CALCULATING EPIDEMIOLOGICAL OUTCOMES FROM SIMULATED LONGITUDINAL DATA

Selina Pi
Department of Biomedical Data Science, School of Medicine
Stanford University
Stanford, CA, USA

Jeremy D. Goldhaber-Fiebert
Department of Health Policy, School of Medicine
Center for Health Policy, Freeman-Spogli Institute for International Studies
Stanford University
Stanford, CA, USA

Fernando Alarid-Escudero[*]
Department of Health Policy, School of Medicine
Center for Health Policy, Freeman-Spogli Institute for International Studies
Stanford University
Stanford, CA, USA

April 30, 2025

ABSTRACT

Microsimulation models generate individual life trajectories that must be summarized as population-level outcomes for model calibration and validation. While there are established formulas to calculate outcomes such as prevalence, incidence, and lifetime risk from cross-sectional and short-term longitudinal studies, limited guidance exists to calculate these outcomes using long-term longitudinal data due to the rarity of large-scale studies covering events across the human lifespan. This technical report presents various methods to calculate epidemiological outcomes from simulated longitudinal data, from replicating a real-world study design to fully incorporating longitudinal disease and exposure durations. We provide an open-source code base with functions in R to calculate the prevalence, incidence, age-conditional risk, lifetime risk, and disease-specific mortality of a condition from individual-level time-to-event data. In addition, we provide guidance and code for calculating cancer-related outcomes from individual-level data, such as the stage distribution at diagnosis, the distribution of precancerous lesion multiplicity, and the mean dwell and sojourn time. Given the various possible formulations for certain outcomes, we call for increased transparency in reporting how summary outcomes are derived from microsimulation model outputs, and we anticipate that this report will facilitate the calculation of epidemiological outcomes in both simulated and real-world data.

***K*eywords** microsimulation · epidemiology · prevalence · incidence

1 Introduction

Calibration and validation of a disease natural history model center around ensuring that the model outputs match relevant epidemiological endpoints, such as incidence, prevalence, lifetime risk, and disease-specific mortality [1]. These statistics may be derived from screening studies or disease surveillance data. While formulas for prevalence

[*] Corresponding author

Variables	
T_X	Start age of one-time condition X
$T_{\bar{X}}$	End age of one-time condition $\bar{X}$
T_{X_i}	Start age of i-th occurrence of recurrent condition X
$T_{\bar{X}_i}$	End age of i-th occurrence of recurrent condition X
T_C	Censor age (e.g., age at which individual is no longer alive, at risk for the condition, or otherwise eligible for the study)
T_S	Study or observation age for cross-sectional study, or beginning of study period for longitudinal study
$T_{\bar{S}}$	End of study period for longitudinal study
General notation	
T_x^n	Value of event time T_x for individual n
T_{xy}	Time from event x to event y (i.e., $T_y - T_x$)
$F_x(t)$	Cumulative distribution function (CDF) of event x over time t
$f_x(t)$	Probability density function (PDF) of event x over time t
$\mathbf{P}(x)$	Probability of event x
$\mathbf{1}\{x\}$	Indicator variable for the condition x, equaling 1 if x is true and 0 otherwise
x^+	Nonnegative part of the quantity x (i.e., $\max(x, 0)$)

Table 1: Notation for events and quantities

and incidence have been described for population-level differential equation models [2, 3, 4, 5, 6], there is limited guidance on calculating such summary outcomes from simulated time-to-event data, possibly because longitudinal data analogous to the life trajectories generated by an individual-level simulation model are rare in practice. Prevalence and incidence have been used as calibration targets for several microsimulation models (e.g., [7, 8, 9, 10, 11, 12, 13]), but their documentation does not specify how these outcomes were calculated from the model outputs. This technical report provides mathematical formulas extrapolating cross-sectional and short-term longitudinal outcome definitions to the lifetime setting for simulated individual trajectories. We demonstrate that there are multiple ways to calculate certain outcomes from lifetime data, with tradeoffs in computational expense, precision, and bias. Therefore, we argue that it is important for researchers to transparently report how these outcomes are calculated for microsimulation models. In addition to prevalence and incidence, we discuss how to calculate age-conditional and lifetime risk, disease-specific mortality, and cancer-related outcomes, such as the stage distribution at diagnosis, the distribution of the number of lesions or tumors, and the mean dwell or sojourn time. Functions to calculate outcomes for one-time conditions from a matrix of life trajectories are provided in an open-source repository (`https://github.com/sjpi22/microsumulation/`).

2 Outcomes

We define variables for event times and other notation used across this report in Table 1. Event episodes for recurrent conditions are assumed not to overlap, so $T_{\bar{X}_i} > T_{X_i}$ and $T_{X_{i+1}} > T_{\bar{X}_i}$. We assume that there are no missing values. For events that would have occurred after an individual's death or events for which a person is not at risk, the event time is assumed to be infinite. Censoring can be due to death, loss to follow-up, or loss of eligibility, for example due to the diagnosis of a condition prior to a screening study of asymptomatic individuals. For the purpose of calibration, when simulating individual life histories, care should be taken to ensure that the simulated population is representative (e.g., has the same baseline characteristics) as that of the empirical studies from which the calibration target of interest is derived. The formulas assume a cohort of N life histories representative of the population or subgroups for which outcomes will be calculated.

2.1 Prevalence

The prevalence of a condition is defined as the proportion of individuals with the condition out of the population at risk [14]. In cross-sectional studies and disease surveillance reporting, prevalence is often stratified by age groups to reflect age-specific differences in risk, though practical and ethical considerations often preclude reporting exact individual ages [15]. In terms of probabilities $\mathbf{P}(x)$, the true prevalence $P(t)$ of a one-time condition X at t according to the notation in Table 1 is:

$$\begin{aligned} P(t) &= \frac{\mathbf{P}\left(T_X \le t < T_{\bar{X}}, T_C > t\right)}{\mathbf{P}\left(T_C > t\right)} \\ &= \frac{\int_0^t \mathbf{P}\left(T_{\bar{X}} > t, T_C > t | T_X = u\right) f_X(u) du}{\mathbf{P}\left(T_C > t\right)} \\ &= \frac{\int_0^t \mathbf{P}\left(\min\left(T_{\bar{X}}, T_C\right) > t | T_X = u\right) f_X(u) du}{\mathbf{P}\left(T_C > t\right)}. \end{aligned} \tag{1}$$

When censoring is independent of the condition and the condition does not end before the censor time, Equation 1 simplifies to $F_X(t)$. Assuming no birth cohort effects, the true prevalence $P(t_1, t_2)$ of the condition from age t_1 to t_2 is the expected value of the single-age prevalence from t_1 to t_2, weighted by the proportion of the population that is uncensored:

$$P\left(t_1, t_2\right) = \frac{\int_{t_1}^{t_2} \left(1 - F_C(u)\right) P(u) du}{\int_{t_1}^{t_2} \left(1 - F_C(u)\right) du}. \tag{2}$$

In this section, we provide three approaches to calculate age-specific disease prevalence from simulated life trajectories and compare the Monte Carlo variation, computation time, and bias from the true prevalence associated with each method.

2.1.1 Cross-sectional formulation (point prevalence)

Age-specific prevalence can be calculated for a simulated cohort in a cross-sectional manner analogous to real-world studies by sampling a study or observation age T_S for each individual, excluding individuals who are censored before T_S or otherwise ineligible for the study, and calculating the proportion of at-risk individuals in the age range of interest with the condition at T_S. T_S can be determined by assigning birth years Y_{birth} representative of the population, setting a study year or sampling from a range of years Y_S, and calculating $T_S = Y_S - Y_{birth}$. Alternately, T_S can be randomly sampled between the age ranges of interest. A uniform distribution between the minimum and maximum ages for which prevalence will be calculated would provide a representative estimate of the population in terms of the relative frequency of individuals at each age who have not been censored. To calculate the cross-sectional or point prevalence P_{cs} of a condition X between age t_1 (inclusive) to t_2 (exclusive), we use the following formula across the simulated life trajectories n and event episodes i:

$$\begin{aligned} P_{cs}(t_1, t_2) &= \frac{N_{cases}}{N_{risk}} \\ &= \frac{\sum_n \sum_i \mathbf{1}\left\{T^n_{X_i} \le T^n_S < T^n_{\bar{X}_i}, T^n_S < T^n_C, t_1 \le T^n_S < t_2\right\}}{\sum_n \mathbf{1}\left\{T^n_S < T^n_C, t_1 \le T^n_S < t_2\right\}} \\ &= \frac{\sum_n \sum_i \mathbf{1}\left\{\max\left(T^n_{X_i}, t_1\right) \le T^n_S < \min\left(T^n_{\bar{X}_i}, T^n_C, t_2\right)\right\}}{\sum_n \mathbf{1}\left\{t_1 \le T^n_S < \min\left(T^n_C, t_2\right)\right\}}, \end{aligned} \tag{3}$$

where N_{cases} indicates the number of individuals with the condition at T_S among the N_{risk} at-risk individuals observed between age t_1 and t_2. For a one-time condition, Equation 3 simplifies to

$$P_{cs}(t_1, t_2) = \frac{\sum_n \mathbf{1}\left\{\max\left(T^n_X,\ t_1\right) \le T^n_S < \min\left(T^n_{\bar{X}}, T^n_C, t_2\right)\right\}}{\sum_n \mathbf{1}\left\{t_1 \le T^n_S < \min\left(T^n_C, t_2\right)\right\}}. \tag{4}$$

As an example, for the prevalence of colorectal cancer (CRC) in a screening study (e.g., among those without diagnosed CRC) of average-risk individuals, T_X would be the age at CRC onset, $T_{\bar{X}}$ would be the age at the symptomatic detection of CRC, and T_C would be the earliest of the age of death and the age of symptomatic CRC diagnosis in Equation 4. We assume that individuals with inflammatory bowel disease, familial adenomatous polyposis, and other conditions associated with an increased risk of CRC are already excluded. In this case, $T_{\bar{X}}$ is redundant to T_C, but this does not always occur; for instance, for the prevalence of precancerous lesions without CRC in the same study, T_X would be the age at which the first lesion develops, $T_{\bar{X}}$ would be the age at onset of CRC (assuming no lesion regression), and T_C would be the same as for preclinical CRC prevalence.

To calculate the prevalence of a condition at a single age t, T_S, t_1, and t_2 are all set to t, and prevalence is calculated as the number of individuals with the condition out of the at-risk individuals at age t, making Equation 3 simplify to:

$$P_{cs}(t) = \frac{\sum_n \sum_i \mathbf{1}\left\{T^n_{X_i} \leq t < \min\left(T^n_{\bar{X}_i}, T^n_C\right)\right\}}{\sum_n \mathbf{1}\left\{t < T^n_C\right\}}. \tag{5}$$

Conceptualizing prevalence as a proportion, the Wilson score interval with continuity correction can be used to calculate confidence intervals (CIs) for cross-sectional prevalence [16]. To estimate the standard error (SE), we can use the binomial approximation:

$$SE = \sqrt{\frac{p(1-p)}{N_{risk}}}, \tag{6}$$

where p is the prevalence estimate and N_{risk} is the number of individuals in the denominator for the estimate.

2.1.2 Longitudinal formulation

Assuming negligible birth cohort effects within the age intervals of interest, we can make more efficient use of the simulated longitudinal data by calculating the prevalence P_{long} of a condition X from age t_1 to t_2 as the proportion of time at risk in the age range that individuals have the condition:

$$P_{long}(t_1, t_2) = \frac{\sum_n \sum_i \left(\min\left(T^n_{\bar{X}_i}, T^n_C, t_2\right) - \max\left(T^n_{X_i}, t_1\right)\right)^+}{\sum_n \left(\min\left(T^n_C, t_2\right) - t_1\right)^+}. \tag{7}$$

2.1.3 Repeated cross-sectional formulation

The repeated cross-sectional (RCS) formulation of prevalence is a hybrid of the cross-sectional and longitudinal approaches that sums the numerators and denominators of the single-age prevalence (Equation 5) at each integer age from t_1 to t_2 and is well-suited to discrete-time microsimulation outputs:

$$P_{rcs}(t_1, t_2) = \frac{\sum_{t=t_1}^{t_2} \sum_n \sum_i \mathbf{1}\left\{T^n_{X_i} \leq t < \min\left(T^n_{\bar{X}_i}, T^n_C\right)\right\}}{\sum_{t=t_1}^{t_2} \sum_n \mathbf{1}\left\{t < T^n_C\right\}}. \tag{8}$$

2.1.4 Comparison of formulations

We compare the three formulations with a simulation study in which half of the population is at risk of developing a condition X whose onset time follows a Weibull distribution with shape 4 and scale 60. Individuals die uniformly between age 30 and 100 independently of the condition, and the condition lasts until death. Therefore, the true prevalence of the condition at any time point is merely half value of the Weibull CDF at that point (see Figure 1), and the true prevalence in an age range is calculated using Equation 2 with $P(u) = F_X(u)$. We simulate 1,000 cohorts of size 100,000 and calculate the cross-sectional, longitudinal, and repeated cross-sectional prevalence of the condition across the population from age 30 to 80 and within 10-year age ranges from 30 to 80. The top two panels of Figure 2 show the mean estimate for each formulation in terms of the percentage difference from the true prevalence (e.g., $(\text{estimated} - \text{true})/\text{true}$), and the error bars represent the standard deviation of the 1,000 estimates as a proportion of the true prevalence. The bottom two panels reflect the mean time per calculation.

To calculate prevalence in a single age range, the longitudinal formulation is fastest with the lowest Monte Carlo variation. The repeated cross-sectional method demonstrates considerable bias for a large age range, as the true

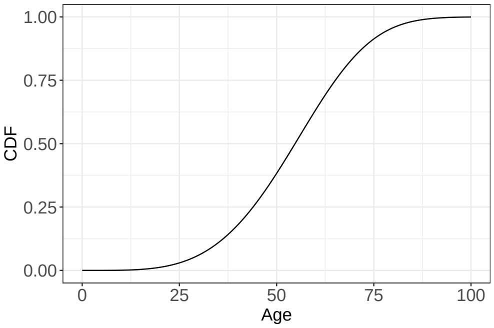


Figure 1: Cumulative distribution function for Weibull distribution with shape 4 and scale 60

prevalence is not within one standard deviation of the mean estimate. For a typical use case, in which prevalence is calculated for several smaller age ranges, the longitudinal formulation exceeds the other two methods in computation time but produces less than half the Monte Carlo variation of the cross-sectional method. The repeated cross-sectional method is faster and has similar variance as the longitudinal method. It is on average slightly biased from the true prevalence, though the magnitude of bias is lower than the standard deviation of the estimates.

Overall, the longitudinal method has the optimal level of Monte Carlo variation and bias, but its computation time increases linearly with the number of age brackets, whereas the computation times of the cross-sectional and repeated cross-sectional methods appear to be stable with the number of age ranges. The repeated cross-sectional method has more precision than the cross-sectional method and is faster than the longitudinal method when there is a sufficient number of age brackets, but it can produce biased estimates when the single-age prevalence function has a high degree of curvature. Given how it is calculated, we hypothesize that the computation time increases linearly with the width of the full age range to calculate prevalence. To choose which formulation to use, we recommend assessing the tradeoffs between the computation time, necessary precision, and amount of bias across the three methods for the specific use case and have provided template code in the repository to reproduce Figure 2 to do so.

2.2 Incidence

Incidence is defined as the number of events per person-year at risk [17]. For age-specific incidence, the denominator is the total person-years at risk within the age range of interest [18]. The time required to observe new cases means that incidence cannot be measured in cross-sectional studies [19]. The Surveillance, Epidemiology, and End Results (SEER) Program calculates the crude incidence rate as the number of new cases or events in a given year divided by the total population, or the total population of a given sex for sex-specific cancers [20]. Cancer recurrences are generally not included as incident cases [21]. For rare conditions with a low incidence, the rate is often reported per 100,000. Here, we show formulas for crude incidence rates, which are acceptable for the narrow age ranges for which cancer incidence targets are typically reported (e.g., 5- or 10-year age ranges). For the modifications needed to obtain age-adjusted incidence rates across a full population, we refer to [20].

With lifetime data, we can calculate the incidence of a condition between age t_1 and t_2 from simulated life trajectories n and event episodes i as:

$$I(t_1, t_2) = \frac{N_{events}}{T_{risk}} = \frac{\sum_n \sum_i \mathbf{1}\left\{t_1 \leq T^n_{X_i} < \min\left(T^n_C, t_2\right)\right\}}{\sum_n \left(\min(T^n_C, t_2) - t_1\right)^+}, \tag{9}$$

where N_{events} is the number of new condition onset events from t_1 to t_2 and T_{risk} is the total person-years at risk from t_1 to t_2. For a one-time condition, Equation 9 simplifies to:

$$I(t_1, t_2) = \frac{\sum_n \mathbf{1}\left\{t_1 \leq T^n_X < \min\left(T^n_C, t_2\right)\right\}}{\sum_n \left(\min(T^n_C, t_2) - t_1\right)^+}. \tag{10}$$

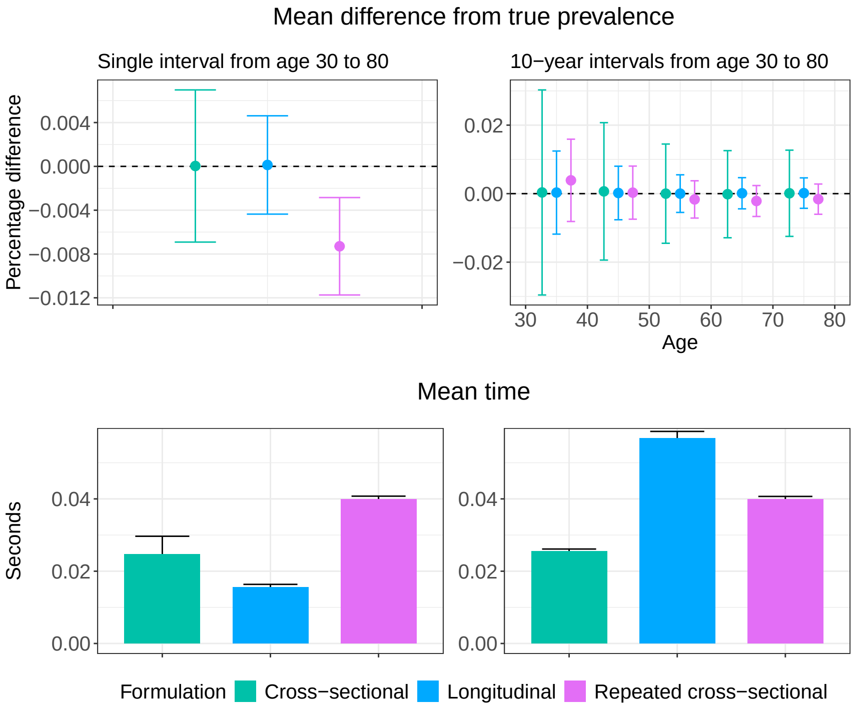


Figure 2: Comparison of formulations for aggregate prevalence from age 30 to 80 (left) and prevalence in 10-year intervals from age 30 to 80 (right), with standard deviation of the estimates shown as error bars

As an example, to calculate the incidence rate of cancer diagnoses among individuals not previously diagnosed, T_X would be the age at cancer diagnosis and T_C would be the earliest of the ages at death and cancer diagnosis in Equation 10. For the incidence of cancer diagnoses out of the total population, T_C would be the age of death.

To create an analog of a real-world longitudinal study, each individual would have a study or observation period from T_S to $T_{\bar{S}}$, and Equation 9 would be modified by bounding the event and exposure times by the observation period, as described in [18]:

$$I(t_1, t_2) = \frac{\sum_n \sum_i \mathbf{1}\left\{\max\left(T_S, t_1\right) \leq T_X^n < \min\left(T_{\bar{S}}, T_C^n, t_2\right)\right\}}{\sum_n \left(\min\left(T_{\bar{S}}, T_C^n, t_2\right) - \max\left(T_S, t_1\right)\right)^+}. \tag{11}$$

Following [20], assuming that the annual rate of cases is Poisson-distributed, the SE of the incidence rate of a one-time condition is calculated as:

$$SE = \frac{\sqrt{N_{events}}}{T_{risk}}, \tag{12}$$

and the bounds of the $(1-\alpha) \times 100\%$ CIs are calculated as:

$$\begin{aligned} CI_{lower} &= \left(\frac{\chi^2\left(\frac{\alpha}{2}, 2N_{events}\right)}{2T_{risk}}\right) \\ CI_{upper} &= \left(\frac{\chi^2\left(1-\frac{\alpha}{2}, 2\left(N_{events}+1\right)\right)}{2T_{risk}}\right), \end{aligned} \tag{13}$$

where N_{events} and T_{risk} are as defined for Equation 9, and $\chi^2\left(\alpha, \nu\right)$ is the value x at which the chi-squared distribution with ν degrees of freedom has a probability α of being greater than x.

2.3 Age-conditional and lifetime risk

The risk of developing a condition X for the first time between age t_1 and t_2 conditional on not having developed the condition before t_1 is calculated as:

$$R(t_1, t_2) = \frac{\sum_n \mathbf{1}\left\{t_1 \leq T_X^n < \min\left(T_C^n, t_2, \right)\right\}}{\sum_n \mathbf{1}\left\{t_1 \leq \min\left(T_X^n, T_C^n\right)\right\}} \tag{14}$$

The lifetime risk of the condition can be calculated by setting $t_1 = 0$ and $t_2 = \infty$, making Equation 14 simplify to:

$$R(0, \infty) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}\left\{T_X^n < T_C^n\right\}. \tag{15}$$

2.4 Disease-specific mortality

For a microsimulation model in which the cause of death is simulated, we can use standard competing risk methods to calculate the disease-specific mortality and survival rates from diagnosis. We fit a Kaplan-Meier curve with event time $T_{XC} = T_C - T_X$ and event indicator D to the population for which $T_X \leq T_C$, where T_X is the time at condition onset, T_C is the time of death, and D is 1 if a simulated individual's death is due to the condition and 0 if the death is due to other causes. The population can first be filtered to subgroups such as individuals diagnosed at a particular age or stage of disease. The Kaplan-Meier curve provides the proportion that have not died of cancer t years after diagnosis.

2.5 Stage distribution of cancer

The stage distribution of cancer is the proportion diagnosed at each stage among individuals diagnosed with cancer in their lifetimes. Letting T_X be the time at the diagnosis of cancer, T_C be the time of death, and $S(T_X)$ be the stage at diagnosis, the proportion of individuals with cancer diagnosed at stage s is calculated as:

$$\mathbf{P}\left(S\left(T_X\right) = s\right) = \frac{\sum_n \mathbf{1}\left\{T_X^n < T_C^n, S^n(T_X^n) = s\right\}}{\sum_n \mathbf{1}\left\{T_X^n < T_C^n\right\}}. \tag{16}$$

2.6 Lesion multiplicity

The distribution of precancerous lesion multiplicity is the proportion of individuals with a given number of concurrent lesions among individuals with at least one lesion. Letting $L(t_1, t_2)$ be the number of concurrent lesions at an observation time between t_1 to t_2, $T_X^{n,j}$ and $T_{\bar{X}}^{n,j}$ respectively be the times at which lesion j in individual n develops and ends, T_S^n be the age at which an individual is observed for lesions (e.g., through screening), and T_C^n be the earliest of the time of death or the time at which cancer first develops from any lesion for individual n, we calculate the proportion of individuals with x lesions among those with any lesions as:

$$\begin{aligned}\mathbf{P}(L(t_1, t_2) = x) &= \frac{\sum_n \mathbf{1}\left\{L^n(t_1, t_2) = x\right\}}{\sum_n \mathbf{1}\left\{L^n(t_1, t_2) \geq 1\right\}}, \\ L^n(t_1, t_2) &= \sum_j \mathbf{1}\left\{\max\left(T_X^{n,j}, t_1\right) \leq T_S^n < \min\left(T_{\bar{X}}^{n,j}, T_C^n, t_2\right)\right\}.\end{aligned} \tag{17}$$

The distribution can also be calculated longitudinally by converting the lesion-level duration data to mutually exclusive episodes with a single number of lesions. Similar calculations can be used for other characteristics of precancerous lesions and tumors, such as the distribution of the largest lesion size.

2.7 Dwell and sojourn time

Modeling may help estimate intermediate outcomes of cancer natural history that are difficult to observe. These outcomes include sojourn time, defined as the time from the onset to the diagnosis of cancer; precancerous lesion dwell time, or the time from precancerous lesion onset to the onset of preclinical cancer; and stage-specific dwell time, or the time from one stage of cancer to the next. The mean duration generally follows the formula:

$$\bar{T}_{X\bar{X}} = \frac{\sum_n \left(T^n_{\bar{X}} - T^n_X\right) \mathbf{1}\left\{T^n_Y < T^n_C, Z^n\right\}}{\sum_n \mathbf{1}\left\{T^n_Y < T^n_C, Z^n\right\}}, \tag{18}$$

where T_Y is an event that must occur before the censor time for an individual to be included in the statistic and Z is a catch-all for any other inclusion criteria. For the mean sojourn time among individuals diagnosed with cancer in their lifetimes, T_X is the onset time of preclinical cancer, $T_{\bar{X}}$ is the time at diagnosis of cancer, T_Y is also the time at diagnosis, T_C is the time of death, and Z contains no restrictions. For cancers that include a precursor lesion state, the mean dwell time of lesions among individuals diagnosed with cancer is calculated with T_X equal to the onset time of the first lesion, $T_{\bar{X}}$ equal to the onset of preclinical cancer assuming no lesion regression, and T_Y, T_C, and Z equal to the same quantities as for mean sojourn time. For the dwell time from stage s to $s+1$ of cancer before diagnosis among individuals diagnosed at or after stage $s+1$, T_X equals the start time of stage s, $T_{\bar{X}}$ equals the start time of stage $s+1$, T_Y is the time at diagnosis, T_C is the time of death, and Z restricts to individuals diagnosed at stage $s+1$ or later.

3 Discussion

In this technical report, we show how to map longitudinal event data as produced by individual-level models to population outcomes that can be compared with statistics typically reported in real-world epidemiological studies, such as prevalence, incidence, and age-conditional risk. In addition, we provide equations for other useful statistics from microsimulation model outputs, such as the stage distribution of cancer and the mean duration of a health state. The formulas for these summary statistics will also facilitate likelihood computation for model calibration. Prevalence can be calculated either by replicating a cross-sectional study or using longitudinal condition and risk durations as inputs, which results in lower Monte Carlo variation but may be more computationally expensive. The precision gains of the longitudinal formulation compared to the cross-sectional formulation essentially come from a longer observation period and from allowing individuals to contribute data to multiple age groups. When calculating analogs of real-world targets with simulated longitudinal data, care should be taken to define event time inputs and other eligibility criteria for the simulated population consistent with the comparator study. Given the multiple possible formulations and the sensitivity of outcomes to the characteristics of the study population, we recommend that modelers report equations, functions, or both for epidemiological outcome calculations used for microsimulation models.

4 Acknowledgments

SP was supported by the US National Institutes of Health (NIH Grant T15LM007033) and the National Science Foundation (NSF) Graduate Research Fellowship Program (Grant DGE-2146755). FAE was supported by grants U01-CA253913 and U01-CA265750 from the National Cancer Institute (NCI) as part of the Cancer Intervention and Surveillance Modeling Network (CISNET). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NIH, NSF, or NCI. Authors have no competing interests to declare.

References

[1] Tazio Vanni, Jonathan Karnon, Jason Madan, Richard G. White, W. John Edmunds, Anna M. Foss, and Rosa Legood. Calibrating Models in Economic Evaluation. *PharmacoEconomics*, 29(1):35–49, January 2011.

[2] S. Haberman. Mathematical treatment of the incidence and prevalence of disease. *Social Science & Medicine. Part A: Medical Psychology & Medical Sociology*, 12:147–152, January 1978.

[3] Jeremy A. Lauer, Klaus Röhrich, Harald Wirth, Claude Charette, Steve Gribble, and Christopher JL Murray. PopMod: a longitudinal population model with two interacting disease states. *Cost Effectiveness and Resource Allocation*, 1(1):6, February 2003.

[4] Elamin H. Elbasha, Erik J. Dasbach, and Ralph P. Insinga. Model for Assessing Human Papillomavirus Vaccination Strategies - Volume 13, Number 1—January 2007 - Emerging Infectious Diseases journal - CDC. 13(1):28, January 2007.

[5] Fernando Alarid-Escudero, Valeria Gracia, Marina Wolf, Ran Zhao, Caleb W Easterly, Jane J Kim, Karen Canfell, Inge M C M de Kok, Ruanne V Barnabas, and Shalini Kulasingam. State-level disparities in cervical cancer prevention and outcomes in the United States: a modeling study. *JNCI: Journal of the National Cancer Institute*, 117(4):737–746, April 2025.

[6] Fernando Alarid-Escudero, Jason R. Andrews, and Jeremy D. Goldhaber-Fiebert. Effects of Mitigation and Control Policies in Realistic Epidemic Models Accounting for Household Transmission Dynamics. *Medical Decision Making*, 44(1):5–17, January 2024. Publisher: SAGE Publications Inc STM.

[7] Megumi Kasajima, Karen Eggleston, Shoki Kusaka, Hiroki Matsui, Tomoki Tanaka, Bo-Kyung Son, Katsuya Iijima, Kazuo Goda, Masaru Kitsuregawa, Jay Bhattacharya, and Hideki Hashimoto. Projecting prevalence of frailty and dementia and the economic cost of care in Japan from 2016 to 2043: a microsimulation modelling study. *The Lancet Public Health*, 7(5):e458–e468, May 2022. Publisher: Elsevier.

[8] Yoon-Sun Jung, Young-Eun Kim, Dun-Sol Go, and Seok-Jun Yoon. Projecting the prevalence of obesity in South Korea through 2040: a microsimulation modelling approach. *BMJ Open*, 10(12):e037629, December 2020. Publisher: British Medical Journal Publishing Group Section: Public health.

[9] Jacek A. Kopec, Eric C. Sayre, Jolanda Cibere, Linda C. Li, Hubert Wong, Anya Okhmatovskaia, and John M. Esdaile. Reducing the burden of low back pain: results from a new microsimulation model. *BMC Musculoskeletal Disorders*, 23(1):804, August 2022.

[10] Scott B. Patten, Lee Gordon-Brown, and Graham Meadows. Simulation studies of age-specific lifetime major depression prevalence. *BMC Psychiatry*, 10(1):85, October 2010.

[11] Bin Lu, Le Wang, Ming Lu, Yuhan Zhang, Jie Cai, Chenyu Luo, Hongda Chen, and Min Dai. Microsimulation Model for Prevention and Intervention of Coloretal Cancer in China (MIMIC-CRC): Development, Calibration, Validation, and Application. *Frontiers in Oncology*, 12, April 2022. Publisher: Frontiers.

[12] Chih-Yuan Cheng, Silvia Calderazzo, Christoph Schramm, and Michael Schlander. Modeling the Natural History and Screening Effects of Colorectal Cancer Using Both Adenoma and Serrated Neoplasia Pathways: The Development, Calibration, and Validation of a Discrete Event Simulation Model. *MDM Policy & Practice*, 8(1):23814683221145701, January 2023. Publisher: SAGE Publications Inc.

[13] Vahab Vahdat, Oguzhan Alagoz, Jing Voon Chen, Leila Saoud, Bijan J. Borah, and Paul J. Limburg. Calibration and Validation of the Colorectal Cancer and Adenoma Incidence and Mortality (CRC-AIM) Microsimulation Model Using Deep Neural Networks. *Medical Decision Making*, 43(6):719–736, August 2023. Publisher: SAGE Publications Inc STM.

[14] Steven Tenny and Mary R. Hoffman. Prevalence. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2025.

[15] Theresa Diaz, Kathleen L. Strong, Bochen Cao, Regina Guthold, Allisyn C. Moran, Ann-Beth Moller, Jennifer Requejo, Ritu Sadana, Jotheeswaran Amuthavalli Thiyagarajan, Emmanuel Adebayo, Elsie Akwara, Agbessi Amouzou, John J. Aponte Varon, Peter S. Azzopardi, Cynthia Boschi-Pinto, Liliana Carvajal, Venkatraman Chandra-Mouli, Sarah Crofts, Saeed Dastgiri, Jeremiah S. Dery, Shatha Elnakib, Lucy Fagan, B. Jane Ferguson, Julia Fitzner, Howard S. Friedman, Ann Hagell, Eduard Jongstra, Laura Kann, Somnath Chatterji, Mike English, Philippe Glaziou, Claudia Hanson, Ahmad R. Hosseinpoor, Andrew Marsh, Alison P. Morgan, Melinda K. Munos, Abdisalan Noor, Boris I. Pavlin, Rich Pereira, Tyler A. Porth, Joanna Schellenberg, Rizwana Siddique, Danzhen You, Lara M. E. Vaz, and Anshu Banerjee. A call for standardised age-disaggregated health data. *The Lancet Healthy Longevity*, 2(7):e436–e443, July 2021. Publisher: Elsevier.

[16] Alan Agresti and Brent A. Coull. Approximate Is Better than "Exact" for Interval Estimation of Binomial Proportions. *The American Statistician*, 52(2):119–126, 1998. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

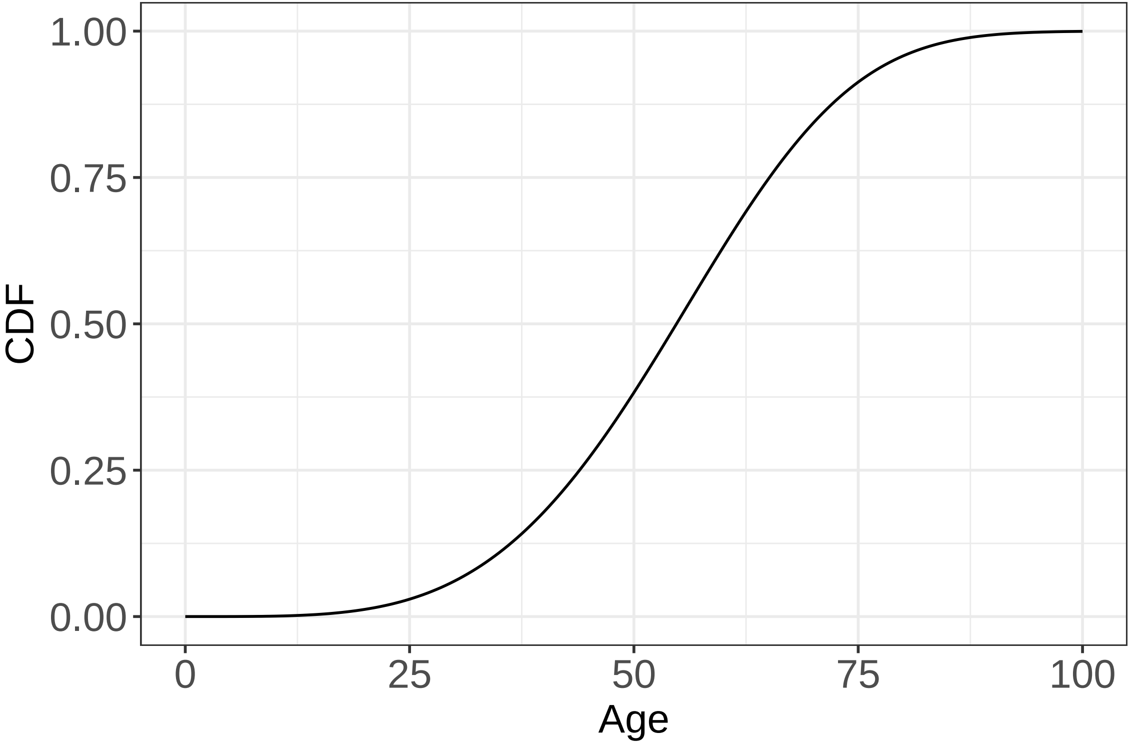
[17] Steven Tenny and Sameh W. Boktor. Incidence. In *StatPearls*. StatPearls Publishing, Treasure Island (FL), 2024.

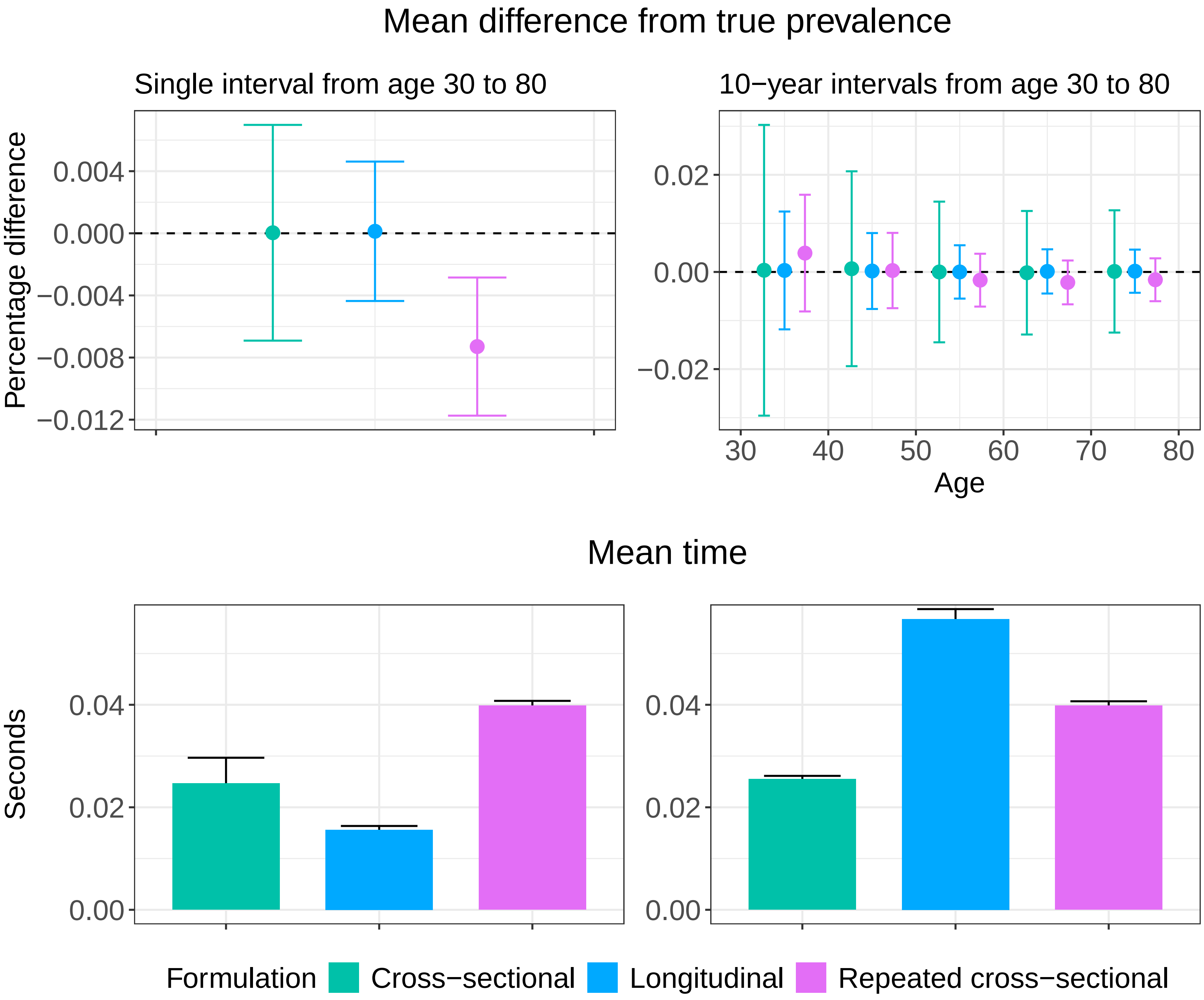
[18] Inger Johanne Bakken, Kari Tveito, Nina Gunnes, Sara Ghaderi, Camilla Stoltenberg, Lill Trogstad, Siri Eldevik H åberg, and Per Magnus. Two age peaks in the incidence of chronic fatigue syndrome/myalgic encephalomyelitis: a population-based registry study from Norway 2008-2012. *BMC Medicine*, 12(1):167, October 2014.

[19] Xiaofeng Wang and Zhenshun Cheng. Cross-Sectional Studies: Strengths, Weaknesses, and Recommendations. *CHEST*, 158(1):S65–S71, July 2020. Publisher: Elsevier.

[20] Surveillance, Epidemiology, and End Results (SEER) Program, National Cancer Institute. Rate Algorithms.

[21] Division of Cancer Control and Population Sciences, National Cancer Institute. Cancer Incidence Statistics.


1.00
0.75
0.50
0.25
0.00
CDF
0
25
50
75
100
Age



Mean difference from true prevalence
Single interval from age 30 to 80
10−year intervals from age 30 to 80
Percentage difference
0.004
0.000
−0.004
−0.008
−0.012
0.02
0.00
−0.02
30
40
50
60
70
80
Age
Mean time
Seconds
0.04
0.02
0.00
0.04
0.02
0.00
Formulation
Cross−sectional
Longitudinal
Repeated cross−sectional