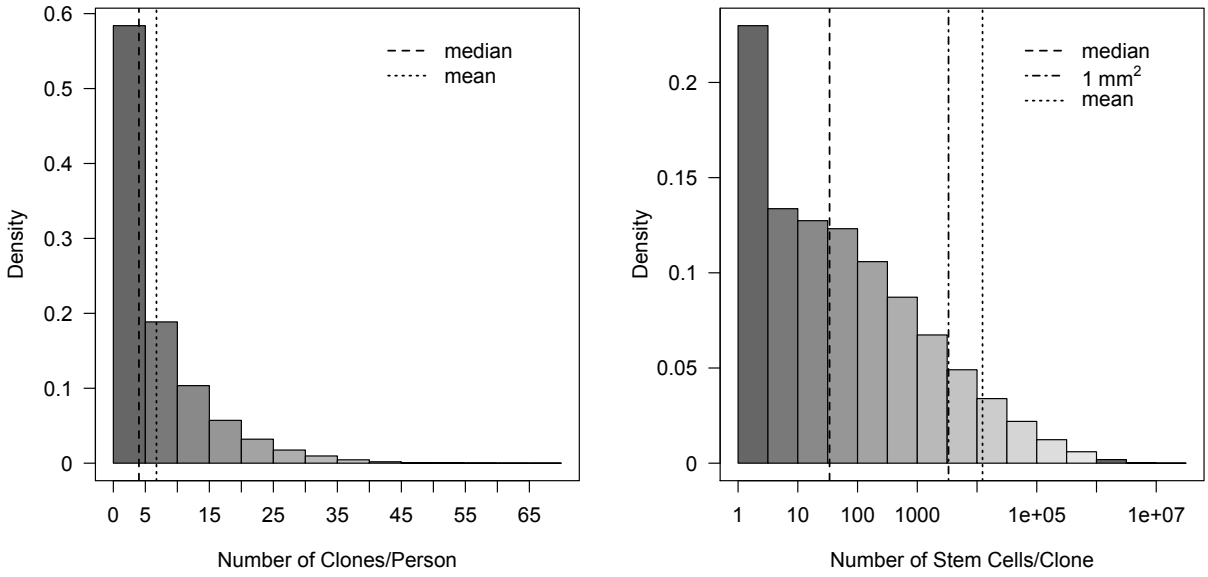

Figure B.4: **Simulated distributions of number and sizes of non-extinct premalignant clones, males.** (Left panel) Histogram of number of non-extinct premalignant clones in BE segment at time of screening. (Right panel) Histogram of number of premalignant stem cells in each of the independent premalignant clones accounted for in left panel. Example shown for 100K males with BE, age 60 from the 1930 birth cohort. Median and mean values are depicted by dashed and dotted lines, respectively. Assuming $\sigma = 3300$ stem cells/mm^2, the dashed-dotted line on the right graph gives the number of cells in a 1mm^2 surface area of tissue.

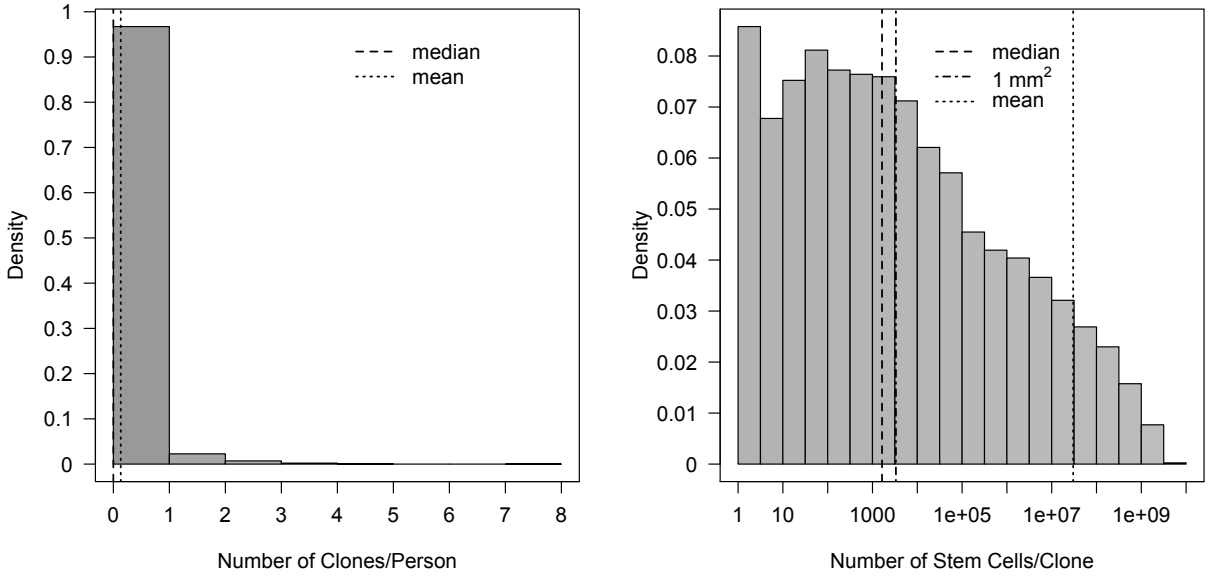

Figure B.5: **Simulated distributions of number and sizes of non-extinct malignant clones, males.** (Left panel) Histogram of number of non-extinct malignant clones in BE segment at time of screening, each originating from unique premalignant ancestor clone. (Right panel) Histogram of number of malignant stem cells in each of the independent malignant clones accounted for in left panel. Example shown for 100K males with BE, age 60 from the 1930 birth cohort. Median and mean values are depicted by dashed and dotted lines, respectively. Assuming $\sigma = 3300$ stem cells/mm^2, the dashed-dotted line on the right graph gives the number of cells in a 1mm^2 surface area of tissue.

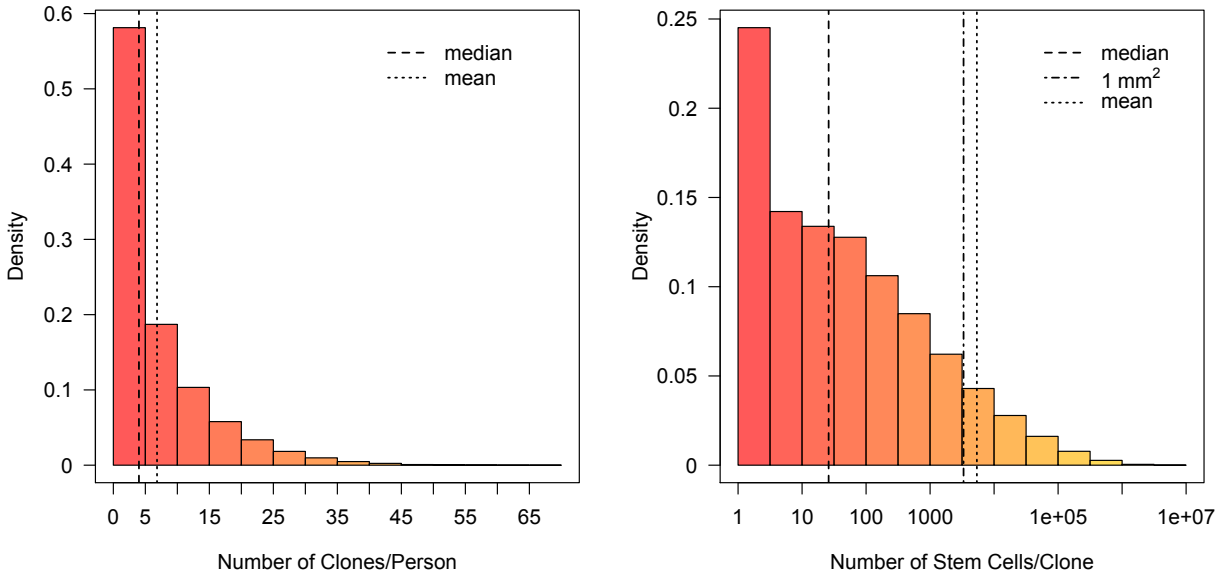

Figure B.6: **Simulated distributions of number and sizes of non-extinct premalignant clones, females.** (Left panel) Histogram of number of non-extinct premalignant clones in BE segment at time of screening. (Right panel) Histogram of number of premalignant stem cells in each of the independent premalignant clones accounted for in left panel. Example shown for 100K females with BE, age 60 from the 1930 birth cohort. Median and mean values are depicted by dashed and dotted lines, respectively. Assuming $\sigma = 3300$ stem cells/mm^2, the dashed-dotted line on the right graph gives the number of cells in a 1mm^2 surface area of tissue.

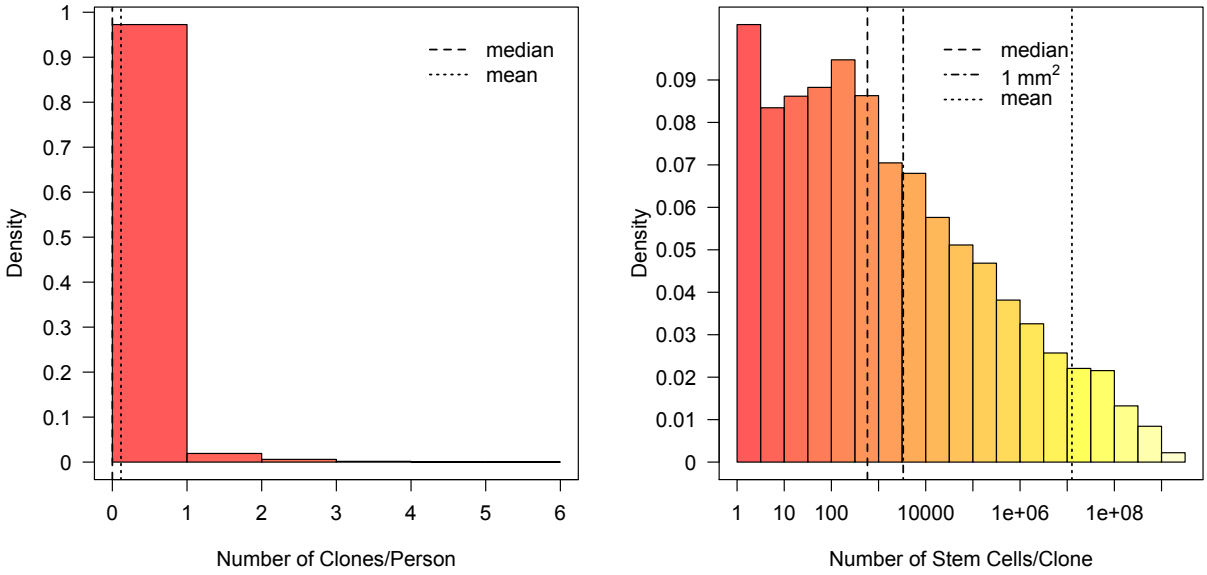

Figure B.7: **Simulated distributions of number and sizes of non-extinct malignant clones, females.** (Left panel) Histogram of number of non-extinct malignant clones in BE segment at time of screening, each originating from unique premalignant ancestor clone. (Right panel) Histogram of number of malignant stem cells in each of the independent malignant clones accounted for in left panel. Example shown for 100K females with BE, age 60 from the 1930 birth cohort. Median and mean values are depicted by dashed and dotted lines, respectively. Assuming $\sigma = 3300$ stem cells/mm^2, the dashed-dotted line on the right graph gives the number of cells in a 1mm^2 surface area of tissue.

Appendix C

Base Case II: Impact of Endoscopic Eradication on EAC

The following are supplemental tables and figures for Chapter 5.

Patient characteristic	Surveillance interval	Source
Surveillance endoscopy interval after CE-IM of HGD patient	3 months for one year then 6 months for one year then annual	[208]
Surveillance endoscopy interval after CE-IM of NDBE patient	every three years	Expert consensus*
Surveillance endoscopy interval after CE-D, non CE-IM of HGD patient	3 months for one year then 6 months for one year then annual	Expert consensus*
Surveillance endoscopy interval after non CE-D, non CE-IM of HGD patient	Every three months	Expert consensus*
Surveillance endoscopy interval after non CE-IM of ND patient	Every three years	Expert consensus*

Table C.1: **Post-treatment surveillance strategies.** *Expert consensus: panel of experts N Shaheen; S Spechler; J Inadomi; C Hur; J Rubenstein.

	Base values		Lower value		Upper value	
Durability of successful treatment						
	Pre-treatment histology					
	NDBE	HGD	NDBE	HGD	NDBE	HGD
Annual Probability of BE recurrence	7.0%	10.0%	3.5%	5.0%	14.0%	20.0%
Efficacy of the initial treatment						
Success of therapy in pre-treatment HGD patients						
CE-IM, CE-D	0.89		0.78		0.94	
non CE-IM, CE-D	0.04		0.07		0.02	
non CE-IM, non CE-D	0.07		0.15		0.04	
Success of therapy in pre-treatment in NDBE patients						
CE-IM	0.97		0.94		0.98	
non CE-IM	0.03		0.06		0.02	
Halting surveillance after a period of observed good health post-treatment						
	Until death or age 80 (follow patients up to age 100)		STOP surveillance when at 5 year remains CE-IM, after CE-IM at initial endoscopic therapy		STOP surveillance when at 10 year remains CE-IM, after CE-IM at initial endoscopic therapy	

Table C.2: **Base Case II input & sensitivity parameters.**

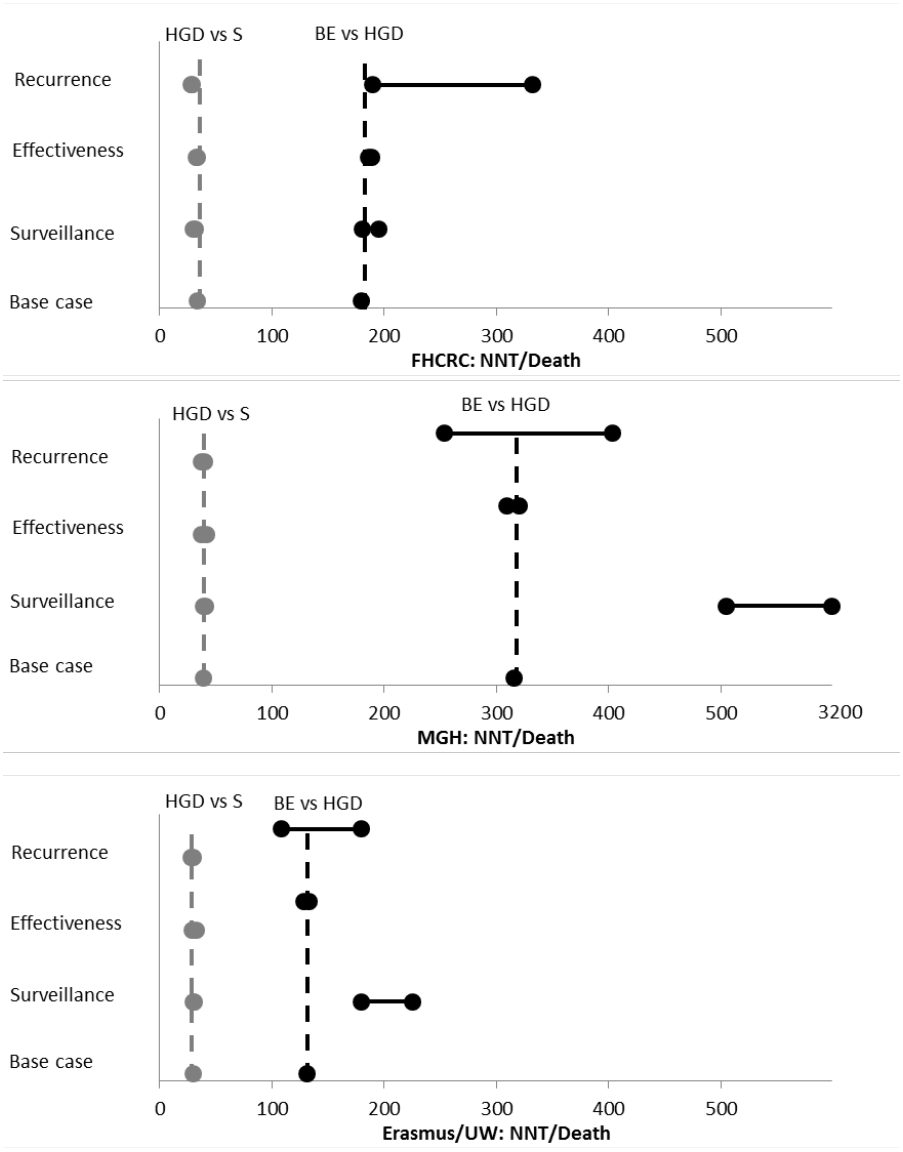

Figure C.1: **Impact of sensitivity analysis on NNT/Death**. — The incremental number needed to treat to prevent 1 EAC death (NNT/Death) for the various sensitivity analyses. The incremental outcomes of two strategies are shown: the additional treatments needed to prevent 1 EAC death in case of HGD treatment compared to the strategy where only surveillance is applied (grey dots). Next to this, the incremental NNT/death when applying treatment to all BE patients compared to treatment for only HGD patients is presented (black dots). Dotted vertical lines represent the base case value. MGH and UW/MISCAN model sensitivity results (including LGD strategy) and are shown for comparison.

Incremental NNT/Death	MSCE-EAC (FHCRC)	
Strategy	Reference strategy	
	S	HGD
HGD	31	
BE	101	177

Table C.3: **Incremental numbers needed to treat to prevent one EAC death per strategy, females.** Results for MSCE-EAC screening model, females, all races, age 60 at index endoscopy.

Age 50-100		
	Reference strategy	
Strategy:	S	HGD
HGD	31	
BE	71	110
Age 60-100		
	Reference strategy	
Strategy:	S	HGD
HGD	34	
BE	99	180
Age 70-100		
	Reference strategy	
Strategy:	S	HGD
HGD	37	
BE	170	412

Table C.4: **Incremental numbers needed to treat to prevent one EAC death per strategy.** MSCE-EAC screening model results for males, all races, surveillance start ages 50, 60 and 70.

Appendix D

Barrett's Esophagus: "How long has that been there?"

The following are supplemental tables referred to in Chapter 7.

Patient.No.	Dataset type	Age at Biopsy	Sex	matched SQ
31	D1: serial	39.00	M	no
32	D1: serial	50.00	M	no
33	D1: serial	44.00	M	no
34	D1: serial	70.00	M	no
35	D1: serial	56.00	M	no
36	D1: serial	55.00	M	no
37	D1: serial	55.00	F	no
38	D1: serial	33.00	M	no
39	D1: serial	35.00	M	no
40	D1: serial	57.00	M	no

Table D.1: **BE patient information for data set D1.**

Patient.No.	Dataset type	Age at Biopsy	Sex	matched SQ
1	D2: cross-sectional	21.00	M	yes
2	D2: cross-sectional	41.00	M	yes
3	D2: cross-sectional	48.00	M	yes
4	D2: cross-sectional	47.00	M	yes
5	D2: cross-sectional	51.00	F	yes
6	D2: cross-sectional	50.00	M	yes
7	D2: cross-sectional	55.00	F	yes
8	D2: cross-sectional	55.00	M	yes
9	D2: cross-sectional	55.00	M	yes
10	D2: cross-sectional	59.00	M	yes
11	D2: cross-sectional	63.00	M	yes
12	D2: cross-sectional	66.00	M	yes
13	D2: cross-sectional	66.00	M	yes
14	D2: cross-sectional	68.00	M	yes
15	D2: cross-sectional	70.00	M	yes
16	D2: cross-sectional	67.00	M	yes
17	D2: cross-sectional	78.00	M	yes
18	D2: cross-sectional	80.00	M	yes
19	D2: cross-sectional	77.00	M	yes
20	D2: cross-sectional	83.00	F	yes
21	D2: cross-sectional	84.00	M	yes
22	D2: cross-sectional	86.00	F	yes
23	D2: cross-sectional	88.00	M	yes
24	D2: cross-sectional	53.00	F	yes
25	D2: cross-sectional	65.00	M	yes
26	D2: cross-sectional	67.00	M	yes
27	D2: cross-sectional	71.00	M	yes
28	D2: cross-sectional	84.00	M	yes
29	D2: cross-sectional	50.00	F	yes
30	D2: cross-sectional	54.00	M	yes

Table D.2: **BE patient information for data set D2.**

Patient.No.	Dataset type	Age at Biopsy	Sex	matched SQ	HGD	EAC
41	D3: familial BE	57.00	M	no		
42	D3: familial BE	71.00	M	no		
43	D3: familial BE	39.00	M	no		
44	D3: familial BE	57.00	M	no		
45	D3: familial BE	70.00	M	no		
46	D3: familial BE	72.00	M	no		
47	D3: familial BE	52.00	M	no		
48	D3: familial BE	84.00	M	no		
49	D3: familial BE	68.00	M	no		
50	D3: familial BE	60.00	F	no		
51	D3: familial BE	56.00	M	no		
52	D3: familial BE	80.00	M	no		
53	D3: familial BE	50.00	M	no		
54	D3: familial BE	66.00	M	no		
55	D3: familial BE	71.00	M	no		
56	D3: familial BE	47.00	M	no		
57	D3: familial BE	44.00	M	no		
58	D3: familial BE	70.00	M	no		
59	D3: familial BE	73.00	M	no		
60	D3: familial BE	72.00	M	no		
61	D3: familial BE	70.00	M	no		
62	D3: familial BE	52.00	M	no		
63	D4: HGD / EAC	49.00	F			yes
64	D4: HGD / EAC	82.00	M		yes	
65	D4: HGD / EAC	79.00	M			yes
66	D4: HGD / EAC	64.00	M			yes
67	D4: HGD / EAC	83.00	M		yes	yes
68	D4: HGD / EAC	53.00	F		yes	yes
69	D4: HGD / EAC	71.00	M	yes	yes	yes
70	D4: HGD / EAC	63.00	M			yes
71	D4: HGD / EAC	77.00	M	yes	yes	yes
72	D4: HGD / EAC	35.00	M	yes	yes	

Table D.3: **BE patient information for data sets D3 and D4.**

	CpG Name	Gene Symbol	ChrX	Island	Promoter	pop. rate	longit. rate	p.value
1	cg03142956	CNTN2	1	NA	TRUE	0.02	0.03	0.00
2	cg02547025	LBH	2	Island	TRUE	0.04	0.04	0.07
3	cg03222971	IGFBP5	2	N_Shore	FALSE	0.03	0.04	0.04
4	cg26988215	GALNT14	2	Island	TRUE	0.04	0.06	0.02
5	cg13501181	EGFEM1P	3	Island	FALSE	0.03	0.03	0.10
6	cg00394358	NRG2	5	Island	FALSE	0.05	0.07	0.01
7	cg11613229	PRR16	5	N_Shore	TRUE	0.03	0.04	0.09
8	cg21986718	SOGA3	6	Island	FALSE	0.02	0.02	0.06
9	cg02215070	AKR1B1	7	Island	TRUE	0.04	0.06	0.05
10	cg14592474	MAGI2	7	N_Shore	TRUE	0.04	0.06	0.05
11	cg19924540	FBXL18	7	Island	FALSE	-0.00	0.03	0.08
12	cg23149470	NA	NA	S_Shore	TRUE	0.01	0.02	0.05
13	cg26473651	CHAT	10	Island	TRUE	0.06	0.06	0.02
14	cg21067341	NA	NA	Island	FALSE	0.05	0.08	0.01
15	cg27319192	NA	NA	Island	FALSE	0.05	0.05	0.03
16	cg05719720	ATP12A	13	Island	FALSE	0.04	0.03	0.02
17	cg08226111	INSM2	14	Island	TRUE	0.06	0.08	0.03
18	cg20312821	IGDCC3	15	Island	FALSE	0.04	0.05	0.07
19	cg02755525	NETO2	16	Island	FALSE	0.07	0.12	0.00
20	cg06755373	NA	NA	Island	FALSE	0.03	0.03	0.09
21	cg07869548	NETO2	16	Island	TRUE	0.04	0.05	0.01
22	cg18671949	LOC728392	17	Island	TRUE	0.03	0.04	0.07
23	cg02724271	TTC9B	19	Island	TRUE	0.06	0.02	0.09
24	cg04572161	ZNF559	19	Island	TRUE	0.04	0.05	0.08
25	cg08375754	TTC9B	19	Island	TRUE	0.02	0.02	0.08
26	cg27388983	ZNF256	19	S_Shore	TRUE	0.05	0.05	0.09
27	cg00992385	HCK	20	Island	TRUE	0.03	0.08	0.01
28	cg24606935	HCK	20	Island	TRUE	0.05	0.07	0.03
29	cg01363596	CACNA1E	1	Island	FALSE	0.03	0.03	0.07
30	cg11882432	GPR88	1	Island	FALSE	0.04	0.03	0.10
31	cg17142105	NA	NA	S_Shore	FALSE	0.02	0.03	0.02
32	cg20785136	PTGER3	1	Island	TRUE	0.02	0.04	0.10
33	cg25582303	NA	NA	Island	FALSE	0.02	0.02	0.07
34	cg26936593	ALPL	1	N_Shore	TRUE	0.00	0.04	0.01
35	cg10279685	NPHP1	2	Island	FALSE	0.03	0.04	0.02

Table D.4: **(Epi)genetic information for BE clock CpG set.** Information for CpGs 1-35 of 70 BE clock set. Reported p-values refer to the longitudinal drift slopes obtained from data set D1.

	CpG Name	Gene Symbol	ChrX	Island	Promoter	pop. rate	longit. rate	p.value
36	cg08514735	WNT5A	3	Island	TRUE	0.04	0.04	0.10
37	cg12268888	FAM198A	3	Island	TRUE	0.03	0.02	0.07
38	cg23162001	CACNA2D2	3	S_Shore	TRUE	0.00	0.04	0.05
39	cg15106296	NA	NA	S_Shore	FALSE	0.01	0.04	0.01
40	cg22830947	NA	NA	S_Shore	FALSE	0.01	0.02	0.10
41	cg00005040	RSPO3	6	N_Shore	TRUE	0.02	0.03	0.04
42	cg05443922	NA	NA	Island	FALSE	0.03	0.03	0.07
43	cg26280666	MSC	8	N_Shore	FALSE	0.03	0.03	0.08
44	cg05090725	RAB39A	11	Island	TRUE	0.03	0.04	0.07
45	cg25603277	PAX6	11	Island	TRUE	0.02	0.02	0.06
46	cg03165343	NA	NA	N_Shore	FALSE	0.04	0.03	0.07
47	cg03272292	C12orf68	12	Island	TRUE	0.01	0.05	0.01
48	cg15441864	NA	NA	Island	FALSE	0.00	0.02	0.04
49	cg18722247	NA	NA	N_Shore	FALSE	0.04	0.04	0.09
50	cg04812484	FAM194B	13	NA	TRUE	0.02	0.03	0.05
51	cg13241077	FAM194B	13	NA	TRUE	0.02	0.03	0.10
52	cg01423820	VASH1	14	Island	TRUE	0.05	0.06	0.00
53	cg07850604	INSM2	14	Island	TRUE	0.04	0.04	0.07
54	cg18489434	VASH1	14	Island	TRUE	0.04	0.05	0.04
55	cg20881054	VASH1	14	Island	TRUE	0.04	0.06	0.00
56	cg24038454	INSM2	14	Island	TRUE	0.04	0.07	0.02
57	cg26140961	NA	NA	NA	FALSE	0.05	0.03	0.11
58	cg21110939	SV2B	15	Island	TRUE	0.04	0.11	0.04
59	cg25475171	SV2B	15	Island	TRUE	0.04	0.08	0.06
60	cg06475764	NETO2	16	Island	FALSE	0.05	0.05	0.03
61	cg16733756	MYH11	16	S_Shore	TRUE	0.02	0.04	0.01
62	cg14355911	NA	NA	Island	FALSE	0.03	0.02	0.09
63	cg21911143	NA	NA	Island	FALSE	0.02	0.03	0.01
64	cg25836567	WNT9B	17	Island	FALSE	0.04	0.04	0.03
65	cg13462028	RAB31	18	Island	TRUE	0.01	0.03	0.03
66	cg00288050	MAST1	19	Island	FALSE	0.04	0.04	0.06
67	cg05119514	ZNF256	19	Island	TRUE	0.05	0.06	0.04
68	cg11409275	NA	NA	Island	FALSE	0.04	0.05	0.02
69	cg13157315	ZNF114	19	N_Shore	TRUE	0.02	0.03	0.10
70	cg22977016	NA	NA	Island	FALSE	0.04	0.08	0.02

Table D.5: **(Epi)genetic information for BE clock CpG set (cont.).** Information for CpGs 36-70 of 70 BE clock set. Reported p-values refer to the longitudinal drift slopes obtained from data set D1.

BIBLIOGRAPHY

[1] Edwards BK, Ward E, Kohler BA, Eheman C, Zauber AG, Anderson RN, et al. Annual report to the nation on the status of cancer, 1975-2006, featuring colorectal cancer trends and impact of interventions (risk factors, screening, and treatment) to reduce future rates. Cancer. 2010;116(3):544–573.

[2] Kohler BA, Ward E, McCarthy BJ, Schymura MJ, Ries LA, Eheman C, et al. Annual report to the nation on the status of cancer, 1975–2007, featuring tumors of the brain and other nervous system. J Natl Cancer I. 2011.

[3] American Cancer Society; 2014. `http://www.cancer.org/`.

[4] Thrift AP, Whiteman DC. The incidence of esophageal adenocarcinoma continues to rise: analysis of period and birth cohort effects on recent trends. Ann Oncol. 2012;23(12):3155–3162.

[5] Hur C, Miller M, Kong CY, Dowling EC, Nattinger KJ, Dunn M, et al. Trends in esophageal adenocarcinoma incidence and mortality. Cancer. 2013;119(6):1149–1158.

[6] Kong CY, Kroep S, Curtius K, Hazelton WD, Jeon J, Meza R, et al. Exploring the Recent Trend in Esophageal Adenocarcinoma Incidence and Mortality Using Comparative Simulation Modeling. Cancer Epidemiol Biomarkers Prev. 2014;23(6):997–1006.

[7] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) SEER*Stat Database: Incidence - SEER 9 Regs Research Data, Nov 2013 Sub (1973-2011) Katrina/Rita Population Adjustment - Linked To County Attributes - Total U.S., 1969-2012 Counties, National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch, released April 2014, based on the November 2013 submission.

[8] BE vs. Normal squamous; 2015. `http://www.eugenegi.com/procedures/treatement-of-barretts-esophagus`.

[9] Howlander N, Noone A, Krapcho M, Garshell J, Miller D, Altekruse S, et al.. SEER cancer statistics review, 1975–2011. Bethesda, MD: National Cancer Institute; 2014.

[10] Bhat SK, McManus DT, Coleman HG, Johnston BT, Cardwell CR, McMenamin Ú, et al. Oesophageal adenocarcinoma and prior diagnosis of Barrett's oesophagus: a population-based study. Gut. 2014;p. gutjnl–2013.

[11] Vaughan TL, Fitzgerald RC. Precision prevention of oesophageal adenocarcinoma. Nat Rev Gastroentero. 2015;12(4):243–248.

[12] Prasad G, Bansal A, Sharma P, Wang K. Predictors of progression in Barrett's esophagus: current knowledge and future directions. Am J Gastroenterol. 2010;105:1490–1502.

[13] Luebeck E, Curtius K, Jeon J, Hazelton W. Impact of tumor progression on cancer incidence curves. Cancer Res. 2013;73(3):1086–1096.

[14] Curtius K, Hazelton W, Jeon J, Luebeck E. A Multiscale Model Evaluates Screening for Neoplasia in Barrett's Esophagus. PLoS Comput Biol. 2015;11(5):e1004272.

[15] Kroep S, Curtius K, Heberle C, Wolf W, Shaheen N, Spechler S, et al. The Impact of Endoscopic Eradication for Barrett's Esophagus on Esophageal Adenocarcinoma Incidence and Mortality: A Comparative Modeling Analysis. 2015 (in preparation).

[16] Curtius K, Luebeck E, Grady W, Hazelton W, Wang K, Chak A, et al. Barrett's Esophagus: "How long has that been there?". 2015 (in preparation).

[17] Force UPST, et al. Screening for breast cancer: US Preventive Services Task Force recommendation statement. Ann Intern Med. 2009;151(10).

[18] Pignone M, Saha S, Hoerger T, Mandelblatt J. Cost-effectiveness analyses of colorectal cancer screening: a systematic review for the US Preventive Services Task Force. Ann Intern Med. 2002;137(2):96–104.

[19] Moyer VA. Screening for lung cancer: US Preventive Services Task Force recommendation statement. Ann Intern Med. 2014;160(5):330–338.

[20] Knudsen AB, Lansdorp-Vogelaar I, Rutter CM, Savarino JE, van Ballegooijen M, Kuntz KM, et al. Cost-effectiveness of computed tomographic colonography screening for colorectal cancer in the medicare population. J Natl Cancer I. 2010;102(16):1238–1252.

[21] Lansdorp-Vogelaar I, Kuntz KM, Knudsen AB, Wilschut JA, Zauber AG, van Ballegooijen M. Stool DNA testing to screen for colorectal cancer in the Medicare population: a cost-effectiveness analysis. Ann Intern Med. 2010;153(6):368–377.

[22] CISNET Model Profiles; 2015. http://cisnet.cancer.gov/.

[23] Hennings H, Shores R, Wenk ML, Spangler EF, Tarone R, Yuspa SH. Malignant conversion of mouse skin tumours is increased by tumour initiators and unaffected by tumour promoters. Nature. 1983;304.

[24] Kopp-Schneider A. Carcinogenesis models for risk assessment. Stat Methods Med Res. 1997;6(4):317–340.

[25] Tan W. Stochastic Models of Carcinogenesis, first edition. New York, NY: Marcel Decker; 1991.

[26] Whittemore A, Keller JB. Quantitative theories of carcinogenesis. SIAM Rev. 1978;20(1):1–30.

[27] Tan WY. Handbook of cancer models with applications. vol. 9. World Scientific; 2008.

[28] Muller HJ. Radiation damage to the genetic material. Am Sci. 1950;38(1).

[29] Nordling C. A new theory on the cancer-inducing mechanism. Brit J Cancer. 1953;7(1).

[30] Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Brit J Cancer. 1954;8(1).

[31] Armitage P, Doll R. A two-stage theory of carcinogenesis in relation to the age distribution of human cancer. Brit J Cancer. 1957;11(2).

[32] Fisher J. Multiple-mutation theory of carcinogenesis. Nature. 1958;181:651–652.

[33] Kendall DG. Birth-and-death processes, and the theory of carcinogenesis. Biometrika. 1960;p. 13–21.

[34] Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A. 1971;68(4):820–823.

[35] Friend SH, Bernards R, Rogelj S, Weinberg RA, Rapaport JM, Albert DM, et al. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. Nature. 1986;323(6089):643–646.

[36] Moolgavkar S, Venzon D. Two-event models for carcinogenesis: incidence curves for childhood and adult tumors. Math Biosci. 1979;47:55–77.

[37] Moolgavkar S, Knudson A. Mutation and cancer: a model for human carcinogenesis. J Natl Cancer I. 1981;66:1037–1052.

[38] Moolgavkar SH, Luebeck EG. Multistage Carcinogenesis: Population-Based Model for Colon Cancer. J Natl Cancer I. 1992;84(8):610–618.

[39] Kai M, Luebeck EG, Moolgavkar SH. Analysis of the incidence of solid cancer among atomic bomb survivors using a two-stage model of carcinogenesis. Radiat Res. 1997;148(4):348–358.

[40] Heidenreich WF, Luebeck EG, Moolgavkar SH. Effects of exposure uncertainties in the TSCE model and application to the Colorado miners data. Radiat Res. 2009;161(1):72–81.

[41] Luebeck E, Moolgavkar S. Multistage carcinogenesis and the incidence of colorectal cancer. Proc Natl Acad Sci U S A. 2002;99:15095–15100.

[42] Hazelton WD, Luebeck EG, Heidenreich WF, Moolgavkar SH. Analysis of a historical cohort of Chinese tin miners with arsenic, radon, cigarette smoke, and pipe smoke exposures using the biologically based two-stage clonal expansion model. Radiat Res. 2001;156:78–94.

[43] Curtis S, Luebeck E, Hazelton W, Moolgavkar S. A new perspective of carcinogenesis from protracted high-LET radiation arises from the two-stage clonal expansion model. Adv Space Res. 2002;30(4):937–944.

[44] Hazelton WD, Clements MS, Moolgavkar SH. Multistage carcinogenesis and lung cancer mortality in three cohorts. Cancer Epidemiol Biomarkers Prev. 2005;14(5):1171–1181.

[45] Meza R, Hazelton WD, Colditz GA, Moolgavkar SH. Analysis of lung cancer incidence in the Nurses' Health and the Health Professionals' Follow-Up Studies using a multistage carcinogenesis model. Cancer Cause Control. 2008;19(3):317–328.

[46] Moolgavkar S, Luebeck E, de Gunst M, Port R, Schwarz M. Quantitative analysis of enzyme-altered foci in rat hepatocarcinogenesis experiments–I. Single agent regimen. Carcinogenesis. 1990;11:1271–1278.

[47] Luebeck EG, Curtis SB, Cross FT, Moolgavkar SH. Two-stage model of radon-induced malignant lung tumors in rats: effects of cell killing. Radiat Res. 1996;145(2):163–173.

[48] Luebeck EG, Buchmann A, Schneider D, Moolgavkar SH, Schwarz M. Modulation of liver tumorigenesis in Connexin32-deficient mouse. Mutat Res-Fund Mol M. 2005;570(1):33–47.

[49] Little M. Are two mutations sufficient to cause cancer? Some generalizations of the two-mutation model of carcinogenesis of Moolgavkar, Venzon, and Knudson, and of the multistage model of Armitage and Doll. Biometrics. 1995;51:1278–1291.

[50] Jeon J, Luebeck E, Moolgavkar S. Age effects and temporal trends in adenocarcinoma of the esophagus and gastric cardia (United States). Cancer Cause Control. 2006;17:971–981.

[51] Meza R, Jeon J, Moolgavkar S, Luebeck E. Age-specific incidence of cancer: Phases, transitions, and biological implications. Proc Natl Acad Sci U S A. 2008;105:16284–16289.

[52] Meza R, Jeon J, Renehan A, Luebeck E. Colorectal cancer incidence trends in the United States and United kingdom: evidence of right- to left-sided biological gradients with implications for screening. Cancer Res. 2010;70:5419–5429.

[53] Dewanji A, Jeon J, Meza R, Luebeck G. Number and Size distribution of Colorectal Adenomas under the Multistage Clonal Expansion Model of Cancer. PLoS Comput Biol. 2011;7(10).

[54] Fakir H, Tan WY, Hlatky L, Hahnfeldt P, Sachs RK. Stochastic population dynamic effects for lung cancer progression. Radiat Res. 2009;172(3):383–393.

[55] Fakir H, Hofmann W, Sachs RK. Modeling progression in radiation-induced lung adenocarcinomas. Radiat Environ Bioph. 2010;49(2):169–176.

[56] Dewanji A, Moolgavkar S, Luebeck E. Two-mutation model for carcinogenesis: joint analysis of premalignant and malignant lesions. Math Biosci. 1991;104:97–109.

[57] Little MP, Heidenreich WF, Li G. Parameter identifiability and redundancy in a general class of stochastic carcinogenesis models. PLoS One. 2009;4(12):e8520.

[58] Cahill D, Kinzler K, Vogelstein B, Lengauer C. Genetic instability and darwinian selection in tumours. Trends Cell Biol. 1999;9:M57–60.

[59] Moffett B, Baban D, Bao L, Tarin D. Fate of clonal lineages during neoplasia and metastasis studied with an incorporated genetic marker. Cancer Res. 1992;52:1737–1743.

[60] Beerenwinkel N, Antal T, Dingli D, Traulsen A, Kinzler KW, Velculescu VE, et al. Genetic Progression and the Waiting Time to Cancer. PLoS Comput Biol. 2007 11;3(11):e225.

[61] Powell SM, Zilz N, Beazer-Barclay Y, Bryan TM, Hamilton SR, Thibodeau SN, et al. APC mutations occur early during colorectal tumorigenesis. Nature. 1992 09;359(6392):235–237.

[62] Karaman A, Kabalar M, Binici DN, Oztürk C, Pirim I. Genetic alterations in gastric precancerous lesions. Genetic counseling (Geneva, Switzerland). 2009;21(4):439–450.

[63] Morton JP, Timpson P, Karim SA, Ridgway RA, Athineos D, Doyle B, et al. Mutant p53 drives metastasis and overcomes growth arrest/senescence in pancreatic cancer. Proc Natl Acad Sci U S A. 2010;107(1):246–251.

[64] Norfadzilah M, Pailoor J, Retneswari M, Chinna K, Noor L. P53 expression in invasive pancreatic adenocarcinoma and precursor lesions. The Malaysian journal of pathology. 2011;33(2):89.

[65] Harris T. The Theory of Branching Processes. New York, NY: Dover Publications, Inc; 1989.

[66] O'Brien CA, Pollett A, Gallinger S, Dick JE. A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. Nature. 2007;445(7123):106–110.

[67] Crump C, Subramaniam R, Van Landingham C. A Numerical Solution to the Nonhomogeneous Two-Stage MVK Model of Cancer. Risk Anal. 2005;25:921–926.

[68] Heidenreich W, Luebeck E, Moolgavkar S. Some Properties of the Hazard Function of the Two-Mutation Clonal Expansion Model. Risk Anal. 1997;17:391–399.

[69] Teixeira CR, Haruma K, Teshima H, Yoshihara M, Sumii K, Kajiyama G. Endoscopic therapy for gastric cancer in patients more than 80 years old. Am J Gastroenterol. 1991;86(6):725–728.

[70] Nishida K, Kaneko T, Yoneda M, Nakagawa S, Ishikawa T, Yamane E, et al. Doubling time of serum CA 19-9 in the clinical course of patients with pancreatic cancer and its significant association with prognosis. J Surg Oncol. 1999;71(3):140–146.

[71] Bolin S, Nilsson E, Sjödahl R. Carcinoma of the colon and rectum–growth rate. Ann Surg. 1983;198(2):151.

[72] Balmain A, Barrett JC, Moses H, Renan MJ. How many mutations are required for tumorigenesis? Implications from human cancer data. Mol Carcinogen. 1993;7(3):139–146.

[73] Luebeck EG, Buchmann A, Stinchcombe S, Moolgavkar SH, Schwarz M. Effects of 2, 3, 7, 8-tetrachlorodibenzo-p-dioxin on initiation and promotion of GST-P-positive foci in rat liver: A quantitative analysis of experimental data using a stochastic model. Toxicol Appl Pharm. 2000;167(1):63–73.

[74] Luebeck EG, Heidenreich WF, Hazelton WD, Paretzke HG, Moolgavkar SH. Biologically based analysis of the data for the Colorado uranium miners cohort: age, dose and dose-rate effects. Radiat Res. 1999;152(4):339–351.

[75] Holford TR, Zhang Z, McKay LA. Estimating age, period and cohort effects using the multistage model for cancer. Stat Med. 1994;13(1):23–41.

[76] Muto T, Bussey H, Morson B. The evolution of cancer of the colon and rectum. Cancer. 1975;36(6):2251–2270.

[77] Brentnall TA, Lai LA, Coleman J, Bronner MP, Pan S, Chen R. Arousal of cancer-associated stroma: overexpression of palladin activates fibroblasts to promote tumor invasion. PLoS One. 2012;7(1):e30219.

[78] Jones S, Chen Wd, Parmigiani G, Diehl F, Beerenwinkel N, Antal T, et al. Comparative lesion sequencing provides insights into tumor evolution. Proc Natl Acad Sci U S A. 2008;105(11):4283–4288.

[79] Moolgavkar SH. Carcinogenesis models: An overview. In: Physical and Chemical Mechanisms in Molecular Radiation Biology. Springer; 1991. p. 387–399.

[80] Xu X. The biological foundation of the Gompertz model. Int J Biomed Comput. 1987;20(1):35–39.

[81] Brosnan JA, Iacobuzio-Donahue CA. A new branch on the tree: next-generation sequencing in the study of cancer evolution. In: Semin Cell Dev Biol. vol. 23. Elsevier; 2012. p. 237–242.

[82] Martens EA, Kostadinov R, Maley CC, Hallatschek O. Spatial structure increases the waiting time for cancer. New J Phys. 2011;13(11):115014.

[83] Cameron A, Zinsmeister A, Ballard D, Carney J. Prevalence of columnar-lined (Barrett's) esophagus. Comparison of population-based clinical and autopsy findings. Gastroenterology. 1990;99(4):918–922.

[84] Taylor J, Rubenstein J. Meta-Analyses of the Effect of Symptoms of Gastroesophageal Reflux on the Risk of Barrett's Esophagus. Am J Gastroenterol. 2010;105(8):1730–1737.

[85] Falk GW, Jacobson BC, Riddell RH, Rubenstein JH, El-Zimaity H, Drewes AM, et al. Barrett's esophagus: prevalence-incidence and etiology-origins. Ann NY Acad Sci. 2011;1232(1):1–17.

[86] Thrift A, Kramer J, Qureshi Z, Richardson P, El-Serag H. Age at Onset of GERD Symptoms Predicts Risk of Barrett's Esophagus. Am J Gastroenterol. 2013;108(6):915–922.

[87] Ruigomez A, Wallander M, Lundborg P, Johansson S, Garcia Rodriguez L. Gastroesophageal reflux disease in children and adolescents in primary care. Scan J Gastroenterology. 2010;45:139–146.

[88] Ruigomez A, Garcia Rodriguez L, Wallander M, Johansson S, Graffner H, Dent J. Natural history of gastro-oesophageal reflux disease diagnosed in general practice. Aliment Pharmacol Ther. 2004;20:751–760.

[89] Moolgavkar S, Dewanji A, Venzon D. A Stochastic Two-Stage Model for Cancer Risk Assessment. II. The Number and Size of Premalignant Clones. The Hazard Function and the Probability of Tumor. Risk Anal. 1988;8(3):383–392.

[90] Hazelton WD, Curtius K, Inadomi JM, Vaughan TL, Meza R, Rubenstein JH, et al. The role of gastroesophageal reflux and other factors during progression to esophageal adenocarcinoma. Cancer Epidemiol Biomarkers Prev. 2015;p. cebp–0323.

[91] Pohl H, Sirovich B, Welch HG. Esophageal Adenocarcinoma Incidence: Are We Reaching the Peak? Cancer Epidemiol Biomarkers Prev. 2010;19(6):1468–1470.

[92] Chu KC, Miller BA, Feuer EJ, Hankey BF. A method for partitioning cancer mortality trends by factors associated with diagnosis: An application to female breat cancer. J Clin Epidemiol. 1994;47(12):1451–1461.

[93] US Census Bureau; 2015. `http://www.census.gov/population/www/projections`.

[94] Luebeck E. R Bhat package vignette; 2015. http://cran.r-project.org/web/packages/Bhat/.

[95] Hvid-Jensen F, Pedersen L, Drewes AM, Sørensen HT, Funch-Jensen P. Incidence of adenocarcinoma among patients with Barrett's esophagus. New Engl J Med. 2011;365(15):1375–1383.

[96] Desai TK, Krishnan K, Samala N, Singh J, Cluley J, Perla S, et al. The incidence of oesophageal adenocarcinoma in non-dysplastic Barrett's oesophagus: a meta-analysis. Gut. 2012;61(7):970–976.

[97] Sharma P, Falk G, Weston A, Reker D, Johnston M, Sampliner R. Dysplasia and Cancer in a Large Multicenter Cohort of Patients With Barrett's Esophagus. Clin Gastroeneterol H. 2006;4:566–572.

[98] Cook MB. Editorial: Optimization and Expansion of Predictive Models for Barrett's Esophagus and Esophageal Adenocarcinoma: Could a Life-Course Exposure History Be Beneficial? Am J Gastroenterol. 2013;108(6):923–925.

[99] Shaheen NJ. Editorial: Predicting the Risk of Barrett's Esophagus: Does the BERET Fit&quest. Am J Gastroenterol. 2013;108(3):363–365.

[100] Reid BJ, Li X, Galipeau PC, Vaughan TL. Barrett's oesophagus and oesophageal adenocarcinoma: time for a new synthesis. Nat Rev Cancer. 2010 2010 Feb;10:87–101.

[101] Barrett M, Sanchez C, Prevo L, Wong D, Galipeau P, Paulson T, et al. Evolution of neoplastic cell lineages in Barrett oesophagus. Nat Genet. 1999;22:106–109.

[102] Maley C. Multistage carcinogenesis in Barrett's esophagus. Cancer Lett. 2007;245:22–32.

[103] Neshat K, Sanchez C, Galipeau P, Cowan D, Ramel S, Levine D, et al. Barrett's esophagus: A model of human neoplastic progression. Cold Spring Harb Symp Quant Biol. 1994;59:577–583.

[104] Paulson T, Maley C, Xiaohong L. Chromosomal Instability and Copy Number Alterations in Barrett's esophagus and Esophageal Adenocarcinoma. Clin Cancer Res. 2009;15(10):3305–3314.

[105] Ronkainen J, Aro P, Storskrubb T, Johansson S, Lind T, Bolling-Sternevald E, et al. Prevalence of Barrett's esophagus in the general population: an endoscopic study. Gastroenterology. 2005;129(6):1825–1831.

[106] Gillespie D. Exact stochastic simulation of coupled chemical reactions. J Phys Chem. 1977;81(25):2340–2361.

[107] Reya T, Clevers H. Wnt signaling in stem cells and cancer. Nature. 2005;434:843–850.

[108] Humphries A, Wright N. Colonic crypt organization and tumorigenesis. Nature. 2008;8:415–424.

[109] Kostadinov RL, Kuhner MK, Li X, Sanchez CA, Galipeau PC, Paulson TG, et al. NSAIDs modulate clonal evolution in Barrett's esophagus. PLoS Genet. 2013;9(6):e1003553.

[110] Nicolas P, Kim K, Shibata D, Tavare S. The Stem Cell Population of the Human Colon Crypt: Analysis via Methylation Patterns. PLoS Comput Biol. 2007;3(3):e28.

[111] Calabrese P, Tavare S, Shibata D. The Stem Cell Population of the Human Colon Crypt: Analysis via Methylation Patterns. Am J Pathol. 2004;164(4).

[112] Baker AM, Cereser B, Melton S, Fletcher AG, Rodriguez-Justo M, Tadrous PJ, et al. Quantification of Crypt and Stem Cell Evolution in the Normal and Neoplastic Human Colon. Cell Rep. 2014 August;8:940–947.

[113] Miros M, Kerlin P, Walker N. Only patients with dysplasia progress to adenocarcinoma in Barrett's oesophagus. Gut. 1991;32:1441–1446.

[114] Abela J, Going J, Mackenzie J, McKernan M, O'Mahoney M, Stuart R. Systematic four-quadrant biopsy detects Barrett's dysplasia in more patients than nonsystematic biopsy. Am J Gastroenterol. 2008;103:850–855.

[115] Peters F, Curvers W, Rosmolen W, de Vries C, Ten Kate F, Krishnadath K, et al. Surveillance history of endoscopically treated patients with early Barrett's neoplasia: nonadherence to the Seattle biopsy protocol leads to sampling error. Dis Esophagus. 2008;21:475–479.

[116] Reid B, Weinstein W, Lewin K, Haggit R, VanDeventer G, DenBesten L, et al. Endoscopic biopsy can detect high-grade dysplasia or early adenocarcinoma in Barrett's esophagus without grossly recognizable neoplastic lesions. Gastroenterology. 1988;94:81–90.

[117] Abrams JA, Kapel RC, Lindberg GM, Saboorian MH, Genta RM, Neugut AI, et al. Adherence to biopsy guidelines for Barrett's esophagus surveillance in the community setting in the United States. Clin Gastroenterol Hepatol. 2009;7(7):736–742.

[118] O'Connor J, Falk G, Richter J. The Incidence of Adenocarcinoma and Dysplasia in Barrett's Esophagus Report on the Cleveland Clinic Barrett's Esophagus Registry. Am J Gastroenterol. 1999;94(8):2037–2042.

[119] Conio M, Blanchi S, Lapertosa G, Ferraris R, Sablich R, Marchi S, et al. Long-Term Endoscopic Surveillance of Patients With Barrett's Esophagus. Incidence of Dysplasia and Adenocarcinoma: A Prospective Study. Am J Gastroenterol. 2003;98(9):1931–1939.

[120] Guardino J, Khandwala F, Lopez R, Wachsberger D, Richter J, Falk G. Barrett's Esophagus at a Tertiary Care Center: Association of Age on Incidence and Prevalence of Dysplasia and Adenocarcinoma. Am J Gastroenterol. 2006;101:2187–2193.

[121] Weston A, Badr A, Hassanein R. Prospective Multivariate Analysis of Clinical, Endoscopic, and Histological Factors Predictive of the Development of Barrett's Multifocal High-Grade Dysplasia or Adenocarcinoma. Am J Gastroenterol. 1999;94(12):3413–3419.

[122] Jeon J, Meza R, Moolgavkar S, Luebeck G. Evaluation of screening strategies for premalignant lesions using a biomathematical approach. Math Biosci. 2008;213:56–70.

[123] Reid B, Blount P, Feng Z. Optimizing endoscopic biopsy detection of early cancers in Barrett's high-grade dysplasia. Am J Gastroenterol. 2000;95(11):3089–3096.

[124] Tschanz E. Do 40% of Patients Resected for Barrett Esophagus With High-Grade Dysplasia Have Unsuspected Adenocarcinoma? Arch Pathol Lab Med. 2005;129:177–180.

[125] Cameron A, Carpenter H. Barrett's esophagus, high-grade dysplasia, and early adenocarcinoma: A pathological study. Am J Gastroenterol. 1997;92:586–591.

[126] Falk G, Rice T, Goldblum J, Richter J. Jumbo biopsy forceps protocol still misses unsuspected cancer in Barrett's esophagus with high-grade dysplasia. Gastrointest Endosc. 1999;49(2):170–176.

[127] Peters JH, Clark G, Ireland AP, Chandrasoma P, Smyrk TC, DeMeester TR. Outcome of adenocarcinoma arising in Barrett's esophagus in endoscopically surveyed and nonsurveyed patients. J Thorac Cardiovasc Surg. 1994;108(5):813–21.

[128] Konda V, Waxman I. Low risk of prevalent submucosal invasive cancer among patients undergoing esophagectomy for treatment of Barrett's esophagus with high grade dysplasia. J Gastrointest Oncol. 2011;2(1):1–4.

[129] Heitmiller R, Redmond M, Hamilton S. Barrett's Esophagus with High-Grade Dysplasia, An Indication for Prophylactic Esophagectomy. Ann Surg. 1996;224(1):66–71.

[130] Zaninotto G, Bennett C. Surveillance for Low-Grade Dysplastic Barrett's Oesophagus: One Size Fits All? World J Surg. 2014;p. 1–8.

[131] Fleischer DE, Overholt BF, Sharma VK, Reymunde A, Kimmey MB, Chuttani R, et al. Endoscopic ablation of Barrett's esophagus: a multicenter study with 2.5-year follow-up. Gastrointest Endosc. 2008;68(5):867–876.

[132] Wani S, Puli SR, Shaheen NJ, Westhoff B, Slehria S, Bansal A, et al. Esophageal adenocarcinoma in Barrett's esophagus after endoscopic ablative therapy: a meta-analysis and systematic review. Am J Gastroenterol. 2009;104(2):502–513.

[133] Rastogi A, Puli S, El-Serag HB, Bansal A, Wani S, Sharma P. Incidence of esophageal adenocarcinoma in patients with Barrett's esophagus and high-grade dysplasia: a meta-analysis. Gastrointest Endosc. 2008;67(3):394–398.

[134] Association AG, et al. American Gastroenterological Association medical position statement on the management of Barrett's esophagus. Gastroenterology. 2011;140(3):1084–1091.

[135] Phoa KN, van Vilsteren FG, Weusten BL, Bisschops R, Schoon EJ, Ragunath K, et al. Radiofrequency ablation vs endoscopic surveillance for patients with Barrett esophagus and low-grade dysplasia: a randomized clinical trial. Jama- J Am Med Assoc. 2014;311(12):1209–1217.

[136] Oliphant Z, Snow A, Knight H, Barr H, Almond L. Endoscopic resection with or without mucosal ablation of high grade dysplasia and early oesophageal adenocarcinoma–Long term follow up from a regional UK centre. Int J Surg. 2014;12(11):1148–1150.

[137] Qumseya B, David W, McCrum M, Dong Y, Raimondo M, Woodward TA, et al. Gender Differences in Remission of Esophageal Intestinal Metaplasia after Radiofrequency Ablation. Am J Gastroenterol. 2014;109(3):369–374.

[138] Ertan A, Zaheer I, Correa AM, Thosani N, Blackmon SH. Photodynamic therapy vs radiofrequency ablation for Barrett's dysplasia: efficacy, safety and cost-comparison. World journal of gastroenterology: WJG. 2013;19(41):7106.

[139] Anders M, Bähr C, El-Masry MA, Marx AH, Koch M, Seewald S, et al. Long-term recurrence of neoplasia and Barrett's epithelium after complete endoscopic resection. Gut. 2014;p. gutjnl–2013.

[140] Orman ES, Li N, Shaheen NJ. Efficacy and durability of radiofrequency ablation for Barrett's Esophagus: Systematic review and meta-analysis. Clin Gastroeneterol H. 2013;11(10):1245–1255.

[141] Haidry RJ, Dunn JM, Butt MA, Burnell MG, Gupta A, Green S, et al. Radiofrequency ablation and endoscopic mucosal resection for dysplastic barrett's esophagus and early esophageal adenocarcinoma: outcomes of the UK National Halo RFA Registry. Gastroenterology. 2013;145(1):87–95.

[142] Gupta M, Iyer PG, Lutzke L, Gorospe EC, Abrams JA, Falk GW, et al. Recurrence of esophageal intestinal metaplasia after endoscopic mucosal resection and radiofrequency ablation of Barrett's esophagus: results from a US Multicenter Consortium. Gastroenterology. 2013;145(1):79–86.

[143] Odze R. Diagnosis and grading of dysplasia in Barrett's oesophagus. J Clin Pathol. 2006;59:1029–1038.

[144] Schlemper R, Riddell R, Kato Yea, Borchard F, Cooper H, Dawsey S, et al. The Vienna classification of gastrointestinal epithelial neoplasia. Gut. 2000;47(2):251–255.

[145] Reid B, Haggitt R, Rubin C, Roth G, Surawicz C, Van Belle G, et al. Observer variation in the diagnosis of dysplasia in Barrett's esophagus. Hum Pathol. 1988;19(2):166–178.

[146] Montgomery E. Is there a way for pathologists to decrease interobserver variability in the diagnosis of dysplasia? Arch Pathol Lab Med. 2005;129(2):174–176.

[147] Shaheen NJ, Overholt BF, Sampliner RE, Wolfsen HC, Wang KK, Fleischer DE, et al. Durability of radiofrequency ablation in Barrett's esophagus with dysplasia. Gastroenterology. 2011;141(2):460–468.

[148] Fleischer D, Overholt B, Sharma V, Reymunde A, Kimmey M, Chuttani R, et al. Endoscopic radiofrequency ablation for Barrett's esophagus: 5-year outcomes from a prospective multicenter trial*. Endoscopy. 2010;42(10):781.

[149] Hur C, Choi SE, Rubenstein JH, Kong CY, Nishioka NS, Provenzale DT, et al. The cost effectiveness of radiofrequency ablation for Barrett's esophagus. Gastroenterology. 2012;143(3):567–575.

[150] Shaheen NJ, Sharma P, Overholt BF, Wolfsen HC, Sampliner RE, Wang KK, et al. Radiofrequency ablation in Barrett's esophagus with dysplasia. New Engl J Med. 2009;360(22):2277–2288.

[151] Wolf WA, Overholt BF, Li N, Lightdale CJ, Cotton CC, Wolfsen HC, et al. 778 Durability of Radiofrequency Ablation (RFA) in Barrett's Esophagus With Dysplasia: the AIM Dysplasia Trial At Five Years. Gastroenterology. 2014;146(5):S–131.

[152] Wolf WA, Pruitt RE, Ertan A, Cotton CC, Komanduri S, Pasricha S, et al. Sa1452 Predictors of Esophageal Adenocarcinoma in Patients With Prior Radiofrequency Ablation (RFA) for Treatment of Barrett's Esophagus: Results From the US RFA Registry. Gastrointest Endosc. 2014;79(5):AB217.

[153] Yu B, Tiwari RC, Cronin KA, McDonald C, Feuer EJ. CANSURV: a Windows program for population-based cancer survival analysis. Comput Meth Prog Bio. 2005;80(3):195–203.

[154] Wani S, Drahos J, Cook MB, Rastogi A, Bansal A, Yen R, et al. Comparison of endoscopic therapies and surgical resection in patients with early esophageal cancer: a population-based study. Gastrointest Endosc. 2014;79(2):224–232.

[155] Arias E. United States life tables, 2008. National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System. 2012;61(3):1–63.

[156] Kastelein F, van Olphen S, Steyerberg E, Sikkema M, Spaander M, Looman C, et al. Surveillance in patients with long-segment Barrett's oesophagus: a cost-effectiveness analysis. Gut. 2014;p. gutjnl–2014.

[157] Gordon LG, Hirst NG, Mayne GC, Watson DI, Bright T, Cai W, et al. Modeling the cost-effectiveness of strategies for treating esophageal adenocarcinoma and high-grade dysplasia. J Gastrointest Surg. 2012;16(8):1451–1461.

[158] Dewanji A, Biswas A. An optimal design for simple illness-death model. J Stat Plan Infer. 2001;96:289–300.

[159] Hanin L, Pavlova L. Optimal screening schedules for prevention of metastatic cancer. Stat Med. 2013;32:206–219.

[160] Eads CA, Lord RV, Kurumboor SK, Wickramasinghe K, Skinner ML, Long TI, et al. Fields of aberrant CpG island hypermethylation in Barrett's esophagus and associated adenocarcinoma. Cancer Research. 2000;60(18):5021–5026.

[161] Barrett MT, Sanchez CA, Prevo LJ, Wong DJ, Galipeau PC, Paulson TG, et al. Evolution of neoplastic cell lineages in Barrett oesophagus. Nat Genet. 1999;22(1):106–109.

[162] Jin Z, Cheng Y, Gu W, Zheng Y, Sato F, Mori Y, et al. A Multicenter, Double-Blinded Validation Study of Methylation Biomarkers for Progression Prediction in Barrett's Esophagus. Cancer Res. 2009;10(69):4112–4115.

[163] Alvarez H, Opalinska J, Zhou L, Sohal D, Fazzari MJ, Yu Y, et al. Widespread hypomethylation occurs early and synergizes with gene amplification during esophageal carcinogenesis. PLoS Genet. 2011;7(3):e1001356.

[164] Kaz A, Grady W. Epigenetic biomarkers in esophageal cancer. Cancer Lett. 2014;342(2):193–199.

[165] Issa JPJ, Ottaviano YL, Celano P, Hamilton SR, Davidson NE, Baylin SB. Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. Nature genetics. 1994;7(4):536–540.

[166] Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. Mol Cell. 2013;49(2):359–367.

[167] Horvath S. DNA methylation age of human tissues and cell types. Genome Biol. 2013;14(R115).

[168] Portale G, Hagen J, Peters J, Chan L, DeMeester S, Gandamihardja T, et al. Modern 5-Year Survival of Resectable Esophageal Adenocarcinoma: Single Institution Experience with 263 Patients. J Am Coll Surg. 2006;202(4):588–596.

[169] Maley CC, Galipeau PC, Li X, Sanchez CA, Paulson TG, Reid BJ. Selectively advantageous mutations and hitchhikers in neoplasms p16 lesions are selected in Barrett's Esophagus. Cancer Res. 2004;64(10):3414–3427.

[170] Reid BJ, Prevo LJ, Galipeau PC, Sanchez CA, Longton G, Levine DS, et al. Predictors of progression in Barrett's esophagus II: baseline 17p (p53) loss of heterozygosity identifies a patient subset at increased risk for neoplastic progression. Am J Gastroenterol. 2001;96(10):2839–2848.

[171] Laird PW. The power and the promise of DNA methylation markers. Nat Rev Cancer. 2003;3(4):253–266.

[172] Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, et al. Genome-wide DNA methylation profiling using Infinium® assay. Epigenomics. 2009;1(1):177–200.

[173] Bock C, Tomazou EM, Brinkman AB, Müller F, Simmer F, Gu H, et al. Quantitative comparison of genome-wide DNA methylation mapping technologies. Nat Biotechnol. 2010;28(10):1106–1114.

[174] Laird PW. Principles and challenges of genome-wide DNA methylation analysis. Nat Rev Genet. 2010;11(3):191–203.

[175] Bird A. DNA methylation patterns and epigenetic memory. Gene Dev. 2002;16(1):6–21.

[176] Lorincz MC, Schübeler D, Hutchinson SR, Dickerson DR, Groudine M. DNA methylation density influences the stability of an epigenetic imprint and Dnmt3a/b-independent de novo methylation. Mol Cell Biol. 2002;22(21):7572–7580.

[177] Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003;33:245–254.

[178] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456(7218):53–59.

[179] Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, et al. High density DNA methylation array with single CpG site resolution. Genomics. 2011;98(4):288–295.

[180] Chak A, Ochs-Balcom H, Falk G, Grady WM, Kinnard M, Willis JE, et al. Familiality in Barrett's esophagus, adenocarcinoma of the esophagus, and adenocarcinoma of the gastroesophageal junction. Cancer Epidemiol Biomarkers Prev. 2006;15(9):1668–1673.

[181] Alisch R, Barwick B, Chopra P, Myrick L, Satten G, Conneely K, et al. Age-associated DNA methylation in pediatric populations. Genome Res. 2012;22:623–632.

[182] Sontag L, Lorincz M, Luebeck E. Dynamics, stability and inheritance of somatic DNA methylation imprints. J Theor Biol. 2006;242(4):890–899.

[183] Siegmund K, Marjorama P, Woob S Y Tavare, Shibata D. Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers. Proc Natl Acad Sci U S A. 2009;106(12):4828–4833.

[184] Miranda TB, Jones PA. DNA methylation: the nuts and bolts of repression. J Cell Physiol. 2007;213(2):384–390.

[185] Przybilla J, Rohlf T, Loeffler M, Galle J. Understanding epigenetic changes in aging stem cells–a computational model approach. Aging Cell. 2014;13(2):320–328.

[186] Toyota M, Issa JPJ. CpG island methylator phenotypes in aging and cancer. In: Seminars in cancer biology. vol. 9. Elsevier; 1999. p. 349–357.

[187] Issa JPJ, Ahuja N, Toyota M, Bronner MP, Brentnall TA. Accelerated age-related CpG island methylation in ulcerative colitis. Cancer Res. 2001;61(9):3573–3577.

[188] Shibata D. Mutation and epigenetic molecular clocks in cancer. Carcinogenesis. 2011;32(2):123–128.

[189] Teschendorff AE, West J, Beck S. Age-associated epigenetic drift: implications, and a case of epigenetic thrift? Hum Mol Genet. 2013;p. ddt375.

[190] Issa JP. Aging and epigenetic drift: a vicious cycle. J Clin Invest. 2014;124(1):24–29.

[191] Kass RE, Raftery AE. Bayes factors. J Am Stat Assoc. 1995;90(430):773–795.

[192] Wu SC, Zhang Y. Active DNA demethylation: many roads lead to Rome. Nat Rev Mol Cell Biol. 2010;11(9):607–620.

[193] Guo JU, Su Y, Zhong C, Ming Gl, Song H. Hydroxylation of 5-methylcytosine by TET1 promotes active DNA demethylation in the adult brain. Cell. 2011;145(3):423–434.

[194] Vilkaitis G, Suetake I, Klimašauskas S, Tajima S. Processive methylation of hemimethylated CpG sites by mouse Dnmt1 DNA methyltransferase. J Biol Chem. 2005;280(1):64–72.

[195] Stairs DB, Nakagawa H, Klein-Szanto A, Mitchell SD, Silberg DG, Tobias JW, et al. Cdx1 and c-Myc foster the initiation of transdifferentiation of the normal esophageal squamous epithelium toward Barrett's esophagus. PLoS One. 2008;3(10):e3534.

[196] Badreddine RJ, Wang KK. Barrett esophagus: an update. Nature Rev Gastroentero. 2010;7(7):369–378.

[197] Price ME, Cotton AM, Lam LL, Farré P, Emberly E, Brown CJ, et al. Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin. 2013;6(4).

[198] Hvid-Jensen F, Pedersen L, Drewes AM, Sørensen HT, Funch-Jensen P. Incidence of adenocarcinoma among patients with Barrett's esophagus. New Engl J Med. 2011;365(15):1375–1383.

[199] Lada MJ, Nieman DR, Han M, Timratana P, Alsalahi O, Peyre CG, et al. Gastroesophageal reflux disease, proton-pump inhibitor use and Barrett's esophagus in esophageal adenocarcinoma: Trends revisited. Surgery. 2013;154(4):856–866.

[200] Wilson M, West C, Wilson G, Roberts S, James R, Schofield P. Intra-tumoral heterogeneity of tumour potential doubling times (Tpot) in colorectal cancer. Brit J Cancer. 1993;68(3):501–506.

[201] Gillespie D. Approximate accelerated stochastic simulation of chemically reacting systems. J Chem Phys. 2001;115(4):1716–1733.

[202] Cao Y, Gillespie D, Petzold L. Avoiding negative populations in explicit Poisson tau-leaping. J Chem Phys. 2005;123(5).

[203] Gillespie D, Petzold L. Improved leap-size selection for accelerated stochastic simulation. J Chem Phys. 2003;119(16):8229–8234.

[204] Kurtz T. In: Approximation of Population Processes. Philadelpha, PA: SIAM; 1981. .

[205] Sehl M, Alekseyenko A, Lange K. Accurate Stochastic Simulation via Step Anticipation τ-Leaping (SAL) Algorithm. J Comput Bio. 2009;16(9):1195–1208.

[206] Sehl M, Zhou H, Sinsheimer J, Lange K. Extinction models for cancer stem cell therapy. Math Biosci. 2011;234:32–146.

[207] Cao Y, Gillespie D, Petzold L. Efficient step size selection for the tau-leaping simulation method. J Chem Phys. 2006;124(044109).

[208] Bedi AO, Kwon RS, Rubenstein JH, Piraka CR, Elta GH, Scheiman JM, et al. A survey of expert follow-up practices after successful endoscopic eradication therapy for Barrett's esophagus with high-grade dysplasia and intramucosal adenocarcinoma. Gastrointest Endosc. 2013;78(5):696–701.