

# Modeling and Calibration for Exposure to Time-Varying, Modifiable Risk Factors: The Example of Smoking Behavior in India

Jeremy D. Goldhaber-Fiebert, PhD, Margaret L. Brandeau, PhD

**Background.** Risk factors increase the incidence and severity of chronic disease. To examine future trends and develop policies addressing chronic diseases, it is important to capture the relationship between exposure and disease development, which is challenging given limited data. **Objective.** To develop parsimonious risk factor models embeddable in chronic disease models, which are useful when longitudinal data are unavailable. **Design.** The model structures encode relevant features of risk factors (e.g., time-varying, modifiable) and can be embedded in chronic disease models. Calibration captures time-varying exposures for the risk factor models using available cross-sectional data. We illustrate feasibility with the policy-relevant example of smoking in India. **Methods.** The model is calibrated to the prevalence of male smoking in 12 Indian regions estimated from the 2009–2010 Indian Global Adult Tobacco Survey. Nelder-Mead searches (250,000 starting locations) identify distributions of starting, quitting, and restarting rates that minimize the difference between modeled and observed age-specific prevalence. We compare

modeled life expectancies to estimates in the absence of time-varying risk exposures and consider gains from hypothetical smoking cessation programs delivered for 1 to 30 years. **Results.** Calibration achieves concordance between modeled and observed outcomes. Probabilities of starting to smoke rise and fall with age, while quitting and restarting probabilities fall with age. Accounting for time-varying smoking exposures is important, as not doing so produces smaller estimates of life expectancy losses. Estimated impacts of smoking cessation programs delivered for different periods depend on the fact that people who have been induced to abstain from smoking longer are less likely to restart. **Conclusions.** The approach described is feasible for important risk factors for numerous chronic diseases. Incorporating exposure-change rates can improve modeled estimates of chronic disease outcomes and of the long-term effects of interventions targeting risk factors. **Key words:** model calibration methods; risk factors; chronic disease; time-varying risks; smoking; India. (*Med Decis Making* 2015;35:196–210)

Changes in the prevalence and intensity of risk factors drive the increasing burden of many chronic diseases, prompting calls to implement mitigation policies focused on risk factors.<sup>1</sup> To understand the likely future trends of chronic diseases and to develop policies that efficiently reduce their burden, it is essential to understand the relationship between risk factor exposure and the development of chronic disease. However, because chronic diseases may not develop until decades after risk exposure, empirical studies of the risk factor–disease relationship and the effects of risk factor interventions may be infeasible. In lieu of such studies, simulation models can be used.

Two methodological challenges arise when developing such models. First, the model must incorporate the relevant disease process as well as exposures to risk factors and resulting effects while remaining computationally tractable and externally interpretable. Second, modeling both a disease and its risks increases data requirements. This can be particularly challenging in developing countries, where risk factor exposure levels are particularly important given rapidly increasing chronic disease rates but where only limited data are available.<sup>2–4</sup>

The risk factors that are most relevant to the global burden of chronic diseases (e.g., smoking and obesity)<sup>5</sup> share specific characteristics that must be incorporated into the model. Namely, they are time varying, are modifiable, and convey risks for incidence and severity that change based on exposure duration. Past work on modeling the prevalence and

effects of risk factors includes closed-form epidemiological estimates based on population-attributable fractions and improvement in outcomes estimates based on risk factor prevalence at baseline.<sup>6,7</sup> Other estimation methods account for movement between risk categories over a lifetime. Direct estimation from longitudinal data of transition rates between risk factor categories is possible with the use of statistical competing risk techniques and time-varying covariate techniques, but such approaches require population-representative longitudinal data.<sup>8–10</sup> Some researchers have developed simulation models that rely on age-specific prevalence to estimate “net transition rates” but with no change in risk categories for individuals when the overall prevalence in risk categories remains stable across ages.<sup>11,12</sup> Such an approach is limited when mortality rates and rates of transition between risk factor categories are affected by an individual’s duration of exposure.

The goal of our study was to develop methods for modeling chronic diseases and their risk factors. We developed model structures that encode the relevant features of risk factors (e.g., time-varying), while being simple enough to embed within a complex

chronic disease model, and employed calibration procedures to capture time-varying exposures that use commonly available, cross-sectional survey data. Our risk factor models allow for the possibility that mortality and transition rates between risk categories can depend on the duration of exposure to a risk factor. We illustrate the feasibility and utility of this approach with the policy-relevant example of a model of smoking behavior among men in India and hypothetical smoking cessation interventions.

## METHODS

### Risk Factor Model Structure

Our risk factor model of smoking among men in India derives from a family of simplified risk factor models (Table 1). Because the goal is to illustrate risk factor models that can be embedded in a chronic disease model and the calibration of such risk factor models, we employ a highly stylized “chronic disease” Markov model (Figure 1A), consisting of 2 states (alive and dead) with an average age-specific mortality rate ( $\mu(a)$  for individuals of age  $a$ ) moving portions of the cohort from alive to dead in each cycle. In practice, the alive state will be divided into more states (e.g., no chronic disease, mild chronic disease, severe chronic disease). In this first model, individuals are not distinguished by risk exposure.

The most important feature for risk factor modeling is whether the risk factor is fixed or time-varying. Risk factors such as sex are represented with fixed strata that remain constant for an individual from birth through death and have a simplified influence via an increased chronic disease risk and consequent increased mortality ( $\mu_{\text{high-risk}}(a) > \mu_{\text{low-risk}}(a)$ ) (Figure 1B). While we illustrate a model with 2 strata, more strata are easily incorporated. When individuals can change their risk status (i.e., becoming lower or higher risk), a time-varying risk factor structure is required (Figure 1C). The model structure can extend to incorporate the fact that the likelihood of changing one’s risk status and the likelihood of disease incidence or death can vary with the duration of high risk exposure (Figure 1D). Likewise, the model can accommodate the case in which people who have been previously high risk but have now lowered their risk are unlike those who have always been low risk because of the impact of past cumulative exposures or because other unrepresented factors that correlate with the modeled risk factors differ between

Received 2 April 2013 from Stanford Health Policy, Centers for Health Policy and Primary Care and Outcomes Research, Stanford University, Stanford, CA, USA (JDG-F); and Department of Management Science and Engineering, Stanford University, Stanford, CA, USA (MLB). An earlier version of this work was presented as an oral abstract at the 2012 Society for Medical Decision Making (SMDM) Annual Meeting in Phoenix, Arizona. The funding for this work comes from the US National Institutes of Health (NIH). The funders had no role in the design or conduct of the research, the results found and conclusions drawn, or the decision to publish. Financial support for this study was provided by the NIH’s National Institute on Aging (K01 AG037593; principal investigator [PI]: Goldhaber-Fiebert) and by Stanford’s Freeman Spogli Institute’s Underdevelopment Action Fund (PI: Goldhaber-Fiebert). Margaret Brandeau was supported by grant number 1-R01-DA15612 from the National Institute on Drug Abuse. The funding agreements ensured the authors’ independence in designing the study, interpreting the data, and writing and publishing the report. The authors gratefully acknowledge multiple formative conversations with Professor Jay Bhattacharya. They also thank the 2012 SMDM Annual Meeting attendees who raised a number of important comments and questions that further developed the abstract into the work presented in this article. Revision accepted for publication 20 November 2013.

Supplementary material for this article is available on the *Medical Decision Making* website at <http://mdm.sagepub.com/supplemental>.

Address correspondence to Jeremy D. Goldhaber-Fiebert, PhD, Stanford Health Policy, Centers for Health Policy and Primary Care and Outcomes Research, Stanford University, 117 Encina Commons, Stanford, CA 94305-6019, USA; telephone: 650-721-2486; fax: 650-723-1919; e-mail: [jeremygf@stanford.edu](mailto:jeremygf@stanford.edu).

**Table 1** Taxonomy of Risk Factor Model Types for Chronic Diseases

Risk Factor Model Description	Risk Factor Examples
Time-invariant risk strata (Figure 1B)	<ul style="list-style-type: none"> <li>• Sex (male, female)</li> <li>• Genotype</li> </ul>
Time-varying risk strata	
Not influenced by duration of status or prior exposures (Figure 1C)	<ul style="list-style-type: none"> <li>• Age and cancer development</li> </ul>
Influenced by duration of status but not prior exposures (Figure 1D)	<ul style="list-style-type: none"> <li>• Obesity and chronic diseases</li> </ul>
Influenced by duration of status and prior exposures (Figure 1E)	<ul style="list-style-type: none"> <li>• Smoking behavior and cancer development</li> <li>• Pollution exposure and cancer development</li> <li>• Development of severe dengue fever in which an initial episode of dengue alters the risk/severity of subsequent exposures</li> </ul>

“previously high risk” and “never high risk” and do not change when people become “previously high risk” (e.g., alcohol consumption patterns among previous v. never smokers) (Figure 1E).

Chronic disease and mortality risks are often related to exposures to more than 1 risk factor. For example, the risk of cardiovascular disease is related to smoking, diet, and physical activity. While modeling a second risk factor is possible using double stratification analogous to Figure 1C and 1D, it is also possible to capture multiple risk factors in a unidimensional scale of overall risk.<sup>13–15</sup> However, complexities such as dependence between changes in risk factor exposures (e.g., an individual who quits smoking may also increase his physical activity) may argue for a microsimulation in such a case. Nonetheless, our model structures are embeddable in both the Markov cohort and microsimulations and are also useful for earlier steps in iterative model development commonly undertaken as part of developing detailed microsimulations that may incorporate multiple risk factors.

### Data Needs and Calibration Methods for Simplified Risk Factor Models

Models with fixed risk strata (e.g., such as that in Figure 1B) only require data on prevalence, potentially stratified by age and sex. Models with time-varying risk factors (e.g., such as those in Figure 1C–E) require transition rates between risk factor categories that increase the data needs beyond the prevalence for these rates to be identifiable. Additionally, it is important to include differential mortality effects due to risk factor exposure that arise both via an increased chronic disease risk and severity and via other channels of increased mortality (e.g., in the

context of a smoking and diabetes model, the effect of smoking on diabetes as well as an elevated cancer risk).

When direct estimation of transition rates between risk factor categories is not feasible due to the limited availability of longitudinal data, calibration methods applied to cross-sectional data such as household surveys provide a useful indirect approach to estimating transition rates between risk factor categories.<sup>16</sup> For the risk factor models in Figure 1C–E, 2 or more sets of transition rates are needed to characterize movement between risk levels. Assuming that these rates depend on age, this implies that the number of calibration targets required to reach identifiability is larger than age-specific prevalence. For example, for the model in Figure 1C, data on age-specific prevalence and the age-specific distribution of the duration of risk factor exposure would generally be sufficient to instantiate the model, provided one is willing to impose a parametric form on how duration alters rates of change in an individual’s current risk level. For example, the probability of quitting smoking in the next period ( $P_{quit}$ ) for someone of a given age could be assumed to decrease with the duration of smoking as follows:

$$P_{quit}(age, duration) = \frac{1}{duration} P_{quit}(age, duration = 0).$$

Such relationships might be estimated from limited longitudinal data or from other contexts in which datasets are more readily available.

We now describe the process of calibrating our risk factor model. Model calibration is the process of systematically varying model uncertainties or unknown inputs until model outputs are consistent with observed data. Its application has become increasingly common in health.<sup>17–20</sup> Current efforts focus

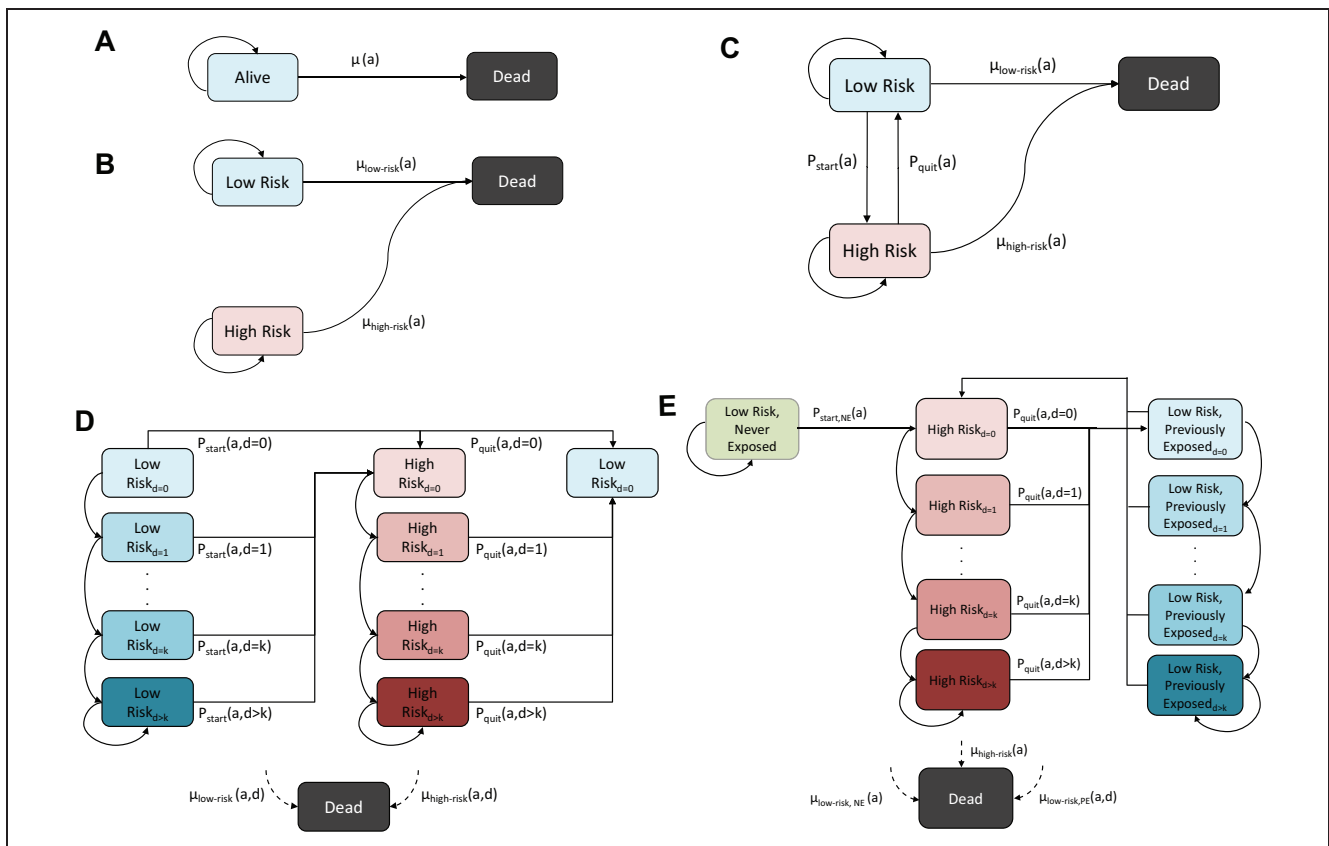


Figure 1 Simplified Markov model structures incorporating different types of risk factors: (A) No risk factors; (B) fixed risk strata; (C) time-varying risk strata, not depending on the duration of status or prior exposure; (D) time-varying risk strata depending on the duration of status; and (E) time-varying risk strata depending on the duration of status and prior exposure. In all panels, transitions can occur between health states (rounded rectangles) along directed black edges at each time step. Transitions can depend on age ( $a$ ) and duration ( $d$ ) of exposure, depending on the model. The proportion of the cohort that does not transition in a given cycle remains in their current health state.

on incorporating calibration into best-practice guidelines,<sup>21</sup> increasing the use of calibration through tutorials that make its methods accessible, and propagating good reporting standards to ensure its rigor and reproducibility.<sup>21–23</sup> There are still many open questions about how best to apply calibration in various modeling situations such as the case that we consider in this article: time-varying risk factors that affect chronic disease outcomes. In general, model calibration involves a standard set of steps,<sup>24</sup> which we apply to our time-varying risk factor model.

**Calibration target estimation.** The risk factor model must be calibrated to a set of target values. We use the age-specific prevalence levels of a risk factor (e.g., smoking), estimated from cross-sectional data with logistic regressions (or, for more than 2 categories, multinomial logistic regressions), predicting age-specific point estimates and bootstrapping age-specific confidence intervals. To estimate

levels of a risk factor across age groups  $j$  with various covariates, we minimize the following:

$$\text{Logit}(Y_i) = B_0 + \sum_j (B_j \alpha_j) + BX + \text{Error}_i,$$

where  $Y_i$  is the outcome of interest (e.g., being a current smoker) for individual  $i$ , and  $\alpha_j$  is an indicator variable for age group  $j$ ,  $B$  is a vector of coefficients, and  $X$  is a vector of other covariates (e.g., geographic region). The equation could also include interactions between age and the vector of covariates. By estimating the regression across age categories and other covariates (e.g., urban/rural location or region), we construct targets with information borrowed across groups, which can then be propagated to inform our calibrated parameters for each group.

Using cross-sectional data collected at a single time point to infer the age-specific patterns of a cohort requires the assumption that there are no strong

secular time trends that affect risk factor exposure (e.g., decreasing smoking prevalence in subsequent birth cohorts). If such an assumption is not realistic in a particular context, one can construct age period-specific patterns of prevalence from diagonal birth cohorts as observed in multiple cross-sectional surveys conducted at different time points (e.g., estimation based on prevalence among 20- to 25-year-old people in 2000, prevalence among 25- to 30-year-old people in 2005, etc.), which is an approach that increases data requirements for successful calibration.

*Objective function.* Ideally, the objective function for the calibration process should minimize the distance between a set of model outputs and the calibration targets in a way that is unitless and incorporates the relative levels of uncertainty in the various calibration targets. For example, the objective function can be defined as a linear combination of the difference between each model output  $i$  ( $M_i$ ), conditional on a given set of inputs and the point estimate of each target ( $T_i$ ), scaled by the standard error of the estimate of each target ( $se_i$ ), and potentially weighted for output  $i$  ( $w_i$ ) based on an additional importance consideration:

$$\text{Objective}(M(\text{inputs}), T) = \sum_{i \in T} \left[ w_i \left( \frac{M_i - T_i}{se_i} \right)^2 \right].$$

For applications in which targets all are in the same units, targets are roughly equally uncertain, and the analyst believes that they are equally important, one might forego  $w_i$  and  $se_i$ , setting both implicitly to 1. When feasible, defining the objective function in terms of the underlying likelihood of the data or a function proportional to the likelihood may have important advantages in terms of linkage to statistical theory, efficiency, and consistency of estimates of parameter uncertainty.

*Search constraints.* Constraints on the values of model inputs are often relevant. In the case of calibrating a discrete-time model's probabilities, the values must fall in the range of [0,1]. While we typically start searches using a uniformly distributed prior over this range, some commonly used search optimization procedures such as Nelder-Mead (described below) do not support search constraints and may explore infeasible points (i.e., those <0 or >1). Instead of implementing the constraints directly in the algorithm, which may be complex and prone to bugs or may yield suboptimal results due to premature collapse of the Nelder-Mead search simplex, we

often use a large penalty term in the objective function if the constraints are violated that grows even larger for larger violations of the constraints (i.e., large penalty for probability = 1.1 and even larger penalty for probability = 2.0), acknowledging that care must be taken depending on the search optimization procedure used to overcome difficulties with convergence.<sup>25–28</sup>

*Search procedure for optimization.* Many well-studied procedures exist for searching for optimum solutions.<sup>26</sup> We often use Nelder-Mead because it is an efficient, directed search technique.<sup>28</sup> Other calibration techniques that exploit the specific features of the problem have been described recently.<sup>29</sup> We start the Nelder-Mead search from many different initial simplexes to reduce the risk of local optima traps. This approach is important because the models that we consider are not necessarily globally concave as a function of their parameters, and thus the objective function is not necessarily concave.

*Identifying the best-fitting input parameter combinations.* Because the search is conducted in a hyperdimensional space, it is not simple to determine whether the points identified as good fitting by multiple searches all correspond to a single optimum even if they have the same goodness-of-fit value as computed by our objective function. Therefore, we treat separately all good-fitting points identified by our multiple calibration searches and use them to define the range of uncertainty in model-predicted outcomes, consistent with the uncertainty in the data to which we calibrated. Specifically, after conducting our calibration, we run the model multiple times (once per good-fitting set of points) to generate a range of predicted outcomes. Although this can be computationally intensive, it protects against averaging across distinct optima and thereby using input values that correspond to no optimum at all.<sup>17–20</sup>

## The Example of Smoking in India

*Risk factor model.* We illustrate this approach using the example of smoking among men in India. The goal is to determine sets of starting, quitting, and restarting smoking rates for various Indian subpopulations of men that can be embedded in models of chronic diseases including tuberculosis,<sup>30</sup> cancer, and other diseases related to smoking. The structure of the model is shown in Figure 2 and is similar to that shown in Figure 1E, except that here, men

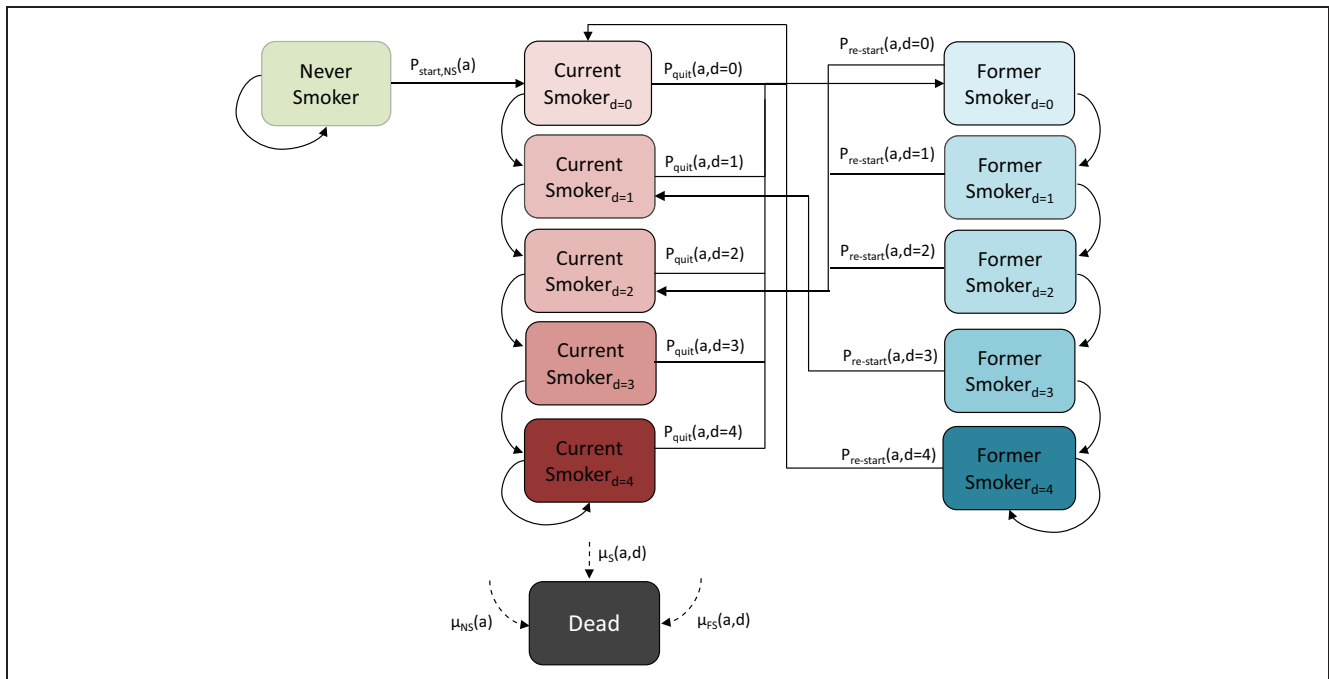


Figure 2 Model of smoking risk behavior in India, illustrated for age group  $a$ . The model schematic represents 3 general categories of smoking-related risk: those who have never smoked, those who are current smokers, and those who previously smoked. These risk categories and transitions between them are stratified by age ( $a$ ) and duration of status ( $d$ ), as appropriate. This enables risks of mortality as well as risks of starting, quitting, and restarting smoking to differ across these characteristics.

who return to high risk (i.e., restart smoking) are distributed among the different categories of high risk as distinguished by duration (details below). Duration is categorized from less than 1 year ( $d = 0$ ) to more than 4 years ( $d = 4$ ).

Given the diversity of age-specific smoking prevalence patterns across the more than 500 million men in India, we separately calibrated 12 models of smoking that share the common structure shown in Figure 2 but represent different population groups defined by urban/rural status and 6 Indian geographic regions (Central, East, North, Northeast, South, West).

**Parameters to be calibrated.** The model simulates men aged  $a = 10, \dots, 90$  years. We calibrated the following 3 parameters:  $p_{start}(a)$ ,  $p_{quit}(a, d = 0)$ , and  $p_{re-start}(a, d = 0)$ . These are, respectively, the probability that a person of age  $a$  who has never smoked begins smoking, the probability that a person of age  $a$  who has smoked for less than 1 year quits smoking, and the probability that a former smoker of age  $a$  who now has not smoked for less than 1 year restarts smoking. The analogous quitting and restarting parameters for longer durations ( $d = 1, 2, 3$ , and  $\geq 4$  years) were not calibrated but were calculated from the calibrated quantities as follows:

$$p_{quit}(a, d) = (\alpha_q)^d p_{quit}(a, d = 0), d = 1, 2, 3, 4 \text{ and}$$

$$p_{re-start}(a, d) = \alpha_{rs}(d) p_{re-start}(a, d = 0), d = 1, 2, 3, 4.$$

We assumed that the probability of restarting smoking for former smokers declines as a function of duration in that risk state and that likewise the probability of quitting smoking for current smokers declines as a function of duration in that state. We estimated the search range for  $p_{start}(a)$  (which accounts for the fact that older individuals who have never smoked are less likely to start smoking and therefore ranges from 0 to a value of  $<1$ ) and the values of the decay parameters  $\alpha_q$  and  $\alpha_{rs}(d)$  ( $q$  referring to quit and  $rs$  referring to restart) from the Trivandrum Oral Cancer Study (TOCS), which is a large longitudinal study that repeatedly measured smoking status in a geographically defined cohort in southern rural India ( $n = 27,243$  for men with 3 measures per individual) (see further details in the Appendix).<sup>31</sup> While we used longitudinal data to directly inform  $\alpha_q$  and  $\alpha_{rs}(d)$ , we note that in the absence of the TOCS data, we could have calibrated these parameters as well. We opted to use the TOCS data primarily for efficiency

reasons but also to illustrate a calibration situation in which some longitudinal data are available.

Evidence suggests that former smokers who have successfully quit for a longer period are more successful in quitting after restarting smoking than former smokers who have quit for shorter periods of time.<sup>32–34</sup> To capture this effect without unduly complicating the model, we made the following assumption. Instead of having all former smokers who restart smoking return to the category of “current smoker with duration  $d = 0$ ,” we assumed that previous smokers who restarted joined “current smoker” duration groups in inverse relationship to their duration of being a previous smoker. Thus, individuals who had been previous smokers for  $\geq 4$  years returned to being current smokers with a duration of 0 to 1 years, whereas those who had been previous smokers for 3 to 4 years returned to being current smokers with 1 to 2 years’ duration. All individuals who had quit less than 3 years previously returned to being current smokers with 3 to 4 years’ duration. The simplifying assumption necessary to allow for differential rates of restarting smoking based on the duration of cessation also results in some smokers of short duration who quickly restart being exposed to mortality and subsequent quitting probabilities in excess of their actual smoking duration.

*Calibration targets.* The calibration procedure used 3 targets for each age group  $a = 10, \dots, 90$ : overall prevalence of smokers among men of age  $a$ , which we denote by  $p_S^{TARGET}(a)$ ; prevalence of long-term smokers ( $\geq 4$  years) among men of age  $a$ , which we denote by  $p_{LT}^{TARGET}(a)$ ; and prevalence of former smokers among men of age  $a$ , which we denote by  $p_{FS}^{TARGET}(a)$ . Because the reported age and number of individuals over the age of 75 years are more imprecise, we report results for targets for individuals up to age 75 years.

We used data from the Indian Global Adult Tobacco Survey (GATS) (2009–2010) to define calibration targets (never, current, and past smokers).<sup>35</sup> We confined our analysis of the GATS to male respondents whose reported age was 15 to 80 years, and we assumed, given difficulties in recalling the exact age for Indians born earlier in the century, that 71 to 80 years represented individuals older than 80 years as well ( $n = 33,413$  men; approximately 200–500 individuals per subgroup defined by age, urban/rural status, and region). We categorized individuals as current long-term smokers (men reporting current smoking of  $\geq 4$  years’ duration), current short-term smokers (men reporting current smoking with a duration  $< 4$  years), former smokers, and never

smokers. We also classified individuals by the Indian region of residence and whether they lived in a rural or urban area (see the GATS questionnaire).<sup>36</sup>

To capture the age-specific proportions of each subpopulation in each smoking category, we used a multinomial logistic regression with an individual’s smoking category as the outcome variable. This allowed us to ensure that the predicted probabilities of being in each category would sum to 100%. Predictors used in the model were indicators for age category (15–20, 21–30, 31–40, 41–50, 51–60, 61–70, 71–80 years), indicators for geographic region, indicators for urban/rural residence, interactions between age and region indicators, and interactions between region and urban indicators. The regression appropriately incorporated survey sample weights of individuals that reflected the survey design. For each combination of these indicators, we predicted the proportion of the male population falling into each group. For example, we predicted the proportion of the male population who were never smokers and who were 31 to 40 years of age and lived in the northern region in urban areas. Likewise, we predicted the proportion of this group who were long-term current smokers and then repeated this for all groups and smoking categories.

Because these estimates represented means for age groups with generally 10-year widths, but our calibration required targets for each year of age ( $a = 10, \dots, 90$ ), we used a piecewise cubic Hermite interpolation of the points predicted from our multinomial logistic regressions (Stata program PCHIPOLATE<sup>37</sup>) in which we assumed that smoking prevalence and former smoking prevalence were both zero at age 10 years and that the means predicted by our regression represented a value near the midpoint of the age group (e.g., age 17 years for the 15- to 20-year category). Interpolations were performed separately for each geographic and urban/rural group to produce separate sets of age-specific calibration targets for each group. Calibration targets were extrapolated beyond the age range used in the GATS analyses for several reasons. For younger children, smoking prevalence is known to be very low, and the experience of smoking between the ages of 10 to 20 years determines the distribution of durations of current and former smokers. For older individuals, exact ages in the GATS were likely imprecise as seen in other household surveys, and the life expectancy contribution of time lived above age 80 years is increasing. Because we use the model to project life expectancy for smoking interventions, we want to ensure that it performs reasonably in the absence of interventions even at older ages.

Our calibration to cross-sectional, age-specific smoking prevalence reflects the limited availability of data and the assumption of no strong birth cohort differences in the risks of starting, quitting, or restarting smoking. Limited data from other cross-sectional surveys conducted in different calendar years including the National Family and Health Surveys,<sup>38</sup> World Health Survey,<sup>39</sup> and District-Level Health Surveys<sup>35</sup> provide some support for this assumption (Appendix).

*Model inputs (mortality rates).* To define mortality-related model inputs, we directly computed age-, urban/rural-, region-, and smoking status-specific mortality rates for men from 2 primary sources. The Indian Sample Registration System<sup>40</sup> provides population-representative life tables and associated mortality rates stratified by these characteristics with the exception of smoking; we denote these mortality rates by  $\mu_{avg}(a)$ . The Indian Million Deaths Study<sup>41</sup> contains sex- and age-specific relative risks of death due to smoking; we denote this by  $RR_s(a)$ . We computed death rates for relevant groups of nonsmokers,  $\mu_{NS}(a)$ , in each of the 12 regions using the following equation:

$$p_S^{TARGET}(a)\mu_{NS}(a)RR_s(a) + p_{NS}^{TARGET}(a)\mu_{NS}(a) = \mu_{avg}(a).$$

In the above equation,  $p_{NS}^{TARGET}(a)$  denotes the prevalence of nonsmokers among men of age  $a$  ( $p_{NS}^{TARGET}(a) = 1 - p_S^{TARGET}(a)$ ); these prevalences were those that we estimated from the GATS. The product  $\mu_{NS}(a)RR_s(a)$  is the death rate for smokers of age  $a$ ,  $\mu_S(a)$ . Mortality calculations use point estimates from these sources to compute rates entered directly into the model.

Mortality risks from current and former smoking depend on the duration of current smoking, cumulative lifetime exposure to smoking, and time since stopping smoking.<sup>42–46</sup> To reflect the attenuation of increased mortality risks as individuals who quit remain nonsmokers, we assumed that the mortality rate for former smokers,  $\mu_{FS}(a)$ , is between those of smokers and nonsmokers and is nearly as high (95% smoker rate + 5% nonsmoker rate) as that of current smokers for the first year after smoking and attenuates to nearly as low (5% smoker rate + 95% nonsmoker rate) as for nonsmokers after having been a former smoker for more than 4 years, consistent with evidence that after about 10 years, the mortality risks of nonsmokers and those who have smoked far back in the past are largely similar.<sup>42,45,46</sup> Specifically, we estimate death rates for each

duration category of nonsmoking  $d$  (1, 2, 3, or  $\geq 4$  years) as the following:

$$\mu_{FS}(a, 1) = .95\mu_S(a) + .05\mu_{NS}(a),$$

$$\mu_{FS}(a, 2) = .65\mu_S(a) + .35\mu_{NS}(a),$$

$$\mu_{FS}(a, 3) = .35\mu_S(a) + .65\mu_{NS}(a), \text{ and}$$

$$\mu_{FS}(a, 4) = .05\mu_S(a) + .95\mu_{NS}(a).$$

*Calibration procedure.* For each of the 12 regional models, we generated random starting simplexes of values for the 3 sets of probabilities to be calibrated ( $p_{start}(a)$ ,  $p_{quit}(a, d = 0)$ , and  $p_{restart}(a, d = 0)$ , for age  $a = 10, \dots, 90$ ). To do so, we selected random values of these 3 parameters for ages  $a = 10, 15, 18, 20, 23, 25, 30, 35, 45, 55, 65, 75, 85$ , and 100. Within each parameter, we linearly interpolated between the values at these ages to develop estimates for the full range of ages  $a = 10, 11, 12, \dots, 90$ . We assumed that at time zero (i.e.,  $a = 10$ ), everyone in the population is a nonsmoker. We projected the model forward for 80 years to populate the various risk groups (smoker, nonsmoker, former smoker for each age group  $a = 10, \dots, 90$ , and for each risk duration  $d = 1, \dots, 4$  for smokers and former smokers). We then compared the model's calculated smoking prevalence values,  $p_S^{MODEL}(a)$ ,  $p_{LT}^{MODEL}(a)$ , and  $p_{FS}^{MODEL}(a)$ , to the target values that we estimated:  $p_S^{TARGET}(a)$ ,  $p_{LT}^{TARGET}(a)$ , and  $p_{FS}^{TARGET}(a)$ . We evaluated each parameter set using the following objective function:

$$J = \sum_a \left[ (p_S^{TARGET}(a) - p_S^{MODEL}(a))^2 + (p_{LT}^{TARGET}(a) - p_{LT}^{MODEL}(a))^2 + (p_{FS}^{TARGET}(a) - p_{FS}^{MODEL}(a))^2 \right].$$

Because all calibration targets were expressed in the same units and had reasonably similar uncertainty ranges in the age ranges used, we chose to use an unweighted least squares approach.

We applied the Nelder-Mead search algorithm to each simplex, selecting new values as directed by the search algorithm after each iteration. We continued the procedure until the objective function changed by less than 0.01% or after 40,000 iterations, whichever came first.

For each subpopulation, we repeated the calibration procedure with 250,000 different random

starting simplexes of values for the 3 probabilities. We then selected the 100 parameter sets with the best fits (minimum objective function value) and used these for model projections to reflect the mean and uncertainty in these projections consistent with the calibration data, weighting the model-projected outcomes by the inverse of the objective function value to emphasize fits from the best of the 100 best-fitting sets.<sup>17,47</sup> Additionally, for the 100 best-fitting parameter sets, we compared the amount of error for each of the 3 parameter values and each age group (difference between model-projected v. target value). We did this to make sure that errors were approximately equal, indicating that a good fit was obtained for all 3 targets and not just for the aggregated objective function  $J$ .

We implemented the models in C++<sup>48</sup> using the GNU Scientific Library implementation of Nelder-Mead searches (*gsl\_multimin\_fminimizer\_nmsimplex2*)<sup>49</sup> and the Mersenne Twister pseudorandom number generator.<sup>50</sup>

*Model projections.* Predictions made with the calibrated models included life expectancy lost due to smoking in each Indian geographic subgroup relative to an otherwise similar population of lifetime nonsmokers and differences in predicted life expectancy lost due to smoking in fixed risk factor strata models versus the time-varying models that we calibrated to the data. We then used the models to consider the effect on life expectancy of hypothetical smoking cessation interventions that caused smokers to quit and prevented former smokers from starting for either 1 year, 15 years, or 30 years. Model projections were made with each of the 100 best-fitting calibrated parameter sets to reflect uncertainty given the empirical data.

## RESULTS

Calibration yielded close matches between modeled and observed smoking prevalence for men in all regions for all parameter sets. Figure 3 shows modeled and observed current smoking prevalence of men for all 12 regions (Figure 3A), prevalence of long-term smoking (Figure 3B), and prevalence of former smokers (Figure 3C). We observe that the confidence intervals of the empirical estimates of smoking prevalence are wider for older men than for younger men because fewer individuals survive to later ages. Even so, the calibrated model's output matches point estimates for all ages quite closely and falls within their confidence bounds

for all of the model's 100 best-fitting parameter sets.

Calibration updates the range of input parameters used from the initial distributions searched, consistent with the empirical data and its uncertainty. Figure 4 shows the range of values for the calibrated input parameters across the 100 best-fitting sets. Notably, starting rates for those who have never smoked before (Figure 4A) rise with earlier ages and are generally more precisely calibrated prior to age 30 years when the pool of former smokers who may otherwise restart is relatively small compared to older age groups; the starting rates apply to a much larger pool of people and are lower than the rates of quitting and restarting (Figure 4B and 4C), which are conditional on being in the subgroups of current or former smokers. Likewise, because we allow for both quitting and restarting smoking, shifts between risk groups can compensate for one another. Thus, it is particularly difficult to disentangle estimates of age-specific quitting and restarting rates from empirical prevalence data. Hence, the calibrated ranges for these parameters are wider relative to those for the age-specific starting rates. Even so, calibration allows us to find reasonable estimates for these parameters across the entire range of age groups that are consistent with the empirical data (Figure 3).

To evaluate the improvement in life expectancy estimates and resulting mortality projections due to our time-varying risk factor model, we compared life expectancy at age 30 years for rural men under 4 different scenarios: 1) all men are nonsmokers; 2) all men are smokers; 3) men aged 30 years are divided into smokers and never smokers, and these risk strata remain constant for their remaining lifetimes (a fixed risk factor model, similar to that in Figure 1B); and 4) men can start smoking, stop smoking, and restart smoking (our time-varying risk factor model). Prevalences of current smokers and former smokers at age 30 years are derived from our calibration targets for the fixed risk strata model. For the time-varying model, these prevalences are the same as they have been calibrated to be. To ensure that the distribution of smoking and former smoking durations is the same, we started the model with a cohort of age 10 years and then computed the remaining life expectancy for those alive at age 30 years. Mortality and mortality relative risks from smoking are consistent in all scenarios. The remaining life expectancy for rural men who had never been smokers through age 30 years was 39.22 years (North), 36.36 years (Central), 37.02 years (East), 33.95 years (Northeast), 36.83 years (West), and 36.77 years (South).

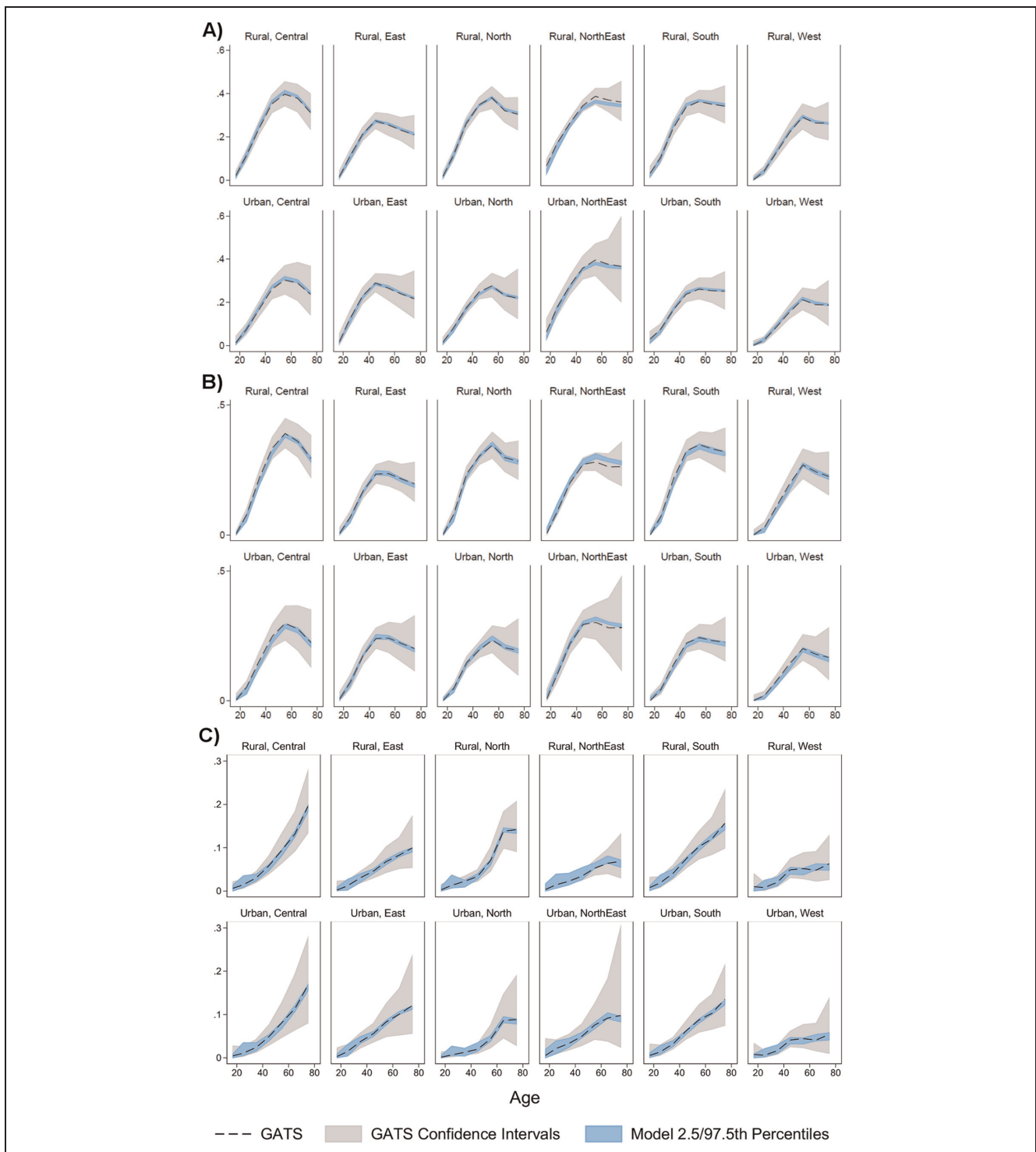


Figure 3 Modeled and observed smoking prevalence among men in all regions of India. (A) Current smoking prevalence. (B) Long-term smoking prevalence ( $\geq 4$  years). (C) Prevalence of former smokers. By geographic region and urbanicity, each panel shows the comparison of the calibrated modeled outputs of age-specific prevalence (blue region is the uncertainty region for model predictions) and prevalences and 95% confidence intervals (dashed black lines and gray regions are both) estimated from the 2009–2010 Indian Global Adult Tobacco Survey (GATS).

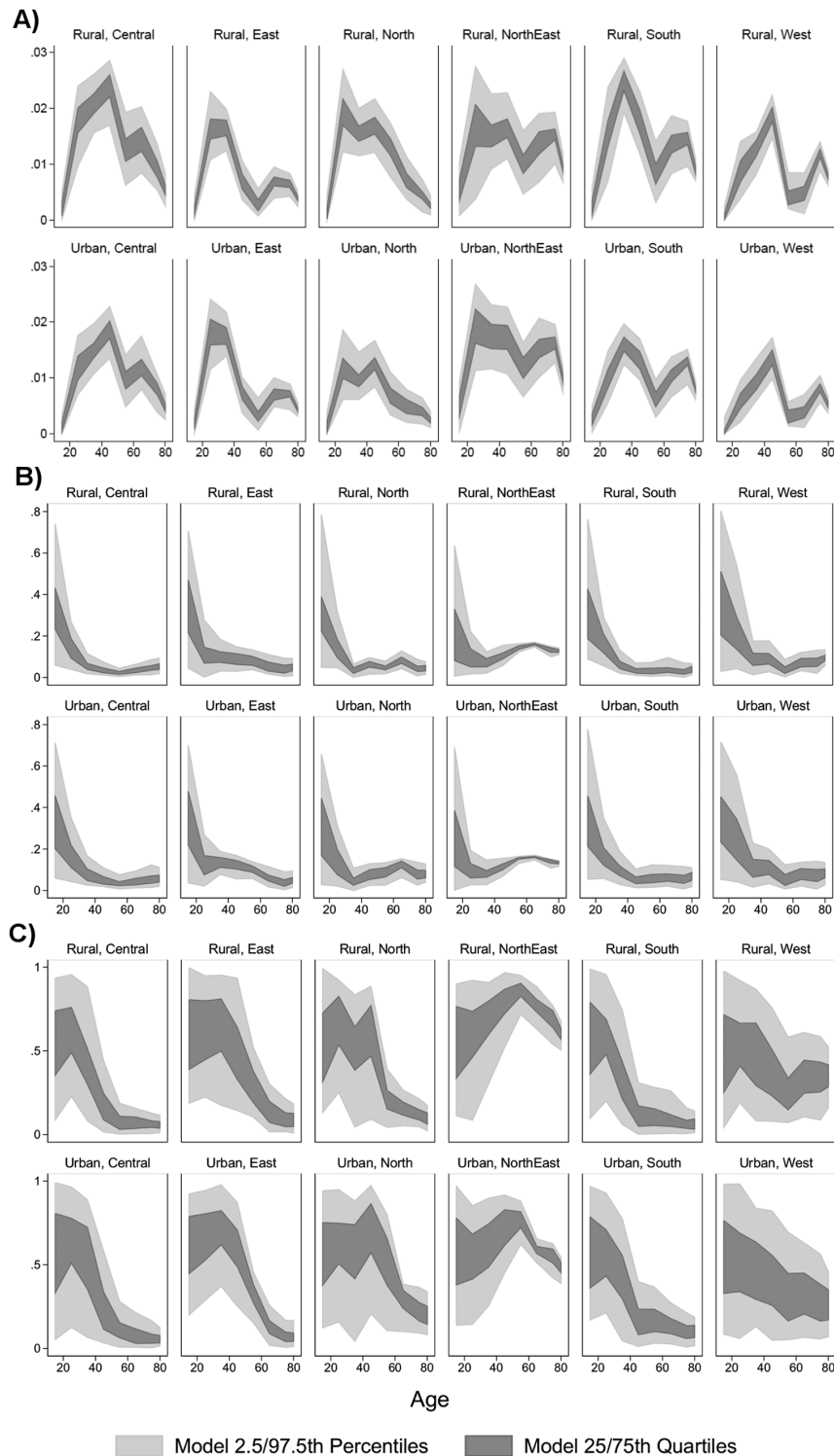


Figure 4 Range of values for the calibrated input parameters across the 100 best-fitting sets found through calibration for men in all regions of India. (A) Values of  $p_{\text{start}}(a)$ . (B) Values of  $p_{\text{quit}}(a, d = 0)$ . (C) Values of  $p_{\text{restart}}(a, d = 0)$ . Dark gray areas represent the 25th through 75th percentiles of values, and light gray areas represent the 2.5th through 97.5th percentiles of values.

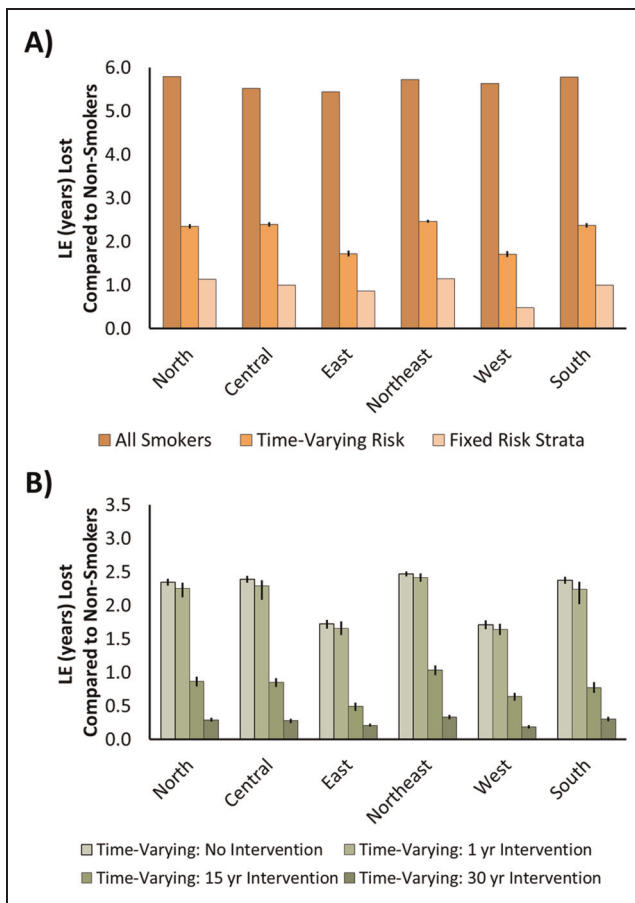


Figure 5 Life expectancy losses from smoking and gains from a hypothetical smoking cessation intervention for populations of rural men aged 30 years in regions of India relative to otherwise similar populations of never smokers. (A) Life expectancy losses for 3 groups: 1) smokers (all men are smokers and remain smokers for life); 2) fixed risk strata (men aged 30 years are divided into smokers and never smokers, and these risk strata remain constant for their lifetimes); and 3) time-varying risk strata (men can start smoking, stop smoking, and restart smoking). Each scenario is shown in a different color bar from left to right for each geographic region. Vertical bars show the uncertainty in life expectancy estimates for our calibrated model (scenario 3) due to the uncertainty in calibrated values of starting, quitting, and restarting smoking. (B) Life expectancy losses as computed using our time-varying risk strata model relative to an otherwise similar population of never smokers under a hypothetical smoking cessation intervention that also prevents restarting smoking for 1, 15, or 30 years. Colored bars show the different intervention program durations, and vertical bars show the uncertainty in life expectancy estimates for our calibrated model due to the uncertainty in calibrated values of starting, quitting, and restarting smoking.

Figure 5A shows life expectancy losses from smoking relative to these estimates of life expectancy for nonsmokers. The largest losses, of nearly 6 years,

occur for a population of smokers who remain smokers for the remainder of their lives. When some men smoke and others do not, life expectancy losses will fall between that of the lifetime smoker group and zero. The use of a fixed risk factor model reduces the estimated life expectancy by about 0.5 to 1.1 years in the different regions (i.e., the loss attributable to smoking assuming fixed risk strata). The use of our time-varying risk factor model further reduces the estimated life expectancy by about 0.9 to 1.4 years (i.e., the loss attributable to smoking assuming that individuals can switch between risk strata). Our time-varying risk factor model estimates a lower life expectancy than the fixed risk factor model because our model accounts for the possibility that some men who do not currently smoke may begin smoking (or restart smoking), thus increasing their mortality, and some men who do smoke may quit smoking, but when they do, they will have a higher mortality rate than men who never smoked, at least for a while. By accounting for these possibilities, the time-varying risk factor model generates life expectancy estimates that have greater face validity given underlying real-world processes than does a fixed risk factor model.

Models that provide improved estimates of life expectancy for time-varying risk factors can improve estimates of the impacts of public health programs and help decision makers form more informed decisions about which programs to invest in. We illustrate this by considering a highly stylized smoking cessation and prevention program targeted to rural men that causes all smokers to quit at age 30 years and prevents recidivism for 1, 15, or 30 years. Compared to no such program, individuals receiving the intervention gain 0.05 to 0.2 years of life expectancy if recidivism occurs in 1 year and 1.5 to 2.0 years of life expectancy if the program prevents recidivism for 30 years (Figure 5B). The total benefit is due both to the prevented years of smoking and corresponding reduced mortality risks; additionally, even after the program terminates, the longer time spent as nonsmokers reduces the subsequent risk of restarting. Our model accurately reflects this additional benefit and shows the nonlinearity of benefits in life expectancy as a function of how long smokers are induced to quit. Thus, for example, the benefit of the 15-year program is nearly that of the 30-year program, showing that even if an intervention's direct effects attenuate over time, it can still have appreciable indirect benefits.

We performed sensitivity analyses to examine how uncertainty about the relative risk of death from smoking may influence our results (Appendix

Figures 2 and 3). The Million Deaths Study provides estimates of age- and sex-specific relative risks of death from smoking used in our model. Those estimates are based on a very large sample, which is reported with 99% confidence intervals that are fairly narrow and consistent with many other large longitudinal studies performed worldwide. Without recalibrating the model, we used relative risks from the high and low ends of the 99% confidence intervals and reran the model to examine 1) how our model fits to the calibration targets changed and 2) how our model predictions compared to a fixed risk strata prediction of life expectancy loss due to smoking. We find that the calibration fits remained within the confidence intervals of our targets and were largely indistinguishable from our main analysis. The only noticeable differences, albeit within the confidence intervals, were for individuals aged 65 years and over, in whom higher relative risks caused our model to underestimate the smoking prevalence and lower relative risks caused it to overestimate prevalence (Appendix Figure 2). While the level of life expectancy loss due to smoking depended on higher versus lower assumed relative risks, our model consistently predicted greater losses due to smoking than a fixed risk strata model (Appendix Figure 3).

## DISCUSSION

Calibrating rates of change in exposures to risk factors using widely available, population-level, cross-sectional data is a feasible and accurate approach to modeling risk factor-dependent mortality in a population. For the example of smoking in India, calibration provided good estimates of the probabilities that men start, quit, and restart smoking across all ages. Finding good values for these parameters would not be feasible via direct empirical analysis, given the lack of longitudinal data. Moreover, the use of a model with time-varying risk factors allowed us to obtain more realistic estimates of life expectancy given underlying real-world processes than would be obtained from a model with fixed risk strata. Such accuracy is important not only for forecasting the chronic disease burden but also for evaluating the potential effects of risk factor mitigation policies (e.g., programs to prevent smoking initiation in youth or smoking cessation programs).

Our analysis has several limitations. Our proposed risk factor models employ discrete risk categories (e.g., current smoker for 1 year, 2 years, etc.), but risk is often continuous (e.g., number of pounds

overweight or amount of exposure to environmental pollutants). When using discrete categories to represent continuous risk, a balance is required between realism (enough risk categories) versus complexity (too many risk categories). We did not model mortality risks as a function of total duration of exposure (e.g., total of 10 lifetime years of smoking) nor of level of exposure (e.g., cigarettes smoked per day). In principle, it would be possible to construct a categorization that captures both the duration of the current episode of smoking and the duration of total exposure and potentially the amount of exposure in the current episode and over all episodes. Such a categorization could improve the accuracy of risk prediction, but at the expense of complexity. Our models could be extended to capture such effects with additional categories but would eventually become intractable, especially with the goal of identifying starting, quitting, and restarting probabilities and embedding the risk factor model within a larger chronic disease model.

We did not model social network effects that may affect risks. Changes in an individual's risk factor exposure may depend on the levels of risk factor exposure in the population (e.g., an individual's chance of being obese may depend on the social acceptability and prevalence of obesity among friends or the entire population).<sup>51–53</sup> In this case, one might model risk factor exposure with transmission dynamics.<sup>53–56</sup> Such a model would be more complex than those we have considered, and would require significantly more data, but could be useful when network effects are important determinants of risk.

Our calibration method uses a multitarget ordinary least squares approach. Our objective function  $J$  could be modified to incorporate the likelihood of the model, providing true values given the empirical data under assumptions about the true process that generates the data.<sup>57</sup> Additionally, our calibration method may not optimally exploit the age-dependent structure of the problem because it assumes the independence of values of the same target at different ages. Thus, our updating of parameters based on calibration may be less statistically efficient or consistent than if the actual underlying likelihood function were specified.

Incorporating the effects of risk factors into chronic disease models is essential for accurately projecting the burden of disease and estimating the potential effects of mitigation policies. Our simplified approach to modeling time-varying risk exposures, along with calibration techniques to estimate key

model parameters from commonly available cross-sectional data, is a useful means of generating accurate estimates of morbidity and mortality in a population due to risk factors.

## REFERENCES

1. World Health Organization. 2008-2013 action plan for the global strategy for the prevention and control of noncommunicable diseases. 2009. Available from: [http://whqlibdoc.who.int/publications/2009/9789241597418\\_eng.pdf](http://whqlibdoc.who.int/publications/2009/9789241597418_eng.pdf)
2. Pradeepa R, Prabhakaran D, Mohan V. Emerging economies and diabetes and cardiovascular disease. *Diabetes Technol Ther*. 2012; 14 Suppl 1:S59–67.
3. Li Y, Zhang M, Jiang Y, Wu F. Co-variations and clustering of chronic disease behavioral risk factors in China: China Chronic Disease and Risk Factor Surveillance, 2007. *PLoS One*. 2012;7(3): e33881.
4. Abegunde DO, Mathers CD, Adam T, Ortegon M, Strong K. The burden and costs of chronic diseases in low-income and middle-income countries. *Lancet*. 2007;370(9603):1929–38.
5. Lim S, Vos T, Flaxman A, et al. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990-2010: a systematic analysis for the Global Burden of Disease Study 2010. *Lancet*. 2012; 381(9867):2224–60.
6. Bruzzi P, Green SB, Byar DP, Brinton LA, Schairer C. Estimating the population attributable risk for multiple risk factors using case-control data. *Am J Epidemiol*. 1985;122(5):904–14.
7. Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. *Am J Public Health*. 1998;88(1):15–9.
8. Prentice RL, Kalbfleisch JD, Peterson AV Jr., Flournoy N, Farewell VT, Breslow NE. The analysis of failure times in the presence of competing risks. *Biometrics*. 1978;34(5):541–55.
9. Lau B, Cole SR, Gange SJ. Competing risk regression models for epidemiologic data. *Am J Epidemiol*. 2009;170(2):244–56.
10. Huang ES, Basu A, O'Grady M, Capretta JC. Projecting the future diabetes population size and related costs for the U.S. *Diabetes Care*. 2009;32(12):2225–9.
11. Boshuizen HC, Lhachimi SK, van Baal PH, et al. The DYNA-MO-HIA model: an efficient implementation of a risk factor/chronic disease Markov model for use in Health Impact Assessment (HIA). *Demography*. 2012;49(4):1259–83.
12. Kasstele JV, Hoogenveen RT, Engelfriet PM, Baal PH, Boshuizen HC. Estimating net transition probabilities from cross-sectional data with application to risk factors in chronic disease modeling. *Stat Med*. 2012;31(6):533–43.
13. Hoerger TJ, Segel JE, Zhang P, Sorensen SW. Validation of the CDC-RTI Diabetes Cost-Effectiveness Model. Research Triangle (NC): RTI Press; 2009.
14. National Cholesterol Education Program. Risk assessment tool for estimating your 10-year risk of having a heart attack. 2013. Available from: <http://hp2010.nhlbi.nih.net/atpiii/calculator.asp>
15. Diabetes Trials Unit, Oxford Centre for Diabetes EaM. UKPDS risk engine. 2012. Available from: <http://www.dtu.ox.ac.uk/riskengine/>
16. Taylor DC, Pawar V, Kruzikas D, et al. Calibrating longitudinal models to cross-sectional data: the effect of temporal changes in health practices. *Value Health*. 2011;14(5):700–4.
17. Goldhaber-Fiebert JD, Stout NK, Ortendahl J, Kuntz KM, Goldie SJ, Salomon JA. Modeling human papillomavirus and cervical cancer in the United States for analyses of screening and vaccination. *Popul Health Metr*. 2007;5:11.
18. Fryback DG, Stout NK, Rosenberg MA, Trentham-Dietz A, Kurchittham V, Remington PL. The Wisconsin Breast Cancer Epidemiology Simulation Model. *J Natl Cancer Inst Monogr*. 2006;(36): 37–47.
19. Malagón T, Joumier V, Boily MC, Van de Velde N, Drolet M, Brisson M. The impact of differential uptake of HPV vaccine by sexual risks on health inequalities: a model-based analysis. *Vaccine*. 2013;31(13):1740–7.
20. Salomon JA, Weinstein MC, Hammitt JK, Goldie SJ. Cost-effectiveness of treatment for chronic hepatitis C infection in an evolving patient population. *JAMA*. 2003;290(2):228–37.
21. Caro JJ, Briggs AH, Siebert U, Kuntz KM; ISPOR-SMDM Modeling Good Research Practices Task Force. Modeling good research practices—overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force—1. *Med Decis Making*. 2012;32(5): 667–677.
22. Stout NK, Knudsen AB, Kong CY, McMahon PM, Gazelle GS. Calibration methods used in cancer simulation models and suggested reporting guidelines. *Pharmacoeconomics*. 2009;27(7): 533–45.
23. Taylor DC, Pawar V, Kruzikas DT, Gilmore KE, Sanon M, Weinstein MC. Incorporating calibrated model parameters into sensitivity analyses: deterministic and probabilistic approaches. *Pharmacoeconomics*. 2012;30(2):119–26.
24. Vanni T, Karnon J, Madan J, et al. Calibrating models in economic evaluation: a seven-step approach. *Pharmacoeconomics*. 2011;29(1):35–49.
25. Box MJ. A new method of constrained optimization and a comparison with other methods. *The Computer Journal*. 1965;8(1): 42–52.
26. Fletcher R. *Practical Methods of Optimization*. London: John Wiley and Sons; 2001.
27. Guin JA. Modification of the complex method of constrained optimization. *The Computer Journal*. 1968;10(4):416–7.
28. Nelder JA, Mead R. A simplex method for function minimization. *The Computer Journal*. 1965;7(4):308–13.
29. Enns EA, Pershing S, Wang Y, Goldhaber-Fiebert JD. Calibration methods for inferring transition probabilities from cross-sectional studies. Working Paper. Stanford (CA): Stanford University; 2013.
30. Suen S, Bendavid E, Goldhaber-Fiebert JD. India's changing multidrug-resistant tuberculosis epidemic and its implications for disease control. Working Paper. Stanford (CA): Stanford University; 2013.
31. Sauvaget C, Ramadas K, Thomas G, Vinoda J, Thara S, Sankaranarayanan R. Body mass index, weight change and mortality risk in a prospective study in India. *Int J Epidemiol*. 2008;37(5): 990–1004.
32. García-Rodríguez O, Secades-Villa R, Flórez-Salamanca L, Okuda M, Liu SM, Blanco C. Probability and predictors of relapse

to smoking: results of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). *Drug Alcohol Depend.* 2013;132(3):479–85.

33. Li L, Feng G, Jiang Y, Yong HH, Borland R, Fong GT. Prospective predictors of quitting behaviours among adult smokers in six cities in China: findings from the International Tobacco Control (ITC) China Survey. *Addiction.* 2011;106(7):1335–45.

34. Vangeli E, Stapleton J, Smit ES, Borland R, West R. Predictors of attempts to stop smoking and their success in adult general population samples: a systematic review. *Addiction.* 2011;106(12):2110–21.

35. World Health Organization. GATS (Global Adult Tobacco Survey), India. 2010. Available from: <http://www.cdc.gov/tobacco/global/>

36. World Health Organization. GATS (Global Adult Tobacco Survey), India: core questionnaire with optional questions. 2010. Available from: [http://www.who.int/tobacco/surveillance/en/tfi\\_gats\\_corequestionnairewithoptionalquestions\\_v2\\_FINAL\\_03\\_Nov2010.pdf](http://www.who.int/tobacco/surveillance/en/tfi_gats_corequestionnairewithoptionalquestions_v2_FINAL_03_Nov2010.pdf)

37. Cox NJ. PCHIPOLATE: Stata module for piecewise cubic Hermite interpolation. 2013. Available from: <http://econpapers.repec.org/software/bocbocode/s457561.htm>

38. International Institute for Population Sciences (IIPS) and Macro International. National Family Health Survey (NFHS-3), 2005–2006, India: key findings. 2007. Available from: <http://www.measuredhs.com/pubs/pdf/SR128/SR128.pdf>

39. World Health Organization. World Health Survey. 2013. Available from: <http://www.who.int/healthinfo/survey/en/index.html>

40. Government of India, Ministry of Home Affairs. Sample Registration System. 2011. Available from: <http://censusindia.gov.in/2011-Common/srs.html>

41. Jha P, Jacob B, Gajalakshmi V, et al. A nationally representative case-control study of smoking and death in India. *N Engl J Med.* 2008;358(11):1137–47.

42. Peto R. Influence of dose and duration of smoking on lung cancer rates. *IARC Sci Publ.* 1986;(74):23–33.

43. Peto R, Lopez AD, Boreham J, Thun M, Heath CJ. Mortality from tobacco in developed countries: indirect estimation from national vital statistics. *Lancet.* 1992;339(8804):1268–78.

44. Ezzati M, Lopez AD. Estimates of global mortality attributable to smoking in 2000. *Lancet.* 2003;362(9387):847–52.

45. Pirie K, Peto R, Reeves G, Green J, Beral V; Million Women Study Collaborators. The 21st century hazards of smoking and benefits of stopping: a prospective study of one million women in the UK. *Lancet.* 2013;381(9861):133–41.

46. Jha P, Ramasundarahettige C, Landsman V, et al. 21st-century hazards of smoking and benefits of cessation in the United States. *N Engl J Med.* 2013;368(4):341–50.

47. Goldhaber-Fiebert JD, Stout NK, Salomon JA, Kuntz KM, Goldie SJ. Cost-effectiveness of cervical cancer screening with human papillomavirus DNA testing and HPV-16,18 vaccination. *J Natl Cancer Inst.* 2008;100(5):308–20.

48. Press WH. *Numerical Recipes in C++ : The Art of Scientific Computing*. 2nd ed. Cambridge: Cambridge University Press; 2002.

49. Galassi M, Davies J, Theiler J, et al. GNU Scientific Library Reference Manual: multidimensional minimization. 2003. Available from: [http://linux.math.tifr.res.in/programming-doc/gsl/gsl-ref\\_34.html](http://linux.math.tifr.res.in/programming-doc/gsl/gsl-ref_34.html)

50. Galassi M, Davies J, Theiler J, et al. GNU Scientific Library Reference Manual: random number generator algorithms. 2003. Available from: [http://www.gnu.org/software/gsl/manual/html\\_node/Random-number-generator-algorithms.html](http://www.gnu.org/software/gsl/manual/html_node/Random-number-generator-algorithms.html)

51. Christakis NA, Fowler JH. The collective dynamics of smoking in a large social network. *N Engl J Med.* 2008;358(21):2249–58.

52. Christakis NA, Fowler JH. The spread of obesity in a large social network over 32 years. *N Engl J Med.* 2007;357(4):370–9.

53. Hill AL, Rand DG, Nowak MA, Christakis NA. Infectious disease modeling of social contagion in networks. *PLoS Comput Biol.* 2010;6(11):e1000968.

54. Rowe DC, Chassin L, Presson CC, Edwards D, Sherman SJ. An “epidemic” model of adolescent cigarette smoking. *J Appl Social Psych.* 1992;22(4):261–85.

55. Blok DJ, Van Empelen P, Van Lenthe FJ, Richardus JH, De Vlas SJ. Unhealthy behaviour is contagious: an invitation to exploit models for infectious diseases. *Epidemiol Infect.* Epub 2012 May 17.

56. Ejima A, Aihara K, Nishiura H. Modeling the obesity epidemic: social contagion and its implications for control. *Theor Biol Med Model.* 2013;10:17.

57. Berry DA, Inoue L, Shen Y, et al. Modeling the impact of treatment and screening on U.S. breast cancer mortality: a Bayesian approach. *J Natl Cancer Inst Monogr.* 2006;(36):30–6.