Impact of Tumor Progression on Cancer Incidence Curves

E. Georg Luebeck1, Kit Curtius3, Jihyoun Jeon2, and William D. Hazelton1

Abstract
Cancer arises through a multistage process, but it is not fully clear how this process influences the age-specific incidence curve. Studies of colorectal and pancreatic cancer using the multistage clonal expansion (MSCE) model have identified two phases of the incidence curves. One phase is linear, beginning about age of 60 years, suggesting that at least two rare rate-limiting mutations occur before clonal expansion of premalignant cells. A second phase is exponential, seen in early-onset cancers occurring before the age of 60 years that are associated with premalignant clonal expansion. Here, we extend the MSCE model to include clonal expansion of malignant cells, an advance that permits study of the effects of tumor growth and extinction on the incidence of colorectal, gastric, pancreatic, and esophageal adenocarcinomas in the digestive tract. After adjusting the age-specific incidence for birth-cohort and calendar-year trends, we found that initiating mutations and premalignant cell kinetics can explain the primary features of the incidence curve. However, we also found that the incidence data of these cancers harbored information on the kinetics of malignant clonal expansion before clinical detection, including tumor growth rates and extinction probabilities on three characteristic time scales for tumor progression. In addition, the data harbored information on the mean sojourn times for premalignant clones until occurrence of either the first malignant cell or the first persistent (surviving) malignant clone. Finally, the data also harbored information on the mean sojourn time of persistent malignant clones to the time of diagnosis. In conclusion, cancer incidence curves can harbor significant information about hidden processes of tumor initiation, premalignant clonal expansion, and malignant transformation, and even some limited information on tumor growth before clinical detection. Cancer Res; 73(3): 1086–96. ©2012 AACR.

Major Findings
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. Our analyses of the incidences of four digestive tract cancers show that the age-specific incidence curves—upon adjustments for secular trends and, in the case of esophageal adenocarcinoma, inclusion of an event describing the conversion of normal squamous to metaplastic Barrett’s epithelium—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. While this sojourn seems very short for pancreatic cancer (<3 years), intermediate for colorectal cancer (5–7 years), it is much longer for gastric cancer and esophageal adenocarcinoma (10–12 years). Furthermore, with the exception of pancreatic cancer, our results are consistent with the assumption of a high (>95%) probability of tumor stem cell extinction or terminal differentiation.

Introduction
Uncontrolled cell proliferation is the sine qua non of carcinogenesis. However, long before symptoms signal cancer growth, several initiating mutations are generally required to overcome normal homeostatic regulation in a tissue allowing the gradual expansion of premalignant clones. Albeit slow and possibly stagnant, this growth enhances the probability that a premalignant cell undergoes malignant transformation generating a clone that either becomes extinct or progresses until clinical detection. Therefore, at least 2 distinct but overlapping clonal expansion processes are likely to occur in a tissue before clinical detection of cancer. In the context of the multistage clonal expansion (MSCE) carcinogenesis model described here, the first clonal expansion begins after normal tissue stem cells acquire 2 rate-limiting mutations or epigenomic changes that lead to abrogation of homeostatic tissue control, causing gradual outgrowth of occult premalignant clones over an extended time period that may range from years to decades (1). Clonal expansion of the premalignant cell population enhances the probability that one or more of these cells suffer...
Impact of Tumor Progression on Cancer Incidence Curves

Quick Guide to Equations

The multistage clonal expansion model (MSCE) approximation yields the following hazard function, which represents the age-specific rate at which cancers occur in a population that had no prior occurrences of that cancer:

\[ h_{\text{MSCE}}(t) \approx \mu_0 X \left( 1 - \frac{q_0 - p_0}{q_0 e^{-\alpha_0 t} - p_0 e^{-\beta_0 t}} \right) ^{\mu_1 / q_1} \]

with

\[ \mu_2^{\text{eff}} = \mu_2 (1 - \beta_2 / \alpha_2) \quad \text{and} \quad t_{\text{lag}} \approx T_2 \approx \frac{-\ln(\alpha_M \beta_M / (\alpha_M - \beta_M))}{\alpha_M - \beta_M} \]

where \( \mu_2^{\text{eff}} \) represents the effective rate of malignant transformations that give rise to persistent tumors, and \( t_{\text{lag}} \) is a time-lag equal to the mean sojourn time of a preclinical cancer clone from its single cell inception to clinical detection. See Fig. 1 and text for a definition of the basic model parameters \( X \) (number of stem cells), \( \alpha_{M(0)} \) (cell division rates), \( \beta_{P(0)} \) (cell death rates), \( \mu_{0,1,2} \) (mutation rates). Furthermore,

\[ p_0 = \frac{1}{2} \left( \frac{\alpha_0 - \beta_0 - \mu_2^{\text{eff}}}{\alpha_0 - \beta_0 - \mu_2^{\text{eff}}} - \sqrt{\left(\frac{\alpha_0 - \beta_0 - \mu_2^{\text{eff}}}{}\right)^2 + 4\alpha_0 \mu_2^{\text{eff}}} \right) \]

\[ q_0 = \frac{1}{2} \left( \frac{\alpha_0 - \beta_0 - \mu_2^{\text{eff}}}{\alpha_0 - \beta_0 - \mu_2^{\text{eff}}} + \sqrt{\left(\frac{\alpha_0 - \beta_0 - \mu_2^{\text{eff}}}{}\right)^2 + 4\alpha_0 \mu_2^{\text{eff}}} \right) \]

where \( -p_0 \) measures approximately the net cell proliferation in premalignant clones and \( q_0 \) approximately the (effective) rate of malignant transformations.

Additional mutations or epigenomic alterations that cause malignant transformation, which enables tumors to accelerate their growth and invade neighboring tissue, a process captured by the second (malignant) clonal expansion in the model.

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? In previous studies (1–4), we identified 2 characteristic features, or phases, in the incidence curves for colorectal cancer (CRC) and pancreatic cancers using data from the Surveillance Epidemiology and End Results (SEER) registries (5). After adjusting for secular trends related to birth-cohort and calendar-year (period), we were able to identify an exponential phase in the incidence curve beginning in early adult life and extending to approximately the age of 60 years and a linearly increasing trend for late-onset cancers extending beyond the age of 60 years. 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 discernable? To address this question, we use a MSCE model that 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. Malignant transformations, however, are likely to occur in altered cells whose initial survival fitness may be compromised by genomic instability (6) and therefore may be prone to extinction despite higher cell proliferation. This is supported by comparative measurements of cell division rates and cell net cell proliferation (using DNA labeling and radiographic imaging of tumors, respectively) in a variety of carcinomas, showing large differences in the 2 rates, which can only be explained by the frequent death of tumor cells (7).

For this model-driven investigation of cancer incidence, we analyze SEER-9 (5) incidence data (1975–2008) for 4 gastrointestinal malignancies: CRC, gastric cancer, pancreatic cancer, and esophageal adenocarcinomas (EAC). 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 3 characteristic times: (i) the mean sojourn times for premalignant clones until occurrence of the first malignant cell regardless of its fate, (ii) the analogous mean sojourn time to appearance of the first surviving (persistent) malignant clone, and (iii) the mean sojourn time of persistent preclinical cancers from first malignant cell to time of cancer diagnosis.
Combined with a mathematical exploration of the MSCE 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 (noninvasive) 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 (8), whereas the transition from a noninvasive to an invasive tumor, which expands with a significantly higher growth rate, constitutes a critical, rate-limiting event.

Materials and Methods

Model assumptions and properties

Tumor initiation. A hallmark of the MSCE model 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. 1). For colon and pancreatic cancer, we inferred previously that it takes 2 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 (1). The 2 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 (9). Inactivation of TP53 is seen during early development of many digestive tract cancers, including gastric cancer (10), pancreatic cancer (11, 12), and EAC (13–15). Inactivation of p16 often occurs early in the development of EAC (16) and other cancers. However, the 2 hits may also represent activation of an oncogene, such as KRAS in combination with gain-of-function mutation in a tumor suppressor gene (12, 17). In addition, EAC is associated with earlier conversion of a section of normal esophageal squamous epithelium to intestinal type Barrett’s metaplasia, called Barrett’s esophagus (BE). We model the transition to Barrett’s esophagus as an additional one-time tissue alteration, which occurs before the 2 initiating events leading to premalignant clonal expansions in the development of EAC (3).

A mathematical consequence of the 2-hit hypothesis for (premalignant) tumor initiation is that the hazard function of the model (which represents the age-specific incidence) has a linearly increasing trend for older ages (1). The presence of such a linear phase in the incidence curves for CRC and pancreatic cancer could indeed be shown by likelihood-based comparisons of models with 2 (or more) hits for initiation. Models with single-hit tumor initiations do not give rise to a linear phase in the hazard function (1, 2).

**Tumor promotion.** Before the transition into the initiation-associated linear phase, the MSCE hazard function increases exponentially with a rate that is approximately given by the net cell proliferation rate of premalignant cells (1). The transition from the exponential phase to the linear phase occurs around the age of 60 years for CRC and pancreatic cancer. Clonal expansion of P cells is represented by a stochastic birth-death-mutation (bdm) process with cell division rate \( \alpha_p \), death-or-differentiation rate \( \beta_p \), and mutation rate \( \mu_p \). The net cell proliferation rate of P cells is given by \( \alpha_p - \beta_p - \mu_p \) and the asymptotic probability of extinction of P cells by the ratio \( \beta_p/\alpha_p \) (18), 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 \( \mu_p \), 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 (6), only 2 initial rate-limiting mutations before clonal expansion seem necessary to adequately describe the main shape of the incidence curve for the 4 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, which 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, which 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 (bd) process with cell division rate \( \alpha_m \) and cell death rate \( \beta_m \). Clinical detection of the tumor is similarly treated as a stochastic event with rate \( \rho \) per cell. This implies that a tumor of size \( n \) cells has probability \( n \rho \Delta t \) to be detected in a time interval \( \Delta t \) short enough for the tumor to be constant in size. We refer to this generalization of the birth–death process as a birth–death-observation (bdo) process. Note, all analyses reported here are with a fixed value of \( \rho = 10^{-7} \). The rationale for this particular value of \( \rho \) 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 (19). Results obtained with other values for \( \rho \) (in the range of \( 10^{-6} - 10^{-8} \)) were similar and did not change our conclusions (see Supplementary Information).

The essential stochastic components of the MSCE model are illustrated in Fig. 1, including separate clonal-expansions for premalignant and malignant cells. 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 2 important aspects: (i) the rate at which P cells suffer a transformation event that gives rise to a detectable cancer is approximated by an "effective" transformation rate, \( \mu_p^{eff} \), and (ii) 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 to achieve parameter identifiability, as discussed in the Results and in the Supplementary Information.

While multistage generalizations of the models shown in Fig. 1 have also been explored by others (20–24), 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 (1–4), especially in regard to the time scales of premalignant and malignant clonal expansion. However, Fakir and colleagues (20, 21) modeled stochastic effects of lung cancer progression by augmenting a similar model with a more realistic progression model, including extinction and/or dormancy, proliferation, and invasive growth.

**MSCE model hazard function.** For the full version depicted in Fig. 1, which includes 2 stochastic clonal expansions, it is straightforward to derive the probability, \( P_{MSCE}(t) \), for a cancer diagnosis/detection to occur by time \( t \). A general approach is to solve the Kolmogorov backward equations for the marginal probability generating functions properly conditioned on no cancer occurring before time \( t \), as shown in the Supplementary Information. This yields the MSCE-survival function \( S_{MSCE}(t) = 1 - P_{MSCE}(t) \). The MSCE hazard function—the rate at which preclinical cancer cells collectively trigger a first clinical observation event—is given by:

\[
\lambda_{MSCE}(t) = -\frac{d}{dt} \log S_{MSCE}(t) = \frac{-S_{MSCE}(t)}{S_{MSCE}(t)}, \quad (1)
\]

(where the dot represents a derivative with respect to \( t \)). \( \lambda_{MSCE}(t) \) can be computed by numerically solving a set of ordinary differential equations as described in the Supplementary Information.
Data
We used data from the SEER database by the National Cancer Institute. Incidence data were for all races by single-years of ages between 10 and 84 years and calendar-year between 1975 and 2008 for males and females (all races) in the original 9 registries (SEER-9; ref. 5). Incident cancers were defined using the International Classification of Diseases for Oncology, Third Edition (ICD-O-3) as follows: CRC (C18-C20); EAC (C150-C159, 8140/3 adenocarcinoma NOS); 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. Methods for adjusting for secular trends (period and cohort effects) are described in the Supplementary Information.

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 3 underlying time scales in carcinogenesis: the mean sojourn time for premalignant clones until occurrence of the first malignant cell (T1), the mean sojourn time for premalignant clones until the first surviving malignant clone (T2), and the mean sojourn time of persistent preclinical cancers from first malignant cell to time of diagnosis (T3). In the following, we show how these time scales contribute to, and are estimable from, the age-specific incidence curves of 4 digestive tract cancers.

Mathematical properties of MSCE-1 approximation
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. 1. This will show 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, \( \mu_2 \), for nonextinction and further models the outgrowth of persistent malignant clones as a constant time-lag, \( t_{\text{lag}} \), 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 \( t_{\text{lag}} \) and the mean sojourn time \( T_2 \) in the MSCE model, we show (proof given in Supplementary Information) that the hazard function of a model with 2 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, that is, replacing \( \mu_2 \rightarrow \mu_2 (1 - S_M(u)) \), where \( 1 - S_M(u) \) is the unnormalized probability of detection of malignant clones a time \( u \) after the malignant clone is seeded. An exact expression for \( S_M(u) \) is given in the Supplementary Information for constant parameters. However, this mathematical “simplification” of the model from a double to a single clonal expansion process comes at the cost of a time-dependent (conditional) mutation rate.

The time-dependence of the conditional mutation rate \( \mu_2 (1 - S_M(u)) \) has 2 main effects: (i) it reduces the effective rate of malignant transformation, and (ii) it creates a time delay for a malignant clone to grow, conditional on its nonextinction, into a detectable tumor. The latter effect is mainly due to a sharp transition of the conditional mutation rate from 0 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 Supplementary Information). Because asymptotically \( S_M(u) \rightarrow \mu_M/a_M \) as \( u \rightarrow \infty \), we define \( \mu_2^{\text{eff}} \equiv \mu_2 (1 - \mu_M/a_M) \) as the effective malignant transformation rate for the reduced (MSCE-1) model. Therefore, the approximation amounts to

\[
\mu_2 \rightarrow \mu_2^{\text{eff}} \\
t_{\text{lag}} \rightarrow T_2.
\]

(MSCE-1) (MSCE) (2)

Again, \( T_2 \) is the mean sojourn time of a surviving malignant clone that avoids stochastic extinction and which, in the absence of death, would eventually be detected as cancer. The mean sojourn time of a premalignant clone to the first malignant cell (\( T_1 \)), the analogous mean sojourn time to the ancestor of the first persistent malignant clone (\( T_2^{\text{eff}} \)), and the mean sojourn time of a persistent malignant tumor (\( T_2 \)) to detection are functions of the cell kinetic parameters and are given by (see Supplementary Information)

\[
T_1 = \int_0^\infty \left( 1 - \frac{1 - S_P(u)}{1 - \frac{\mu_2^{\text{eff}}}{\alpha_\mu}} \right) du \approx -\frac{\ln(\alpha_\mu \mu_2^{\text{eff}}/(\alpha_\mu - \beta_\mu)^2)}{\alpha_\mu - \beta_\mu},
\]

\[
T_2^{\text{eff}} = \int_0^\infty \left( 1 - \frac{1 - S_P^{\text{eff}}(u)}{1 - \frac{\mu_2^{\text{eff}}}{\alpha_\mu}} \right) du \approx -\frac{\ln(\alpha_\mu \mu_2^{\text{eff}}/(\alpha_\mu - \beta_\mu)^2)}{\alpha_\mu - \beta_\mu},
\]

\[
T_2 = \int_0^\infty \left( 1 - \frac{1 - S_M(u)}{1 - \frac{\mu_2}{\alpha_M}} \right) du \approx -\frac{\ln(\alpha_M \mu_2/(\alpha_M - \beta_M)^2)}{\alpha_M - \beta_M},
\]

respectively, where \( S_P \) and \( S_P^{\text{eff}} \) are survival functions defined analogously to \( S_M \) (see Supplementary Information).

Parameter identifiability and sensitivity
Not all of the MSCE model parameters are identifiable from incidence data—some parameters must be fixed initially to achieve parameter identifiability (see Heidenreich and colleagues: 25). Furthermore, for estimability, the exponential-thin-linear characteristic of the multistage hazard function (see Supplementary Information) suggest a parametrization that involves the slope of the linear phase \( \lambda \equiv \mu_M a_M p_m \) and the growth parameter of the exponential phase \( g_\alpha \equiv \alpha_\alpha - \beta_\alpha - \mu_2 \) (1). Note, the rates \( \mu_\alpha \) and \( \mu_1 \) cannot be estimated separately because the slope \( \lambda \) depends on their product. Analogous to premalignant growth, we introduce the malignant growth parameter \( g_\alpha \equiv \alpha_M - \beta_M - \rho \). To identify \( \mu_2 \) and \( \rho \), we find it necessary to fix the cell division rates \( \alpha_\gamma \) and \( \alpha_\mu \). Although the product \( \alpha_M p_m \) is mathematically identifiable, we were not able to obtain stable estimates and therefore also fixed the (per cell) cancer detection parameter \( \rho \) (see Materials and Methods). Otherwise, the biological model parameters were estimated using a Markov-Chain Monte Carlo (MCMC) method (see...
Supplementary Information). Supplementary Fig. S1a–h shows scatterplots of the MCMC samples obtained for all four cancers studied here, separately by gender. For EAC, an additional parameter is included representing the rate of normal squamous tissue conversion to Barrett’s metaplasia.

To explore the dependence of our parameter estimates on the fixed parameters $\alpha_0, \alpha_3$, and the detection rate $\rho$, 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 $\lambda, g_0, g_3$, and $\mu_{\text{eff}}^3$, assuming constant birth-cohort and calendar-year effects) for each fixed parameter are given in Supplementary Table S1. This analysis (although limited to CRC) shows that the estimates of $g_3$ and therefore the mean sojourn time $T_2$, vary only slightly when $\alpha_0$ and $\alpha_3$ are perturbed but may vary up to 20% as the detection rate $\rho$ changes an order of magnitude. Therefore, the dependence of the preclinical cancer sojourn time on $\rho$ is modest but does not change our conclusions.

**Low fitness of malignant cells**

We use the above results to gain insight into the importance of clonal extinction. Figure 2 shows fits obtained with the MSCE model (solid line) to SEER incidence data for (i) CRC, and (ii) gastric cancers. These fits include adjustments of the model-generated hazard function for secular trends (for details see Supplementary Information). 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 $\mu_2$ into the MSCE-1 approximation and ignoring the lag-time, that is, $t_{\text{lag}} = 0$ (dotted line in Fig. 2). 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, reintroducing the effects of extinction by replacing $\mu_2$ with $\mu_2^{\text{eff}}$ (without a lag-time) restores the general shape of the incidence curve (dot-dash line) except for cancers occurring too early. Finally, reintroducing 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 to 40 years for gastric cancer (see Table 1), as long as 20 years for CRC and EAC, 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 despite their aggressive and invasive behavior.
In contrast, the estimated mean sojourn times $T_2$ of persistent malignant clones vary from 10 to 12 years for gastric cancer and EAC, 5 to 7 years for CRC, and down to less than 1 year for pancreatic cancer (Table 1). The latter is consistent with the observation that most pancreatic carcinomas are diagnosed at an advanced metastatic stage. Note, however, that $T_2^{eff}$, 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 for males, 56.3 vs. 48.7 in females). This suggests that premalignant precursor lesions in pancreas, such as pancreatic intraepithelial neoplasia (PanIN), may be present for many years before a stable malignant transformation occurs.

For EAC, we also estimate a (constant) tissue conversion rate $v_{BE}$ from normal esophageal tissue to the metastatic tissue of Barrett’s esophagus. The age-specific prevalence of Barrett’s esophagus is therefore approximately $v_{BE} \times \text{age}$ and seems to be subject to strong period effects (3). Although our MCMC-based estimates for $v_{BE}$ and the slope parameter $\lambda$ representing initiation of premalignant clones are highly anticorrelated (see Supplementary Fig. S1g–h), the predicted Barrett’s esophagus prevalence (about 1.5% for males and 0.5% for females at age of 60 years in the year 2000) are consistent with the range of epidemiologic estimates obtained from studies in comparable populations (26).

### Tumor growth rates
We find highly stable estimates for the net cell proliferation rate $g_M$ of premalignant cells, based on the posterior distributions of the identifiable MSCE model parameters given the observed cancer incidences in SEER (see Supplementary Fig. S1). The reason for this stability seems to lie in the prominence of the exponential phase of the incidence curve and the resulting linear behavior of the log-likelihood (see Supplementary Fig. S3). 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, whereas 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 and esophageal cancers to rates as high as 30 per year for pancreatic cancer (see Table 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 (27), the latter appears too fast, but not inconsistent with tumor marker doubling times. For example, using the pancreatic tumor marker CA19-9, Nishida and colleagues (28) 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 and colleagues (29), who followed 27 carcinomas radiographically in the colon and rectum, measuring a median of 130 days with a range of 53 to 1,570 days. Despite 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 our claim that carefully collected incidence data harbor quantitative information about the natural history of a tumor, from initiation to promotion to malignant tumor progression.

### Discussion
Early models of carcinogenesis recognized the importance of rate-limiting mutations but provided only crude fits to cancer incidence and mortality (30). Subsequent incorporation of cell proliferation made it possible to account for effects, such as the initiation/promotion effects seen in chemical carcinogenesis (31, 32) or the inverse dose-rate effect for high linear energy transfer radiation (33), that were more difficult to explain with models that did not include clonal expansion. More recently, multistage extensions of the original 2-stage clonal expansion model by Moolgavkar and Knudson (34) and

| Table 1. MCMC-based estimates of various tumor promotion and progression time scales |
|---------------------------------|-----------------|-----------------|-----------------|-----------------|
|                                | $T_1$ (95% CI)  | $T_1^{eff}$ (95% CI) | $T_2$ (95% CI)  | $T_{lag}$ (95% CI) |
| **Males**                      |                |                  |                |                 |
| 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 (8.8–14.7)| 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)   |
| EAC                            | 15.0 (7.2–24.7)  | 39.2 (34.2–44.0) | 12.0 (6.3–16.8)| 12.7 (3.1–16.5) |
| **Females**                    |                |                  |                |                 |
| 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)   |
| EAC                            | 15.8 (13.8–18.4) | 37.9 (31.8–45.1) | 10.6 (7.8–13.8)| 10.2 (7.2–13.2) |

NOTE: Estimates represent medians and 95% credibility regions of the marginal posterior distribution for each quantity listed. All units are in years.

Abbreviations: GaC, gastric cancer; PaC, pancreatic cancer.
Moolgavkar and Venzon (35) 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 underlying biologic processes (3, 4, 36). 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 lifestyle factors on cancer incidence, in this study, we focus on nonspecific 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 4 gastrointestinal cancers (CRC, gastric cancer, pancreatic cancer, and EAC) 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 1 and 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 \( T_{1,eff} \), that is, the time of appearance of the first persistent malignant clone that started with a single premalignant cell, seems to be much longer than estimated from clinical data. For colon, clinical estimates range from 20 to 25 years (37). 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 patients with FAP even though the diagnosis of polyposis may not occur until later. From the age distribution of cancer with polyposis in patients with FAP (see ref. 37), which peaks around the age of 40 years, we conclude that the mean sojourn of an adenoma that has the potential to progress to cancer is likely longer than 40 years, as this time generally represents the time of 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 to 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), which is adjusted for clonal extinction and delayed by a lag-time representing the mean sojourn time \( T_2 \) of the surviving malignancy (see Fig. 1). 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) detection \( \rho \). However, assuming plausible values for the cell division rates \( \alpha_M \) and a (per cell) cancer detection rate \( \beta_M \) (see sensitivity analysis), we do obtain estimates for the net cell proliferation rate \( \epsilon_M \) 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 noninvasive precursors, such as the PanINs, interact with stromal components such as myofibroblasts that facilitate invasion and metastatic colonization (38). 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 clonal expansions. However, as shown by Meza and colleagues (1) 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 seems 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 and colleagues (39) 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 (39).

More recently, Yachida and colleagues (40) undertook a similar study for pancreatic cancer sequencing the genomes of 7 metastatic lesions to evaluate the clonal relationships among primary and metastatic cancers (see also ref. 41). They estimated 6.8 years for the length of time from the appearance of subclones in the primary tumor with metastatic potential to the seeding of the index metastasis and additional 2.7 years to detection. However, our \( T_2 \) estimates for pancreatic cancer are inconsistent with those derived by Yachida and colleagues (see the MCMC-based posterior distributions for \( T_2 \) in Supplementary Fig. S1). Remarkably, we find shorter times that suggest (see Discussion earlier), that the subclones found by Yachida and colleagues in the primary may already have been present in a slow growing precursor lesion. The question is therefore whether metastatic dissemination in pancreas can occur before the primary tumor undergoes a drastic transformation into a rapidly growing tumor.

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 (42, 43)], the 2-stage clonal expansion (TSCE) model (44, 45), and with biologically motivated extensions of the TSCE model (1–4). Although the latter usually provide superior fits to cancer incidence data as compared with the former (1, 2, 4), 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 biologic 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 biologic parameters, whereas in this study, we use a statistical approach [the age-period-cohort model (3, 4, 36)] to effectively adjust cancer incidence for secular trends. 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 (46). 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 (47). 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 and colleagues; ref. 48) 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 2 (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 despite 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; whereas screening for EAC may include assessment of the presence or absence of dysplasia and/or chromosomal abnormalities in endoscopic biopsies and surveillance for early cancer. Mechanistic models such as the MSCE model may use these different outcomes to enhance our understanding of tumor initiation, growth, persistence, and preclinical sojourn.

In this study, we show 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 3 characteristic times scales of carcinogenesis: the mean sojourn time from premalignant cell to first
malignant cell, \(T_T\); the mean sojourn time from premalignant cell to first malignant ancestor that generates a persistent clone, \(T_1\); 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, pancreatic, and esophageal cancers.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Authors' Contributions

Conception and design: E.G. Luebeck, W.D. Hazelton

Development of methodology: E.G. Luebeck, K. Curtius, W.D. Hazelton

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): W.D. Hazelton

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): E.G. Luebeck, K. Curtius, J. Jeon, W.D. Hazelton

Writing, review, and/or revision of the manuscript: E.G. Luebeck, K. Curtius, J. Jeon, W.D. Hazelton

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): K. Curtius

Study supervision: E.G. Luebeck

Acknowledgments

The authors thank Dr. Rafael Meza (University of Michigan) for helpful discussions.

Grant Support

This research was supported by the NCI under grants R01 CA107028 and U01 CA217155 (to E.G. Luebeck, J. Jeon, and W.D. Hazelton) and by the National Science Foundation (NSF) under grant no. DGE-0718124 (to K. Curtius).

The costs of publication of this article were deferred in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received June 5, 2012; revised August 31, 2012; accepted September 18, 2012; published OnlineFirst October 10, 2012.

References


