Appendix 1 for manuscript titled: A two-step Markov processes approach for parameterization of cancer state-transition models for low- and middle- income countries

Technical Appendix

I. Technical proofs related to estimation of disease onset rates (θ_a) in Model 1 of main text

As discussed in the main text π_k are defined as the elements of the steady-state distribution vector $\boldsymbol{\pi}$ of the Markov process $\boldsymbol{Y}$ (Model 1), for state $k,\ k\epsilon\Omega$,

$$\pi_k = \sum_{j\epsilon\Omega,} \pi_j P_{jk} \ ; 0 \leq \pi_k \leq 1; \sum_{k\epsilon\Omega} \pi_k = 1 \qquad (1)$$

where, P_{jk} are the probabilities of transitioning from state j to state k, i.e., elements of the matrix $\mathbb{P}$ defined in the main paper. Our prime element of interest in this Markov process is $P_{H_aU_a}$, the risk or probability of developing the disease in age a, i.e., an element of $\mathbb{P}$ representing the probability of transitioning from H_a to U_a. Below we discuss its relation to the rates of disease onset θ_a and derivation of $P_{H_aU_a}$

Remark S1: Using the standard definition of risk to rate conversion that assumes that the underlying distributions governing transition probabilities are exponential, the rate of disease onset in age group a can be written as $\theta_a = -\ln(1 - P_{H_aU_a})$.

The rest of this subsection relates to derivation of an analytical expression and iterative process for estimation of $P_{H_aU_a}$

Assumption S1: We assume that the system is in steady state, i.e., the distribution **π** does not change over time.

Lemma S1*: Steady-state probabilities π_{D_a} can be expressed as a function of $P_{H_aU_a}$ as*

$$\pi_{D_a} - \pi_{D_{a-1}}\left(1 - P_{D_{a-1}M_{a-1}}\right) = \left[\pi_{H_a}P_{H_aU_a} + \pi_{U_{a-1}}P_{U_{a-1}U_a}\right]P_{U_aD_a} \quad (2)$$

Proof: We show the derivation when $a = 3$, which can be extended to other values. Using (1), $\pi_{D_3} = \pi_{D_2}\left(1 - P_{D_2M_2}\right) + \pi_{U_3}P_{U_3D_3}$. Expanding π_{U_3} again using (1)

$\pi_{D_3} = \pi_{D_2}\left(1 - P_{D_2M_2}\right) + [\pi_{H_3}P_{H_3U_3} + \pi_{U_2}P_{U_2U_3}]P_{U_3D_3}$, which after rearranging is equation (2) for $a = 3$. This completes the proof.

<u>Below, we show derivations of the different terms in (2) expressing</u> $\pi_{U_{a-1}}P_{U_{a-1}U_a}$ in (3), π_{H_a} in (4), $\pi_{D_a} - \pi_{D_{a-1}}\left(1 - P_{D_{a-1}M_{a-1}}\right)$ in (5) and <u>finally present our proposition for</u> $P_{H_aU_a}$ in (6)

Lemma S2: *For any given age a, the term $\pi_{U_{a-1}}P_{U_{a-1}U_a}$ in (2) can be expressed as*

$$\pi_{U_{a-1}}P_{U_{a-1}U_a} = \sum_{k=0}^{a-1}\left(\pi_{H_k}P_{H_kU_k}P(T \geq a-k)P(S \geq a-k),\right) \quad (3)$$

where T is the random variable denoting the time taken to transition to clinical disease from the time of disease onset (sojourn time), and S is the random variable denoting natural survival time from the age at disease onset (i.e., the length of time before a person dies from other causes, given that the person does not die from this disease). Thus, $\pi_{U_{a-1}}P_{U_{a-1}U_a}$ can be interpreted as the weighted probability of aging one additional year in preclinical cancer state (instead of transitioning to clinical disease or

mortality states), weighting by the chances that cancer onset occurred at age $(\forall k < a-1)$.

Proof: We show derivation for $a-1=3$ by repeatedly expanding terms $\pi_{U_{a-1}}$ using (1), starting with $\pi_{U_{a-1}}P_{U_{a-1}U_a}$

$$\pi_{U_3}P_{U_3U_4} = \left(\pi_{H_3}P_{H_3U_3} + \pi_{U_2}P_{U_2U_3}\right)P_{U_3U_4} = \left(\pi_{H_3}P_{H_3U_3} + \left(\pi_{H_2}P_{H_2U_2} + \pi_{U_1}P_{U_1U_2}\right)P_{U_2U_3}\right)P_{U_3U_4}$$

$$= \left(\pi_{H_3}P_{H_3U_3}P_{U_3U_4} + \left(\pi_{H_2}P_{H_2U_2}P_{U_2U_3}P_{U_3U_4} + \pi_{U_1}P_{U_1U_2}P_{U_2U_3}P_{U_3U_4}\right)\right)$$

$$= \pi_{H_3}P_{H_3U_3}P_{U_3U_4} + \pi_{H_2}P_{H_2U_2}P_{U_2U_3}P_{U_3U_4}$$

$$+ \left(\pi_{H_1}P_{H_1U_1} + \pi_{U_0}P_{U_0U_1}\right)P_{U_1U_2}P_{U_2U_3}P_{U_3U_4}$$

$$= \pi_{H_3}P_{H_3U_3}P_{U_3U_4} + \pi_{H_2}P_{H_2U_2}P_{U_2U_3}P_{U_3U_4} + \pi_{H_1}P_{H_1U_1}P_{U_1U_2}P_{U_2U_3}P_{U_3U_4}$$

(the above results from setting $\pi_{U_0}=0$, as 1 represents the youngest age for developing the disease)

$$= \pi_{H_3}P_{H_3U_3}P_{U_3U_4|H_3U_3} + \pi_{H_2}P_{H_2U_2}P_{U_2U_3|H_2U_2}P_{U_3U_4|H_2U_2}$$

$$+ \pi_{H_1}P_{H_1U_1}P_{U_1U_2|H_1U_1}P_{U_2U_3|H_1U_1}P_{U_3U_4|H_1U_1}$$

The above results from $P_{U_iU_{i+1}|H_jU_j} = P_{U_iU_{i+1}}$, for some $j \leq i$; the left-hand side interpreted as the probability of transitioning from current state U_i to state U_{i+1} (i.e., aging from i to $i+1$ in pre-clinical stage, say $U_iU_{i+1}$is event 1) given transition from healthy (H_j) to preclinical stage (U_j) occurred at some age-group $j \leq i$, (say H_jU_j is event 2); the equality results from the property of the Markov chain that transitions are dependent on only the current state, i.e., event 1 is independent of event 2.

Writing a general expression for any age a,

$$\pi_{U_{a-1}} P_{U_{a-1}U_a} = \sum_{k=0}^{a-1} \left(\pi_{H_k} P_{H_k U_k} \prod_{j=k}^{a-1} P_{U_j U_{j+1}|H_k U_k} \right)$$

We can replace,

$\prod_{j=k}^{a-1} P_{U_j U_{j+1}|H_k U_k} = \prod_{j=k}^{a-1} \big(P(T \geq\ j+1-k) P(S \geq j+1-k) \big) = P(T \geq\ a-k) P(S \geq a-k)$.

The first equality originates from the interpretation of $P_{U_j U_{j+1}|H_k U_k}$ as the probability of transitioning from U_j to U_{j+1}, i.e., transition to clinical disease is at an age greater than $j+1$, given $H_k U_k$, i.e. disease onset at age k, thus implying that, from the age of disease onset, the time to clinical diagnosis $T \geq\ j+1-k$ and survival (or no mortality from all other causes) $S \geq j+1-k$. The second equality originates from the fact that $P(T \geq\ j) P(T \geq\ j+1) = P(T \geq\ j+1)$

This completes the proof.

Assumption S2*: We set* $A_a = \pi_{H_a} + \pi_{U_a} \approx$ proportion of the population in age a - π_{D_a}, i.e., persons not living with a disease diagnosis (not in clinical state D_a) are either in healthy state H_a or in pre-clinical state U_a. We assume that estimates for A_a are available, here we obtain these data from the Global Burden of Disease study and population censuses.(2)

Remark S2*:* Using Assumption S2 and expanding $\pi_{U_a} = \pi_{H_a} P_{H_a U_a} + \pi_{U_{a-1}} P_{U_{a-1} U_a}$ using (1), we can write $A_a = \pi_{H_a} + \pi_{H_a} P_{H_a U_a} + \pi_{U_{a-1}} P_{U_{a-1} U_a}$ or

$$\pi_{H_a} = \frac{A_a - \pi_{U_{a-1}} P_{U_{a-1}U_a}}{1 + P_{H_a U_a}} \qquad (4)$$

Assumption S3: We assume that data for incidence estimates, defined as $I_{D_a} = \frac{Number\ of\ new\ cases\ of\ cancer\ diagnoses\ in\ age\ a}{Number\ of population\ in\ age\ a}$, are available from the Global Burden of Disease study, (2) and data for $c_a =$ proportion of the population in age a are available from population census.

Remark S3: By definition of π_{U_a}(i.e., steady-state probability) and $P_{U_a D_a}$(i.e., transition probability) we can interpret

$\pi_{U_a} P_{U_a D_a} = \frac{Number\ of\ persons\ in\ U_a}{Total\ population} * probability\ of\ transitioning\ from\ U_a\ to\ D_a$

$= \frac{Number\ of\ new\ cases\ of\ cancer\ diagnoses\ in\ age\ a}{Total\ population}$

$= \frac{Number\ of\ new\ cases\ of\ cancer\ diagnoses\ in\ age\ a}{Number\ of\ people\ in\ age\ a} * \frac{Number\ of\ people\ in\ age\ a}{Total\ population} = I_{D_a} c_a$, the last equality follows from definition in Assumption S3.

Remark S4: From (1), we can expand $\pi_{D_a} = \pi_{D_{a-1}}\left(1 - P_{D_{a-1}M_{a-1}}\right) + \pi_{U_a} P_{U_a D_a}$, or $\pi_{D_a} - \pi_{D_{a-1}}\left(1 - P_{D_{a-1}M_{a-1}}\right) = \pi_{U_a} P_{U_a D_a}$, thus

$$\pi_{D_a} - \pi_{D_{a-1}}\left(1 - P_{D_{a-1}M_{a-1}}\right) = I_{D_a} c_a, \qquad (5)$$

which follows from Remark S3.

Finally, we present a proposition for risk of disease onset

Proposition S1*: We can express the risk of disease onset at any specific age a as*

$$P_{H_aU_a} = \frac{I_{D_a}c_a - \sum_{k=0}^{a-1}\left(\pi_{H_k}P_{H_kU_k}\left[\sum_i s_i\left(1-e^{-(a-k)\lambda_i}\right) - \sum_i s_i\left(1-e^{-(a-1-k)\lambda_i}\right)\right]\left(\prod_{j=k:a+1} e^{-\mu_j}\right)\right)}{A_a\left[\sum_i s_i\left(1-e^{-\lambda_i}\right)\right]\left(e^{-\mu_a}\right) - I_{D_a}c_a} \quad (6)$$

***Proof**:*

Substituting (4) and (5) into (2) we get

$$I_{D_a}c_a = \left[\frac{A_a - \pi_{U_{a-1}}P_{U_{a-1}U_a}}{1+P_{H_aU_a}}P_{H_aU_a} + \pi_{U_{a-1}}P_{U_{a-1}U_a}\right]P_{U_aD_a}$$

The next few steps are algebraic operations on the above,

$$I_{D_a}c_a = \left[\frac{\left(A_a - \pi_{U_{a-1}}P_{U_{a-1}U_a}\right)P_{H_aU_a} + \left(1+P_{H_aU_a}\right)\left(\pi_{U_{a-1}}P_{U_{a-1}U_a}\right)}{1+P_{H_aU_a}}\right]P_{U_aD_a}$$

$$I_{D_a}c_a = \left[\frac{\left(A_aP_{H_aU_a} - \cancel{\pi_{U_{a-1}}P_{U_{a-1}U_a}P_{H_aU_a}}\right) + \left(\pi_{U_{a-1}}P_{U_{a-1}U_a}\right) + \cancel{\left(P_{H_aU_a}\right)\left(\pi_{U_{a-1}}P_{U_{a-1}U_a}\right)}}{1+P_{H_aU_a}}\right]P_{U_aD_a}$$

$$I_{D_a}c_a = \left[\frac{\left(A_aP_{H_aU_a}\right)P_{U_aD_a} + \left(\pi_{U_{a-1}}P_{U_{a-1}U_a}\right)P_{U_aD_a}}{1+P_{H_aU_a}}\right]$$

$$I_{D_a}c_a + I_{D_a}c_aP_{H_aU_a} = \left(A_aP_{H_aU_a}\right)P_{U_aD_a} + \left(\pi_{U_{a-1}}P_{U_{a-1}U_a}\right)P_{U_aD_a}$$

Rearranging, $P_{H_aU_a} = \frac{I_{D_a}c_a - \pi_{U_{a-1}}P_{U_{a-1}U_a}P_{U_aD_a}}{A_aP_{U_aD_a} - I_{D_a}c_a}$

From (3) $\pi_{U_{a-1}}P_{U_{a-1}U_a} = \sum_{k=0}^{a-1}\left(\pi_{H_k}P_{H_kU_k}P(T \geq a-k)P(S \geq a-k)\right)$, implies

$$P_{H_aU_a} = \frac{I_{D_a}c_a - \sum_{k=0}^{a-1}\left(\pi_{H_k}P_{H_kU_k}P(T \geq a-k)P(S \geq a-k)\right)P_{U_aD_a}}{A_aP_{U_aD_a} - I_{D_a}c_a}$$

$P_{H_a U_a}$

$$= \frac{I_{D_a} c_a - \sum_{k=0}^{a-1} \left(\pi_{H_k} P_{H_k U_k} P(T \geq a-k) P(S \geq a-k) P_{U_a D_a | H_k U_k} \right)}{A_a P_{U_a D_a | H_a U_a} - I_{D_a} c_a} \quad (7)$$

Expanding $P_{U_a D_a | H_k U_k} = P(T < a+1-k) P(S \geq a+1-k)$

$$P_{U_a D_a | H_a U_a} = P(T < 1) P(S \geq 1)$$

($P_{U_a D_a | H_k U_k}$ can be interpreted as the probability of transitioning from U_a to D_a, i.e., transition to clinical disease is at an age less than $a+1$, given $H_k U_k$, i.e. disease onset at age k)

we can write (7) as

$P_{H_a U_a}$

$$= \frac{I_{D_a} c_a - \sum_{k=0}^{a-1} \left(\pi_{H_k} P_{H_k U_k} P(a-k \leq T < a+1-k) P(S \geq a+1-k) \right)}{A_a P(T < 1) P(S > 1) - I_{D_a} c_a}$$

Further,

$$P(a-k \leq T < a+1-k) P(S > a+1-k)$$

$$= \left[\sum_i s_i \left(1 - e^{-(a-k)\lambda_i}\right) - \sum_i s_i \left(1 - e^{-(a-1-k)\lambda_i}\right) \right] \left(\prod_{j=k:a+1} e^{-\mu_j} \right)$$

$$P(T \leq 1) P(S > 1) = \sum_i s_i \left(1 - e^{-\lambda_i}\right) \left(e^{-\mu_a}\right)$$

The above 2 equations are from assuming $T \sim hyperexponential(\lambda_1, s_1, \ldots, \lambda_4, s_4)$, s_i is the probability that T will take the form of the exponential distribution with rate λ_i if diagnosis was at stage i and from conversion of annual mortality rate at age j (μ_j) into a

probability. We assume s_i is the probability of diagnosis at stage i, and $\frac{1}{\lambda_i} = \sum_{j=0}^{i} \frac{1}{p_j}$, where p_j is the natural disease progression rates from preclinical stage j to $j+1$, which we assume do not change by population and utilize values from the literature models. (see Section II below for details of the hyperexponential distribution for T). If p_j is a function of age a then we use $\frac{1}{\lambda_{i,a}} = \sum_{j=0}^{i} \frac{1}{p_{j,a}}$ ($p_{j,a}$ are also used in Model 2). Values for μ_a, which are mortality rates from all other causes at age a, are based on the demographics data for the country.

This completes the proof.

We summarize the algorithm for estimation of onset rates $\theta_a$in Table 2 of the main paper.

II. <u>Estimating parameters for the probability distribution of clinical disease transition time (sojourn time *T*)</u>

As described in the main paper, let T be the time to transition to clinical disease from the time of disease onset (sojourn time). We discuss below the estimation of the distribution of T and corresponding distribution parameters. For clarity of notations, we drop the subscript for age a in all notations. Let,

Y_i= dwell time in stage i, i.e., the duration of time in this stage in the absence of diagnosis before progression to stage $i+1$; and $\bar{y}_i$ are the average dwell times, i.e., $\bar{y}_i = E[Y_i] = \frac{1}{p_i}$; p_i are the natural progression rates from preclinical stage i to $i+1$, which we assume do not vary by population and are available in the literature from other

models (see individual cancer appendix) and Y_i are exponentially distributed (in the main text we use $p_{i,a}$ to indicate that p_i may be a function of age).

X_i= time spent in pre-clinical disease stage for a person diagnosed in stage i; and $\bar{x}_i = E[X_i] = \frac{1}{\lambda_i}$.

Then, we can approximate $\bar{x}_i = \sum_{j=0}^{i} \bar{y}_i$ and X_i to be exponentially distributed by assuming that $Y_0, Y_1, Y_2, ..$ are dependent variables. That is, suppose $\tilde{x}_{i_k}$ is some k^{th} sample from X_i, then $\tilde{x}_{i_k} = F_{Y_0}^{-1}(u_k) + F_{Y_1}^{-1}(u_k) + .. + F_{Y_i}^{-1}(u_k)$, where $F_{Y_j}(t)$ is the cumulative distribution function of Y_j, i.e., $\Pr(Y_j \leq t) = F_{Y_j}(t)$, such that $F_{Y_j}^{-1}(u_k)$ is a sample from Y_j corresponding to the same probability $u_k = Continuous\ Uniform\ [0,1]$ for all $j = 0, .., i$. This dependency between $Y_0, Y_1, Y_2, ..$ can be interpreted as follows. If disease progression in cancer stage 1 is faster for person A than person B, then it is likely that even in subsequent stages of cancer (prior to diagnosis) the progression is faster for A than for B.

Then, T will be a hyperexponential distribution with

$$E[T] = \sum_i s_i \bar{x}_i = \sum_i s_i \frac{1}{\lambda_i} = \sum_i \left(s_i \sum_{j=0}^{i} \bar{y}_j \right) = \sum_i \left(s_i \sum_{j=0}^{i} \frac{1}{p_j} \right)$$

and $\Pr(T \leq t) = \sum_i s_i F_{X_i}(t) = \sum_i s_i [1 - e^{-\lambda_i t}]$; $\frac{1}{\lambda_i} = \sum_{j=0}^{i} \frac{1}{p_j}$

where, s_i is the probability that T will follow the exponential distribution of the X_i^{th} variable and can be approximated as the proportion of cases diagnosed at stage i. If p_i

is a function of age we add a subscript a to the rate parameters to indicate this, i.e., $p_{i,a}, \lambda_{i,a}$.

III. Estimating mortality rates

We estimate the percent increase in mortalities due to cancer relative to region-specific cancer-free mortalities by age. With the assumption that the relative values are constant for all countries, we can then determine cancer mortalities for any country if the base mortalities are known. Specifically, we represent,

$$\bar{\mathrm{R}}_{i,a} = \frac{\bar{\mu}_{i,a} - \mu_a}{\mu_a} X\ 100\ ; \bar{\bar{R}}_{i,a} = \frac{\bar{\mu}_{i,a} r_i - \mu_a}{\mu_a}\ X\ 100; \bar{\bar{\mu}}_{i,a} = \bar{\mu}_{i,a} r_i \qquad (1)$$

where,

$\bar{\mathrm{R}}_{i,a}$ are the percent increase in cancer mortality, in stage i and age a, under the absence of treatment,

$\bar{\bar{R}}_{i,a}$ are the percent increase in cancer mortality on treatment for persons in age a who were diagnosed at stage i,

μ_a are the region-specific cancer-free mortality at age a, obtained from the demographical projections (DemProj) module in the Spectrum software,

r_i are the ratio of case fatality rates on treatment compared to not on treatment (Table 1),

$\bar{\mu}_{i,a}$ are the mortality rates in cancer stage i and age a when not on treatment, and

$\bar{\bar{\mu}}_{i,a}$ are the mortality rates in cancer stage i and age a on treatment.

We estimate $\mathrm{R}_{i,a}$ as below. Let $s_{i,a}$ be the ratio of the proportion of people with and without disease surviving past time t, i.e.,

$s_{i,a} = \frac{1-\left(1-e^{-\bar{\mu}_{i,a}\,t}\right)}{1-(1-e^{-\mu_a t})} = \frac{\left(e^{-\bar{\mu}_{i,a}\,t}\right)}{(e^{-\mu_a t})}$, then (2)

$-\ln\left(s_{i,a}\right) = (\bar{\mu}_{i,a} - \mu_a)\, t$, and

from (1), $\bar{\mathrm{R}}_{i,a} = \frac{\bar{\mu}_{i,a}-\mu_a}{\mu_a} = \frac{-\ln(s_{i,a})}{\mu_a t}$

We estimate $s_{i,a}$ by simulating two cohorts of people for $t = 10$ years, one cohort transitioning through the preclinical stages, and death (i.e., under assumption of no treatment) and the other cohort, who represent a cancer free population, transitioning to death under the cancer-free rates of mortality. For example, to estimate the relative risk for stage local, cohort 1, consisting of a 1000 people in stage local and age a at year zero, were simulated for 10 years by transitioning them through age and, using progression rates ($p_{i,a}$) (data in individual Cancer Appendix), through stages local, regional, distant, and death. Cohort 2, consisting of a 1000 people of age a, were simulated for 10 years by transitioning them through age and generating deaths at rate μ_a. Then $s_{i,a}$ = the proportion who survived in cohort 1 at the end of 10 years divided by the proportion who survived in cohort 2 at the end of the 10 years. This was repeated for all age-groups.

IV. Sensitivity analyses of steady-state assumptions for the Markov process models:

In estimation of the onset rates, the diagnostic rates, and the distribution of the population in disease states we make 2 main assumptions that theoretically guarantee *unique* solution for the estimates. The first assumption is that the Markov chain is stationary, i.e., the values in the transition probability matrix and generator matrix in Models 1 and 2 of the main paper do not change over time. The interpretation of this is that the disease has not evolved over time, e.g., due to changes in environmental factors, population diet, or access to care, and thus, the risk of disease, natural progression of disease, and chances of diagnosis due to symptoms have remained constant over time. In the implementation of the model to a country we continue to make this assumption (except for the case of screening that increases diagnosis), which is similar to most models in the literature as there are not enough data to estimate the evolution of the disease over time. The second assumption is that births are set equal to deaths to generate a *regular* Markov chain such that there exists a unique steady-state distribution. The interpretation of this is that, irrespective of the distribution we assume in the beginning of the model, the final distribution after long-run simulation will be the same. This feature can be exploited to estimate the unknown distribution and parameters.

The disease concept related to the first assumption is beyond the scope of our study, and there are insufficient data on these changes. We test the sensitivity of the second assumption, which is an assumption in Model 2 of the main paper. Specifically, for some randomly chosen value of diagnostic rates, we compare outcomes (in Figure S1) from using 3 different assumptions for births: births=deaths (used in Model 2 for generating the steady-state model), constant births, and constant birth rate. Outcomes

include age-normalized distribution of persons in clinical and preclinical stages, i.e., prevalence by age-group (prevalence in healthy stage not shown)(including health stage, prevalence adds to 1 in each age-group) (see Remark 1 in main paper). The last two assumptions for births modify the population size over time, i.e., the system is not in steady-state, as expected (Figure S1). The population size at the end of the simulation relative to the population in year zero was about 2.5 times with the assumption of constant births and exploded to about 1.1 million times with the assumption of constant birth rate (results not shown). However, even with these extremities, the age-normalized stage distributions were the same in all three models (Figure S1, A and B), though the distribution of the overall population by age group were different (Figure S1, C). This indicates that, because of the approach of normalizing within each age-group, the outputs are not sensitive to steady state assumptions.

V. Conditions for global optimality of diagnostic rates in Model 2 of the main paper:

If an objective function is convex, a local optimal solution is the optimal solution (global) to the problem. Therefore, we test for convexity of (1). As we do not know the analytical form of the function we cannot theoretically test for convexity conditions. Therefore, we ran the optimization search algorithm multiple (30) times with different initial points (arbitrary initial values for d_a) in each run. The optimal value obtained under all runs were the same indicating that it is likely to be a global optimal. As these are only empirical results, to generalize that this will hold over the entire function domain, we augment the empirical results with partial theoretical analyses to test for convexity as follows.

We first state some assumptions and, by showing $Minimize_{d_{i,a}} \|\bar{\boldsymbol{I}} - \boldsymbol{I}\|_2$, $d_{i,a} \geq 0$ (objective function in (1)) is approximately separable by age-group a, present a proposition to solve this sequentially for each age-group. This 'approximation' method transforms the problem to multiple sub-problems whose analytical forms can be empirically determined through function fitting and tested for convexity.

Assumption S4: The chance of being symptomatic increases as cancer advances, therefore, $d_{0,a} < d_{1,a} < d_{2,a} < d_{3,a} < d_{4,a}$ for any specific age $a = 1, \dots, m$. We assume that the chance or risk of developing symptoms in any stage $\mathrm{i} = \{0,1,2,3,4\}$ relative to the last stage, say $\mathrm{i} = 4$, *(*that we will refer to as *relative risk of symptoms)* does not vary by age, i.e.,

$\frac{d_{i,1}}{d_{4,1}} = \frac{d_{i,2}}{d_{4,2}} = \cdots = \frac{d_{i,m}}{d_{4,m}}$ for $a = \{1, \dots, m\}$ and for any stage i

Assumption S5: We assume that the relative risk of symptoms can be approximated as $\frac{d_{i,a}}{d_{4,a}} = \sum_{j=0}^{i} s_j$, where, s_i are the proportion of cases diagnosed at stage i and $\sum_{i=0}^{4} s_i = 1$, i.e., we can write $\frac{d_{0,a}}{d_{4,a}} = s_0$; $\frac{d_{1,a}}{d_{4,a}} = s_0 + s_1$; $\frac{d_{2,a}}{d_{4,a}} = s_0 + s_1 + s_2$; $\frac{d_{3,a}}{d_{4,a}} = s_0 + s_1 + s_2 + s_3$ or equivalently, setting $d_{4,a} = d_a$, write, $d_{0,a} = d_a s_0$; $d_{1,a} = d_a(s_0 + s_1)$; $d_{2,a} = d_a(s_0 + s_1 + s_2)$; $d_{3,a} = d_a(s_0 + s_1 + s_2 + s_4)$.

Proposition S2: *The problem in (9)* $Minimize_{d_{i,a}} \|\bar{\boldsymbol{I}} - \boldsymbol{I}\|_2$, $d_{i,a} \geq 0$ *can be approximated as* m *number of separable sub-problems*

minimize $(I_a - \bar{I}_a)^2$ *; s.t.* $d_a \geq 0$

that can be solved for $a = 1, \dots, m$*, sequentially, starting with* $a = 1$

Proof:

From assumptions S4 and S5, as values of s_i are known, we can drop the subscripts i in $d_{i,a}$ and the optimization model in (9) reduces to

minimize $(I_{a=1} - \bar{I}_{a=1})^2 + (I_{a=2} - \bar{I}_{a=2})^2 + \cdots + (I_{a=m} - \bar{I}_{a=m})^2$; s.t. $d_a \geq 0, \forall a$;

$\bar{I}_a$, i.e., clinical incidence of cancers, are a function of d_a. Specifically, as persons do not transition from age a to $a-1$, $\bar{I}_a = f(d_a|d_1, \ldots., d_{a-1})$, i.e., $\bar{I}_1 = f(d_1)$, $\bar{I}_2 = f(d_2|d_1)$, and so on. Further, as shown above in sensitivity analyses using prevalence as example, due to normalization of results (including $\bar{I}_a$) within age-group, results are not sensitive to changes in steady-state distribution across age-groups. So the results in older age-groups, which are not yet fit have little impact on the results in younger age-groups, e.g., when estimating d_1 to Minimize $(I_{a=1} - \bar{I}_{a=1})^2$, estimates of $(I_{a>1} - \bar{I}_{a>1})^2$ have no impact. Therefore, by sequentially solving for d_a starting with $a = 1$, the objective function in (9) is separable. We also empirically tested that the solution from separately solving for each a, i.e., $Minimize_{d_a} (I_a - \bar{I}_a)^2$ are identical to solving the full model $Minimize_{d_a} \|\bar{\boldsymbol{I}} - \boldsymbol{I}\|_2$, $a = 1, .., m$;

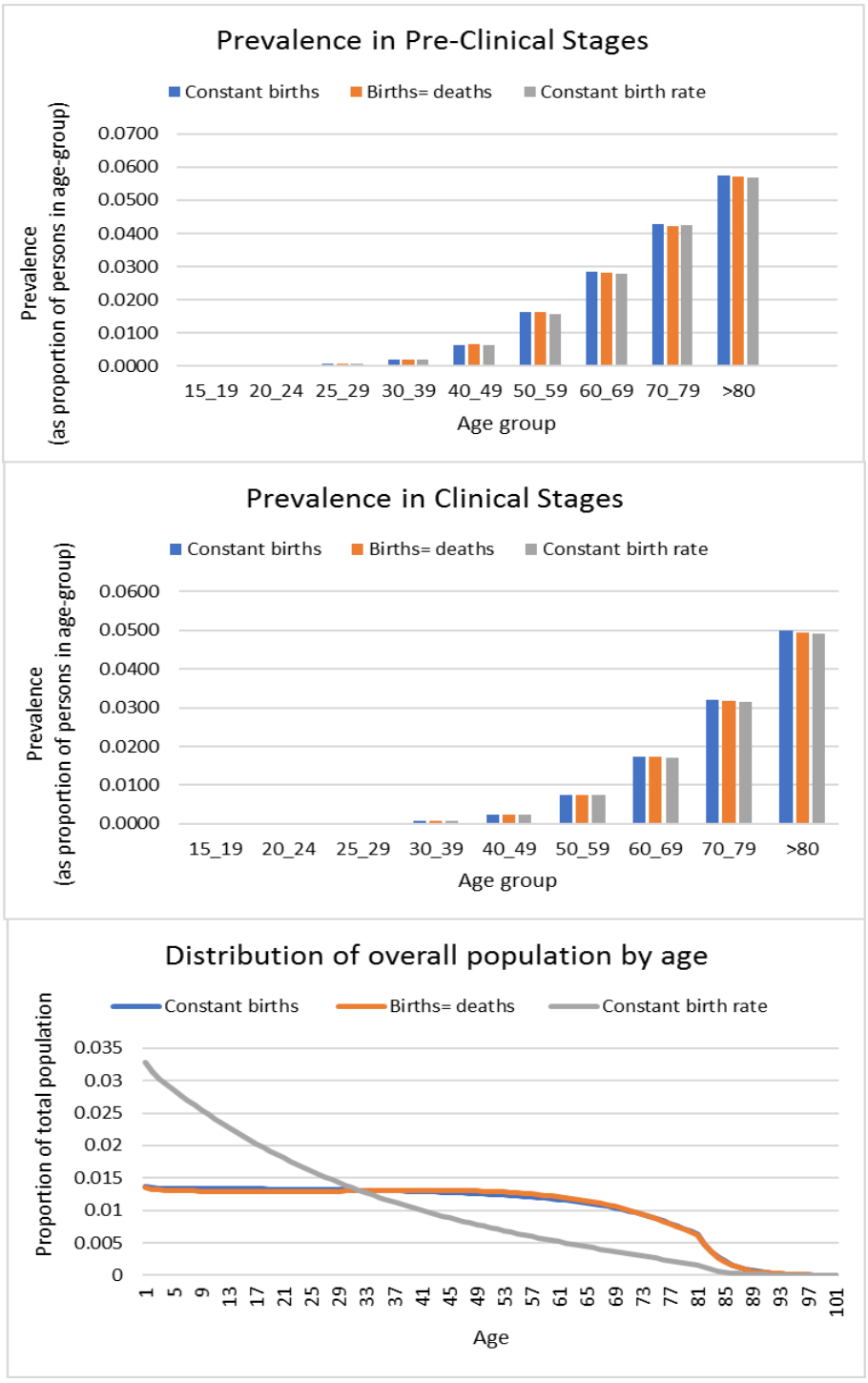
Testing for convexity of objective functions in above sub-problems by function fitting to empirical data

If $f(d_a) = (I_a - \bar{I}_a)^2$ is positive semi-definite, i.e., the second derivative $f''(d_a) \geq 0$ at all points of $d_a \in \{0,1\}$, then $(I_a - \bar{I}_a)^2$ is a convex function on d_a. This will then guarantee global solutions to d_a, if we were to sequentially solve for d_a starting with $a = 1$, However, we do not know the analytical form of I_a to calculate $f''(d_a)$.

However, the method in Proposition S2 simplifies the problem into multiple sub-problems with simpler objective functions whose analytical forms can be approximated

through empirical function fitting. We demonstrate this taking breast cancer in Southeast Asia region as an example (see Figure S2). As the diagnostic rate increases, cancer incidence increases, therefore, $\bar{I}_a$, for a specific age a, is a non-decreasing function on $d_a \in [0,1]$, with incidence =0 when $\mathrm{d_a} = 0$ and reaching a maximum when $\mathrm{d_a} = 1$, and I_a is a constant. Further, empirical data provided a good fit for a model of the form $\bar{I}_a \sim ln(d_a)$. Then, writing $x = d_a$ $f''(x) = \frac{d^2}{dx^2}(ln(x) - \bar{I}_a)^2 = \frac{2(\bar{I}_a - ln(x)+1)}{x^2} > 0$, thus proving that $f(d_a)$ is convex. When applying the model to individual countries, we expect to test these conditions again as data used in fitting analytical function for $\bar{I}_a$ will vary.

Figure S1: Keeping diagnostic rates fixed*, prevalence in pre-clinical stages, prevalence in clinical stages, and the objective function values (Equation 9) are all similar under three different assumptions for births, though the overall population distributions may be different; indicating that error generated from not using actual birth rates are minimal.



* We chose arbitrary diagnostic rates to empirically test the sensitivity of assumptions made. Sample shown here is for breast cancer.

Figure S2: Logarithmic function is a good fit for incidence as function of diagnostic rates