Effects of Categorizing Continuous Variables in Decision-Analytic Models

Tanya G. K. Bentley, Milton C. Weinstein and Karen M. Kuntz

Med Decis Making 2009 29: 549 originally published online 13 July 2009
DOI: 10.1177/0272989X09340238

Effects of Categorizing Continuous Variables in Decision-Analytic Models

Tanya G. K. Bentley, PhD, Milton C. Weinstein, PhD, Karen M. Kuntz, ScD

Purpose. *When using continuous predictor variables in discrete-state Markov modeling, it is necessary to create categories of risk and assume homogeneous disease risk within categories, which may bias model outcomes. This analysis assessed the tradeoffs between model bias and complexity and/or data limitations when categorizing continuous risk factors in Markov models.* ***Methods***. *The authors developed a generic Markov cohort model of disease, defining bias as the percentage change in life expectancy gain from a hypothetical intervention when using 2 to 15 risk factor categories as compared with modeling the risk factor as a continuous variable. They evaluated the magnitude and sign of bias as a function of disease incidence, disease-specific mortality, and relative difference in risk among categories.* ***Results***. *Bias was positive in the base case, indicating that categorization overestimated life expectancy gains. The bias approached zero as the number of risk factor categories increased and did not exceed 4% for any parameter combinations or numbers of categories considered. For any given disease-specific mortality and disease incidence, bias increased with relative risk of disease. For any given relative risk, the relationship between bias and parameters such as disease-specific mortality or disease incidence was not always monotonic.* ***Conclusions***. *Under the assumption of a normally distributed risk factor and reasonable assumption regarding disease risk and moderate values for the relative risk of disease given risk factor category, categorizing continuously valued risk factors in Markov models is associated with less than 4% absolute bias when at least 2 categories are used.* ***Key words:*** *Markov models; Monte Carlo models; bias. (****Med Decis Making 2009;29:549–556)***

Markov models, or state transition models, can be used to model the risk of disease or event over time for a hypothetical population. Often, this risk depends on certain predictor variables, many of which are measured on a continuous scale. For example, blood pressure is a risk factor for coronary heart disease, folate intake can affect the risk of neural tube defects (NTDs), and CD4 counts can predict the risk for opportunistic infections in people living with HIV or AIDS. When using such continuous predictor variables in a discrete-state Markov model, it is necessary to create categories of risk. In so doing, one makes assumptions about the homogeneity of persons residing within each categorical health state, such that all persons face the same risk of developing the disease or event of interest. To the extent that the relationship between the predictor variable and the disease outcome can be defined as a continuous relationship, the assumption of homogeneity within risk category will result in biased model output.

Kuntz and Goldie[1] identified another bias related to the assumption of homogeneity in Markov models. They looked at the impact of an unobservable dichotomous factor that may affect disease risk and demonstrated that failure to incorporate these differential disease risks would bias model outcomes because the model does not account for the fact that higher risk individuals get disease more rapidly than do those with lower risk. The magnitude of this ''heterogeneity bias'' was found to depend on disease risk and the

Received 4 December 2007 from the Faculty of Arts and Sciences, Harvard University, Cambridge, Massachusetts (TGKB) and the Department of Health Policy and Management, Harvard School of Public Health, Boston, Massachusetts (MCW, KMK). This paper was presented at the 27th annual meeting of the Society for Medical Decision Making, 21–24 October 2005, San Francisco, California. Dr. Bentley was a predoctoral fellow in the Dana-Farber/Harvard Cancer Center Program in Cancer Outcomes Research Training. This research was supported by the National Cancer Institute (grant U01 CA88204). Revision accepted for publication 28 May 2009.

Address correspondence to Karen M. Kuntz, ScD, Division of Health Policy and Management, School of Public Health, University of Minnesota, Mayo Mail Code 729, 420 Delaware Street SE, Minneapolis, MN 55455; phone: (612) 625-9333; fax: (612) 624-2196; e-mail: kmkuntz@umn.edu.

DOI: 10.1177/0272989X09340238

relative risk of disease and could be as large as 50%.[1] Similar conclusions were drawn in Zaric's analytic perspective of the issue.[2]

A bias can likewise occur when categorizing continuous predictor variables in Markov models. Because individuals at the lower ends of each category face a different disease risk than those at the higher ends, they should be exiting the disease-free Markov states at a faster or slower pace than those at the higher ends of each category. The relative proportion of higher risk individuals within each category who consequently remain in the disease-free states should decrease over time, whereas that of lower risk individuals should increase. Because categorization does not specifically adjust for these changes within groups over time, the model is unable to account for this differential disease risk and will produce biased results.

Categorizing continuous variables therefore involves tradeoffs between the effects of this bias and model complexity and/or data limitations. For example, if ''too few'' categories are used, there is potential for biased results, yet ''too many'' categories may increase model complexity unnecessarily or increase uncertainty in model estimate due to sparse data. A gold-standard model is one that can incorporate the risk factor as a continuous variable such as with Monte Carlo simulation.[3] However, Monte Carlo models can be difficult to debug, entail long computing time, and preclude probabilistic sensitivity analysis.[3] To examine the tradeoffs between the magnitude of bias v. model complexity and data limitations, we used simple decision-analytic models in which we considered the potential influence of categorizing continuous predictor variables on model outcomes, labeling the effect *categorization bias*. We compared outcomes in the categorized model with those in the ''gold standard'' of the Monte Carlo and quantified the degree and direction of bias as a function of key disease outcome parameters.

METHODS

We developed generic Markov cohort and Monte Carlo simulation models to evaluate the tradeoffs between choosing fewer v. more categories of a continuously valued risk factor. In the cohort model, subjects were distributed into disease-free states defined by a categorized risk factor. From each disease-free state, we assigned a unique probability of developing disease, based on the mean value with respect to the distribution of the risk factor within the category.

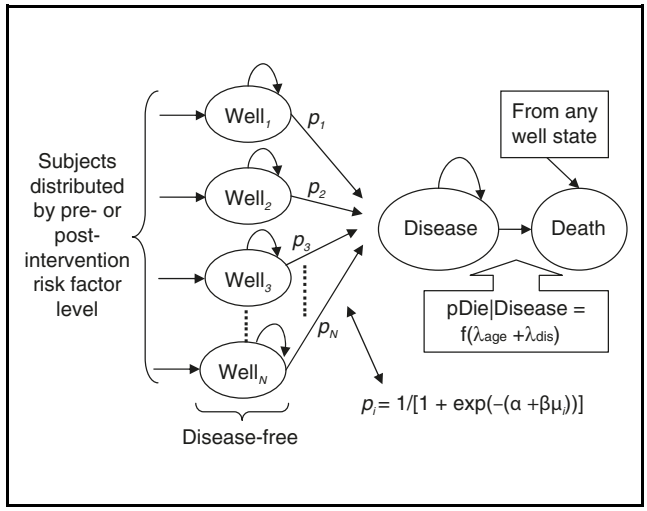

Figure 1 Schematic diagram of the generic Markov cohort model, where μ_i is the risk factor level associated with the mean disease risk for each state, λ_{age} is the age-specific non-disease-related mortality rate, and λ_{dis} is the disease-specific mortality rate.

Initially, we assumed that the disease probability was a logistic function of the risk factor. We chose a logistic function because it is bounded by 0 and 1 and is the basis of risk functions fit via logistic regression analysis. We also considered an assumption of linear risk. Persons could die either from disease or from other causes, and the outcome of interest was life expectancy with v. without a hypothetical intervention. We ran the model while varying the number of risk factor categories from 2 to 15. The Monte Carlo simulation model used the same underlying model parameters and assumptions while modeling the risk factor as a continuous variable.

Generic Model

Figure 1 shows the structure of the generic Markov cohort model, which begins with a cohort of disease-free persons distributed into one of N ''Well$_i$'' states, where $i = 1$ to N and $N = 2, 3, 4, \ldots, 15$. Each state is defined by a categorized, generic risk factor. Individuals in category i face an annual risk of disease (in the absence of intervention) based on our assumption of the following logistic risk function:

$$\Pr(\text{Disease}|\text{category } i, \text{ no intervention}) = \frac{1}{1+\exp(-(\alpha+\beta\mu_i))}, \quad (1)$$

where α is a measure of overall disease risk, β characterizes the relative difference in risk between risk factor values, and μ_i is the risk factor level associated with the mean disease risk for category i. We also varied this assumption in sensitivity analyses to evaluate the following linear risk function:

$$\begin{aligned}\Pr(\text{Disease}|\text{category } i,\\ \text{no intervention}) &= 0, \text{ if } \alpha+\beta\mu_i<0\\ &= 1, \text{ if } \alpha+\beta\mu_i>1\\ &= \alpha+\beta\mu_i, \text{ otherwise.}\end{aligned} \quad (2)$$

We assumed a protective risk factor, such that β is negative and higher risk factor values imply lower disease risk. The primary difference among states is disease risk, which we assumed to be homogeneous within categories. Although the probability of developing disease depends on the risk factor category, the overall disease risk for the initial cohort is the same for all values of N and is equal to the weighted average of the category-specific disease risks, weighted by the proportions of the cohort starting in each category.

Both the Markov cohort and the Monte Carlo simulation models considered a population of 30-year-old adults, for whom age-specific risk of all-cause mortality was based on data from the US Vital Statistics.[4] We modeled the probability of dying from disease as a function of the age-specific non-disease-related mortality rate (λ_{age}) and the disease-specific mortality rate (λ_{dis}) using the following function:

$$\Pr(\text{Die}) = 1-\exp(-(\lambda_{age}+\lambda_{dis})).$$

The goal of our analysis was to quantify the difference in cohort model results for varying values of N compared to those from the Monte Carlo model, where the risk factor was modeled as a continuous variable. All cohort models were run using Treeage Pro (Williamstown, MA) software, and the Monte Carlo model used C programming.

We assumed that the risk factor values were distributed normally with a mean of 100 and a standard deviation of 10 and that an individual's risk factor level did not change with age. The choice of parameter values for the normal distribution was fairly arbitrary but does not influence our analysis because one can transform this distribution into any other normal distribution with a different mean and variance. In the categorized Markov cohort model, we defined the ranges for the N categories of risk by assuming that each category had equal probability:

$$\text{Category } i = \left[\left(\frac{i-1}{N}\right)\%\text{ile}, \left(\frac{i}{N}\right)\%\text{ile}\right], \text{ for } i=1, 2, \ldots, N,$$

where the percentile (%ile) refers to the percentile of the risk factor distribution. Individuals were assigned a constant probability of disease based on their category's mean risk, calculated using the associated risk factor level μ_i and the α and β parameters in the risk function (equation (1) or (2)). When using a linear function of disease risk, this average probability of disease for each category is equal to the value of the risk function at the category's average risk factor level.

We estimated life expectancy gains associated with a hypothetical intervention that decreases the risk of disease and compared our results varying the number of risk factor categories. Bias was defined as the percentage change in life expectancy gain (LEGain) when using N categories ($N=2$–15) of risk factor to that when using the Monte Carlo (MC) gold standard:

$$Bias = \frac{\text{LEGain}_{\text{Categ}(N)} - \text{LEGain}_{\text{MC}}}{\text{LEGain}_{\text{MC}}}.$$

A positive bias would imply that fewer categories overestimated the life expectancy gains with intervention compared with more categories, whereas a negative bias would indicate that these gains were underestimated.

We assumed that the hypothetical intervention reduced disease risk by a percentage φ, a method that is consistent with those used in prior heterogeneity bias analyses.[1,2] To incorporate the intervention, the probability of disease for the logistic (base case) and linear risk functions was as follows:

Logistic:

$$\Pr(\text{Disease}|\text{category } i, \text{intervention}) = \frac{1}{1+\exp(-(\alpha+\beta\mu_i))}*(1-\varphi),$$

Linear:

$$\Pr(\text{Disease}|\text{category } i, \text{intervention}) = (\alpha+\beta\mu_i)*(1-\varphi),$$

where φ equaled 0 with no intervention and 10% with intervention.

We evaluated the predicted life expectancy benefits—and thus the magnitude and sign of bias—as a function of disease risk, disease-specific mortality, and relative difference in risk between 2 equal categories (relative risk [RR]). Table 1 shows our base case parameter values and assumptions as

Table 1 Model Parameters and Assumptions

Variable	Base Case Estimate	Range
Age	30	30–80
Disease risk function	Logistic	Linear
Underlying risk factor distribution	Normal [100, 10]	—
Annual disease risk without intervention, %[a]	3	1–12
Relative risk of disease[b]	2	1.2–2.5
Disease-specific mortality rate (λ_{dis})	0.03	0.01–0.10
Treatment effect (φ), %	10	1–50

a. Represents the average disease risk for a population with the specified risk factor distribution and is derived from risk function using values of α and β such that we get a specified disease risk and relative risk of disease.
b. Represents the ratio of high-risk v. low-risk categories in the 2-category model and is derived from the value of β by dividing the overall risk of the high-risk group by that of the low-risk group to achieve the desired relative risk value.

well as the ranges considered for sensitivity analyses. Both disease risk and RR were dependent on the underlying risk factor distribution (normal) and disease risk function (logistic or linear) and were thus derived by varying the risk function parameters α and β to obtain the desired disease risk and RR values. Note that the parameter values of the normal distribution do not influence the analysis because a change in these parameters would simply affect the values of α and β but would not alter the model estimates for the annual disease risk or RR. To provide a clearer interpretation of β, we present RR as the ratio of the probabilities of disease for the high- v. low-risk groups in the 2-category model with a specified set of risk function parameters. For example, if annual disease risk is equal to 3% and RR is equal to 2.0, then this would imply that the low-risk group has an annual disease risk of 2% and the high-risk group has an annual disease risk of 4%.

Worked Example

We evaluated the bias in a simple worked example of using an antihypertensive agent to lower diastolic blood pressure (DBP) and reduce the risk of coronary heart disease (CHD). We derived the DBP distribution from a US population of 65- to 74-year-old diabetic, smoking men, based on 1976–1980 data from the National Health and Nutrition Examination Surveys.[5] We fit a normal distribution with mean of 83 and standard deviation of 11.5. Annual CHD risk was predicted using published risk equations of a parametric statistical model from the Framingham Heart Study[6] for a cohort of 70-year-old diabetic, smoking men with average cholesterol levels. Calculation details for the prediction coefficients are described in Anderson and others.[6] CHD risks were based on the mean value with respect to the DBP distribution within each of 1, 2, and 10 categories. We evaluated life expectancy gains from a hypothetical intervention that reduced DBP by 15 points and used the 10-category model as a proxy for the gold standard.

RESULTS

Table 2 shows the results of our base case analysis. Given an average annual disease risk of 3%, RR of 2, annual disease-specific mortality of 3%, and an intervention that decreases risk by 10%, the model predicted that the intervention would increase life expectancy by just over 8 months. This benefit ranged from 8 months when the risk factor was modeled as a continuous variable in the Monte Carlo simulation to 8 months plus 1 week when only 2 categories were used. Comparing projected gains with the use of 2 to 15 risk factor categories with those when using a continuous risk factor, the categorization bias was 2.4% for the 2-category model and 0.05% for 15 categories. In the base case, the bias was always positive, approached 0 as N increased, and never exceeded 4% when the risk factor was categorized or 7.5% when pooled (e.g., 1 category).

We evaluated the degree and sign of bias as a function of average annual disease risk, RR between 2 categories, disease-specific mortality, and number of categories (Figure 2). For any levels of disease-specific mortality and disease risk, bias increased with RR. For example, in our base case of disease-specific mortality and disease risk both at 3%, the 2-category bias was 0.9%, 2.4%, and 3.9% for RRs of 1.5, 2, and 2.5, respectively. For any given RR, the relationship between bias and disease risk was not always monotonic. For example, for all combinations of RR and disease-specific mortality, the 2-category bias increased with disease risks of 1% and 3% and then decreased when risk increased further to 5%. Holding all other parameters constant, the magnitude of bias was not sensitive to disease-specific mortality.

Figure 3 shows bias given a linear function of disease risk, using the base case model parameters and assumptions and using the 15-category model as

Table 2 Life Expectancy Gain from Hypothetical Intervention and Percent Bias by Number of Risk Factor Categories

	Life Expectancy, y			
# Categories	No Intervention	Intervention	Months Gained	% Bias
1 (pooled)	40.56	41.07	8.61	7.50
2	40.66	41.35	8.20	2.42
3	40.74	41.41	8.10	1.25
4	40.77	41.44	8.07	0.77
5	40.79	41.46	8.05	0.52
6	40.80	41.47	8.04	0.38
7	40.81	41.48	8.03	0.28
8	40.81	41.48	8.02	0.22
9	40.82	41.49	8.02	0.17
10	40.82	41.49	8.02	0.14
11	40.82	41.49	8.01	0.11
12	40.83	41.49	8.01	0.09
13	40.83	41.49	8.01	0.07
14	40.83	41.50	8.01	0.06
15	40.83	41.50	8.01	0.05
MC	40.84	41.51	8.01	0

MC, Monte Carlo model result.

a proxy for the gold standard in defining bias (note: the Monte Carlo model is very computer time intensive). Given an average disease risk of 3%, an intervention that decreased risk by 10%, and an RR of 2, the linear model predicted the intervention to increase life expectancy by over a year. The bias was highest for 2 categories at 3.2% and increased with higher levels of disease risk. We also used the linear model with a 15-category model as a proxy for the gold standard to further explore the relationship between bias and disease risk, shown in Figure 4 for a constant RR of 1.2. When considering a wider range of disease risk values, we found that magnitude of bias did increase with disease risks of up to 3%, at which point bias decreased and eventually became increasingly negative at disease risks of greater than 8%.

When we varied other parameters using the logistic model, the range of categorization bias by number of categories did not change substantially. For treatment efficacy of 1% to 50% (from a base case value of 10%), the bias decreased with higher efficacy but changed minimally—still ranging between 2.2% and 0.02% (2–15 categories, respectively) from a base case range of 2.4% to 0.05%. For 30- to 80-year-olds, the bias decreased with age—and therefore with higher rates of all-cause mortality—ranging from 1% to 0.005% for 2 to 15 categories in people aged 80+ years.

In our worked example, for a US population of 65- to 74-year-old smoking, diabetic men (1976–1980), an average annual disease probability of 2.03%, an annual disease-specific mortality of 3%, and an intervention that decreases DBP by 15 points among patients with DBP ≥ 90, the model predicted that the intervention would increase life expectancy by approximately 2 weeks. Comparing projected gains with the use of 1 to 2 DBP categories with those when using a 10-category model, the categorization bias was 6.2% for the 1-category model and 3.8% for 2 categories. As with the results from our base case analysis, the bias was always positive, approached 0 as N increased, and was overall very small for any number of categories.

DISCUSSION

We evaluated the tradeoffs involved in categorizing continuous risk factors in decision-analytic modeling. Although one of the benefits of evaluating interventions with Markov cohort models is the ability to simulate disease risk over time, to do so, it is necessary to categorize model inputs and thus assume homogeneous risk within each category. Model outcomes may consequently be biased because of the changing distribution of risk across categories over time, as the higher risk individuals within each category progress to disease more rapidly than their lower risk counterparts. Knowing the degree to which such a bias might affect results for a particular disease and intervention can be important for modelers when making categorization decisions. This analysis provides a methodological contribution to that knowledge beyond that from previous research on heterogeneity bias[1,2] by exploring the impact that assumptions of homogeneity within risk groups may have on model outcomes when more than 2 groups are considered, when using a continuously valued risk factor as a gold standard for defining bias, and when bias challenges a priori expectations that it would be positive because of the heterogeneity effect.[1]

We evaluated the degree and direction of categorization bias in simplified and generic Markov cohort and Monte Carlo simulation models and as a function of overall annual disease risk, disease-specific mortality, risk factor effect, and number of risk factor categories. These parameters were considered important so that modelers can determine the impact of categorization decisions for varying degrees of

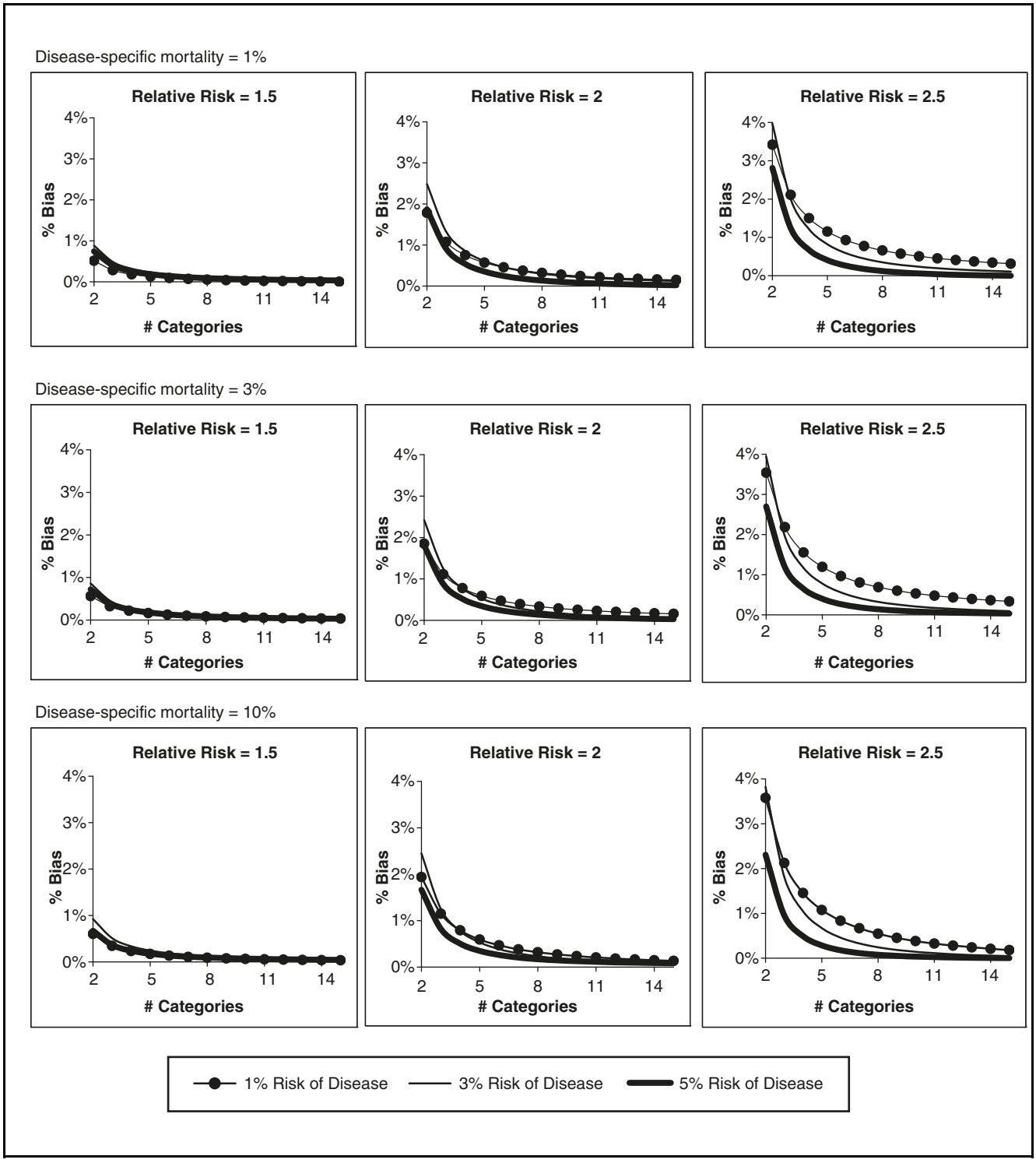

Figure 2 Categorization bias for varying disease risks (1%, 3%, 5%), relative risks between 2 categories (1.5, 2, 2.5), number of risk factor categories (2–15), and disease-specific mortality (1%, 3% 10%), using the model with the continuously valued risk factor as the gold standard.

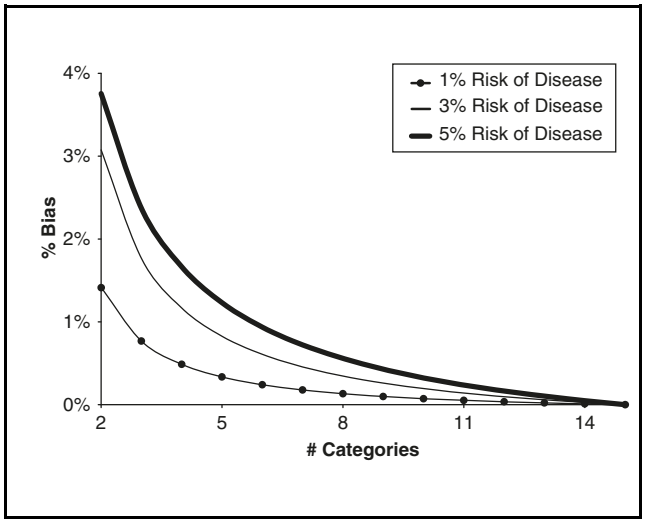

Figure 3 Categorization bias with a linear risk function for varying disease risk and number of risk factor categories, using the 15-category model as comparator, a disease-specific mortality of 3%, and relative risk between 2 categories of 2.

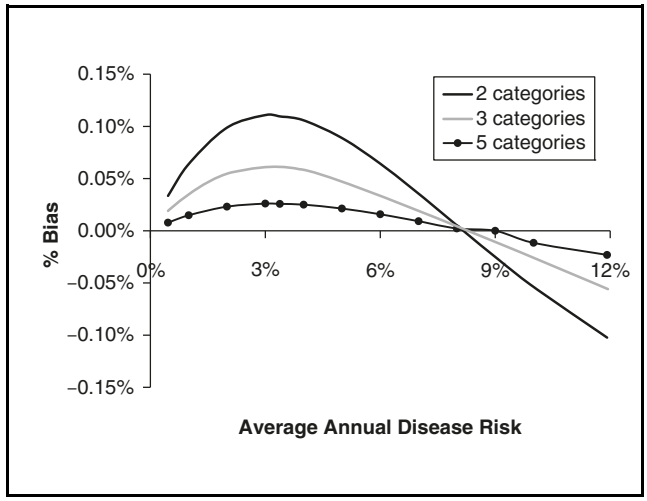

Figure 4 Categorization bias with the linear risk function using 2, 3, and 5 risk factor categories as compared with 15, for disease risks of 0.5% to 12%, a disease-specific mortality of 10%, and relative risk between 2 categories of 1.15.

disease prevalence and mortality, as well as for risk factors of substantial or negligible importance. Categorization bias was defined as the percentage change in life expectancy gain when using 2 to 15 categories of risk factor compared with that when using a Monte Carlo simulation that modeled the risk factor as a continuous variable. A positive bias therefore indicates that categorization overestimates the benefits gained from the intervention, whereas a negative bias indicates an underestimation. For average disease risk of 3%, disease-specific mortality of 0.03, and an intervention that reduces disease risk by 10%, the bias was only 3.9% even when the risk factor effect was large, and it never exceeded 4% for any parameter combinations evaluated. With our base case assumptions, the bias was always positive and consistently remained at less than 1.5% when at least 3 categories were used. Although these base case assumptions may be somewhat restrictive, the results of our worked example of reducing CHD risk through a blood pressure intervention indicate that even when using more general forms of such parameters, the bias still remained at less than 4% when at least 2 categories were used. Similarly, when we evaluated bias in a sensitivity analysis using a log-normal distribution, results were very similar to those in the base case, with bias being positive, approaching 0 with an increasing number of categories, and not exceeding 1% with the use of at least 3 categories. The magnitude of categorization bias is thus relatively small, given that previous research indicates a range of 50% to 75% for heterogeneity bias.[1]

We found that when disease risk is sufficiently high and relative risk is small, the categorization bias can be negative (Figure 4). In other words, using fewer categories can result in an *under*estimation of the life expectancy gains under these conditions. This is opposite to the expected effect induced by an overestimation of disease risk—and risk reduction due to intervention—when fewer categories are used and has been previously reported in the 2-category v. 1-category setting by Zaric.[2] The explanation for this finding is that the overestimation of disease risk (with fewer categories) results in an underestimation of life expectancy *with or without intervention*. The negative bias associated with underestimation of life expectancy (and life expectancy gain) competes with the positive bias associated with overestimating the effect of the intervention on disease risk and is most prominent when overall disease risk is high. With a very high risk of disease, most of the cohort gets the disease within a short period of time, and thus the disease prevention aspect of the intervention is outweighed by the reduction of life expectancy for the cohort with slightly higher disease risk (i.e., the cohort in the model with fewer categories). This effect also explains the nonsequential ordering of bias in relation to disease risk, as bias increases with risk up to a threshold, at which point the

negative influence begins to dominate the positive and causes bias to subsequently *decrease* with increased disease risk.

The results of this analysis must be considered in light of its limitations. To most clearly demonstrate the effects of categorizing continuous variables on model outcomes, we used generic disease prevention models and made simplifying assumptions about the relationship between risk factor and disease and about methods for modeling interventions. Although we evaluated results for both logistic and linear functions of disease risk, these may not be clinically accurate, as the nature of such dose-response relationships between risk factor and disease may be highly irregular, unpredictable, and/or nonparametric.

Similarly, our assumption of a normal underlying risk factor distribution may not accurately represent true risk factor levels in a population. For example, folate intake—a predictor variable for neural tube defect–affected pregnancies as well as possibly for colon cancer and heart disease[7-18]—has been shown to have a lognormal population distribution.[19,20] Modeling folate intake as a normal distribution could affect model outcomes beyond the effects of categorization or heterogeneity bias, such as by changing the proportions of the population estimated to be within disease risk categories. Our analysis also only considered risk factors that do not change with age—which may not be accurate in clinical situations involving such risk factors as body weight or diet—and that are categorized by assigning equal probability to each category.

Markov state transition models offer a valuable method of evaluating health outcomes and disease prevention interventions, yet their very nature requires the categorization of continuously valued predictor variables and the assumption of homogeneous risk within each health state. Although our analysis indicates that the magnitude of the consequential categorization bias is small and somewhat predictable, modelers should consider the potential for error when designing such models, and researchers and policy makers need to be cognizant of these issues in interpreting model results.

REFERENCES

1. Kuntz KM, Goldie SJ. Assessing the sensitivity of decision-analytic results to unobserved markers of risk: defining the effects of heterogeneity bias. Med Decis Making. 2002;22:218–27.

2. Zaric GS. The impact of ignoring population heterogeneity when markov models are used in cost-effectiveness analysis. Med Decis Making. 2003;23:379–96.

3. Gold MR, Siegel JE, Russell LB, Weinstein MC, eds. Cost-Effectiveness in Health and Medicine. New York: Oxford University Press; 1996.

4. Arias E. United States life tables, 2001. Natl Vital Stat Rep. 2004;53:1–38.

5. Drizd T, Dannenberg AL, Engel A. Blood pressure levels in persons 18–74 years of age in 1976-80, and trends in blood pressure from 1960 to 1980 in the United States. Vital Health Stat 11. 1986;(234):1–68.

6. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. Am Heart J. 1991;121:293–8.

7. Abby SL, Harris IM, Harris KM. Homocysteine and cardiovascular disease. J Am Board Fam Pract. 1998;11:391–8.

8. Giovannucci E, Rimm EB, Ascherio A, Stampfer MJ, Colditz GA, Willett WC. Alcohol, low-methionine: low-folate diets, and risk of colon cancer in men. J Natl Cancer Inst. 1995;87:265–73.

9. Giovannucci E, Stampfer MJ, Colditz GA, et al. Multivitamin use, folate, and colon cancer in women in the Nurses' Health Study. Ann Intern Med. 1998;129:517–24.

10. Glynn SA, Albanes D, Pietinen P, et al. Colorectal cancer and folate status: a nested case-control study among male smokers. Cancer Epidemiol Biomarkers Prev. 1996;5:487–94.

11. Freudenheim JL, Graham S, Marshall JR, Haughey BP, Cholewinski S, Wilkinson G. Folate intake and carcinogenesis of the colon and rectum. Int J Epidemiol. 1991;20:368–74.

12. Kato I, Dnistrian AM, Schwartz M, et al. Serum folate, homocysteine and colorectal cancer risk in women: a nested case-control study. Br J Cancer. 1999;79:1917–22.

13. Morris CD, Carson S. Routine vitamin supplementation to prevent cardiovascular disease: a summary of the evidence for the U.S. Preventive Services Task Force. Ann Intern Med. 2003;139:56–70.

14. Rimm EB, Willett WC, Hu FB, et al. Folate and vitamin B6 from diet and supplements in relation to risk of coronary heart disease among women. JAMA. 1998;279:359–64.

15. Su LJ, Arab L. Nutritional status of folate and colon cancer risk: evidence from NHANES I epidemiologic follow-up study. Ann Epidemiol. 2001;11:65–72.

16. Thun MJ, Calle EE, Namboodiri MM, et al. Risk factors for fatal colon cancer in a large prospective study. J Natl Cancer Inst. 1992;84:1491–500.

17. White E, Shannon JS, Patterson RE. Relationship between vitamin and calcium supplement use and colon cancer. Cancer Epidemiol Biomarkers Prev. 1997;6:769–74.

18. Willett WC. Diet, nutrition, and avoidable cancer. Environ Health Perspect. 1995;103:165–70.

19. Bentley TGK, Willett W, Weinstein M, Kuntz KM. Population-level changes in folate intake by age, gender, and race/ethnicity after folic acid fortification. Am J Public Health. 2006;96:2040–7.

20. Lewis CJ, Crane NT, Wilson DB, Yetley EA. Estimated folate intakes: data updated to reflect food fortification, increased bioavailability, and dietary supplement use. Am J Clin Nutr. 1999;70:198–207.