A Simple 3-Parameter Model for Cancer Incidences

Xiaoxiao Zhang1,2,3, Holger Fröhlich1,4, Dima Grigoriev5, Sergey Vakulenko6,7, Jörg Zimmermann1,4 & Andreas Günter Weber8

We propose a simple 3-parameter model that provides very good fits for incidence curves of 18 common solid cancers even when variations due to different locations, races, or periods are taken into account. From a data perspective, we use model selection (Akaike information criterion) to show that this model, which is based on the Weibull distribution, outperforms other simple models like the Gamma distribution. From a modeling perspective, the Weibull distribution can be justified as modeling the accumulation of driver events, which establishes a link to stem cell division based cancer development models and a connection to a recursion formula for intrinsic cancer risk published by Wu et al. For the recursion formula a closed form solution is given, which will help to simplify future analyses. Additionally, we perform a sensitivity analysis for the parameters, showing that two of the three parameters can vary over several orders of magnitude. However, the shape parameter of the Weibull distribution can be robustly estimated from epidemiological data.

Cancers arise after accumulating epigenetic and genetic aberrations1–3. Earlier studies established a power law model on the basis of multi-stage somatic mutation theory to explain age-dependent incidences4–6 for several cancer types. As noted by Hornsby et al.7 in the context of classical epidemiological studies most cancers occur with the same characteristic pattern of incidence, and the simplicity of this pattern is in contrast to the perceived complexity of carcinogenesis. Orthogonal to these age stratification of different cancer types, Tomasetti and Vogelstein8 (with follow-ups4,5,6) reported a significant association between life time cancer risk and stem cell divisions and concluded the latter substantially contributes to the former. Challenging the conclusion of Tomasetti and Vogelstein8 of a high-intrinsic cancer risk Wu et al.9 subdivided cancer risk into extrinsic and intrinsic risk, arguing extrinsic factors contribute more to cancer risks than intrinsic factors do. Based on a mechanistic model of accumulated mutations, these authors provided a recursion formula for theoretical life time intrinsic risk (tLIR) parameterized by age a. This recursion formula has the closed form solution tLIR(a) = 1 − (1 − (1 − (1 − r)^\frac{\log_S}{d_k} - d_k)^S), where S can be interpreted as the numbers of stem cell, d as the stem cell division rate, k as number of driver events required for cancer onset and r as the mutation rate per division. They reported that tLIR goes outside of the plausible range of empirical cancer risks by studying several pairs of values for two parameters (mutation rate and driver gene mutations) concluding that there is a substantial contribution of extrinsic risk factors to cancer development. However, this conclusion only holds in the studied parameter space and when parameters for all cancer types are treated uniformly. By performing a systematic grid search in the space of biologically plausible parameter values we showed that tLIR can be close to empirical risk for different cancer types (R^2 > 0.85). If the extrinsic risk factor is computed by simply setting it to a complement of 1 for the intrinsic risk factor as performed by Wu et al. it will be concluded that there is a possibility of high intrinsic risk, so that one of the presented arguments by Wu et al.9 is fallacious.

On a pure mathematical side, we show that a scaled Weibull function with 3 parameters approximates the 4-parameter mechanistic tLIR model. On an epidemiological data analytical side, this simple 3-parameter model excellently agrees with age-dependent cancer incidence curves among 18 common solid cancers even when variations due to different locations, races, or periods are taken into account. With this model, we study the...
relationship between cancer risk and stem cell divisions, the high correlation between the two entities reported by previous studies breaks down when considering age stratified data.

Results

Approximation of tLIR model by a scaled Weibull function. As is derived in the Materials and Methods the 4-parameter mechanistic tLIR model can be approximated by a scaled Weibull function with 3 parameters:

\[ \lambda \approx aP(ka) Weibull(,)(1) \]

assuming that \( \lambda \) is defined by

\[ \lambda = Sr(dP)kk(2) \]

Here Weibull(, ) is the cumulative distribution function of the Weibull distribution, and \( P \) is the number of independent parallel processes, which e.g. can be interpreted as cell population at risk. Whether the total tissue cells or only a fraction of stem cells are susceptible for cancer risk is unclear. If one sets \( P = S \) then \( r(d) = \lambda P \). However, other possible choices for \( P \) allow to account for other factors such as the selection of mutations, the stem cell microenvironment, and tissue architecture, or effects of clonal expansion. Models incorporating clonal expansion have additional parameters such as the number of clonal copies. Reducing the dimensions of such complexed models results in tLIR, in which \( S \) is interpreted as number of independent clusters after clonal expansion rather than the number of stem cells, \( r \) and \( d \) denote “net” mutation and division rate of independent clusters at average level rather than those of single cells. Whereas a precise analysis of models for clonal expansion will be the topic of future work, these considerations show that when using the scaled Weibull distribution, prior knowledge on the parameter ranges is not necessary. This is indeed one benefit of scaled Weibull function comparing to tLIR model which requires a biologically reasonable guessing on stem cell numbers, mutation rate, cell division rate and number of driver mutations. The Weibull distribution is a special case of the generalized extreme value distribution (GEV). The GEV distribution plays the same role within extreme value statistics as the normal distribution does in average value statistics. It results in the limit distribution being maximized over many independent and identically distributed random variables, thus becoming the default model for the accumulation of micro events which finally leads to a macro event. The GEV is the limit distribution when one takes the maximum (and not the sum) of many independent and identically distributed random variables, thus being the default model for the accumulation of micro events which finally lead to a macro event. Accordingly, the Weibull distribution is not just a distribution providing a good empirical fit, but can be seen as justifiable for use in a plausible causative model of cancer genesis.

Fitting empirical incidence rates with scaled Weibull function. We performed extensive simulations and parameter fittings for the empirical incidence cuminc(,)(a) of cancer type \( c \) at age \( a \) using the scaled Weibull function: cuminc(,)(a) \( \approx P \cdot Weibull(,)(a) \). The model agrees excellently with age-dependent age incidences of 18 common solid cancers (\( R^2 > 0.99 \), Fig. 1).

Goodness of fit maintains when parameters \( P \) and \( \lambda \), varying roughly two orders of magnitudes (Fig. 2). This finding suggests that many parameter combinations provide similar dynamics that are consistent with empirical data. So any interpretations of \( P \) and \( \lambda \) have to take into account this considerable uncertainty. Nevertheless, the estimates for \( P \) are several orders of magnitude smaller than the realistic number of stem cells provided by Tomasetti and Vogelstein, yielding evidence supporting the above statement that the number of independent local processes is not equal to the number of stem cells.

Figure 1. Empirical cumulative cancer incidence data are consistent with the Weibull cumulative probability function in 18 cancers (data for ages up to 85 years old). Empirical (blue line) and Weibull function-fitted (red line) cancer cumulative incidence curves for 18 tissues, goodness of fit is reported in each subplot. The 18 cancers exhibit a good goodness of fit when using \( R^2 \) between model-reported age incidence and the empirical cumulative cancer incidence are used as metrics.
The estimates of parameter $k_c$, which corresponds to the number of driver events in the mechanistic model, are robust against variations of parameters $P_c$ and $\lambda_c$ (Fig. 2). Moreover, the estimates of $k_c$ are robust against race, sex, period and location (Fig. 3). In Supplemental Fig. 2 the best fits of shape parameters are plotted against the best fits of scale parameters for 694 time series.

Figure 2. Sensitivity analysis of parameter estimates using the scaled Weibull function for exemplary 14 cancer types. Whereas the estimates $P_c$ for the cell population at risk and the scale parameter $\lambda_c$ can vary over two order of magnitude, the estimates of the shape parameter $k_c$ are within about $\pm1$. Notice that the shape parameter allows interpretation as the number of limiting events.

Figure 3. Shape parameters estimated by fitting empirical cancer incidence data using the Weibull function (data with ages for up to 85 years). Cancer patients are grouped by year of diagnosis, race and registry. Cancers are ordered by median values of shape. Shapes are uniform regardless of risk factors, which is consistent with intuitive expectations: race and environmental changes are less likely to alter the number of driver events for cancer onsets.

The estimates of parameter $k_c$, which corresponds to the number of driver events in the mechanistic model, are robust against variations of parameters $P_c$ and $\lambda_c$ (Fig. 2). Moreover, the estimates of $k_c$ are robust against race, sex, period and location (Fig. 3). In Supplemental Fig. 2 the best fits of shape parameters are plotted against the best fits of scale parameters for 694 time series.

Relationship between cancer incidence and stem cell divisions. Tomasetti and Vogelstein\(^8\) suggested that the variation in cancer risk among tissues can be explained by the number of stem cell divisions. They reported that the tissue-specific cancer risk is strongly correlated (0.81) with life-time stem cell divisions (LSCD). These authors stated that the total number of stem cell divisions is a causative factor of cancer risk. This assumption yields a prediction on age structured data: for tissue type $c$ the number of stem cell divisions up to age $a$, which we will denote by $\text{LCSD}_c(a)$, should then be strongly correlated with $\text{cuminc}_c(a)$. However, using age incidence data obtained from the SEER-database\(^27\) we found that the regression lines for most tissue types $c$ for age data of 40, 50, 60, 70, and 80 years of $\text{cuminc}_c(a)$ plotted against $\text{LCSD}_c(a)$ in a log-log-scale are much steeper than the ones of the regression lines for different $c$ and $\text{cuminc}_c(80)$—using 80 as average life span as was done by Tomasetti and Vogelstein\(^8\) (see Fig. 4).
This "life time cancer risk" moderately associates with age-dependent stem cell divisions, if one takes a life-time that is less than 70 years (Fig. 5).

Overall, age-dependent stem cell divisions (using ages 40–80 years) is modestly correlated to age-dependent cancer risk for the 31 cancer types considered by Tomasetti and Vogelstein using the SEER database and the estimates of stem cell divisions given therein (Pearson correlation coefficient $\rho = 0.51$).

Hence, the strong correlation for (life-time) tissue-specific cancer risk with life-time stem cell divisions (LSCD) cannot be explained by the simple causative factor (involving the product of the number of stem cells and the number of divisions of each stem cell) suggested by Tomasetti and Vogelstein. A causal explanation on cancer risk should at least shows that the association between cancer risk and risk factor observed at overall level is reproducible on age stratified data. However, one caveat to such explanation, co-factors of risk factors might not be appreciated.

In our 3-parameter model, which gives good fits for age dependent cancer risks, several relations between the model parameters and cancer risks at a certain age can be observed. For instance in our parameter estimates good fits are possible when taking the inverse of the lifetime cancer risk $P_c \approx 1/\text{cuminc}$ (85). However, we will not suggest that the number $P_c$ of cell population at risk is an explanation for the variation of cancer risks among tissues: as the range of $P_c$ yielding good fits varies by two orders of magnitude and independently determining this number is difficult to achieve, a corresponding hypothesis is difficult to verify or to falsify.

As the sensitivity analysis for the scale parameter $\lambda_c$ (Fig. 2) shows that this parameter varies over several order of magnitudes, still yielding very good fits ($R^2 > 0.99$), the corresponding estimates for the mutation rate $r$ in the tLIR model using the approximation (1) and relation (2) are also very uncertain, even when fixing values of $S$ and $d$ and leaving out the considerable uncertainty of these. Nevertheless, when using estimates of $S$ and $d$ taken from the literature the obtained ranges of values of $r$ using relation (2) for several cancer types do not intersect the range $[10^{-10}, 10^{-6}]$ of "plausible values" of $r$ suggested by Wu et al. If we extend the analysis to allow "good fits" by setting a threshold $R^2 > 0.85$, then good fit of the tLIR model with $r \in [10^{-10}, 10^{-6}]$ are possible to achieve (Table 1).

### Testing performance of our 3-parameter against other simpler model.

For testing the performance of our 3-parameter model against other simpler models, we compared the fitting performance of the scaled Weibull function to that of 2-parameter power law model arising as the simplest instance from multistage theory. The empirical time series for different locations, periods and races were fitted (694 time series all together) using both models, the power law model had a goodness of fit of $R^2 < 0.90$ for 90 time series (13.0%), $R^2 < 0.95$ for 257 time series (37.0%), and $R^2 < 0.98$ for 366 time series (52.7%). In contrast, our 3-parameter scaled Weibull model resulted in $R^2 > 0.9$ for all time series, $R^2 > 0.98$ for 686 (=98.8%) of the time series, and $R^2 > 0.99$ for 679 (=97.8%) of them (Fig. 6(a)).

We compare the fitting performance of the scaled Weibull function against that of the scaled Gamma function. Although both functions fit data equivalently well in most cases, the scaled Weibull function outperforms the scaled Gamma function in several time series. (Fig. 6(b)) displays $R^2$ reporting goodness of fit for the two functions. We also calculate the Aikake information criterion (AIC), a likelihood based measurement. A lower AIC value indicates a better fit. Table 2 reports the AIC for 18 cancer types, the AIC values for Weibull function are lower than those of Gamma function in 15 cancer types.
Estimating the Number of Driver Mutations for Cancer Onset. In our model the shape parameter $k_c$ reflects the number of mutations required for cancer onset. The values of this parameter are, however, higher than the number of mutations estimated from sequencing data by Vogelstein et al.\textsuperscript{29}. Vogelstein et al. suggested technical issues as an explanation for the inconsistency between estimates from epidemiological data and sequencing data. Notably, our $k_c$ estimates and the number of driver mutations estimated from a classical power law model are roughly in the same numerical range (Fig. 7). Since we obtain better and more robust fits than the power law model, we believe that our estimated driver mutation numbers are more trustworthy.

Discussion

In this study we connected the mechanism-based cancer development LIR model to the Weibull distribution function. We tested its validity by fitting a 3-parameter Weibull function to data from 18 common solid cancer types, consisting of more than 600 time series. The scaled Weibull function fits well with age dependent incidence curves of all studied cancers and outperforms other models, such as the commonly used 2-parameter power law model and a 3-parameter scaled Gamma function model. With the scaled Weibull function, we can estimate the number of driver mutations required for cancer onset in individual cancer types. To our knowledge, this is the

| age | Spearman’s rho | Pearson’s linear correlation |
|-----|----------------|----------------------------|
| 40  | 0.28           | 0.25                       |
| 50  | 0.49           | 0.46                       |
| 60  | 0.57           | 0.58                       |
| 70  | 0.65           | 0.65                       |
| 80  | 0.74           | 0.69                       |

Figure 5. Relationship between cumulative cancer incidences up to age 40, 50, 60, 70, 80 years old and life time stem cell divisions.
first work matching pan-cancer incidence curves with a statistical distribution function that is partially biologically informative.

Compared to the tLIR model developed by Wu et al. we see two technical benefits of our suggested approach:

First, the scaled Weibull function involves less parameters than tLIR, but it remains to be biologically interpretable. The tLIR model includes several details of the multi-staged process of cancer development, e.g. the number of steps required for transforming a normal cell to malignancy, the number of stem cells in a tissue and division rate of stem cells. Although the tLIR model indeed provides useful insights into linking age-dependent somatic mutations to cancer risk, it has also limitations. For example, it ignores the effects of clonal expansion. Another issue is that most parameters in the tLIR model are difficult to measure accurately in practice. Following Tomasetti and Vogelstein the number of stem cell divisions can be estimated, but the accuracy has been criticized. In contrast, our suggested model requires less specific assumptions about the parameters to be measured in practice.

### Table 1

| Cancer | k | r     | R²  | Stem cell | Division rate | Generation | Risk   |
|--------|---|-------|-----|-----------|---------------|------------|--------|
| AML    | 4.8 | 1.000e−06 | 1.00 | 3.5e+08   | 12.000        | 1047.01    | 4.651e−03 |
| BCC    | 4.5 | 1.000e−06 | 0.98 | 5.82e+09  | 7.600         | 678.44     | 2.181e−04 |
| CLL    | 4.9 | 1.000e−06 | 0.99 | 1.35e+08  | 12.000        | 1047.01    | 6.925e−03 |
| COAD   | 5.5 | 5.012e−07 | 1.00 | 2.00e+08  | 73.000        | 6232.58    | 5.677e−02 |
| DUAD   | 5.3 | 1.000e−06 | 1.00 | 4.00e+06  | 24.000        | 2061.93    | 3.714e−04 |
| ESCA   | 4.5 | 5.012e−07 | 0.99 | 8.64e+05  | 17.400        | 1498.72    | 3.106e−03 |
| GBNPAD | 3.5 | 1.000e−06 | 0.85 | 1.60e+06  | 5.84e+07      | 70.25      | 1.896e−03 |
| GBM    | 1.35e+08 | 0.000  | 1.00 | 2.480      | 242.62       | 3.038e−02 |
| HNSC   | 3.8 | 1.995e−07 | 0.99 | 1.85e+07  | 21.500        | 1851.64    | 1.730e−02 |
| LIHCA  | 3.8 | 1.000e−06 | 0.94 | 3.01e+09  | 8912.05      | 109.05     | 7.079e−03 |
| LUAD   | 2.8 | 7.943e−08 | 0.88 | 1.22e+09  | 0.070        | 36.14      | 2.304e−02 |
| MBM    | 1.36e+08 | 0.000  | 1.00 | 2.480      | 242.62       | 3.038e−02 |
| SKCM   | 3.8 | 1.000e−06 | 1.00 | 3.80e+09  | 2.480        | 242.62     | 3.038e−02 |
| OSARC  | 1.0 | 1.585e−08 | 0.96 | 4.18e+06  | 0.067        | 27.69      | 2.696e−04 |
| OSARCA | 1.0 | 6.310e−07 | 0.96 | 6.50e+05  | 0.067        | 25.01      | 2.527e−05 |
| OSARCH | 3.0 | 1.000e−06 | 0.99 | 8.60e+05  | 0.067        | 25.41      | 1.660e−05 |
| OSARCL | 1.0 | 3.981e−07 | 0.96 | 1.59e+06  | 0.067        | 26.30      | 1.312e−04 |
| OSARCP | 3.1 | 1.000e−06 | 0.91 | 4.50e+05  | 0.067        | 24.47      | 3.229e−05 |
| OVGC   | 1.10e+07 | 0.000  | 1.00 | 116.96     | 1.016e−02    | 23.39      | 7.638e−05 |
| PDAD   | 3.8 | 1.000e−06 | 0.92 | 4.18e+09  | 1.000        | 116.96     | 1.016e−02 |
| PECA   | 3.8 | 1.000e−06 | 0.99 | 7.40e+07  | 1.000        | 111.14     | 1.498e−04 |
| SIAD   | 5.0 | 5.012e−07 | 1.00 | 1.00e+08  | 36.000       | 3086.58    | 8.013e−04 |
| TGCC   | 1.9 | 7.943e−07 | 0.96 | 7.20e+06  | 5.800        | 515.78     | 2.244e−03 |
| TPCF   | 3.1 | 1.000e−06 | 0.98 | 6.50e+06  | 0.087        | 33.35      | 6.922e−03 |
| TMCA   | 3.1 | 7.943e−07 | 0.93 | 6.50e+06  | 0.087        | 30.03      | 8.707e−05 |

* Cancers of which parameter estimates are impossible because division rate is 0.

**Figure 6.** Goodness of fit for scaled Weibull function versus that of power law function (a), and scaled Gamma function (b). Each dot represents R² for one cancer subtype defined by the combination of cancer type and one factor such as diagnosis year, race, location and sex. Cancer types are color coded.

**first work matching pan-cancer incidence curves with a statistical distribution function that is partially biologically informative.**

**Compared to the tLIR model developed by Wu et al. we see two technical benefits of our suggested approach:**

First, the scaled Weibull function involves less parameters than tLIR, but it remains to be biologically interpretable. The tLIR model includes several details of the multi-staged process of cancer development, e.g. the number of steps required for transforming a normal cell to malignancy, the number of stem cells in a tissue and division rate of stem cells. Although the tLIR model indeed provides useful insights into linking age-dependent somatic mutations to cancer risk, it has also limitations. For example, it ignores the effects of clonal expansion. Another issue is that most parameters in the tLIR model are difficult to measure accurately in practice. Following Tomasetti and Vogelstein the number of stem cell divisions can be estimated, but the accuracy has been criticized. In contrast, our suggested model requires less specific assumptions about the parameters to be measured in practice.
Moreover, the Weibull distribution is a special case of the generalized extreme value distribution (GEV) which is well connected to classical statistical approaches to describe rare events26.

Our analysis results mostly agree with those provided by Wu et al.11. They defined intrinsic cancer risk as the probability that one tissue transforms from normal to tumor because of accumulated mutations, and extrinsic cancer risk as 1–intrinsic cancer risk. They quantified upper bounds to intrinsic cancer risk by tLIR but did not properly fit tLIR to epidemiological data, concluding intrinsic factor insignificantly contributes to cancer. According to our understanding their argument mainly results from insufficient exploration of parameter space and implicitly assumes that all tissues require the same number of driver mutation to initiate cancer. Our results suggest that the contribution of extrinsic factors to cancer is overestimated by Wu et al.11. However, one should note that the excellent agreement between the scaled Weibull distribution function and empirical data does not necessarily exclude that in addition to intrinsic there are further extrinsic and unknown risk factors. In that context it is worthwhile to mention that our estimated number of driver mutations required for cancer onset differs from tissue to tissue. Although the exact number is not validated by biological experiment, this observation is consistent with findings in genetic studies29.

One interesting observation is that all non-reproductive tissues have a similar cancer risk accumulation pattern. Cancer incidence rates increase dramatically at about 40–50 years, peaking at about 55–70 years and then decrease. This pattern matches findings reported by Podolskiy et al.31. A question for future work is whether mutation load agrees with the scaled Weibull function or age-specific mutational signatures32–34. Another interesting observation is that testicular germ cell cancer incidence peaks at younger age compared to other cancer types, which might be explained by accelerated aging of testis31. Altogether we believe that our suggested approach provides insights into cancer development by providing a link between empirical data and a mechanism-based model.

| Cancer                      | Gamma          | Weibull         |
|-----------------------------|----------------|-----------------|
| LUSC lung squamous cell carcinoma | 2467641.74     | 2439830.08      |
| LUAD lung adenocarcinoma    | 1773103.18     | 1757914.86      |
| KIPAN pan – kidney cohort (kich + kirc + kirp) | 1086052.13     | 1074221.46      |
| BLCA bladder urothelial carcinoma | 1397493.34     | 1364300.97      |
| THCA thyroid carcinoma      | 970479.60      | 966651.56       |
| PAAD pancreatic adenocarcinoma | 684511.50     | 680353.75       |
| ESCA esophageal carcinoma   | 426913.10      | 424019.74       |
| OV ovarian adenocarcinoma   | 249900.19      | 248282.96       |
| SKCM skin cutaneous melanoma | 3200014.23     | 3147559.74      |
| STAD stomach adenocarcinoma | 438574.51      | 431801.81       |
| PRAD prostate adenocarcinoma | 5411659.76     | 5425983.72      |
| COADREAD colorectal adenocarcinoma | 3498130.92 | 3467722.58      |
| GBMLGG glioma               | 433655.37      | 413968.25       |
| BRCA breast invasive carcinoma | 6654036.96     | 6705665.03      |
| SARC sarcoma                | 248310.46      | 238833.58       |
| TGCT testicular germ cell tumors | 128517.41     | 128593.83       |
| HNSC head and neck squamous cell carcinoma | 627828.32 | 626670.00       |
| LIHC liver hepatocellular carcinoma | 510889.67 | 505268.40       |

Table 2. AIC of the scaled Gamma function and the scaled Weibull function.
Methods

Fitting cumulative cancer incidence with a model for theoretical intrinsic cancer risk (tLIR). Wu et al.\(^\text{11}\) provided the following recursion formula to compute the chance that a single stem cell acquires \(k\) mutation hits after \(g\) divisions given a mutation rate \(r\).

\[
\begin{align*}
P(X_{g+1} = i) &= \sum_{j=0}^{i} P(X_{g+1} = i; X_{g} = j)P(X_{g} = j) \\
&= \sum_{j=0}^{i} \binom{k-j}{i-j} r^{i-j}(1-r)^{k-i}P(X_{g} = j) \\
&= (1-r)^{g}P(X_{g} = 0) \\
P(X_{g+1} = k) &= \sum_{j=0}^{k} r^{j-1}P(X_{g} = j)
\end{align*}
\]

(3)

given the initial cell state at generation 0:

\[
P(X_0 = 0) = 1; P(X_0 = 1) = \ldots; P(X_0 = k) = 0.
\]

Here \(X_g\) is accumulated driver mutations at generation \(g\), \(i\) and \(j\) represents accumulated driver mutations at generation \(g\) and \(g+1\), respectively. A fully developed tissue with \(S\) stem cells must go through \(n = \log_2 S + d \cdot a\) rounds of divisions, assuming division rate is \(d\) and age \(a\). With this transition probability (3), the theoretical lifetime intrinsic cancer risk (tLIR) is formulated as

\[
tLIR = 1 - (1 - P(X_n = k))^S
\]

(4)

Although the recursion formula being dependent on more than one parameter cannot directly be solved in closed form by standard algorithmic techniques, it has nevertheless a simple closed form solution, which was derived by hand computations and verified by standard symbolic computations (using the computer algebra system Maple 2015.2):

\[
(1 - (1 - r)^S)^k
\]

The formula for the age-parameterized theoretical lifetime intrinsic cancer risk (tLIR) hence has the following simple closed form solution, which allows much faster and hence more extensive computations and extends the range of admissible values of \(k\) from the positive integers to the positive real numbers:

\[
tLIR(a) = 1 - (1 - (1 - r)^{\log_2 S + d \cdot a})^S.
\]

(5)

Notice that our result basically coincides with the one obtained by Calabrese and Shibata\(^\text{35}\) that was obtained by a direct probabilistic reasoning.

Relating the tLIR model to a scaled Weibull function. We found a connection between

\[
tLIR(a) = 1 - (1 - (1 - r)^{\log_2 S + d \cdot a})^S,
\]

and the scaled Weibull function

\[
1 - (1 - \text{Weibull}(\lambda, k)(\alpha))^\beta,
\]

(7)

where \(P\) is the cell population at risk.

To see this connection, we assume that \(r \ll 1\). Then

\[
f_0 = (1 - (1 - r)^{\log_2 S + d \cdot a})^k = (1 - \exp(1 - r)^{(\log_2 S + d \cdot a)})^k,
\]

and using the Taylor series for \(\log\) and \(\exp\), we obtain

\[
f_0 \approx (rd)^k (d^{-1} \log_2 S + a)^k.
\]

(8)

We have

\[
tLIR(a) = 1 - f_0^S.
\]

(9)

Comparing (7) and (9) we observe that these relations coincide if

\[
(1 - f_0)^S = (1 - \text{Weibull}(\lambda, k)(\alpha))^\beta = \left(\exp\left(-\frac{d \cdot k}{\lambda}\right)\right)^\beta.
\]

Since for small \(f_0 > 0\) we have \(1 - f_0 = \exp(-f_0)\), the last equation can rewritten as
Using that relation and (8) one finds

\[
(rd)^2 (\log S/d + a)^k \approx \left( \frac{a}{\lambda} \right)^{lp}.
\]

So we have obtained a shifted Weibull distribution. However, if we remove \(d^{-1} \log S\) from the left hand side of the last equality assuming that

\[
d \cdot a \gg \log S
\]

we obtain an unshifted one. This condition admits a transparent interpretation, namely, the number of stem cell divisions (for a fixed cell) should be more than the logarithm of stem cell number. Then we have that the tLIR incidence approximately equals to the scaled Weibull incidence if the parameters satisfy

\[
S(rd)^k = \lambda^{-lp}. \tag{10}
\]
Notice that using a Poisson approximation \([12, p. 104]\) we finally obtain
\[
\lambda = \lambda_0 - \lambda_1 P_k a \leq 1 - W_{\text{eibull}}(\lambda_0, \lambda_1) \leq 1 - W_{\text{eibull}}(\lambda_0, \lambda_1)(1 - W_{\text{eibull}}(\lambda_0, \lambda_1)).
\]

\(\text{Stem cell data.}\) Tomasetti and Vogelstein\(^8\) collected stem cell information for 31 cancer types, including stem cell division rate, stem cell number, tissue total cell number. We excluded 6 from 31 cancer types due to lack of age incidence data: colorectal adenocarcinoma in familial adenomatous polyposis (FAP) patients, colorectal adenocarcinoma in patients with hereditary non-polyposis colorectal cancer (HNPCC, also called lynch syndrome), duodenal adenocarcinoma in FAP patients, head and neck squamous cell carcinoma with human papillomavirus (HPV), hepatocellular carcinoma with hepatitis C virus infection (HCV), lung adenocarcinoma in smokers.

Among the 25 remaining cancer types, stem cell information were obtained from supplementary materials of Tomasetti and Vogelstein\(^8\). We discuss life time stem cell division (LSCD) hypothesis and extrinsic risk factor hypothesis for 25 remained cancers: AML, acute myeloid leukemia; BCC, basal cell carcinoma; CLL, chronic lymphocytic leukemia; COAD, colorectal adenocarcinoma; DUAD, duodenum adenocarcinoma; ESCA, esophageal squamous cell carcinoma; GBNPAD, gallbladder non papillary adenocarcinoma; GBM, glioblastoma; HNSC, head and neck squamous cell carcinoma; LHCA, hepatocellular carcinoma; LUAD, lung adenocarcinoma; MBM, medulloblastoma; SKCM, melanoma; OSARC, osteosarcoma; OSARCA, osteosarcoma of the arms; OSARCH, osteosarcoma of the head; OSARCL, osteosarcoma of the legs; OVGC, ovarian germ cell; PDAD, pancreatic ductal adenocarcinoma; PECA, pancreatic endocrine (islet cell) carcinoma; SIAD, small intestine adenocarcinoma; TGCC, testicular germ cell cancer; TPFC, thyroid papillary or follicular carcinoma; TMCA, thyroid medullary carcinoma.

\(\text{Cancer incidence data.}\) SEER-9 registries (1973–2013), SEER-4 registries (1992–2013), SEER-5 registries (2000–2013) data were downloaded from Surveillance, Epidemiology, and End Results Program (SEER) database\(^27\). SEER database covers about 28% USA population, involving more than 100 features such as race, sex,
period, location, histology and ICD (international classification of disease) code. These data were stored in ASCII file, we used the SEERaBomb R package to parse them into sqlite file facilitating data manipulation.

Cancer names provided by Tomasetti and Vogelstein\(^8\) cannot be directly mapped into those in SEER database. We addressed this difficulty by two steps: first, annotate tumor primary site to (international classification of disease- oncology 3) ICD-O-3 code based on the literal sense of site in Tomasetti and Vogelstein\(^8\); second, annotate histology to ICD-O-3 code based on the literal sense of cancer histology by Tomasetti and Vogelstein\(^8\). For instance, primary site of lung adenocarcinoma is lung, corresponding to ICD-O-3 site code: C340, C341, C342, C343, C348, C349; adenocarcinoma of lung cancer corresponds to ICD-O-3 histology code 8140, 8141, 8143, 8147, 8570, 8571, 8572, 8573, 8574, 8575, 8576. The dictionary needed for mapping step (we call it ICD dictionary) can be found in http://seer.cancer.gov/icd-o-3/. Osteosarcoma definition can be found in ICD dictionary, it is a subtype of malignant bone neoplasm, corresponding ICD-O-3 histology code: 9180–9189. However, the ICD dictionary does not differentiate between osteosarcoma detected in the head, leg, or arm. The ICD9Data database (http://www.icd9data.com/) defines bone cancer using ICD9 code 1700–1709, bone cancer in head, arms, legs, pelvis using ICD9 code 1700, 1704–1705, 1707–1708, 1706 respectively. Head and neck squamous cell carcinoma involves tumors located in many sites, ICD dictionary fails to provide its definition. Liao et al.\(^9\) provided ICD9 site code: 1400–1419, 1430–1499, 1600–1619, 1600–1619, we then used ICD-O-3 histology code: 8070–8076, 8087 to select squamous cell carcinoma. More detailed cancer definitions using ICD code can be found in Table 3. Two hematopoietic cancers: acute myeloid leukemia and chronic lymphocytic leukemia, are defined using site recode ICD-O-3/WHO 2008 definition (http://seer.cancer.gov/siterecode/icdo3_dwhoheme/index.html).

Although we carefully annotated 25 cancer definitions using ICD code, we can not avoid misclassifications. Because annotation needs several data sources of which information confidential levels differ from each other. The Cancer Genome Atlas (TCGA) program\(^3\) is a flag project of cancer research hosted by National Institutes of Health, it provides comprehensive, high-quality molecular and clinical data. Cancer definitions are well annotated using ICD code in TCGA clinic documents. We therefore assume TCGA cancer definitions are precise and extracted definitions of 18 solid tumors (Table 4). With 18 cancer definitions, we selected patients who were diagnosed with cancer after 2000 from SEER-9 registries, SEER-4 registries, SEER-5 registries data to form SEER-18 registries data. As the highest time resolution of SEER data is 1 year, for each year, we took middle age for fitting models, for example, 0 year-old is modified as 0.5 years-old.

For robustness analysis of parameter estimates we classified each cancer into subgroups based on location, period and race, data of subgroups were separately fitted to the mathematical models.

**Fitting the models to empirical cancer incidence data.** As was done in previous work\(^18\), empirical cancer incidence \(I(a)\) was calculated by

\[
I(a) = 1 - \prod_{i=0}^{n} (1 - p_i),
\]

where \(p_i\) is frequency of people diagnosed with cancer at age \(i\).

We performed grid search on an extensive parameter space to fit the tLIR model using \(R^2 = \left( \frac{\sum_{i=0}^{n} \sum_{j=0}^{n} (x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_{i=0}^{n} (x_i - \bar{x})^2 \sum_{j=0}^{n} (y_j - \bar{y})^2}} \right)^2\) as the metrics for goodness of fit, where \(x_i\) and \(y_j\) is empirical and model-derived cancer incidence respectively, \(\bar{x}\) and \(\bar{y}\) respectively denotes mean value of \(x\) and \(y\). Results of fits are given in Table 1 showing that there are biologically reasonable parameter combinations that can yield good fits of the tLIR model for most cancer types.

**Data availability.** All data used in this study are publicly available. The sources are detailed in the section on methods.

**References**

1. Feinberg, A. P., Ohiolsson, R., & Henikoff, S. The epigenetic progenitor origin of human cancer. Nat Rev Genet 7, 21–33 (2006).
2. Feinberg, A. P., Koldobskiy, M. A., & Gondor, A. Epigenetic modulators, modifiers and mediators in cancer aetiology and progression. Nat Rev Genet 17, 284–299 (2016).
3. Stratton, M. R., Campbell, P. J., & Futreal, P. A. The cancer genome. Nature 458, 719–724 (2009).
4. Nordling, C. O. A new theory on the cancer-inducing mechanism. British Journal of Cancer 7, 68–72 (1953).
5. Armitage, P., & Doll, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. British Journal of Cancer 8, 1–12 (1954).
6. Fisher, J. Multiple-mutation theory of carcinogenesis. Nature 181, 651–652 (1958).
7. Hornsby, C., Page, K. M., & Tomlinson, I. P. What can we learn from the population incidence of cancer? Armitage and Doll revisited. Lancet Oncol 8, 1030–1038, https://doi.org/10.1016/S1470-2045(07)70343-1 (2007).
8. Tomasetti, C., & Vogelstein, B. Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science 347, 78–81, https://doi.org/10.1126/science.1260825 (2015).
9. Tomasetti, C., & Vogelstein, B. Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science 355, 1330–1334, https://doi.org/10.1126/science.aaf9011 (2017).
10. Tomasetti, C., & Vogelstein, B. On the slope of the regression between stem cell divisions and cancer risk, and the lack of correlation between stem cell divisions and environmental factors-associated cancer risk, PLoS One 12, 1–5, https://doi.org/10.1371/journal. pone.0175535 (2017).
11. Wu, S., Powers, S., Zhu, W., & Hannun, Y. A. Substantial contribution of extrinsic risk factors to cancer development. Nature 529, 43–47, https://doi.org/10.1038/nature16166 (2016).
12. Frank, S. A. Dynamics of Cancer–Incidence, Inheritance, and Evolution http://www.ncbi.nlm.nih.gov/books/NBK1568/pdf/ Bookshef_NBK1568.pdf (2007).
13. Beerwinkel, N., Schwartz, R. F., Gerstung, M., & Markowetz, F. Cancer evolution: Mathematical models and computational inference. Syst. Biol. 64, e1–e25, https://doi.org/10.1093/sysbio/syu081 (2014).
14. Beerli, N. et al. Genetic progression and the waiting time to cancer. *PLoS Comput Biol* 3, e225, https://doi.org/10.1371/journal.pcbi.0030225 (2007).
15. Koer, R. K. & Loeb, L. A. Significance of multiple mutations in cancer. *Carcinogenesis* 21, 379–385, https://doi.org/10.1093/carcin/21.3.379 (2000).
16. Jackson, A. L. & Loeb, L. A. The mutation rate and cancer. *Genetics* 148, 1483–1490 (1998).
17. Albini, A. & Sporn, M. B. The tumour microenvironment as a target for chemoprevention. *Nat Rev Cancer* 7, 139–147, https://doi.org/10.1038/nrc2067 (2007).
18. Borovski, T., De Sousa E Melo, F., Vermeulen, L. & Medema, J. P. Cancer stem cell niche: The place to be. *Cancer Research* 71, 634–639, https://doi.org/10.1158/0008-5472.CAN-10-3220 (2011).
19. Vermeulen, L. et al. Defining stem cell dynamics in models of intestinal tumour initiation. *Science* 342, 995–998, https://doi.org/10.1126/science.1243148 (2013).
20. Frank, S. A. Somatic mutation: Early cancer steps depend on tissue architecture. *Current Biology* 13, R261–R263, https://doi.org/10.1016/S0960-9822(03)00195-7 (2003).
21. Nowak, M. A., Michor, F. & Iwasa, Y. The linear process of somatic evolution. *Proceedings of the National Academy of Sciences of the United States of America* 100, 14966–14969, https://doi.org/10.1073/pnas.2535419100 (2003).
22. Waclaw, B. et al. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature* 525, 261–264, https://doi.org/10.1038/nature14971 (2015).
23. Durrett, R. & Moseley, S. Evolution of resistance and progression to disease during clonal expansion of cancer. *Theor. Popul. Biol.* 77, 42–48, https://doi.org/10.1016/j.tpb.2009.10.008 (2010).
24. Yamamoto, K. N., Nakamura, A. & Haeno, H. The evolution of tumor metastasis during clonal expansion with alterations in metastasis driver genes. *Sci. Rep.* 5, 15886, https://doi.org/10.1038/srep15886 (2015).
25. Nowak, M. A. & Waclaw, B. Genes, environment, and “bad luck.” Science 355, 1266–1267, https://doi.org/10.1126/science.aam9746 (2017).
26. Coles, S. An Introduction to Statistical Modeling of Extreme Values (Springer, 2001).
27. National Cancer Institute, DCCPS, Surveillance Research Program, Surveillance Systems Branch. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1973–2012) (2015).
28. Anderson, D. R. *Model Based Inference in the Life Sciences* (Springer, 2008).
29. Vogelstein, B. et al. Cancer genome landscapes. *Science* 339, 1546–1558, https://doi.org/10.1126/science.1235122 (2013).
30. O’Callaghan, M. Cancer risk: Accuracy of literature. *Nature Communications* 7, 1–12, https://doi.org/10.1038/ncomms12157 (2016).
31. Podolsky, D. I., Lobanov, A. V., Kryukov, G. V. & Gladyshev, V. N. Analysis of cancer genomes reveals basic features of human aging and its role in cancer development. *Nature Communications* 7, 1–12, https://doi.org/10.1038/ncomms12157 (2016).
32. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* 500, 415–421, https://doi.org/10.1038/nature12477, 23945592[pmid] (2013).
33. Alexandrov, L. B. et al. A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature* 525, 261–264, https://doi.org/10.1038/nature14971 (2015).
34. Calabrese, P. & Shibata, D. A simple algebraic cancer equation: calculating how cancers may arise with normal mutation rates. *BMC Cancer* 10, 3, https://doi.org/10.1186/1471-2407-10-3 (2010).
35. Liao, L.-J., Chou, H.-W., Wang, C.-T., Chung, C.-S. & Lai, M.-S. The impact of second primary malignancies on head and neck cancer survivors: A nationwide cohort study. *PloS ONE* 8, e62116, https://doi.org/10.1371/journal.pone.0062116 (2013).
36. TCGA working group. TCGA data portal https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm (2016).
37. Borovski, T., De Sousa E Melo, F., Vermeulen, L. & Medema, J. P. Cancer stem cell niche: The place to be. *Cancer Research* 71, 634–639, https://doi.org/10.1158/0008-5472.CAN-10-3220 (2011).